again focuses on London teenager usage. Callahan (2004) explores Spanish-English code switching using a corpus comprised of 30 fictional works from 24 Latino authors published in the United States, between 1970 and 2000. Callahan shows that written codeswitching follows for the most part the same syntactic patterns as its spoken counterpart. Her corpus findings also point to the use of non-standard English, which appears in 53% of the corpus in the forms of African-American Vernacular English and certain varieties of New York English. Lapidus and Otheguy (2005), in another New York corpus-based study, look at language contact in the context of English and Spanish. They focus on the use of non-specific *ellos* (English equivalent: *they*). One of Lapidus and Otheguy's main conclusions is that the susceptibility of language varieties to contact influence is primarily at the discourse-pragmatic level. Corpora have had a major influence in the areas of discourse and pragmatics also and throughout this book we will draw on examples of such work.

1.8 How have corpora influenced language teaching?

As we discussed above, the processes of dictionary-making have been revolutionised by the use of language corpora and this obviously feeds into language teaching materials. All major learners' dictionaries of English are now based on constantly updated multi-million word databases of language. Fundamentally, corpora have provided evidence for our intuitions about language and very often they have shown that these can be faulty when it comes to issues such as semantics and grammar. As we noted earlier, we now increasingly base our major grammars, like dictionaries, on large language corpora. The contribution of corpus linguistics, therefore, to the description of the language we teach is difficult to dispute. According to McCarthy (2001: 125) corpus linguistics represents cutting-edge change in terms of scientific techniques and methods and probably foreshadows even more profound technological shifts that will 'impinge upon our long-held notions of education, roles of teachers, the cultural context of the delivery of educational services and the mediation of theory and technique'.

As well as providing an empirical basis for checking our intuitions about language, corpora have also brought to light features about language which had eluded our intuition (e.g. the frequency of ready-assembled chunks; see chapter 3). In terms of what we actually teach, numerous studies have shown us that the language presented in textbooks is frequently still based on intuitions about how we use language, rather than actual evidence of use. While there are often sound pedagogical reasons for using scripted dialogues, their status as a vehicle for enhancing conversation skills has been challenged in recent years (Carter 1998; Burns 2001; Burns, Joyce and Gollin 2001; McCarthy and O'Keeffe 2004; Thornbury and Slade 2006). Burns (2001) notes that scripted dialogues rarely reflect the unpredictability and dynamism of conversation, or the features and structures of natural spoken discourse, and argues that students who encounter only scripted spoken language have less opportunity to extend their linguistic repertoires in ways that prepare them for unforeseeable interactions outside of the classroom. Holmes (1988: 40), for example, looked at epistemic modality in ESL textbooks as compared with corpus data and found that many textbooks devoted an

unjustifiably large amount of attention to modal verbs, at the expense of alternative linguistic strategies. Boxer and Pickering (1995) showed contrast between speech acts in textbook dialogues with real spontaneous encounters found in a corpus. Carter (1998) compares real data from the Cambridge and Nottingham Corpus of Discourse in English (CANCODE, see appendix 1) with dialogues from textbooks and finds that the dialogues lack core spoken language features such as discourse markers, vague language, ellipsis and hedges. Gilmore (2004) examines the discourse features of seven dialogues published in course books between 1981 and 1997, and contrasts them with comparable authentic interactions in a corpus. He finds that the textbook dialogues differ considerably from their naturally-occurring equivalents across a range of discourse features including turn length and patterns, lexical density, number of false starts and repetitions, pausing, frequency of terminal overlap or latching, and the use of hesitation devices and response tokens. He looks at dialogues from more recent course books and finds that there is evidence that they are beginning to incorporate more natural discourse features. The *Touchstone* series (McCarthy, McCarten and Sandiford 2005a and b, 2006a and b) is an attempt to show how course book dialogues, and even entire syllabi, can be informed by corpus data. In addition to the conventional four-skills syllabus strands of speaking, listening, reading and writing, the *Touchstone* authors provide a syllabus of conversational strategies, based on the most common words and phrases in the North American spoken segment of the CIC. The strategies recur throughout the four levels of the multi-skills programme and are graded. An example is given in figure 11, where the discourse marker *I mean* is exploited.

Figure 11: Extract from the *Touchstone* series (McCarthy, McCarten and Sandiford 2005a: 49)

2 Strategy plus *I mean*

You can use ***I mean*** to repeat your ideas or to say more about something.

Where do you go? I mean, do you go somewhere nice?

Do you know Fabio's? It's OK. I mean, the food's good, . . .

In conversation . . .

I mean is one of the top 15 expressions.

A Complete the questions or answers with your own ideas. Compare with a partner. Do you have any of the same ideas?

1. *A* Do you ever go out after class?
   B Well, not very often. I mean, I usually go straight home .
2. *A* How do you like the restaurants in your neighborhood?
   B They're not bad. I mean, they're _____ .
3. *A* Are you busy in the evening? I mean, do you _____ ?
   B Well, I take a lot of classes.
4. *A* What do you do in your free time?
   B Well, I don't have a lot of free time. I mean, _____ .

About you → **B** ***Pair work*** Ask and answer the questions. Give your own answers.

Kettemann (1995) highlights the mismatch between actual language use and the prescription often found in pedagogical grammars that reported speech involves the 'backshift rule' for tenses in the reported speech constructions (see also Baynham 1991, 1996; McCarthy 1998). Hughes and McCarthy (1998) look at the use of past perfect verb forms and find that, across a wide range of speakers in the CANCODE corpus, the past perfect has a broader and more complex function in spoken discourse than hitherto described. Corpus descriptions have also enhanced our understandings of units of fixed phrasing, collocation, and more extended language patterns (Sinclair 1991a, 2003a, 2004; Svartvik 1991; Aston 1995; McCarthy and Carter 2002; Biber et al. 2004; Schmitt 2004; Thornbury and Slade 2006). Throughout the chapters that follow, we will survey and build on relevant findings from corpus research and tease out the implications these have for language teaching.

Corpora of learner languages are a relatively recent, but very important development. Granger (2003), a forerunner in the area, defines a learner corpus as an electronic collection of authentic texts produced by foreign or second language learners. She notes that, in the early 1990s, publishers and academics started, independently but concurrently, to gather and analyse learner data. The International Corpus of Learner English (ICLE, see Granger 1993, 1994, 1996, 1998a; Granger et al. 2002), initiated around that time, currently contains over two million words of writing by learners of English from 19 different mother tongue backgrounds. The writing in the corpus (essays) has been contributed by advanced learners of English as a foreign language rather than as a second language and is made up of 19 distinct sub-corpora, each containing one language variety (English to French, English to German, English to Swedish, etc.). This corpus is error-coded, which allows for invaluable research into typical learner error patterns (see Dagneaux et al. 1996; De Cock et al. 1998). Findings from research into learner corpora can be addressed in materials design, including the development of Computer Assisted Language Learning (CALL) applications. For example, Altenberg and Granger (2001), looking at Swedish- and French-speaking learners, examine the use of high frequency verbs, and in particular use of the verb *make*. As well as looking at the role of transfer in the misuse of these verbs relative to native-speaker norms, they investigate whether learners tend to over- or underuse these verbs and whether high frequency verbs are error-prone or safe. They find that EFL learners, even at an advanced proficiency level, have great difficulty with high frequency verbs such as *make*. They suggest that concordance-based exercises (see Data-driven learning below) can help raise awareness of the complexity of high frequency verbs. Learner spoken data have also been collected, a notable example being the Louvain International Database of Spoken English Interlanguage (LINDSEI) set up in 1995 (see De Cock 1998, 2000). This provides spoken data for the analysis of the speech of second language learners (see also Granger et al. 2002). Numerous other studies have been conducted using learner corpora, including Granger (1996, 1997, 1998a, 1998b, 1998c, 1999, 2002, 2003, 2004), De Cock and Granger (2004), Meunier (2002a, 2002b), Gilquin (2003) and Cosme (2004).

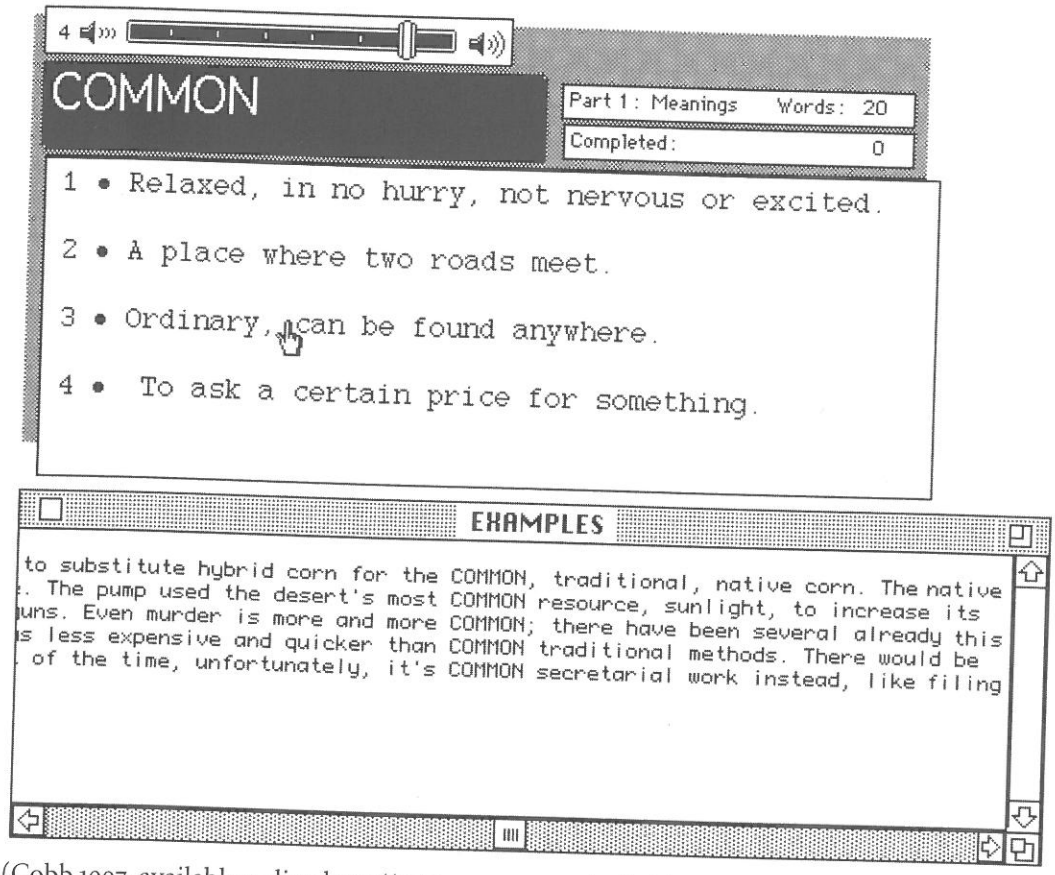
Data-driven learning

Computer Assisted Language Learning (CALL), among many other applications, includes the use of language corpora, where learners get hands-on experience of using a corpus through guided tasks or through materials based on corpus evidence, such as concordance lines on handouts (see Johns 1991a). Here an inductive approach relies on an 'ability to see patterning in the target language and to form generalisations' about language form and use (Johns 1991a: 2). This activity is commonly referred to as 'data-driven learning' (DDL) after Johns (1986 and 1991a). Johns (2002: 108) sees DDL as a process which 'confront(s) the learner as directly as possible with the data', 'to make the learner a linguistic researcher' where 'every student is Sherlock Holmes'. Over the years Johns, among others, has developed the idea and contributed many teaching materials based on the DDL approach (see Johns 1988, 2002; Stevens 1991; Wichmann 1995; Fox 1998; Kettemann 1995; Tribble and Jones 1990; 1997; Flowerdew 1993, 1996; Gavioli 1996; Wichmann et al. 1997; Tribble 2000, 2003; Aston 2001). A basic internet search will bring up numerous homepages dedicated to DDL, which provide many useful links to resources (such as online corpora and concordancers), research findings and materials. Such a search is also evidence of the popularity of DDL among language teachers, many of whom post their materials online and conduct action research into the classroom application of these materials. DDL, like corpus linguistics in general, is not without its critics (see Widdowson 1991, 2000; Prodromou 1996, 1997a, 1997b; Owen 1996; Seidlhofer 1999; Bernardi 2000; see below for further discussion of issues and debates). Many also question the application of DDL to lower-level learners, though some studies provide evidence of its use at lower levels (see Johns 1988, 2002; St John 2001; Kennedy and Miceli 2002).

Chambers, who has been involved in the development of a one-million word corpus of journalistic French (see appendix 1: Chambers-Rostand Corpus of Journalistic French; Chambers and Rostand 2005), provides a number of illustrations of how DDL can be used in the context of teaching French and how it can facilitate the development of learner autonomy (see Chambers and Kelly 2002, 2004; Chambers and O'Sullivan 2004; Chambers 2005; Braun and Chambers 2006; Chambers in press; O'Sullivan and Chambers in press). Chambers and Kelly (2002) note that the pedagogical context of DDL brings together constructivist theories of learning, the communicative approach to language teaching and developments within the area of learner autonomy. Cobb (1997) points to the potential of DDL to provide multiple contextual encounters for the acquisition of new vocabulary. The literature on vocabulary acquisition, according to Cobb, is virtually unanimous on the value of learning words through several contextual encounters (Mezynski 1983; Stahl and Fairbanks 1986; Krashen 1989; Nation 1990). Language learners are advised to read more (see Krashen 1989) so as to facilitate multi-contextual lexical acquisition. In reality, Cobb notes that few language learners have time to do enough reading for natural, multi-contextual lexical acquisition. DDL may have a role in rationalizing and shortening this learning process by providing a rich source of embodiments and contexts from new vocabulary. Empirical studies on the learning benefits of DDL are relatively few, but they do show positive results (see for example Cobb 1997; Turnbull and Burston 1998; Kennedy and Miceli 2001; Lenko-Szymanska 2002). Cobb (1997) reports on his longitudinal study of vocabulary

acquisition using concordance line tasks. This study provides interesting examples (with screen shots) of a variety of sequential DDL activities which draw on a specially designed corpus of 10,000 words (comprised of 20 texts of about 500 words each, assembled from the students' reading materials). Figure 12 shows the opening task:

Figure 12: Example of DDL task from Cobb (1997)
Part 1: Choosing a meaning. The learner is presented with a small concordance of four to seven lines, in KWIC format with the to-be-learned word at the centre, and uses this information to select a suitable short definition for the word from one correct and three randomly generated choices.



(Cobb 1997, available online http://www.er.uqam.ca/nobel/r21270/cv/Hands_on.html).

1.9 Issues and debates in the use of corpora in language teaching

Authenticity of materials for language teaching and learning

As we have seen, collecting data for use in a corpus means collecting examples of language as it is actually used in authentic contexts. Debate over the extent to which authentic

language should form the basis of language courses has been taking place for the last thirty years or so (Canale and Swain 1980; Breen 1983; Van Lier 1996; Rost 2002) but it has been re-energised by the availability of corpus data.

It is often argued that, in language teaching, examples drawn from corpus sources should form the basis for the material used to exemplify the language and that an aim of language teaching should be to produce learners who are able to communicate effectively and competently. In order for this to happen, it is argued further, learners need to experience authentic rather than contrived examples of data; by 'contrived' is meant examples of language that are specially made up or invented for the pedagogic purposes of illustrating a particular feature or rule of the language. One problem is that the terms 'contrived' and 'authentic' have become emotionally charged and in opposition to each other.

The availability of corpus examples has produced a different perspective since we can find in corpora numerous examples of texts that are free-standing, in so far as they are independent of any language learning task. They are in their own authentic context, and they are composed for a particular audience (which tends to be different to that of the language learner). Thus, when they are presented with corpus examples, learners encounter real language as it is actually used, and in this sense it is 'authentic'. However, the language has been wrenched from its original context, and so, in one sense, is 'decontextualised'. This position suggests that as soon as texts are extracted from the context in which they first appeared, are stored in large electronic databases, and are reproduced for the teaching context, they are effectively removed from an authentic environment. The learner, then, has to process such texts with reference to a different context than the one in which they originated, a context which may not reflect his or her communicative goals in the classroom context. Furthermore, one can argue that authentic texts are embedded in particular cultures and may thus be culturally opaque to those outside that (usually western) culture, and that it may, as a result, be next to impossible for learners to 'authenticate' such texts for themselves on this basis. Authenticity should therefore preferably be defined as a relationship between a text and the response that it triggers in its immediate audience (see for example Lee, 1995; Widdowson 1996, 1998). Consequently, there is among many a preference for contrivance and the deliberate use of culturally 'neutral' examples as a more solid basis for a pedagogy that is sensitive to learners' needs. Such contrived texts also allow for material to be more easily graded for learners at different levels of competence. Another non-corpus-based option is to use texts suggested or provided by the learners themselves, which will, by definition, be potentially maximally authentic.

Supporters of the view that there should be more authentic material available in classrooms argue, on the other hand, that naturally-occurring data can be carefully chosen and mediated, that it can be contextualised for the learner, that learners are no different from other human beings, who have a natural proclivity to contextualise language data for themselves, and that the use of such data in the classroom can actually facilitate discussion of cultural background, as well as provide more grounded motivation because the text is so obviously a 'real' example of the target language (Peacock 1997). To deprive learners of such experiences for ideological reasons without consulting them is,

in the opinion of the present authors, patronising and self-defeating. Others advance a related argument that tasks can be graded according to the nature of the authentic material (Willis and Willis 1996; Bygate et al. 2001; Willis 2003). The latter position would also seem to be an argument for a more careful pedagogic selection of materials from authentic sources. In our experience, corpora, both spoken and written, do indeed contain many texts that are obscure and culturally opaque, but they also contain numerous texts that are transparent, easily contextualised and interpretable by any mature human being. It is simply a matter of how carefully one selects the material, who the end-users are and what they want and expect from a language programme. For centuries, language teachers have plucked written texts out of the contexts in which they were originally produced and imported them into the classroom, carefully selecting and mediating them for their students; we see the use of corpora in this connection as an example of historical continuity which harnesses the technical possibilities of speeding up searches for useful and usable material. Many teachers are now using the world's biggest corpus, the internet, and its associated search engines, in just this way.

These issues are addressed in several places in this book. Our basic position is that for most pedagogic purposes in most contexts of teaching and learning a language, it is preferable to have naturally-occurring, corpus-based examples than contrived or unreal examples, but always in the context of freedom of choice and careful mediation by teachers and/or materials writers who know their own local contexts. For further reading on the debate that surrounds this see Sinclair (1991a, 1991b), Aston (1995), Carter and McCarthy (1995), Prodromou (1996), Owen (1996), Carter (1998), Cook (1998), Seidlhofer (1999), Widdowson (2000, 2001).

The 'native speaker' and the classroom

Authentic language invariably invokes the idea of language drawn from sources supplied by native speakers and recent research has shown that language learners often regard the approximation to native speaker English as a main goal in the language learning process (Timmis 2002). While the notion of the native speaker of English tends to be used to refer to those whose first language is English, the concept is a complex one (Roberts 2005), as there are, as Rampton (1990) and others have demonstrated, non-native speakers who have great affiliation to a language and are more competent in that language than native speakers. The vast number of different varieties of 'native speaker' English (e.g. American, British, Irish, Australian, South African, Singaporean) means that this notion cannot easily be translated, or modelled, into one particular standard for the language classroom, although international publishers tend to focus on either American or British English as a model.

Whether we are referring to contrived, invented or naturally-occurring samples of English, the choice of a particular variety for the ELT context, even down to fine-grained choices of a particular regional or local variety, is inevitably to some degree a matter of ideology and invariably a political issue. At the same time, it is acknowledged that the proportion of English exchanged daily between non-native speakers is growing rapidly, with an overall increase in globalisation and internationalisation (see Crystal 1997) to the point

where non-native users of English far outnumber native speakers of English (Graddol 1998), undermining, for some, any privileging of native speaker discourse.

At the same time this raises the further question whether native-speaker models are the most appropriate basis for language learners, who may predominantly use their L2 to operate in an international, rather than a 'native' context. This state of affairs has led some to propose that English as a Lingua Franca (ELF) is more significant internationally than English as a first or second language and that consequently, corpora of non-native Englishes are needed in order to help us identify the kinds of English crucial to communication in such ELF contexts (see below) and to use such evidence as a preferred basis for classroom teaching and learning (see Medgyes, 1994; Braine 1999; Oda 1999, 2000; Jenkins 2000; Tajino and Tajino 2000; Seidlhofer 2001a; Carter and Fung (forthcoming) for further discussion on native versus non-native speaking teachers).

ELF: English as a lingua franca

Seidlhofer (2001a: 143–4) notes that while learner corpora (see above) have their use as a 'sophisticated tool for analysing learner language . . . some of the data in the learner corpora could also contribute to a better understanding of English as a lingua franca'. Seidlhofer goes on to detail a corpus development which she has championed: The Vienna-Oxford International Corpus of English (VOICE), a collection English as a Lingua Franca (ELF) currently under construction. Here lingua franca is defined as an additionally acquired language system that serves as a means of communication for speakers from different speech communities, who use it to communicate with each other but for whom it is not their native language. It is 'a language which has no native speakers' (Seidlhofer 2001a: 146) (see also Malmkjær 1991; House 1999, 2002, 2003; James 2000). The initial target for the VOICE corpus is to collect around half a million words of spoken data from speakers whose first language is not English and whose primary and secondary education did not take place in English, but who make use of English as a lingua franca (ELF) (see Seidlhofer 2004). In a parallel development, Mauranen (2003) reports on a corpus of ELF in academic settings (EFLA) at the Tampere Technology University, Finland. Its initial target is to collect half a million words of spoken data from two university settings. Both Seidlhofer and Mauranen aim, through empirical investigations of ELF, to show that a sophisticated and versatile form of language can develop which is *not* a native language (Seidlhofer 2001b; Mauranen 2003). Seidlhofer (2001a) argues that this is a much-needed development to fill the conceptual gap between the growing recognition and meta-linguistic discussions about global English and the existence of a codified form which eventually might have pedagogical applications in the identification of the most efficient forms of communication in the domain of ELF. With this in mind, the corpus may establish 'something like an index of communicative redundancy' (Seidlhofer 2001a: 147). Early findings from the VOICE corpus (see Seidlhofer 2004) tentatively identify a number of features which point to systematic lexico-grammatical differences between native-speaker English and ELF, for example dropping the third person present tense 's' (e.g. *she look*), omitting definite and indefinite articles, insertion of prepositions (e.g. *can we discuss about this issue*). These features often

involve typical errors which most English teachers would correct and remediate. However, Seidlhofer points out that they appear to be generally unproblematic and do not cause an obstacle to communicative success in ELF. The work of Jenkins (1996, 2000, 2004, 2005) has also been very influential here in relation to the teaching of pronunciation for ELF. She makes a parallel argument relating to ELF phonology. Her research finds that a number of items common to most native-speaker varieties of English were not necessary in successful ELF interactions; for example, the absence of weak forms in words like *from* and *for*; and the substitution of voiceless and voiced *th* with /t/ or /s/ and /d/ or /z/ (e.g. *think* became *sink* or *tink*, and *this* became *dis* or *zis*). Jenkins argues that such features occur regularly in ELF interactions and do not cause intelligibility problems.

Developments in and findings from corpus-based ELF studies further the debate about 'ownership' and function of a language like English and their empirical findings put forward ELF as a pedagogical model which challenges the accepted native-speaker-based norms of EFL. However, great uncertainties remain in this area, not least whether the object of description is a *function* of English rather than a codifiable variety, that is to say a way in which people adapt differently to every different circumstance and make greater or lesser use of their communicative repertoire depending on the exigencies of each individual interaction. Mauranen (2003) confidently labels ELF as a variety, but much discussion is still needed as to what, exactly is meant by 'variety' here. Other problems arise in the (perhaps unfair) equation between a reduced or 'stripped down' ELF syllabus and an impoverished experience of the L2. Indeed, it could be argued that learners of any language always end up producing less than the input they are exposed to, and that if that input itself is deliberately restricted, then even less will be the outcome, and so on. Lastly, the evidence so far as to what exactly ELF is is rather scant, and there is reason to believe that East Asian ELF, for example (e.g. a Chinese speaker interacting in English with a Korean speaker) may be very different from European ELF (e.g. a Danish speaker using English with a Dutch speaker) and we may need to describe many 'ELFs' to get anywhere near an accurate picture of the global uses of English. What the present authors do support, however, is the way native-speaker corpora of spoken language, with all their attendant shortcomings, have sparked a lively if sometimes heated debate as to the most suitable models of English for pedagogy. This is a step forward from the days when southern-England, middle-class English was unquestioned as the pedagogical model in most parts of the world (the situation which pertained when two of the present authors began their teaching careers). We also support the move to build more and yet more useful corpora from a wider range of different settings.

SUEs or Successful Users of English

Rather than continuing to focus solely on the native speaker, we should begin to look much more closely at the notion of the 'expert user' and at ideas advanced by Prodromou and others (Prodromou 2003a, 2005) concerning what he terms SUEs (or Successful Users of English). As we discuss in chapter 4, Prodromou (2005) takes idiomaticity as a paradoxical example of something which, for native speakers, makes life easy, enabling fluent production

of deeply culturally-embedded chunks heard and rehearsed since childhood. These same idiomatic chunks seem to place impossible obstacles in the path of the non-native speaker, however proficient. SUEs are highly successful L2 communicators, but they will achieve this goal by strategic use of their resources in ways different from those of native speakers. It makes more sense, therefore, not to see SUEs as failed native speakers, but to look upon all successful users of a language, whether native- or non-native-speaking, as 'expert users'.

A spoken corpus can underline for us how important it is to look closely at what speakers and listeners do, whoever those speakers are, whether they are native or non-native. Such research shows that our ability to interact with others is an important part of what makes us successful users of the language and is, we believe (and this is confirmed by research that is reported throughout this book), what learners of English aspire to know about and do in and with a language, and for the very reason that they know that this is what they do successfully in their first language. We will never meet those needs just by introspecting on what we *think* we say, nor by feeding our learners an impoverished diet of what we think they need based on those intuitions; only by respecting learners' and teachers' choices and aspirations within their own local contexts will we best serve them.

When we do look at what speakers and listeners do, we may not hear native speakers as we might want to hear them or as how we might have learned to expect to hear them. But we do hear real people interacting with one another, working at full stretch with the language, adjusting millisecond by millisecond to the interactive context they are in, playing with the language, being creative, being affective, being interpersonal and, above all, expressing themselves as they engage with the processes of communication which are most central to our lives. It is hard to imagine any learner of a second language not wanting to be a good, human communicator in that second language, whether they are going to use it with native speakers or with any other human beings. Language teaching can only benefit from even closer inspection of such fundamentally human processes. And the road from corpus to pedagogy, upon which we take tentative, sometimes faltering steps in this book, is an essential part of that process.