Citation: Caines, A., McCarthy, M.J. and **O'Keeffe, A.** (2016) "Spoken language corpora and pedagogic applications". In F. Farr and L. Murray (Eds) *Routledge Handbook of Language Learning and Technology.* London: Routledge, pp. 348 - 361.

Chapter: Spoken language corpora and pedagogic applications

Authors: Andrew Caines, Michael McCarthy & Anne O'Keeffe

In comparison with written corpora, spoken corpora have not developed at the same rate. The reasons for this are largely to do with the huge costs and time involved in compilation and transcription, as well as access to recordable data. What has developed over the last 20 years, however, is an acknowledgement of the importance of spoken corpora in creating a fuller understanding of everyday spoken language, especially casual conversation. Whereas spoken corpora were initially small appendages to much larger written corpora, they are now increasingly valued and created in their own right. Two broad types of spoken corpora are of relevance to language pedagogy: large, demographically sampled corpora which attempt to grab a snapshot of a language as a whole (e.g. for contemporary British English the British National Corpus, hereafter BNC) and carefully targeted corpora aimed at collecting data for more specialised purposes such as spoken business language, spoken academic language, teenage language, spoken language in the broadcast media, etc. These latter corpora are often smaller, yet nonetheless yield invaluable insights into particular kinds of speaking. Sizeable research output has accrued over the years into the two types of spoken corpora and this has enhanced our understanding of the differences between spoken and written language in general, as well as offering insights into variation on a number of parameters. However, the pedagogical potential of these research findings has not always been fully exploited. This chapter reviews key findings from research into spoken corpora and current pedagogical applications and discusses how spoken-corpus-informed pedagogy might be expanded and brought further into the domains of conventional classrooms and blended and online learning.

1) SPOKEN CORPORA: WHAT ARE THEY?

Defining spoken corpora

Spoken corpora are collections of recordings of speaking which have been transcribed to form a database or corpus. A distinction is generally made between spoken corpora and '*speech corpora',* which are usually collections of speech (such as recordings of people reading out loud) that are compiled for purposes such as the analysis of the phonetic substance of speaking or the creation of voice-to-text applications and telephone technology (Harrington, 2010). The growth in the number of spoken corpora can be seen as having gone in tandem with the emergence and development of recording technology (see Murphy and Riordan this volume).

The evolution of large spoken corpora

McCarthy and O'Keeffe (2008) note that many early spoken corpora were developed as add-ons to much larger written corpora. This is a function of the time and expense involved in collecting spoken data relative to written texts. For example, the BNC (Crowdy 1993) contains over 100 million words of data, with the spoken component accounting for only ten percent of this. The 10 million words of spoken data comprise informal conversations recorded by volunteers selected from different ages, regions and social classes in a demographically balanced way. Also important in the evolution of spoken corpora is the ICE (International Corpus of English) project, designed 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).

Other notable large-scale spoken corpora that were developed internationally include the five-million word Longman Spoken American Corpus (see Chafe, Du Bois and Thompson 1991). By the turn of the millennium, the American National Corpus (ANC) was set up as a comparative corpus to the BNC (Ide and Macleod 2001). It is available as an online resource comprising a total of over 14.5 million words, 3.2

million of which are spoken data, (see http://www.anc.org/data/oanc/contents/).

The largest available online corpus of one variety of English now available is the 450-million word Corpus of Contemporary American English (COCA), which includes 85 million words of spoken data, including unscripted conversation from nearly 150 different TV and radio programs (Davies 2010). Despite the availability of substantial amounts of spoken American English data, there is a dearth of spontaneous face-to-face conversation. An exception to this is the Santa Barbara Corpus of Spoken American English (SBCSAE), a collection of approximately 249,000 words of recordings of natural speech, representing a wide variety of speakers (Du Bois et al 2003).

English spoken corpora still tend to dominate but spoken corpora for many other languages now exist, including Bulgarian, French (both European and Canadian), Mandarin Chinese, Vietnamese, Egyptian Arabic, Farsi, German, Greek, amongst others. Many of these are available from the Linguistic Data Consortium at the University of Pennsylvania (see www.ldc.upenn.edu). ELDA, the Evaluations and Language resources Distribution Agency in Europe also makes available a number of spoken corpus resources in different languages (see www.elda.org).

The evolution of smaller spoken corpora

Apart from the large-scale corpora of particular languages, there has also been a parallel growth in the development of smaller, specialised or domain-specific corpora. These are often designed to meet a particular research need where a research question focuses on one particular context of use. Specialised corpora are usually quite small (around or less than one million words). Some of the major developments in specialised spoken corpora have taken place in the domain of academic discourse and include, for example, the Michigan Corpus of Academic Spoken English (MICASE), and its British counterpart, the British Academic Spoken English corpus (BASE). A sub-category of these are learner corpora. While most learner corpora consist of written texts, some spoken learner corpora exist, for instance the Louvain International Database of Spoken English Interlanguage (LINDSEI) (Gilquin *et al*, 2010, see also Muenier this volume). Other interesting developments have been

corpora of expert users of English such as the Vienna-Oxford International Corpus of English (VOICE) (Breiteneder *et al*, 2006) and the English as a Lingua Franca in Academic Settings (ELFA) Corpus (Mauranen, 2003)

Research into small, specialised spoken corpora has been particularly fruitful in the area of pragmatics (see O'Keeffe *et al* 2011). Small domain-specific corpora allow for concentrated patterns of use to emerge, particularly those features which have become pragmatically specialised. Some examples of domain-specific studies that have yielded insights into pragmatic specialisations include Koester (2006), who looks at office talk, while Adolphs *et al*. (2007) explore health communication. Cotterill (2003) looks at the language of courtrooms, O'Keeffe (2006) examines radio phone-ins, while academic seminars are explored by Evison *et al*. (2007).

2) KEY FINDINGS FROM RESEARCH INTO SPOKEN CORPORA

Studies of lexical frequency

Leech *et al* (2001) present comprehensive English word lists for the spoken and written components of the BNC, showing which items are significantly more frequent in speaking or in writing. In addition to back-channel items such as *er, erm,* and *mm*, words such as *yeah, oh* and *no*, the verbs *know, think* and *mean* rank very high among the items distinctively characteristic of the spoken language, owing to their frequency of occurrence in the discourse-marking items *you know, I think* and *I mean*. Other discourse markers such as *well, right, okay, really* and *actually* also achieve high ranks in the spoken list. These items reveal a lot about the nature of everyday spoken English interaction: *you know*, *I think* and *I mean* all form part of the web of interpersonal relations and the monitoring of shared and non-shared knowledge. Other items in Leech *et al*'s (ibid.) list include hedges such as *just, sort of* and *a bit.* The prominence of all of these items in the lexis of spoken English is indicative of the real-time, constant monitoring of the interpersonal stratum that speakers engage in as the discourse unfolds (see also Stenström, 1990). Notably, the computer gives to vocalisations such as *er* and *mm* the same status as conventional words, thus underpinning the work of conversation analysts on back-channel behaviour and reinforcing the ubiquity of such 'non-words' in conversation.

Carter and McCarthy (2006: 830-831) provide lists of lexical chunks for written and spoken English of up to five words long. Their written lists are dominated by prepositional phrases and noun-phrase elements with embedded prepositional phrases, in contrast with the spoken lists, which remain characterised by the presence of verbs such as *know* and *mean* and vague expressions such as *and that sort of thing* and *it's a bit of a*. Other studies have also attempted to assess the use of recurring clusters or chunks and the general conclusion is that chunks are an important characteristic of speaking (Altenberg, 1991; Biber et al, 1999; Erman and Warren, 2000; McCarthy and Carter, 2002; Sinclair and Mauranen, 2006).

Buttery and McCarthy (2012) compared the top 2,000 words in the spoken list of the BNC with the top 2,000 words in its written list. They found that approximately 65% of the words were common to both lists, leaving some 35% of words that were unique to either the spoken or written list. On examination of those unique to the spoken list, Buttery and McCarthy noted that, as reported in McCarthy and Carter (1997a), the spoken list was characterised by words that support face-to-face interaction (including pragmatic markers such as *well*, *like* and *right*), as well as informal words (e.g. *y*-suffix adjectives such as *yucky, stroppy* and *comfy*).

Research on items characteristic of speaking

Alongside bird's-eye-view studies of spoken corpora as a whole are studies of individual items that are frequent in spoken data, especially everyday conversation. These items tend to be words and multi-word strings of high frequency in spoken data and/or items notably higher in frequency than in comparable written data. Tottie (1991) investigated backchannel behaviour in British and American English spoken data, and looked at vocalisations such as *mm, mhm* and *uh-(h)uh* alongside 'bona fide words and phrases' (Tottie, 1991: 255). Tottie's work underpinned the body of back-channel research in corpus linguistics and conversation analysis (Yngve 1970; Gardner 1997, 1998) and shed light on the problem of establishing boundaries between vocalisations, short responsive turns and full, floor-grabbing turns (Duncan and Niederehe, 1974; Zimmerman 1993; Tottie, 2011).

Aijmer (2002) examined 'discourse particles' (e.g. *now, oh, just, sort of, and that sort of thing, actually*), and focused on contextual cues such as text type, position in the talk, prosody and collocation. Aijmer's numerous studies of pragmatic markers, including common words and phrases such as *actually, well, of course, it's okay, I think, sort of* and *kind of* (Aijmer, 1984, 1996, 2001, 2003) reveal items of high frequency in talk which are not easily amenable to reflection and objective analysis without the evidence of corpus data. All these studies stand as a powerful counter-action against public prejudice about the use of many of the pragmatic markers, which may be perceived as of low status and negatively evaluated (Watts 1989).

Spoken grammar

Biber et al (1999) investigated differences of distribution and function of grammatical items as between written registers (fiction writing, news writing and academic writing) and conversation. Carter and McCarthy (1995) had also listed common grammatical features found in their conversational corpus that were rare or which functioned differently in writing; they also drew attention to cases where grammatical items and features are particularly associated with either speaking or writing (Carter and McCarthy, 2006). Rühlemann (2007:11) notes that much of the work done on grammar in spoken corpora should perhaps be better termed conversational grammar, since it is there that the most outstanding differences between speaking and writing have been brought to light. Leech (2000) also notes how it is often conversational data which stands out as different from the rest. Leech discusses the fact that conversational speaking reflects its online, linear nature in the brevity of utterances (where words and phrases, rather than long clauses or heavily embedded structures, predominate).

Situational ellipsis is a good example of how spoken grammar reflects the conditions under which spontaneous speech occurs. In informal English conversation, pronouns, copular and auxiliary verbs and articles may be regularly absent from places where they would be considered obligatory in most forms of writing (Quirk et al, 1985: 895-900; Carter and McCarthy, 2006: 181ff). Rühlemann (2007: 55-58) sees situational ellipsis as reflecting the shared context of face-to-face conversation and real-time processing factors. Caines and Buttery (2010) report that, in their British English corpus, in 27% of questions with second-person subjects involving progressive aspect

(e.g. *What you doing? You been working?*), the auxiliary was not used, as compared to only 5.4% of occurrences in comparable written data. They demonstrate that ellipsis of this kind is not random, and surrounding grammatical contextual features (e.g. subject pronoun type, tense-aspect configuration) can be used in the creation of a predictive model for training computers in natural language processing, resulting in a high level of success in automated searching which may relieve the drudgery of manual analysis.

Other spoken grammatical features reflecting the conditions under which conversational speaking occurs include pre-posed and post-posed items, sometimes referred to as left- and right-dislocated items (Geluykens, 1992) or, by Carter and McCarthy (2006:782-783) as headers and tails. These are features not totally excluded or proscribed from the grammar of writing, but rather ones which are overwhelmingly preferred in speaking. Extract 1 is a typical example of the way a noun phrase or phrases focusing in on the topic may occur before the main subject of the verb (in this case the pronoun *he*), forming the header, a sort of lead-in for the listener (marked in bold). In formal written grammar, the pronoun *he* would be considered unnecessary or even ill-formed:

Extract 1
[The speaker is reminiscing on his years working in the maritime lighthouse service]
And er when the anchor man always had his hand on the rope you know and you'd hear him saying, "Anchor coming home sir anchor coming home sir." And **the engine man he** was on his knees beside the engine … [BNC[1]]

Extract 2 exemplifies the post-posed tail phenomenon, where a pronoun is later reiterated in the form of its fully lexical noun-phrase referent (marked in bold).

Extract 2
[speaker is talking about a wrist watch which she changed for a smaller one because the face was too big]

[1] All rights in the texts cited from the BNC are reserved (Oxford University Computing Services on behalf of the BNC Consortium).

<$1> Oh that's beautiful isn't it?
<$2> Yeah, got a small little face and+
<$1> It's gorgeous that
<$2> +it had to change that for me, cos **it** was so big **the other one**
<$1> Mm (BNC)

McCarthy and Carter (1997b) found that tails of this kind correlated strongly with evaluative contexts (see also Aijmer 1989). Headers and tails are indicative of the real-time, online construction that is characteristic of spoken grammar, where items may be only loosely related in terms of conventional written structures and where the grammatical output is essentially linear and listener-sensitive.

Phenomena such as headers and tails and particular types of ellipsis, e.g. ellipsis of determiners, existential *there*, conditional *if* (see Carter and McCarthy, 2006: 185-187 for examples), because of their low likelihood of occurrence in written corpora, can easily be overlooked or relegated to non-standard or low-status usage. Their presence in spoken corpora from the mouths of speakers of all regional and social backgrounds, ages, and educational achievements show them to be anything but rare or non-standard. In this respect, one of the achievements of spoken corpus analysis has been to raise questions about the nature of "standard" grammar, and the sources from which the grammatical canon is conventionally derived (see in particular the papers by Carter and Cheshire in Bex and Watts, 1999).

Discourse and pragmatics

A notable pragmatic feature that spoken corpus research has brought to light is the ubiquity of vague language in conversation, with vague category markers such as *and things like that, or something, or whatever, and that sort/kind of thing* (see Carter and McCarthy, 2006: 835-836 for examples) revealing how speakers project assumptions of shared context, shared knowledge, shared meanings and world views among interlocutors (O'Keeffe 2003, Cutting 2007, Evison *et al* 2007).

Another area of spoken corpus analysis that has been fruitful is the study of turn-

construction. Although the onset and construction of any individual turn in informal conversation may be quite unpredictable (apart from highly ritualised turns such as greetings, congratulations, thanks, and so on), corpus analysis shows that a surprisingly small repertoire of words can account for a large number of turn-openings. Tao (2003) searched in his corpus for the words immediately following a new speaker tag and found a notable consistency in how speakers opened their turns. Turn-openings, in Tao's data, utilise a small repertoire of items (e.g. *yeah, uh-huh, oh, and, well, so, right, okay, no* and personal pronouns). Tao's overall conclusion, that turn-openers are syntactically free forms that function as links or bridges between turns, is a powerful indicator of the way speakers work to create continuity or "flow" in conversation, a phenomenon which McCarthy (2010) refers to as "confluence". McCarthy (2002; 2003) had already noted how, in conversational corpora, responsive turns routinely consisted of single-word adjectives or adverbs such as f*ine, great, absolutely, definitely*, along with responses consisting of clusters of such items (e.g. *Okay, great, fine!*) or reduplications of particular items (*Good, good*.) (see also O'Keeffe and Adolphs 2008).

Research into spoken corpora, both general and specialised, has brought to light features of everyday interaction that are difficult to access through intuition or reflection alone. Spoken corpus investigations have often served to add large-scale, quantitative underpinning to the explanations and insights of conversation analysts, discourse analysts and pragmaticians, as well as offering ways of investigating phenomena such as speaking turns or problematic grammatical phenomena such as ellipsis. Discourse analysts, pragmaticians and conversation analysts have much to gain from large-scale corpus analysis through the ratification of or challenge to findings based on small amounts or individual pieces of data. In the realm of grammar, spoken corpora might be said to have disturbed the soil more fundamentally, raising debates that will no doubt continue for some time, while in the lexical domain, confirmation and hard evidence of the ubiquity of chunking has provided a new perspective and a renewed interest in the vocabulary of conversation.

3) KEY PEDAGOGICAL IMPLICATIONS FROM THE RESEARCH INTO

SPOKEN CORPORA

Lack of application

Römer (2008) provides a survey of direct and indirect influences of corpora on language teaching. There is no gainsaying that the influence of corpora has been extensive and is increasing. However, many of the research findings outlined in the section above are, at the time of writing, poorly represented in language teaching materials. Generally, in the teaching of oral skills, most attention is given to the more mono-directional notion of 'speaking skills' as opposed to bi- or multi-directional conversation skills. Equally, when textbooks focus on 'listening skills' they usually separate them from the concept of speaking and focus almost solely on developing listening comprehension skills, where students listen and then complete content-related questions about what they have just listened to. As the research from spoken corpora illustrates, real conversational listening involves responding and co-constructing, with a speaker, across turns. One possible reason for the lack of widespread application of the findings of research into spoken corpora could be their general absence from language teacher education programmes as discussed by O'Keeffe and Farr (2003) and McCarthy (2008).

Examples of materials which have applied research findings

A small number have taken on the challenge of translating spoken corpus findings into classroom materials, for example the *Touchstone* and *Viewpoint* series (McCarthy, McCarten and Sandiford 2005-2011; 2012-2013). The syllabus of this English for adults series, which covers the main language teaching levels from false beginner to advanced, has a strong focus on conversation and includes input and practice in conversation strategies in every unit at every level, based on insights from spoken corpus data. Learners are presented with conversational extracts based on spoken corpus evidence to illustrate target items. Frequency patterns are explicitly presented (e.g. adjectives most frequently used after *That's* in response tokens, which grammatical pattern is most common in speaking, e.g. the choice between *isn't* and *'s not* as the negative of *be*). One of the main tenets underpinning the course is the notion of promoting 'noticing' (Schmidt 1990, 1993), based on the belief that learners generally need to be assisted in developing observation and awareness of spoken

features, which are unlikely to simply come to them as second nature without pedagogical intervention and input enhancement (Sharwood Smith 1993). The *Touchstone/Viewpoint* series also attempts to bring together the skills of speaking and listening by highlighting appropriate responses to incoming talk and giving learners opportunities not only to listen to and comprehend audio input but also to react and respond in a contextually suitable manner.

The concept of spoken grammar has also become established in grammar reference books written (partly or wholly) with second-language learners in mind. Biber *et al* (1999) and Carter and McCarthy (2006) clearly affirm the distinction between spoken and written grammar and bring spoken corpus research insights into the purview of language teaching. Other grammars aimed at learners, and/or supplemented with exercises, which also feature corpus-based material on spoken language, include Carter et al (2011a and 2011b) and Bunting et al (2013).

4) A CASE STUDY OF A KEY FINDING FROM SPOKEN CORPORA AND HOW IT MIGHT BE APPLIED PEDAGOGICALLY

Case study: ellipsis in the context of 'zero auxiliary' progressive

We now turn to findings from a corpus study and consider how these might be applied to a teaching context. Our case study features an example of ellipsis in British English - omission of the auxiliary verb in progressive (continuous) aspect constructions - the so-called 'zero auxiliary' progressive. Such constructions do not feature a tensed auxiliary verb, as in (1a) where forms of *BE* and *HAVE* which would be obligatory in formal writing and formal speaking are not used (cf. 1b):

(1a) What you doing? Who you looking for? You been working?
(1b) What are you doing? Who are you looking for? Have you been working?

According to standard grammatical conventions, especially those derived from writing, the auxiliary verb is an obligatory feature of such constructions. But the ‘rule’ is not always adhered to in the production of informal spoken language, as shown by

(2)-(4) from the BNC[2]:

(2) How you feeling now? KBK 3474
(3) You not having any cake? KBW 13888
(4) What you been buying? KPV 5313

However, non-use of the auxiliary gains no mention in one of the major reference works of recent times on the grammar of English (Huddleston and Pullum 2002) and is only given passing mention in the footnotes of another (Quirk *et al* 1985). With the evidence of spoken corpora, we consider the progressive construction in a new light and point out the pedagogical implications of the corpus statistics presented here.

Corpus study

We extracted every progressive construction from the 10 million word spoken section of the BNC (sBNC). Auxiliary realisation (full, contracted or zero) was noted along with various linguistic and extra-linguistic properties at situational, clausal and lexical levels. These included subject type (pronoun, other noun, or 'zero') and subject person (1st-3rd, singular or plural, or 'zero'), clause type (declarative or interrogative), clause tense, clause polarity, and finally 'spontaneity' level (i.e. the formality of the recording context, from sermons to meetings to casual conversations).

As a result, we had a subcorpus of 93,253 annotated sentences, in which the majority of progressive constructions have pronominal subjects, most frequently third person singular, the clause is an un-negated present tense declarative and the auxiliary is contracted. The majority of the progressives occur at the informal spontaneity level even though overall this makes up only 4 million of the 10 million words contained in sBNC.

When we overlay these variables and investigate the interaction of factors, we find that the zero auxiliary occurs in almost every context available, with the exception of interrogatives at the formal, scripted level, for which frequencies are very low anyway. That is, the zero auxiliary is near-ubiquitous in terms of the contexts in

[2] Each extract is followed by a unique text identifier and sentence number.

which it can occur, although it only occurs at low frequencies, proportionally-speaking, for all but the zero subject and interrogatives at the informal spontaneity level. Here in the interrogatives we find that the zero auxiliary is most frequent in the second person (at 34.1%), followed by the first person plural (23.6%) and then third person plural pronouns (20.2%). These three construction types are exemplified below:

(5) You still using that monitor? (KD5 9846)
(6) You been waiting long? (KDK 510)
(7) Who we talking about? (KBW 15230)
(8) We opening them now? (KD0 5133)
(9) They rising to the top a lot Zoe? (KB6 478)
(10) What they charging him with? (KDP 556)

We also found that zero subject + zero auxiliary constructions are a highly frequent type, and this is the main way in which declarative zero auxiliaries occur. For all spontaneity levels the proportion of zero subject progressives without an auxiliary is at least 50%. This type of zero auxiliary is pervasive across registers; from the most formal to the least formal, the zero subject + zero auxiliary is found.

(11) Yeah hold on just looking at something (KD1 920)
(12) Trying to decide whether to take them down off my windows and put some poles up (KCX 1178)
(13) It's pretty decent. Thinking of buying myself one (KNY 565)

On the other hand, for zero auxiliaries with a subject noun or pronoun, we can infer a stylistic dimension to their use. In the more formal registers the zero auxiliary is a rare occurrence, whereas at the informal spontaneity level its use, especially for interrogatives, rises markedly. This stylistic dimension is one that we will return to when considering the pedagogical implications of our findings below.

A final point from our corpus study is that certain constructions correlate strongly with zero auxiliary use. We used the 'collostructional' statistical method

(Stefanowitsch and Gries 2003) to identify the verbs most 'attracted' to and 'repulsed' by the zero auxiliary. On the plus side, *doing, going/gonna* and *laughing* were attracted to a statistically significant degree, whilst *saying, taking, working* and *happening* were strongly repulsed. Further investigation reveals that a small number of constructional patterns account for more than half of the second person zero auxiliary interrogatives, as shown in Table 1.

Constructional patterns	Frequency	% accounted for
wh- *you going/gonna* + V	189	14
what you doing	185	14
you going/gonna + V	132	10
where you going	101	7.5
how you doing	49	3.5
what you looking for/at	24	2
what you talking about	23	2
what you having	11	1
what you laughing for/at	7	0.5
Subtotal	721	54.5
2nd person interrogative zero auxiliary	1330	100

Table 1: Highly frequent second person interrogative zero auxiliary patterns in the spoken section of the BNC

In this section, we have shown that a feature of English that is thought to be obligatory – the progressive aspect auxiliary verbs *BE* and *HAVE* – is in fact at times omitted by native speakers of English. Furthermore, our BNC survey demonstrates that such omission occurs more frequently in certain lexico-syntactic contexts. In the

next section, we take the observations from this corpus study into account in our discussion of pedagogical implications.

Pedagogical implications

These results bring to light an issue which is both methodologically challenging for corpus linguists and English language teachers in that they relate to researching and teaching a feature which involves variable *absence*. The case study shows how corpus research into spoken data can unearth the key patterns of ellipsis in spoken language, in this case in relation to auxiliary verbs, and can dispel erroneous intuitions about what we feel is 'obligatory'. In the sBNC data, for the participants at least, nothing is 'missing' and the utterances are perfectly grammatical.

Our case study shows that a grammatical item thought to be obligatory in progressive constructions - the auxiliary verb *BE,* or *HAVE* in perfect constructions - is in fact not always used by native speakers, especially in interrogative clauses, in zero subject constructions, and even more so in less formal registers. Since these auxiliaries are thought to be obligatory, they will generally always have been taught as such, both in first and second language teaching.

The challenge for the teacher then is how to 'teach' something which is absent. The first step is 'noticing' (Schmidt 1993). A simple drill such as the following would bring to students' attention what typically happens in spoken language:

Below are real examples of what people say, taken from recordings of conversations. How would these differ if they were written rather than spoken? What is the effect of the changes in the written versions?
1. How you doing?
2. [Talking about food] What you having?
3. Think you don't need one.
4. Where you going Mum?
5. Trying to think. [BNC data]

Secondly, once the students' attention has been brought to some of the typical differences between (casual) spoken English and (formal) written English, the teacher may encourage appropriate practice in non-use of the auxiliary: *i.e.* in less formal

situations, when asking questions, and with second person subjects above all. Patterns such as *what/how you doing, where you going,* and *(wh-) you going/gonna* + V are especially appropriate skills to teach, as these are the types of zero auxiliary most frequently used by native speakers and most likely to be heard by students in encounters with native usage in films, internet chat and other forms of media, as well as in face-to-face encounters.

The teacher could then move on to similar examples of informal ellipsis in appropriate situations: for instance with the omission of copular *be,* omission of subject pronouns, or the omission of determiners in spoken English. As with the zero auxiliaries above, corpus resources could be used to demonstrate and enhance such teaching.

5 LOOKING TO THE FUTURE

In this chapter, we have attempted to demonstrate the potential use and benefit of spoken corpus research in language teaching. In order that this potential may be fully realised, there is a need for greater awareness of corpus resources and new findings from corpus research among language teachers. Academic networks such as the *English Profile* Project (http://www.englishprofile.org/) can help in this regard, as can journals aimed at language teachers, such as *Language Teaching, ELT Journal,* and *Language Teaching Research.*

To underpin any such development, major corpora themselves need to be enhanced with a greater amount of spoken data, as well as greater coverage of different contexts and registers. Equally, corpus linguists need to work with increased zeal towards making their findings accessible and transferable to pedagogy. As advocated in O'Keeffe and Farr (2003), corpus linguists need to present their research at teacher conferences and in language teaching journals in greater numbers than at the time of writing. This would add weight to the importance of including corpus linguistics as part of language teacher education programmes.

There also needs to be a wider availability of corpus-based teaching tools which allow

for live concordance searches, visualisation of corpus frequencies, and audio examples, at the very least. While there is a substantial amount of spoken data available online, for example via the COCA corpus, it is often scripted or media data. This is not quite representative of the most frequent human activity, that of everyday face-to-face conversation. In this regard, multimodal corpora, where audio and video data are combined and analysed in tandem, offer the prospect of further enhancing the language learning experience with video examples allowing for an additional focus on gesture and body language. Focusing on an innovative tool developed to make corpus use easier to access for language teaching, Farr (2010) details the potential of the SACODEYL (System Aided Compilation and Open Distribution of European Youth Language, a European Commission-funded project) corpus. This is a corpus of interviews with teenagers in seven different languages, available as a multi-modal corpus (audio files, video files and transcriptions). This, and similar teacher-led innovations, will be key to bringing the benefits of spoken corpora directly to the language classroom.

Recently, too, debates have arisen over the status and positioning of speaking in blended learning and online environments. Decisions on which aspects of the classroom to 'flip' to a computer-mediated environment and which to retain in the face to face classroom could be positively underpinned by an awareness of the findings of spoken corpus investigations (for example, the need to offer opportunities for active, responsive listening, as discussed above, or the need to develop noticing skills). At the time of writing, computer-mediated learning activities offer limited resources for recreating face-to-face spoken interaction in terms of controlled exercises, though sophisticated adaptive learning technologies may, in the future, replicate more convincingly the experience of listener feedback and bi-directional conversational flow (the 'confluence' referred to above) . However, the addition to the blended learning environment of online social networking in the form of blogs, wikis, email exchanges or synchronous computer-mediated chat (SCMC) does offer contexts in which the patterns of informal dialogue immanent in spoken corpus data are seen as both natural and appropriate to the process of engaging with one's peers and teachers when learning a language (see Stevenson and Liu 2010 for a discussion of social networking in online language learning).

Further reading

McCarthy, M. J. (1998) *Spoken Language and Applied Linguistics,* Cambridge: Cambridge University Press.
McCarthy describes the genesis of the five-million word CANCODE spoken corpus, offers corpus-informed answers to the question of what can and should be taught about the spoken language and reports on findings relevant to the teaching of grammar, vocabulary and other features of speaking.

O'Keeffe, A., McCarthy, M. J. and Carter, R. A. (2007) *From Corpus to Classroom: Language Use and Language Teaching,* Cambridge: Cambridge University Press.
This book gives and introduction to corpora for language teachers. It brings together the findings from corpus research and their applications in the language classroom, with an extensive focus on spoken data and suggestions for practical applications in relation to some of the features in the present chapter.

Aijmer, K. and Stenström, A.-B. (2005) Approaches to spoken interaction, *Journal of Pragmatics,* 37: 1743-1751.
This paper forms the introduction to a series of articles on spoken interaction as evidenced in spoken corpora. It explores how spoken corpus linguistics relates to other linguistic sub-disciplines such as conversation analysis and discourse analysis and provides a useful overview of the rest of the equally important papers in the journal's special issue.

Reppen, R. (2010) *Using Corpora in the Language Classroom,* Cambridge: Cambridge University Press.
This book is designed to help teachers and teacher trainers better to understand corpus linguistics and the ways in which corpus resources can be brought into the classroom. It features a directory of available corpus resources and comes with an informative companion website.

References

Adolphs, S., Atkins, S. and Harvey, K. (2007) 'Caught between professional requirements and interpersonal needs: vague language in healthcare contexts', in J. Cutting (ed.) *Vague Language Explored*, Basingstoke: Palgrave Macmillan: 62-78.

Aijmer, K. (1984) '*Sort of* and *kind of* in English conversation', *Studia Linguistica,* 38: 118-28.

Aijmer, K. (1989) 'Themes and tails: the discourse function of dislocated elements', *Nordic Journal of Linguistics,* 12 (2): 137-54.

Aijmer, K. (2002) *English discourse particles: Evidence from a corpus*, Amsterdam/Philadelphia: John Benjamins.

Aijmer, K. (1996) 'I think— an English modal particle', in T. Swan and O. J. Westvik (eds.) *Modality in Germanic languages. Historical and comparative perspectives*, Berlin: Mouton de Gruyter: 1-47.

Aijmer, K. (2001) 'It's okay', in M. Ljung (ed.) *Language Structure and Variation*, University of Stockholm: Stockholm Studies in English 92: 1-17.

Aijmer, K. (2003) 'Discourse particles in contrast: the case of *in fact* and *actually*', in A. Wilson, P. Rayson and T. McEnery (eds.) *Corpus Linguistics by the Lune. A Festschrift for Geoffrey Leech*, Bern: Peter Lang: 23-35.

Altenberg, B. (1991) 'Amplifier Collocations in spoken English', in S. Johansson and A-B Stenström, *English Computer Corpora: Selected Papers and Research Guide,* The Hague: Mouton de Gruyter: 127-47.

Bex, T. and Watts, R. J. (eds.) (1999) *Standard English: The Widening Debate*,

London: Routledge.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) The Longman *grammar of spoken and written English*, London: Longman.

Breiteneder, A. Pitzl, M.-L., Majewski, S. and Klimpfinger, T. (2006) 'VOICE recording - Methodological challenges in the compilation of a corpus of spoken ELF', *Nordic Journal of English Studies,* 5 (2): 161-88. (http://hdl.handle.net/2077/3153)

Bunting, J. and Diniz, L. with Reppen, R. (2013) *Grammar and Beyond.* Level 4, Cambridge: Cambridge University Press.

Buttery, P and McCarthy, M. J. (2012) 'Lexis in spoken discourse', in J. Gee and M. Handford (eds.) *The Routledge Handbook of Discourse Analysis*, Abingdon, Oxon and New York: Routledge: 285-300.

Caines, A. P. and Buttery, P. J. (2010) 'You Talking to Me? A predictive model for zero-auxiliary constructions', *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground, ACL-2010*, Uppsala: Association for Computational Linguistics: 43-51.

Carter, R. A. and McCarthy, M. J. (1995) 'Grammar and the spoken language', *Applied Linguistics,* 16 (2): 141-58.

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

Carter, R. A., McCarthy, M. J., Mark, G. and O'Keeffe, A. (2011a) *English Grammar Today,* Cambridge: Cambridge University Press.

Carter, R. A., McCarthy, M. J., Mark, G. and O'Keeffe, A. (2011b) *English Grammar Today Workbook,* Cambridge: Cambridge University Press.

Chafe W., Du Bois J. and Thompson S. (1991) 'Towards a new corpus of spoken American English', in K. Aijmer and B. Altenberg (eds) E*nglish Corpus Linguistics*,

London: Longman: 64-82.

Cotterill, J. (2003) *Language and Power in Court: A Linguistic Analysis of the O. J. Simpson Trial,* Basingstoke: Palgrave Macmillan.

Crowdy, S. (1993) 'Spoken corpus design', *Literary and Linguistic Computing*, 8: 259–65.

Cutting, J. (ed.) (2007) *Vague Language Explored*, Basingstoke: Palgrave Macmillan.

Davies, M. (2010) 'The Corpus of Contemporary American English as the first reliable monitor corpus of English', *Literary and Linguistic Computing,* 25 (4): 447–65.

Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A. and Martey, N. (2003) S*anta Barbara Corpus of Spoken American English Part II,* Philadelphia: Linguistic Data Consortium.

Duncan, S. and Niederehe, G. (1974) 'On signalling that it's your turn to speak', *Journal of Experimental Social Psychology,* 10 (3): 234-47.

Erman, B. and Warren, B. (2000) 'The idiom principle and the open choice principle', *Text,* 20 (1): 29-62.

Evison, J. M., McCarthy, M. J. and O'Keeffe, A. (2007) 'Looking out for love and all the rest of it: vague category markers as shared social space', in J. Cutting (ed.) *Vague Language Explored*, Basingstoke: Palgrave Macmillan: 138-57.

Farr, F. (2010) 'How can corpora be used in teacher education?', in A. O'Keeffe and M. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*, London: Routledge: 620-32.

Gardner, R. (1997) 'The listener and minimal responses in conversational interaction', *Prospect,* 12 (2): 12-32.

Gardner, R. (1998) 'Between speaking and listening: the vocalisation of

understandings', *Applied Linguistics,* 19 (2): 204-24.

Geluykens, R. (1992) *From Discourse Process to Grammatical Construction: on Left-dislocation in English*, Amsterdam: John Benjamins.

Gilquin, G., De Cock, S. and Granger, S. (2010) *Louvain International Database of Spoken English Interlanguage*, Louvain-la-Neuve: Presses Universitaires de Louvain. (http://www.uclouvain.be/en-352660.html)

Harrington, J. (2010) *Phonetic Analysis of Speech Corpora*, Oxford: Wiley-Blackwell.

Huddleston, R. and Pullum, G. K. (2002) *The Cambridge Grammar of the English Languag*e. Cambridge: Cambridge University Press.

Ide, N. and Macleod, C. (2001) 'The American National Corpus: A Standardized Resource of American English', in P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja (eds.) *Proceedings of Corpus Linguistics 2001,* Vol. 13*,* Lancaster: University of Lancaster: 274-80.

Koester, A. (2006) *Investigating Workplace Discourse,* Abingdon, Oxon: Routledge.

Leech, G. (2000) 'Grammars of spoken English: New outcomes of corpus-oriented research', *Language Learning,* 50 (4): 675-724.

Leech, G., Rayson, P. and Wilson, A. (2001) *Word Frequencies in Written and Spoken English: based on the British National Corpus,* London: Longman.

Mauranen, A. (2003) 'The Corpus of English as Lingua Franca in Academic Settings', *TESOL Quarterly,* 37 (3): 513-27.

McCarthy, M. J. (1998) *Spoken Language and Applied Linguistics*, Cambridge: Cambridge University Press.

McCarthy, M. J. (2008) 'Accessing and interpreting corpus information in the teacher education context', *Language Teaching*, 41 (4): 563-74.

McCarthy, M. J. (2002) ‘Good listenership made plain: British and American non-minimal response tokens in everyday conversation’, in R. Reppen, S. Fitzmaurice and D. Biber (eds.) *Using Corpora to Explore Linguistic Variation,* Amsterdam: John Benjamins: 49-71.

McCarthy, M. J. (2003) ‘Talking back: “small” interactional response tokens in everyday conversation’, *Research on Language in Social Interaction,* 36 (1): 33-63.

McCarthy, M. J. (2010) ‘Spoken fluency revisited’, *English Profile Journal,* 1 (1): e4.

McCarthy, M. J. and Carter, R. A. (1997a) ‘Written and spoken vocabulary’, in N. Schmitt and M.J. McCarthy (eds.) *Vocabulary: description, acquisition, pedagogy,* Cambridge: Cambridge University Press: 20-39.

McCarthy, M. J., and Carter, R. A. (1997b) ‘Grammar, tails and affect: constructing expressive choices in discourse’, *Text,* 17 (3): 405–429.

McCarthy, M. J. McCarten, J. and Sandiford, H. (2005-2011) *Touchstone*, Student Books 1-4, and blended/online program, Cambridge: Cambridge University Press.

McCarthy, M. J. McCarten, J. and Sandiford, H. (2012-2013) *Viewpoint,* Student Books 1 and 2, Cambridge: Cambridge University Press.

McCarthy, M. J. and O'Keeffe, A. (2008) ‘Corpora and the Study of Spoken Language’, in A, Ludeling, K. Merja and T. McEnery (eds.) *Handbook of Corpus Linguistics,* Berlin: Mouton de Gruyter: 1-16.

McCarthy, M. J., and Carter, R. A. (2002) ‘*This that and the other*: multi-word clusters in spoken English as visible patterns of interaction’, *Teanga* (Yearbook of the Irish Association for Applied Linguistics), 21: 30-52.

Nelson, G. (1996) ‘The Design of the Corpus’, in S. Greenbaum (ed.) *Comparing English Worldwide: the International Corpus of English*, Oxford: Oxford University Press: 27-35.

O'Keeffe, A. and F. Farr (2003) 'Using language corpora in language teacher education: pedagogic, linguistic and cultural insights', *TESOL Quarterly,* 37 (3): 389–418.

O'Keeffe, A. (2003) '*Like the wise virgins and all that jazz* – using a corpus to examine vague language and shared knowledge', in U. Connor and T. A. Upton (eds.) *Applied Corpus Linguistics: a multidimensional perspective,* Amsterdam: Rodopi: 1-20.

O'Keeffe, A. (2006) *Investigating Media Discourse,* Abingdon, Oxon: Routledge.

O'Keeffe, A. and Adolphs, S. (2008) 'Using a corpus to look at variational pragmatics: response tokens in British and Irish discourse', in K.P. Schneider and A. Barron (eds.) *Variational Pragmatics,* Amsterdam: John Benjamins: 69-98.

O'Keeffe, A., Clancy, B. and Adolphs, S. (2011) *Introducing Pragmatics in Use*, London: Routledge.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985) *A Comprehensive Grammar of the English Language*, London: Longman.

Römer, U. (2008) 'Corpora and language teaching', in A. Lüdeling and Merja K. (eds.) *Corpus Linguistics. An International Handbook (volume 1)*. Berlin: Mouton de Gruyter, 112-130.

Rühlemann, C. (2007) *Conversation in Context: a corpus-driven approach*, London: Continuum.

Schmidt, R. (1993) 'Awareness and second language acquisition', *Annual Review of Applied Linguistics,* 13: 206-26.

Sharwood Smith, M. (1993) 'Input enhancement in instructed SLA: theoretical bases', *Studies in Second Language Acquisition,* 15: 165–179.

Sinclair, J. M. and Mauranen, A. (2006) *Linear Unit Grammar: integrating speech*

and writing, Amsterdam: John Benjamins.

Stefanowitsch, A. and Gries, S. T. (2003) 'Collostructions: investigating the interaction between words and constructions', *International Journal of Corpus Linguistics,* 8 (2): 209-43.

Stenström, A.-B. (1990) 'Lexical items peculiar to spoken discourse', in J. Svartvik (ed.) *The London-Lund Corpus of Spoken English,* Lund: Lund University Press: 137-75.

Stevenson, M. P. and Liu, M. (2010) 'Learning a language with Web 2.0: exploring the use of social networking features of foreign language learning websites', *CALICO Journal,* 27 (2): 233-259.

Tao, H. (2003) 'Turn initiators in spoken English: A corpus-based approach to interaction and grammar', in P. Leistyna and C. F. Meyer (eds.) *Corpus Analysis: language structure and language use,* Amsterdam: Rodopi: 187-207.

Timmis, I. (2010) ' "Tails" of linguistic survival', *Applied Linguistics*, 31 (3): 325–45.

Tottie, G. (1991) 'Conversational style in British and American English, the case of backchannels', in K. Aijmer and B. Altenberg (eds.) *English Corpus Linguistics,* London: Longman: 254-71.

Tottie, G. (2011) '*Uh* and *um* as sociolinguistic markers in British English', *International Journal of Corpus Linguistics,* 16 (2): 173-97.

Watts, R. J. (1989) 'Taking the pitcher to the "well": native speakers' perception of their use of discourse markers in conversation', *Journal of Pragmatics,* 13: 203-37.

Yngve, V. (1970) 'On getting a word in edgewise', *Papers from the 6th Regional Meeting, Chicago Linguistic Society,* Chicago: Chicago Linguistic Society.

Zimmerman, D. (1993) 'Acknowledgement tokens and speakership incipiency

revisited', *Research on Language and Social Interaction,* 26 (2): 179-94.

Notes on Contributors (include affiliation, approx 40 words each)

Andrew Caines is a postdoctoral researcher in the Institute for Research in Automated Language Teaching and Assessment at the University of Cambridge. He has published research on the effect of document length and the effect of topic on language features in learner corpora. His doctoral dissertation was on zero auxiliaries in spoken British English.

Michael McCarthy is Emeritus Professor of Applied Linguistics at the University of Nottingham, UK. He is author/editor and co-author/co-editor of more than 50 books, including *The Routledge Handbook of Corpus Linguistics* (Routledge, 2010) and more than 100 articles on language teaching and on vocabulary, grammar, spoken discourse and spoken corpus linguistics.

Anne O'Keeffe is Senior Lecturer in Applied Linguistics at Mary Immaculate College, University of Limerick, Ireland. She is author of numerous publications on corpus linguistics, media discourse and on language teaching. She has published six books and has co-edited *The Routledge Handbook of Corpus Linguistics* (Routledge, 2010).

List of abbreviations and acronyms

ANC American National Corpus

BASE British Academic Spoken English corpus

BNC British National Corpus

CANCODE Cambridge and Nottingham Corpus of Discourse in English

COCA Corpus of Contemporary American English

ELDA Evaluations and Language resources Distribution Agency

ELFA English as a Lingua Franca in Academic Settings corpus

ICE International Corpus of English

LINDSEI Louvain International Database of Spoken English Interlanguage

MICASE Michigan Corpus of Academic Spoken English

SACODEYL System Aided Compilation and Open Distribution of European Youth Language

SBCSAE Santa Barbara Corpus of Spoken American English

SCMC synchronous computer-mediated chat

VOICE Vienna-Oxford International Corpus of English