59. Corpora and spoken language

Authors: Michael McCarthy & Anne O'Keeffe

University of Nottingham Mary Immaculate College
University of Limerick

Contact details:

Main contact:
Prof. Michael McCarthy
School of English Studies
University of Nottingham
Nottingham NG7 2RD
UK
Tel: +4401159515902
Fax: +4401159515924
Email: mactoft@dial.pipex.com

Dr Anne O'Keeffe
Mary Immaculate College – University of Limerick
South Circular Road
Limerick,
Ireland
Tel: +353 61 204957
Fax: +353 61 323632
Email: anne.okeeffe@mic.ul.ie

59. Corpora and spoken language

35

1. Introduction: evolution of spoken corpora

Spoken corpora have evolved over the last four decades from early attempts at corpus-building for the purposes of better understanding such phenomena as first-language acquisition, social variation and conversational structure, to the large, general spoken corpora of today, which have found applications in a variety of contexts from speech recognition, lexicography, sociolinguistics and first and second language acquisition. In this article we focus on spoken corpora and their applications in linguistics and applied linguistics, rather than on 'speech corpora', which are typically collected for the purposes of improving technology, a distinction discussed at greater length by Wichmann in article 15; see also article 32.

Some of the earliest spoken corpora were developed within the field of child language acquisition, an example of which was the child-language word-frequency analyses described in Beier, Starkweather and Miller (1967). Another example, which included informal spoken language by adults, as well as by selected age groups of children from six years upwards in a corpus of some 84,000 words, is described in Carterette/Jones (1974). A notable early spoken corpus project of the kind that has since become quite common was the *Oral*

Vocabulary of the Australian Worker (OVAW), for which Schonell et al. (1956) give a full account of the data and its collection. The *OVAW* corpus consisted of some 500,000 words of spoken language and was used for, among other things, the study of idiomatic words and phrases in speech. 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 presented statistics for spoken usage. An influential early spoken corpus of British English was the London-Lund corpus (LLC). This corpus derives from two projects: the Survey of English Usage (SEU) at University College London, launched in 1959 by Randolph Quirk. The second project is the Survey of Spoken English (SSE), which was started by Jan Svartvik at Lund University in 1975. The London-Lund corpus, which is the spoken component of the Survey of English Usage, comprises half a million words. Its goal is to provide a resource for accurate descriptions of the grammar of adult educated speakers of English (Svartvik 1990; Edwards/Lampert 1993). The spoken English component comprises both dialogue and monologue and was collected over a 30-year period towards the end of the last century.

Several other early spoken corpora for English were developed as appendages to much larger written corpora, a reflection of the time and expense

involved in collecting such data relative to written texts. Major corpus projects such as the COBUILD Bank of English (see Moon 1997) and the British National Corpus (see Crowdy 1993, 1994; Rundell 1995a, b) contain large spoken segments, including broadcast speech as well as everyday unrehearsed conversation. The British National Corpus (BNC) contains over 100 million words of data, with the spoken component accounting for around ten million words. The spoken data consist of unscripted informal conversations recorded by volunteers selected from different ages, regions and social classes in a demographically balanced way. It is designed to represent as wide a range of modern British English as possible.

In the USA, work by Chafe and his colleagues, initially based on the British London-Lund spoken corpus design (Chafe/Du Bois/Thompson 1991), developed into larger corpus enterprises such as the five-million word Longman Spoken American Corpus (see Stern 1997). Informal Australian spoken English has also been subjected to corpus analysis more recently by Eggins/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).

In 1984, Knowles, Alderson, Williams, Taylor, Leech and Kaye embarked on a joint research project between the University of Lancaster and the Speech Research Group at IBM UK Scientific Centre into the automatic assignment of intonation. The first aim of the project was to collect samples of natural spoken British English which could be used as a database for analysis and for testing the intonation assignment programs. The result was the Spoken English Corpus (SEC), a machine-readable corpus of approximately 52,000 words of spoken British English. The majority of texts in the corpus were obtained from the BBC and include news broadcasts, commentary, religious broadcast, magazine-style reporting as well as fiction, poetry and dialogue (see Knowles 1990). Leech (2000) notes that while the LLC and the SEC benefited from careful and detailed prosodic transcription, they suffer from restrictions owing to the data they contain. The LLC, for example, used heavy reel-to-reel tape recorders, and a considerable portion of the spontaneous dialogue data is restricted to academic settings among staff and students at London University, and so academic topics of conversation prevail, while the SEC is even narrower in range as its

recordings are mostly confined to scripted speech such as radio broadcasts.

In 1993, Stenström and Breivik set up the Bergen Corpus of London Teenage Language (COLT). The aim of the project was to create a corpus of British English teenage talk and make it available for research. The corpus designers believed that studying spontaneous teenage talk would yield insights into language development and language change, especially as regards grammaticalisation (see Breivik/Hasselgren 2002). The reason for restricting the corpus collection to London was the assumption that new trends predominate among teenagers in the capital, from where they can be expected to spread to the rest of the country, and even further afield. Stenström/Andersen/Hasund (2002) provide an extensive study of the COLT data, outlining the most prominent features of the teenagers' talk including 'slanguage', speech reporting, non-standard grammatical features, intensifiers, tags, and interactional behaviour in terms of conflict talk.

In 2000, Leech noted that 'it may seem strange that the United States, where the age of English electronic corpora began with the Brown Corpus (in 1961), has held back from the development of a wide-coverage spoken corpora' (Leech 200, 684). This may be due to the long shadow cast by the general rejection of

corpus data by Chomskyan linguists, Leech (2000) surmises. Offsetting the earlier lack of a major spoken corpus project, the American National Corpus (ANC) was set up as a comparative corpus to the BNC (Ide/Macleod 2001; Ide/ Reppen/Suderman 2002). In 2003, a pilot sample of 11 million words was released. This comprised over 3 million words of spoken data and over 8 million words of written texts. The spoken data came from three sources: 1.5% from 'Callhome' (10 minute segments of telephone conversations), 95% from 'Switchboard' (2320 spontaneous telephone conversations averaging six minutes in length and comprising about 3 million words by over 500 speakers) and 3.5% from 'Charlotte Narratives' (95 narratives, conversations and interviews representative of the residents of Mecklenburg County, North Carolina and surrounding North Carolina communities). The full corpus, consisting of (at least) 100 million words annotated for part of speech, together with search and retrieval software, was expected to be in place in the fall of 2005 (see ANC website http://americannationalcorpus.org/ 2004).

As in so many other aspects of linguistic study, English tended to dominate spoken corpus building in the earlier years, but spoken corpora for many other languages now exist, including Bulgarian, French (both European and Canadian), Mandarin

Chinese, Vietnamese, Egyptian Arabic, Farsi, German, Greek, European Spanish, Hindi, Japanese, and Korean, Tamil, Vietnamese, 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), see also section 5.

2. Corpora for studying language varieties and types of discourse

The International Corpus of English (ICE) project was launched in 1991 by Sidney Greenbaum (see Greenbaum 1991, 1992). His initial goal was to gather at least 15 regional components from countries where English was the 'native language' as well as countries where it was an 'official non-native language – India, Nigeria and the Philippines' (Greenbaum 1992, 171). Each corpus would comprise one million words of spoken and written material and the same template would be used throughout in the compilation and collection of data. The goal was 'to provide the means for comparative studies' and for the first time provide 'the resources for systematic study of the national variety as an end in itself' (Greenbaum 1992: 171). This project has led to the collection of spoken and written data

for the Englishes of Hong Kong (Bolton et al. 2003), New Zealand (Holmes 1996), Singapore (Ooi 1997), Great Britain (Nelson/Wallis/Aarts 2002), Ireland (Kallen/Kirk 2001), Nigeria (Banjo 1996), East Africa (Schmied/Hudson-Ettle 1996) and the Caribbean (Nero 2000), with others under development.

In recent years a number of spoken corpora have been assembled with the express purpose of the study of aspects of spoken discourse in both formal and informal settings. The design principles of such corpora differ from spoken corpora collected for more general purposes. One such example is the Cambridge and Nottingham Corpus of Discourse in English (CANCODE), a five million word collection of spoken data. It is designed so as to represent spoken language in different contexts of use, genres of speech and between different speaker relationships across the islands of Britain and Ireland (see McCarthy 1998). The corpus design focuses on representing a range of discourse contexts and speech genres across different speaker relationships with the aim of informing research and language pedagogy in the fields of lexis, grammar and discourse. Using the same design matrix, the Limerick Corpus of Irish English (LCIE) comprises one million words of Irish English conversations (see Farr/Murphy/O'Keeffe 2002). Other discourse-oriented corpora include that described by Cheng and Warren, who oversaw the collection of the two-

million-word Hong Kong Corpus of Spoken English (HKCSE) (see Cheng/Warren 1999, 2000).

Spoken corpora focusing on institutional settings include the Michigan Corpus of Academic Spoken English (MICASE) (Simpson/Lucka/Ovens 2000), offering by 2004 online access to more than 150 transcripts of academic speech events recorded at the University of Michigan, USA (totalling 1.8 million words). MICASE was established in 1997 with the goal of describing the characteristics of contemporary academic speech and any potential differences across academic disciplines and different classes of speakers. The MICASE data consist of speech within the microcosm of the University of Michigan at Ann Arbor. Speakers represented in the corpus include faculty, staff, and all levels of students, and both native and non-native speakers. The contexts in which the recordings were made include large lectures, discussions, seminars, student presentations, advising sessions, dissertation defences, interviews, meetings, office hours, service encounters, study groups, tours and tutorials. Farr (2003) looks at a corpus of spoken encounters in the context of teacher education consisting of post-observation trainer-trainee interactions (the POTTI corpus) in a university setting. The Cambridge and Nottingham Business English Corpus (CANBEC), a one million word corpus of conversations in business contexts (see

Handford/McCarthy 2004, O'Keeffe/McCarthy/Carter 2007), and the Corpus of Spoken Professional American English (CSPAE) a two million-word corpus, consisting of 50 per cent White House press briefings and 50 per cent university academic council meetings (Barlow 1998) are also recent examples of specialised, targeted spoken corpora. Within the field of language pedagogy, learner spoken data have been collected, a notable example being the Louvain International Database of Spoken English Interlanguage (LINDSEI) set up in 1995 (see De Cock, 1998, 2000), which provides spoken data for the analysis of the speech of second language learners (see also Granger/Hung/Petch-Tyson 2002).

3. Size, representativeness, transcription, and other issues

Spoken corpora, because of collection and transcription problems and financial and time constraints, inevitably tend to be much smaller than general written corpora. However, this is not always necessarily seen as a disadvantage. Leech (2000) notes that more important than size for assessing the research value of a corpus is its composition in terms of genres and other design features. Furthermore, a number of researchers have noted the value of small corpora for particular kinds of research (Carter/McCarthy 1995; McCarthy/Carter 2001a; Cameron/Deignan

2003). O'Keeffe/Farr (2003) suggest the following guidelines: for spoken corpora anything over one million words is considered to be moving into the 'larger' range, for written anything below five million is quite small. McCarthy/Carter (2001a), arguing for more qualitative research (as opposed to the quantitative tradition) in corpus linguistics support the view that small spoken corpora can be used to great effect, especially where high-frequency linguistic items and features are concerned.

Various perspectives on how a corpus should be designed concur that it should be a principled collection of texts that is assembled for a specific purpose. Sinclair (1995) sees a corpus as a something that is not a random assortment of data but a collection of pieces of language that are selected and ordered to explicit linguistic criteria to be later used as a sample of the language in question (see also Francis 1982; Atkins/Clear/Ostler 1992; Crowdy 1994; Biber /Conrad/Reppen 1998; Tognini-Bonelli 2001). Three criteria generally prevail in the literature as regards good corpus design: 1) authenticity of the texts, 2) representativeness of language included in corpus and 3) sampling criteria used in the selection of texts (Tognini-Bonelli 2001, 54). Hunston (article 11) offers, in addition to the criteria of size and the problematic notion of representativeness, the criterion of

balance, that is to say ensuring equality in the sizes of the sub-corpora that make up the whole corpus (see Hunston's discussion of the MICASE spoken academic corpus). Decisions regarding the representativeness and balance of written corpora may be largely resolved by recourse to text typologies (see Crystal 1995; Lee 2001, Aston 2001) and ensuring that the corpus includes a broad coverage of text types in substantial and balanced quantities. In the case of the design of spoken corpora, however, not least of the problems is deciding precisely what constitutes a text. Written texts have clear orthographic boundaries, which spoken texts do not. And in the case of casual conversation, topical segments blend into one another, paragraphs and sentences are a mere artefact of transcription and, except in the case of extended monologue, more than one speaker contributes to the text, often simultaneously. Two main non-text-based solutions to these problems are therefore commonly pursued. One is to collect demographically stratified samples of undetermined (or arbitrarily chosen) length which may be to a greater or lesser extent clearly delineated in terms of boundary phenomena such as conversational openings and closings. For example some spoken corpora aim to represent a language variety, e.g. the British National Corpus (BNC), and therefore need to give careful attention to the collection of data across a representative balance of standard

demographic sampling variables for example gender, age, region, social class, etc. The Corpus of London Teenage speech (COLT) on the other hand only sought to represent one age group in one region, so while COLT modelled its design principles on the BNC, it limited its scope to a sample of teenagers in the London area. During a three-week period, using a network of London schools, students carried a small recording device and a lapel microphone for a few days and recorded all the conversations they took part in, with friends of the same age who were not supposed to be aware of the recording. The recruits were also equipped with a logbook and instructed to write down information about the co-speaker(s) and the setting. In three weeks all 0.5 million words of spoken language were collected (see also article 15). The other, not mutually exclusive solution to the problem of delineating data samples is to take a context- or genre-based approach, in which spoken samples are collected based on a pre-determined set of situational parameters. Corpora such as CANCODE, LCIE and HKCSE set out to examine English spoken discourse in specified contexts rather than to describe a language variety. In such cases, a highly representative corpus is not necessarily one which adheres to demographic sampling principles, but rather one which is based on representing the genres of spoken language itself (article 15 gives further examples of genre-based

approaches to spoken corpus design). The five-million word CANCODE spoken corpus, for example, was designed so as to represent everyday spoken language across different genres and speaker relationships. The design of CANCODE as described in McCarthy (1998) was based on a matrix with two axes for classification: *context type* and *interaction type*. Context type distinguished texts that were predominantly collaborative and those that were non-collaborative. The collaborative types were classified as *ideas* (e.g. exchanging opinions) and *tasks* (engagement in some physical task, e.g. doing the washing up) whereas the non-collaborative types were more asymmetrical and were classified as *information provision*. The interaction types reflected the relationship between the conversational participants. These fell into five broad categories: intimate, socialising, professional, transactional and pedagogic. LCIE used the same design principle with the same goal, and because these two corpora use the same design principles they have lent themselves to comparisons across two varieties (e.g. McCarthy/O'Keeffe, 2003). The HKCSE is also genre-based and includes Hong Kong Chinese speakers of English and native speakers of English. It is made up of four sub-corpora each comprising 0.5 million words, under the headings of conversations, academic discourses, business discourses, and public discourses. The data are transcribed

both orthographically and prosodically.

Transcription of spoken corpora is as Holmes et al. put it the art of making the ephemeral tangible in a consistent and practical manner (Holmes/Vine/Johnson, 1998). In reality the spoken word is very difficult to make tangible in written form as one immediately loses the audio and visual component in which it had its existence. Transcription has been the cause of much discussion and debate (see for example Ochs 1979; Edwards 1991; Cook 1990; Edwards/Lampert 1993; Bucholtz 2000; Hepburn 2004). Duranti (1997) suggests that transcripts are inherently incomplete and should be continuously revised to display features of an interaction that have been illuminated by a particular analysis and allow for new insights that might lead to a new analysis. Much of the already extant detailed work by conversation analysts has informed corpus transcription techniques over the years. For example, Jefferson (1985) provided a comprehensive account of laughter using a corpus of phone calls to a child protection helpline; Hepburn (2004), building on the work of Jefferson, examines crying using a corpus of calls to Child Protection Officers at the British National Society for the Prevention of Cruelty to Children. Hitherto, she points out, crying was considered as a unitary and self-evident category where it was uncommon for transcription to try and capture its

different elements. Her work makes explicit some different elements of crying and shows how these elements can be represented in transcription, for example sniffing, wobbly voice, high pitch, aspiration, sobbing and silence. Despite such detailed attention to potential features for transcription, however, large corpora tend to remain only broadly transcribed. Leech (2000) notes that many of the large spoken corpora were built primarily for the purpose of English language dictionaries and were transcribed quickly and at a low unit cost, which means a simple orthographic transcription. One of the consequences of such ' "basic" transcriptions' (Leech 2000, 678) is that while lexis and grammar can be investigated, key aspects of spoken language such as prosody and discourse cannot, due to the absence of accurate and detailed phonological, contextual and turn-taking information. For this reason Leech (ibid: 678) notes that 'even at a time when the availability of machine-readable corpora has brought a vast increase of knowledge about the spoken language within our grasp, the influence and limitations of the written language continue to impinge on the spoken medium'. Cheng/Warren (2002) also note that while the orthographic transcription of spoken data is well established and the conventions quite well-known, the number of spoken corpora that are also prosodically transcribed is very small, a well-known exception being the London-Lund Corpus of Spoken

English (Svartvik/Quirk 1980; Svartvik 1990). Cheng and Warren point out that the representation of prosodic features is less standardized, that it is notoriously difficult and time-consuming to prosodically transcribe naturally-occurring data, and that it ideally requires inter-rater reliability measures to ensure the quality of the transcription. Articles 15 and 32 offer further discussion of transcription and annotation issues, especially those generated by the different purposes and applications which speech databases and spoken corpora, as demarcated at the beginning of this article, typically serve

4. Research: important findings

One far-reaching impact of the availability of spoken corpora can be seen in the attempts to elaborate independent descriptions of spoken grammar (Leech 2000). The availability of spoken corpus data brought to light the fact that written models were not always adequate to describe spoken usage. While, in the case of English and many other languages, the actual forms of grammar are to a very great extent shared between the spoken and written media, and while, potentially, any grammatical form may occur in either medium, the distribution of forms in actual fact is often markedly different across the two media (see Blanche-Benveniste 1995; Fonseca-Greber/Waugh

2003 for examples from French). Phenomena such as so-called left- and right-dislocated items (otherwise known as pre- and post-posed items) and situational ellipsis (e.g. non-use of otherwise obligatory forms such as verb subjects or determiners) are common in casual spoken data but extremely rare in most kinds of formal writing (Carter/McCarthy 1995, 2006). In this extract from the CANCODE spoken corpus, the speakers are looking at photographs; ellipsis of *the* occurs before *same* in <$1>'s turn, ellipsis of *you've* occurs before *seen* in <$4>'s turn, and *that* in <$5>'s turn is post-posed:

[<$#> = speaker number in order of speaking; <$?F> = speaker unidentifiable, probably female; <$E> <\$E> beginning and end of non-verbal event (e.g. laughter)]

<$4> Oh. I'm like my father there aren't I.

<$?F> You can't do anything about that now.

<$1> Yes. Mm. Same eyes look. Same shape.

<$4> Seen that one of Jim haven't you.

<$?F> <$E> laughs <\$E>

<$5> Yeah. It's a good one that.

Furthermore, the descriptive apparatus and terminology itself is called into question in the face of spoken corpus evidence. The notion of 'subordination' as it has derived from the intuition of grammarians or from the observation of written texts has come under close scrutiny (Blanche-Benveniste 1982). McCarthy (2001: 128) points out that

metaphors such as ‘left’ and right’ (as used to refer to dislocated elements) are western-culture, page-driven ones, and that a different metalanguage is called for when spoken data is described. Similarly, ‘ellipsis’ is based on a notion of the absence of obligatory items, whereas face-to-face interaction proceeds unproblematically with often only minimal use of so-called ‘obligatory’ elements. Leech (2000) however cautions against assuming that the grammars of spoken and written language are radically different. He argues that spoken and written language utilise the same basic grammatical repertoire, though its implementation may differ. Speech, according to Leech, shows a tendency to simplified, loosely integrated and disjunctive construction (see Chafe 1982, Halliday 1989), giving grammatical structure a lesser role in the overall communication process than is characteristic of writing, something which can only be fully implemented by corpus research.

Alongside and emerging from grammatical research, studies of the spoken lexicon have suggested that the core, heavy-duty vocabulary of everyday spoken interaction is smaller than that of mainstream written texts, but that, importantly, the phenomenon of ‘chunking’ (i.e. recurrence of strings of two or more words) is more widespread in spoken data. Chunks, or lexical bundles (Biber/Conrad/Reppen 1999, McCarthy and Carter 2002) are also

different in kind across spoken and written corpora. While both types of corpora throw up common chunks characterised by syntactic fragments functioning as clause- or sentence frames (e.g. *I don't know if …, at a time when …*), there are notable differences between spoken and written data. Predominantly, the two-, three-, four- and five-word chunks found in written corpora tend to be prepositional phrases referring to basic notional categories such as time, place, manner, etc., or else determiner phrases (e.g. *one of the*), or adverbial phrases expressing various inter-clausal relations (e.g. *on the other hand*). Spoken chunks are dominated by interactional discourse marking expressions such as *you know what I mean* and vague expressions such as *or something like that* (McCarthy/Carter 2001b; McCarthy/Carter 2002; O'Keeffe 2004). The ubiquitous evidence of chunking in spoken corpora has contributed to debates on key aspects of language processing and the notion of fluency, not only in monolingual contexts (Schmitt and Carter, 2004) but also across languages (Spöttl/McCarthy 2004).

Grammatical and lexical studies based on spoken corpora have developed in tandem with studies of discoursal and pragmatic aspects of spoken language. Difficulties persist in areas such as the automatic coding and retrieval of features such as speech acts and figures

of speech, but, nonetheless, spoken corpora have been effectively exploited to investigate the reality of the everyday performance of common speech acts (Aijmer 1996), in contrast to the previous tradition within pragmatics of using intuitive data or elicitation instruments such as discourse completion tasks (DCTs). Aspects of turn-management have been investigated quantitatively by Tao (2003), and vocative address terms have been described, based on corpus evidence (Leech 1999; McCarthy/O'Keeffe 2003). McCarthy (2003) used the CANCODE corpus to investigate short listener responses (e.g. *right, fine, great, that's true*), a discourse feature in large part automatically retrievable by searching for single-word or very brief speaker turns. Meanwhile Aijmer (2002) used the LLC to examine 'discourse particles' (e.g. *now, oh, just, sort of, and that sort of thing, actually*), showing how the methods and tools of corpus analysis can sharpen their description. Aijmer illustrated the importance of linguistic and contextual cues such as text type, position in the discourse, prosody and collocation in the analysis of these items, hence the need to use a corpus which incorporated a detailed prosodic transcription system. Farr/O'Keeffe (2002) looked at the pragmatics of hedging in spoken Irish English. More diffuse but equally fundamental linguistic phenomena such as metaphor (Cameron/Deignan 2003), irony (Clift 1999), hyperbole (McCarthy/Carter 2004) and general conversational

creativity (Carter/McCarthy, 2004) have also been investigated and described using spoken corpora analysed through a combination of automatic retrieval of items (e.g. transcribed laughter, coded turn-overlaps, etc.) and manual searching, see O'Keeffe/McCarthy/Carter (2007) for specific examples.

5. Applications of spoken corpus research

Spoken corpora are increasingly used in diverse areas. These include forensic linguistics, for example in relation to forensic phonetics (e.g. speaker identification), the language of police confession, interrogation and deception (Shuy 1998), courtroom discourse (Cotterill 2002a,b, 2003, 2004). Boucher (2005), in his analysis of features of deceit in recounting, compared a corpus of 200 three-to-five minute discourses where half represented truthful and half inaccurate accounts. He was able to statistically describe significant differences in variables such as hesitation, lexical repetition and utterance length.

Given that corpora can be built around variables such as age, gender, level of education and socio-economic background, the area of sociolinguistics, not surprising, is one where there is increasing use for spoken corpora. For example, Ihalainen (1991a) looked at regional variation in verb patterns in south-western British English, while Ihalainen (1991b)

compared the grammatical subject in educated and dialectal English in the London-Lund and the Helsinki Corpus of modern English dialects. Kirk (1992, 1999) and Kallen/Kirk (2001) look at languages in contact in the context of Northern Ireland and Irish English, Ulster Scots, Irish and Scots Gaelic using a corpus-based approach.

Age-related research is prevalent especially in the context of teenager language. The Corpus of London Teenage Language (COLT) (see Haslerud/Stenström 1995; Stenström 1998) has provided the basis for numerous studies. Features such as discourse markers have been given particular attention, for example Andersen (1997a, 1997b) on the use of *like* in London teenage speech, Stenström (1995, 1997a) and Stenström/Andersen/Hasund (2002) on the use of tags and taboo language, Hasund (1998) on class-determined variation in the verbal disputes of London teenage girls, Hasund/Stenström (1997) on conflict talk using a corpus-based comparison of the verbal disputes of adolescent females. Other corpus-based studies on language and gender include Aijmer (1995) which looks at apologies, Holmes (2001) which examines linguistic sexism and Mondorf (2002), a study of gender differences in English syntax.

Lapidus/Otheguy (2005), in a 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.

In the second language pedagogical context, studies often illustrate how far the spoken language presented in textbooks for learners can be at odds with evidence from spoken corpora. Boxer/Pickering (1995), for example, looked at speech acts in textbook dialogues in comparison with real spontaneous encounters found in a corpus, while Carter (1998) found that textbook dialogues lacked core spoken language features such as discourse markers, vague language, ellipsis and hedges when compared to spoken corpus data (see also Gilmore 2004). Likewise, Hughes/McCarthy (1998) look at a range of grammatical items from the stock-in-trade of English as a Second Language pedagogy and argue that their distributions and functions in spoken language, based on corpus evidence are often different from those focused on in pedagogy.

Recent years have seen a debate over the use of native-speaker corpora versus learner corpora and non-native speaker corpora in the pedagogy of English as a second language (Prodromou 1997, 2003; Seidlhofer 2001; Gut 2006).

Written corpora tend to be more homogenous and usually include texts aimed at a very wide readership, whereas spoken corpora (especially informal conversational ones) inevitably reflect

very localised conditions and reflect the high levels of context-dependence and shared understandings typical of face-to-face speech. In the case of English, the issue is further complicated by the fact that the language has acquired the status of an international lingua franca, where users are not necessarily interested in modelling their talk on native speaker norms. There have, as a result, been arguments presented in favour of non-native, lingua franca spoken corpora. Prodromou (1997), arguing from the evidence of his mixed native- and non-native spoken English corpus of some 200,000 words, pointed to the potentially undermining effect of native-speaker English corpora on non-native-speakers faced with the many varieties and cultures of the target language as captured in the extant native-speaker corpora. Reacting to similar concerns, Seidlhofer proposed a spoken corpus of English as a Lingua Franca (ELF) to profile ELF as robust and independent of English as a native language with pedagogical applications (Seidlhofer 2001).

It is worth pointing out that many of the large spoken language corpora are collected not primarily for linguistic research but for speech technology projects. While English data dominates both types of spoken corpora, there is a growing number of non-English corpora. For example, Portuguese: *Português Falado - Documentos Autênticos: Gravações áudio com transcrição alinhada*

(Bacelar do nascimento 2001), which includes Portuguese varieties spoken in Portugal, Brazil, Goa and African countries; Italian: *Banca dati dell'italiano parlato,* which hosts the 490,000 word LIP corpus (Pusch 2002; Voghera 1996, Cresti, E. 2000); Basque: Basque Spoken Corpus, a collection of forty two narratives (Aske 1997); Spanish: The *Corpus Oral de Referencia del Español Contemporáneo* (Ballester/Santamaria/Marcos-Marin 1993), over one million words of spoken Spanish and *Corpus de Referencia del Español Actual*, a 133 million word corpus, 10% of which comprises spoken data (see http://corpus.rae.es/creanet.html); Czech: 800,000 words of spontaneous spoken language(•ermák 1997).

6. Directions in spoken corpus linguistics

At the present time, projects are underway to combine different media in the construction and exploitation of spoken corpora. Cauldwell (2002) combines sound files with on-screen textual displays of natural data, while the Kids' Audio Speech Corpus at the University of Colorado, Boulder, USA combines audio and video data with the aim of enabling the development of auditory and visual recognition systems (see http://cslr.colorado.edu/beginweb/reading/data_collection.html). The British Academic Spoken English (BASE),

assembled at the Universities of Warwick and Reading in Great Britain, under the directorship of Nesi and Thompson, is a companion corpus to MICASE (see above) (see Creer/Thompson 2004 for further details and see http://www.rdg.ac.uk/AcaDepts/ll/base_corpus/). The majority of the BASE recordings are on digital video. The corpus team also plans to edit and compress the video recordings, and to link transcripts and video/audio files on CD-ROM. The corpus construction aims to facilitate the analysis of features such as the pace, density and delivery styles of academic lectures and the discourse function of intonation. Alongside these, the Multimedia Adult ESL Learner Corpus (MAELC) at Portland State University, Portland, Oregon, USA promises for 2006 an audio and video corpus of some 5,000 hours of classroom interaction where transcripts, audio files and video clips will be available for research into second language acquisition (Reder/ Harris/Setzler 2003). Further developments in voice recognition may lead to effective automatic transcription of spoken data, and shortcomings in automatic tagging and parsing may be expected to be resolved as techniques advance, and as the need for spoken corpora increases with the extension of research and applications in areas such as voice recognition for the control of machine- and computer-processes, spoken databanks that are accessed automatically in service contexts such as tourism,

financial services, telecommunications, and so on.

7. Literature

Aijmer, K. (1995), Do women apologise more than men? In: Melchers, G. & Warren, B. (eds) *Studies in Anglistics*. Stockholm: Almqvist and Wiksell, 59-69.

Aijmer, K. (1996), *Conversational Routines in English.* London: Longman.

Aijmer, K. (2002), *English Discourse Particles - Evidence from a Corpus*. Amsterdam: John Benjamins.

Andersen, G. (1997a), They gave us these yeah, and they like wanna see like how we talk and all that' The use of *like* and other discourse markers in London teenage speech. In: Kotsinas, U.-B., Stenström, A.-B. & Karlsson, A.-M. (eds) *Ungdomsspråk i Norden*. Stockholm: MINS 43, 82-95.

Andersen, G. (1997b), 'They like wanna see like how we talk and all that'. The use of *like* as a discourse marker in London teenage speech. In: Ljung, M. (ed.) *Corpus-based studies in English*. Amsterdam: Rodopi, 37-48.

Aske, Jon. 1997. *Basque word order and disorder: Principles, variation, and prospects*. Ph. D. dissertation, Department of Linguistics, University of California, Berkeley

Aston, G. (2001), Text Categories and Corpus Users: a Response to David Dee. In: *Language Learning and Technology*. 5(3), 37-72 (available at http://llt.msu.edu/vol5num3/pdf/aston.pdf)

Atkins, S./Clear J./Ostler N. (1992), Corpus Design Criteria. In: *Literary and Linguistic Computing* 7(1), 1-16.

Bacelar do nascimento, F. (2001),*Português Falado, Documentos Autênticos, Gravações audio com transcrições alinhadas* (CD-ROM), Lisboa, Centro de Linguística da Universidade de Lisboa e Instituto Camões.

Ballester, A./Santamaria, C./Marcos-Marin, F. A. (1993), Transcription Conventions used for the Corpus of Spoken Contemporary. In: *Spanish Literary and Linguistic Computing* 8(4), 283-292.

Banjo, A. (1996), The Sociolinguistics of English in Nigeria and the ICE project. In Greenbaum S. (ed.) *Comparing English Word-Wide: The International Corpus of English.* Oxford: Oxford University Press, 239-248.

Barlow, M. (1998), *Corpus of Spoken Professional American English* (CSPAE, CD ROM, Athelstan: Houston.

Beier E./Starkweather J./Miller D.

(1967), Analysis of Word Frequencies in Spoken Language of Children. In: *Language and Speech* 10, 217-227.

Biber, D./Conrad S./Reppen R. (1998), *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Biber, D./Johansson S./Leech, G./Conrad S./Finegan E. (1999), *Longman Grammar of Spoken and Written English.* London: Longman.

Blanche-Benveniste, C. (1982), Examen de la Notion de Subordination. In : *Recherche sur le Français Parlé.* 4, 71-115.

Blanche-Benveniste, C. (1995), De la Rareté de Certains Phénomènes Syntaxiques en Français Parlé. In: *French Language Studies* 5(1), 17-29.

Bolton, K./Gisborne, N./Hung, J./Nelson, G. (2003), *The International Corpus of English Project in Hong Kong.* Amsterdam: John Benjamins.

Boucher, V. J. (2005), On the measurable linguistic correlates of deceit in recounting passed events. Paper presented to *International Association of Forensic Linguists 7th Biennial Conference on Forensic Linguistics/Language and Law*, Cardiff University, UK, 1st - 4th July 2005.

Boxer, D./Pickering L. (1995),

Problems in the Presentation of Speech Acts in ELT Materials: the Case of Complaints. In: *English Language Teaching Journal* 49, 99-158.

Breivik, L. E./Hasselgren A. (2002), From the Colt's Mouth ... And Others. Amsterdam: Rodopi.

Bucholtz, M. (2000), The Politics of Transcription. In: *Journal of Pragmatics* 32, 1439-1465.

Cameron, L./Deignan, A. (2003), Combining Large and Small Corpora to Investigate Tuning Devices around Metaphor in Spoken Language. In: Metaphor and Symbol. 18(3), 149-160.

Carter, R. (1998), Orders of Reality: CANCODE, Communication and Culture. In: English Language Teaching Journal 52, 43-56.

Carter, R.A./McCarthy, M.J. (1995), Grammar and the Spoken Language. In: *Applied Linguistics* 16(2), 141-158.

Carter, R.A./McCarthy, M.J. (2004), Talking, Creating: Interactional Language, Creativity and Context. In: *Applied Linguistics* 25(1), 62-88.

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

Carterette, E./Jones M.H. (1974),

Informal Speech. Berkeley and Los Angeles: University of California Press.

Cauldwell, R. (2002), *Streaming Speech: Listening and Pronunciation for Advanced Learners of English.* Birmingham: Speechinaction. CD-ROM, ISBN 0-9543447-0-7

•ermák, F. (1997), Czech National Corpus: A Case in Many Contexts. In: *International Journal of Corpus Linguistics* 2 (2), 181-197.

Chafe, W. (1982), Integration and Involvement in Speaking, Writing, and Oral Literature. In: Tannen D. (ed.) *Spoken and Written Language: Exploring Orality and Literacy*. Norwood, New Jersey: Ablex Publishing Corporation, 35-53.

Chafe W./Du Bois J./Thompson S. (1991), Towards a New Corpus of Spoken American English. In: Aijmer K. & Altenberg, B. (eds) *English Corpus Linguistics*. London: Longman, 64-82.

Cheng, W./Warren, M. (1999), Facilitating a Description of Intercultural Conversations: the Hong Kong Corpus of Conversational English. In: *ICAME Journal* 23, 5-20.

Cheng, W./Warren, M. (2000), The Hong Kong Corpus of Spoken English: Language Learning through Language Description. In: Burnard, L. & McEnery, T. *Rethinking Language Pedagogy from a Corpus*

Perspective. Frankfurt am Main: Peter Lang, 133-144.

Cheng, W./Warren, M. (2002), // ↘↗ beef ball // → you like //: The Intonation of Declarative-Mood Questions in a Corpus of Hong Kong English. In: *Teanga* 21, 151-165.

Clift, R. (1999), Irony in Conversation. In: *Language in Society* 28, 523-553.

Cook, G. (1990), Transcribing Infinity: Problems of Context Presentation. In: *Journal of Pragmatics* 14, 1-24.

Cotterill, J. (2003), *Language and Power in Court*. Basingstoke: Palgrave.

Cotterill, J. (ed.) (2002a), *Language in the Legal Process*. Basingstoke: Palgrave.

Cotterill, J. (2002b), *Language and Power in Court, a linguistic analysis of the O. J. Simpson trial*. Basingstoke: Palgrave.

Cotterill, J. (2004), Collocation, Connotation, and Courtroom Semantics: Lawyers' Control of Witness Testimony through Lexical Negotiation. In: *Applied Linguistics* 25 (4), 513-537.

Creer, S./Thompson, P. (2004), Processing Spoken Language Data: The BASE Experience. Workshop on Compiling

and Processing Spoken Language Corpora, 24th May, LREC 2004 (available at http://www.rdg.ac.uk/AcaDepts/ll/base_corpus/creer_thompson_final.pdf)

Cresti, E. (2000), *Corpus di italiano parlato.* Firenze: Accademia della Crusca.

Crowdy, S. (1993), Spoken Corpus Design. In: *Literary and Linguistic Computing* 8(2), 259-265.

Crowdy, S. (1994), Spoken Corpus Transcription. In: *Literary and Linguistic Computing* 9(1), 25-28.

Crystal, D. (1995), Refining Stylistic Discourse Categories. In: G. Melchers & Warren, B. (eds) *Studies in Anglistics.* Stockholm: Almqvist and Wiksell International, 35-46.

De Cock, S. (1998), A Recurrent Word Combination Approach to the Study of Formulae in the Speech of Native and Non-Native Speakers of English. In: *International Journal of Corpus Linguistics* 3, 59-80.

De Cock, S. (2000), Repetitive Phrasal Chunkiness and Advanced EFL Speech and Writing. In: Mair, C. & Hundt, M. (eds) *Corpus Linguistics and Linguistic Theory. Papers from the Twentieth International Conference on English Language Research on Computerized*

Corpora (ICAME 20), Freiburg im Breisgau 1999, Amsterdam: Rodopi 51-68.

Duranti, A. (1997), *Linguistic Anthropology*. Cambridge: Cambridge University Press.

Edwards , J. (1991), Transcription of Discourse. In: Bright, W. (ed.) *Oxford International Encyclopedia of Linguistics, Vol. 1.* Oxford: Oxford University Press. 367-371.

Edwards, J. A./Lampert, M. D. (eds) (1993), *Talking Data: Transcription and Coding in Discourse Research*. Hillsdale, New Jersey: Lawrence Erlbaum.

Fang, A.C. (1995), Distribution of Infinitives in Contemporary British English: a Study Based on the British ICE Corpus. In: *Literary and Linguistic Computing* 10(4), 247-57.

Farr, F. (2003), Engaged Listenership in Spoken Academic Discourse. In: *Journal of English for Academic Purposes* 2(1), 67-85.

Farr, F./O'Keeffe, A. (2002), *Would* as a Hedging Device in an Irish Context: an Intra-Varietal Comparison of Institutionalised Spoken Interaction. In: Reppen, R./Fitzmaurice S. & Biber, D. (eds), *Using Corpora to Explore*

Linguistic Variation, Amsterdam: John Benjamins, 25-48.

Farr, F./Murphy, B./O'Keeffe, A. (2002), The Limerick Corpus of Irish English: Design, Description and Application. In: *Teanga* 21, 5-29.

Fonseca-Greber, B./Waugh, L. R. (2003), On the Radical Difference between the Subject Personal Pronouns in Written and Spoken European French. In: Leistyna, P. & Meyer, C. (eds) *Corpus Analysis. Language Structure and Language Use*. Amsterdam: Rodopi, 225-240.

Francis, N. (1982), Problems of Assembling and Computerizing Large Corpora. In: Johansson, S. (ed.) *Computer Corpora in English Language Research*. Bergen: Norwegian Computing Centre for the Humanities, 7-24.

Gilmore A. (2004) A comparison of textbook and authentic interactions. In: *English Language Teaching Journal* 58(4), 363-374.

Granger, S./Hung, J./Petch-Tyson, S. (eds) (2002), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Amsterdam: John Benjamins.

Greenbaum, S. (1991), ICE: the International Corpus of English. In: *English Today* 28, 3-7.

Greenbaum, S. (1992), A New Corpus of English: ICE. In: Svartvik, J. (ed.) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82,* Stockholm 4-8 August 1991. Berlin: Mouton de Gruyter, 171-179.

Gut, U. (2006), Learner Speech Corpora in Language Teaching. In: Braun, S., Kohn, K. and Mukherjee, J. (eds.) *Corpus Technology and Language Pedagogy.* Frankfurt: Lang, 69-86.

Halliday, MAK (1989), *Spoken and Written Language*. Oxford: Oxford University Press.

Handford. M./McCarthy, M. J. (2004), 'Invisible to Us': A Preliminary Corpus-Based Study of Spoken Business English. In: Connor, U. & Upton, T. (eds) *Discourse in the Professions: Perspectives from Corpus Linguistics.* Amsterdam: Benjamins, 167-201.

Haslerud, V./Stenström, A-B. (1995), The Bergen Corpus of London Teenager Language (COLT). In: Leech, G., Myers, G. & Thomas, J. (eds) *Spoken English on Computer*. London: Longman, 235-242.

Hasund, K. (1998), From woman's place to women's places: class-determined variation in the verbal disputes of London teenage girls. In: Despard, A. (ed.) *A Woman's Place: Women, Domesticity and Private Life.* Kristiansand: Norwegian Academic Press, 187-199.

Hepburn, A. (2004), Crying: Notes on Description, Transcription, and Interaction. In: *Research on Language and Social Interaction* 37(3), 251–290.

Holmes, J. (1996), The New Zealand Spoken Component of ICE: Some Methodological Challenges. In Greenbaum, S. (ed.) *Comparing English World-Wide: The International Corpus of English.* Oxford: Oxford University Press, 163–178.

Holmes, J. (2001), Ladies and gentlemen: corpus analysis and linguistic sexism. In: Mair, C. & Hundt, M. (eds). *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi, 141–156.

Holmes, J./Vine, B./Johnson, G. (1998), *Guide to the Wellington Corpus of Spoken New Zealand English.* Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.

Howes, D.H. (1966), A Word Count of Spoken English. In: *Journal of Verbal Learning and Verbal Behaviour* 5, 572–606.

Hughes, R./McCarthy, M.J. (1998), From Sentence to Discourse: Discourse Grammar and English Language Teaching. In: *TESOL Quarterly* 32, 263–287.

Ide, N./Macleod, C. (2001), The

American National Corpus: A Standardized Resource of American English. In: Rayson, P./Wilson, A./McEnery, T./Hardie, A. & Khoja, S. (eds) *Proceedings of Corpus Linguistics 2001,* Vol 13, Lancaster: University of Lancaster..

Ide, N./Reppen, R./Suderman, K. (2002), The American National Corpus: More than the Web Can Provide . In: *Proceedings of the Third Language Resources and Evaluation Conference (LREC), Las Palmas, Canary Islands, Spain, 2002*, 839-44. Available at: http://americannationalcorpus.org/pubs.html.

Ihalainen, O. (1991a), A point of verb syntax in south-western British English: an analysis of a dialect continuum. In: Aijmer, K. & Altenberg, B. (eds) *English Corpus Linguistics*. London: Longman, 290-302.

Ihalainen, O. (1991b), The grammatical subject in educated and dialectal English: comparing the London-Lund Corpus and the Helsinki Corpus of modern English dialects. In: Johansson, S. & Stenström, A.-B. (eds) *English Computer Corpora: Selected Papers and Research Guide*. Berlin: Mouton de Gruyter, 201-214.

Jefferson, G. (1985), An Exercise in the Transcription and Analysis of Laughter. In: Van Dijk, T. (ed.) *Handbook of Discourse Analysis* (Vol. 3).

London: Academic Press, 25–34.

Kirk, J. M. (1992), The Northern Ireland Transcribed Corpus of Speech. In: Leitner, G. (ed.) *New Directions in English Language Corpora*. Berlin: Mouton de Gruyter, 65–73

Kirk, J. M. (1999), The Dialect Vocabulary of Ulster. In: *Cuadernos de Filología Inglesa* 8, 305–334.

Kallen, J.L./Kirk, J.M. (2001), Convergence and Divergence in the Verb Phrase in Irish Standard English: a Corpus-Based Approach. In: Kirk, J. M. & Ó Baoill, D. P. (eds) *Language Links: the Languages of Scotland and Ireland*. Belfast: Cló Ollscoil na Banríona, 59–79.

Knowles, G. (1990), The Use of Spoken and Written Corpora in the Teaching of Language and Linguistics. In: *Literary and Linguistic Computing* 5(1), 45–8.

Lapidus, N./Otheguy, R. (2005), Contact Induced Change? Overt Nonspecific *Ellos* in Spanish in New York. In: Sayahi, L. & Westmoreland, M. (eds) *Selected Proceedings of the Second Workshop on Spanish Sociolinguistics*. Somerville, MA: Cascadilla Proceedings Project, 67–75. Available at http://www.lingref.com/cpp/wss/2/paper1141.pdf.

Lee, D. (2001), Genres, Registers, Text Types, Domains, and Styles:

Clarifying the Concepts and Navigating a Path through The BNC Jungle. In: *Language Learning & Technology* 5(3), 37-72 (available at http://llt.msu.edu/vol5num3/lee/)

Leech, G. (2000), Grammars of Spoken English: New Outcomes of Corpus-Oriented Research. In: *Language Learning* 50(4), 675-724.

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

McCarthy, M.J. (2001), *Issues in Applied Linguistics.* Cambridge: Cambridge University Press.

McCarthy, M.J. (2003), Talking Back: 'Small' Interactional Response Tokens in Everyday Conversation. In: *Research on Language in Social Interaction* 36(1), 33-63.

McCarthy, M.J./Carter, R.A. (2001a), Size Isn't Everything: Spoken English, Corpus and the Classroom. In: *TESOL Quarterly* 35(2), 337-340.

McCarthy, M.J./Carter, R.A. (2001b), Ten Criteria for a Spoken Grammar. In: Hinkel E. & Fotos S. (eds) *New Perspectives on Grammar Teaching in Second Language Classrooms*. Mahwah, New Jersey: Lawrence Erlbaum Associates, 51-75.

McCarthy, M.J./Carter, R.A. (2002), *This That and The Other*: Multi-word Clusters in Spoken English as Visible Patterns of Interaction. In: *Teanga* 21, 30-52.

McCarthy, M.J./Carter, R.A. (2004), 'There's Millions of Them': Hyperbole in Everyday Conversation. In: *Journal of Pragmatics* 36, 149-184.

McCarthy, M.J./O'Keeffe, A. (2003), 'What's in a name?' - Vocatives in Casual Conversations and Radio Phone-in Calls. In: Leistyna, P. & Meyer, C. (eds) *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi, 153-185.

Mondorf, B. (2002), Gender differences in English syntax. In: *Journal of English Linguistics*. 30 (2), 158-180.

Moon, R. (1997), Vocabulary Connections: Multi-Word Items in English. In: Schmitt, N. & McCarthy, M.J. (eds) *Second Language Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press, 40-63.

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

Nelson, G./Wallis, S./Aarts, B. (2002), *Exploring Natural Language:*

Working with the British Component of the International Corpus of English. Amsterdam: John Benjamins.

Nero, S.J. (2000), The Changing Faces of English: a Caribbean Perspective. In: *TESOL Quarterly* 34(3), 483–510.

Ochs, E. (1979), Transcription as Theory. In: Ochs, E. & Schieffelin, B. B. (eds) *Developmental Pragmatics.* New York: Academic Press, 43-72.

O'Keeffe, A. (2004), 'Like the Wise Virgins and All that Jazz': Using a Corpus to Examine Vague Categorisation and Shared Knowledge. In: *Language and Computers* 52(1), 1-20.

O'Keeffe, A./Farr, F. (2003), Using Language Corpora in Language Teacher Education: Pedagogic, Linguistic and Cultural Insights. In: *TESOL Quarterly* 37(3), 389-418.

O'Keeffe, A./ McCarthy, M.J./ Carter, R. (2007), *From Corpus to Classroom.* Cambridge: Cambridge University Press.

Ooi, V. (1997), Analysing the Singapore ICE Corpus for Lexicographic Evidence. In: Ljung, M. (ed) *Corpus-Based Studies in English.* Amsterdam: Rodopi, 245–260.

Prodromou, L. (1997), Global English and Its Struggle against the Octopus.

In: *IATEFL Newsletter 135*, 12-14.

Prodromou, L. (2003), In Search of the Successful User of English. In: *Modern English Teacher* 12, 5-14.

Pusch, C.D. (2002), A survey of spoken language corpora in Romance. In: Pusch, C.D.and Raible, W. (eds) *Romanistische Korpuslinguistik.*Tübingen: Narr, 245-264.

Reder, S./Harris, K./Setzler, K. (2003), The Multimedia Adult ESL Learner Corpus. In: *TESOL Quarterly* 37(3), 546-557.

Rundell, M. (1995a), The BNC: A Spoken Corpus. In: *Modern English Teacher* 4(2), 13-15.

Rundell, M. (1995b), The Word on the Street. In: *English Today* 11(3), 29-35.

Schmied, J./Hudson-Ettle, D. (1996), Analysing the Style of East African Newspapers in English. In: *World Englishes* 15(1), 103-113.

Schmitt, N./Carter, R.A. (2004), Formulaic Sequences in Action: An Introduction. In: Schmitt, N. (ed.) *Formulaic Sequences*. Amsterdam: John Benjamins, 1-22.

Schonell, F./Meddleton, I./Shaw, B./Routh, M/Popham, D./Gill, G./Mackrell, G./ Stephens, C. (1956), *A*

Study of the Oral Vocabulary of Adults. Brisbane and London: University of Queensland Press/University of London Press.

Seidlhofer, B. (2001), Closing a Conceptual Gap: the Case for a Description of English as a Lingua Franca. In: *International Journal of Applied Linguistics* 11, 133-158.

Shuy, R (1998), *The Language of Confession, Interrogation and Deception.* London: Sage.

Simpson, R.C./Lucka, B./Ovens, J. (2000), Methodological Challenges of Planning a Spoken Corpus with Pedagogical Outcomes. In: Burnard, L. & McEnery, T. (eds) *Rethinking Language Pedagogy from a Corpus Perspective*: P*apers from the Third International Conference on Teaching and Language Corpora* (TALC). Frankfurt: Peter Lang, 43-49.

Sinclair, J. (1995), Corpus Typology- a Framework for Classification. In Melchers, G. & Warren, B. (eds) *Studies in Anglistics.* Stockholm: Almqvist and Wiksell International, 17-34.

Spöttl, C./McCarthy, M.J. (2004), Comparing Knowledge of Formulaic Sequences across L1, L2, L3, and L4. In: Schmitt, N. (ed.) *Formulaic Sequences*. Amsterdam: John Benjamins, 191-225.

Stenström, A.-B. (1995), Taboos in teenage talk. In: Melchers, G. and Warren, B. (eds) *Studies in Anglistics*. Stockholm: Almqvist and Wiksell International, 71-80.

Stenström, A.-B. (1997a), Tags in Teenage Talk. In: Fries, U., Müller, V. & Schneider, P. (eds) *From Ælfric to the New York Times. Studies in English Corpus Linguistics*. Amsterdam: Rodopi, 139-148.

Stenström, A.-B. (1997b), 'Can I have a chips please? - Just tell me what one you want' Nonstandard grammatical features in London teenage talk. In: Aarts, J., de Mönninck, I. & Wekker, H. (eds). *Studies in English Language and Teaching*. Amsterdam: Rodopi, 141-152.

Stenström, A.-B. (1998), From sentence to discourse: *cos(because)* in teenage talk. In: Jucker, A. & Ziv, Y. (eds). *Discourse Markers: Descriptions and Theory*. Amsterdam: John Benjamins, 127-146.

Stenström, A.-B./Andersen, G./Hasund, I.K. (2002), *Trends in Teenage Talk*. Amsterdam: John Benjamins.

Stern, K. (1997), The Longman Spoken American Corpus: Providing an In-Depth Analysis of Everyday English. In: *Longman Language Review* 3, 14-17.

Svartvik, J. (ed) (1990), *The London-Lund Corpus of Spoken English:*

Description and Research. Lund: Lund University Press.

Svartvik, J./Quirk, R. (1980), *A Corpus of English Conversation*. Lund: Liberläromedel.

Svartvik, J. (ed) (1990), *The London Corpus of Spoken English: Description and Research.* Lund Studies in English 82. Lund: Lund University Press.

Tao, H. (2003), Turn Initiators in Spoken English: a Corpus-Based Approach to Interaction and Grammar. In: Leistyna P. & Meyer, C. (eds) *Corpus Analysis: Language Structure and Language Use.* Amsterdam: Rodopi, 187-207.

Tognini Bonelli, E. (2001), *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Voghera, M. (1996), Corpora dell'italiano. In : *Revue Française de* Linguistique Appliquée (1), 131–134.