Corpus types and uses

CORPUS TYPES AND USES

ABSTRACT

This chapter provides a broad overview of corpus types and uses. It surveys five types of corpora: General, Parallel, Historical, Multimodal and Specialised. In each section, we provide a description of the corpus type, the key issues associated with the type as well as its applications in pedagogical contexts. The overview is not meant to be exhaustive as there are many more corpora than we have space to mention. However, our aim is to introduce the main types and uses so that readers may then seek to explore the types themselves more fully depending on their interests (see appendix for further information).

1 GENERAL CORPORA

General corpora, or reference corpora, can be spoken, written or both, and aim 'to provide information about language as a whole, showing how it is generally used in speech and writing of various kinds' (Kübler and Aston 2010: 504). Baker (2010: 12) suggests that this 'could be seen as a prototypical corpus in that it is normally very large, consisting of millions of words, and texts collected from a wide range of sources representing many language contexts'. There are three generations of general corpora, the first of which is represented by the Brown family. The BROWN corpus is a one million-word collection of written American English from 1961 (Kucěra and Francis 1967). Its British counterpart, the London-Oslo/Bergen (LOB) corpus is one million words of written texts also collected in 1961 (Johansson, Leech and Goodluck 1978). Both BROWN and LOB are otherwise known as synchronic corpora, as their texts stem from one period of time. Some years later, two new corpora joined this family; the

Freiberg Brown Corpus of American English (FROWN), consisting of one million words from the 1990s (Hundt, Sand and Skandera 1999), and the Freiberg London-Oslo/Bergen (FLOB) corpus of one million words of British English from the 1990s (Hundt, Sand and Siemund 1998). Another example of this first generation is the International Corpus of English (ICE), which includes a number of one million-word corpora collected from 1990-94 in countries where English is a first or official language (Nelson 1996).

The second generation of corpora grew in size and an example is the British National Corpus (BNC), a 100 million-word corpus of spoken and written English (Aston and Burnard 1998). The American National Corpus (ANC) is designed on the same principle (Reppen and Ide 2004), but with two differences, namely that data from the ANC stems from 1990 onwards, whereas data in the BNC is from 1960-93, and there are newer text types in the ANC such as blogs and web pages (Reppen 2009). Another corpus based on the design of the BNC is the Turkish National Corpus (TNC) of 50 million words (Askan *et al*. 2012). Further examples of approximately 100-million word corpora include the Corpus di Italiano Scritto (CORIS) (Rossini Favretti *et al.* 2004), the Corpus del Español (Davies 2002) and the Russian Reference Corpus (Sharoff 2004). The third generation of general corpora are even bigger in size, for instance, the Bulgarian National Corpus contains 1.2 billion words (Koeva *et al.* 2012). Also, the Bank of English (BoE), which emerged as part of the COBUILD project (Sinclair 1987), contains 650 million words of spoken and written texts, and is constantly being updated, making it a monitor corpus (Clear 1987), in that texts are continuously added to the corpus and changes can be tracked using software. Another example is the Corpus of Contemporary American English (COCA), the largest online freely-available spoken and written corpus at 450 million words collected since 1990 (Davies 2010).

These corpora have a number of applications, for example, they can be used to offer valuable information about how language, or a variety of language can be used, or they can be used as a reference for comparison purposes. For instance, linguistic analyses have included the examination of collocation in BROWN (Kjellmer 1994), and modality in the Brown family (Degani 2009). Furthermore, different varieties of one language have been examined in ICE (Hundt and Gut 2012). Studies of a more sociolinguistic nature include exploring lexical change using BROWN and the ANC (Fengxiang 2012), taboo language in the BNC (McEnery and Xiao 2004), and gender differences in specialised corpora and the BNC (Schmid 2003; see Baker 2010 for more about corpora in sociolinguistics). Pragmatics has also been examined, for example the use of apologies in the BNC (Deutschmann 2003) and laughter in the BNC and other corpora (Partington 2006; see also Caines *et al.*, this volume). General corpora are also increasingly being used for lexicography (see Hanks 2009). Another application is in the field of language teaching and learning (see O'Keeffe, McCarthy, and Carter 2007; Reppen 2009; 2010a). Using large corpora, teachers can, for example, study specific linguistic items rather than using their intuition (Sinclair 1997), students can use corpora for data-driven learning (Johns 1991; see also Warren Chapter 24, this volume), for access to authentic language (Aston 1995), and as a source of reference (Chambers 2005). Large corpora can also support the creation of text books (McCarten 2010), or grammar books (Biber *et al.* 1999; Carter and McCarthy 2006; see also this volume

Caines *et al*., Chapter 25 on spoken corpora; Chambers, Chapter 26 on written corpora; and Chapter 6 for courseware design using other digital resources).

Although the uses are plentiful, a major issue in corpus linguistics is the ability for users to interpret the findings (see O'Keeffe and Farr 2003). As general corpora are large, users must accustom themselves to working with an abundance of data, which requires skills in which the user may need training (see O'Keeffe and Farr 2003; Sinclair 2003). When comparing a smaller corpus with a general corpus, it must also be acknowledged that different sized corpora are not comparable, and therefore, in order to draw conclusions, the rule of thumb is to calculate figures in words per million. Also, the size of a corpus is important when considering the focus of the investigation, for example, while corpora of a million words are useful for grammatical co-occurrence patterns, they might not be useful for lexical studies (Reppen 2010a; see also Chapter 34, this volume for details on the use of CALL for lexico-grammatical acquisition).

2 SPECIALISED CORPORA

In contrast to general corpora, a specialised corpus is more restricted and may be regarded as 'specialised' if it involves any or all of the following criteria as outlined by Flowerdew (2004: 21): a) it has been compiled for a specific purpose (for example, to investigate a particular item; b) it represents a particular context (for example, setting, participants and communicative purpose); c) it represents a genre (for example, sales letters); d) it includes a particular type of text/discourse (for example, biology textbooks); e) it represents a subject matter/topic (for example, economics); and/or f) it represents a variety of English (for example, Learner English). Corpora which have emerged so far and can be classified as specialised emanate from various contexts such as:

- *Education*: the International Corpus of Learner English (ICLE; see Meunier, Chapter 27, this volume), the Michigan Corpus of Academic Spoken English (MICASE) (Simpson-Vlach and Leicher 2006), the British Academic Written English (BAWE) corpus (Nesi 2012); English as a Lingua Franca in Academic settings (ELFA) (Mauranen 2012); the Michigan Corpus of Upper-level Student Papers (MICUSP) (O'Donnell and Römer 2012);
- *Business*: the Cambridge and Nottingham Business English Corpus (CANBEC) (Handford 2010);
- *Law*: the Cambridge Corpus of Legal English (CCLE);
- *Professional English*: the Corpus of Spoken Professional American English (CSPAE) (Yaguchi *et al*. 2004);
- *Society*: the Corpus of London Teenage English (COLT) (Stenström, Andersen, and Hasund 2002).
- *Internet*: the internet has also been used as a specialised corpus (see Renouf 2002) and as a source for building specialised corpora (Hundt, Nesselhauf and Biewer 2009: 1-7; see also Kilgarriff 2001; Kilgarriff and Grefenstette 2003 for more on the use of the web as a corpus). Online corpora such as the Enron email corpus and the Cambridge and Nottingham E-Language Corpus (CANELC) also exist.

Although specialised corpora are normally smaller than general corpora precisely because of their narrower focus (Lee 2010: 114), they have been criticised because of their size (Sinclair 2004). However, research has shown that they can yield reliable results when investigating high frequency items and that a corpus does not always need to consist of millions of words and a large number of texts (Biber 1990). The message is clear that while small corpora are not suitable for all types of studies (Koester 2010: 77), they do have advantages over larger corpora. For instance, they are not de-contextualised and as a result, allow the researcher to explore a much closer link between the corpus and the contexts in which the texts are produced (Koester 2010: 74; O'Keeffe 2007). The size of the corpus means that each occurrence of a particular form can be explored, and not just a random sample, which is common when working with general corpora. They also provide insights into patterns of language use in particular settings and as the corpus compiler is often the analyst, they usually have a high degree of familiarity with the context which assists the interpretation of the data, in a way that is not often possible when dealing with larger corpora (see Koester 2010; Handford 2010). However, it is worth noting that not all specialised corpora have to be small and indeed as highlighted by O'Keeffe *et al*. (2007), a specialised corpus can be defined as large if it contains a million words or more. Handford (2010: 258) lists CANBEC as one such example and another is the 1.9 billion word Corpus of Global Web-based English (GloWbE), compiled by Davies (2013).

In a language teaching and learning context (see Warren, this volume), Tribble (2002) argues for the use of small specialised corpora to inform pedagogy (Johns 1991; Flowerdew 2004; Reppen 2010a). He claims that large corpora do not meet the needs of teachers and learners in ESP/EAP, for instance, as they either provide 'too much data across too large a spectrum or too little focused data to be directly helpful with EAP'

(2002: 132). Smaller corpora, on the other hand, yield more insights which are directly relevant for teaching and learning (Flowerdew 2004). Aston (1997) highlights that small specialised corpora are not only a valuable asset in their own right as a means of discovering the characteristics of a particular area of language but also useful in helping and training students to use bigger corpora more appropriately. Reppen (2010b) highlights that when used in a teaching and learning context, specialised corpora can help to identify unfamiliar/high frequency words, provide concordance lines from which to develop class activities, identify word senses and practise inferencing strategies. Reppen shows how she used a small specialised corpus in her own teaching context by collecting a set of class papers from an elementary writing group. The writing was coded for three types of errors: noun morphology, verb morphology, and subject/verb agreement. Reppen then used the corpus to generate a list of errors to inform instruction and as a source of classroom activities (see also Reppen, Chapter 29, this volume). The challenges, however, of using small corpora in the classroom have not gone unnoticed. Gavioli (2002), for example, highlights the practical difficulties of balancing the materials provided to students which, on the one hand, need to be limited and controlled for teaching, but on the other need to be plentiful in order to allow the students enough data to work on for the facilitation of confident linguistic hypotheses. She claims that particular teaching/learning needs may not always align with practical issues in an ESP context (see Flowerdew 2009 for a more critical account of corpora in ESP; Gavioli 2005).

3 PARALLEL CORPORA

While a monolingual corpus contains one language, a multilingual corpus contains two or more languages, and the latter can be divided into two categories, parallel and comparable. Parallel corpora are designed based on the relationship of translation between texts, thus having an original group of texts and translations of those texts (Tognini-Bonelli and Sinclair 2006). A comparable corpus does not contain translations of texts, but rather texts collected in a number of languages, and based on the same communicative function (Kenning 2010), much like the BACKBONE corpus discussed later. The first parallel corpus was the Canadian Hansard Corpus (Tognini-Bonelli and Sinclair 2006), which consists of government documents in English and Canadian French. One of the best known parallel corpora is The English-Norwegian Parallel Corpus (ENPC), containing original and translated texts in both languages (2.6 million words), therefore making it bidirectional (Johansson, Ebeling and Oksefjell 2002). The Oslo Multilingual Corpus (OMC) is an extension of the ENPC, including English, Norwegian, German, Finnish, Swedish, Dutch, French, and Portuguese texts (OMC 2010). Based on a similar design, the English-Swedish Parallel Corpus (ESPC) consists of 2.8 million words of bidirectional English and Swedish texts (Altenberg, Aijmer and Svensson 2001). An online freely-available parallel corpus is the Open Parallel Corpus (OPUS), which is a growing collection of translated texts from the web (Tiedemann 2012).

One thing that sets parallel corpora aside from other corpora is that they have bilingual concordances, where all occurrences of a search word in both languages are found and presented alongside each other. This concordancer ‘trawls thorough all the parts of a parallel corpus, retrieving not only all the occurrences of the search item in context, but also the sentences that contain the corresponding segments in the other language/languages' (Kenning 2010: 491). The applications of parallel corpora are varied, and Bowker and Pearson (2002) categorise their users into three domains. Firstly, language teachers and learners can use parallel corpora as a dictionary which offers multiple examples of context, and to examine how words are translated across languages. Students can also analyse specific language features across languages, or identify how cultural references are dealt with during translation (Bowker and Pearson 2002). The second group of users are translators and translation students. They can use parallel corpora for the same reasons as above, but also to examine what happens during a translation (Bowker and Pearson 2002), assisting with both practical and research-based translation (Kenning 2010). It has been suggested that 'each translator's dream is a resource which instantly provides reliable candidate translations, and this is what a parallel corpus ideally offers' (Kübler and Aston 2010: 510; see also Chapter 39, this volume for other translation technologies). Baker (2000) examines individual translators' styles in the Translational English Corpus (TEC – two million words at the time of her analysis), and Xiao and Yue (2009) examined some translation universals in a 200,000-word sample of the Lancaster Corpus of Mandarin Chinese (LCMC) and the one million-word Contemporary Chinese Translated Fiction Corpus (CCTFC) (for more on parallel corpora for translation studies, see Véronis 2000 and Xiao and Yue 2009). The third group of users are computational linguists, who can use parallel corpora to test alignment software, and to give further insights into machine translation (Bowker and Pearson 2002; see Caines and Buttery 2010 for more on training computers in NLP). Of course, lexicographers also use parallel corpora for bilingual lexicography, and contrastive linguists use them to describe a given language, and explore the similarities and differences between languages (Kenning 2010).

Issues to consider with parallel corpora include the fact that one needs pairs of texts in two or more languages for the creation of a corpus, and multilingual texts are harder to find that monolingual ones. The web helps in that many texts are now in electronic format, therefore the user does not have to scan the texts to be exploited for analysis. Texts, however, need to be pre-processed to prepare them for alignment, which is the creation of links between texts so they can be used for later investigations (see Bowker and Pearson 2002 and Kenning 2010). Lastly, there are not a lot of publically available parallel corpora because of the complexity in getting permission to use a text and its translation (Kübler and Aston 2010).

4 HISTORICAL CORPORA

The earliest historical electronic resources emerged in the 1980s with the Dictionary of Old English database prepared in Toronto and the Augustan Prose Sample and the Century of Prose Corpus compiled at Cleveland State University (Rissanen 2000: 7). Since that time, corpus linguistics has continued to make its mark on the history of English through the growing number of historical corpora representing various periods, genres, dialects, registers and social strata of English (Kytö 2012). Claridge (2008: 242) defines a historical corpus as one which has been intentionally created to represent past stages of a language and/or to study language change. Developments in historical corpus linguistics have been loosely grouped into four categories (see Rissanen 2000: 8-13): (i) multi-purpose corpora e.g. the widely-known c. 1.5 million-word Helsinki

Corpus (c. 730-1710) and the c. 1.7 million-word ARCHER corpus (c. 1650-1900) (Kytö and Pahta 2012; Yáñez Bouza 2011), which together extend over several centuries and a wide range of genres; (ii) Old and Middle English: general and author based corpora e.g. the c. 3.5 million-word Toronto Dictionary of Old English Corpus in Electronic Form, which consists of practically all extant Old English writings (with the exception of some parallel manuscripts) (Healey 1999); (iii) Middle and Modern English: genre and regional varieties corpora including the c. 2.6 million-word Corpus of Early English Correspondence (CEEC) (Nevalainen and Raumolin-Brunberg 1996) from 1417-1681 and the c. 1.5 million-word Corpus of Early English Medical Writing (CEEM) spanning 1375 to 1750 (Taavitsainen and Pahta 2010); (iv) Renaissance and Twentieth Century English such as the Lampeter Corpus (see Kytö and Pahta 2012: 128-131; Claridge 2000). While 'long and thin' corpora (Rissanen 2000: 10) such as the Helsinki Corpus have been the norm, advances in historical corpus linguistics have witnessed the emergence of much larger corpora such as the 400 million-word Corpus of Historical American English (COHA) (Davies 2012), which when added to Rissanen's (2000) list marks a period of movement in the approach taken to the compilation of historical corpora. Its online accessibility and availability means that it will have a considerable impact on research in the area of historical corpus linguistics and in some way provides an insight into its future.

The history of English has been revolutionised by corpus linguistics (Lee 2010: 113-14) and indeed, Rissanen (2012) claims that if it had not been for corpus linguistics, evidence-based historical linguistics might not have survived, let alone experience the Renaissance it did (Kytö 2012: 3). Its merits include the fact that corpus linguistics has provided researchers with the tools to collect, sort and analyse large quantities of data with speed and accuracy (Rissanen 2012) (see also section on Specialised Corpora in this chapter). Corpus methods have also helped to eliminate the idea of fragmentation which often occurs in historical linguistics and have facilitated the replicability and accuracy of linguistic results (see Kyto and Pahta 2012). However, the literature has also highlighted the challenges involved in historical corpus linguistics (see Claridge 2008). For example, the transference of text from handwritten or printed into computerised format presents an edited truth of the language used in the original, and means that the nature of the editorial process and involvement of researchers' time is crucial for the reliability of the corpus data. In terms of sampling, there is a clear imbalance of gender representation with most texts being produced by men as women did not have opportunities for formal education to the extent that men had up until the 1800s or later. Also, very few texts have been preserved from representatives of the less educated social classes (see Rissanen 2008). In addition, as corpora often span several centuries, the definition of genre for certain periods does not always hold true for others and this gives rise to difficulties in corpus compilation, which require careful consideration (see Rissanen 2008). Therefore, like all other corpus linguists, scholars of historical corpus linguistics need to be especially aware of how the corpora have been compiled, how they can be used and what their limitations are.

In language teaching and learning contexts, scholars such as Curzan (2008) have discussed how historical corpus linguistics has been incorporated into pedagogy. Corpora such as COHA mean that students have immediate access to data which act as rich sources of linguistic evidence and the time previously spent tracking down and collecting data has been considerably reduced. Students can pursue their own questions about language and linguistic change and engage more interactively and holistically than before with historical change across morphological, syntactic, semantic and orthographic levels as well as different varieties and registers (see also Biber, Conrad and Reppen 1998). Brinton (2012) also highlights the potential for pragmatic and discourse-based analyses of the history of English (see Culpeper 2010). However, Rissanen (2008: 65; see 2012) highlights the need to fully understand the language form studied and the main characteristics of the literary, political, social, geographical and cultural background from which the texts arise. Otherwise, he claims that fatal misinterpretations of textual evidence may take place. A more recent shift in corpus linguistics is the development of multimodal corpora, outlined below.

5 MULTIMODAL CORPORA

A multimodal corpus has 'transcripts that are aligned or synchronised with the original audio or visual recordings' (Lee 2010: 114). This type of corpus, while still in its infancy (Knight 2011), involves both textual and non-textual data. It has been acknowledged that one shortcoming of spoken corpora is that they lack visual representations by showing speech in textual format (Knight and Adolphs 2007; Knight and Tennent 2008), which the multimodal approach is attempting to tackle, by depicting communication in its 'entire complexity' (Blache *et al.* 2009: 38). One example is the Nottingham Multimodal Corpus (NMMC), a 250,000-word corpus with recordings and transcriptions collected from single speaker and dyadic conversations in an academic context (Knight *et al.* 2008). Another is the SACODEYL corpus, which includes transcribed interviews with British, German, French, Italian Spanish, Lithuanian, and

Romanian adolescents between 13 and 18 years of age (Hoffstaedter and Kohn 2009). Each language contains 20 to 25 video-recorded interviews, which have been transcribed and stored as corpora, and then thematically and linguistically annotated (Hoffstaedter and Kohn 2009). The annotated sections are time-stamped so there is synchronisation between the transcripts and the accompanying audio files (Widmann, Kohn, and Ziai 2011). A corpus based on the same premise is the BACKBONE corpus, which contains data collected from adults who speak regional varieties of languages, as well as lesser taught languages (British-English, Irish-English, German, French, Spanish, Turkish, Polish and manifestations of English as a Lingua Franca). The Santa Barbara Corpus of Spoken American English (249,000 words) can also be read online with the transcripts and audio files synchronised (for more multimodal corpora see Knight 2011).

Multimodal corpora move beyond the field of traditional corpus linguistics because they are 'potentially useful to many other fields in linguistics, including pragmatics, conversation analysis, discourse analysis, sociolinguistics, as well as language technologists working on speech recognition, audio (visual) file search technologies, and in some cases, natural language processing' (Haugh 2009: 76). For example, verbal and non-verbal behaviour can be examined (Knight and Tennent 2008), head, eye, hand and body movements can be analysed (Allwood 2009), as well as lexical, prosodic and gestural features (Knight 2011; Allwood 2009; Blache *et al.* 2009), thus facilitating a deeper understanding of context (Adolphs *et al.* 2011). For instance, Knight and Adolphs (2007) studied head-nod behaviour and verbal backchannels on a sub-set of NMMC, and Dahlmann and Adolphs (2009) analysed the relationship between the multi-word unit *I think* and pauses in the English Native Speaker Interview Corpus (ENSIC). As well as this, Allwood (2009) notes that multimodal corpora can be used to examine, and in turn improve, any kind of communicative behaviour such as, presentation techniques, teaching-related communication, and doctor-patient communication.

Shortcomings of multimodal corpora include the fact that they are not generally available (although SACODEYL and BACKBONE are), many are in the thousands of word size (compared to the vast general corpora mentioned earlier), and some are not yet transcribed (Knight 2011). Technical issues also need consideration, for example the data needs to be collected and transcribed, which is much more time-consuming and expensive than what is involved in compiling other corpora. Furthermore, a timeline is required to align text with speech, power is needed for the algorithm to show gestures, and the storage required for the files is very large (Knight and Tennent 2008). Moreover gestures need to be coded, which is a complex process (Knight 2011; see also Blache *et al.* 2009 for more on annotation). Therefore, while some considerations are often similar to other types of corpus compilation such as what to record, how to record, storing recordings, transcribing recordings, storing/saving transcriptions, what should be analysed and how can it be done (Allwood 2009; see also Reppen, Chapter 29, this volume), for multimodal corpora, technical issues regarding recording, lighting, placing, and type of equipment need further attention. However, this type of corpus is a significant move in corpus linguistics, and over time the limitations should be reduced (see Chapters 5 and 37, this volume for more multimodal technologies). The remainder of this chapter discusses the future implications of the types of corpora we have outlined.

6 FUTURE DIRECTIONS

Many types of corpora have emerged and the trend looks set to continue largely because research within the corpus paradigm has proven so fruitful (Lee 2010: 107). The future therefore looks promising in terms of the kinds of innovation we can expect and how they might benefit pedagogical contexts. In this final section, we highlight some issues related to the future advancement of each of the corpus types:

(i) General corpora should continue to grow and reach trillion-word size (Baker 2010: 12), and similar to the emergence of mega corpora such as COCA, they are expected to be more freely available. With the availability and diversity of texts online, it is likely that we will witness more contemporary text types being included, representing the emergence of digital communication, such as the Birmingham Blog Corpus (Kehoe and Gee 2012).

(ii) Specialised corpora require more attention to context and the mark-up of contextual features and the co-textual environment in order to facilitate the interpretation of smaller specialised data-sets (see Flowerdew 2009: 411-12). Also, the need to continue to explore how specialised corpora can be used in pedagogy and how challenges can be overcome remains a valid future line of enquiry.

(iii) Parallel corpora need to address issues of representativeness. In their current state, they 'span relatively few genres (mainly fiction, parliamentary proceedings, technical manuals), [and] a limited set of languages' (Kenning 2010: 488). It is thought that a more representative spectrum would greatly enhance innovation and insight in the area of parallel corpora.

(iv) Historical corpora need more attention to synergy between resources, research agendas and collaboration across interdisciplinary borders (Kytö 2011: 443-444). Kytö lists three over-arching categories for future directions; i) enhancing and adding to resources and methodologies for studying long-term and recent change; ii) ensuring comparability and links across corpora, and other electronic resources, and software; and iii) increasing our knowledge of the sociohistorical and cultural context of corpus texts, with special reference to interdisciplinary considerations.

(v) Multimodal corpora need to be larger, more representative and include a range of media via digital modes of communication. Knight (2011) highlights the need to improve technical devices and suggests as more investigations are implemented, limitations such as coding gestures will be reduced.

This chapter has provided a brief overview of the main types of corpora which exist as an introduction for scholars who are new to corpora and corpus linguistics. The pedagogic applications may be examined more closely in other chapters in this volume which focus on specific corpus types (see Caines *et al.* Chapter 25*;* Chambers Chapter 26*;* Meunier Chapter 27).

FURTHER READING

Davies, M. (2012) 'The 400 million word Corpus of Historical American English (1810-2009)' in I. Hegedus and Fodor, A. (eds) English Historical Linguistics 2010: Selected Papers from the Sixteenth International Conference on English Historical Linguistics (ICEHL), Pecs, 23-27 August 2010: 231-262.

Friginal, E. and Hardy, J. (2013) *Corpus-based sociolinguistics: A guide for students*, Abingdon, Oxon: Routledge.

Kipp, M. Martin, J. C., Paggio, P. and Heylen, D. (2009) (eds) *Multimodal Corpora*, Berlin: Springer.

McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus-Based Language Studies,* London: Routledge (and companion web source)

http://cw.routledge.com/textbooks/0415286239/resources/corpa.htm

Xiao, R. and Yue, M. (2009) 'Using corpora in translation studies: The state of the art' in P. Baker (ed.) *Contemporary Corpus Linguistics*, London: Continuum: 237-61.

Appendix: Websites

Corpus Archives		**Details**
Corpus BYU	http://corpus.byu.edu/	Free online search facility (registration required)
Lextutor	http://www.lextutor.ca	Free
Linguistic Data Consortium (LDC)	http://www.ldc.upenn.edu/	Membership fee
Oxford Text Archive	http://ota.ahds.ac.uk/	Free
General Corpora		
American National Corpus (ANC)	http://www.americannationalcorpus.org/	Free
Bank of English (BoE)/ Wordbanks	http://www.mycobuild.com/about-collins-corpus.aspx http://www.collinslanguage.com/content-solutions/wordbanks	Subscription fee
British National Corpus (BNC)	http://www.natcorp.ox.ac.uk/	Free online search facility Licence fee for corpus
BROWN Corpus	http://icame.uib.no/brown/bcm.html	Subscription fee or can be purchased with ICAME
Bulgarian National Corpus	http://ibl.bas.bg/en/BGNC_en.htm	Free online search facility
Corpus del Español	http://www.corpusdelespanol.org/	Free online search facility
Corpus di Italiano Scritto (CORIS)	http://corpora.dslo.unibo.it/coris_eng.ht	Free online search facility

	ml	
Corpus of Contemporary American English (COCA)	http://corpus.byu.edu/coca/	Free online search facility
Freiberg Brown Corpus of American English (FROWN)	http://khnt.hit.uib.no/icame/manuals/frown/	Subscription fee or can be purchased with ICAME
Freiberg London-Oslo/Bergen Corpus (FLOB)	http://icame.uib.no/flob/	Subscription fee or can be purchased with ICAME
International Corpus of English (ICE)	http://ice-corpora.net/ice/	Some corpora are freely available for research purposes
London-Oslo/Bergen Corpus (LOB)	http://khnt.hit.uib.no/icame/manuals/lob/	Subscription fee or can be purchased with ICAME
Russian Reference Corpus	http://bokrcorpora.narod.ru/index-en.html	Free online search facility of pilot version
Turkish National Corpus (TNC)	http://www.tnc.org.tr/index.php/en/	Free online search facility (registration required)
Parallel Corpora		
Canadian Hansard Corpus	http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC95T20	Subscription fee
Contemporary Chinese Translated Fiction Corpus (CCTFC)	http://www.bfsu-corpus.org/static/cctfc/	Free online search facility
English-Norwegian Parallel Corpus (ENPC)	http://www.hf.uio.no/ilos/english/services/omc/enpc/	Not publically available beyond creator institution
English-Swedish Parallel Corpus (ESPC)	http://www.sol.lu.se/engelska/corpus/corpus/espc.html	Not publically available beyond creator institution
Lancaster Corpus of Mandarin Chinese (LCMC)	http://www.lancs.ac.uk/fass/projects/corpus/LCMC/	Available through the Oxford Text Archive
Open Parallel Corpus (OPUS)	http://opus.lingfil.uu.se/	Free
Oslo Multilingual Corpus (OMC)	http://www.hf.uio.no/ilos/english/services/omc/	Not publically available beyond creator institution
Translational English Corpus (TEC)	http://ronaldo.cs.tcd.ie/tec2/jnlp/	Free
Historical Corpora		
ARCHER Corpus	http://www.alc.manchester.ac.uk/subjects/lel/research/projects/archer/	Free online search facility once user agreement has been signed
Corpus of Early English Correspondence (CEEC)	http://www.helsinki.fi/varieng/domains/CEEC.html	Available through the Oxford Text Archive
Corpus of Historical American English (COHA)	http://corpus.byu.edu/coha/	Free online search facility
Dictionary of Old English Corpus in Electronic Form	http://www.ota.ox.ac.uk/desc/2488	Available through the Oxford Text Archive
Early English Medical Writing	http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/EMEMTindex.html	CD-ROM by John Benjamins
Helsinki Corpus	http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/	Available through the Oxford Text Archive
Lampeter corpus	http://khnt.hit.uib.no/icame/manuals/LAMPETER/LAMPHOME.HTM	Available through the Oxford Text Archive
Multimodal Corpora		
BACKBONE Corpus	http://webapps.ael.uni-tuebingen.de/backbone-search/faces/initialize.jsp	Free
Nottingham Multimodal Corpus (NMMC)	http://www.cs.nott.ac.uk/~axc/DReSS/LRECw08.pdf	Not publically available beyond creator institution
SACODEYL Corpus	http://sacodeyl.inf.um.es/sacodeyl-search2/	Free
Santa Barbara Corpus of Spoken American English	http://www.linguistics.ucsb.edu/research/santa-barbara-corpus	Free online search facility

Specialised Corpora		
British Academic Written English (BAWE)	http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/	Available through the Oxford Text Archive
Cambridge and Nottingham Business English Corpus (CANBEC)	http://www.cambridge.org/gb/elt/catalogue/subject/custom/item3646597/Cambridge-English-Corpus-Business-English/?site_locale=en_GB	Not publically available
Cambridge and Nottingham E-Language Corpus (CANELC)	http://www.ncl.ac.uk/linguistics/research/publication/178260	Not publically available
Cambridge Corpus of Legal English (CCLE)	http://www.cambridge.org/gb/elt/catalogue/subject/custom/item3646600/Cambridge-English-Corpus-Cambridge-Corpus-of-Legal-English/?site_locale=en_GB	Not publically available
Corpus of London Teenage English (COLT)	http://www.hd.uib.no/colt/	Licence fee or can be purchased with ICAME
Corpus of Spoken Professional American English (CSPAE)	http://www.athel.com/cspa.html	Licence fee
English as a Lingua Franca Corpus (ELFA)	http://www.helsinki.fi/englanti/elfa/	Freely available once user agreement has been signed
Enron Email Corpus	https://www.cs.cmu.edu/~enron/	Free
International Corpus of Learner English (ICLE)	http://www.uclouvain.be/en-cecl-icle.html	Licence fee
Michigan Corpus of Academic Spoken English (MICASE)	http://quod.lib.umich.edu/m/micase/	Free online search facility
Michigan Corpus of Upper-level Student Papers (MICUSP)	http://micusp.elicorpora.info/	Free online search facility
WebCorp	http://www.webcorp.org.uk/live/	Free online search facility

References

Adolphs, S., Knight, D. and Carter, R. (2011) ‘Capturing context for heterogeneous corpus analysis: Some first steps’, *International Journal of Corpus Linguistics,* 16 (3): 305-24.

Allwood, J. (2009) ‘Multimodal corpora’, in A. Lüdeling and M. Kytö (eds) *Corpus Linguistics. An International Handbook*, Berlin: Mouton de Gruyter: 207-25.

Altenberg, B., Aijmer, K. and Svensson, M. (2001) *The English-Swedish Parallel Corpus (ESPC) Manuel,* Sweden: Department of English, University of Lund and Department of English, University of Göteborg.

Aksan, Y., Askan, M., Koltuksuz, A., Sezer, T., Mersinli, Ü., Demirhan, U., Yilmazer, H., Atasoy, G., Öz, S., Yildiz, İ., Kurtoğlu, Ö. (2012). ‘Construction of the Turkish National Corpus (TNC)’, *Proceedings of the Eight International Conference on*

Language Resources and Evaluation (LREC 2012), İstanbul, Turkiye, [online], available: http://www.lrec-conf.org/proceedings/lrec2012/papers.html [accessed 20/02/14].

Aston, G. (1995) ‘Corpora in language pedagogy: Matching theory and practice’, in G. Cook and B. Seidlhofer (eds) *Principle and Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson*, Oxford: Oxford University Press: 257-70.

Aston, G. (1997) ‘Small and large corpora in language learning’, in B. Lewandowska-Tomaszczyk and P. Melia (eds) *PALC 97: Practical Applications in Language Corpora*, Lodz: Lodz University Press: 51-62.

Aston, G. and Burnard, L. (1998) *The BNC Handbook*, Edinburgh: Edinburgh University Press.

Baker, M. (2000) ‘Towards a methodology for investigating the style of a literary translator’, *Target,* 12 (2): 241-66.

Baker, P. (2010) *Sociolinguistics and Corpus Linguistics,* Edinburgh: Edinburgh University Press.

Biber, D. (1990) ‘Methodological issues regarding corpus-based analyses of linguistic variation’, *Literary and Linguistic Computing* 5 (4): 257-269.

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

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

Blache, P., Bertrand, R. and Ferré, G. (2009) ‘Creating and exploiting multimodal annotated corpora: The ToMA project’, in M. Kipp, J. C. Martin, P. Paggio, and D. Heylen (eds) *Multimodal Corpora*, Berlin: Springer: 38-53.

Bowker, L. and Pearson, J. (2002) *Working with Specialized Language. A Practical Guide to Using Corpora,* London: Routledge.

Brinton, L. J. (2012) 'Historical pragmatics and corpus linguistics: Problems and strategies', in M. Kytö (ed) *English Corpus Linguistics*: *Crossing Paths* (*Language and Computers – Studies in Practical Linguistics* 76), Amsterdam: Rodopi: 101-131.

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. and McCarthy, M. (2006) *The Cambridge Grammar of English,* Cambridge: Cambridge University Press.

Chambers, A. (2005) 'Integrating corpus consultation into language studies', *Language Learning and Technology*, 19 (2): 111-125, [online], available: http://llt.msu.edu/vol9num2/chambers/default.html [accessed 25/08/12].

Claridge, C. (2008) 'Historical corpora', in A. Lüdeling and M. Kytö (eds) *Corpus Linguistics: An International Handbook Vol. 1*, Berlin: Walter de Gruyter: 242-59.

Claridge, C. (2000) *Multi-word Verbs in Early Modern English. A Corpus-based Study,* Amsterdam: Rodopi.

Clear, J. (1987) 'Trawling the language: Monitor corpora', in M. Snell-Hornby (ed.) *ZuriLEX Proceedings*, Tubingen: Francke: 383-89.

Culpeper, J. (2010) 'Historical Pragmatics', in L. Cummings (ed) Th*e Pragmatics Encyclopedia*, London: Routledge: 188-192.

Curzan, A. (2008) ‘Historical corpus linguistics and evidence of change’, in A. Lüdeling and M. Kytö (eds) *Corpus Linguistics: An International Handbook Vol. 1*, Berlin: Walter de Gruyter: 1091-1109.

Dahlmann, I. and Adolphs, S. (2009) ‘Spoken corpus analysis: Multimodal approaches to language description’ in P. Baker (ed.) *Contemporary Corpus Linguistics*, London: Continuum: 125-39.

Davies, M. (2002). *Corpus del Español: 100 million words, 1200s-1900s*, [online], available: http://www.crpusdelespanol.org.

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.

Davies, M. (2012) ‘Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English’, *Corpora* 7: 121-157.

Davies, M. (2013) *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries*. Available online at http://corpus2.byu.edu/glowbe/.

Degani, M. (2009) ‘Re-analysing the semi-modal ought to: An investigation of its use in the LOB, FLOB, Brown and Frown corpora’, *Language and Computers,* 69 (1): 327-46.

Deutschmann, M. (2003) *Apologising in British English,* Umeå: Umeå University Press.

Fengxiang, F. (2012) ‘A quantitative study on the lexical change of American English’, *Journal of Quantitative Linguistics,* 19 (3): 171-80.

Flowerdew, L. (2009) ‘Applying corpus linguistics to pedagogy. A critical evaluation’, *International Journal of Corpus Linguistics* 14 (3): 393-417.

Flowerdew, L. (2004) ‘The argument for using English specialised corpora to understand academic and professional settings’, in U. Connor and T. Upton (eds)

Discourse in the Professions: Perspectives from Corpus Linguistics, Amsterdam: John Benjamins: 11-33.

Gavioli, L. (2002) 'Some thoughts on the problem of representing ESP through small corpora', in B. Kettemann and G. Marko (eds) Teaching *and Learning by Doing Corpus Analysis*, Amsterdam: Rodopi: 293-303.

Gavioli, L. (2005) *Exploring Corpora for ESP*, Amsterdam: John Benjamins.

Handford, M. (2010) 'What can a corpus tell us about specialist genres?', in A. O'Keeffe and M. McCarthy (eds) *The Routledge Handbook of Corpus Linguistics*, London: Routledge: 255-269.

Hanks, P. (2009) 'The impact of corpora on dictionaries', in P. Baker (ed.) *Contemporary Corpus Linguistics*, London: Continuum: 214-36.

Haugh, M. (2009) 'Designing a multimodal spoken component of the Australian National Corpus', in M. Haugh, K. Burridge, J. Mulder and P. Peters (eds) *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, Somerville, MA: Cascadilla Proceedings Project: 74-86.

Healey, A. (1999) 'The Dictionary of Old English Corpus on the World Wide Web', *Medieval English Studies Newsletter,* 40: 2-10.

Hoffstaedter, P. (2010) 'BACKBONE: E-learning for modern language teaching', *Transfer: The Steinbeis Magazine*.

Hoffstaedter, P. and Kohn, K. (2009) 'Real language and relevant language learning activities: Insights from the SACODEYL project', in A. Kirchhofer and J. Schwarzkopf (eds) *The Workings of the Anglosphere. Contributions to the Study of British and US-American Cultures*, Trier: WVT.

Hundt, M., Nesselhauf, N. and Biewer, C. (eds) (2009) *Corpus Linguistics and the Web.* Amsterdam: Rodopi: 1-7.

Hundt, M., Sand, A. and Siemund, R. (1998) *Manual of Information to Accompany the Freiburg-LOB Corpus of British English ("FLOB"),* [online], available: http://www.hit.uib.no/icame/flob/index.htm [accessed 06/06/13].

Hundt, M., Sand, A. and Skandera, P. (1999) *Manual of Information to Accompany the Freiburg-Brown Corpus of American English (Frown),* [online], available: http://khnt.hit.uib.no/icame/manuals/frown/INDEX.HTM [accessed 06/06/13].

Hundt, M. and Gut, U. (eds) (2012) *Mapping Unity and Diversity Worldwide: Corpus-based Studies of New Englishes,* Amsterdam: John Benjamins.

Johansson, S., Leech, G. and Goodluck, H. (1978), *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*, Oslo: University of Oslo.

Johansson, S., Ebeling, J. and Oksefjell, S. (2002) *The English-Norwegian Parallel Corpus Manual,* Oslo: Department of British and American Studies University of Oslo.

Johns, T. (1991) 'Should you be persuaded. Two samples of data-driven learning materials', *English Language Research Journal*, 1-14.

Kehoe, A. and Gee, M. (2012) 'Reader comments as an aboutness indicator in online texts: Introducing the Birmingham Blog Corpus', *Studies in Variation, Contacts, and Change in English Vol. 12. Aspects of Corpus Linguistics: Compilation, Annotation and Analysis* [online], available

http://www.helsinki.fi/varieng/series/volumes/12/kehoe_gee/ [accessed 26/02/14].

Kenning, M. M. (2010) ‘What are parallel and comparable corpora and how can we use them?’, in A. O'Keeffe and M. McCarthy (eds) *The Routledge Handbook of Corpus Linguistics,* London: Routledge: 487-500.

Kilgarriff, A. (2001) ‘Web as corpus’, in P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja (eds) *Proceedings of the Corpus Linguistics 2001 Conference*, Lancaster: UCREL: 342-44.

Kilgarriff, A. and Grefenstette, G. (2003) ‘Introduction to the special issue on the web as corpus’, *Computational Linguistics,* 29 (3): 333-48.

Kjellmer, G. (1994) *A Dictionary of English Collocations Based on the Brown Corpus,* Oxford: Clarendon Press.

Koester, A. (2010) ‘Building small specialised corpora’, in A. O’Keeffe and M. McCarthy (eds) *The Routledge Handbook of Corpus Linguistics*, London: Routledge: 66-79.

Koeva, S., Stoyanova, I., Leseva, S., Dimitrova, T., Dekova, R. and Tarpomanova, E. (2012) ‘The Bulgarian National Corpus: Theory and Practice in Corpus Design’, *Journal of Language Modelling* 1: 65-110.

Kohn, K., Hoffstaedter, P. and Widmann, J. (2009) *‘*BACKBONE – Pedagogic corpora for content and language integrated learning*’*, paper presented at *EuroCALL 2009 New trends in CALL: Working together*, Universidad Politécnica de Valencia (Spain), 2009.

Knight, D. (2011) ‘The future of multimodal corpora’, *RBLA, Belo Horizonte,* 11 (2): 391-415.

Knight, D. and Tennent, P. (2008) ‘Introducing DRS (The Digital Replay System): A tool for the future of Corpus Linguistic research and analysis’, paper presented at the

6th *Language Resources and Evaluation Conference*, Palais des Congrés, Mansour Eddahbi, Marrakech, Morocco, May 2008.

Knight, D., Adolphs, S., Tennent, P. and Carter, R. (2008) ‘The Nottingham Multi-Modal Corpus: A Demonstration’, paper presented at the 6th *Language Resources and Evaluation Conference*, Palais des Congrés, Mansour Eddahbi, Marrakech, Morocco, May 2008.

Knight, D. and Adolphs, S. (2007) ‘Pragmatics and corpus linguistics: A mutualistic entente’, in J. Romero-Trillo (ed.) *Corpus and Pragmatics*, Berlin: Mouton de Gruyter: 172-90.

Kübler, N. and Aston, G. (2010) ‘Using corpora in translation’, in A. O'Keeffe and M. McCarthy (eds) *The Routledge Handbook of Corpus Linguistics*, London: Routledge: 501-15.

Kucĕra, H. and Francis, W. (1967) *Computational Analysis of Present-day English*, Providence: Brown University Press.

Kytö, M. (2012) ‘Introduction’, in M. Kytö (ed) *English Corpus Linguistics: Crossing Paths* (*Language and Computers – Studies in Practical Linguistics*76), Amsterdam: Rodopi: 1-6.

Kytö, M. (2011) ‘Corpora and historical linguistics’, *Revista Brasileira de Linguistica Aplicada* 11 (2): 391-415.

Kytö, M. and Pahta, P. (2012) ‘Evidence from historical corpora up to the twentieth century’, in T. Nevalainen and E. C. Traugott (eds) *The Oxford Handbook of the History of English*, Oxford: Oxford University Press: 123-33.

Lee, D. Y. W. (2010) ‘What corpora are available?’, in A. O’Keeffe and M. McCarthy (eds) *The Routledge Handbook of Corpus Linguistics*, London: Routledge: 107-21.

McCarten, J. (2010) ‘Corpus-informed course book design’, in A. O’Keeffe and M. McCarthy (eds) *The Routledge Handbook of Corpus Linguistics*, London: Routledge: 413-27.

McEnery, A. and Xiao, R. (2004) ‘Swearing in modern British English: The case of *fuck* in the BNC’, *Language and Literature,* 13 (3): 235-68.

Mauranen, A. (2012) *Exploring ELF: academic English shaped by non-native speakers*, Cambridge: Cambridge University Press.

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

Nesi, H. (2012) ‘Laughter in university lectures’, *Journal of English for Academic Purposes* 11 (2): 79-89.

Nevalainen, T, and Raumolin-Brunberg, H. (eds) (1996) *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*, Amsterdam: Rodopi.

O’Donnell, M. and Römer, U. (2012) ‘From student hard drive to web corpus (part 2): The annotation and online distribution of MICUSP’, *Corpora* 7: 1-18.

O’Keeffe, A. (2007) ‘The Pragmatics of Corpus Linguistics’, keynote paper presented at the fourth *Corpus Linguistics Conference* held at the University of Birmingham, Birmingham, July 2007.

O'Keeffe, A. and Farr, F. (2003) ‘Using language corpora in initial teacher education: Pedagogic issues and practical applications’, *TESOL Quarterly,* 37 (3): 389-418.

O'Keeffe, A., McCarthy, M. and Carter, R. (2007) *From Corpus to Classroom. Language Use and Language Teaching,* Cambridge: Cambridge University Press.

OMC (2010) *Oslo Multilingual Corpus*, [online], available: http://www.hf.uio.no/ilos/english/services/omc/ [accessed 12/06/13].

Partington, A. (2006) *The Linguistics of Laughter: A Corpus-Assisted Study of Laughter-Talk,* London: Routledge.

Renouf, A. (2002) 'WebCorp: Providing a renewable data source for corpus linguists', in S. Granger and S. Petch-Tyson (eds) *Extending the scope of corpus-based research*, Amsterdam: Rodopi, 39-58.

Reppen, R. (2009) 'English language teaching and corpus linguistics: Lessons learned from the American National Corpus', in P. Baker (ed.) *Contemporary Corpus Linguistics*, London: Continuum: 204-13.

Reppen, R. (2010a) *Using Corpora in the Language Classroom*, Cambridge: Cambridge University Press.

Reppen, R. (2010b) 'Corpora in the classroom: Forging new paths', paper presented at *Annual TESOL convention*, Denver, CO.

Reppen, R. and Ide, N. (2004) 'The American National Corpus: Overall goals and the first release', *Journal of English Linguistics,* 32 (2): 105-13.

Rissanen, M. (2000) 'The world of English historical corpora', *Journal of English Linguistics,* 28 (1): 7-20.

Rissanen, M. (2008) 'Corpus linguistics and historical linguistics', in A. Lüdeling and M. Kytö (eds) *Corpus Linguistics: An International Handbook Vol. 1*, Berlin: Walter de Gruyter: 53-68.

Rissanen, M (2012) 'Corpora and the study of the history of English', in M. Kytö (ed) *English Corpus Linguistics: Crossing Paths* (*Language and Computers – Studies in Practical Linguistics*76), Amsterdam: Rodopi: 197-220.

Rossini Favretti, R., Tamburini, F., and De Santis, C. (2004). ‘A corpus of written Italian: a defined and a dynamic model’, in A. Wilson, P. Rayson & T. McEnery (eds) *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, Munich: Lincom-Europa: 27-38.

Schmid, H. J. (2003) ‘Do men and women really live in different cultures? Evidence from the BNC’, in A. Wilson, P. Rayson and T. McEnery (eds) *Corpus Linguistics by the Lune*, Frankfurt: Peter Lang: 185-221.

Sharoff, S. (2004) ‘Methods and tools for development of the Russian Reference Corpus’, in D. Archer, A. Wilson, and P. Rayson (eds) *Corpus Linguistics Around the World*, Amsterdam: Rodopi: 167-180.

Sinclair, J. (ed.) (1987) *Looking Up*, London: HarperCollins.

Sinclair, J. (1997) ‘Corpus evidence in language description’, in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds) *Teaching and Language Corpora*, London: Longman: 27-39.

Sinclair, J. (2003) *Reading Concordances,* London: Longman.

Sinclair, J. (2004) *Trust the Text: Language, Corpus and Discourse,* London: Routledge.

Simpson-Vlach, R. and Leicher, S. (2006) *The MICASE Handbook: A Resource for Users of the Michigan Corpus of Academic Spoken English*, Michigan: Michigan University Press.

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

Taavitsainen, I and Pahta, P. (eds) (2010) *Early Modern English Medical Texts. Corpus Description and Studies*, Amsterdam: John Benjamins.

Tiedemann, J. (2012) ‘Parallel Data, Tools and Interfaces in OPUS’, paper presented at the 8th *International Conference on Language Resources and Evaluation* (LREC'2012), Istanbul, Turkey, 2012.

Tognini-Bonelli, E. and Sinclair, J. (2006) ‘Corpora’, in K. Brown (ed.) *Encyclopedia of Language and Linguistics*, Amsterdam: Elsevier: 216-19.

Tribble, C. (2002) ‘Corpora and corpus analysis: New windows on academic writing’, in J. Flowerdew (ed.) *Academic Discourse*, London: Longman: 131-49.

Véronis, J. (ed.) (2000) *Parallel Text Processing. Alignment and Use of Translation Corpora,* Amsterdam: Kluwer Academic Publishers.

Widmann, J., Kohn, K. and Ziai, R. (2011) ‘The SACODEYL search tool: exploiting corpora for language learning purposes’, in A. Frankenberg Garcia, L. Flowerdew and G. Aston (eds) *New Trends in Corpora and Language Learning,* London: Continuum: 167-180.

Yaguchi, M, Iyeiri, Y. and Okabe, H. (2004) ‘Style and gender differences in formal contexts: An analysis of *sort of* and *kind of* appearing in the CSPAE’, paper presented at the 5th *Annual Wenshan Conference on ELT*, Literature, and Linguistics.

Yáñez-Bouza, N. 2011. ‘ARCHER past and present (1990-2010)’, *ICAME Journal* 35: 205-236.

Xiao, R. and Yue, M. (2009) ‘Using corpora in translation studies: The state of the art’, in P. Baker (ed.) *Contemporary Corpus Linguistics*, London: Continuum: 237-61.