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 (Kucéra 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
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 than 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 publicly 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


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


Appendix: Websites

<table>
<thead>
<tr>
<th>Corpus Archives</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus BYU</td>
<td><a href="http://corpus.byu.edu/">http://corpus.byu.edu/</a> Free online search facility (registration required)</td>
</tr>
<tr>
<td>Lextutor</td>
<td><a href="http://www.lextutor.ca">http://www.lextutor.ca</a> Free</td>
</tr>
<tr>
<td>Linguistic Data Consortium (LDC)</td>
<td><a href="http://www.ldc.upenn.edu/">http://www.ldc.upenn.edu/</a> Membership fee</td>
</tr>
<tr>
<td>Oxford Text Archive</td>
<td><a href="http://ota.ahds.ac.uk/">http://ota.ahds.ac.uk/</a> Free</td>
</tr>
<tr>
<td><strong>General Corpora</strong></td>
<td></td>
</tr>
<tr>
<td>American National Corpus (ANC)</td>
<td><a href="http://www.americannationalcorpus.org/">http://www.americannationalcorpus.org/</a> Free</td>
</tr>
<tr>
<td>British National Corpus (BNC)</td>
<td><a href="http://www.natcorp.ox.ac.uk/">http://www.natcorp.ox.ac.uk/</a> Free online search facility Licence fee for corpus</td>
</tr>
<tr>
<td>BROWN Corpus</td>
<td><a href="http://icame.uib.no/brown/bcm.html">http://icame.uib.no/brown/bcm.html</a> Subscription fee or can be purchased with ICAME</td>
</tr>
<tr>
<td>Bulgarian National Corpus</td>
<td><a href="http://ibl.bas.bg/en/BGNC_en.htm">http://ibl.bas.bg/en/BGNC_en.htm</a> Free online search facility</td>
</tr>
<tr>
<td>Corpus del Español</td>
<td><a href="http://www.corpusdelespanol.org/">http://www.corpusdelespanol.org/</a> Free online search facility</td>
</tr>
<tr>
<td>Corpus di Italiano Scritto (CORIS)</td>
<td><a href="http://corpora.dslo.unibo.it/coris_eng.ht">http://corpora.dslo.unibo.it/coris_eng.ht</a> Free online search facility</td>
</tr>
<tr>
<td>Corpus</td>
<td>Website</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Corpus of Contemporary American English</td>
<td><a href="http://corpus.byu.edu/coca/">http://corpus.byu.edu/coca/</a></td>
</tr>
<tr>
<td>(COCA)</td>
<td></td>
</tr>
<tr>
<td>Freiberg Brown Corpus of American English</td>
<td><a href="http://khnt.hit.uib.no/icame/manuals/frown/">http://khnt.hit.uib.no/icame/manuals/frown/</a></td>
</tr>
<tr>
<td>(FROWN)</td>
<td></td>
</tr>
<tr>
<td>Freiberg London-Oslo/Bergen Corpus (FLOB)</td>
<td><a href="http://icame.uib.no/flob/">http://icame.uib.no/flob/</a></td>
</tr>
<tr>
<td>International Corpus of English (ICE)</td>
<td><a href="http://ice-corpora.net/ice/">http://ice-corpora.net/ice/</a></td>
</tr>
<tr>
<td>London-Oslo/Bergen Corpus (LOB)</td>
<td><a href="http://khnt.hit.uib.no/icame/manuals/lob/">http://khnt.hit.uib.no/icame/manuals/lob/</a></td>
</tr>
<tr>
<td><strong>Parallel Corpora</strong></td>
<td></td>
</tr>
<tr>
<td>Canadian Hansard Corpus</td>
<td><a href="http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC95T20">http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC95T20</a></td>
</tr>
<tr>
<td>Contemporary Chinese Translated Fiction</td>
<td><a href="http://www.bfsu-corpus.org/static/cctfc/">http://www.bfsu-corpus.org/static/cctfc/</a></td>
</tr>
<tr>
<td>Corpus (CCTFC)</td>
<td></td>
</tr>
<tr>
<td>English-Norwegian Parallel Corpus (ENPC)</td>
<td><a href="http://www.hf.uio.no/ilos/english/services/omc/enpc/">http://www.hf.uio.no/ilos/english/services/omc/enpc/</a></td>
</tr>
<tr>
<td>English-Swedish Parallel Corpus (ESPC)</td>
<td><a href="http://www.sol.lu.se/engelska/corpus/esc.html">http://www.sol.lu.se/engelska/corpus/esc.html</a></td>
</tr>
<tr>
<td>Lancaster Corpus of Mandarin Chinese (LCMC)</td>
<td><a href="http://www.lancs.ac.uk/fass/projects/corpora/LCMC/">http://www.lancs.ac.uk/fass/projects/corpora/LCMC/</a></td>
</tr>
<tr>
<td>Open Parallel Corpus (OPUS)</td>
<td><a href="http://opus.lingfil.uu.se/">http://opus.lingfil.uu.se/</a></td>
</tr>
<tr>
<td>Oslo Multilingual Corpus (OMC)</td>
<td><a href="http://www.hf.uio.no/ilos/english/services/omc/">http://www.hf.uio.no/ilos/english/services/omc/</a></td>
</tr>
<tr>
<td>Translational English Corpus (TEC)</td>
<td><a href="http://ronaldo.cs.tcd.ie/tec2/tmlp/">http://ronaldo.cs.tcd.ie/tec2/tmlp/</a></td>
</tr>
<tr>
<td><strong>Historical Corpora</strong></td>
<td></td>
</tr>
<tr>
<td>ARCHER Corpus</td>
<td><a href="http://www.alc.manchester.ac.uk/subject/lel/research/projects/archer/">http://www.alc.manchester.ac.uk/subject/lel/research/projects/archer/</a></td>
</tr>
<tr>
<td>Corpus of Early English Correspondence</td>
<td><a href="http://www.helsinki.fi/varieng/domains/CEEC.html">http://www.helsinki.fi/varieng/domains/CEEC.html</a></td>
</tr>
<tr>
<td>(CEEC)</td>
<td></td>
</tr>
<tr>
<td>Corpus of Historical American English (COHA)</td>
<td><a href="http://corpus.byu.edu/coha/">http://corpus.byu.edu/coha/</a></td>
</tr>
<tr>
<td>Dictionary of Old English Corpus in Electronic Form</td>
<td><a href="http://www.ota.ox.ac.uk/desc/2488">http://www.ota.ox.ac.uk/desc/2488</a></td>
</tr>
<tr>
<td>Lampeter corpus</td>
<td><a href="http://khnt.hit.uib.no/icame/manuals/LAMPHETER/LAMPHOME.HTM">http://khnt.hit.uib.no/icame/manuals/LAMPHETER/LAMPHOME.HTM</a></td>
</tr>
<tr>
<td><strong>Multimodal Corpora</strong></td>
<td></td>
</tr>
<tr>
<td>BACKBONE Corpus</td>
<td><a href="http://webapps.ael.uni-tuebingen.de/backbone-search/faces/initialize.jsp">http://webapps.ael.uni-tuebingen.de/backbone-search/faces/initialize.jsp</a></td>
</tr>
<tr>
<td>Nottingham Multimodal Corpus (NMMC)</td>
<td><a href="http://www.cs.nott.ac.uk/~axc/DReSS/RECw08.pdf">http://www.cs.nott.ac.uk/~axc/DReSS/RECw08.pdf</a></td>
</tr>
<tr>
<td>SACODEYL Corpus</td>
<td><a href="http://sacodeyl.inf.um.es/sacodeyl-search2/">http://sacodeyl.inf.um.es/sacodeyl-search2/</a></td>
</tr>
<tr>
<td>Specialised Corpora</td>
<td>Available through the Oxford Text Archive</td>
</tr>
<tr>
<td>--------------------------------------------------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>British Academic Written English (BAWE)</td>
<td><a href="http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/">http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/</a></td>
</tr>
<tr>
<td>Cambridge and Nottingham E-Language Corpus (CANELC)</td>
<td><a href="http://www.ncl.ac.uk/linguistics/research/publication/178260">http://www.ncl.ac.uk/linguistics/research/publication/178260</a></td>
</tr>
<tr>
<td>Corpus of London Teenage English (COLT)</td>
<td><a href="http://www.hd.uib.no/colt/">http://www.hd.uib.no/colt/</a></td>
</tr>
<tr>
<td>Corpus of Spoken Professional American English (CSPAE)</td>
<td><a href="http://www.athel.com/cspa.html">http://www.athel.com/cspa.html</a></td>
</tr>
<tr>
<td>English as a Lingua Franca Corpus (ELFA)</td>
<td><a href="http://www.helsinki.fi/englanti/elfa/">http://www.helsinki.fi/englanti/elfa/</a></td>
</tr>
<tr>
<td>Enron Email Corpus</td>
<td><a href="https://www.cs.cmu.edu/~enron/">https://www.cs.cmu.edu/~enron/</a></td>
</tr>
<tr>
<td>Michigan Corpus of Academic Spoken English (MICASE)</td>
<td><a href="http://quod.lib.umich.edu/m/micase/">http://quod.lib.umich.edu/m/micase/</a></td>
</tr>
<tr>
<td>Michigan Corpus of Upper-level Student Papers (MICUSP)</td>
<td><a href="http://micusp">http://micusp</a> elicorpora.info/</td>
</tr>
<tr>
<td>WebCorp</td>
<td><a href="http://www.webcorp.org.uk/live/">http://www.webcorp.org.uk/live/</a></td>
</tr>
</tbody>
</table>

**References**


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


