Corpora: English-Language

SEBASTIAN HOFFMANN

Introduction

This entry provides an overview of a selection of electronic language corpora that are available for the study of English at the time of writing (early 2010), including information about their availability and pricing as well as any specific corpus tools that are required to access the data, where applicable. Before taking a look at individual corpora, a few introductory comments are in order.

First, it is useful to distinguish two senses in which the word "corpus" is used. The more conventional sense is the one described by Sinclair (1996): A corpus "is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language." In other words, a corpus is a carefully crafted compilation of instances of authentic language use rather than just any large number of electronically stored texts. In recent years, however, some linguists have started to take a more generous view of what constitutes a corpus. This particularly refers to the whole range of data that is available in virtually limitless quantities on the Internet, but it also applies to specialized electronic text archives (e.g., of literary texts or newspapers) and electronic dictionaries (e.g., the *Oxford English Dictionary*). While such data sets no doubt represent interesting sources of data, the present overview will use the more conventional definition and concentrate exclusively on traditional corpora that are compatible with Sinclair's definition given above.

Second, since specialized corpora are treated in a separate entry, this overview will be restricted to what is commonly termed "general corpora." This label refers to a balanced collection of electronic texts that is intended to be representative of the whole range of language contexts available to speakers of a particular language variety (e.g., British or Australian English). For very large general corpora, the term "reference corpus" is sometimes used. While some general corpora cover both spoken and written data, the term is also applicable if only one of the two modes is represented.

Third, it must be stressed that the brief descriptions of corpora offered here are no more than introductory in character. Since no fully objective criteria for corpus representativeness exist, reliable interpretations of the linguistic patterns observed in the data greatly benefit from a full awareness of the sampling strategies employed by the corpus compilers. Readers are therefore invited to consult as much additional information as is available for the corpora before using them to conduct serious research. A useful resource for this purpose is the Corpus Resource Database (CoRD—see http://www.helsinki.fi/varieng/CoRD/index.html), which provides first-hand information on a range of English-language corpora; all information was submitted or approved by the compilers of each corpus.

Finally, a compact overview of this type must necessarily be selective in its coverage. As a matter of course, the omission of a general corpus should not be seen as an indication that it would not be a useful resource for the study of the English language. A more complete and annotated overview of corpus resources can be found in Lee (*n.d.*).

The Encyclopedia of Applied Linguistics, Edited by Carol A. Chapelle.
 Published 2013 by Blackwell Publishing Ltd.
DOI: 10.1002/9781405198431.wbeal0229

Overview of General Corpora

British National Corpus

The British National Corpus (BNC—see www.natcorp.ox.ac.uk; Burnard, 2007) was created to be a balanced reference corpus of late 20th-century British English. It contains almost 100 million words, about 10% of which are transcriptions of spoken data. All in all, the BNC consists of over 4,000 text samples of varying length (ranging from a few hundred words to several tens of thousands of words). All texts were automatically part-of-speech tagged with CLAWS (see http://ucrel.lancs.ac.uk/claws/; Garside, 1987), making use of the C5 tag set (also known as the "BNC Basic Tag Set"). In addition, all word forms in the corpus are associated with their corresponding headwords (e.g., *GIVE* for the verb forms *give*, *gives*, *gave*, *given*, and *giving*).

The texts to be included were selected according to specified selection criteria in order to mirror the language of the time. For the written component of the corpus, these criteria were "text domain" (e.g., world affairs, leisure, arts), "time period" (mostly 1985–93, but including some older texts dating back as far as the 1960s), and "medium" (e.g., book, periodical). For the spoken component, two separate strategies were employed. On the one hand, a selection of about 150 recruited speakers recorded all their conversations over a given period of time. The selection of these so-called respondents was based on socio-demographic criteria (age, sex, social class, and geographical region) and mirrored the structure of British society at the time. The 4.2 million words recorded in this fashion form the demographically sampled part of the corpus (DS). In contrast, the 6.2 million words of the context-governed part (CG) were chosen to represent particular settings or contexts of language use, for instance meetings, radio broadcasts, lectures, and tutorials. Texts from four broad domains were selected in roughly equal proportions: "business," "educational and informative," "institutional," and "leisure." Unlike the DS part of the corpus, the CG part also contains monologues (approximately 25%). On the whole, the CG part of the corpus tends to represent more formal language use than the spontaneous conversations of the DS part; this difference is not categorical, however.

Once the material was selected, it was included in the corpus together with whatever supplementary information about it was available (e.g., age, sex, domicile of author or speaker, perceived level of difficulty, genre). The proportions of the corpus annotated with these descriptive features are thus accidental rather than the result of the corpus compilers' conscious objectives to mirror general language use as closely as possible. For example, about 70% of CG is produced by men, while in DS an equal proportion of words is spoken by women and men.

The transcription of the spoken data is fairly broad and purely orthographic, but it includes such metatextual information as indications of pause length, overlap of speakers, and speech quality (e.g., whispering). No prosodic information is available. Some of the original tape recordings were deposited with the British Library, and at the time of writing a project is under way that will prepare and release the sound recordings of the demographically sampled part of the corpus to the public. Significantly, the project also involves alignment of the transcription with the sound.

Since its first release in 1995, the BNC has been revised twice. No new texts were added to the corpus, but some known errors were fixed (e.g., duplicate texts removed) and the annotation of metatextual information was improved. The texts of the third edition of the corpus (released in 2007) are formatted in XML. The corpus can be searched with a variety of tools. It is distributed with XAIRA (see xaira.sourceforge.net), an open-source corpus tool that can also be used to search other corpora in XML format. An alternative is offered by BNCweb (see bncweb.info), a web-based interface integrated with the

powerful CQP query software (see Hoffmann, Evert, Smith, Lee, & Berglund Prytz, 2008). There are also a number of free web-based services available that offer restricted access to the corpus (e.g., with limited context of query results); these include Mark Davies's BYU-BNC (see http://corpus.byu.edu/bnc) and Bill Fletcher's Phrases in English or PIE (see http://phrasesinenglish.org/).

The BNC is distributed by the University of Oxford. A personal license is priced at £75 while an institutional license costs £500 (plus VAT, if applicable). Two smaller subsets of BNC texts, BNC Baby (4 million words) and the BNC Sampler (2 million words) are available through the same channels.

Corpus of Contemporary American English

For over a decade, no corpus of American English existed that could match the BNC as regards size and range of data. In the year 2008, this gap was to a large extent filled by the Corpus of Contemporary American English (COCA—see www.americancorpus.org; Davies, 2009), a balanced corpus containing over 400 million words of both spoken and written data. In contrast to the BNC, the contents of COCA are not fixed but updated on a twice-yearly basis, adding approximately 20 million words every year, thus making it possible to investigate very recent developments in American English. A further difference to the BNC is that virtually all texts in COCA are complete texts rather than text samples of varying sizes. The texts in the corpus were automatically part-of-speech tagged with CLAWS. The tag set that was used is the more detailed C7 tag set (see http://ucrel.lancs.ac.uk/claws7tags.html); however, since both COCA and the BNC were tagged with the same tagger, it is still possible to compare tag-based findings across these data sets.

Like the BNC, COCA contains a range of different domains or genres, but the choice of categories and their proportions are somewhat different, covering the five genres of spoken, fiction, popular magazines, newspapers, and academic journals in equal quantities. For each of these genres, care was taken to balance their contents not only overall but also for each year. For example, the genre "popular magazines" contains data from almost 100 different sources that cover a large range of different topics and target audiences, including news, health, home and gardening, women, finance, religion, and sports. Similarly, the texts for the "academic journals" genre are taken from the entire range of the Library of Congress classification system, and again this applies as far as possible to each individual year for which data were collected. An Excel spreadsheet containing detailed information about the composition of the corpus can be downloaded from the COCA Web site (see http://corpus.byu.edu/coca/files/texts.zip).

Since sampling strategies and selection criteria can vary considerably, it is always a somewhat problematic exercise to compare findings from different corpora. The COCA genre for which direct comparability with BNC data is potentially most challenging is the spoken data. This is because all spoken texts are derived from transcripts of unscripted conversation on TV and radio programs (e.g., *Newshour* [PBS], *Good Morning America* [ABC], *Today Show* [NBC], and *Larry King Live* [CNN]). Although there can be little doubt that these interactions are indeed relevant examples of spoken interaction, their nature is necessarily different from the carefully compiled and demographically sampled set of spontaneous conversations in the spoken component of the BNC. For researchers whose focus is primarily on spoken interaction, the use of additional American sources (e.g., the Santa Barbara Corpus, see below) is therefore recommended to reduce the danger that findings may be significantly influenced by issues relating to corpus compilation rather than by actual differences between British and American English.

For copyright reasons, COCA cannot be distributed to users but is exclusively accessible through its Web-based interface. This tool offers the typical features of a concordancer,

including flexible ways of restricting searches to individual sections of the corpus (e.g., time spans or genres) and displaying results in tabular and graphical formats. The display of the immediate context of search results is restricted to a few words before and after the query match. While this will have no impact on a large range of research questions, this limitation may nevertheless at times be frustrating for users wishing to investigate corpus findings in a more qualitative way (e.g., to conduct research on pragmatics).

The Bank of English

Like COCA, the Bank of English is a corpus whose contents are not static but instead updated on a regular basis. Its compilation was started in the 1980s under the direction of John Sinclair as part of the COBUILD (Collins Birmingham University International Language Database) project, resulting in a corpus of initially 8 million words of English text. In forming the basis for the *Collins COBUILD English Language Dictionary* (1987), it set the stage for a new approach to lexicography and dictionary making that was explicitly corpus-based. Since then, the Bank of English has grown to a size of over 500 million words, containing data from a wide range of sources (written and spoken) from eight different (native) varieties of English. The corpus is available to subscribers as part of Collins Wordbanks*Online* (see http://www.collinslanguage.com/wordbanks/default.aspx) via a Web-based interface (from £695, excluding VAT; free one-month trials are available). A 56-million word subset of the corpus can be accessed for free, but the output is restricted to a maximum of 40 concordance lines.

The Brown Family of Corpora

The three general corpora presented so far belong to a category that is often referred to as "mega-corpora." However, there are a number of smaller corpora that also clearly deserve the label "general corpus." A set of these is known as the Brown family of corpora, whose name derives from the very first electronic corpus of English, the Standard Sample of Present-Day American English, later simply known as the Brown Corpus, referring to the university where it was compiled (Francis, 1965; Francis & Kučera, 1979). It consists of 500 written text samples of 2,000 words dating from the year 1961, thus 1 million words. For the selection of texts, the compilers applied a sampling frame that was intended to capture a very wide range of (written) language use: The corpus is divided into two major parts, termed "informative" (approximately 75%) and "imaginative" (approximately 25%), which are in turn subdivided into a total of 15 different genres of varying proportions (e.g., press: editorial, popular lore, romance, and love story). The actual text samples to be included for each section were chosen randomly, for instance from a list of all available publications in a particular subject field, rather than based on any text-internal criteria.

About a decade later, the Brown Corpus was twinned by a matching selection of texts of British English dating from the year 1961, the 1-million-word Lancaster–Oslo–Bergen Corpus (LOB; Johansson, Leech, & Goodluck, 1978), to permit direct quantitative comparisons of written English as it is used on both sides of the Atlantic. In the early 1990s, a team at Freiburg University, Germany, then compiled two further analogues of the original Brown Corpus, the Freiburg–Brown Corpus (abbreviated Frown; Hundt, Sand, & Skandera, 1999) and the Freiburg–LOB Corpus (abbreviated FLOB; Hundt, Sand, & Siemund, 1998) containing texts from 1991 to 1992. The Brown family of corpora thus makes it possible to investigate and compare recent language variation and change in British and American English over a period of 30 years.

All four corpora are distributed on the ICAME CD-ROM at a cost of NOK 3,500 (approximately €440 in spring 2010) for individual users and NOK 8,000 (approximately €1,000) for institutional licenses. The CD-ROM contains a range of other English-language

corpora, some of which are also mentioned in the present overview. The corpora are in a simple text-only format, with minimal mark-up using simplified SGML tags (e.g., indicating paragraph boundaries or foreign words) in the case of Frown and FLOB, and can thus be searched by means of standard corpus tools (e.g., AntConc [see http://www.antlab.sci.waseda.ac.jp/software.html] or WordSmith Tools [see http://www.lexically.net/wordsmith/]). The Brown Corpus is also distributed in XML format on the BNC Baby CD (see above). Finally, both the Brown Corpus and LOB are also available in part-of-speech tagged format on the ICAME CD-ROM (see http://icame.uib.no/newcd.htm and http://khnt.hit.uib.no/icame/manuals/).

Three further corpora modeled on the Brown format were compiled for different varieties of English: the Kolhapur Corpus of Written Indian English (Shastri, 1986), the Australian Corpus of English (ACE; Peters, *n.d.*), and the Wellington Corpus of New Zealand English (Bauer, 1993), containing data published in 1978 (Kolhapur) and 1986 (ACE and Wellington) respectively. Furthermore, at the time of writing, several projects are under way or have recently been completed that extend the Brown family of corpora to include both older and more recent data sets. Thus, there are now two additional corpora for British English: BLOB-1931 (for "before LOB"; Leech & Smith, 2005), containing material published around the year 1931, and BE06 (Baker, 2009), consisting of a comparable set of texts from the year 2006, the latter compiled from Internet sources. These corpora are currently not publicly distributed; however, it is anticipated that BLOB-1931 will become available soon via a Web-based search interface. Text collection for 1901 British and 1931 American analogues is still in progress.

By the standards of the early 21st century, the size of the Brown family of corpora is relatively small. However, given their virtually exact match in terms of sampling frames, they nevertheless represent highly valuable sources of data, particularly where medium- to high-frequency phenomena are concerned.

International Corpus of English

One further set of small-scale corpora deserves mention in the current context: the International Corpus of English (ICE—see http://ice-corpora.net/ice). Its aim is to provide researchers with a collection of 1-million-word comparable corpora of more than 20 varieties of English worldwide. However, each of its components is also designed to be a general corpus of its variety in its own right. Each corpus is compiled following the same strategies; again, 500 text samples of 2,000 words each are included. In contrast to the Brown family of corpora, however, ICE corpora contain 60% of spoken material. At the time of writing, about half the corpora are publicly available, including data sets for four native varieties of English (Canada, Great Britain, Ireland, and New Zealand).

All ICE corpora follow the same annotation scheme, involving for example markup for overlapping speech, foreign or indigenous words, and typographic features such as boldface font and underlining. In addition, all corpora will eventually be part-of-speech tagged with the same automatic tagger, thus improving the opportunities for comparing the use of grammatical structures across different varieties. Furthermore, ICE-GB has already been syntactically parsed at the phrase, clause, and sentence level.

The majority of available ICE components can be either downloaded free of charge from the ICE Web site (after submitting a signed license agreement), or accessed for a nominal fee from the corresponding ICE team. The exception is ICE-GB, with prices between about £350 and £750, depending on location and type of license; a single-user student license is available for £25, however. The original audio recordings for ICE-GB can be obtained for an additional £250–500. With the exception of ICE-GB and its complex syntactic-parsing annotation scheme, all ICE corpora can be searched with standard corpus tools. For ICE-GB,

the (Windows-only) tool ICECUP is provided to make full use of the grammatical information encoded in the corpus.

Spoken General Corpora

Since the transcription of conversation is a much more time-consuming task than collecting written data, the number of available general spoken corpora is fairly limited, and none exceeds even half the size of the 10-million-word spoken component of the BNC. For British English, the London–Lund Corpus (LLC; Greenbaum & Svartvik, 1990) is a widely used corpus containing approximately 500,000 words representing various types of spoken language (e.g., spontaneous face-to-face conversations or prepared monologues) recorded mostly during the 1960s and 1970s, but with a few texts dating as far back as 1953 and 11 texts recorded in the 1980s. Some of the recordings were made surreptitiously, without the knowledge of (at least some of) the people involved, which ensures a high level of authenticity of the material, but would no longer be allowed today.

The LLC is distributed on the ICAME CD-ROM. All recordings were carefully transcribed and contain detailed prosodic information. While this is no doubt a great asset for researchers of speech, the format makes a reliable search for individual words or phrases with a standard corpus tool next to impossible. As a case in point, consider the four lines below, reproduced from the LLC. The word *wiggle* occurs twice, but with different types of intonation, indicated by the word-internal symbols "\/" (i.e., fall–rise) and "\" (i.e., fall), respectively. A search for *wiggle* with a standard corpus tool would match neither of these instances.

```
111b 37 8750 1 1 A 11 2^and !then . 'do a :w\/iggle# - - /
111b 37 8760 1 1 A 12 2+^so as _to+ - ++^s/orry#++ /
111b 37 8770 1 1 B 11 2+((it ^always !w\as#))+ /
111b 37 8780 1 1 B 11 2an ^old 'old ++w\iggle#++ /
```

Various informal versions of the LLC stripped of its prosodic annotation exist, but it appears that only one of them is officially distributed (see DCPSE below).

For spoken American English, the readily available corpus data is sparse. The only general corpus of spoken interactions is the Santa Barbara Corpus (see http://www.linguistics.ucsb.edu/research/sbcorpus.html), with 249,000 words in 60 different texts. The data contain interactions from various regions in the USA; the majority of contributions appear to be by speakers of educated Standard American English. The corpus is distributed by the Linguistic Data Consortium in four separate volumes and includes the original sound files; prices for individual volumes range from US$100 to US$200 each. A much larger corpus is the Longman Spoken American Corpus, containing 5 million words of spontaneous conversations of more than 1,000 speakers from a diverse set of sociodemographic backgrounds. However, virtually no documentation exists and access to the corpus unfortunately appears to be restricted to Longman in-house use, where it is for example employed in the compilation and validation of modern grammars of English.

Diachronic Corpora

The final section of general corpora to be presented here deals with text collections whose contents are intended to represent language use from different time periods and are thus primarily suited for the investigation of language change. Perhaps the best-known of these is the diachronic part of the Helsinki Corpus of English Texts (or Helsinki Corpus for short; Kytö, 1996), which covers almost 1,000 years in the history of English (ca. 750 to ca. 1700). It contains 1,572,800 words in three major sections (Old/Middle/Early Modern

English), which are further subdivided into a total of 13 subsections. Most texts included are samples (2,000 to 10,000 words) rather than complete texts. In the selection of texts, the compilers took care to cover a wide variety of language use available through records from each time period; however, given the relatively small overall size of the corpus, the representativeness of individual registers within each section must necessarily be evaluated with caution. For many types of investigation, the Helsinki Corpus offers a convenient starting point that can then be complemented by research on the basis of more specialized corpora covering much shorter periods of time.

Each text of the Helsinki Corpus contains a header with detailed information relating to such parameters as geographical dialect, genre, and sociolinguistic features such as age, sex, and social rank of the author, if available. The file format allows searches with standard corpus tools; however, users need to be aware of the conventions employed to encode certain typical features of older manuscripts. For example, superscript is indicated with an equal sign (e.g., y = t = for y^t) and letters such as ash and thorn are indicated with compound characters (e.g., + t for þ). The Helsinki Corpus is distributed on the ICAME CD-ROM.

For researchers wishing to investigate the full length of the history of English, the Helsinki Corpus can be complemented with ARCHER (short for A Representative Corpus of Historical English Registers—see http://www.llc.manchester.ac.uk/research/projects/archer/; Biber, Finegan, & Atkinson, 1994), which contains about 1.8 million words sampled from seven 50-year periods (1650–1990) covering nine different registers. A total of 70% of the corpus represents British English; for American English, only three time periods (1750–99, 1850–99, and 1950–90) are included, but plans are in place to fill the gaps in the American data in the near future. The files are in a simple text format and can be searched with standard corpus tools. ARCHER is maintained by a team of researchers involving 15 different universities. For copyright reasons, it is not publicly available; however, interested researchers can gain local access to the corpus at any of the 15 universities involved.

Given the ephemeral nature of speech, very few collections of historical spoken data exist, and even fewer data sets belong to the category of "general corpora." There is one notable exception, though, involving major parts of two corpora that have already been mentioned here: the Diachronic Corpus of Present-Day Spoken English (DCPSE—see http://www.ucl.ac.uk/english-usage/projects/dcpse/index.htm). It consists of equal proportions of data from the LLC and ICE-GB, amounting to a total of about 875,000 words. Like ICE-GB, the texts from the LLC have been syntactically parsed, thus allowing direct diachronic comparisons of grammatical, and not merely lexical, features of spoken British English across a time span of about 30 years. The corpus can be searched with ICECUP; various licensing options are available and prices range from about £350 to £720; a student license is available for £25.

Finally, at the time of writing, a promising new source of diachronic data for American English was about to be released for public access: the Corpus of Historical American English (COHA—see http://corpus.byu.edu/coha.asp), a balanced collection of 400 million words covering the four genres of fiction, popular magazines, newspapers, and academic prose, and thus nicely complementing the present-day English data contained in COCA (see above).

Conclusion

The use of electronic corpora has had an enormous impact on the field of applied linguistics. Today's scholar can access large amounts of authentic language use by way of a few mouse clicks. Furthermore, sophisticated corpus tools have been developed that support both novices and experts in the field in their analysis of the patterns that emerge from the

data. In particular, English corpora have become increasingly influential in language teaching, both when it comes to informing the creation of modern pedagogical materials and when used as active tools in the classroom (see, e.g., Hunston, 2002; papers in Quereda, Santana, & Hidalgo, 2006).

As will be apparent from this overview of general corpora, researchers can select from a wide range of options to find the corpus that is best suited for answering a particular research question. However, it will also be clear that the selection criteria employed by the compilers of these general corpora vary quite considerably. Apart from the obvious difference between corpora incorporating spoken language, written language, or both, more subtle differentiations arise due to the choice and proportions of text types included. As already alluded to in the introduction, it is necessary for researchers to be thoroughly acquainted with the corpus that they are searching in order to make full sense of the patterns that can be observed. This is even more necessary if findings from different corpora are compared.

Fascinating new avenues of research are opening up in corpus linguistics due to the use of virtually limitless quantities of Internet-derived data. Still, for many types of applications, recourse to a data set that was compiled on the basis of principled decisions in order to be representative of a well-defined population of language users is recommended.

SEE ALSO: Corpora: Specialized; Corpus Analysis of English as a World Language; Corpus Analysis of the World Wide Web; Francis, Nelson; Greenbaum, Sidney; Kučera, Henry; Sinclair, John

References

Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics, 14*(3), 312–37.

Biber, D., Finegan, E., & Atkinson, D. (1994). ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers. In U. Fries, G. Tottie, & P. Schneider (Eds.), *Creating and using English language corpora* (pp. 1–14). Amsterdam, Netherlands: Rodopi.

Collins Cobuild English Language Dictionary. (1987). London, England: Collins.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*(2), 159–90.

Francis, W. N. (1965). A standard corpus of edited present-day American English. *College English, 26*, 267–73.

Garside, R. (1987). The CLAWS word-tagging system. In R. Garside, G. Leech, & G. Sampson (Eds.), *The computational analysis of English: A corpus-based approach* (pp. 30–41). London, England: Longman.

Greenbaum, S., & Svartvik, J. (1990). The London–Lund Corpus of Spoken English. In J. Svartvik (Ed.), *The London–Lund Corpus of Spoken English: Description and research*. (*Lund studies in English, 82*, pp. 11–45). Lund, Sweden: Lund University Press. Retrieved March 14, 2010 from http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM

Hoffmann, S., Evert, S., Smith, N., Lee D., & Berglund Prytz, Y. (2008). *Corpus linguistics with BNCweb—a practical guide*. Frankfurt, Germany: Peter Lang.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge, England: Cambridge University Press.

Leech, G., & Smith, N. (2005). Extending the possibilities of corpus-based research on English in the twentieth century: A prequel to LOB and FLOB. *ICAME Journal, 29*, 83–98.

Quereda, L., Santana, J., & Hidalgo, E. (Eds.). (2006). *Corpora in the foreign language classroom*. Amsterdam, Netherlands: Rodopi.

Suggested Readings

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243–57.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the Web* (pp. 133–49). Amsterdam, Netherlands: Rodopi.

Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). *Change in contemporary English: A grammatical study*. Cambridge, England: Cambridge University Press.

Rissanen, M. (2000). The world of English historical corpora: From Cædmon to the computer age. *Journal of English Linguistics, 28*(1), 7–20.

Xiao, R. Z. (2007). Well-known and influential corpora. In A. Lüdeling & M. Kytö (Eds.), *Handbook on corpus linguistics* (*Handbooks of linguistics and communication science,* pp. 383–457). Berlin, Germany: De Gruyter.

Online Resources

Bauer, L. (1993). *Manual of information to accompany the Wellington Corpus of Written New Zealand English*. Wellington, New Zealand: Victoria University of Wellington. Retrieved March 14, 2010 from http://khnt.hit.uib.no/icame/manuals/wellman/INDEX.HTM

Burnard, L. (2007). *Reference guide for the British National Corpus (XML edition)*. Retrieved March 14, 2010 from http://www.natcorp.ox.ac.uk/docs/URG/

Francis, W. N., & Kučera, H. (1979). *Manual of information to accompany a Standard Corpus of Present-Day Edited American English, for use with digital computers*. Providence, RI: Brown University, Department of Linguistics. Retrieved March 14, 2010 from http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM

Hundt, M., Sand, A., & Siemund, R. (1998). *Manual of information to accompany the Freiburg–LOB Corpus of British English ("FLOB")*. Freiburg, Germany: Albert-Ludwigs-Universität Freiburg. Retrieved March 14, 2010 from http://khnt.hit.uib.no/icame/manuals/flob/INDEX.HTM

Hundt, M., Sand, A., & Skandera, P. (1999). *Manual of information to accompany the Freiburg-Brown Corpus of American English ("Frown")*. Freiburg, Germany: Albert-Ludwigs-Universität Freiburg. Retrieved March 14, 2010 from http://khnt.hit.uib.no/icame/manuals/frown/INDEX.HTM

Johansson, S., Leech, G., & Goodluck, H. (1978). *Manual of information to accompany the Lancaster–Oslo–Bergen Corpus of British English, for use with digital computers*. Oslo, Norway: University of Oslo. Retrieved March 14, 2010 from http://khnt.hit.uib.no/icame/manuals/lobman/INDEX.HTM

Kytö, M. (1996). *Manual to the diachronic part of the Helsinki Corpus of English Texts: Coding conventions and lists of source texts*. Helsinki, Finland: University of Helsinki. Retrieved March 14, 2010 from http://khnt.hit.uib.no/icame/manuals/HC/INDEX.HTM

Lee, D. (*n.d.*). *Bookmarks for corpus-based linguistics*. Retrieved February 24, 2011 from http://tiny.cc/corpora

Peters, P. (*n.d.*). *Manual of information to accompany the Australian Corpus of English*. Sydney, Australia: Macquarie University. Retrieved March 14, 2010 from http://khnt.hit.uib.no/icame/manuals/ace/INDEX.HTM

Shastri, S. V. (1986). *Manual of information to accompany the Kolhapur Corpus of Indian English, for use with digital computers*. Kolhapur, India: Shivaji University. Retrieved March 14, 2010 from http://khnt.hit.uib.no/icame/manuals/kolhapur/INDEX.HTM

Sinclair, J. McH. (1996). *Corpus* and *computer corpus*. In *EAGLES: Preliminary recommendations on corpus typology/Definitions/Corpus and computer corpus*. Retrieved March 14, 2010 from http://www.ilc.cnr.it/EAGLES96/corpustyp/node5.html