Chapter 13:
Creating and using corpora

Stefan Th. Gries and John Newman

1. Introduction

Over the last few decades, corpus-linguistic methods have established themselves as among the most powerful and versatile tools to study language acquisition, processing, variation, and change. This development has been driven in particular by the considerations in (1).

(1) a. technological progress (e.g., processor speeds as well as hard drive and RAM sizes);
b. methodological progress (e.g., the development of software tools, programming languages, and statistical methods);
c. a growing desire by many linguists for (more) objective, quantifiable, and replicable findings as an alternative to, or at least as an addition to, intuitive acceptability judgments (see Chapter 3);
d. theoretical developments such as the growing interest in cognitively- and psycholinguistically-motivated approaches to language in which frequency of (co-) occurrence plays an important role for language acquisition, processing, use, and change.

In this chapter, we will discuss a necessarily small selection of issues regarding (i) the creation, or compilation, of new corpora and (ii) the use of corpora once they have been compiled. Although this chapter encompasses both the creation and use of corpora, there is no expectation that any individual researcher would be engaged in both these kinds of activities. Different skills are called for when it comes to creating and using corpora, a point noted by Sinclair (2005: 1) who draws attention to the potential pitfalls of a corpus analyst building a corpus, specifically, the danger that the corpus will be constructed in a way that can only serve to confirm the analyst's pre-existing expectations. Some of the issues addressed in this chapter are also dealt with in Wynne (2005), McEnery, Xiao, and Tono (2006), and McEnery and Hardie (2012) in a fairly succinct way and more thoroughly in Lüdeling and Kytö (2008a, 2008b) and Beal, Corrigan, and Moisl (2007a, 2007b).[1]

2. Creating corpora

2.1 The notion of a 'corpus': A prototype and dimensions of variation

The notion of a *corpus* can best be defined as a category organized around a prototype. Most generally, a corpus can be described as 'a body of naturally occurring language' (McEnery, Xiao, and Tono 2006: 4), thereby distinguishing a corpus from word lists, dictionaries, databases, etc. These days, the prototypical corpus is a *machine-readable* collection of language used in authentic settings/contexts: one that is intended to be *representative* for a particular language, variety, or register (in the sense of reflecting all the possible parts of the intended

[1] Details of corpora and software mentioned in this chapter are provided in Appendices 1 and 2 (URLs accessed 28 April 2012). These are rapidly developing domains and information provided here is naturally only current as at the time of writing. Updated lists are available on the companion website for this volume.

language/variety/register) and that is intended to be *balanced* such that the sizes of the parts of the corpus correspond to the proportion these parts make up in the language/variety/register (cf. McEnery et al 2006: 5, Hunston 2008: 160-166, Gries 2009: Ch. 1). However, many corpora differ from an ideal design along these (and other) parameters; in fact there is disagreement as to whether just *any* body of naturally occurring language can be called a corpus. Kilgarriff and Grefenstette (2006: 334), by way of introducing and advocating the study of data from the World Wide Web, adopt a definition of a corpus as 'a collection of texts when considered as an object of language or literary study.' On the other hand, Sinclair (2005: 15) explicitly excludes a number of categories from linguistic corpora, e.g., a single text, an archive, and, in particular, the World Wide Web. Beyond being a body of naturally occurring language, then, it is difficult to agree on any more particular definition of what a corpus is or is not. Note, too, that some collections of language can diverge from the prototypical property of being 'naturally occurring language' and yet are still happily referred to as corpora by their creators. As an example, consider the TIMIT Acoustic-Phonetic Continuous Speech Corpus, made up of audio recordings of 630 speakers of eight major dialects of American English. For these recordings, each speaker read ten 'phonetically rich' sentences – a uniquely valuable resource for the study of acoustic properties of American English, but not what one would consider naturally occurring language.

A detailed overview of corpora, illustrating the range of types of corpora that are being studied within linguistics, can be found in the chapters of Lüdeling and Kytö (2008a: 154-483). Apart from the above criteria defining prototypical corpora, one can distinguish corpus types by the media that hold the data: written text (web, text documents, historical manuscripts, see Chapter 11 for details on the use of diachronic corpora); audio; video and audio; audio and transcribed spoken texts based on the audio, etc. There is often an assumption that a corpus will include written language or transcriptions of spoken language (which arguably represents the prototypical kind of language use), but it is important to appreciate that collections of naturally occurring speech in the form of audio files ('speech corpora' as opposed to transcriptions of spoken language) are valid corpora. Ostler (2008: 459) remarks on the artificiality of distinctions between speech-based and text-based corpora in light of the increasing use of multi-tiered annotations of audio and video data (see Chapter 12 for details on transcription and multi-tier annotation). One may also choose to distinguish corpus types by content or source: synchronic vs. historical, national corpora, learner corpora, academic discourse, children's language, interviews, static vs. monitor corpora, multilingual, web-based, etc. Corpora, as used in linguistics, are created with particular purposes of study in mind and the variety of corpus types should not be surprising – it is no more than a reflection of the richness and multi-facetedness of language use and the many perspectives one can bring to the study of language. One can therefore not speak of a 'standard' in corpus construction or design in the sense of a set of protocols that must be adhered to in order for the corpus to be admissible in corpus linguistics; the conception of 'corpus' as a category around a prototype is more appropriate (cf. Gilquin and Gries 2009: Section 2). Further information on these corpora can be found in Appendix 1.

There are now many large corpora of high quality available, where 'large' means, say, 100 million words or more. We emphasize, though, that smaller corpora also have their place alongside the larger corpora. The key consideration is to have an appropriate match of research goal and corpus type/size and, for some research goals, even quite a small corpus constructed by a researcher can yield insightful results. Berkenfield (2001), using a corpus of just 10,640 words, was able to carry out research on phonetic reduction of *that* in spoken English; Thompson and Hopper (2001) successfully explored transitivity in a corpus of multi-party conversations consisting of just 446 clauses; Fiorentino (2009) studied ordering of adverbial and main clauses in an Italian corpus consisting of 26,000 words for the written part of the corpus and 32,000 for the spoken part. Smaller corpora such as these can suffice when the focus of the study is a relatively frequent phenomenon, but would not be advisable if the focus is a relatively rare phenomenon. Granath (2007), reflecting on the different results obtained from searching for an English inversion structure like *Thus ended his dreams*, found reason to appreciate both the one-million word corpora and the 50 million word corpora used in the study: '…in the end, combining evidence from large and small corpora can give us information that neither type of corpus could provide on its own' (Granath 2007: 183).

2.2 Collecting the corpus data

In this and the following section, we describe the main steps involved in preparing and annotating a new corpus, before reviewing readily available corpora in section 2.4. It is fair to say that most corpora are created with the expectation that they are, in some sense, representative of something larger than themselves – what we referred to as the prototypical corpus in section 2.1 – rather than the ultra-pragmatic view of a corpus held by Kilgarriff and Grefenstette. Consequently, an initial and profound decision relates to exactly what the corpus is supposed to be representative of and what sampling technique is to be used (see Chapter 5 for a more general discussion of sampling). One very basic kind of decision guiding the collection of language data concerns the categories that form the basis of the sampling: categories of language users (e.g., gender, age, socio-economic class, geographical location), categories of the language products (e.g., spoken language, written language, register of language use, text type, formality of the language), or a combination of both of these. A noteworthy example of how categories of language users can figure prominently in corpus data is the sub-corpus of the Uppsala Learner English Corpus used in Johansson and Geisler (2011). For the purposes of their study of the syntax of Swedish learners of English, the authors carefully chose learners' essays to balance the numbers of boys and girls, and the levels of the school year, as summarized in Table 1.

		boys		girls	
Level	school year	mean essay length in words	number of essays	mean essay length in words	number of essays
Junior high	Year 7	228.0	5	217.0	5
	Year 9	221.8	5	234.0	5
Senior high	Year 1	220.8	5	190.0	5
	Year 3	277.8	5	245.0	5
Total		237.1	20	221.5	20

Table 1: A subset of the Uppsala Learner English Corpus (adapted from Table 1 in Johansson and Geisler 2011: 140)

Typically, it is categories such as register (i.e., categories relating to properties of the product rather than the user) that are the preferred basis for structuring the more common corpora in use (see the examples of widely used corpora in section 2.4). This is in part due to the unavailability of socio-demographic data on speakers and writers in the case of many texts (as retrieved, for example, from the World Wide Web), but it may also be due to the view that the variation between, say, spoken and written modalities is far more significant than variation between male and female speech or writing. The approach adopted in creating the Canadian component of the International Corpus of English (ICE-CAN offers a practical way of proceeding: data is basically sampled on the basis of categories of register (broadly understood), such as spoken vs. written, spoken dialogue vs. spoken monologue, spoken dialogue private vs. spoken dialogue public, written printed vs. non-printed, but some attempt is made to balance the numbers of male and female speakers in the data collection. The *metadata* on speakers contributing to the spoken part of ICE-CAN and available as part of the distribution of the corpus, summarized in (2), is in fact extensive enough for a sociolinguistically oriented use of the corpus:

(2) a. date of recording
b. place of recording
c. gender
d. age
e. mother tongue
f. other languages spoken
g. self-reported ethnicity
h. occupation
i. educational profile
j. professional training
k. overseas experience

The decision as to what the corpus should be representative of will always have a huge impact on the how the corpus data will be collected: recordings of natural conversation, recorded interviews, conversation from TV programs, fictional texts or journalese (from the web or processed by optical character recognition (OCR) software), blogs and chatroom data, general content crawled/collected from the web are but a few possible data sources, and careful decisions as to what can and must be included are required and will, realistically, often have to be balanced with what is possible within the restrictions of particular research agendas and goals. Sometimes there can be hidden biases in making decisions about representativeness, skewing the data collected in unintended ways. A typical bias may favor a 'standard' or better known variety of language over less prestigious (dialectal, colloquial) varieties or favor the collection of data from more educated speakers. Newman and Columbus (2009), for example, found an (unintended) over-representation of vocabulary relating to the education domain in a number of the conversational corpora in the International Corpus of English project, most likely a consequence of the easy availability of speakers from the education sector as contributors of data. Of course, the researcher may quite consciously opt for data specifically restricted to a standard variety, educated speakers, or other factors, but it should not be thought that a corpus must be restricted in this way. In addition, there is a variety of further restrictions on the collection of data which often have to do with what speakers/writers allow to be done with their speech/texts. For example, for reasons of copyright or the traditions of speech communities, not everything that can be found on the web can be added to a corpus that is intended for use by others.

These days, the World Wide Web offers a useful starting point for obtaining text which can be utilized for the construction of corpora. Collections of published materials (out of the range of current copyright) such as Project Gutenberg provide a wealth of literary texts in many languages that can be exploited for the creation of customized digital corpora. But as already indicated above, there is an abundance of material available for downloading apart from literary texts: newspaper collections, Wikipedia entries, university lectures, film scripts, translations of the Bible, blogs, etc. Oral history projects provide opportunities for the creation of spoken corpora. Consider, as just one example, the Southern Oral History Program which began in 1973 with the aim of documenting the life of the American South in tapes, videos, and transcripts. According to the website, this project will ultimately make 500 oral history interviews available over the internet (400 are already available), selected from the 4,000 or so oral history interviews carried out over thirty years. The interviews cover a variety of topics in recent North Carolina history, particularly civil rights, politics, and women's issues. As of writing, the index contains a list of 496 topics. Interviews can be read as text transcript, listened to (or downloaded) with a media player, or both simultaneously. Note, also, that applications such as HTTrack (for Windows/Unix) or Sitesucker (for Mac) can currently be used with many sites enabling an automated mirroring of whole websites.

Our emphasis in this chapter is on creating and using corpora as written or transcribed texts, but some comments on collecting spoken data are in order (see Chapters 9 and 11), for many observations directly relevant here). One issue immediately confronting a researcher collecting data directly from a speaker is how to minimize observer effects. Inconspicuousness and versatility are two key goals in managing the collection of speech data (intended to reflect natural, non-self-conscious use of language), as discussed in Chapters 6 and 9. The CallHome American English Speech corpus, for example, follows a procedure which is likely to reduce any observer effect. The corpus is based on recorded telephone conversations lasting up to 30 minutes, where the participants are fully aware that they are being recorded. The transcripts which derive from these recordings, however, are based only on 10 contiguous minutes from within those 30 minutes. While this strategy does not exclude some self-consciousness on the part of the speakers, it does serve to lessen any such effect since the speakers cannot know in advance which 10 minutes is being utilized for the transcript. A second issue surrounding the collection of spoken data concerns the quality of the audio/video recording. Needless to say, one aims for the best quality possible (WAV rather than MP3 format for audio files, for example), though sometimes a lesser quality may suffice. The corpora in the International Corpus of English project, for example, are designed primarily for distribution as corpora in the form of text files where the spoken data have been transcribed into regular English orthography. In such cases, the quality of the recording must be good enough for reliable transcription even if it falls short of what a researcher carrying out a fine acoustic analysis requires. Finally, creating a speech corpus in which the acoustic characteristics are of importance leads naturally to additional kinds of metadata compared with those in (2) above. (3) summarizes the metadata available in the CallHome American English Speech corpus.

(3) Metadata for a conversation recording:

a. total number of speakers

b. number of females and males

d. number of speakers per channel and number of males/females per channel

e. difficulty (overall quality of the channel in terms of number of speakers, background noise, channel noise, speed, accent, articulation)

f. background noise (amount of sound not made by the speakers, e.g., baby crying, television, radio, etc.)

g. distortion (echo and other types of recording problems)

h. crosstalk (audibility of the channel A speaker on channel B, and vice-versa)

Metadata for the caller:
i. gender
j. age
k. years of education
l. where the caller grew up
m. telephone number called

Once first versions of video/sound/text files have been obtained, typically one or more follow-up steps are necessary, which are discussed in the following section.

2.3 Preparing the corpus data

The first versions of files obtained in the first collection step hardly ever correspond to the desired final versions. Rather, such files typically require two additional steps before they can be used and made available as corpus files: they virtually always need to be cleaned up and standardized, and they often need to be marked up and annotated. In today's age of increased data sharing, it is important to standardize corpus files to facilitate later use by other researchers with different goals.

2.3.1 Cleaning up and standardizing

The first versions of files typically need to be cleaned of any undesired information they may contain. Files which include information that is protected for privacy reasons need to have such information edited in some way (see Chapters 2 and 12). For example, if one gathers recordings of authentic conversation, it is often necessary to protect the speakers' privacy as well as the privacy of those who a speaker talks about in their absence. (Imagine a case where, during a recording, a speaker mentions that her neighbor cheated on last year's tax report or that her brother's visa has expired.) Data like these require careful consideration of how much one can and must anonymize the data. In ICE-CANCanada, for example, names other than those of public figures, were anonymized through the use of pseudonyms.

Files obtained from the Internet or other sources can be in one of any number of formats (.txt, .html, .xml, .pdf, .doc etc.) and will almost invariably require some editing for them to be used most effectively. In using files from the web as a convenient example, editing *may* include, but is not limited to the tasks listed in (4).

(4) a. converting all files into one and the same interoperable file format and language encoding (e.g., converting data into Unicode text files);
b. removing and or standardizing unwanted elements (e.g., deleting unwanted HTML tags such as image references, title, body, table, and other tags, links, scripts, etc.);
c. standardizing different spellings and character representations (e.g., standardizing ü and `&uuml;` into ü, etc.);
d. identifying files downloaded more than once and deleting copies.

This kind of editing typically requires ready-made tools with particular features or, better, the use of a programming language. An example of a ready-made application at the time of writing is the free cross-platform Java-based text editor jEdit. While jEdit has many attractive features, it includes the three key features relevant to formatting texts for corpus-based research: (i) it accepts a wide range of *language encodings*, including UTF-8 and UTF-16; (ii) it allows for search and replace over multiple files; (iii) it features search and replacement operations using *regular expressions*, which are a method to describe simple or very complicated sequences of characters in files (see Table 11 below). Software like jEdit and other text editors intended for programmers force the user to be more attuned to properties of files which become important in working with corpus tools, such as language encodings and (Unix- vs. Windows- vs. Mac-style) line breaks. Regular expressions increase the power of editing considerably, allowing options such as finding and deleting all annotation contained within angular brackets, adding an annotation at the beginning of each line, removing some variable number of lines of text at the beginning of a file such as all text within `<teiHeader>…</teiHeader>`, features that are not necessarily available in typical word processing software.

2.3.2 Marking up and annotating

Once one has files that are cleaned up and standardized as desired, a second preparatory step usually involves enriching these with desired information they do not yet contain. Such information serves to facilitate the retrieval of linguistic patterns and their co-occurrence with other (linguistic or extra-linguistic) data. Usually, one distinguishes markup and annotation.

In the case of written or transcribed data, the *markup* section of a file refers to metadata about the file and might include information such as when the data in the file was collected, a description of the source of the data, when the file was prepared, basic social information about participants if relevant, and other such details. Figure 1 shows an example of markup from the beginning of the Extensible Markup Language (XML) version of the Brown corpus, distributed as part of Baby BNC v.2. The elements of markup conform to the specifications laid down by the Text Encoding Initiative (TEI), a consortium of interested parties, which are concerned with establishing standards for sharing documents. Angled brackets ‘<‘ and ‘>‘ demarcate the tags which enclose metadata; a ‘/’ indicates a closing tag. All the information in the TEI header, for example, is found between the opening tag `<teiHeader>` and the closing tag `</teiHeader>`; the `header` in turn, consists of a file description within the `<fileDesc>` tags; a title statement within the `<titleStmt>` tags; an edition statement within the `<editionStmt>` tags, and so on, as seen in Figure 1. The TEI guidelines for markup of texts are intended to apply to all kinds of texts and are not designed specifically for the files of a linguistic corpus. An extension of the TEI guidelines specifically intended for corpus markup (and annotation) is the Corpus Encoding Standard (CES) and the more recent version of these standards designed for XML, namely Extensible Corpus Encoding Standard (ECES).

```
<teiHeader>
        <fileDesc>
                <titleStmt>
                        <title>Sample A01 from  The Atlanta Constitution</title>
                        <title type="sub"> November 4, 1961, p.1 "Atlanta Primary ..."
                        "Hartsfield Files"
                        August 17, 1961, "Urged strongly ..."
                        "Sam Caldwell Joins"
                        March 6,1961, p.1 "Legislators Are Moving" by Reg Murphy
                        "Legislator to fight" by Richard Ashworth
                        "House Due Bid..."
                        p.18 "Harry Miller Wins..."
                        </title>
```

```
            </titleStmt>
            <editionStmt>
                  <edition>A part  of the XML version of the Brown Corpus</edition>
            </editionStmt>
            <extent>1,988 words 431 (21.7%) quotes 2 symbols</extent>
            <publicationStmt>
                  <idno>A01</idno>
                  <availability><p>Used by permission of The Atlanta Constitution State News
                  Service (H), and Reg Murphy (E).</p></availability>
            </publicationStmt>
            <sourceDesc>
                  <bibl> The Atlanta Constitution</bibl>
            </sourceDesc>
      </fileDesc>
      <encodingDesc>
            <p>Arbitrary Hyphen: multi-million [0520]</p>
      </encodingDesc>
      <revisionDesc>
            <change when="2008-04-27">Header auto-generated for TEI version</change>
      </revisionDesc>
</teiHeader>
```

Figure 1: Markup in the TEI Header of file A01 in the XML Brown Corpus

The *annotation* part of a file refers to elements added to provide specifically linguistic information, e.g., part of speech, semantic information, and pragmatic information. Most commonly, annotation takes the form of part-of-speech tagging of words. The first sentence of the Brown Corpus is shown in a parts-of-speech annotated form in (5a). The tags used in this sentence are explained in (5b) – full details can be found in the Brown Corpus Manual (icame.uib.no/Brown/bcm.html#bc6). Other tagsets are the various versions of Constituent Likelihood Automatic Word-tagging System (CLAWS) and the University of Pennsylvania (Penn) Treebank Tagset. Figure 2 shows the same annotated sentence in an XML format.

(5) a. The/at Fulton/np-tl County/nn-tl Grand/jj-tl Jury/nn-tl said/vbd Friday/nr an/at investigation/nn of/in Atlanta's/np$ recent/jj primary/nn election/nn produced/vbd ``/`` no/at evidence/nn "/" that/cs any/dti irregularities/nns took/vbd place/nn ./.

b. at = article; np-tl = proper noun, also appearing in title (of the newspaper article, in this case); nn-tl = singular common noun, also appearing in the title; jj-tl = adjective, also appearing in the title; vbd = past tense of verb; nr = adverbial noun; nn = singular common noun; in = preposition; np$ = possessive proper noun; jj = adjective; cs = subordinating conjunction; dti = singular or plural determiner/quantifier; nns = plural common noun; . = sentence closer; " = punctuation

```
<p
      <s n="1"
            <w type="AT"The</w
            <w type="NP" subtype="TL"Fulton</w
            <w type="NN" subtype="TL"County</w
            <w type="JJ" subtype="TL"Grand</w
            <w type="NN" subtype="TL"Jury</w
            <w type="VBD"said</w
            <w type="NR"Friday</w
            <w type="AT"an</w
            <w type="NN"investigation</w
            <w type="IN"of</w
            <w type="NP$"Atlanta's</w
            <w type="JJ"recent</w
            <w type="NN"primary</w
            <w type="NN"election</w
            <w type="VBD"produced</w
            <c type="pct"``</c
            <w type="AT"no</w
            <w type="NN"evidence</w
            <c type="pct"''</c
            <w type="CS"that</w
```

```
                <w type="DTI"any</w
                <w type="NNS"irregularities</w
                <w type="VBD"took</w
                <w type="NN"place</w
                <c type="pct".</c
        </s
</p
```

Figure 2: The first sentence (and paragraph) in the text body of file a01 in the XML Brown Corpus (the tags beginning with p, s, and w mark the paragraph, sentence, and each word respectively)

Sometimes, a tagging system allows for multiple tags to be associated with one and the same word. In general, the CLAWS tagger assigns to each word in a text one or more tags (regardless of the context in which it occurs) and then tries to identify the one best tag based on the frequency of word-tag combinations in the immediate context. However, sometimes the algorithm is unable to clearly identify one and only one tag and uses a hyphenated tag such as VVG-NN1 instead (as when *singing* in the sentence *She says she couldn't stop singing* is tagged VVG-NN1). The hyphenated tag in this case, as used in the British National Corpus (BNC), indicates that the algorithm was unable to decide between the VVG (the *-ing* form of a verb) and NN1 (the singular of a common noun), but the preference is for the VVG tag.

Hyphenated tags are employed by Meurman-Solin (2007) as a way of indicating the range of different functions that can be expressed by the word in a diachronic corpus of English, creating, in effect, tags which embody grammaticalization facts. Certainly, there should be no expectation that part-of-speech tagging algorithms will produce identical results. Consider the tags assigned to *rid* in the three sentences in Table 2, based on four automatic tagging programs, where it can be seen that there is no uniform assignment of the part of speech of *rid* in any of the three sentences given. Here we see indications of a re-grammaticalization of a past participle as an adjective, just one example of how any part-of-speech system needs to be critically assessed.

	I am now completely rid of such things.	*You are well rid of him.*	*I got rid of the rubbish.*
CLAWS tagger	past participle	past participle	past participle
Infogistics	verb base	verb base	past participle
FreeLing	adjective	verb base	past participle
(Brill-based) GoTagger	adjective	adjective	adjective

Table 2: Four tagging solutions for English *rid*

Another way in which multiple tags can refer to one word involves multi-word units. For instance, the complex preposition *in terms of* is tagged in the BNC XML as shown in Figure 3 (for expository reasons, we have added line breaks to highlight the annotation's structure).

```
<mw c5="PRP">
        <w c5="PRP" hw="in" pos="PREP">in </w>
        <w c5="NN2" hw="term" pos="SUBST">terms </w>
        <w c5="PRF" hw="of" pos="PREP">of </w>
</mw>
```

Figure 3: The annotation of *in terms of* as a multi-word unit in the BNC XML

Transcription of spoken language presents considerable challenges, at least if one wishes to faithfully highlight features of spoken language (cf. Newman 2008, see also Chapter 12). The annotated transcription in (6), a sample of transcribed spoken language taken from ICE-CANADA illustrates some of this complexity. Overlapping strings are indicated by <[>…</[>, with the complete set of overlapping strings contained within <{>…</{>, stretching across both speaker A and speaker B. The tags <}>…</}> indicate a 'normative replacement' where a repetition of *they* (in casual, face-to-face conversation) is indicated. This annotation allows for searching on the raw data (containing the original two instances of *they*) or on the normalized version (containing one instance of *they* within <=…></=>). The example in (6) illustrates only a tiny fraction of the challenges presented by spoken language. The Great Britain component of the International Corpus of English (ICE-GB) contains syntactic parses for all the data, which make the annotation even more complex.

(6) <$A> <ICE-CAN:S1A-001#34:1:A> I think some of the trippers actually do a bit of the portaging by themselves <}> <-> they> </-> <=> they </=> </}> bring it to the other end and they come back to help the kids with <{> <[> their packs </[>
<$B> <ICE-CAN:S1A-001#35:1:B> <[> I see </[> </{>

The advent of extremely large multimodal corpora such as the corpus created through the Human Speechome Project (90,000 hours of video and 140,000 hours of audio recordings) takes the problems of dealing with audio and video to another level altogether, requiring the development of new kinds of tools to manage the extraordinary amount of data involved (Roy 2009).

Just as with cleaning up and standardizing data, the processes of marking up and annotating typically require more sophisticated tools than mere word-processing tools. For some tasks (e.g., straightforward replacement operations), general-purpose applications such as sophisticated text editors may be sufficient. For some more specialized tasks, ready-made applications with a graphical user interface are available. For example, language-encoding converters (Encoding Master for Windows/Mac, iconv for Unix/Linux, at the time of writing) and annotation software such as ELAN, Transcriber, Soundscriber (Windows) are available (see Chapter 12 on transcription). Some larger and more automatic processes such as part-of-speech tagging, however, would normally be carried out by running scripts in a programming environment, though some Graphical User Interface (GUI) applications are also available (e.g., GoTagger for English and the Windows interface to TreeTagger for English and other languages).

To exemplify at least one application here, TreeTagger is a suite of scripts (currently available for Linux, Windows, and Mac) that would suit the needs of most researchers wanting to tag a corpus for part of speech. Some basic knowledge of programming environments is required to run these scripts, though running them is not a daunting task. To illustrate what is involved, (7) shows the one-line command needed to tag an English sentence with the output directed to the screen as three columns (each word in the input, a tag, and a lemmatized form of the word). The tags are based on the Penn Treebank tagset. In this example, DT = determiner, VBP = non-[3rd person singular present] of a verb, NNS = plural common noun, WDT = Wh-determiner, NN = singular common noun, SENT = sentence closer. It is equally straightforward to tag a whole file or a directory of files. The tagging requires language-specific parameter files which are available for a dozen or so languages (including English, German, Italian, Dutch, Spanish, Bulgarian, Russian, French, Mandarin). TreeTagger includes a training module which allows one to create a new parameter file for any language, trained on a lexicon and a training corpus. A 'chunker' script outputs the tagged words plus some grouping into syntactic constituents. When run on the sentence in (7), for example, the chunker script would insert noun clusters (NC) tags around *some words* and *a sentence* and verb cluster (VC) tags around the one-word verb clusters *are* and *make*. As reported by Schmidt (1994), using TreeTagger to tag for parts of speech in an English corpus achieved over 95% accuracy.

```
(7)   $   echo   'These   are   some   words   which   make   a   sentence.'   |
          cmd/tree-tagger-english
          reading parameters …
          tagging ...
          finished.
             These      DT      these
             are        VBP     be
             some       DT      some
             words      NNS     word
             which      WDT     which
             make       VBP     make
             a          DT      a
             sentence   NN      sentence
             .          SENT    .
```

2.4 Several widely-used corpora

Before turning to how corpora are used, we briefly present here a few widely used corpora with an eye to showcasing different kinds of data and annotation (see Appendix 1 for more information on access to these corpora). Readers should be aware that the Linguistic Data Consortium (LDC, www.ldc.upenn.edu) makes available many high-quality corpora, some free to non-members and others available through an annual subscription. It is also worth mentioning the Child Language Data Exchange System (CHILDES) database and associated tools, the child language component of the TalkBank project. Between them, CHILDES and TalkBank offer a great variety of freely available adult and child language corpora available in various media, with an option of playing streaming audio and video through the internet. TalkBank, for example, includes corpora designed for the study of aphasia, dementia, second language acquisition, conversation analysis, and sociolinguistics. The CHILDES system of transcription and coding has in turn given rise to the Language Interaction Data Exchange System (LIDES) which aims to standardize transcription and coding for spoken multilingual data (LIPPS 2000; Gardner-Chloros, Moyer, and Sebba 2007).

The Brown corpus (Kučera and Francis 1967) holds a unique place in the history of corpus linguistics. It represents the first systematic and, at the time, large-scale attempt to sample written American English containing material which first appeared in print in the year 1961. The corpus, described as a 'Standard Corpus of Present-Day American English' by the authors, has become known as the Brown corpus since it was created at Brown University. The corpus contains approximately one million words in 500 samples of 2000+ words each, divided into fifteen sub-categories shown in Table 3. There is quite a spread of writing styles represented in the corpus, with written language being the clear guiding principle in the collection of data. Drama writing, for example, was excluded on the basis of belonging more to the realm of spoken discourse. Fiction writing was included, as long as there was no more than 50% dialogue. The design of the Brown corpus has been adopted in the creation of a number of other one-million-word English corpora: the Lancaster-Oslo-Bergen corpus (LOB), the Freiburg Brown corpus (FROWN), the Freiburg LOB corpus (FLOB), among others. The corpora mentioned here enable corpus-based comparative studies of American and British written English in 1961 (Brown, LOB), American English in 1961 and 1991 (Brown, FROWN) and British English in 1961 and

1991 (LOB, FLOB).

Genre	Words	% of Total
News	88,000	8.8
Editorials	54,000	5.4
Reviews	34,000	3.4
Religion	34,000	3.4
Skills and Hobbies	72,000	7.2
Lore	96,000	9.6
Belles lettres	150,000	15
Miscellaneous	60,000	6
Learned	160,000	16
General fiction	58,000	5.8
Mystery	48,000	4.8
Science fiction	12,000	12
Adventure	58,000	5.8
Romance	58,000	5.8
Humor	18,000	18
Total	1,000,000	

Table 3: Sub-corpora of the Brown written corpus

The International Corpus of English (ICE) has been mentioned already: It is a global project whereby English language materials from many national varieties of English are being collected and marked up according to common guidelines. The primary aim of ICE is to collect material for comparative studies of English worldwide, based on the adoption of a common corpus size (approximately one million words) and design. As of April 2012, there were 24 varieties of English represented in the project, according to the ICE website. These varieties include better known ones such as Great Britain and USA, as well as lesser known ones such as Malta, Philippines, and Sri Lanka. A full description of the project, as originally conceived, is given in Greenbaum (1996) and Greenbaum and Nelson (1996). A breakdown of the sub-parts of an ICE corpus can be seen in Table 4.

Mode	Genre	Words	% of Total
Spoken (60%)	Private	200000	20
	Public	160000	16
	Unscripted	140000	14
	Scripted	100000	10
Written (40%)	Student Writing	40000	4
	Letters	60000	6
	Academic Writing	80000	8
	Popular Writing	80000	8
	Reportage	40000	4
	Instructional Writing	40000	4
	Persuasive Writing	20000	2
	Creative Writing	40000	4
	Total	1000000	

Table 4: Sub-corpora of the ICE corpora

The Michigan Corpus of Academic Spoken English (MICASE) is a corpus of spoken academic English as recorded at the University of Michigan (Simpson, Briggs, Ovens, and Swales 2002) between 1997 and 2002. It consists of transcriptions of almost 200 hours of recordings, amounting to about 1.8 million words (according to the MICASE website). Individual speech events range in length from 19 to 178 minutes, with word counts ranging from 2,805 words to 30,328 words. Table 5 provides word counts for an untagged version of MICASE in which hyphenated parts of a word and parts of a word separated by apostrophes count as one word.

Genre	Words	% of Total
Small Lectures	333,338	19.7
Large Lectures	251,632	14.8
Discussion Sections	74,904	4.4
Lab Sections	73,815	4.4
Seminars	138,626	7.7
Student Presentations	143,369	8.5
Advising Sessions	35,275	2.1
Dissertation Defenses	56,837	3.4
Interviews	13,015	0.8
Meetings	70,038	4.1
Office Hours	171,188	8.2
Service Encounters	24,691	1.5
Study Groups	129,725	7.7
Tours	21,768	1.3
Colloquia	157,333	9.3
Total	1,695,554	

Table 5: Sub-corpora of the MICASE spoken corpus

The BNC contains a collection of written and transcribed spoken samples of British English reflecting a wide range of language use and totaling about 100 million words. The corpus has been published in various editions: the two most widely used ones (containing the same samples) being the BNC World Edition (2001), marked up in the Standard Generalized Markup Language (SGML), and the BNC XML Edition (2007). Most of the language samples date from the years 1985-1993, but some written language samples were taken from the years 1960-1984. For the 'context-governed' part of the spoken component, data was collected based on particular domains of language usage; for the 'spoken demographic' part, conversations were collected by 124 volunteers recruited by the British Market Research Bureau, with equal numbers of men and women, approximately equal numbers from each age group, and equal numbers from each social grouping. Table 6 provides a breakdown of the sub-parts of the BNC, with size in terms of 'w-units', where a 'w-unit' is similar to an orthographic word of English but may also include some multi-word units, i.e. sequences of orthographic words, such as *a priori*, *of course*, *all of a sudden* etc.

The Corpus of Contemporary American English (COCA) is a corpus of contemporary American English sampled from the years 1990 on (see Davies 2008-, Davies 2011), which is only available via a web interface. The corpus is being added to each year, i.e., it is a 'monitor corpus'. At the time of writing it contains more than 437 million words of text, equally divided among spoken, fiction, popular magazines, newspapers, and academic texts, as shown in Table 7. The spoken samples are taken from transcripts of unscripted conversation from more than 150 different TV and radio programs. The Corpus of Historical American English (COHA) is an equally impressive historical corpus of American English sampled from the period 1810-2009, consisting of more than 400 million words, with the same kind of interface as COCA.

The Buckeye corpus of conversational speech was created primarily to support the study of phonological variation in American English speech (Pitt et al. 2005, 2007). The corpus consists of 40 'talkers' from Columbus, Ohio, who were each interviewed at Ohio State University in 1999-2000. The interviewees were told prior to the interview that the purpose of the interview was 'to learn how people express 'everyday' opinions in conversation, and that the actual topic was not important' (Pitt et al. 2005: 91). Debriefing on the true purpose of the interview and obtaining further consent of the interviewee were carried out after the interview had taken place. The target length of each interview was 60 minutes. The corpus includes high-fidelity WAV files, consists of a total of 305,652 words, and comes with phonemic labeling and orthographic transcription.

Mode	Genre	'w-units'	% of Total 'w-units'
Written (87.9%)	Imaginative	16496420	16.8
	Informative: natural and pure science	3821902	3.9
	Informative: applied science	7174152	7.3
	Informative: social science	14025537	14.3
	Informative: world affairs	17244534	17.5
	Informative: commerce and finance	7341163	7.5
	Informative: arts	6574857	6.7
	Informative: belief and thought	3037533	3.1
	Informative: leisure	12237834	12.4
Spoken: context-governed (6.1%)	Educational/Informative	1646380	1.7
	Business	1282416	1.3
	Public/Institutional	1672658	1.7
	Leisure	1574442	1.6
Spoken: spoken demographic (4.2%)	Respondent Age 0-14	267005	0.3
	Respondent Age 15-24	665358	0.7
	Respondent Age 25-34	853832	0.9
	Respondent Age 35-44	845153	0.9
	Respondent Age 45-59	963483	1.0
	Respondent Age 60+	639124	0.6
		98363783	

Table 6: Sub-corpora of the BNC

Genre	Sub-genre	Words	% of Total
Spoken (20%)	Spoken	90,065,764	20.6
Written (80%)	Fiction	84,965,507	19.4
	Magazine	90,292,046	20.6
	Newspaper	86,670,481	19.8
	Academic	85,791,918	19.6
	Total	437,785,716	100

Table 7: Sub-corpora of the written component of COCA, as of April 2011

The six corpora singled out for discussion here give some sense of the kind of material that linguists work with as corpora. Clearly, there is considerable variation along many parameters as one compares these corpora: specialized (English as spoken in an academic context, informal interview speech, historical data) vs. general (spoken and written language in a variety of contexts); written language vs. speech; relative balance in the size of the main sub-parts of a corpus, as in COCA, vs. skewing in the size of the main sub-parts, as in the BNC; single medium such as electronic texts vs. multi-media. This variability in design also points to a need for caution when making direct comparison across the corpora or when a researcher relies solely upon a particular corpus with its own idiosyncratic design to establish 'baseline' frequencies of occurrence of words or patterns.

Obviously, many more corpora than those mentioned above are available. For instance, Xiao (2008) refers, by our count, to more than 130. Even the category of 'national' corpora alone, i.e., corpora designed to be representative of a range of usage of a national language by native-speakers, includes more than twenty (three for Polish alone) and that number has likely increased in the years since Xiao's overview was published. One particularly important desideratum for the future of corpus linguistics and the neighboring field of natural language processing is to recognize resources in language other than English and to appreciate the need to develop tools and software applicable to all the languages of the world.

3. Using corpora

The previous section discussed a variety of topics concerned with how to create corpora. In this section, we will turn to how to study corpora. In section 3.1, we will briefly introduce the three main corpus-linguistic methods, and in section 3.2, we will discuss the kinds of applications and tools that corpus linguists use in their research.

3.1 Analytical tools of corpus linguistics

Corpus linguistics is inherently a distributional discipline because, essentially, corpora only offer answers to the questions in (8) regarding the distributions of linguistic items:

(8) a. how often and where does something occur in a corpus?
    b. how often do linguistic expressions occur in close proximity to other linguistic expressions?
    c. how are linguistic elements used in their actual contexts?

The following three sections will discuss each of these methods in turn.

3.1.1 Frequency lists and dispersion

Frequency lists are the most basic corpus-linguistic tool. They usually indicate how frequent each word or each *n*-gram (a chain of *n* adjacent words) is in a (part of a) corpus. Examples are shown in the three panels of Table 8.

Words	Frequency
the	62580
of	35958
and	27789

Words	Frequency
yllufdaerd	80
yllufecaep	1
yllufecarg	5

Words	Frequency
of the	4892
in the	3006
to the	1751

to	25600	*yllufecruoser*	8	*on the*	1228
a	21843	*yllufeelg*	1	*and the*	1114
in	19446	*yllufeow*	1	*for the*	906
that	10296	*ylluf*	2	*at the*	832
is	9938	*yllufepoh*	8	*to be*	799
was	9740	*ylluferac*	87	*with the*	783
for	8799	*yllufesoprup*	1	*from the*	720

Table 8: Frequency lists (left panel: words sorted according to frequency; center panel: reversed words sorted alphabetically; right panel: 2-grams sorted according to frequency)

Crucially, this method assumes a working definition of what a word is, which is less straightforward than one may think and less straightforward than many corpus programs' default settings reveal: how many words are *John's book* and *John's at home*, or *isn't it*?

There are a variety of ways in which frequency lists are used and/or modified. First, one has to decide whether one needs the frequency lists of word forms or lemmas: should *run*, *runs*, *running*, and *ran* all be grouped together under the lemma RUN or not? Second, in order to be able to compare frequencies of words from corpora of different sizes, frequencies are often normalized as a ratio of occurrences per million words. Third, comparisons of frequency lists can give rise to interesting data, as when a frequency list of a (usually smaller) specialized corpus is compared to one of a (usually larger) general reference corpus. For example, one can compute for each word in a corpus *w* the percentage p_1 that it makes up of a corpus *c*1 and divide it by the percentage p_2 that *it* makes up in a different corpus c_2, and when you order the resulting relative frequency ratios by their size, the top and the bottom will reveal the words most strongly associated with c_1 and c_2.

It is important to realize how such lists decontextualize each use: one only sees how often, say, *and*, *gracefully*, and *in the* appear, but not where in the file or in which context(s). One way to obtain some information about where in a (part of a) corpus a word occurs is by exploring the *dispersion* of a word. In the left panel of Figure 4, the *x*-axis represents the distribution of the word *perl* in the Wikipedia entry for 'Perl' and each occurrence of the word *perl* is indicated by a vertical line. It is very obvious that the highest density of occurrence occurs at the end of the file (where the reference section is located). In the right panel, the corpus has been divided into 10 equally-sized parts and a barplot represents the frequencies of *perl* in the 10 bins. Again, *perl* is particularly clustered in the final 10% of the file. Also, the dispersion of a word in a corpus can be quantified, and the right panel provides two such measures of dispersion, Juilland's *D* and chi-square. Such measures are particularly useful because two words may have (about) the same frequency of occurrence but one of them may be evenly spread out through the corpus (reflecting its status as a common word) while the other may be much more unevenly distributed (reflecting its status as a more specialized word that is just very frequent in particular registers or topics). An example would be the words *having* and *government*, which occur roughly equally frequently in the BNC Baby, but the former is much more evenly spread out throughout the corpus. Similarly, words may be very unequal in frequency but still equally dispersed; for instance, *any* and *the* have very different frequencies in the BNC Baby corpus (4563 and 201,940 respectively), but dispersion measures reflect that both of them are function words; see Gries (2008) for more discussion.

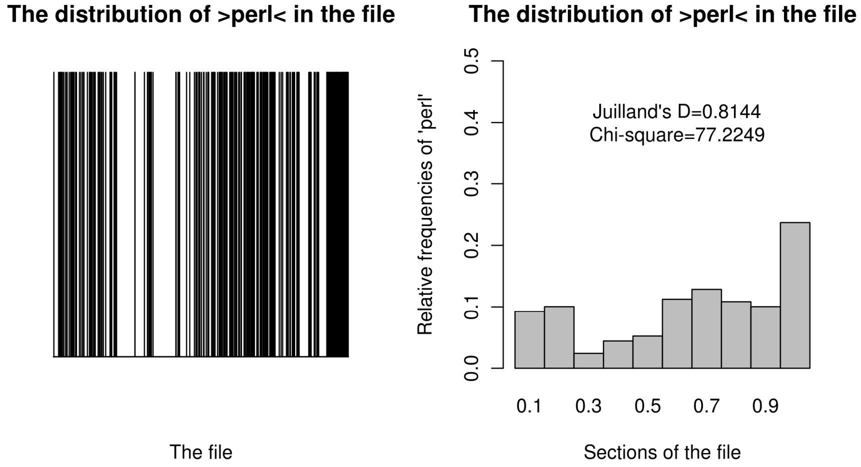

Figure 4: Two ways of representing the dispersion of a word (*perl*) in a file

3.1.2 Collocations

Just like dispersion plots, the second most basic corpus-linguistic tool focuses on a particular linguistic element *w* (typically a word) and provides some information on where *w* is used. However, unlike dispersion plots, the information about where *w* occurs does not use the location in the file/corpus as a reference, but lists which words are most frequently found around *w*. The standard format in which collocations are displayed is exemplified in Table 9. Such tables are read vertically – not horizontally – such that the frequencies listed reveal how often a word occurs in a particular position around the node word, here *general* or *generally*. You can immediately see how words are used and which larger expression it enters into: meaningful collocations such as *General Motors* (found 31 times), *Attorney General* (23), *Secretary General* (16), *General Assembly* (15), and others immediately stand out.

Left 2	Freq L2	Left 1	Freq L1	Node	Right 1	Freq R1	Right 2	Freq R2
the	53	*the*	121		*motors*	31	*of*	52
of	28	*in*	54		*and*	15	*the*	30
and	20	*a*	40		*assembly*	15	*and*	25
to	20	*of*	31		*the*	14	*in*	12
in	15	*attorney*	23		*of*	12	*to*	12
a	13	*and*	19		*public*	12	*that*	11
it	12	*secretary*	16		*business*	10	*as*	11
by	9	*is*	12		*s*	10	*with*	8
is	8	*more*	10		*ized*	9	*for*	8
be	7	*was*	10		*izations*	7	*a*	8

Table 9: Excerpt of a collocate display of *general*/*generally*

In a small table like Table 9, these few interesting collocations can immediately be identified, but it is also obvious that many collocations involve function words (*the*, *and*, *in*, *to*, *a*, …) that are so widely dispersed that they will show up in every word's vicinity. Corpus linguists have therefore developed a variety of so-called association measures, most of which essentially quantify how much more frequent a collocate is around a word of interest *w* than one would expect given *w*'s and that collocate's overall frequency in a corpus. In such an approach, collocates are then ranked by their association strength rather than their overall frequency; widely used measures are Mutual Information *MI*, *t*, the log-likelihood ratio, and the Fisher-Yates exact test. Space does not permit us to discuss this in more detail, but see Wiechmann (2008) for a comprehensive discussion.

3.1.3 Concordances

Probably, the most common corpus-linguistic tool currently used is the KWIC (for 'key word in context') concordance, i.e. a display of the word of interest in its immediate context. Consider Table 10 for part of a KWIC concordance of *alphabetic* and *alphabetical*.

File	Line	Preceding context	Match	Subsequent context
A6S	687	and the invention of	alphabetic	writing.
BN9	81	and seven first-class counties taken in	alphabetical	order of rotation.
H99	1583	seeks to negotiate the problems of the	alphabetical	subject approach as outlined in
EES	788	a word is a contiguous sequence of	alphabetic	characters.
B2M	196	provided the basis for an	alphabetical	sort within each functional category.
CHA	3656	and then put them into	alphabetical	order.'
EA3	516	to isolate the cultural consequences of	alphabetic	literacy'(ibid. p. 42).
F7G	656	But you would put it in	alphabetical	order
CLH	1422	most languages with writing systems	alphabetic	fingerspelling has been available for over
KCY	2439	again I can put the type in	alphabetical	ascending order

Table 10: Excerpt of a concordance display of *alphabetic* and *alphabetical*

This is the most comprehensive display, showing exactly how the two adjectives are used, but the large amount of information comes at the cost that this display usually needs a human analyst for interpretation whereas frequencies and collocate displays can often be processed further automatically. This type of table would normally be saved into a tab-delimited text file, which can then be opened with a spreadsheet software (e.g., LibreOffice Calc) so that every match, i.e. every row, can be annotated for linguistic variables of interest. The resulting file would exhibit the case-by-variable format discussed in Chapter 15 and can then be loaded into statistics software and analyzed as discussed there.

With increasingly complex use of concordancing, it quickly becomes necessary to learn about regular expressions, mentioned earlier. This is because while one can search for the two forms of *alphabetic* and *alphabetical* separately, the manual spelling-out of search patterns becomes cumbersome if many thousands of verb lemmas are being retrieved. Even worse, there are many applications where the desired result cannot even be spelt out *a priori*: if you want to find all words ending in *-ic* or *-ical*, then you cannot always predict which forms might exist in usage in a given corpus; the same holds if you want to find all verbs ending in *ing* or *in'*. Regular expressions, a technique for describing (sets of) character sequences, can handle such cases. Table 11 lists a few simple examples that showcase the potential of regular expressions (examples are based on SGML/XML annotation of the BNC).

<table>
<tr><th>Regular expression</th><th>'Translation'</th></tr>
<tr><td><code>colou?r</code></td><td>finds both <i>color</i> and <i>colour</i> because the <code>u</code> is made optional by the <code>?</code></td></tr>
<tr><td><code>smokin[g']</code></td><td>finds both <i>smoking</i> and <i>smokin'</i> because either <code>g</code> or <code>'</code> are allowed after the <code>n</code></td></tr>
<tr><td><code>\bg[eo]t(s|t(ing|en))?\b</code></td><td>finds at least <i>get</i>, <i>gets</i>, <i>getting</i>, <i>got</i>, and <i>gotten</i> as individual words</td></tr>
</table>

`[-\w]+ly\b`	sequences of word characters and hyphens ending in *ly*
`<w (dtq|pnq).*?<w prp[^<]*?<c pun\?`	*wh*-words followed by other words until a preposition before a question mark (to find cases of preposition stranding, such as *What are you talking about?*)

Table 11: Examples of regular expressions

3.2 Tools for analysis in corpus linguistics

We have come to expect a range of basic features relevant to a corpus-based analysis, as listed in (9), and consequently there is an expectation that software tools will incorporate some selection of these.

(9) a. open multiple files
b. accept a variety of language encodings, especially unicode
c. calculate frequency of words, parts of words, sequences of words etc.
d. calculate frequency of parts of speech in a part-of-speech tagged corpus
e. calculate frequency of patterns allowing for wild card searches
f. return concordance lines for a search pattern (word, phrase, part of speech)
g. return concordance lines with variable length of lines
h. return collocates of a search pattern (word, phrase)
i. calculate measures of strength of association between words
j. return a list of *n*-grams
k. save and export results

Four different kinds of approaches are available to corpus linguists, only the fourth of which covers all the functionality mentioned in (9) and more.

The most restricted of these approaches arises when a corpus is only available via a web interface, as is currently the case with BNCWeb, MICASE, COCA, and many others. Here, the user is completely dependent on the functionality made available in the interface and the correctness of what is made available. While the search facilities of many online corpora are far-reaching, studies that require extensive frequency information or large amounts of contexts usually cannot be undertaken with such corpora.

Second, a situation often more useful to the analyst arises when a corpus can be installed on one's own hard drive and comes with a specific software to explore that corpus. For example, the ICE-GB comes with a tool designed specifically for it (ICECUP III, see Nelson, Wallis, and Aarts 2002) and which allows inspection of many features of the corpus. As another example, the BNC XML edition currently comes with Xaira searching software. In such cases, the advantages are that the linguist has the whole corpus available for more individual queries and that the corpus software is tailored to the precise format of the corpus. However, this type of corpus software is sometimes not as user-friendly as it could be, users are still restricted to the functionality of the program, and the ability to work with one corpus software does not transfer to other corpora.

Third, and perhaps most widely used, the corpus linguist has the corpus on his/her hard drive and uses a ready-made general corpus program for retrieval and other operations. Apart from some commercial applications that are restricted to the Microsoft Windows operating system (e.g., Wordsmith), several free alternatives are available, the most useful of which is perhaps AntConc, because it is the only tool we are aware of that runs on the three major operating systems, is good at handling different encodings, and possesses powerful regular expressions that, unlike nearly all other currently available tools (including the commercial ones), allow it to handle many kinds of annotations flexibly. AntConc has a built-in Keywords feature which identifies words overused in one corpus by reference to another corpus. While corpus tools like AntConc allow parallel analysis of disparate corpora, users are still dependent on the functionality that is included in the programs. This also means, for example, that hardly any of the widely-used ready-made programs can read CHAT files well, an annotation format widely used in language acquisition research and the CHILDES database mentioned above.

The fourth and final scenario, one that is becoming increasingly common, involves researchers having corpora on their hard drive and using general purpose programming languages to process, manipulate, and search files. We devote the next section to this topic.

3.3 Programming tools for corpus linguistics

The huge advantage of programming languages is that they are immensely more versatile and powerful than any ready-made software. This allows researchers to pursue research more efficiently, creatively, and within one environment (as opposed to having to learn and using different applications for, say, web-crawling, cleaning up files, standardizing them, retrieving concordances, annotating them, analyzing them statistically, and plotting some graphs). There is a well-known downside to using programming languages and that is the learning curve for the novice user. However, the potential benefits to be gained from persevering and achieving a basic and comfortable literacy in a programming language far outweigh, in our opinion, any learning pains. And there are two additional considerations to bear in mind when thinking about the pros and cons of investing time in learning programming languages: (1) there is a vast number of ways in which programming knowledge can be put to good use in dealing with digital information quite apart from corpus linguistics; (2) once you have learned one programming language, like R, then you typically have some advantage when it comes to learning another one.

Typically, programming languages can be installed on any modern computer desktop computer or laptop; they may have to be installed as stand-alone applications or they may be already included as part of the computer's installed software, e.g., Perl and Python are bundled with the Mac OS. Examples of well-known programming languages are: Perl, C#, Java, PHP, Python, and Ruby. While Perl was probably the most widely used programming language for many years, an increasing number of researchers are now using Python and R, which therefore deserve brief exemplification here. Both Python and R are freely downloadable and available as cross-platform installations (Linux/Unix, Mac OS, Windows). A researcher can choose one or more GUIs for each of these languages to create a more friendly or helpful interface (e.g., color coding in the script, help or documentation available through pull-down menus, etc.).

For the purposes of corpus linguistics, the comprehensive package of Python tools known as the Natural Language Toolkit (NLTK) has many attractive features. The best introduction to NLTK is Bird, Klein, and Loper (2009), also currently available as a free online eBook at the NLTK website; Perkins (2010) is a useful additional text. Figure 5 shows a log of a session working with NLTK and illustrates just a sample of the functions which are available in this module. In this session, a directory of two English .txt files (downloaded from Project Gutenberg and pre-processed using jEdit) is loaded as a corpus with the name 'MyFiles' (lines 3-4). One can obtain a list of all the files that make up the corpus (line 5). In this case, there are just two files: one being the Project Gutenberg file for the novel *Emma* and another for the novel *Pride* *and Prejudice*, both by Jane Austen. The corpus consisting of these two files is broken down into a list of words (line 6) and then a list of the first ten words can be displayed (line 7). As can be seen from the display of the first ten words, the files have not been pre-processed and some metadata about Project Gutenberg appears as the first ten words. Similarly, one can break the corpus into sentences (line 8) and view the first three sentences (line 9), or paragraphs (line 10) and view the first three paragraphs (line 11). Further commands can produce a set of the first ten concordance lines based on the search term *friend* (line 13), words which occur in similar contexts as *friend* (line 14), and significant bigrams (line 15). It is possible to add part-of-speech tags (not always accurate) to create a tagged corpus MyFiles_tag (line 16) and print out the first ten words and punctuation marks of the first tagged sentence (= sentence 13 of the corpus) of Jane Austen's *Emma*.

```
1.  >>> import nltk
2.  >>> from nltk.corpus import PlaintextCorpusReader
3.  >>> corpus_root = '/Users/Myname/Desktop/MyFiles/'
4.  >>> MyFiles = PlaintextCorpusReader(corpus_root, '.*.txt')
5.  >>> MyFiles.fileids()
    ['Emma.txt', 'Pride_and_Prejudice.txt']
6.  >>> words = MyFiles.words()
7.  >>> words[:10]
    ['The', 'Project', 'Gutenberg', 'EBook', 'of', 'Emma', ',', 'by', 'Jane', 'Austen']
8.  >>> sents = MyFiles.sents()
9.  >>> sents[:3]
    [['The', 'Project', 'Gutenberg', 'EBook', 'of', 'Emma', ',', 'by', 'Jane', 'Austen'],
    ['This', 'eBook', 'is', 'for', 'the', 'use', 'of', 'anyone', 'anywhere', 'at', 'no', 'cost',
    'and', 'with', 'almost', 'no', 'restrictions', 'whatsoever', '.'], ['You', 'may', 'copy',
    'it', ',', 'give', 'it', 'away', 'or', 're', '-', 'use', 'it', 'under', 'the', 'terms', 'of',
    'the', 'Project', 'Gutenberg', 'License', 'included', 'with', 'this', 'eBook', 'or',
    'online', 'at', 'www', '.', 'gutenberg', '.', 'org']]
10. >>> paras = MyFiles.paras()
11. >>> paras[:3]
    (results omitted here)
12. >>> words1 = nltk.Text(words)
13. >>> words1.concordance("friend", lines = 10)
        Building index...
        Displaying 10 of 289 matches:
        family , less as a governess than a friend , very fond of both daughters , but
         , they had been living together as friend and friend very mutually attached ,
         been living together as friend and friend very mutually attached , and Emma d
        n the wedding - day of this beloved friend that Emma first sat in mournful tho
         every promise of happiness for her friend . Mr . Weston was a man of unexcept
        derer recollection . She had been a friend and companion such as few possessed
        the change ?-- It was true that her friend was going only half a mile from the
        as not only a very old and intimate friend of the family , but particularly co
        el so much pain as pleasure . Every friend of Miss Taylor must be glad to have
        t Smith ' s being exactly the young friend she wanted -- exactly the something
14.  >>> words1.similar('friend')
        Building word-context index...
        father sister mother own family daughter letter mind time brother aunt
        wife life and heart way side cousin eyes feelings
15.  >>> words1.collocations()
        Building collocations list
        Frank Churchill; Lady Catherine; Miss Woodhouse; Project Gutenberg;
        young man; Miss Bates; Miss Fairfax; every thing; Jane Fairfax; great
        deal; dare say; every body; Sir William; Miss Bingley; John Knightley;
        Maple Grove; Miss Smith; Miss Taylor; Robert Martin; Colonel
        Fitzwilliam
16.  >>> MyFiles_tag=[nltk.pos_tag(sent) for sent in sents]
17.  >>> MyFiles_tag[13][:10]
        [('Emma', 'NNP'), ('Woodhouse', 'NNP'), (',', ','), ('handsome', 'NN'), (',',
        ','),  ('clever', 'RB'), (',', ','), ('and', 'CC'), ('rich', 'JJ'), (',', ',')]
```

Figure 5: Python session illustrating some functions in NLTK (See Bird, Klein, and Loper 2009 and Perkins 2010)

R is an open-source programming language and environment originally designed for statistical computing and graphics, but with all the functionality of 'normal' multi-purpose programming languages including loops, conditional expressions, text processing with and without regular expressions, etc. Figure 6 exemplifies how very easily a rough frequency list can be created in just four lines of code in a short R session: first, a corpus file is loaded (line 1), then it is split up into words (in a somewhat simplistic way, line 2), then R computes a sorted frequency list of the whole file (line 3) and prints out the 30 most frequent words and frequencies (line 4). Then, two of Zipf's laws are tested by (i) plotting words' lengths against their frequencies (line 5; note that the frequencies are logged in order to better represent the distribution of frequencies in a corpus) and adding a summary line (line 6), and by (ii) plotting words' frequency ranks against their (logged) frequencies (line 7) and adding a summary line (line 8).

```
1.  corpus.file <- scan("Brown1_G.txt", what=character(0), sep="\n")
2.  words <- unlist(strsplit(corpus.file, "\\W+", perl=TRUE))
3.  freq.list <- sort(table(words), decreasing=TRUE)
4.  freq.list[1:30]
     words
      the    of   and    to     a    in  that    is   was   his   for    he    as    it  with
     9790  6363  4320  4116  3319  3100  1905  1795  1467  1342  1199  1182  1159  1069  1063
     The      s     I    be   not   had    by    on which  from   are    at  have  this    or
     948    929   871   846   819   804   797   768   679   651   647   633   628   627   588
5.  plot(nchar(names(freq.list)) ~ log(freq.list), xlab="Log word frequency", ylab="Word length
     in characters")
6.  lines(lowess(nchar(names(freq.list)) ~ log(freq.list)))
7.  plot(log(rank(-freq.list)) ~ log(freq.list), xlab="Log word frequency", ylab="Log rank
     frequency")
8.  lines(lowess(log(rank(-freq.list)) ~ log(freq.list)))
```

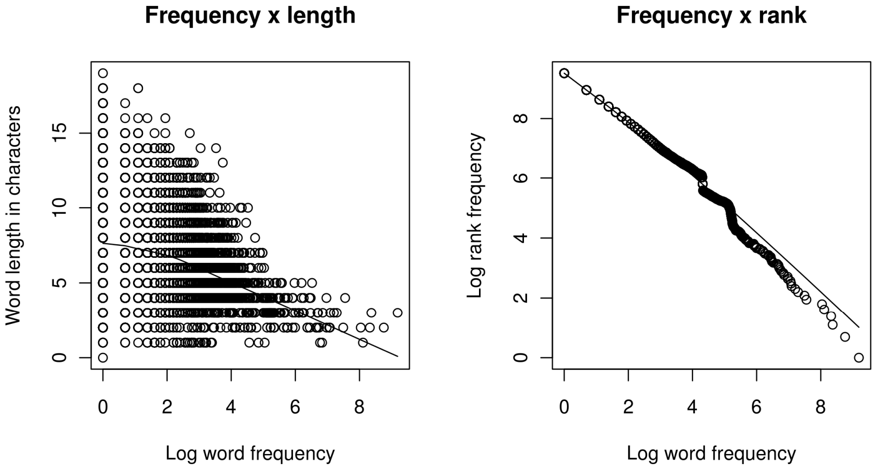

Figure 6: R session to create a frequency list of a file from the Brown corpus and the resulting plots (See Gries 2009 for an introduction).

Given that corpora continuously increase in size and diversity, it is becoming increasingly important that corpus linguists use tools that are not restricted to particular formats, encodings, sizes, or other design factors, and recent changes show that the field is making great strides to this end. If this trend continues, the field will transform into an even more exciting discipline and contribute more than its share to insightful studies of all aspects of language.

References

Beal, Joan C., Karen P. Corrigan, and Hermann L. Moisl (eds.) 2007a. *Creating and digitizing* *language corpora*. Volume 1: *Synchronic databases*. Basingstoke, U.K. and New York: Palgrave Macmillan.

Beal, Joan C., Karen P. Corrigan, and Hermann L. Moisl (eds.) 2007b. *Creating and digitizing language corpora*. Volume 2: *Diachronic databases*. Basingstoke, U.K. and New York: Palgrave Macmillan.

Bird, Steven, Ewan Klein, and Edward Loper 2009. *Natural language processing with Python: Analyzing text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly Media.

Berkenfield, Catie 2001. The role of frequency in the realization of English *that*. In Joan L. Bybee and Paul J. Hopper (eds.) *Frequency and the emergence of linguistic structure*, 281-307. Philadelphia: John Benjamins.

Davies, Mark 2008-. *The Corpus of Contemporary American English (COCA).* Available online at www.americancorpus.org.

Davies, Mark 2011. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25(4): 447-65.

Fiorentino, Giuliana 2009. The ordering of adverbial and main clauses in spoken and written Italian. In Barbara Lewandowska-Tomaszczyk and Katarzyna Dziwirek (eds.), *Studies in Cognitive Corpus Linguistics*, 207-22. Frankfurt am Main: Peter Lang.

Gardner-Chloros, Penelope, Melissa Moyer, and Mark Sebba 2007. In Beal, Corrigan, and Moisl (eds.), Volume 1, 91-120.

Gilquin, Gaëtanelle and Stefan Th. Gries 2009. Corpora and experimental methods: a state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1): 1-26.

Granath, Solveig 2007. Size matters – or thus can be meaningful structures be revealed in large corpora. In Roberta Facchinetti (ed.), *Corpus linguistics 25 years on*, 169-185. Amsterdam and New York: Rodopi.

Greenbaum, Sidney (ed.) 1996. *Comparing English worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Greenbaum, Sidney and Gerald Nelson 1996. The International Corpus of English (ICE) Project. *World Englishes* 15(1): 3-15.

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4): 403-37.

Gries, Stefan Th. 2009. *Quantitative corpus linguistics with R: a practical introduction*. London and New York: Routledge, Taylor and Francis Group.

Hunston, Susan 2008. Collection strategies and design decisions. In Lüdeling and Kytö (eds.), Volume 1, 154-68.

Johansson, Christine and Christer Geisler 2009. The Uppsala Learner English Corpus: A new corpus of Swedish high school students' writing. In A. Saxena and Å. Viberg (eds.), *Multilingualism: proceedings of the 23rd Scandinavian conference of linguistics,* 181-90. Uppsala: Acta Universitatis Upsaliensis.

Johansson, Christine and Christer Geisler 2011. Syntactic aspects of the writing of Swedish L2 learners of English. In John Newman, Harald Baayen, and Sally Rice (eds.), *Corpus-based Studies in Language Use, Language Learning, and Language* Documentation, 139-73. Amsterdam: Rodopi Press.

Kilgarriff, Adam and Gregory Grefenstette 2006. Introduction to the special issue on the web as corpus. *Computational Linguistics* 29(3): 333-47.

Kučera, H. and W. Francis. 1967. *Computational analysis of present-day English*. Providence: Brown University Press.

LIPPS - Language Interaction in Plurilingual and Plurilectal Speakers Group (2000). A Document for Preparing and Analysing Language Interaction Data. Special issue of the *International Journal of Bilingualism* 4.2.

Lüdeling, Anke and Merja Kytö (eds.). 2008a. *Corpus linguistics: An international handbook.* Volume 1. Berlin and New York: Mouton de Gruyter.

Lüdeling, Anke and Merja Kytö (eds.). 2008b. *Corpus linguistics: An international handbook.* Volume 2. Berlin and New York: Mouton de Gruyter.

McEnery, Tony, Richard Xiao, and Yukio Tono 2006. *Corpus-based language studies: An advanced resource book*. Milton Park: Routledge.

McEnery, Tony, and Andrew Hardie. 2012. *Corpus linguistics: Method, theory, and practice.* Cambridge: Cambridge University Press.

Meurman-Solin, Anneli 2007. The manuscript-based diachronic corpus of Scottish correspondence. In Beal, Corrigan, and Moisl (eds.), Volume 2, 127-47.

Nelson, G., S. Wallis, and B. Aarts (2002). *Exploring natural language: working with the British component of the international corpus of English*. Amsterdam and Philadelphia: John Benjamins.

Newman, John 2008. Spoken corpora: Rationale and application. *Taiwan Journal of Linguistics* 6(2): 27-58.

Newman, John and Georgie Columbus. 2009. Education as an over-represented topic in the ICE corpora [Part II]? Presentation for the 15th International Conference of the International Association for World Englishes (IAWE), Cebu City, Philippines.

Newman, John and Georgie Columbus (2010). *The ICE-Canada Corpus. Version 1*. Retrieved from http://ice-corpora.net/ice/download.htm.

Ostler, Nicholas 2008. Corpora of less studied languages. In Lüdeling and Kytö (eds.), Volume 1, 457-83. Berlin and New York: Mouton de Gruyter.

Perkins, Jacob 2010. *Python text processing with NLTK 2.0 cookbook*. Birmingham, UK: Packt Publishing.

Pitt, Mark A., Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond 2005. The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication* 45(1): 89-95.

Pitt, M.A., L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier 2007. Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu]. Columbus, OH: Department of Psychology, Ohio State University (Distributor).

Roy, Deb 2009. New horizons in the study of child language acquisition. In *Proceedings of Interspeech 2009*. Brighton, England. Retrieved April 7, 2011 from www. media.mit.edu/cogmac/publications/Roy_interspeech_keynote.pdf.

Schmid, Helmut 1994. Probabilistic part-of-speech tagging using decision trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Simpson, Rita C., S. L. Briggs, J. Ovens, and John M. Swales 2002. *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

Sinclair, John 2005. Corpus and text – basic principles. In Martin Wynne (ed.), *Developing linguistic corpora: A guide to good practice*, 1-16. Oxford: Oxbow Books.

Thompson, Sandra A. and Paul J. Hopper 2001. Transitivity, clause structure and argument structure. In Joan L. Bybee and Paul J. Hopper (eds.) *Frequency and the emergence of linguistic structure*, 27-60. Philadelphia: John Benjamins.

Wiechmann, Daniel. 2008. On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory* 4(2): 253-90.

Wynne, Martin (ed.) 2005. *Developing linguistic corpora: A guide to good practice*. Oxford: Oxbow Books. Retrieved April 16, 2011 from www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm.

Xiao, Richard 2006. Xaira – an XML-aware indexing and retrieval architecture. *Corpora* 1(1): 99-103.

Xiao, Richard 2008. Well-known and influential corpora. In Lüdeling and Kytö (eds.), Volume 1, 383-457.

Appendix 1. Corpora referred to in this chapter

Baby BNC. Details of this collection of corpora, with XAIRA, can be found at www.natcorp.ox.ac.uk/corpus/babyinfo.html. Payment required.

BNC. The British National Corpus can be accessed online at no cost through two interfaces: Mark Davies' website at corpus.byu.edu/bnc and William Fletcher's Phrases in English site at phrasesinenglish.org/. Information on purchasing the corpus (and other releases of samples of the BNC) may be found at www.natcorp.ox.ac.uk. Online access to the BNC is also provided for BNC licensees. A full description of the BNC can be found in the Reference Guide for the British National Corpus (XML Edition) at www.natcorp.ox.ac.uk/docs/URG/.

Brown. The Brown corpus may be downloaded at no cost through the 'language commons' collection at http://www.archive.org/details/BrownCorpus and the nltk package of Python at www.nltk.org. It can be searched online through the LDC at online.ldc.upenn.edu/login.html and the Corpus Concordance English at www.lextutor.ca/concordancers/concord_e.html. The corpus is included in the ICAME Corpus Collection available on CD-ROM through ICAME at www.icame.uib.no/newcd.htm. Different versions of the corpus may segment the corpus differently. The language commons version contains the 500 x 2,000 word samples as separate files; the ICAME version contains fifteen files reflecting the sub-categories in Table 1. Both tagged and untagged versions of the corpus are included in the ICAME Corpus Collection; an XML tagged version of the Brown is included as part of BabyBNC v.2 which is available at www.natcorp.ox.ac.uk.

Buckeye. The Buckeye corpus, together with a manual, may be obtained at no cost by following instructions on the homepage of the project at buckeyecorpus.osu.edu.

CallHome American English Speech corpus is available at cost through the Linguistic Data Consortium at www.ldc.upenn.edu.

CHILDES. The Child Language Data Exchange System, developed by Brian MacWhinney, is accessed freely at childes.psy.cmu.edu.

COCA. The Corpus of Contemporary American English is freely accessible online at www.americancorpus.org/ but not distributed as a corpus. A full description of the corpus can be found at this website.

COHA. The Corpus of Historical American English is freely accessible online at corpus.byu.edu/coha/ but not distributed as a corpus. A full description of the corpus can be found at this website.

FLOB. The Freiburg LOB corpus is included in the ICAME Corpus Collection and is described in the accompanying manual. Available for purchase from khnt.aksis.uib.no/icame.

FROWN. The Freiburg BROWN corpus is included in the ICAME Corpus Collection and is described in the accompanying manual. Available for purchase from khnt.aksis.uib.no/icame.

ICAME. The International Computer Archive of Modern and Medieval English s Collection is available for purchase on CD-ROM at icame.uib.no/newcd.htm.

ICE. Information on obtaining corpora of the International Corpus of English is available through the ICE website at ice-corpora.net/ice/index.htm. At the time of writing, ICE corpora for Canada, Jamaica, Hong Kong, East Africa, India, Singapore, and Philippines are available at no cost and can be downloaded from the ICE website; ICE corpora for Great Britain, New Zealand, and Ireland are available on CD ROM at relatively low cost.

ICE-CANADA. The Canadian component of the International Corpus of English is freely available at at ice-corpora.net/ice/index.htm and is described more fully in Newman and Columbus (2010).

LOB. The Lancaster-Bergen-Oslo Corpus (written) corpus is included in the ICAME Corpus Collection and is described in the accompanying manual. Available for purchase from khnt.hit.uib.no/icame.

MICASE. The Michigan Corpus of Academic Spoken English is freely accessed online at www.quod.lib.umich.edu/m/micase. A full description of the MICASE project and the corpus can be found in the MICASE manual available at www.micase.elicorpora.info. Individual XML transcripts of the files can be downloaded at no cost. A version of the whole corpus can also be purchased through the MICASE website.

Uppsala Learner English Corpus. This corpus is described in Johansson and Geisler (2009, 2011).

TalkBank. This collection of corpora and transcripts is accessed freely at talkbank.org.

TIMIT Acoustic-Phonetic Continuous Speech Corpus. Available for purchase through the Linguistic Data Consortium.

Appendix 2. Tools/software referred to in this chapter

AntConc Concordancer. www.antlab.sci.waseda.ac.jp/software.html

CES. Corpus Encoding Standard. www.cs.vassar.edu/CES

CLAWS. The Constituent Likelihood Automatic Word-tagging System (CLAWS) tagset(s). ucrel.lancs.ac.uk/claws

ELAN. EUDICO Linguistic Annotator software. www.lat-mpi.eu/tools/elan

FreeLing. nlp.lsi.upc.edu/freeling

GoTagger. web4u.setsunan.ac.jp/Website/GoTagger.htm
For notes in English on this Windows-only tagger:
hi.baidu.com/seanxpq/blog/item/7aa9db03f8bffc0f738da50e.html

HTTrack. www.httrack.com

Infogistics. www.infogistics.com

jEdit. www.jedit.org

LibreOffice Calc. www.libreoffice.org

NLTK. Natural Language Toolkit. www.nltk.org. An electronic version of the accompanying book (Bird, Klein, and Loper 2009) is also available at this site.

Penn Treebank Tagset. www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html

Project Gutenberg. www.gutenberg.org

R. www.R-project.org

Sitesucker. www.sitesucker.us/mac.html

Southern Oral History Program. docsouth.unc.edu/sohp

Transcriber. trans.sourceforge.net

TreeTagger. www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
For the Windows interface to TreeTagger:
www.smo.uhi.ac.uk/~oduibhin/oideasra/interfaces/winttinterface.htm

Wordsmith. Corpus linguistic software available for purchase at www.lexically.net/wordsmith/

XCES. Corpus Encoding Standard in XML format. www.xces.org