III. Existing corpora

20. Well-known and influential corpora

1. Introduction
2. National corpora
3. Monitor corpora
4. Corpora of the Brown family
5. Synchronic corpora
6. Diachronic corpora
7. Spoken corpora
8. Academic and professional English corpora
9. Parsed corpora
10. Developmental and learner corpora
11. Multilingual corpora
12. Non-English monolingual corpora
13. Well-known distributors of corpus resources
14. Conclusion
15. Appendix: URLs
16. Literature

1. Introduction

As corpus building is an activity that takes times and costs money, readers may wish to use ready-made corpora to carry out their work. However, as a corpus is always designed for a particular purpose, the usefulness of a ready-made corpus must be judged with regard to the purpose to which a user intends to put it. There are thousands of corpora in the world, but most of them are created for specific research projects and are not publicly available. This article introduces well-known and influential corpora, which are grouped in terms of their primary uses so that readers will find it easier to choose corpus resources suitable for their particular research questions. Note, however, that overlaps are inevitable in our classification. It is used in this article simply to give a better account of the primary uses of the relevant corpora. The higher number of English corpora covered here might reflect the fact that English was the forerunner in corpus research, though as we will see shortly, many other languages are catching up. Information on the web site addresses for the corpora discussed in this article are given in the appendix.

2. National corpora

National corpora are normally general reference corpora which are supposed to represent the national language of a country. They are balanced with regard to genres and domains that typically represent the language under consideration. While an ideal national corpus should cover proportionally both written and spoken language, most exist-

ing national corpora and those under construction consist only of written data, as spoken data is much more difficult and expensive to capture than written data. This section introduces a number of major national corpora.

2.1. The British National Corpus

The first and best-known national corpus is perhaps the British National Corpus (BNC), which is designed to represent as wide a range of modern British English as possible so as to "make it possible to say something about language in general" (Burnard 2002, 56). The BNC comprises approximately 100 million words of written texts (90%) and transcripts of speech (10%) in modern British English. Written texts were selected using three criteria: "domain", "time" and "medium". Domain refers to the content type (i. e. subject field) of the text; time refers to the period of text production, while medium refers to the type of text publication such as books, periodicals or unpublished manuscripts. Table 20.1 summarizes the distribution of these criteria (see Aston/Burnard 1998, 29–30).

Tab. 20.1: Composition of the written BNC

Domain	%	Date	%	Medium	%
Imaginative	21.91	1960–74	2.26	Book	58.58
Arts	8.08	1975–93	89.23	Periodical	31.08
Belief and thought	3.40	Unclassified	8.49	Misc. published	4.38
Commerce/Finance	7.93			Misc. unpublished	4.00
Leisure	11.13			To-be-spoken	1.52
Natural/pure science	4.18			Unclassified	0.40
Applied science	8.21				
Social science	14.80				
World affairs	18.39				
Unclassified	1.93				

The spoken data in the BNC was collected on the basis of two criteria: "demographic" and "context-governed". The demographic component is composed of informal encounters recorded by 124 volunteer respondents selected by age group, sex, social class and geographical region, while the context-governed component consists of more formal encounters such as meetings, lectures and radio broadcasts recorded in four broad context categories. The two components of spoken data complement each other, as many types of spoken text would not have been covered if demographic sampling techniques alone were used in data collection. Table 20.2 summarizes the composition of the spoken BNC. Note that in the table, the first two columns apply to both demographic and context-governed components while the third column refers to the latter component alone.

In addition to part-of-speech (POS) information (see article 24 for POS tagging), the BNC is annotated with rich metadata (i. e. contextual information) encoded according

Tab. 20.2: Composition of the spoken BNC

Region	%	Interaction type	%	Context-governed	%
South	45.61	Monologue	18.64	Educational/informative	20.56
Midlands	23.33	Dialogue	74.87	Business	21.47
North	25.43	Unclassified	6.48	Institutional	21.86
Unclassified	5.61			Leisure	23.71
				Unclassified	12.38

to the TEI guidelines, using ISO standard 8879 (i. e. SGML, see article 22). Because of its generality, as well as the use of internationally agreed standards for its encoding, the BNC corpus is a useful resource for a very wide variety of research purposes, in fields as distinct as lexicography, artificial intelligence, speech recognition and synthesis, literary studies and, of course, linguistics. There are a number of ways one can access the BNC corpus. It can be accessed online remotely using the BNC Online service (see appendix), the BNCWeb interface (see appendix), or more recently the BNCWeb CQP Edition (see appendix), which integrates the strengths of both BNCWeb and CQP (Corpus Query Processor) (cf Hoffmann/Evert 2006). Another online interface to the BNC (World Edition) is "Variation in English Words and Phrases" (i. e. VIEW, see appendix). Users interested in phrases in English can also access the BNC (World Edition) via "Phrase in English" (PIE, see appendix), a simple yet powerful interface that allows users to study words and phrases up to six words long. Alternatively, if a local copy the corpus is available, the BNC can be explored using corpus exploration tools such as WordSmith (Scott 2004) and Xaira (see below).

The current version of the full release of the BNC is BNC-2, the World Edition. This version has removed a small number of texts (less than 50) which restrict the worldwide distribution of the corpus. The BNC World has also corrected errors relating to mislabeled texts and indeterminate part-of-speech codes in the first version, and has included a classification system of genre labels developed by Lee (2001) at Lancaster. The World Edition, originally marked up in TEI-compliant SGML, has now been replaced with the BNC XML Edition. With a few exceptions, the XML edition contains almost the same texts as in the previous World Edition, but this latest release has corrected many known errors and inconsistencies and included lemma information together with some other improvements. The BNC XML Edition is released on two DVD-ROMs, including the XML-aware corpus indexing and exploration tool Xaira (see appendix) as well as an indexed version of the BNC corpus ready for use. As a prelude to this full release of the XML version, a four-million-word subset of the BNC − BNC Baby − was released in October 2004 together with the XML-aware corpus tool Xaira (see appendix). BNC Baby was originally developed as a manageable subcorpus from the BNC for use in the language classroom, consisting of comparable samples for four kinds of English − unscripted conversation, newspapers, academic prose and written fiction (Burnard 2003).

The BNC model for achieving corpus balance and representativeness has been followed by a number of national corpus projects including, for example, the American National Corpus, the Polish National Corpus and the Russian National Corpus.

2.2. The American National Corpus

The American National Corpus (ANC) project was initiated in 1998 with the aim of building a corpus comparable to the BNC. While the ANC follows the general design of the BNC, there are differences with regard to its sampling period and text categories. The ANC only samples language data produced from 1990 onward whereas the sampling period for the BNC is 1960–1993. This time frame has enabled the ANC to cover text categories which have developed recently and thus were not included in the BNC, e. g. e-mails, web blogs, web pages, and chat room talks, as shown in Table 20.3. In addition to the BNC-like core, the ANC will also include specialized "satellite" corpora (cf. Reppen/Ide 2004, 106–107).

Tab. 20.3: Text categories in the ANC (Second release)

Channel	Text category	%
Written	Books (informative texts for various domains and imaginative texts of various types)	22
	Newspapers, magazines and journals	38
	Electronic (e-mails, web pages etc.)	18
	Miscellaneous (published and unpublished)	5
Spoken	Face-to-face/phone conversations, speech, meetings	17

The ANC corpus is encoded in XML, following the guidelines of the XML version of the Corpus Encoding Standard (XCES, see article 22). The standalone annotation, i. e. with the primary data and annotations kept in separate documents but linked with pointers, has enabled the corpus to be POS tagged using different tagsets (e. g. Biber's (1988) tags, the CLAWS C5/C7 tagsets (Garside/Leech/Sampson 1987) and the Penn tags (see Marcus/Santorini/Marcinkiewicz 1993) to suit the needs of different users.

The second release of the ANC, which has become available since October 2006, contains 22 million words of written and spoken data (18.5 million words for writing and 3.9 million words for speech, but not balanced for genre). In subsequent releases a target size of 100 million words is expected to be reached. The corpus is currently annotated for lemma, part-of-speech, noun chunks, and verb chunks. All data in the second release is POS tagged applying the Penn tagset while many documents are also tagged using Biber's tagset. The ANC corpus is distributed by the Linguistic Data Consortium (LDC, see appendix). In addition, the ANC website offers the Open ANC, which comprises approximately 14 million words (3.2 spoken and 11.4 written) from the second LDC release. The Open ANC is licensed free of charge and can be downloaded from the corpus website.

Another sizeable corpus worth mentioning is the BYU Corpus of American English, which currently contains 360 million words equally sampled from five sources (spoken, fiction, popular magazines, newspapers, and academic journals), with 20 million words for each of the 17 years during the period 1990–2007. The corpus is designed following a dynamic model, meaning that it will grow by 20 million words each year. The BYU corpus website (see appendix) provides a freely accessible interface with the same functionalities as the BNC VIEW interface, allowing users to search by word, phrase, substring, part-of-speech, collocates, etc.

2.3. Reference Corpora of Polish

This section introduces three large reference corpora of Polish, the PELCRA Reference Corpus of Polish, the PWN Corpus of Polish, and the IPI PAN Corpus of Polish. While three corpora are introduced in this section, there is presently no agreed upon national corpus of Polish, though the aim of the PELCRA project (Polish and English Language Corpora for Research and Application) was to build the Polish National Corpus – "with a well planned structure, balanced data etc., replicating the structure of the BNC" (Barbara Lewandowska Tomaszczyk, personal communication), and the PELCRA Reference Corpus was named as such from the outset (see Lewandowska-Tomaszczyk 2003, 106). The controversy over the title of the Polish national corpus has now turned into an interesting cooperative project, with the aim of creating a truly national corpus of Polish, by a consortium including the creators of the three Polish corpora mentioned above (Adam Przepiórkowski, personal communication).

As noted, the PELCRA Reference Corpus of Polish is created at the PELCRA project, which is undertaken jointly by the Universities of Łódź and Lancaster. The project aims to develop a large, fully annotated reference corpus of native Polish, "mirroring the BNC in terms of genres and its coverage of written and spoken language" (Lewandowska-Tomaszczyk 2003, 106). The corpus consists of 100 million words of running text, which covers genres and styles comparable in proportions to those included the BNC. An important difference lies in that the PELCRA corpus contains whole texts whereas the BNC is composed of text samples. The PELCRA corpus is TEI-compliant and is annotated for part-of-speech. Presently, it can be accessed at the corpus website via an online interface that can be used to search not only for simple words and phrases but for frequencies and occurrences of morphologically related words as well, using the so called "Inflections Concordancer". In addition, the interface allows for statistical analysis such as computation of word frequency and collocation.

The PWN corpus is a corpus of native Polish which is used by the Polish scientific publishers PWN primarily in dictionary making. The corpus consists of 60 million words of texts from the 20th and 21st centuries together with three million words from earlier periods. It keeps growing over time. An online version of the corpus, which is available for access for a fee, consists of 40 million words of samples taken from books, press publications, web pages and advertising leaflets and other ephemera, as well as transcripts of spoken data. A demonstration version of this online corpus (totaling 7.5 million words) is also accessible free of charge at the PWN site (see appendix).

The IPI PAN Corpus is presently the largest corpus of Polish, consisting of over 250 million segments (approximately 200 million orthographic words) in its second edition. Developed at the Institute of Computer Science (IPI) of the Polish Academy of Sciences (PAN), this giant corpus is morpho-syntactically annotated. While the corpus as a whole is not balanced, a balanced sample of 30 million segments is also available, which is composed of contemporary prose (10%), older prose (10%), non-fiction (10%), newspapers (50%), parliamentary proceedings (15%) and law (5%). The IPI PAN corpus, as well as its balanced component, is available to the public free of charge. Both can be downloaded together with the accompanying concordancer (i.e. Poliqarp) from the IPI PAN corpus website (see appendix), which also provides the online searching function.

2.4. The Czech National Corpus

The Czech National Corpus (CNC) consists of two sections: synchronous and diachronic. The synchronous section is designed to include written and spoken components. The texts in this section stem from a collection of electronic documents in legacy formats. There are plans for dialectal components of both the synchronous and the diachronic section, which currently are hardly more than blueprints for future work. We will thus only introduce the written and spoken components in the synchronous section.

The written component of the synchronous section, which contains 100 million words, was completed in 2000 and thus named SYN2000. SYN2000 includes both imaginative (15%) and informative (85%) texts, each being divided into a number of text categories, as shown in Table 20.4 (see Kučera 2002, 247−248). The technical and specialized texts in the corpus proportionally cover nine domains: lifestyle (5.55%), technology (4.61%), social sciences (3.67%), arts (3.48%), natural sciences (3.37%), economics/management (2.27%), law/security (0.82%), belief/religion (0.74%) and administrative texts (0.49%).

The spoken component of the synchronous section, the so-called Prague Spoken Corpus (PMK), contains 800,000 words of transcription of authentic spoken language sampled in a balanced way according to four sociolinguistic criteria: speaker sex, age, educational level and discourse type, as shown in Table 20.5. The data contained in the Prague

Tab. 20.4: Design of SYN2000

Major category	Genre	%
Imaginative (15%)	Fiction	11.02
	Poetry	0.81
	Drama	0.21
	Other literary texts	0.36
	Transitional text types	2.6
Informative (85%)	Journal	60
	Technical/specialized texts	25

Tab. 20.5: Sampling frame of the Prague Spoken Corpus

Criterion	Type	Proportion
Speaker sex	Male	50%
	Female	50%
Speaker age	21−35	50%
	35+	50%
Education level	Secondary school	50%
	University	50%
Discourse type	Formal	50%
	Informal	50%

Spoken Corpus consists exclusively of impromptu spoken language (roughly equivalent to the demographically sampled component in the BNC). Texts representing various blends of written and spoken language such as lectures, political speeches and play scripts are included in a special section in the written corpus (cf. Kučera 2002, 248, 253).

Both SYN2000 and the Prague Spoken Corpus are marked up in TEI-compliant SGML and tagged to show part-of-speech categories. SYN2000 is licensed free of charge for non-commercial use. A scaled-down version of SYN2000, PUBLIC, which contains 20 million words with the same genre distribution, is accessible online at the corpus website (see appendix).

2.5. The Hungarian National Corpus

The Hungarian National Corpus (HNC, see Váradi 2002) is a balanced reference corpus of present-day Hungarian. The corpus contains 187.6 million words of texts produced from the mid-1990s onwards, which are divided into five subcorpora, each representing a written text type: press, literature, science, official, and personal (electronic forum discussions). It is particularly interesting to note that in addition to stratification in genres, the HNC is also regionally stratified, covering language variants beyond the border of Hungary such as those from Slovakia, Subcarpathia, Transylvania and Vojvodina. Table 20.6 shows the sizes of these components.

Tab. 20.6: Components of the HNC corpus

	Hungary	Slovakia	Subcarpathia	Transylvania	Vojvodina	Total
Press	71.0	5.7	0.7	5.5	1.5	84.5
Literature	35.5	1.4	0.4	0.8	0.2	38.2
Science	20.5	2.3	0.7	1.6	0.3	25.5
Official	19.9	0.2	0.3	0.6	0.1	20.9
Personal	17.8	0.0	0.4	0.4	0.1	18.6
Total	164.7	9.5	2.5	8.9	2.0	187.6

The HNC is encoded in SGML in compliance with the Corpus Encoding Standard (CES) and annotated with part-of-speech and morphological information. The corpus can be accessed free of charge after registration via a sophisticated online query system at the corpus site (see appendix), which uses CQP functionality to enhance the search engine.

2.6. The Russian National Corpus

The Russian National Corpus (RNC or Natsionalny Korpus Russkogo Yazyka in Russian), which was known as the Russian Reference Corpus (BOKR), is designed as a Russian match for the BNC (see Sharoff 2006, where the corpus design is discussed under its pilot project name BOKR). The corpus contains over 147 million words, of which the BNC-comparable modern language subset amounts to 109 million words, with the rest coming from the 18–19th centuries and the first half of the 20th century. The

Tab. 20.7: Text categories covered in the BNC and RNC corpora

Text category	BNC	RNC
Spoken	10.7%	5%
Life (imaginative texts in the BNC)	16.7%	30%
Natural sciences	3.8%	5%
Applied sciences	7.2%	10%
Social sciences	14.2%	12%
Politics (world affairs in the BNC)	18.9%	15%
Commerce	7.6%	5%
Arts	6.8%	5%
Religion and philosophy (belief and thought in the BNC)	3.1%	3%
Leisure	11.2%	10%

modern language component includes 39.7% of imaginative writing, 56.4% of informative writing and 3.9% of spoken data (see RNC in appendix). Table 20.7 compares text categories covered in the BNC and RNC corpora (cf. Sharoff 2006, 172).

The RNC corpus is encoded in TEI-compliant SGML and annotated for part-of-speech. As Russian is a highly inflective language, the technique used in annotating English corpora with complex POS tags is impractical for Russian, because that would entail thousands of tags which would make corpus exploration ineffective, if not completely impossible. Hence in the Russian National Corpus, each word is annotated with a bundle of lexical and syntactic features such as part-of-speech, aspect, transitivity, voice, gender, number and tense. Separate features from a feature bundle associated with each word can be selected in a window in the query interface. The corpus was completed in 2005 (Serge Sharoff, personal communication) and a pilot version is now accessible online (see Ruscorpora in appendix).

2.7. The CORIS corpus

The CORIS (Corpus di Italiano Scritto) corpus is a general reference corpus of present-day Italian. It contains 100 million words of written Italian sampled from five text categories, which constitute five subcorpora, as shown in Table 20.8.

Unlike most national corpora that are sample corpora, the CORIS corpus follows a dynamic corpus model, which will be updated every two years by means of a built-in monitor corpus (Rossini Favretti/Tamburini/de Santis 2004). The current version of the corpus can be accessed online free of charge via a web-based query system at the corpus website (see appendix).

2.8. The Hellenic National Corpus

The Hellenic National Corpus is a 47-million-word corpus of written Modern Greek sampled from several publication media covering various genres (articles, essays, literary

Tab. 20.8: Components of the CORIS corpus

Category	Subcategory	Words (million)
Press	Newspapers, periodicals, supplement	38
Fiction	Novel, short stories	25
Academic prose	Human sciences, natural sciences, physics, experimental sciences	12
Legal and administrative prose	Legal, bureaucratic, administrative documents	10
Miscellaneous	Books on religion, travel, cookery, hobbies, etc.	10
Ephemeral	Letters, leaflets, instructions	5
	Total size	100

works, reports, biographies etc.) and domains (economy, medicine, leisure, art, human sciences etc.) published from 1976 onwards. Of the five types of medium, newspapers account for 62.83 % of the total texts, books 8.23 %, magazines 5.19 %, Internet texts 0.3 %, and miscellaneous (leaflets pamphlets, typed material, reports etc.) 23.49 %. In terms of genres, the HNC covers a wide variety including fiction, non-fiction, feature articles, informative writing, official documents, as well as advertising, private material and discussion.

The text classification with regard to medium, genre and domain follows the standards established on the PAROLE project (see section 11.11.). This taxonomy information, together with the bibliographic information, is encoded in TEI-compliant SGML (cf Hatzigeorgiu et al. 2000, 1737). The corpus is constantly being updated and can be accessed online at the corpus site (see appendix), where users can make queries concerning the lexicon, morphology, syntax and usage of Modern Greek (e. g. words, lemmas, part-of-speech categories or combinations of the three).

2.9. The DWDS corpus

The DWDS corpus is a product of the DWDS (Digital Dictionary of the 20th Century German Language) project. The corpus is divided into two parts, a 100-million-word balanced core and a much larger opportunistic subcorpus with a target size of 500 million words. This section introduces the core corpus, which is roughly comparable to the British National Corpus, covering the whole 20th century (1900−2000). Table 20.9 shows the text categories covered in the corpus.

The metadata such as genre information is encoded in XML. Linguistic annotation consists basically of lemmatization, part-of-speech and semantic annotation on the word level, as well as prepositional phrase and noun phrase recognition on the phrase level (Cavar/Geyken/Neumann 2000). The core corpus is available for online search after free-of-charge registration at the DWDS website (see appendix), which provides quite sophisticated tools for searching and presenting search results. These tools are primarily focused on lexicological research.

Tab. 20.9: Design criteria of the DWDS Corpus

Text category	Proportion
Literature	26%
Journalistic prose	27%
Scientific texts	22%
Specialized texts (advert, manuals, etc.)	20%
Spoken (everyday language, televised debates, dialect, etc.)	5%

2.10. The Slovak National Corpus

The Slovak National Corpus (SNK) is a large database of contemporary Slovak texts covering a broad range of genres and language styles. The latest release of the corpus, version 3.0, contains 339 million tokens taken from journalistic text (60.6%), fiction (17.5%), specialized texts (11.6%) and other sources (10.3%) published since 1955. In-addition to the SNK as a whole, there is a cleaned up version of the corpus of approximately 319 million tokens which contains only non-linguistic texts of standard quality (correct diacritics, standard contemporary language from the territory of Slovakia).

The texts in the Slovak National Corpus are morphologically annotated and lemmatized automatically. There is also a small subcorpus (approximately half a million tokens) with hand-crafted morphological annotation and lemmatization, which can be used to train morphological taggers (cf. Gianitsová 2005; Šimková 2005; Garabík 2006). In addition to such linguistic analyses, the texts are encoded with source (bibliographical and style-genre) information.

The SNK corpus is available to the public for research, educational, and other strictly non-commercial purposes. Users can access the corpus using a simple online query system at the SNK website (see appendix). More complex searches require the corpus manager client program (Bonito), which supports regular expressions and can be downloaded at the same website. Online registration is also required in order to login using the client program.

We have so far introduced national corpora for European languages. The next two sections will introduce two national corpora of Asian languages, namely Chinese and Korean; Japanese corpora are discussed in article 21.

2.11. The Modern Chinese Language Corpus

The Modern Chinese Language Corpus (MCLC) is China's national corpus built under the auspices of the National Language Committee of China. The corpus contains 100 million Chinese characters of systematically sampled texts produced during 1919–2002, with the majority of texts produced after 1977. 1919 is generally considered as the beginning of modern Chinese. The corpus covers three large categories (humanities/social sciences, natural sciences, and miscellaneous text categories such as official documents, ceremony speech and ephemera) including more than 40 subcategories. Text categories containing over five million characters include literature, society, economics, newspapers,

Tab. 20.10: Components of the MCLC corpus

Domain	Category
Humanities and social sciences (8 categories)	Politics and laws, history, society, economics, arts, literature, military and physical education, life
Natural sciences (6 categories)	Mathematics and physics, biology and chemistry, astronomy and geography, oceanology and meteorology, agriculture and forestry, medical and health
Miscellaneous (6 categories)	Official documents, regulations, judicial documents, business documents, ceremonial speech, ephemera

miscellaneous, and legal texts, with literary texts accounting for the largest proportion (nearly 30 million characters). Table 20.10 shows the components of the corpus. Most samples in the corpus are approximately 2,000 characters in length, with the exception of samples taken from books, which may contain up to 10,000 characters. The digitalized texts were proofread three times so that errors are less than 0.02 % (see Wang 2001, 283).

All text samples in the MCLC corpus are encoded with detailed bibliographic information (up to 24 items) in the corpus header. A core component of the corpus, which is composed of 50 million Chinese characters, has been tokenized (with an error rate of 0.5 ‰) and POS-tagged (with an error rate of 0.5 %), while a small part of it (one million characters, in 50,000 sentences) has been built into a treebank.

Presently, a scaled down version of the corpus, which contains 20 million characters proportionally sampled from the larger corpus, has been made available to the public free of charge for online access at the MCLC website (see appendix).

2.12. The Korean National Corpus

The Korean National Corpus (KNC) is under construction on the 21st Century Sejong project, which was launched in 1998 as a ten-year development project to build various kinds of language resources including Korean corpora and Korean electronic dictionaries. One of the goals of the project is to construct a balanced national corpus comparable to the BNC (Kang/Kim 2004). The initial target size was 500 million eojuls (similar to tokens but different from English words – an eojul is a morpho-syntactic combination of a word plus particle(s), or a word plus ending(s), or simply a word alone). However since the annotated part of the corpus has been increased, the current goal of the corpus size is 200 million eojuls.

The KNC consists of two components, namely, the primary corpus and the specialized corpus. The primary corpus division deals with the Korean language as used in South Korea, with some parts annotated with various types of linguistic information. The current version of the primary corpus consists of four components – raw corpus (63,899,412 eojuls), grammatically tagged corpus (15,226,186 eojuls), parsed corpus (570,064 eojuls), and semantically tagged corpus (10,132,348 eojuls) – totaling 89,830,015 eojuls. The raw corpus contains data from a range of genres including news-

papers (20%), magazines (10%), academic writing (35%), literary works (20%), quasi-spoken data (10%) and others (5%). The special corpus division is largely concerned with language variation and parallel corpus construction. Presently the following specialized corpora are available: a corpus of transcribed colloquial discourse (3,390,533 eojuls), a historical corpus (5,291,215 eojuls), a corpus of international Korean (9,096,159 eojuls), and a Korean-English parallel corpus (5,616,313 eojuls).

Texts in the Korean National Corpus are encoded in SGML, applying the TEI (Text Encoding Initiative) P3 standard. The simplified TEI header shows details such as bibliography, text category, history of computerization, and record of correction. As the Text Encoding Initiative has been updated to TEI P5 based on XML, there is a plan to migrate the KNC corpus to the new standard (see Kim 2006).

The corpus is accessible over the Internet after registration at the corpus site (see KNC in appendix).

2.13. Other National corpora

In addition to those introduced above, there are a number of nation-level corpora which are either already available or are under construction. They include, for example, the FRANTEXT Database for French (see appendix), the Croatian National Corpus (101.3 million tokens, see appendix), Korpus 2000 for Danish (28 million words, see appendix), the National Corpus of Irish (30 million words, see appendix). A number of corpora representing other national languages are also under construction, including, for example, Norwegian (Choukri 2003), Dutch (Wittenburg/Brugman/Broeder 2000), Maltese (Dalli 2001), Basque (Aduriz et al. 2003), Kurdish (Gautier 1998), Nepali (Glover 1998), Tamil (Malten 1998) and Indonesian (Riza 1999).

3. Monitor corpora

While most of the national corpora introduced in section 2 follow a static sample corpus model, there are also corpora which are constantly updated to track rapid language change, such as the development and the life cycle of neologisms. Corpora of this type are referred to as monitor corpora.

The best-known monitor corpus is the Bank of English (BoE), which was initiated in 1991 on the COBUILD (Collins Birmingham University International Language Database) project. The corpus was designed to represent standard English as it was relevant to the needs of learners, teachers and other users, while also being of use to researchers in present-day English language. Written texts (75%) come from newspapers, magazines, fiction and non-fiction books, brochures, reports, and websites while spoken data (25%) consists of transcripts of television and radio broadcasts, meetings, interviews, discussions, and conversations. The majority of the material in the corpus represents British English (70%) while American English and other varieties account for 20% and 10% respectively. Presently the BoE contains 524 million words of written and spoken English. The corpus keeps growing with the constant addition of new material (cf the BoE website at Collins, see appendix).

The BoE corpus is particularly useful for lexical and lexicographic studies, for example, tracking new words, new uses or meanings of old words, and words falling out of use. A 56-million-word sampler of the corpus can be accessed online free of charge at the corpus website. Access to larger corpora is granted by special arrangement.

Another corpus of the monitor type is the Global English Monitor Corpus, which was started in late 2001 as an electronic archive of the world's leading newspapers in English. The corpus aims at monitoring language use and semantic change in English as reflected in newspapers so as to allow for research into whether the English language discourses in Britain, the United States, Australia, Pakistan and South Africa have changed in the same way or differently. As the Global English Monitor Corpus will monitor as accurately as conceivable all relevant changes of attitudes and beliefs, it will prove a useful tool not only for lexicographers, historical linguists and semanticists, but also for those interested in social and political studies all over the world. With its first results having become available at the end of 2003, the corpus is expected to reach billions of words within a few years (cf. the corpus website, see appendix).

4. Corpora of the Brown family

The first modern corpus of English, the Brown University Standard Corpus of Present-day American English (i. e. the Brown corpus, see Kučera/Francis 1967), was built in the early 1960s for written American English. The population from which samples for this pioneering corpus were drawn was written English text published in the United States in 1961, while its sampling frame was a list of the collection of books and periodicals in the Brown University Library and the Providence Athenaeum. The target population was first grouped into 15 text categories, from which 500 samples of approximately 2,000 words were then drawn proportionally from each text category, totaling roughly one million words.

The Brown corpus was constructed with comparative studies in mind, in the hope of setting the standard for the preparation and presentation of further bodies of data in English or in other languages. This expectation has now been realized. Since its completion, the Brown corpus model has been followed in the construction of a number of corpora for synchronic and diachronic studies as well as for cross-linguistic contrast. Table 20.11 shows a brief comparison of these corpora.

As can be seen, these corpora are roughly comparable but have sampled different languages or language varieties. Their sampling periods are either similar for the purposes of synchronic comparison or distanced by about three decades for the purposes of diachronic comparison. For example, the Brown and LOB (the Lancaster/Oslo-Bergen corpus of British English, see Johansson/Leech/Goodluck 1978) can be used to compare American and British English as used in the early 1960s. The updated versions of the two corpora, Frown (see Hundt/Sand/Skandera 1999) and FLOB (see Hundt/Sand/Siemund 1998) can be used to compare the two major varieties of English as used in the early 1990s. Other corpora of a similar sampling period, such as ACE (the Australian Corpus of English, also known as the Macquarie corpus), WWC (the Wellington Corpus of Written New Zealand English) and Kolhapur (the Kolhapur Corpus of Indian English), together with FLOB and Frown, allow for comparison of "world Englishes". For

Tab. 20.11: Corpora of the Brown family

Corpus	Language variety	Period	Samples	Words (million)
Brown	American English	1961	500	One
Frown	American English	1991−1992	500	One
LOB	British English	1961	500	One
Lancaster 1931	British English	1931 +/− 3 years	500	One
FLOB	British English	1991−1992	500	One
Kolhapur	Indian English	1978	500	One
ACE	Australian English	1986	500	One
WWC	New Zealand English	1986−1990	500	One
LCMC	Mandarin Chinese	1991 +/− 3 years	500	One

diachronic studies, the Brown vs. Frown on the one hand, and the Lancaster 1931 (see Leech/Smith 2005), LOB and FLOB corpora on the other hand, provide a reliable basis for tracking recent language change over 30-year periods. The LCMC corpus (the Lancaster Corpus of Mandarin Chinese, see McEnery/Xiao/Mo 2003), when used in combination with the FLOB/Frown corpora, provides a valuable resource for contrastive studies between Chinese and two major varieties of English (see LCMC in appendix).

In comparing these corpora synchronically, caution must be exercised to ensure that the sampling periods are similar. For example, comparing the Brown corpus with FLOB would involve not only language varieties but also language change. Also, as the Brown model may have been modified slightly in some of these corpora, account must be taken of such variation in comparing these corpora across text categories by normalizing the raw frequencies to a common basis. Table 20.12 compares the text categories and number of samples for each category in these corpora.

It can be seen from the table that the two American English corpora (Brown and Frown) have the same numbers of samples for each of the 15 text categories while the British English corpora share the same proportions. The two groups differ in the numbers of samples for categories E, F, and G. The WWC and LCMC corpora follow the model of FLOB. There are important differences between the Kolhapur corpus and others in both sampling periods and the proportions of text categories. The ACE corpus covers 17 text categories instead of 15. All of these differences should be taken into account when comparing these corpora.

With the exceptions of the Lancaster 1931 corpus, which has not been released to the public presently, and LCMC, which is distributed by the European Language Resources Association (ELRA, see appendix), all of the corpora of the Brown family are available from the International Computer Archive of Modern and Medieval English (ICAME, see appendix).

The corpora of the Brown family are balanced corpora representing a static snapshot of a language or language variety in a certain period. While they can be used for synchronic and diachronic studies, more appropriate resources for these kinds of research are synchronic and diachronic corpora, which will be introduced in the following two sections.

Tab. 20.12: Text categories m the corpora of the Brown family

Code	Text category	Brown	Frown	LOB	FLOB	Lancaster 1931	Kolhapur	ACE	WWC	LCMC
A	Press reportage	44	44	44	44	44	44	44	44	44
B	Press editorials	27	27	27	27	27	27	27	27	27
C	Press reviews	17	17	17	17	17	17	17	17	17
D	Religion	17	17	17	17	17	17	17	17	17
E	Skills, trades and hobbies	36	36	38	38	38	38	38	38	38
F	Popular lore	48	48	44	44	44	44	44	44	44
G	Biographies and essays	75	75	77	77	77	70	77	77	77
H	Miscellaneous (reports, official documents)	30	30	30	30	30	37	30	30	30
J	Science (academic prose)	80	80	80	80	80	80	80	80	80
K	General fiction	29	29	29	29	29	59	29	29	29
L	Mystery and detective fiction	24	24	24	24	24	24	15	24	24
M	Science fiction	6	6	6	6	6	2	7	6	6
N	Western and adventure fiction	29	29	29	29	29	15	8	29	29
P	Romantic fiction	29	29	29	29	29	18	15	29	29
R	Humour	9	9	9	9	9	9	15	9	9
S	Historical fiction	–	–	–	–	–	–	22	–	–
W	Women's fiction	–	–	–	–	–	–	15	–	–

5. Synchronic corpora

While the corpora of the Brown family are generally good for comparing language varieties such as world Englishes, the results from such a comparison must be interpreted with caution if the corpora under examination were built for different periods or the Brown model has been modified. A more reliable basis for comparing language varieties is a synchronic corpus.

5.1. The International Corpus of English

A typical corpus of this type is the International Corpus of English (ICE), which is specifically designed for the synchronic study of world Englishes. The ICE corpus consists of a collection of twenty corpora of one million words each, each composed of written and spoken English produced during 1990−1994 in countries or regions in which English is a first or official language (e. g. Australia, Canada, East Africa, Hong Kong as well as Great Britain and the USA). As the primary aim of ICE is to facilitate comparative studies of English worldwide, each component follows a common corpus design as well as a common scheme for grammatical annotation to ensure direct comparability among the component corpora. All ICE corpora contain 500 texts of approximately 2,000 words each, sampled from a wide range of spoken (60%) and written (40%) genres, as shown in Table 20.13 (see Nelson 1996, 29−30).

The ICE corpora are marked up and annotated at various levels. In written texts, features of the original layout are marked, including sentence and paragraph boundaries, headings, deletions, and typographic features, while spoken texts are transcribed orthographically, and are marked for pauses, overlapping strings, discourse phenomena such as false starts and hesitations, and speaker turns. The bibliographic markup, which gives a complete description (e. g. text category, date, and publisher) of each text, is stored in the corpus header of each file. While uniform criteria for data collection and markup style have been applied for all ICE corpora, different levels of linguistic annotation have been undertaken for different components. While the British component (ICE-GB) is POS tagged and parsed (see section 9.6. for further discussion), others are currently available as unannotated lexical corpora, e. g. the components for East Africa, Hong Kong, India, the Philippines, New Zealand, and Singapore. With the exceptions of ICE-GB and ICE-New Zealand, which can be ordered on CD-ROMs, all other currently available ICE corpora are licensed free of charge and can be downloaded at the ICE website (see appendix), but users must sign a license agreement in order to receive the passwords to decompress the corpus files.

5.2. The Longman/Lancaster Corpus

The Longman/Lancaster Corpus consists of about 30 million words of published English. British data takes up 50% and American data 40% while the other 10% represents other varieties such as Australian, African and Irish English. One half of the samples were selected randomly ("microcosmic texts") and the other half selected by a panel of experts

Tab. 20.13: Corpus design of ICE

Spoken (300)	Dialogues (180)	Private (100)	Conversations (90) Phone calls (10)
		Public (80)	Class lessons (20) Broadcast discussions (20) Broadcast interviews (10) Parliamentary debates (10) Cross-examinations (10) Business transactions (10)
	Monologues (120)	Unscripted (70)	Commentaries (20) Unscripted speeches (30) Demonstrations (10) Legal presentations (10)
		Scripted (50)	Broadcast news (20) Broadcast talks (20) Non-broadcast talks (10)
Written (200)	Non-printed (50)	Student writing (20)	Student essays (10) Exam scripts (10)
		Letters (30)	Social letters (15) Business letters (15)
	Printed (150)	Academic (40)	Humanities (10) Social sciences (10) Natural sciences (10) Technology (10)
		Popular (40)	Humanities (10) Social sciences (10) Natural sciences (10) Technology (10)
		Reportage (20)	Press reports (20)
		Instructional (20)	Administrative writing (10) Skills/hobbies (10)
		Persuasive (10)	Editorials (10)
		Creative (20)	Novels (20)

("selective texts"). Most texts in the corpus are about 40,000 words long but no whole texts are used.

Both imaginative and informative text categories are included. Imaginative texts come from well-known literary works and works randomly sampled from books in print; informative texts come from natural and social sciences, world affairs, commerce and finance, the arts, leisure, and so on. Imaginative texts are mainly works of fiction in book form while informative texts comprise books, newspapers and journals, unpublished and ephemera. Four external criteria have been used in text selection (see Holmes-Higgin/Abidi/Ahmad 1994): "region" (language varieties), "time" (1900s−1980s), "medium" (books 80%, periodicals 13.3% and ephemera 6.7%), and "level" (literary, middle and popular for imaginative texts, and technical, lay and popular for informative texts). As

part of the Longman Corpus Network (see appendix), the Longman/Lancaster Corpus is not available for public access.

5.3. The Longman Written American Corpus

The Longman Written American Corpus contains over 100 million words of running texts taken from newspapers, journals, magazines, best-selling novels, technical and scientific writing, and coffee-table books. The design of the Longman Written American Corpus is based on the general design principles of the Longman/Lancaster Corpus and the written section of the BNC. The corpus is dynamically refined and keeps growing with the constant addition of new materials. Like the other components of the Longman Corpus Network (cf. the corpus website), this corpus does not appear to allow public access.

5.4. The CREA corpus of Spanish

The CREA (Corpus de Referencia del Español Actual) is a corpus of standard varieties of Spanish. The corpus currently contains 150 million words sampled from a wide range of written and spoken text categories produced in all Spanish speaking countries (European Spanish 83 million words and American Spanish 67 million words). The domains covered in the corpus include science and technology, social sciences, religion and thought, politics and economics, arts, leisure and ordinary life, health, and fiction.

The CREA was designed as a monitor corpus which is continually updated so that it always represents the last twenty-five years of the history of Spanish. New data is added proportionally to maintain the corpus balance and to ensure that the various trends in current Spanish are represented.

The CREA corpus is marked up in SGML. Bibliographic and taxonomic information is encoded in the corpus header of each file. For written texts, both structural (paragraph and page number) and intratextual (notes, formulas, tables, quotations, foreign words etc.) marks are encoded. For spoken texts, the markup scheme indicates structural (speech turns) and non-structural (overlapping, tottering, anacoluthon, etc.) marks (cf. Guerra 1998).

The modular structure of the CREA corpus allows for flexible searches using geographical, generic, temporal, and thematic criteria. The corpus is accessible on the Internet (see appendix).

5.5. The LIVAC corpus of Chinese

The LIVAC (Linguistic Variation in Chinese Speech Communities) project started in 1993 with the aim of building a synchronous corpus for studying varieties of Mandarin Chinese. For this purpose, data has been collected regularly and simultaneously, once every four days since July 1995, from representative Mandarin Chinese newspapers and

the electronic media of six Chinese speaking communities: Hong Kong, Taiwan, Beijing, Shanghai, Macau and Singapore. The contents of these texts typically include the editorial, and all the articles on the front page, international and local news pages, as well as features and reviews. The corpus is planned to cover a 10-year period between July 1995 and June 2005, capturing salient pre- and post-millennium evolving cultural and social fabrics of the diverse Chinese speech communities (Tsou et al. 2000). The collection of materials from these diverse communities is synchronized with uniform calendar reference points so that all of the components are comparable. The LIVAC corpus contains over 150 million Chinese characters, with 720,000 word types in its lexicon.

All of the corpus texts in LIVAC are segmented automatically and checked by hand. In addition to the corpus, a lexical database is derived from the segmented texts, which includes, apart from ordinary words, those expressing new concepts or undergoing sense shifts, as well as region specific words from the six communities. The database is thus a rich resource for research into linguistics, sociolinguistics, and Chinese language and society.

As LIVAC captures the social, cultural, and linguistic developments of the six Chinese speaking communities within a decade, it allows for a wide range of comparative studies on linguistic variation in Mandarin Chinese. The corpus also provides an important resource for tracking lexical development such as the evolution of new concepts and their expressions in present-day Chinese. While the access to the entire corpus is restricted to registered users only, a sample (covering the period from 1 July 1995 to 30 June 1997) can be searched using the online query system at the LIVAC site (see appendix), which shows KWIC concordances as well as frequency distributions across the six speech communities.

6. Diachronic corpora

Another way to explore language variation is from a diachronic perspective using diachronic corpora. A diachronic (or historical) corpus contains texts from the same language gathered from different time periods. Typically that period is far more extensive than that covered by Brown/Frown and LOB/FLOB or a monitor corpus such as the Bank of English. Diachronic corpora are used to track changes in language evolution. This section introduces a number of corpora of this kind.

6.1. The Helsinki Corpus of English Texts

Perhaps the best-known historical corpus is the diachronic part of the Helsinki Corpus of English Texts (i. e. the Helsinki Corpus), which consists of approximately 1.5 million words of English in the form of 400 text samples, dating from the 8th to the 18th centuries. The corpus is divided into three periods (Old, Middle, and Early Modern English) and eleven subperiods, as shown in Table 20.14 (cf. Kytö 1996).

In addition to the basic selection of texts as indicated in the table, there is a supplementary part in the Corpus, which focuses on regional varieties. This part consists of 834,000 words of Older Scots (in international distribution) and 300,000 words of early

Tab. 20.14: Periods covered in the Helsinki Diachronic Corpus

Period	Subperiod	Words	Percent	Overall
Old English	I. −850	2,190	0.5	413,250
	II. 850−950	92,050	22.3	
	III. 950−1050	251,630	60.9	
	IV. 1050−1150	67,380	16.3	
	Total	413,250	100	26.27%
Middle English	I. 1150−1250	113,010	18.6	608,570
	II. 1250−1350	97,480	16.0	
	III. 1350−1420	184,230	30.3	
	IV. 1420−1500	213,850	35.1	
	Total	608,570	100	38.70%
Early Modern English	I. 1500−1570	190,160	34.5	
	II. 1570−1640	189,800	34.5	
	III. 1640−1710	171,040	31.0	
	Total	551,000	100	35.03%
Total		1,572,820		100%

American English (in compilation). While the primary selectional criteria are the dates of texts, the Helsinki Corpus has sought to reflect socio-historical variation (e. g. author sex, age and social rank) and a wide range of text types (e. g. law, handbooks, science, trials, sermons, diaries, documents, plays, private and official correspondence, etc.) for each specific period. The textual markup scheme includes more than thirty genre labels, which indicate, whenever available, parameter values for the dialect and the level of formality of the text, the relationship between the writer and the receiver as well as the author's age, sex, and social rank (Rissanen 2000).

As the Helsinki Corpus has not only sampled different periods covering one millennium, but also encoded genre and sociolinguistic information, this corpus allows researchers to go beyond simply dating and reporting language change, by combining diachronic, sociolinguistic and genre studies. The Helsinki Corpus can be ordered from ICAME (see appendix) or the Oxford Text Archive (OTA, see appendix).

6.2. The ARCHER corpus

ARCHER, an acronym for "A Representative Corpus of Historical English Registers", contains 1.7 million words of data in the form of 1,037 texts sampled from seven 50-year historical periods covering both Early and Late Modern English (1650−1990). The corpus is designed as a balanced representation of seven written genres (journal-diaries, letters, fiction, news, science, etc.) and three speech-based ones (fictional conversation, drama and sermons-homilies) in British (two thirds of the corpus) and American (one

third, data available only for the periods 1750–1799, 1850–1899, 1950–1990) English. Each 50-year subcorpus includes 20,000–30,000 words per register, typically containing ten texts of approximately 2,000–3,000 words each (cf. Biber/Finegan/Atkinson 1994). ARCHER is tagged for grammatical/functional categories. It allows for a wide variety of investigations on recent linguistic change and change in discourse and genre conventions. The corpus is presently being expanded with more American texts to make the American and British data comparable. The expanded version will also enable a systematic comparison of the two varieties of English diachronically. However, because of copyright problems, ARCHER is not publicly available at the moment.

In addition to the Helsinki and ARCHER corpora, which cover many centuries, there are a number of well-known historical corpora focusing on a particular period or a specific domain or genre, which will be introduced in the following sections.

6.3. The Lampeter Corpus of Early Modern English Tracts

The Lampeter Corpus of Early Modern English Tracts is a balanced corpus covering one century between 1640 and 1740, which is divided into ten decades. Each decade consists of data sampled from six domains (religion, politics, economics/trade, science, law and miscellaneous). Two complete texts, ranging from 3,000 to 20,000 words, are included for each domain within each decade, totaling approximately 1.1 million words (Schmied 1994).

The Lampeter corpus is encoded in TEI-compliant SGML. The TEI headers provide the framework for historical, sociolinguistic and stylistic investigations, including information regarding authors (name, age, sex, place of residence, education, social status, political affiliation), printers/publishers, place and date of print, publication format, text characteristics and bibliographical sources. As the corpus includes whole texts rather than smaller samples, the corpus is also useful for study of textual organization in Early Modern English. The Lampeter corpus can be ordered from ICAME or OTA (see appendix).

6.4. The Dictionary of Old English Corpus in Electronic Form

The Dictionary of Old English Corpus in Electronic Form (DOEC, the 2004 release) contains 3,047 texts of Old English, totaling four million words, in addition to two million words of Latin. The texts in the corpus are practically all extant Old English writings. The DOEC corpus includes at least one copy of each surviving text in Old English while in cases where it is significant because of dialect or date, more than one copy is included. These texts cover six text categories: poetry, prose, interlinear glosses, glossaries, runic inscriptions, and inscriptions in the Latin alphabet. In the prose category in particular, a wide range of text types are covered which include, for example, saints' lives, sermons, biblical translations, penitential writings, laws, charters and wills, records (of manumissions, land grants, land sales, land surveys), chronicles, a set of tables for computing the moveable feasts of the Church calendar and for astrological calculations, medical texts, prognostics (the Anglo-Saxon equivalent of the horoscope),

charms (such as those for a toothache or for an easy labour), and even cryptograms (cf. the corpus website). The texts in the corpus are encoded in HTML, TEI-compliant SGML, and XML. The DOEC corpus can be ordered on CDs or assessed online by institutional site license at the corpus website (see appendix). The web-based query system allows for searches by single words, word combinations, word proximity and bibliographic sources.

6.5. The EEBO and ECCO databases

Early English Books Online (EEBO) is a joint effort launched in 1999 between the University of Michigan, Oxford University and ProQuest Information and Learning to create a full-text archive of Early English. From the first book published in English through the age of Spenser and Shakespeare, the EEBO collection now contains about 100,000 of over 125,000 titles listed in Pollard & Redgrave's *Short-Title Catalogue (1475–1640)* and Wing's *Short-Title Catalogue (1641–1700)* and their revised editions, as well as the *Thomason Tracts (1640–1661)* collection and the *Early English Books Tract Supplement*, covering a wide range of domains including, for example, English literature, history, philosophy, linguistics, theology, music, fine arts, education, mathematics and science (cf. the corpus website). The remaining titles will be digitized and added to the database in the near future. The database can be accessed online at the EEBO website (see appendix).

Note that Early English Books Online is more of an archive than a corpus. Another similar database is the Eighteenth Century Collections Online (ECCO), which claims to be the most ambitious single digitization project ever undertaken. It includes all significant English-language and foreign-language titles printed in Great Britain during the eighteenth century as well as thousands of important works from the Americas, covering a great variety of materials ranging from books and directories, Bibles, sheet music and sermons to advertisements, and amounting to more than 26 million pages in 150,000 printed volumes. This database is available for a free trial at the ECCO website (see appendix).

6.6. The Corpus of Early English Correspondence

The Corpus of Early English Correspondence (CEEC) is nowadays a cover term for a family of corpora. The full version completed in 1998 consists of 96 collections of 6,039 personal letters written by 778 people (women accounting for 20%) between 1417 and 1681, totaling 2.7 million words. The corpus is accompanied by a sender database, which offers users easy access to various sociolinguistic variables, including writer age, gender, place of birth, education, occupation, social rank, domicile and the relationship with the addressee. CEEC is a balanced corpus which can be neatly divided into two parts, both covering chronologically fairly equal periods: the first from ca. 1417 to 1550 and the second from 1551 to 1680 (cf. Laitinen 2002). Table 20.15 shows the proportions in terms of writers' social ranks and domiciles (see Nevalainen 2000, 40).

As the copyright problem has prevented public access to the full release of the CEEC corpus, a CEEC sampler (CEECS) was published by ICAME, which represents the non-

Tab. 20.15: The CEEC corpus by rank and domicile

Rank	%	Domicile	%
Royalty	2.4	Court	7.8
Nobility	14.7	London	13.9
Gentry	39.3	East Anglia	17.1
Clergy	13.6	North	12.5
Professionals	11.2	Other regions	48.6
Merchants	8.4		
Other nongentry	9.4		

copyrighted materials included in CEEC. The sampler reflects the structure of the full CEEC only in some respects. The time covered is nearly the same (1418–1680), which is divided into two parts. CEECS1 (246,055 words) covers the 15th and 16th centuries while CEECS2 (204,030 words) covers the 17th century. The sampler corpus consists of 23 collections of 1,147 letters with 194 informants, totaling 450,085 words. The CEEC sampler is available from ICAME or the Oxford Text Archive.

A more recent release of the annotated CEEC corpus (PCEEC) is now available from the Oxford Text Archive, which includes the bulk of the original corpus, with 2.2 million words in total (4,979 letters from 657 writers). This new release is part-of-speech tagged and fully parsed.

CEEC is now supplemented by an extension (CEECE, 1681–1800, ca. 2.2 million words) and a supplement (CEECSU 1402–1663, 0.44 million words). These two corpora are currently not available for public use.

6.7. The Zurich English Newspaper Corpus

The Zurich English Newspaper Corpus (ZEN) is a 1.2-million-word collection of newspapers in Early English, covering 120 years (from 1671 to 1791) of British newspaper history. To achieve a representative coverage, a wide variety of newspapers were included. Up to ten issues per newspaper were selected at ten-year intervals throughout the whole period. With the exception of stock market reports, lottery figures, long lists of names and poetry, the whole newspapers were included in the corpus. The news stories are grouped into two major categories: foreign news and home news, with each news category further classified according to its own text genre definition (cf. Fries/Schneider 2000). The corpus is split into four periods in order to track potential language change, as shown in Table 20.16 (see Schneider 2002, 202).

The ZEN corpus is SGML-conformant. It not only allows for linguistic analysis of different types of news stories in the 17th and 18th centuries, it has also made it possible to compare news texts in Early English with modern newspaper language. The ZEN query system (see appendix) allows restricted access to the online database.

Tab. 20.16: The ZEN corpus

Section	Period	Words	Sentences
A	1670–1709	242758	7642
B	1710–1739	347825	12163
C	1740–1769	339362	14112
D	1770–1799	298249	11843
Total		1228194	45760

6.8. The Innsbruck Computer Archive of Machine-Readable English Texts

The Innsbruck Computer Archive of Machine-Readable English Texts (ICAMET) contains ca. 500 Middle English texts totaling 5.7 million words. The database comprises three parts, namely, the Prose Corpus (129 texts written during 1100–1500, accounting for two thirds of the total), the Letter Corpus (254 letters written during 1386–1688, arranged in diachronic order), and the Prose Varia Corpus (mainly translations or normalized versions of Middle English texts). An advantage of ICAMET is that the database consists of complete texts instead of extracts, which allows literary, historical and topical analyses of various kinds, particularly studies of cultural history (Marcus 1999). Nevertheless, the copyright issue has restricted public access to many prose texts in the corpus. A sampler containing half of the prose texts and all letters is available from ICAME.

6.9. The Corpus of English Dialogues

The Corpus of English Dialogues (CED) contains 1.2 million words of Early Modern English dialogue texts produced over a 200-year period between 1560 and 1760. This sampling period is divided into five time spans of 40 years, with each including approximately 200,000 words. While the spoken language of the past is inaccessible directly to modern speakers, it is recorded in speech related texts. The CED corpus has sampled from six such text categories, including trial proceedings, witness depositions, drama, fictional dialogues, didactic works in dialogue form comprising the sub-categories of language teaching texts and other didactic works, and a residual category of miscellaneous texts (cf. Culpeper/Kytö 1997, 2000, and forthcoming).

The focus on dialogue will allow insight into the nature of impromptu speech and interactive two-way communication in the Early Modern English period – aspects which have received little research attention. The CED corpus was released by the Universities of Uppsala and Lancaster in spring 2006. A user's guide accompanying the corpus (Kytö/Walker 2006) is also available.

6.10. A Corpus of Late Eighteenth-Century Prose

A Corpus of Late Eighteenth-Century Prose contains 30,000 words of unpublished letters transcribed from the originals dated from the period 1761–1790. The corpus is

distributed in both plain text (extended ASCII) and HTML versions. The text version can be used with a concordancer while the HTML version facilitates viewing the corpus in a browser. The plain text version is marked up in the COCOA format, giving information on writer, date and page breaks, etc. The corpus is intended to complement major diachronic corpora like the Helsinki Corpus, which stop in the early eighteenth century. Another aim of the corpus is "to illustrate non-literary English and English relatively uninfluenced by prescriptivist ideas, in the belief that it might help with research into change in (ordinary, spoken) language in the late Modern English period" (van Bergen/ Denison 2004, 228). The corpus is by no means uniform, nor is it balanced. Nevertheless, because of the nature of the material, it is of great use to both linguists and historians. The corpus can be ordered from the Oxford Text Archive, free of charge, for use in education and research.

6.11. A Corpus of Late Modern English Prose

A Corpus of Late Modern English Prose contains 10,000 words of informal private letters written by British writers between 1861 and 1919. All decades in this period are represented, with about 6,000 words for the decade 1880−1889, 13,000 words for 1890−1899 and 20,000 words for the other four decades each. These blocks of texts are sampled from five sources.

Stored in seven extended (8-bit) ASCII text files, the corpus is marked up following the conventions used in the Helsinki Corpus, with information on writer, recipient, relationship, date, genre, and page etc. encoded in COCOA-style brackets (see Denison 1994). The corpus can be ordered at no cost from the Oxford Text Archive.

In addition to the diachronic corpora introduced in the previous sections, there are a number of online databases which are accessible on the Internet, for example, Michigan Early Modern English Materials (MEMEM, see appendix), the Corpus of Middle English Prose and Verse (CME, see appendix), the Middle English Collection (MidEng, see appendix), and the Korpus of Early Modern Playtexts in English (KEMPE, see appendix).

6.12. Corpus del Espahol

Corpus del Espanol contains 100 million words of Spanish texts covering periods from the 1200s to post-1900s which are distributed as follows: 20 million words from the 1200s-1400s, 40 million from the 1500s-1700s, 40 million from the 1800s-1900s, and 20 million from the 1900s. The data from the 1900s is divided equally among literature, oral texts, and newspapers/encyclopedias.

A unique feature of this corpus lies in its use of relational databases, which are used to store texts and various types of annotation. These databases are linked to a very powerful web interface that supports different types of search such as patterns/wildcards, collocations, word forms, lemmas, synonyms, and grammatical categories. The corpus can be accessed online over the Web (see *Corpus del Español* in appendix).

6.13. Corpus do Português

Corpus do Portugues contains more than 45 million words in more than 50,000 Portuguese texts from the 1300s to the 1900s. The structure of the corpus allows for easy comparison of the frequency and distribution of words, phrases, and grammatical constructions across different parameters, e. g. historical period (data for each century forms a subcorpus), dialect (Portugal and Brazil Portuguese), and register (oral, fiction, newspaper, and academic writing).

The corpus is open to the public free of charge via the Web interface (see *Corpus do Português* in appendix), which allows users to search for exact words or phrases, substrings, lemmas, part-of-speech, or any combinations of these. It is also possible to search for collocates within a ten-word window (five to the left and five to the right of the search term).

7. Spoken corpora

While general corpora like national corpora may contain spoken material, there are a number of well-known publicity available spoken English corpora, which will be introduced in this section.

7.1. The London-Lund Corpus

The London-Lund Corpus (LLC), as the first electronic corpus of spontaneous language, is a corpus of spoken British English recorded from 1953−1987. The corpus derived from two projects: the Survey of English Usage (SEU) at University College London and the Survey of Spoken English (SSE) at Lund University. There are two versions of LLC, the original version consisting of 87 transcripts from SSE totaling 435,000 words, and the complete version, which has been augmented by 13 supplementary transcripts from SEU. The full LLC corpus comprises 100 texts, each of 5,000 words, totaling half a million running words. A distinction is made between dialogue (e. g. face-to-face conversations, telephone conversations, and public discussion) and monologue (both spontaneous and prepared) in the organization of the corpus (cf. Greenbaum/Svartvik 1990). This textual information is encoded together with speaker information (e. g. gender, age, occupation). The texts in the corpus are transcribed orthographically, with detailed prosodic annotation. The LLC corpus is available from ICAME.

7.2. SEC, MARSEC and Aix-MARSEC

The Lancaster/IBM Spoken English Corpus (SEC) consists of approximately 53,000 words of spoken British English, mainly taken from radio broadcasts dating between 1984 and 1991. For a corpus of this size, it is impossible to include samples of every style of spoken English. The SEC corpus has been designed to cover speech categories suitable for speech synthesis, as shown in Table 20.17 (see Taylor/Knowles 1988).

Tab. 20.17: The SEC categories

Code	Category	Words	Proportion
A	Commentary	9066	17%
B	News broadcast	5235	10%
C	Lecture aimed at general audience	4471	8%
D	Lecture aimed at restricted audience	7451	14%
E	Religious broadcast including liturgy		
F	Magazine-style reporting	4710	9%
G	Fiction	7299	14%
H	Poetry	1292	2%
J	Dialogue	6826	13%
K	Propaganda	1432	3%
M	Miscellaneous	3352	6%
Total		52637	c. 100%

In the SEC corpus, efforts have been made to achieve a balance between the highly stylized texts (e. g. poetry, religious broadcast, propaganda) and dialogue, and between male and female speakers. Of the 53 speakers in the corpus, 17 are female, representing 30% of the corpus. The higher proportions of male speakers in the news and commentary categories reflect the tendency of the BBC (at the time when the texts in the corpus originated) to use mainly male speakers in these types of programmes.

SEC is available in orthographic, prosodic, grammatically tagged and treebank versions, which should prove most useful to those who research in the speech synthesis or speech recognition fields. The corpus can be ordered from ICAME.

The Machine Readable Spoken English Corpus (MARSEC) is an extension of SEC in which the original acoustic recordings were digitalized, and word-level time-alignment between the transcripts and the acoustic signals was included. Tonetic stress marks were also converted into ASCII symbols to make the corpus machine-readable. The prosodically annotated word-level alignment files are available at the MARSEC website (see appendix).

The Aix-MARSEC database is a further development of MARSEC. The database consists of two major components: the digitalized recordings from MARSEC and the annotations. Annotations have so far been undertaken at nine levels such as phonemes, syllables, words, stress feet, rhythm units, and minor and major turn units. Two supplementary levels, the grammatical annotation by CLAWS and a Property Grammar system developed at Aix-en-Provence, are to be integrated soon (cf. Auran/Bouzon/Hirst 2004). The database, together with tools, is available under GNU GPL licensing at the Aix-MARSEC project site (see appendix).

7.3. The Bergen Corpus of London Teenage Language

The Bergen Corpus of London Teenage Language (COLT) is the first large English corpus focusing on the speech of teenagers. It contains half a million words (about 55

hours of recording) of orthographically transcribed spontaneous teenage talk recorded in 1993 by 31 volunteer recruits from five socially different school boroughs. The speakers in the corpus are classified into six age groups: preadolescence (0−9 years old), early adolescence (10−13), middle adolescence (14−16), late adolescence (17−19), young adults (20−29) and older adults (30+). As the name of corpus suggests, the core of the corpus represents teenagers. The early, middle and late adolescence groups account respectively for 24%, 61% and 9%, totaling 94% of the corpus. The older adult group, mostly parents and teachers, takes up 6%. As regards speaker gender, girls and boys contributed roughly the same amount of text: the male speakers about 51.8% (230,616 words) and the female speakers 48.2% (214,215 words). In terms of social class, only about 50% of the corpus material can be assigned a social group value. The material that has been classified is evenly distributed across the three social groups: high, middle, and low. While a wide range of settings are present in the COLT corpus, settings in connection with school (48%) and home (32%) are the most common. Speaker-specific information (speaker age, gender, social class, etc.) and conversation-specific information (location and setting) is encoded in the header of each corpus text. In the body of the text, paralinguistic features and non-verbal sounds are also marked up (cf Haslerud/ Stenström 1995).

The corpus constitutes part of the British National Corpus. In addition, COLT is released in both orthographically transcribed (pure text) and tagged versions (using the CLAWS C7 tagset). A prosodically annotated version (a representative selection amounting to approximately 150,000 words) is also available. The corpus is free for non-commercial purposes and can be accessed online by registered users (COLT, see appendix) or ordered form ICAME.

7.4. The Cambridge and Nottingham Corpus of Discourse in English

The Cambridge and Nottingham Corpus of Discourse in English (CANCODE) is part of the Cambridge International Corpus (CIC, see appendix). The corpus comprises five million words of transcribed spontaneous speech recorded in Britain and Ireland between 1994 and 2001, covering a wide variety of mostly informal settings: casual conversation, people working together, people shopping, people finding out information, discussions and many other types of interaction. As CANCODE is designed as a contextually and interactively differentiated corpus, the data has been carefully collected and sociolinguistically profiled with reference to a range of different speech genres and with an emphasis on everyday communication.

A unique feature of CANCODE is that the corpus has been coded with information pertaining to the relationship between the speakers: whether they are intimates (living together), casual acquaintances, colleagues at work, or strangers. For this purpose, CANCODE is organized along two main axes: context-type and interaction-type. Alongside the axis of context-type are, on the cline from "public" to "private", transactional, professional, socializing and intimate. Alongside the axis of interaction-type are, on the cline from "collaborative" to "non-collaborative", information provision, collaborative idea, and collaborative work. The interactions between the two axes, together with typi-

Tab. 20.18: CANCODE text types

Context-type Information provision	Interaction-type		
	Collaborative idea	Collaborative work	
Transactional	commentary by museum guide	chatting with hairdresser	choosing and buying a television
Professional	oral report at group meeting	planning meeting at place of work	colleagues window-dressing
Socializing	telling jokes to friends	reminiscing with friends	friends working together
Intimate	partner relating the story to a film seen	siblings discussing their childhood	couple decorating a room

cal settings, are shown in Table 20.18 (see Carter/McCarthy 2004, 67). This coding allows users to look more closely at how different levels of familiarity (formality) affect the way in which people speak to each other. The corpus is not currently available to the public.

7.5. The Spoken Corpus of the Survey of English Dialects

A corpus that was built specifically for the study of English dialects is the spoken corpus of the Survey of English Dialects (SED, see Beare/Scott 1999). The Survey of English Dialects was started in 1948 by Harold Orton at the University of Leeds. The initial work comprised a questionnaire-based survey of traditional dialects based on extensive interviews of about 1,000 people from 313 locations all over rural England. During the survey, a number of recordings were made as well as the detailed interviews. The recordings, which were made during 1948−1961, consist of about 60 hours of dialogue of people aged 60 or above talking about their memories, families, work and the folklore of the countryside from a century ago. Elderly people were chosen as subjects because they were most likely to speak the traditional, "uncontaminated" dialect of their area.

The spoken corpus derived from SED consists of transcripts of 314 recordings from 289 (out of the 313) SED localities in England, totaling roughly 800,000 running words. The original recordings were transcribed, with sound files linked to transcripts. The corpus is marked up in TEI-compliant SGML and POS tagged using CLAWS.

While the spoken corpus of SED comprises data invariably produced by elderly people, as the survey was conducted nationwide, covering every county of England, it has, for the first time, made it possible to conduct a detailed study of the regional variation in English dialects on a national level. Also, as the data reflects a society which was different in many ways from today, the corpus is a valuable resource for dialectologists and historical linguistics as well as historians. The CD-ROM of the spoken corpus is published by Routledge, London.

7.6. The Intonational Variation in English Corpus

The Intonational Variation in English (IViE) corpus was constructed for the investigation of cross-varietal and stylistic variation in British English intonation, focusing on

nine urban varieties of English spoken in the British Isles, i. e. Belfast, Bradford, Cambridge, Cardiff, Dublin, Liverpool, Leeds, London, and Newcastle. The corpus comprises 36 hours of speech data in five different speaking styles: phonetically controlled sentences (statements, questions without morpho-syntactic markers, WH-questions, inversion questions, coordination structures), a read text (the fairy tale *Cinderella*), a retold version of the same text, a map task ("find your way around a small town") and free conversations (on the assigned topic of smoking). The data was collected in urban secondary schools, and the speakers were 16 years old at the time when the recordings were made. A minimum of six male and six female speakers from each variety were recorded, though more speakers were included for some of the varieties, totaling 116 speakers in all (cf. Grabe/Post/Nolan 2001). The corpus is available free of charge for non-commercial use only. Orthographic and prosodic transcriptions, together with digitalized sound files can be ordered on CDs or downloaded from the corpus website (see appendix).

7.7. The Longman British Spoken Corpus

The Longman British Spoken Corpus contains 10 million words of natural, spontaneous conversations from a representative sample of the population in terms of speaker age, gender, social group and region, and from the language of lectures, business meetings, after dinner speeches and chat shows. The design criteria are discussed in detail in Crowdy (1993). The Longman British Spoken Corpus is the first large scale attempt to collect spoken data in a systematic way. The corpus is part of the spoken section of the British National Corpus (see section 2.1.).

7.8. The Longman Spoken American Corpus

The Longman Spoken American Corpus comprises five million words of spoken data collected from everyday conversations of more than 1,000 Americans of various age groups, levels of education, and ethnicity from over 30 US States. Equal numbers of participants were chosen from each region, and a balance was struck between the numbers of participants from rural and city areas within those regions. Recordings were made of four-hour chunks of the normal daily conversations of each participant over periods of at least four days. The participants were chosen to be representative for gender, age, ethnicity and education, as shown by the latest US demographic census statistics (Table 20.19, see Stern 1997). As part of the Longman Corpus Network (see appendix), the Longman Spoken American Corpus is a property of the Longman publishers for in-house use only.

Tab. 20.19: Demographic distribution of the Longman Spoken American Corpus

Variable	Proportions
Gender	Male: 50%; Female: 50%
Age	18–24: 20%; 25–34: 20%; 35–44: 20%; 45–60: 20%; 60+: 20%
Ethnicity	White: 75%; Black: 13%; Hispanic: 8%; Asian: 4%
Education	Degree/Higher degree: 33%; College: 33%; High school: 33%

7.9. The Santa Barbara Corpus of Spoken American English

The Santa Barbara Corpus of Spoken American English (SBCSAE) is based on hundreds of recordings of spontaneous speech from all over the United States, representing a wide variety of people of different regional origins, ages, occupations, and ethnic and social backgrounds. It reflects the many ways that people use language in their lives: conversation, gossip, arguments, on-the-job talk, card games, city council meetings, sales pitches, classroom lectures, political speeches, bedtime stories, sermons, weddings, etc. (cf. Du Bois et al. 2000−2004).

The corpus is particularly useful for research into speech recognition as each speech file is accompanied by a transcript in which phrases are time-stamped to allow them to be linked with the audio recording from which the transcription was produced. Personal names, place names, phone numbers, etc. in the transcripts have been altered to preserve the anonymity of the speakers and their acquaintances, and the audio files have been filtered to make these portions of the recordings unrecognizable. The SBCSAE corpus is distributed by the LDC in five parts, the first four of which have become available. The corpus (including both transcripts and digital audio files) can also be downloaded freely at the TalkBank site (see TalkBank SBCSAE in appendix).

7.10. The Saarbrucken Corpus of Spoken English

The Saarbrücken Corpus of Spoken English (SCoSE) consists of five parts. Parts 1 (Stories) and 3 (Jokes) comprise excerpts transcribed from audio-taped real conversations among family members and friends, fellow students and colleagues at Northern Illinois University and at Saarland University. Part 2 (Indianapolis Interviews) includes transcripts of stories recorded in interviews with senior citizens aged 80 and older in a retirement community in Indianapolis, Indiana in the summer of 2002. Conversations included in Part 4 (Complete Conversations) are transcripts from recordings made by two students in their junior year at a large state university near Chicago as a class assignment to record family and friends in natural settings during their Thanksgiving break at the end of November. The final part (Drawing Experiment) of the corpus consists of transcripts of conversations in which one subject describes a drawing to the other in the same pair who has not seen the picture. All subjects are young students aged 20−25 at various colleges near Chicago. In all of these parts, speech turns are indicated. The hard copy of the corpus (in PDF format), together with a description of transcription conventions, is available at no cost at the corpus site (cf. see SCoSE in appendix). The electronic copy of the corpus, together with digitalized audio files, is downloadable at the TalkBank site (see appendix).

7.11. The Switchboard Corpus

The Switchboard Corpus (SWB) is a corpus of 2,438 spontaneous telephone conversations, averaging 6 minutes in length, recorded for over 542 speakers of both sexes from every major dialect of American English in the early 1990s. The transcripts total three

Tab. 20.20: The Switchboard corpus

Dialect	Speaker age	Speaker sex	Education
South Midland (155) Western (85) North Midland (77) Northern (75) Southern (56) NYC (33) Mixed (26) New England (21)	20−29 (140) 30−39 (179) 40−49 (112) 50−59 (87) 60−69 (13)	Male (292) Female (239)	High school − (14) College − (39) College (309) College + (176) Unknown (4)

million words (over 240 hours of recordings). Information relevant to speakers' sex, year of birth, education level and dialect region is available in the documentation accompanying the corpus. Table 20.20 shows the distribution of major sociolinguistic variables (see Linguistic Data Consortium (1995), the Switchboard User's Manual).

As each transcript in the corpus is time-aligned at the word level, the corpus is useful for sociolinguistic studies as well as for speech recognition. The corpus is distributed by the LDC. A subcorpus annotated with various types of linguistic information (e. g. POS tagging and parsing) is also freely available for download at the TalkBank site (see TalkBank SWB in appendix). It can also be searched over the Web (see SWB online in appendix).

7.12. The Wellington Corpus of Spoken New Zealand English

The Wellington Corpus of Spoken New Zealand English (WSC) comprises one million words of spoken New Zealand English in the form of 551 2,000-word extracts collected between 1988 and 1994 (99% of the data from 1990−1994, the exception being eight private interviews). A very stringent criterion was adopted to ensure the integrity of the New Zealand samples included in the corpus. Data was collected only from those who had lived in New Zealand since before the age of 10, had spent less than 10 years (or half their lifetime, whichever was greater) abroad, and had not made an overseas trip during the year before data collection. The extracts are classified into 15 text categories covering a wide range of contexts in which each style of speech is found, as shown in Table 20.21 (cf. Holmes/Vine/Johnson 1998).

The formal speech section (12%) in the WSC corpus includes all monologue categories and "parliamentary debate" in the public dialogue category. The semi-formal section (13%) includes the three types of interview (both public and private). All of the other text categories make up the informal speech section (75%), with private conversation alone accounting for 50% of the corpus. In terms of speaker gender, women contributed 52% and men 48% of the final transcribed words, reflecting the New Zealand population balance. With regard to speaker age, data for the age group 20−24 accounts for more than 20% of the corpus, and the proportions for age groups 45−49 and 40−44 both exceed 10% while there is little data for those aged over 70. The distribution across different age groups generally mirrors the population structure in New Zealand. The corpus data also reflects the distribution of population across ethnic groups, with data

Tab. 20.21: Composition of the WSC corpus

Category	Text category	Words
Monologue: Public scripted, broadcast	Broadcast news	28,929
	Broadcast monologue	11,205
	Broadcast weather	3,641
Monologue: Public unscripted	Sports commentary	26,010
	Judge's summation	4,489
	Lecture	30,406
	Teacher monologue	12,496
Dialogue: Private	Conversation	500,363
	Telephone conversation	70,156
	Oral history interview	21,972
	Social dialect interview	31,058
Dialogue: Public	Radio talkback	84,321
	Broadcast interview	96,775
	Parliamentary debate	22,446
	Transactions and meetings	102,332
Total		1,046,599

collected for Pakeha accounting for 76%, and for Maori 18%. Every speech sample included in the corpus is described as fully as possible in terms of sociolinguistic variables such as the gender, age, regional origin, social class, level of education and occupation of its contributor.

The unusually high proportion of private material and the rich sociolinguistic variation make the WSC corpus a valuable resource for research into informal spoken registers as well as for sociolinguistic studies. The corpus is available from ICAME.

7.13. The Limerick Corpus of Irish English

The Limerick Corpus of Irish English (L-CIE) comprises one million words in the form of 375 transcripts of naturally occurring conversations recorded in a wide variety of speech contexts throughout Ireland (excluding Northern Ireland). Speakers range from 14 to 78 years of age and there is an equal representation of both male and female speakers. While the corpus consists mainly of casual conversation, it also has over 200,000 words of professional, transactional and pedagogic Irish English which, along with the casual conversation data, were carefully collected with reference to a range of different speech genres. The corpus follows the design of CANCODE by organizing the corpus alongside the axes of context type and interaction type, as shown in Table 20.22 (cf. Farr/Murphy/O'Keeffe 2004).

Tab. 20.22: Design of the L-CIE corpus

	Information provision	Collaborative idea	Collaborative task
Pedagogic	80,253 words e. g. linguistics lecture	60,473 words e. g. English poetry tutorial	10,000 words e. g. one-to-one computer lesson
Professional	145,000 words e. g. real-estate office talk	100,000 words e. g. team meeting	60,000 words e. g. waitresses washing dishes
Socializing	50,000 words e. g. describing a new bar	54,356 words e. g. friends discussing college	30,000 words e. g. friends assembling a bed
Intimate	60,000 word e. g. mother storytelling	266,000 words e. g. partners making holiday plans	60,000 word e. g. family preparing dinner
Transactional	5,000 words e. g. product presentation	10,000 words e. g. chatting in a taxi	1,000 words e. g. eye examination

While it is not designed to be geographically representative – it does not include data from every county in the Republic of Ireland, the L-CIE corpus has developed a careful sociolinguistic classification scheme which facilitates inter-corpus comparisons, especially with regard to linguistic choices and the relationships that hold between the speakers. The corpus website (see appendix) allows online access by registered users.

7.14. The Hong Kong Corpus of Spoken English

The Hong Kong Corpus of Spoken English (HKCSE) comprises 200 hours of orthographically transcribed recordings. The corpus is divided into four subcorpora (conversations, academic discourses, business discourses and public discourses, with about 50 hours of recordings for each), amounting to approximately two million words. The four subcorpora represent the main overarching spoken English discourses in Hong Kong. The compilation work began in the mid-1990s when half a million words of natural

Tab. 20.23: Structure of HKCSE

Subcorpus	Speech type	Size
Academic discourse	Lectures, seminars, student presentations, tutorials and supervisions, workshops for staff	28 hours 30 minutes
Business discourse	Service encounters, meetings, interviews, presentations and announcements, conference calls and videoconferencing, informal office talks, workplace phone calls	29 hours and 14 minutes
Public discourse	Speeches, talks plus interaction, press briefings with/without interaction, TV/radio interviews, discussion forums	25 hours
Conversation	Naturally occurring conversations	27 hours

conversations were recorded between Hong Kong Chinese and non-Cantonese speakers (mostly native speakers of English).

In addition to the orthographic transcription, part of the corpus has been annotated prosodically to enable the examination of the communicative role of intonation. Presently, 1.06 million words have been annotated, covering 53 % of the orthographic version of the corpus. Table 20.23 shows the contents of the HKCSE corpus (cf. Cheng/Greaves/Warren 2005).

Presently the prosodic version of HKCSE is perhaps the largest English corpus which has been annotated with prosodic details. In addition, a computer program (iConc) is specifically developed for the prosodic version of the corpus, which can search for tags for various prosodic features such as tone unit, tones, prominence, termination and key. The prosodic version of the corpus is expected to be released on CD-ROMs.

8. Academic and professional English corpora

As language may vary considerably across genre and domain, specialized corpora provide valuable resources for investigations in the relevant genres and domains. Unsurprisingly, there has recently been much interest in the creation and exploitation of specialized corpora in academic or professional settings. This section introduces a number of well-known English corpora of this kind.

8.1. The MICASE and MICUSP corpora

The Michigan Corpus of Academic Spoken English (MICASE) contains approximately 1.8 million words in the form of 152 transcripts of nearly 200 hours of recordings of 1,571 speakers, focusing on contemporary university speech within the domain of the University of Michigan. Table 20.24 shows the structure of the corpus (cf. English Language Institute (2003), the MICASE Manual).

Tab. 20.24: The MICASE corpus

Criterion	Distribution
Speaker gender	Male (46 %) Female (54 %)
Academic role	Faculty (49 %) Students (44 %) Other (7 %)
Language status	Native speakers (88 %) Non-native speakers (12 %)
Academic division	Humanities & Arts (26 %) Social Sciences & Education (25 %) Biological & Health Sciences (19 %) Physical Sciences & Engineering (21 %) Other (9 %)
Primary discourse mode	Monologue (33 %) Panel (8 %) Interactive (42 %) Mixed (17 %)
Speech event type	Advising (3.5 %) Colloquia (8.9 %) Discussion sections (4.4 %) Dissertation defenses (3.4 %) Interviews (0.8 %) Labs (4.4 %) Large lectures (15.2 %) Small lectures (18.9 %) Meetings (4.1 %) Office hours (7.1 %) Seminars (8.9 %) Study groups (7.7 %) Student presentations (8.5 %) Service encounters (1.5 %) Tours (1.3 %) Tutorials (1.6 %)

In the MICASE corpus, speakers are divided into four age groups: 17−23, 24−30, 31−50, and 51+. In terms of academic role, they are classified into a number of categories: junior and senior undergraduates, junior and senior postgraduates, junior and senior faculty and researchers, etc. The language status can be native speaker (North American English), other native speaker (non-American English), near native speaker, and non-native speaker.

The MICASE corpus was originally marked up in TEI-compliant SGML. All of the SGML files have now been converted to the XML format in order to meet the requirements for further corpus development including a web-based search interface and the streaming web delivery of the sound recordings, synchronized with the transcripts. At present, only the orthographically transcribed version of the corpus is available, though future releases will include various kinds of annotations such as part-of-speech, lemmas and discourse-pragmatic categories. The MICASE corpus can be searched online free of charge or ordered at a nominal fee at the corpus website (see appendix).

MICUSP is an acronym for Michigan Corpus of Upper-level Student Papers, an ongoing project which aims to compile a 1.6-million-word collection of 500 to 1,000 word samples of writing by students at different stages of undergraduate and graduate level study in both humanities and science subjects, both native and non-native speakers, from across the University of Michigan (see the MICUSP project site for updates).

8.2. The British Academic Spoken English corpus

The British Academic Spoken English (BASE) corpus, which is designed as a British counterpart to MICASE, was constructed jointly by the Universities of Warwick and Reading. The corpus comprises a collection of recordings and marked up transcripts of 160 lectures and 40 seminars, totaling approximately 196 hours of recordings and 1.6 million words. The lectures and seminars spread evenly across four subject areas, as shown in Table 20.25.

Tab. 20.25: Components of the BASE corpus

Subject area	Lectures	Seminars
Arts and Humanities	42	10
Social Studies and Sciences	40	11
Physical Sciences	40	9
Life and Medical Sciences	39	10
Total	161	40

Unlike MICASE, the BASE corpus only covers two types of speech event, lectures and seminars. Most of the recordings were made on digital video instead of audiotapes. All of these recordings have been transcribed and marked up in TEI-compliant XML. The corpus will not only enable research into spoken academic English at the lexical and structural levels, it will also make it possible, when used in combination with MICASE, to compare academic spoken English in British and US university settings. The British

Academic Spoken English data (transcripts in plain text and XML formats and audio/video files) can be downloaded at the BASE website (see BASE in appendix), but only authorized users can access them.

The British Academic Written English (BAWE) corpus of student writing is a British cousin of MICUSP. This is a selection of about 3,000 student assignments from four disciplinary groupings (Arts and Humanities, Life Sciences, Physical Sciences, and Social Sciences), which are sampled in 28 departments across the three British universities (Oxford Brookes, Reading and Warwick). These samples represent both undergraduate work (typically three years of study) and postgraduate work (typically one year of study) of a high quality (graded II.i (B+) or above). The corpus is marked up in TEI-compliant XML, with metadata such as student gender, year of birth, first language, course of study, year of study, module name and code, etc. recorded in the corpus header. Textual organization in each assignment is also marked up to show title and title page, table of contents, abstract or summary, section headings, figures and diagrams, lists (simple, bulleted and ordered), quotations, bibliography, and appendices. Boundaries for paragraphs and sentences are also marked up. The structure of the BAWE corpus together with the metadata encoded make it possible to compare textual organization across years, text types, disciplines and disciplinary groupings. This corpus is freely available to researchers who agree to the license conditions (see BAWE in appendix).

8.3. The Reading Academic Text corpus

The Reading Academic Text (RAT) corpus is a collection of academic texts written by academic staff and research students at the University of Reading. The initial corpus was composed of twenty research articles written by staff and a small number of PhD theses contributed by successful doctoral candidates in the Faculty of Agriculture, totaling nearly a million words. The theses included in the corpus are all written by native speakers. Since the corpus was created in 1995, the number of theses has increased from 8 to 38. The corpus is still expanding further to represent the discourses of a greater range of disciplines covering both the natural and social sciences as well as a wider range of text types including dissertations, projects, laboratory reports, and samples of textbook readings for Master's courses. In addition to the original files, the texts have been converted to an HTML version which allows the full text to be viewed in a browser, and a plain text version used for linguistic analysis and for the coding of the corpus (see Thompson 2001). The RAT corpus has been used to study text construction practices in academic settings such as the organization of theses in different disciplines as well as the various uses of citations. At present, access to the corpus is restricted to the staff and researchers at the School of Linguistics and Applied Language Studies of Reading University, though it is possible for other users to access the corpus on a Research Attachment arrangement.

8.4. The Academic Corpus

The Academic Corpus is a written corpus of academic English developed at Victoria University of Wellington. The corpus contains approximately 3.5 million words, covering 28 subject areas from four faculty sections (arts, commerce, law, and science), as shown in Table 20.26 (cf. Coxhead 2000, 220).

Tab. 20.26: Subject areas in the Academic Corpus

Faculty	Arts	Commerce	Law	Science	Total
Texts	122	107	72	113	414
Words	883,214	879,547	874,723	875,846	3,513,330
Subject areas	Education History Linguistics Philosophy Politics Psychology Sociology	Accounting Economics Finance Industrial relations Management Marketing Public policy	Constitutional law Criminal law Family law and medicolegal International law Pure commercial law Quasi-commercial law Rights and remedies	Biology Chemistry Computer science Geography Geology Mathematics Physics	

Each of these faculty sections is divided into seven subject areas of ca. 125,000 words, totaling 875,000 words for each section. The corpus comprises 414 academic texts by more than 400 authors which were sampled from journal articles, book chapters, course workbooks, laboratory manuals, course notes and the Internet. With the exception of 41 excerpts from the Brown corpus, 31 excerpts from LOB and 42 excerpts from the Wellington Corpus of Written New Zealand English, full texts (excluding bibliographies) are included. The majority of the texts were written for an international audience, with 64% sourced in New Zealand, 20% in Britain, 13% in the United States, 2% in Canada and 1% in Australia. The texts were selected according to whether they were of suitable length (over 2,000 running words) and were representative of the academic genre in that they were written for an academic audience. Efforts have also been made to balance the corpus with respect to the number of short (2,000−5,000 words), medium-length (5,000−10,000 words) and long (over 10,000 words) texts in the four faculty sections.

The corpus has been used to develop an Academic Word List (AWL) containing 570 word families (see Coxhead 2000), which is available at the AWL site (see appendix).

8.5. The Corpus of Spoken Professional American English

The Corpus of Spoken Professional American English (CSPAE) has been constructed using a selection of transcripts of interactions of various types occurring in professional settings recorded during 1994−1998. The corpus contains two million words of speech involving over 400 speakers. The CSPAE corpus has two main components. The first component is made up of transcripts (0.9 million words) of press conferences from the White House, which contains almost exclusively question and answer sessions in addition to some policy statements by politicians and White House officials. The second component consists of transcripts (1.1 million words) of faculty meetings and committee meetings related to national tests, which involve statements and discussions, as well as questions (cf. Barlow 1998).

The transcripts in the corpus have been marked up in a minimal but consistent way. The markup scheme only indicates speech turns by identifying the last name of the speaker (or VOICE if the name is unknown) with the <SP> element, and puts non-verbal events such as laughter in the brackets. Two versions of the corpus are available,

a raw text version and an annotated version tagged by the Lancaster CLAWS tagger. Both versions can be ordered from the corpus website (see CSPAE in appendix).

8.6. The Corpus of Professional English

A much more ambitious project has been initiated by the Professional English Research Consortium (PERC), which aims to create a 100-million-word Corpus of Professional English (CPE). The corpus is expected to include both spoken and written discourse used by working professionals and professionals-in-training, covering a wide range of domains such as science, engineering, technology, law, medicine, finance and other professions. The CPE corpus is designed as a balanced representation of professional English via texts published between 1995 and 2001 by over 1,000 major review and research journals, trade magazines, and textbooks, in American and British English, based on selection criteria such as impact factors provided by the *Journal of Citation Reports*, and other pertinent criteria (cf. Rayson et al. 2005).

The Corpus of Professional English is marked up in XML. Contextual information such as author's name, title, publication year and journal title is stored in the corpus header. Structural information is also encoded to show paragraphs, sections, headings and similar features in written texts. Linguistic annotations such as POS and semantic tagging will be carried out on the corpus using tools developed at Lancaster University.

The CPE corpus can be used for linguistic research as well as for the development of educational resources, such as specialized dictionaries, handbooks, language tests, and other materials that will be useful to working professionals and professionals-in-training. The corpus, when completed, will be made available to consortium members for online access at the PERC website (see PERC in appendix).

9. Parsed corpora

Parsing, also called treebanking, is a form of corpus annotation (see article 13). It is independent of corpus design criteria. Hence, a corpus, whether balanced or specialized, whether written or spoken, can be syntactically parsed. However, as parsing is a much more challenging task which often necessitates human correction, parsed corpora are typically very small in size. Of the corpora we have introduced so far, only ICE-GB is parsed. This section introduces a number of well-known parsed corpora.

9.1. The Lancaster-Leeds Treebank

The Lancaster-Leeds Treebank is perhaps the first syntactically parsed corpus. The corpus is a subset of 45,000 words taken from all text categories in the LOB corpus, which was parsed manually by Geoffrey Sampson using a specially devised surface-level phrase structure grammar compatible with the CLAWS word-tagging scheme (cf. Sampson 1987). The annotation scheme used in the Lancaster-Leeds Treebank, which consisted

of 47 labels for daughter nodes (14 phrase and clause classes, 28 word classes and five classes of punctuation marks), represented surface grammar only, without indications of logical form. This hand-crafted treebank provided training data for the automatic probabilistic parser which was used to analyze the Lancaster Parsed Corpus. The corpus was not published but is available from UCREL at Lancaster University (see appendix).

9.2. The Lancaster Parsed Corpus

The Lancaster Parsed Corpus (LPC) is a much larger sample of approximately 144,000 words taken from the LOB corpus that has been parsed. Except for categories M (science fiction, six samples) and R (humor, nine samples), which are included in their entirety, LPC takes the first 10 samples from each of the other 13 text categories in LOB, totaling 145 files which account for 13.29% of the full LOB corpus. Even in these 145 samples, longer sentences have been excluded from the parsed corpus because the parser was unable to process sentences over 20−25 words in length, with the result that the parsed corpus no longer contains LOB text extracts in their entirety. The errors resulting from automatic parsing were corrected by hand to ensure the corpus is reasonably error free (cf Garside/Leech/Váradi 1992).

The Lancaster Parsed Corpus can be regarded as a treebank broadly representative of the syntax of written English across a great variety of styles and text types. It provides a testbed for wide-coverage general-purpose grammars and parsers of English and a valuable resource for quantitative linguistic studies of English syntax. The corpus is available through ICAME.

9.3. The SUSANNE corpus

The SUSANNE (an acronym for "surface and underlying structural analysis of natural English") is a 130,000-word sub-sample taken from the Brown corpus of American English that has been parsed. The parsed corpus comprises 64 text samples, with 16 taken from each of the four text categories: A (press reportage), G (belles-letters, biography and memoir), J (learned writing) and N (adventure and Western fiction).

The parsing was largely undertaken manually in accordance with the SUSANNE analytic scheme developed by Geoffrey Sampson in collaboration with Geoffrey Leech on the basis of samples from written British and American English. The SUSANNE scheme is perhaps the first serious attempt to produce a comprehensive, fully explicit annotation scheme for English grammatical structure.

In SUSANNE, a parse tree is represented as a bracketed string, with the labels of non-terminal nodes inserted between opening and closing brackets. There are three types of information in the parsing scheme: a form tag, a function tag and an index. The hierarchy of form tag ranks (word, phrase, clause and root) defines the shape of a parse tree. The function tags identify surface roles such as surface and logical subject, agent of passive, and time and place adjuncts. An index shows referential identity between nodes (cf. Sampson 1995).

The SUSANNE corpus was first released in 1992 and its latest version, Release 5, was published in 2000. Each successive release has corrected errors found in earlier

releases. The latest release, together with the documentation accompanying the corpus, is distributed free of charge at the SUSANNE website (see appendix).

More recently two derivations of the SUSANNE Treebank have been produced. One is SEMiSUSANNE, which covers 33 of the 64 texts from SUSANNE, and supplements the grammatical annotations of the SUSANNE scheme with semantic annotations identifying the WordNet (1.6) senses in which vocabulary items are used. SEMiSUSANNE is freely available at the SUSANNE website. The other recent version of SUSANNE is in XML-based GXL (Graph eXchange Language), which can be downloaded freely at the Indogram (Induction of Document Grammar for Webgenre Representation) project site (see XGL in appendix).

9.4. The CHRISTINE corpus

The CHRISTINE corpus is a spoken counterpart to SUSANNE, developed by Geoffrey Sampson and his team. It is one of the first treebanks of spontaneous speech. The CHRISTINE analytic scheme includes explicit extensions to the SUSANNE annotation which are designed to handle speech phenomena such as pauses, discourse items and speech repairs. The first stage of CHRISTINE (CHRISTINE/I), which was released in 1999, is based on 40 extracts chosen at random from the demographically sampled component in the spoken BNC and other sources, totaling approximately 80,500 words of spoken data representing 147 identified speakers in addition to a great number of unidentifiable speakers. The information about speakers and the metadata originally contained in the BNC corpus header were converted into database files accompanying the corpus (cf Sampson 2000).

The full version of the CHRISTINE corpus includes 66 further texts drawn from the spoken BNC and other sources. The overall proportion of the BNC data accounts for 50% of the full CHRISTINE corpus, with 40% from the London-Lund corpus and 10% from the Reading Emotional Speech Corpus (see Stibbard 2001 for a description). The full release also incorporates a minor change in the distribution of analytic information between the fields to make it more compatible with SUSANNE and easier to read. This version became available in 2000. At present CHRISTINE in plain text and XML can be downloaded at the corpus website (see CHRISTINE and XGL in appendix).

9.5. The LUCY corpus

The LUCY corpus is the third in Sampson's series of treebanks. This corpus represents written English in modern Britain, ranging from published prose to the less skilled writing of young adults, and spontaneous writing by children aged 9–12. To deal with writing of this latter type, the LUCY parsing scheme contains some further extensions to the SUSANNE scheme which can identify cases where an unskilled writer fails to put words together in a meaningful way (cf. Sampson 2005).

There are 239 text files in LUCY, amounting to 165,000 words. The corpus consists of three sections: polished writing (41 text files, 102,000 words), young adult writing (48 text files, 33,000 words), and child writing (150 files, 30,000 words). The polished texts are taken from both informative and imaginative categories in the written section of the

British National Corpus. The young adult writing comprises three groups, namely, A-level general study scripts, access-course coursework, and first-year undergraduate essays. The child writing section is composed of material from the Nuffield corpus, a collection of writing by children aged between 9 and 12 years in 1965.

In addition to providing a valuable source of information on the realities of skilled written usage in modern Britain, LUCY holds the promise to support study of the process through which English-speaking children acquire writing skills. The current version of the corpus is Release 2, which became available in late 2005, and has corrected a number of errors in the initial release of 2003. As the data from the BNC (about half of the corpus) is copyright protected, a copyright free edition and an unreduced edition were prepared. The only difference is that in the copyright free edition, for those files where copyright is an issue, the words of the original texts are replaced by abbreviations. While these abbreviations may be recoverable to human eyes, they are by no means recoverable computationally. This reduced version is available from LUCY website (see appendix). The unreduced edition is only available to those who have purchased a copy of the BNC corpus. The XML version is also available (see XGL in appendix).

9.6. ICE-GB

The British component of the International Corpus of English (ICE-GB) is the first corpus that has been completed in the ICE series. Like all of the ICE components, ICE-GB comprises 300 spoken and 200 written texts from 32 categories, amounting to one million words. As noted in section 5.1., this corpus is not only POS tagged but also fully parsed and hand checked. The corpus contains 83,394 parse trees, including 59,640 in the spoken part of the corpus. Each node in the tree is labelled with up to three types of information: word class/syntactic category, syntactic function and features (e. g. transitivity), the latter being optional (cf Nelson/Wallis/Aarts 2002).

Unlike the SUSANNE, CHRISTINE and LUCY corpora, which come without retrieval software, ICE-GB is distributed together with a utility program, ICECUP, which allows very complex queries of various kinds, e. g. markup queries, exact and inexact grammatical node queries, text fragment queries, Fuzzy Tree Fragment (FTF) queries, and sociolinguistic variable queries.

The second full release of the corpus and ICECUP can be ordered on CD-ROMs from the ICE-GB website (see appendix). The ICE-GB sampler, which includes ICECUP and ten ICE-GB texts, is also available free of charge at the site. The digitized speech recordings of the spoken part of the corpus, aligned with the text, can be ordered as an option (11 CD-ROMs) with Release 2 of ICE-GB, which also includes an updated version of ICECUP (3.1) that can play audio files. This feature allows researchers to hear the original source of what they see on-screen. In addition to the online help included in ICECUP, Nelson/Wallis/Aarts (2002) provides a comprehensive reference guide to both corpus and software.

9.7. The Diachronic Corpus of Present-Day Spoken English

The Diachronic Corpus of Present-Day Spoken English (DCPSE) is composed of 400,000 words from the spoken section of ICE-GB (collected in the early 1990s, see

section 9.6.) and 400,000 words from the London-Lund Corpus (late 1960s−early 1980s, see section 7.1.). The corpus DCPSE was parsed using ICE-GB as a gold standard, and the parsing has been corrected by a variety of methods to ensure a high quality. The corpus is particularly useful in research of recent change in grammar of spoken English. This resource is available for order on CD-ROM (see DCPSE in appendix), which comes with ICEUP, the same exploration tool as ICE-GB.

9.8. The Penn Treebank

The Penn Treebank (PTB) is an example of skeleton parsing. Three releases of the treebank have so far been published by the LDC. The original release (Penn Treebank I, 1992) contains over 4.5 million words of American English data. The whole corpus is POS tagged while two thirds of the data is parsed. All of this material has been corrected by hand after automatic processing. Table 20.27 shows the components of Penn Treebank Release I (cf. Marcus/Santorini/Marcinkiewicz 1993).

Tab. 20.27: Penn Treebank Release 1

Component	Parsed words	Parsed words
Dow Jones news stories	3,065,776	1,061,166
Brown corpus retagged	1,172,041	1,172,041
Dept. of Energy abstract	231,404	231,404
MUC-3 messages	111,828	111,828
Library of America texts	105,652	105,652
IBM manuals	89,121	89,121
Dept. of Agriculture bulletins	78,555	78,555
ATIS sentences	19,832	19,832
WBUR radio transcripts	11,589	11,589
Total	4,855,798	2,881,188

Penn Treebank Release I applies a parsing scheme which is extended and modified on the basis of the Lancaster parsing scheme. While both annotation schemes employ a phrase structure grammar which covers noun, verb, adjective, adverbial and prepositional phrases, the Lancaster scheme also distinguishes between different clause types such as adverbial clause, comparative clause, nominal clause and relative clause whereas the Penn Treebank scheme differentiates between different types of *wh*-clauses (e. g. noun, adverb and prepositional phrases). The latter also includes a variety of null elements which indicate, for example, the understood subject of infinitive or imperative verbs, and its zero variant in subordinate clauses.

Penn Treebank Release 2, which was published in 1995, features the new Treebank II bracketing style. The new bracketing style is designed to facilitate the extraction of

simple predicate/argument structure (see Marcus et al. 1994). Penn Treebank Release 2 contains one million words of 1989 *Wall Street Journal* material and a small sample of ATIS-3 material annotated in Treebank II style in addition to a cleaned copy of the Release 1 material annotated in Treebank I style. Penn Treebank Release 3 (1999) includes tagged and parsed Switchboard transcripts which are also dysfluency-annotated, as well as the parsed texts from the Brown corpus. The Penn Treebank can be ordered on CD-ROM from the LDC. The corpus is also searchable free of charge via the LDC Online (see Penn Treebank in appendix).

9.9. Parsed historical corpora

In addition to the treebanks of present-day English introduced above, this section introduces a number of parsed historical corpora. These corpora are largely based on the diachronic part of the Helsinki Corpus.

The Penn-Helsinki Parsed Corpus of Middle English version 2 (PPCME2) is a corpus of prose text samples of Middle English, annotated for syntactic structure to allow searching not only for words and word sequences but also for syntactic structures. Based on the Middle English section of the Helsinki Corpus (with additions and deletions), PPCME2 comprises 55 text samples amounting to 1.3 million words. The annotation scheme for the corpus follows the basic formatting conventions of the Penn Treebank (Kroch/Taylor 2000). PPCME2 is an improved and extended version of an earlier corpus, PPCME1, which was smaller (510,000 words) and which used a simpler annotation scheme (no POS tagging, no indication of the internal structure of noun phrases, less detailed annotation of several complex sentence and phrase types). Both versions of the corpus are available at the corpus website (see appendix). PPCME1 is free for downloading while PPCME2 can be ordered on CD-ROM at a nominal cost. The corpus search tool, CorpusSearch, is freely available.

The York-Helsinki Parsed Corpus of Old English Poetry is a selection of poetic texts from the Old English Section of the Helsinki Corpus which have been annotated to facilitate searches on lexical items and syntactic structures. The corpus contains 71,490 words of Old English text samples ranging from 4,000 to 17,000 words in length. The Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English is a selection of texts from the Old English Section of the Helsinki Corpus. The corpus contains 106,210 words of Old English text samples, ranging 5,000 to 10,000 words in length, which represent a range of dates of composition, authors and genres. A much larger corpus with much more detailed annotation is the York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE), which contains 1.5 million words of Old English prose texts taken from the Toronto Dictionary of Old English Corpus, with special formatting which has made it possible to search conveniently for syntactic structures using a computer search engine. These corpora apply the PPCME2 annotation scheme. They are available at no cost for non-commercial use at the corpus website (see appendix) or via OTA.

10. Developmental and learner corpora

Two types of corpora are particularly relevant to language learning: developmental corpora and learner corpora. A learner corpus is a collection of the writing or speech pro-

duced by learners acquiring a second language (L2). The term is used here as opposed to a developmental corpus, which consists of data produced by children acquiring their first language (L1). This section introduces well-known corpora of these two types.

10.1. The Child Language Data Exchange System

The Child Language Data Exchange System (CHILDES) is an international database organized for the study of first and second language acquisition. The database consists of three parts: Codes for the Human Analysis of Transcripts (CHAT), Computerized Language Analysis (CLAN), and a database. The CHILDES database contains transcripts of data collected from children and adults who are learning both first and second languages. The total size of the database is now approximately 300 million characters. The database, which includes a wide variety of language samples from a wide range of ages and situations, consists of five major components: English data, non-English data, narrative data, data from clinical populations, data from bilinguals and second-language acquisition. Some files have associated audio and video recordings. The transcripts from English-speaking children constitute over half of the total CHILDES database, but up to 26 languages are currently covered. All of the data is transcribed in the CHAT format and can be analyzed using the CLAN programs, which support four basic types of linguistic analysis: lexical analysis, morpho-syntactic analysis, discourse analysis, and phonological analysis (cf MacWhinney 1995).

The CHILDES database has been used in a wide range of research of normal and abnormal child language. The database and computer programs are freely available for research at the CHILDES website (see appendix).

10.2. The Louvain Corpus of Native English Essays

The Louvain Corpus of Native English Essays (LOCNESS) is a corpus of argumentative essays on a great variety of topics written by native British and American university students (cf. Granger/Tyson 1996). The LOCNESS corpus comprises three parts, 114 British pupils' A-Level essays (60,209 words), 90 British university students' essays (95,695 words), and 232 American university students' essays (168,400 words), totaling 324,304 words. As the age group of those students is comparable to that of the non-native EFL students in the International Corpus of Learner English (ICLE, see section 10.4.), LOCNESS provides control data in comparing writings of native and non-native learners. The corpus can be ordered from the Centre for English Corpus Linguistics at the University of Louvain (CECL, see appendix).

10.3. The Polytechnic of Wales corpus

The Polytechnic of Wales (POW) corpus contains 65,000 words of informal conversations of about 120 6 to 12-year-old children, which were collected between 1978 and 1984 in South Wales. The children were selected in order to minimize any Welsh or other

second language influence and divided into four groups of 30, each within three months of the ages 6, 8, 10, and 12. These groups were subdivided by sex (boys, girls) and socio-economic class (A, B, C, D). The corpus is fully parsed by hand using a Systemic Functional Grammar with rich syntactico-semantic categories, capable of handling raising, dummy subject clauses, ellipsis, and replacement strings (cf. Souter 1993). The corpus contains 11,396 parse trees in 184 files, each file with a reference header which identifies the age, sex and social class of the child, and whether the text is from a play session or an interview. Only the parsed corpus is available in machine readable form via ICAME or OTA. The recorded tapes and 4-volume transcripts with intonation contours are available in hard copy from the British Library.

10.4. The International Corpus of Learner English

The first and best-known learner corpus is the International Corpus of Learner English (i. e. ICLE). The corpus comprises argumentative essays written by advanced learners of English, i. e. university students of English as a foreign language (EFL) in their 3rd or 4th year of study. The primary goal of ICLE is the investigation of the interlanguage of the foreign language learner (cf. Granger 2003).

ICLE version 1.1, published on CD-ROM in 2002, contained over 2.5 million words in the form of 3,640 texts ranging between 500−1,000 words in length written by EFL learners from 11 mother tongue backgrounds, namely, Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, and Swedish. The corpus is still expanding with additional subcorpora (each containing 200,000 words) of ten other L1 backgrounds including Brazilian Portuguese, Chinese, Greek, Japanese, Lithuanian, Norwegian, Portuguese (Portugal), Slovene, South African (Setswana) and Turkish (cf. the ICLE website, see appendix). ICLE published on CD-ROM (version 1.1) is not tagged for part-of-speech or learner errors. The error and POS-tagged versions of the corpus are expected to become available in the near future.

In addition to allowing the comparison of the writing of learners from different backgrounds, the corpus can be used in combination with LOCNESS to compare native and learner English. The ICLE corpus is available for linguistic research but cannot be used for commercial purposes. The ICLE corpus (version 1.1) on CD-ROM accompanied by a handbook can be ordered by following the link at the website of the Centre for English Corpus Linguistics (see CECL in appendix).

10.5. The LINDSEI corpus

The Louvain International Database of Spoken English Interlanguage (LINDSEI) is a spoken counterpart to ICLE. Each subcorpus represents an L1 background and comprises transcripts of fifty 15-minute interviews with 3rd and 4th year university students. The first component of LINDSEI contains transcripts of interviews with 30 female and 20 male French learners of English, totaling ca. 100,000 words. The database has now been expanded with additional components representing other L1 backgrounds including Bulgarian, Chinese, Dutch, German, Greek, Italian, Japanese, Polish, Spanish, and Swedish (see LINDSEI in appendix). As most learner corpora have used written data

only, this type of data allows new research into a wide range of features of oral interlanguage by comparing learner data with comparable native speaker data or data produced by learners from different L1 backgrounds.

10.6. The Longman Learners' Corpus

The Longman Learners' Corpus contains ten million words of essays written during 1990–2002 by students of English at a range of levels of proficiency from 20 different L1 backgrounds. The elicitation tasks varied, ranging from in-class essays with or without the use of a dictionary to exam essays or assignments. Each script in the corpus is coded for the student's L1 background, proficiency level, text type (essay, letter, exam script, etc.), target variety (British, American or Australian English), and for the country of residence. This corpus has been designed to provide balanced and representative coverage for each of these categories (cf. Gillard/Gadsby 1998, 160). Taken as a whole it offers a multi-faceted picture of interlanguage, which can be explored in a variety of ways. The Longman Learners' Corpus is not POS tagged, but part of the corpus has been error-tagged manually, although this portion is only for internal use by the Longman publishers. The Longman Learners' Corpus is a commercial corpus, but it is also available for academic use. At present around 10 million words can be supplied. Users can also order a subcorpus for a certain proficiency level or L1 background. For details, see the Longman website (see appendix).

10.7. The Cambridge Learner Corpus

As part of the Cambridge International Corpus (CIC), the Cambridge Learner Corpus (CLC) is a large collection of examples of English writing from learners of English all over the world. The English in the CLC comes from anonymized exam scripts written by students taking Cambridge ESOL English exams worldwide. The corpus currently contains over 22 million words in the form of 85,000 scripts from 180 countries (representing 100 different L1 backgrounds) and it is expanding continually. Each script is coded with information about the student's first language, nationality, level of English, age, etc. Over twelve million words (or about 35,000 scripts) have been coded for errors using the Learner Error Coding system developed by Cambridge University Press. CLC is a commercial corpus. Currently the corpus can only be accessed by authors and writers working for Cambridge University Press and by members of staff at Cambridge ESOL (cf. the CLC site, see appendix).

10.8. Other learner corpora

In addition to the corpora which cover multiple L1 backgrounds as introduced above, there are a number of learner corpora specific to one particular mother tongue.

The HKUST Corpus of Learner English is one such example. The corpus contains 25 million words of essays and exam scripts of upper-secondary and tertiary-level Chi-

nese learners of English in Hong Kong (mainly Cantonese speakers). The average length of these essays is 1,000 words. The corpus is partly tagged for part-of-speech and learner error (see Milton/Chowdhury 1994). The HKUST learner corpus is available to the public for use in research on a collaborative basis.

The Chinese Learner English Corpus (CLEC) contains one million words of writing produced by Chinese learners of English from five proficiency levels: high-school students, junior and senior non-English majors, and junior and senior English majors. The five types of learners are equally represented in the corpus. The CLEC material includes writings for tests, guided writings and free writings. The corpus is not POS tagged, but it is fully annotated with learner errors using an annotation scheme which consists of 61 error types clustered in 11 categories (see Gui/Yang 2002). The CLEC corpus can be searched online at the CLEC website (see appendix).

The JEFLL (Japanese EFL Learner) is a 700,000-word collection of spontaneous compositions (without the help of dictionaries or any careful revision, completed in 20 minutes) written by more than 10,000 Japanese learners of English at beginning and intermediate levels, covering mainly junior and senior high school students in Japan. The essay task used in data collection is carefully controlled so that each subcorpus can be comparable across topics, proficiency, school years, and school types, among others. JEFLL is POS tagged and tagged for learner errors. The corpus is made publicly available for free online access via the Shogakukan Corpus Network (see appendix).

The Standard Speaking Test (SST) corpus, also known as the NICT JLE (Japanese Learner English) Corpus, contains one million words of error-tagged spoken English produced by Japanese learners. Based entirely upon the audio-recordings of an English oral proficiency interview test called the Standard Speaking Test (SST), the corpus comprises 1,200 samples transcribed from 15-minute oral interview tests (around 300 hours of recording in total). This is the largest spoken learner corpus which has been built to date. The subjects are classified into nine SST proficiency levels, thus making it possible to compare speech across different learner proficiency groups. Two types of tagging have been used in the SST corpus: discourse tagging and error tagging. The tags are XML-compliant. More than 30 basic tags are used to mark up discourse phenomena in the learners' utterances, which are clustered into four main categories: tags for representing the structure of the entire transcription file, tags for the interviewee's profile, tags for speaker turns, and tags for representing utterance phenomena such as fillers and repetitions (see Izumi/Uchimoto/Isahara 2004, 34). The error tagging scheme consists of 47 tags. Each tag shows three types of information: part-of-speech, a grammatical/lexical rule, and a corrected form (cf Izumi/Isahara 2004). More details on the SST corpus can be found at the SST corpus website (see appendix).

The Thai English Learner Corpus (TELC) currently contains 1.5 million words of writings by Thai learners of English. One half of the materials were taken from university entrance exams at the Institute for English Language Education (IELE, Assumption University) and the other half came from writings by 4th-year undergraduate students of EFL at Assumption University. The TELC corpus is tagged for part-of-speech and lemma. The corpus is presently not open to the public.

The Uppsala Student English (USE) corpus contains 1.2 million words in the form of 1,489 essays written during 1999–2001 by 440 Swedish university students of English at three different levels, the majority in their first term of full-time studies. These essays

were written out of class, against a deadline of 2−3 weeks, with length limitations imposed (usually 700−800 words), and suitable text structure suggested. There are a variety of essay types in the corpus, including evaluation, argumentation, and discussion, etc. The corpus is available for non-commercial research and educational use only. More information about the corpus is available at the USE site (see USE in appendix) and the corpus can be ordered via the Oxford Text Archive (see OTA in appendix).

The Polish Learner English Corpus is designed by the PELCRA project (see section 2.3.) as a half-a-million-word corpus of written learner data produced by Polish learners of English from a range of learner styles at different proficiency levels, from beginning learners to post-advanced learners (cf. Lewandowska-Tomaszczyk 2003, 107). The data was collected between 1998 and 2000 from the exam essays of Polish learners of English at the Institute of English Studies in Łódź and two teacher-training colleges affiliated with the University of Łódź. Each data file contains a "TEI lite" conformant header. The corpus is tagged using CLAWS with the standard C7 tagset. Learner errors are identified by comparing the questionable language portions in the learner corpus with materials from native English corpora (e. g. the BNC and ANC) on the one hand, and the PELCRA corpus of native Polish on the other. Some sample files are available at the PELCRA project site (see appendix).

The JPU (Janus Pannonius University) learner corpus contains 300,000 words of essays and research papers by advanced level Hungarian university students, which were collected from 1992 to 1998. JPU has five subcorpora: Postgraduate, Writing and Research Skills, Language Practice, Electives and Russian Retraining (cf. Horváth 1999). The essays are available at the JPU corpus site (see appendix) while the whole corpus is searchable via the website Lexical Tutor (see appendix).

All of the learner corpora introduced above are for English, given the status of English as an international language. There are, however, a number of existing interlanguage corpora for other languages. For example, the Progression Corpus is a longitudinal corpus of French which was developed to investigate progression in foreign language learning in the early years of secondary schooling, with specific reference to French as a first foreign language. The corpus contains around 200 hours of spoken French produced by a cohort of 60 children who were tracked through two years (six terms) of classroom French, from the second term of year 7 until the first term of year 9 inclusive. The corpus is encoded following the CHAT standards of the CHILDES system (see section 10.1.). The transcripts and audio files of the corpus can be accessed at website of the French Progression Corpus (see appendix). Another learner corpus of spoken French is the Linguistic Development Corpus, which was created to complement the Progression Corpus. This is a cross-sectional corpus which is composed of 240 digitally recorded sound files, as well as their transcripts and tagged files (also in CHAT format). The data contained in this corpus was produced by children of years 9, 10 and 11 of secondary (aged 13−16) education in the UK context. The corpus is accessible at the French Development Corpus website (see appendix).

The Chinese Interlanguage Corpus contains over one million Chinese characters in the form of 1,731 writing samples produced by 740 learners of Mandarin Chinese as a foreign language in nine universities in China. The data was sampled from 5,574 compositions and exercises (totaling over 3.5 million Chinese characters) produced by 1,635 Chinese learners from 96 countries and regions. The corpus is richly encoded with 23 items of metadata information including for example, learner's name, sex, age, national-

ity, native language, education level, textbooks used, and date of writing. Errors of various types (mainly lexical errors) are also tagged and indexed for easy retrieval. The corpus comes with an integrated system that allows users to perform tasks such as keyword and full text searching, as well as data browsing and processing. The corpus is currently open to on-site use only.

11. Multilingual corpora

We have so far introduced major monolingual corpora of English and a number of other languages. This section introduces multilingual corpora. The term multilingual is used here in a broad sense to include bilingual corpora. Multilingual corpora can be parallel or comparable. Corpora of this kind are particularly useful in translation and con-trastive studies.

11.1. The Canadian Hansard Corpus

The earliest and perhaps best-known parallel corpus is the Canadian Hansard Corpus, which consists of debates from the Canadian Parliament published in the country's official languages, English and French. While its content is limited to legislative discourse, the corpus covers a broad range of topics and styles, e. g. spontaneous discussion, written correspondence, as well as prepared speeches.

There are several versions of the Canadian Hansard parallel corpus. The USC version comprises 1.3 million pairs of aligned text chunks (i. e. sentences or smaller fragments) from the official records (*Hansards*) of the 36th Canadian Parliament (1997−2000) with ca. 2 million words in English and French each. This version is freely downloadable at the USC site (USC Hansard, see appendix). TransSearch (see appendix) offers an online service which allows subscribed users to access all of the Hansard texts from 1986 to February 2003 (approximately 235 million words). The LDC released a collection of Hansard parallel texts in 1995, covering a time span from the mid-1970s through 1988. This version is available on CD-ROM from the LDC. The Canadian Hansard Treebank contains 750,000 words of skeleton-parsed texts from proceedings in the Canadian Parliament, which is available from UCREL of Lancaster University.

11.2. The English-Norwegian Parallel Corpus

The English-Norwegian Parallel Corpus (ENPC) is one of the earliest and best-known parallel corpora. The corpus is bi-directional in that it contains both original and translated texts in the two languages. ENPC consists of 100 original texts between 10,000 to 15,000 words in length in English and Norwegian together with their corresponding translations in the two languages, totaling 2.6 million words. Unlike most parallel corpora which are limited to a particular domain or text type, efforts have been made to balance the ENPC corpus. Both fiction (30 originals plus translations in each language)

and non-fiction (20 originals plus translations in each language) texts are sampled. Fiction texts include children's fiction, detective fiction and general fiction. Non-fiction texts cover religion, social sciences, law, natural sciences, medicine, arts, and geography/history (see Johansson/Ebeling/Oksefjell 2002). ENPC is marked up in TEI-compliant SGML. Both English and Norwegian texts in the corpus are POS tagged and lemmatized. The corpus is aligned at the sentence level. The ENPC corpus is available for non-commercial research. Registered users can access the corpus online. See the corpus homepage (ENPC, see appendix) for details on registration.

11.3. The English-Swedish Parallel Corpus

The English-Swedish Parallel Corpus (ESPC) follows ENPC in its design. The corpus consists of 64 English text samples and their translations into Swedish and 72 Swedish text samples and their translations into English, amounting to 2.8 million words. The samples from each language have been drawn from two main text categories, fiction and non-fiction. The fiction categories include children's fiction, crime and mystery fiction, and general fiction, while non-fiction texts cover memoirs and biography, geography, humanities, natural sciences, social sciences, applied sciences, legal documents, and prepared speech. The text types of the originals from both languages are comparable in terms of genre, subject matter, type of audience and register (cf. Altenberg/Aijmer/Svensson 2001). ESPC is aligned at the sentence level and marked up in TEI-compliant SGML. The corpus is for non-commercial research and only registered users can access it. See the ESPC site (see appendix) for contact details.

11.4. The Oslo Multilingual Corpus

The Oslo Multilingual Corpus (OMC) is an extension of ENPC which covers more languages including, in addition to English and Norwegian, also German, French, Swedish, Dutch, Finnish and Portuguese. The corpus is composed of many subcorpora that differ in composition with regard to languages and number of texts included. Apart from ENPC and ESPC, the corpus currently includes a French-Norwegian subcorpus Corpus (FNPC, ca. 0.86 million words), a German-Norwegian subcorpus (GNPC, ca. 1.3 million words), an English-German-English subcorpus (En-Ge-En, 1.5 million words), a German-Norwegian-German subcorpus (Ge-No-Ge, 1.8 million words), a Norwegian-English-German subcorpus (En-Ge-No, 289,230 words of Norwegian original texts, 432,500 words of English original texts, and 287,400 words of German original texts, plus the translations in the other two languages), an English-Dutch subcorpus (En-Du, 0.3 million words), an English-Norwegian-Portuguese subcorpus (En-No-Po, 0.6 million words), a Norwegian-French-German subcorpus (No-Fr-Ge, 1.5 million words), a Norwegian-English-French-German subcorpus (No-En-Fr-Ge, 1.7 million words), and an English-Finnish subcorpus (0.3 million words).

OMC has been constructed following the same principles as ENPC; and like ENPC, the corpus is coded and marked up in TEI-compliant SGML. The OMC corpus is for academic, non-commercial purposes but it can be accessed only by registered users. See the OMC homepage (see appendix) for the current status of the corpus.

11.5. The ET10/63 and ITU/CRATER parallel corpora

ET10/63 is a bilingual parallel corpus of English and French, containing ca. one million words of EC official documents on telecommunications in each language. The corpus is POS tagged and also lemmatized. This bilingual parallel corpus has been extended to include Spanish on the Corpus Resources and Terminology Extraction project. The extension is thus named the CRATER parallel corpus, which contains one million words in each of the three languages. The corpus is sentence aligned and tagged with part-of-speech in all three languages (cf. Garside et al. 1994). An expanded version of the CRATER corpus, CRATER 2, has increased the size of the English and French components of the parallel corpus from one million to 1.5 million words. Both versions of CRATER are available via ELRA. The corpus can also been accessed online or downloaded via FTP at the CRATER site (see appendix).

11.6. The IJS-ELAN Slovene-English Parallel Corpus

The Slovene-English Parallel Corpus (IJS-ELAN) contains one million words from 15 terminology-rich bilingual texts produced in the 1990s. One half of the corpus (in terms of text size) consists of 11 Slovene texts and their English translations while the other half comprises four English texts and their Slovene translations. The corpus is aligned at the sentence level (cf. Erjavec 2002). Two versions of the IJS-ELAN corpus are available, with one version marked up in SGML/TEI P3 and the other encoded in XML/TEI P4 and lemmatized and POS tagged. Both versions are freely available for downloading at the corpus website (see appendix), which also allows free online access.

11.7. JRC-ACQUIS Multilingual Parallel Corpus

The JRC-ACQUIS Multilingual Parallel Corpus is a truly multilingual corpus, covering 22 European languages: Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Maltese, Dutch, Polish, Portuguese, Romanian, Slovak, Slovene, and Swedish. The current version 3.0 contains 463,792 texts of EU legislation between the 1950s and 2006, totaling over one billion words. The current release contains pairwise alignment for 231 language pairs. It has also made some corrections in the Bulgarian subcorpus. The corpus is marked up in XML/TEI P4 and currently aligned at the paragraph level. The corpus is distributed for research purposes and can be downloaded at the corpus website (see ACQUIS in appendix).

11.8. The CLUVI parallel corpus

The CLUVI (Linguistic Corpus of the University of Vigo) parallel corpus is an open textual corpus of specialized registers (taken from fiction, computing, journalism and legal and administrative fields), totaling eight million words of running texts. The corpus

currently comprises seven main sections. They are the LEGA parallel corpus of Galician-Spanish legal texts (6.33 million words), the UNESCO parallel corpus of English-Galician-French-Spanish scientific-technical divulgation (3.72 million words), the LOGALIZA corpus of English-Galician software localization (1.98 million words), the TECTRA parallel corpus of English-Galician literary texts (1.47 million words), the FEGA Corpus of French-Galician literary texts (1.27 million words), the CONSUMER corpus of Spanish-Galician-Catalan-Basque consumer information (5.59 million words), and the LEGE-BI corpus of Basque-Spanish legal texts (2.38 million words). The corpus is being expanded with six additional sections: Galician-Spanish economy texts, English-Portuguese literary texts, English-Spanish literary texts, German-Galician literary texts, English-Galician film subtitling, and Portuguese-Spanish postcolonial literature (cf. Gómez Guinovart/Sacau Fontenla 2004). The completed sections of the corpus are freely accessible at the CLUVI website (see appendix), which permits both simple and very complex searches of isolated words or sequences of words.

11.9. European Corpus Initiative Multilingual Corpus I

European Corpus Initiative Multilingual Corpus I (ECI/MCI) was released in 1994 by ELSNET (see section 13). The corpus contains 98 million words of texts from 27 languages, covering most of the major European languages as well as some non-European languages such as Chinese, Japanese and Malay. The corpus has 48 components, 12 of which are parallel corpora composed of 2−9 subcorpora. It also includes a great diversity of text types such as newspapers, novels and stories, technical papers and dictionaries and wordlists, though most components are quite homogeneous in contents (cf. Armstrong-Warwick et al. 1994).

ECI/MCI is marked up in TEI P2 conformant SGML, but the markup has been undertaken in such a way that users can also get easy access to the source text without markup. The corpus is available from ELSNET (see appendix) or the LDC.

11.10. The MULTEXT corpora

Multilingual Tools and Corpora (MULTEXT) is a series of projects whose aims are to develop standards and specifications for the encoding and processing of linguistic corpora, and to develop tools, corpora and linguistic resources embodying these standards. The multilingual corpus used for developing linguistic tools is the JOC (*Official Journal of European Community*) corpus, which comprises 40 files in five languages: English, German, Italian, Spanish and French. Of these, ten files in five languages (English, French, German, Spanish and Italian) are POS tagged and 10 files in four language pairs (English-French, English-German, English-Italian and English-Spanish) are aligned at the sentence level. The corpus is conformant with the Corpus Encoding Standard (CES, see article 22). The availability of the corpus is unknown, but some samples can be downloaded at the MULTEXT website (see appendix).

MULTEXT-East is a project which is intended to extend the scope of MULTEXT by transferring MULTEXT's expertise, methodologies, and tools to Central and Eastern European countries, thus enabling the extension and validation of these methodologies

and tools on a new range of languages. The latest release of MULTEXT-East resources, version 3, became available in July 2004. It is marked up in XML/TEI P4.

The MULTEXT-East dataset has four components: morpho-syntactic lexica, a parallel corpus, a spoken corpus, and a comparable corpus. The parallel corpus consists of the English original of George Orwell's *Nineteen Eighty-Four* (100,000 words) together with its translations into the nine project languages: Bulgarian, Czech, Estonian, Hungarian, Lithuanian, Romanian, Russian, Serbian, and Slovene. The translations of *Nineteen Eighty-Four* are POS tagged manually and sentence aligned with the English original, with tagging and alignment validated by hand. The spoken corpus, which covers seven languages (Romanian, Slovene, Estonian, Hungarian, English, Czech, Bulgarian), is composed of the translations (from English) of forty short passages of five thematically connected sentences. For the first four languages in the above list, the texts have also been read, recorded and included in the distribution. The MULTEXT-East multilingual comparable corpus comprises a fiction subset and a news subset of at least 100,000 words each, for each of the six project languages (Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene). Each language component is comparable in terms of the number and size of texts. The multilingual comparable corpus is marked up in CES format with over 40 different elements (see Erjavec 2004). The comparable spoken corpus (transcripts and audio files) is freely downloadable, while the parallel and comparable corpora, together with other MULTEXT-East language resources, are subject to license and restricted to research use only. Licensed users can browse or download full resources. Registrations can be made on the MULTEXT-East website (see appendix).

11.11. The PAROLE corpora

PAROLE (Preparatory Action for Linguistic Resources Organization for Language Engineering) represents a large-scale harmonized effort to create comparable text corpora and lexica for EU languages. Fourteen languages are involved on the PAROLE project: Belgian French, Catalan, Danish, Dutch, English, French, Finnish, German, Greek, Irish, Italian, Norwegian, Portuguese and Swedish. Corpora containing 20 million words and lexica containing 20,000 entries were constructed for each of these languages using the same design and composition principles during 1996–1998. These corpora all include specific proportions of texts from the categories book (20%), newspaper (65%), periodical (5%) and miscellaneous (10%) within a settled range.

The PAROLE corpora are marked up according to CES-conformant PAROLE DTD (Document Type Declaration). An equal proportion of the texts (up to 250,000 running words) in each PAROLE corpus was POS tagged according to a common PAROLE tagset and morpho-syntactic annotation standards. Part of the tagged data was validated: 50,000 words checked for maximum granularity and 200,000 for part-of-speech. For some PAROLE corpora, only a copyright-free subset is available to the public. The PAROLE corpora that are currently available are distributed by ELRA.

11.12. Multilingual Corpora for Cooperation

Multilingual Corpora for Cooperation (MLCC) is a corpus acquisition project which aims to collect a set of texts representing a substantial improvement in range, quantity

and quality of corpus material available. The MLCC multilingual data consists of the Multilingual Parallel Corpus and the comparable Polylingual Document Collection. The parallel corpus comprises translated data in nine European languages: Danish, Dutch, English, French, German, Greek, Italian, Portuguese and Spanish. This corpus has two datasets, with one set taken from the *Official Journal of the European Commission*, C Series: Written Questions 1993, totaling approximately 10.2 million words (1.1 million words per language), and the other set taken from the *Official Journal of the European Commission*, Annex: Debates of the European Parliament 1992−1994, with 5−8 million words for each language. The comparable corpus includes financial newspaper articles from the early 1990s in six European languages: Dutch (8.5 million words), English (30 million words), French (10 million words), German (33 million words), Italian (1.88 million words), and Spanish (10 million words). The MLCC multilingual and parallel corpora are marked up in TEI-compliant SGML (cf. Armstrong et al. 1998). The resources are available via ELRA.

We have so far introduced multilingual corpora of European languages. The following sections are concerned with corpora involving other languages.

11.13. The EMILLE Corpus

The EMILLE Corpus is a product of the Enabling Minority Language Engineering project which develops language resources for South Asian languages. Two versions of the EMILLE Corpus are available: the EMILLE/CIIL Corpus distributed free of charge for non-commercial research, and the EMILLE/Lancaster Corpus for commercial use only.

The EMILLE/CIIL Corpus consists of three components: monolingual, parallel and annotated corpora. There are fourteen monolingual corpora, including both written and (for some languages) spoken data for fourteen South Asian languages: Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sinhala, Tamil, Telegu and Urdu. The EMILLE monolingual corpora contain approximately 92,799,000 words (including 2,627,000 words of transcribed spoken data for Bengali, Gujarati, Hindi, Punjabi and Urdu). The parallel corpus consists of 200,000 words of text in English and its accompanying translations in Hindi, Bengali, Punjabi, Gujarati and Urdu. The annotated component includes the Urdu monolingual and parallel corpora annotated for part-of-speech, together with twenty written Hindi corpus files annotated to show the nature of demonstrative use. The EMILLE/Lancaster Corpus consists of three components: monolingual, parallel and annotated corpora. This version differs from the EMILLE/CIIL Corpus in its monolingual component, which consists of monolingual corpora covering seven South Asian languages (Bengali, Gujarati, Hindi, Punjabi, Sinhala, Tamil, and Urdu), totaling approximately 58,880,000 words (including 2,627,000 words of transcribed spoken data for Bengali, Gujarati, Hindi, Punjabi and Urdu). The parallel and annotated components are the same as in the EMILLE/CIIL Corpus (cf. Baker et al. 2004).

The EMILLE Corpus is marked up using CES-compliant SGML, and encoded using Unicode. More information about the corpus is available on the EMILLE website (see appendix). Both versions of the corpus are distributed via ELRA.

11.14. The BFSU Chinese-English Parallel Corpus

The BFSU (Beijing Foreign Studies University) Chinese-English Parallel Corpus contains 30 million words. Presently it is the largest parallel corpus of English and Chinese. The corpus is composed of four subcorpora, i. e. Balanced Corpus, Translation Corpus, Bilingual Sentences Corpus and Corpus for Specific Purpose. The bidirectional parallel corpus includes both literary (fiction, prose and play scripts) and non-literary texts, which are sampled from 12 text categories covering three major domains: humanities, social sciences and natural sciences. The Chinese-English and English-Chinese texts account for 40% and 60% respectively while literary and non-literary texts account for 55% and 45% respectively. The BFSU parallel corpus is automatically sentence aligned and hand validated. It has been annotated in such a way as to allow concordances of words, phrases, collocations, and sentence patterns (cf. Wang 2004). The corpus is available from the China National Research Centre for Foreign Language Education (see Sinotefl in appendix).

11.15. The Babel Chinese-English parallel corpora

The PKU Babel Chinese-English Parallel Corpus hosted at Peking University contains 20 million Chinese characters and 10 million English words of bilingual texts sampled from a great variety of text categories including government documents, news, academic prose, fiction, play scripts, and speech, among other text categories. It is designed as a balanced corpus covering three styles (literature, practical writing and news), six fields (arts, business/economics, politics, science, sports, and society/culture), two modes (written, spoken), and four periods (ancient, early modern, modern, and contemporary for Chinese texts, and Old English, Middle English, Early Modern English and present-day English for English texts). Presently only contemporary/present-day written texts are included, and about 400,000 sentence pairs have been aligned (cf. Bai/Chang/Zhan 2002). The Babel parallel corpus is marked up in XML. Each document has two parts, the text header and the text body. The header part shows Chinese and English titles, author, translator, style, field, mode and period. The text body is annotated for paragraphs, aligned anchoring points, sentences, and words. The Chinese texts in the corpus are tokenized and POS tagged while the English texts are POS tagged and lemmatized. The completed part of the corpus can be accessed online at the PKU Babel website (see appendix).

The Babel English-Chinese parallel corpus hosted at Lancaster University consists of 327 English articles and their translations in Mandarin Chinese collected from two online bilingual magazines in 2000 and 2001, totaling half a million words. The corpus is tagged for part-of-speech for both English and Chinese and is aligned at the sentence level. It is also marked for paragraph and sentence boundaries. The corpus can be accessed online at the Lancaster Babel website.

11.16. Hong Kong Parallel Text

Hong Kong Parallel Text is a large parallel corpus released by the LDC in 2004. The corpus contains approximately 59 million English words and 49 million Chinese words

(or 98 million Chinese characters). It consists of the updates of three parallel corpora published in 2000: Hong Kong Hansards, Hong Kong Laws, and Hong Kong News. The Hong Kong Hansards component contains excerpts from the Official Record of Proceedings of the Legislative Council of the HKSAR from October 1985 to April 2003, totaling 36,140,737 English words and 56,618,181 Chinese characters. The Hong Kong Laws component contains statute laws of Hong Kong in English and Chinese, constitutional instruments, national laws and other relevant instruments published by the Department of Justice of the HKSAR up to the year 2000, amounting to 8,396,243 English words and 14,868,621 Chinese characters. The Hong Kong News component contains press releases from the Information Services Department of the HKSAR between July 1997 and October 2003, amounting to 14,798,671 English words and 26,677,514 Chinese characters. All of the three components in the Hong Kong Parallel Text corpus are aligned at the sentence level. The English and Chinese texts are kept in separate files, with alignment indicated by corresponding sentence numbers. The corpus is available from the LDC.

11.17. The OPUS parallel corpus

OPUS is a publicly available, open-source parallel corpus which consists of translated texts collected from the Web. It covers not only European languages but also Asian languages such as Chinese and Japanese. The corpus is constantly growing with fresh data. The current version has seven components.

The OpenSubtitles corpus comprises 361 bitexts in 30 languages, amounting to 23.4 million tokens in 20,722 files. The subcorpus of European constitution is composed of 210 aligned bitexts in 21 European languages, totaling 987 files and three million tokens. The OpenOffice corpus consists of 2,014 documents in English original (about half a million words) and their partial translations into five languages (French, Spanish, Swedish, German, and Japanese), totaling 10,983 files and 2.6 million tokens. All documents in this subcorpus are tokenized and, except for the Spanish part, tagged with part-of-speech, while the English part is marked with syntactic chunks as well. The subcorpus of KDE system messages consists of 1,830 bitexts in 61 languages, totaling 24,586 files and 20 million tokens. The KDE manual corpus has 226 bitexts in 24 languages with a total of 3.8 million words in 3,736 files. The EUROPARL subcorpus comprises 55 bitexts of European Parliament Proceedings (1996−2003) in 11 languages, amounting to 296 million tokens in 5,214 files. Finally, the PHP manual corpus consists of 231 bitexts in 22 languages, totaling 3.3 million tokens in 71,518 files.

This OPUS parallel corpus is marked up in XML. All components of the corpus, as well as some corpus building tools, can be downloaded at the OPUS site (see appendix), which also provides a multilingual and a bilingual query interface.

12. Non-English monolingual corpora

We have so far been concerned with well-known and influential English corpora and multilingual corpora involving English, in addition to some national corpora. This section introduces a number of major monolingual corpora of other languages.

12.1. The COSMAS corpora

COSMAS (Corpus Search, Management and Analysis System) is a large collection of German text corpora developed at the Mannheim IDS (Institut für deutsche Sprache). With a size of almost two billion words, this is the world's largest, ever-growing collection of German online corpora for linguistic research. The collection covers a wide variety of sources, e. g. classic literary texts, national and regional newspapers, transcribed spoken language, morpho-syntactically annotated texts and several unique corpora.

The copyright free part of the COSMAS collection (over 1.1 billion words) is publicly available free of charge for searching via the COSMAS online toolbox (see appendix), which allows complex queries, collocation analysis, clustering, and virtual corpus composition, etc. The COSMAS corpora are only available for non-commercial use and anonymous COSMAS sessions are limited to 60 minutes.

12.2. The CETEMPúblico Corpus

The CETEMPublico (Corpus de Extractos de Textos Electronicos MCT/Pμblico) corpus includes the text of around 2,600 editions of the Portuguese daily newspaper *Público*, written between 1991 and 1999, amounting to approximately 180 million words. The corpus is marked up in SGML. Having removed some repeated extracts from version 1.0, CETEMPublico version 1.7 consists of over 1.5 million extracts. The first million words (8,043 extracts) have been parsed. This subset represents a balanced selection from the whole period (1991−1999) rather than early years alone. It also covers all of the categories included in the full corpus (cf. Santos/Rocha 2001). CETEMPublico can be used for research and technological development, but direct commercial exploitation is not permitted. There are a number of ways to access the corpus: CD-ROM from the LDC, FTP download, and online access at the corpus website (see appendix).

12.3. The INL corpora

The Institute for Dutch Lexicology (INL) has offered three corpora over the Web. The Five Million Words Corpus 1994 has diversified compositions. It comprises texts of present-day Dutch derived from 17 text sources dating from 1989−1994, including books, magazines, newspapers and TV broadcasts which cover topics such as journalism, politics, environment, linguistics, leisure and business/employment (see Kruyt 1995). The 27 Million Words Dutch Newspaper Corpus 1995 consists of newspaper texts derived from issues published in 1994−1995 by a major national newspaper, NRC (see Kruyt et al. 1996). The 38 Million Words Corpus 1996 has three main components: a component with varied composition (books, magazines, newspaper texts, TV broadcasts, parliamentary reports, 1970−1995, 12.7 million words), a newspaper component (*Meppeler Courant*, 1992−1995, 12.4 million words), and a legal component (Dutch legal texts operative in 1989, with some dating back as early as 1814, 12.9 million words) (see Kruyt/Dutilh 1997). All three corpora are lemmatized and tagged for part-of-speech and users can define subcorpora using the parameters encoded therein. They are available for non-

commercial research purposes only. Access to these corpora is free of charge but subject to an individual user agreement, which can be obtained from the INL website (see appendix).

12.4. The CEG corpus

The CEG (Cronfa Electroneg o Gymraeg) corpus contains one million words of written Welsh prose. The corpus is designed as a Welsh parallel to the Brown and LOB corpora, consisting of five hundred 2,000-word samples selected from a representative range of text types to illustrate modern (mainly post 1970) Welsh prose writing. However, the text categories and their proportions in the corpus are different from those in Brown and LOB. The texts in CEG are grouped into two broad categories: factual prose and fiction. There are seven types of fiction such as novels and short stories, while the factual prose is further divided into 22 categories such as various types of press material, administrative documents, academic texts and biography (see Ellis et al. 2001).

The corpus is of value for lexical and syntactic analyses of modern Welsh prose. It is available as both raw and annotated texts. Annotations include lemmatization and POS tagging. Both versions are available at the CEG website (see appendix).

12.5. The Scottish Corpus of Texts and Speech

The Scottish Corpus of Texts and Speech (SCOTS) is an ongoing project which aims to build a large electronic corpus of both written and spoken texts for the languages of Scotland, aiming to cover the period from 1945 to the present day (with most spoken texts recorded since 2000). Currently the corpus (SCOTS Dataset 12) consists of 1,175 documents (4,024,343 words), which include written, spoken and visual materials from a range of genres such as conversation, interviews, correspondence, poetry, fiction and prose. The corpus includes samples from Scots and Scottish English, in addition to a small number of Scottish Gaelic texts. Great efforts have been made on the SCOTS project to render the corpus as balanced as practically possible, by including a wide range of texts of different language varieties, genres and registers, and including speakers and writers from a wide range of geographical locations, backgrounds, age and gender groups as well as occupations, etc. However, SCOTS does not claim to be a truly representative corpus because some genres (e. g. newspaper articles, personal diaries, business correspondence) are not covered owing to practical issues such as permissions, copyright and text availability.

The SCOTS corpus is marked up in SGML. The extensive sociolinguistic metadata includes, for example, resource type, text type, setting, medium, audience, text details, author/speaker details (gender, age, geographic region, education, occupation, religious background, languages used, etc.), and copyright information (see SCOTS in appendix). The current version of the SCOTS corpus is not linguistically annotated, but the transcripts of spoken data are aligned with digital audio/video recordings. The available texts can be searched at the SCOTS project site (see appendix).

12.6. The FIDA Corpus of Slovenian

FIDA is a monolingual reference corpus of the Slovene language which contains just over 100 million words of contemporary Slovene texts. The corpus covers a broad range of Slovene language variants and registers as found in the Slovene press, complemented by some texts from the Internet and speech transcripts. Both literary (poetry, prose and drama, 5.9%) and non-literary (scientific texts in both natural and social sciences as well as non-scientific writing, 94.1%) texts are included. In terms of media, newspapers (46.6%), journals (23.9%) and books (22.7%) account for over 93.2% of the whole corpus.

FIDA can be searched via its web-based interface at the corpus site (see FIDA in appendix), but the access is not free of charge. Only users with a valid account can access the corpus, though a guest account is also available for users to test-run the query system.

12.7. The Nova Beseda corpus of Slovenian

Nova Beseda is another large collection of Slovenian texts which started with web presentation of a three-million-word electronic collection of Slovenian fiction in 1999. In the years that followed the corpus grew in both size and diversity so that it has increased to 240 million words, with most texts taken from publications from the 1990s. All of the texts are marked on the sentence level.

Tab. 20.28: Structure of the Nova Beseda corpus

Part	Contents	Words (million)	Proportion
A	Fiction in Slovenian	12	5%
B	Non-fiction in Slovenian	2	0.8%
C	Scientific and technical publications	3	1.3%
D	Delo Slovenian daily (1998–2007)	169	70.4%
G	Slovenian National Assembly session transcripts (1996–2007)	31	12.9%
P	Delo FT, Jana, Mladina, Monitor, National Geographic, Viva magazines	21	8.8%
S	Republic of Slovenia Legislation	2	0.8%
	Total	240	

The corpus has seven parts as shown in Table 20.28. As can be seen, newspaper texts account for 70% of the total number of tokens. The corpus can be searched online freely at the Nova Beseda website (see appendix).

12.8. The Prague Dependency Treebank

The Prague Dependency Treebank (PDT, version 2.0) contains two million words of texts drawn from the Czech National Corpus (see section 2.4.) which have been annota-

ted morphologically and syntactically. Of the texts included in the treebank, general newspaper articles related to politics, sports, culture, hobbies, etc. account for 60 %, economic news and analyses 20 %, and popular science magazines 20 %. PDT version 2.0 is marked up in XML. The annotation scheme consists of three levels. The morphological level assigns a lemma and a morphological tag to each token. The analytical level uses dependency grammar to annotate the structure of the parse tree and the analytical function of every node, which determines the relationship between the dependent node and its governing node one level higher in the tree. The highest level of annotation, the tectogrammatical level, uses the dependency framework to describe the linguistic meaning of a sentence (see Böhmová et al. 2003). The third level annotation has been added in PDT version 2.0. The same texts are annotated on all three levels, but the amount of annotated material decreases with the complexity of the levels, specifically about two million tokens on the morphological level, about 1.5 million tokens at the analytical level, and 0.8 million tokens on the tectogrammatical level (see Hajič 2004). The Prague Dependency Treebank version 2.0 is available on CD-ROM from the LDC. It can also be accessed at the PDT website (see appendix) using an online tool which allows users to search for and view parse trees.

12.9. The Sinica corpora of Chinese

The Academia Sinica Balanced Corpus (ASBC) is the first annotated corpus of modern Chinese. The corpus is a representative sample of Mandarin Chinese as used in Taiwan. The current version (3.1) of the corpus contains five million words of texts sampled from different areas and classified according to five criteria: genre, style, mode, topic, and source. Table 20.29 (cf. Huang/Chen 1995/1998) shows the proportions of texts and categories in terms of these criteria.

The values of these parameters, together with bibliographic information, are encoded at the beginning of each text in the corpus. The whole corpus is tagged for part-of-

Tab. 20.29: Composition of ASBC (version 3.1)

Criterion	Proportions
Genre	Press reportage: 56.25 %, Press review: 10.01 %, Advert: 0.59 %, Letter: 1.29 %, Fiction: 10.12 %, Essay: 8.48 %, Biography and diary: 0.50 %, Poetry: 0.29 %, Quotes: 0.03 %, Manual: 2.03 %, Play script: 0.05 %, Public speech: 8.19 %, Conversation: 1.34 %, Meeting minutes: 0.11 %
Style	Narrative texts: 70.66 %, Argumentative texts: 12.24 %, Expository texts: 14.72 %, Descriptive texts: 2.83 %
Mode	Written: 90.14 %, Written-to-be-read: 1.38 %, Written-to-be-spoken: 0.82 %, Spoken: 7.29 %, Spoken-to-be-read: 0.35 %
Topic	Philosophy: 8.68 %, Natural science: 12.97 %, Social science: 34.99 %, Arts: 9.28 %, General/leisure: 17.89 %, Literature: 16.20 %
Source	Newspaper: 31.28 %, General magazine: 29.18 %, Academic journal: 0.70 %, Textbook: 4.08 %, Reference book: 0.13 %, Thesis: 1.36 %, General book: 8.45 %, Audio/video medium: 22.83 %, Conversation/interview: 1.63 %, Public speech: 0.25 %

speech and a range of linguistic features such as nominalization and reduplication. The Sinica corpus is accessible online at the ASBC website (see appendix) using the query system which also allows users to define subcorpora.

A new version with a target size of 10 million words has been completed and is expected to become available soon (cf. Huang 2006). The texts in the new release have been sampled mainly from 1996 onwards, in a balanced way in terms of five criteria: genre, style, mode, topic, and source. In this version, the proportions for each topic are as follows: philosophy (10%), science (10%), society (35%), art (5%), life (10%), and literature (20%). As in the current version, the new release is tokenized and POS tagged.

The Academia Sinica Tagged Corpus of Early Mandarin Chinese is another Chinese corpus built by the Academia Sinica. This corpus is divided into three subcorpora according to stages of grammatical developments: Old Chinese (from Pre-Qin to Pre-Han, 5,128,068 characters), Middle Chinese (from Late Han to the Six Dynasties, 8,101,662 characters), and Early Mandarin Chinese (from Tang to Qing, 4,406,381 characters). Presently most parts of the subcorpora for Old Chinese and Early Modern Chinese have been tokenised and POS tagged. The corpus is accessible using the online query system at the corpus website (see Early Mandarin in appendix), which permits keyword searching, statistics, and collocation analysis.

12.10. The Sinica Treebank

The Sinica Treebank (version 3.0) contains 361,834 words extracted from the ASBC corpus, covering subject areas such as politics, travel, sports, finance and society. There are 61,087 structural trees in the treebank. Like the Prague Dependency Treebank, the thematic relation between a predicate and an argument is marked in addition to grammatical categories in the Sinica Treebank. Six non-terminal phrasal categories are annotated in the treebank: S (a complete tree headed by a predicate), VP (a verb phrase headed by a predicate), NP (a noun phrase headed by a noun), GP (a phrase headed by locational noun or locational adjunct), PP (a prepositional phrase headed by a preposition), and XP (a conjunctive phrase that is headed by a conjunction). There are three different kinds of grammatical heads: Head, head and DUMMY. Head indicates a grammatical head in a phrasal category; head indicates a semantic head which does not simultaneously function as a syntactic head; and DUMMY indicates the semantic head(s) whose categorical or thematic identity cannot be locally determined. A total of 63 thematic roles are annotated in the treebank including, for example, agent, causer, condition and instrument for verbs, and time and location for nouns (see Huang et al. 2000). The Sinica Treebank can be accessed online (see appendix) using the web-based interface which allows users to search the treebank and view diagrammatical parse trees. A sample of 1,000 syntactic tree structures is available for free download.

12.11. The Penn Chinese Treebank

The current version (version 6.0) of the Penn Chinese Treebank (CTB) consists of 780,000 words (over 1.28 million Chinese characters) that are segmented, part-of-speech tagged and fully bracketed. A total of 2,036 text files are included in this release, covering

newswire texts from Xinhua News Agency, articles from Sinorama Magazine, news stories from the website of the Hong Kong Special Administrative Region, and transcripts from various news broadcast programs.

The annotation format of CTB follows that of the Penn English treebank. The formal structural properties are represented with structural labels (such as NP, VP) in brackets while the functional properties are represented with functional labels such as -ADV, -TMP, and -SBJ. Six main grammatical relations are represented in the Chinese treebank, with complementation, adjunction and coordination represented structurally, while predication, modification and apposition are represented non-configurationally (see Xue et al. 2005). There are 28,295 parsed sentences in the treebank. The corpus is available from the LDC.

12.12. The Spoken Chinese Corpus of Situated Discourse

The Spoken Chinese Corpus of Situated Discourse (SCCSD) is an ongoing project under the auspices of the Chinese Academy of Social Science which aims to collect 1,000 hours of recordings of Mandarin Chinese spoken in China. The corpus consists of three subcorpora, one for workshop discourse, one for major dialects in China, and one for speeches. At present, 600 hours of audio and 50 hours of video recordings have been collected. The sampling frame for the societal discourse was established sociologically on the basis of a yellow book while the familial discourse was defined in terms of habitation and occupation, as shown in Table 20.30 (cf. Gu 2002).

Tab. 20.30: Discourse types in SCCSD

Category	Subcategory	Example
Societal	Major activities of organization	government and political discourse, business discourse, educational and academic discourse, legal and mediatory discourse, mass media discourse, discourse of medicine and health, discourse of sports, public service discourse, public welfare discourse, religious and superstitious discourse
	Activities common to organization	administrative discourse, banquet discourse, discourse of celebration and ceremony, discourse of entertainment and leisure, office discourse, political study discourse, telephone discourse
	Special discourse	pathological discourse, criminal discourse, military discourse, miscellaneous
Familial discourse	Family discourse in a metropolis	family of high-ranking officials, family of entrepreneurs, family of businessmen, family of academics, family of white collar, family of blue collar, family of suburb farmers, family of immigrant labor
	Family discourse in a small town	family of academics, family of white collar

The corpus is presently being transcribed and annotated, with segmented audio/video chunks linked to the corresponding transcripts. When the corpus is completed, about 50−100 hours will be mounted at the SCCSD website (see appendix) and made available on the Internet in a multimedia form.

12.13. The PKU Chinese corpora

The PKU-CCL-Corpus has two components, one for Modern Chinese and the other for Ancient Chinese. The Modern Chinese subcorpus has reached a size of 264 million Chinese characters, covering a variety of genres such as newspapers, magazines, literary texts, applied writing, and speeches, while the ancient Chinese subcorpus comprises 84 million Chinese characters of running texts sampled from different historical periods from Old Chinese to Early Modern Chinese. The texts in the corpus are not tokenized or POS tagged, but basic bibliographic information such as title and author is provided. Both Modern and Ancient Chinese components of the corpus can be searched via the online query system at the corpus website (see PKU-CCL-Corpus in appendix).

In addition to the unannotated Chinese corpus introduced above, Peking University has also been developing a Chinese treebank PKU-CTB, with a target size of one million words of Chinese texts with syntactic bracketing. The corpus consists of four parts: (1) Chinese government white papers, (2) newspaper articles, (3) Chinese textbooks of primary/middle/high schools, and (4) test sentences used for machine translation evaluation. At present, a total of 271,460 words in 19,560 sentences have been collected, which are parsed into 207,539 phrases tagged with 22 phrases categories. PKU-CTB differs from the Penn Chinese Treebank in that their parsing schemes contain different numbers and types of phrase labels, and more importantly, different approaches to phrase bracketing have been used. Phrase bracketing in Penn Chinese Treebank is based on generative grammar whereas PKU-CTB is bracketed under the paradigm of traditional structuralism, especially the method of immediate constituent analysis (cf. Zhan et al. 2006). As the corpus is still under construction, its availability is unknown at the time of writing.

13. Well-known distributors of corpus resources

While many corpora introduced in this article are made available at individual project or corpus websites, there are a number of organizations which aim at creating, collecting and distributing corpus resources. The best-known of these include CSLU, ELRA/ELD A, ELSNET, ENABLER, ICAME, the LDC, OTA, and TELRI/TRACTOR.

CSLU (Centre for Spoken Language Understanding) is a research centre at the Oregon Graduate Institute of Science and Technology (OGI) that focuses on spoken language technologies. The centre offers a range of products and services. For non-commercial purposes (educational, research, personal and evaluation), most products are freely available. Some products (generally source codes) are also available for commercial use via a membership agreement. CSLU has created, collected and distributed telephone and cellular speech data in over 20 languages for use in the area of voice processing. A

description of the corpora currently available from the centre is available at the CSLU website (see appendix).

ELRA (The European Language Resources Association) is a non-profit organization established in 1995 with the goal of promoting the creation, validation, standardization, and distribution of language resources (LRs) for the Human Language Technology (HLT) community, and evaluating language engineering technologies. Many of these tasks are carried out by ELRA's operational body ELDA (Evaluations and Language Resources Distribution Agency), which is set up to identify, classify, collect, validate and produce language resources. The language resources available from ELRA are classified into four major categories: spoken LRs (telephone/microphone recordings, speech related resources), written LRs (corpora, monolingual and multilingual lexicons), terminological resources (monolingual, bilingual and multilingual), and multimodal/multimedia LRs. See the ELRA catalogue for the available language resources.

ELSNET (European Network in Language and Speech) is a Europe-based forum which aims to advance human language technologies in a broad sense in Europe, by bringing together Europe's key players in research, development, integration or deployment in the field of language and speech technology and neighboring areas. See the ELSNET resources page for the corpora made available or supported by ELSNET.

The ENABLER (European National Activities for Basic Language Resources) Network aims at improving cooperation among the national activities which provide language resources for their respective languages. ENABLER has worked in close collaboration with ELSNET to develop the Language Resources Roadmap and the Language Resources Landscape. Resources offered by ENABLER include written, spoken, and multimodal corpora, as well as lexical resources. See the ENABLER catalogue (see appendix) for a list of available corpora.

ICAME (International Computer Archive of Modern and Medieval English) is an international organization of linguists and information scientists working with English corpora. The aim of the organization is to collect and distribute information on English language material available for computer processing, and on linguistic research completed or in progress on the material, to compile an archive of English text corpora in machine-readable form, and to make material available to research institutions. About 20 corpora amounting to 17 million words are currently available on CDs from ICAME.

The LDC (Linguistic Data Consortium) is an open consortium of universities, companies and government research laboratories which creates, collects and distributes speech and text databases, lexicons, and other resources for research and development purposes. The LDC is the largest distributor of corpus resources, but most LDC resources are specialized corpora which are more geared towards language engineering than linguistic analysis. See the LDC catalogue for a list of available corpora.

OTA (Oxford Text Archive) is one of the oldest and best-known electronic text centres in the world. It works closely with members of the Arts and Humanities academic community to collect, catalogue, and preserve high-quality electronic texts for research and teaching. OTA currently distributes more than 2,500 resources in over 25 different languages covering all areas of literary and linguistic studies, which include a great variety of language corpora in addition to electronic editions of works by individual authors, manuscript transcriptions and reference works. See the OTA catalogue for available resources.

TRACTOR is the TELRI (Trans-European Language Resources Infrastructure) Research Archive of Computational Tools and Resources, which aims at collecting, pro-

moting, and making available monolingual and multilingual language resources and tools for the extraction of language data and linguistic knowledge, with a special focus on Central and Eastern European languages. The TRACTOR archive features monolingual and multilingual corpora as well as lexicons in a wide variety of languages, currently including Bulgarian, Croatian, Czech, Dutch, English, Estonian, French, Finnish, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Polish, Romanian, Russian, Serbian, Slovak, Slovene, Swedish, Turkish, Ukrainian and Uzbek. Resources distributed through TRACTOR (see appendix) are available for non-commercial use only, but TRACTOR aims to promote and foster commercial links between academic and industrial researchers.

14. Conclusion

This article introduced well-known and influential corpora for various research purposes, including national corpora, monitor corpora, corpora of the Brown family, synchronic corpora, diachronic corpora, spoken corpora, academic/professional corpora, parsed corpora, developmental/learner corpora, multilingual corpora, and non-English monolingual corpora. This discussion, however, only covers a very small proportion of the available corpus resources. The classification used in this article was for illustrative purposes only. The distinctions given have been forced by the purpose of this introductory article. It is not unusual to find that any given corpus will be a blend of many of the features introduced here.

With a few exceptions, most of the corpora introduced in this article are publicly available, either free of charge or at an affordable cost. Many of these corpora are searchable or downloadable over the Internet. This article has introduced well-known and influential corpora for English as well as a wide range of European and Asian languages such as French, German, Spanish and Chinese, most of which have been subject to much study. The next article introduces corpus resources for less studied languages.

15. Appendix: URLs

URLs were accessed on 12 February 2008.

ACQUIS: http://wt.jrc.it/lt/Acquis/
Aix-MARSEC: http://aune.lpl.univ-aix.fr/~EPGA/en_marsec.html
ASBC: http://www.sinica.edu.tw/SinicaCorpus/index.html
AWL: http://language.massey.ac.nz/staff/awl/awlinfo.shtml
Bank of English: http://www.collins.co.uk/books.aspx?group= 153
BASE: http://www2.warwick.ac.uk/fac/soc/celte/research/base/
BAWE: http://www2.warwick.ac.uk/fac/soc/celte/research/bawe
BNC Baby: http://www.natcorp.ox.ac.uk/corpus/baby/index.html
BNC Online: http://www.natcorp.ox.ac.uk/using/index.xml.ID=online
BNC PIE: http://pie.usna.edu/
BNC VIEW: http://corpus.byu.edu/bnc/
BNCWeb: http://escorp.unizh.ch/

BNCWeb CQP Edition: http://es-corp.unizh.ch/
BNC XML Edition: http://www.natcorp.ox.ac.uk/XMLedition/
BYU corpora: http://view.byu.edu/
CECL: http://cecl.fltr.ucl.ac.be/
CEG: http://www.bangor.ac.uk/ar/cb/ceg.php.en
CETEMPublico: http://acdc.linguateca.pt/cetempublico/whatisCETEMP.html
CHILDES: http://childes.psy.cmu.edu/
CHRISTINE: http://www.grsampson.net/ChrisDoc.html
CIC: http://www.cambridge.org/elt/corpus/international_corpus.htm
CLC: http://www.cambridge.org/elt/corpus/learner_corpus.htm
CLEC: http://www.clal.org.cn/corpus/EngSearchEngine.aspx
CLUVI: http://sli.uvigo.es/CLUVI/index_en.html
CME: http://www.hti.umich.edu/c/cme/
COLT: http://www.hf.uib.no/i/Engelsk/COLT/
CORIS: http://corpora.dslo.unibo.it/coris_eng.html
COSMAS: http://corpora.ids-mannheim.de/~cosmas/
Corpus del Espanol: http://www.corpusdelespanol.org/
Corpus do Portugues: http://www.corpusdoportugues.org/
CRATER: http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html
CREA: http://corpus.rae.es/creanet.html
Croatian National Corpus: http://www.hnk.ffzg.hr/default_en.htm
CSLU: http://cslu.cse.ogi.edu/corpora/corpCurrent.html
CSPAE: http://www.athel.com/cspa.html
Czech National Corpus: http://ucnk.ff.cuni.cz/english/
DCPSE: http://www.ucl.ac.uk/english-usage/projects/dcpse/index.htm
DOEC: http://www.doe.utoronto.ca/pub/corpus.html
Early Mandarin: http://www.sinica.edu.tw/Early_Mandarin/
ECCO: http://gale.cengage.com/EighteenthCentury/index.htm
EEBO: http://eebo.chadwyck.com/home
ELRA: http://www.elra.info/
ELSNET: http://www.elsnet.org/
EMILLE: http://www.lancs.ac.uk/fass/projects/corpus/emille
ENABLER: http://www.ilsp.gr/enabler/search_sel.asp
ENPC: http://www.hf.uio.no/iba/prosjekt/
ESPC: http://www.englund.lu.se/corpus/corpus/espc.html
FIDA: http://www.fida.net/eng/index.html
FRANTEXT Database: http://www.lib.uchicago.edu/efts/ARTFL/databases/TLF/
French Development Corpus: http://www.flloc.soton.ac.uk/LDC.html
French Progression Corpus: http://www.flloc.soton.ac.uk/ProgC.html
DWDS corpus: http://www.dwds.de/cgi-bin/rest/loginstart
Global English Monitor Corpus: http://www.corpus.bham.ac.uk/ccl/global.htm
Hellenic National Corpus: http://hnc.ilsp.gr/find.asp
Hungarian National Corpus: http://corpus.nytud.hu/mnsz/index_eng.html
ICAME: http://nora.hd.uib.no/icame.html
ICE: http://www.ucl.ac.uk/english-usage/ice/
ICE-GB: http://www.ucl.ac.uk/english-usage/ice-gb/
ICLE: http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm
IJS-ELAN: http://nl.ijs.si/elan/
INL: http://www.inl.nl
IPI PAN Corpus: http://korpus.pl/index.php?lang=en&page=welcome
IViE: http://www.phon.ox.ac.uk/~esther/ivyweb/
JPU: http://joeandco.blogspot.com/

KEMPE: http://corp.hum.sdu.dk/cqp.en.html
KNC: http://www.sejong.or.kr/
Korpus: http://corp.hum.sdu.dk/corpustop.en.html
Korpus 2000: http://korpus.dsl.dk/korpus2000/engelsk_summary.php?lang=uk
L-CIE: http://www.ul.ie/~lcie/homepage.htm
Lancaster Babel: http://www.ling.lancs.ac.uk/corplang/babel/babel.htm
LCMC: http://www.elda.org/catalogue/en/text/W0039.html
LDC: http://www.ldc.upenn.edu/
LDC Online: http://www.ldc.upenn.edu/ldc/online/
Lexical Tutor: http://www.lextutor.ca/
LINDSEI: http://cecl.fltr.ucl.ac.be/Cecl-Projects/Lindsei/lindsei.htm
LIVAC: http://www.livac.org/
Longman Corpus Network: http://www.longman.com/dictionaries/corpus/index.html
Longman Learners' Corpus: http://www.longman.com/dictionaries/corpus/learners.html
LUCY: http://www.grsampson.net/LucyDoc.html
MARSEC: http://www.rdg.ac.uk/AcaDepts/ll/speechlab/marsec/
MCLC: http://www.clr.org.en/retrieval
MEMEM: http://www.hti.umich.edu/m/mememl/
MICASE: http://lw.lsa.umich.edu/eli/micase/index.htm
MICUSP: http://lw.lsa.umich.edu/eli/eli 1/micusp/index.htm
MidEng: http://etext.virginia.edu/mideng.browse.html
MULTEXT: http://www.lpl.univ-aix.fr/projects/multext/MUL4.html
MUTEXT-East (version 3): http://nl.ijs.si/ME/V3/
National Corpus of Irish: http://www.focloir.ie/corpus/
Nova Beseda: http://bos.zrc-sazu.si/a_beseda.html
OMC: http://www.hf.uio.no/ilos/OMC/English/index_e.html
OTA: http://ota.ahds.ac.uk/
OPUS: http://urd.let.rug.nl/tiedeman/OPUS/
PAROLE: http://www.elda.org/catalogue/en/text/doc/parole.html
Parsed historical corpora: http://www-users.york.ac.uk/~lang18/pcorpus.html
PDT (2.0): http://ufal.mff.cuni.cz/pdt2.0/
PELCRA: http://pelcra.ia.uni.lodz.pl/index_en.php
Penn Treebank: http://www.ldc.upenn.edu/ldc/online/treebank/index.html
PERC: http://www.perc21.org/menu.html
PKU-CCL-Corpus: http://ccl.pku.edu.cn/YuLiao_Contents.Asp
PKU Babel: http://www.icl.pku.edu.cn/icl_groups/parallel/default.htm
PPCME: http://www.ling.upenn.edu/hist-corpora/
PWN Corpus of Polish: http://korpus.pwn.pl/szukaj_en.php
RNC: http://www.ruscorpora.ru
Ruscorpora: http://corpus.leeds.ac.uk/ruscorpora.html
SCCSD: http://ling.cass.cn/dangdai/corpus.htm
SCoSE: http://www.uni-saarland.de/fak4/norrick/scose.html
SCOTS: http://www.scottishcorpus.ac.uk/
Shogakukan Corpus Network: http://www.corpora.jp
Sinica Treebank: http://treebank.sinica.edu.tw/
Sinotefl: http://www.sinotefl.ac.cn/english.asp
Slovak National Corpus: http://korpus.juls.savba.sk/index.en.html
SST: http://leo.meikai.ac.jp/~tono/sst/index.html
SUSANNE: http://www.grsampson.net/SueDoc.html
SWB online: http://www.ldc.upenn.edu/cgi-bin/lol/swb/speechcorpus?&corpus= swb
Talkbank: http://www.talkbank.org/
TalkBank SBCSAE: http://www.talkbank.org/data/Conversation/SBCSAE-zipped.zip

TalkBank SWB: http://www.talkbank.org/media/SWB/
TRACTOR: http://tractor.bham.ac.uk/tractor/catalogue.html
TransSearch: http://www.terminotix.com/eng/index.htm
UCREL: http://ucrel.lancs.ac.uk/
USC Hansard: http://www.isi.edu/natural-language/download/hansard/
USE: http://www.engelska.uu.se/use.html
Xaira: http://www.xaira.org
XGL: http://ariadne.coli.uni-bielefeld.de/indogram/component/option,com_vfm/Itemid,33/dir,Corpora/
ZEN: http://escorp.unizh.ch/

16. Literature

URLs were accessed on 12 February 2008.

Aduriz, I./Aldezabal, I./Alegria, I./Arriola, J./Diaz de Ilarraza, A./Ezeiza, N./Gojenola, K. (2003), Finite State Applications for Basque. In: *Proceedings of'EACL'2003 Workshop on Finite-state Methods in Natural Language Processing*. Budapest, 13−14 April 2003. Available online at: http://citeseer.ist.psu.edu/623753.html.

Altenberg, B./Aijmer, K./Svensson, M. (2001), *The English-Swedish Parallel Corpus (ESPC): Manual of Enlarged Version*. Lund and Göteborg: Universities of Lund and Göteborg.

Armstrong, S./Kempen, M./McKelvie, *D.*/ Petitpierre, D./Rapp, R./Thompson, H. (1998), Multilingual Corpora for Cooperation. In: Proceedings of the *First International Conference on Language Resources and Evaluation (LREC)*. Granada, Spain, 975−980.

Armstrong-Warwick, S./Thompson, H./McKelvie, D./Petitpierre, D. (1994), Data in your Language: The ECI Multilingual Corpus 1. In: *Proceedings of the International Workshop on Shareable Natural Language Resources*. Nara, Japan. Available at: http://citeseer.ist.psu.edu/205355.html.

Aston, G./Burnard, L. (1998), *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Auran, C./Bouzon, C./Hirst, D. (2004), The Aix-MARSEC Project: An Evolutive Database of Spoken British English. In: *Proceedings of the Second International Conference on Speech Prosody*. Nara, Japan, 561−564.

Bai, X./Chang, B./Zhan, W. (2002), Building a Large Chinese−English Parallel Corpus. In: Huang, H. (ed.), *Proceedings of the National Symposium on Machine Translation 2002*. Beijing: Electronic Industry Press, 124−131.

Baker, P./Hardie, A./McEnery, T./Xiao, R./Bontcheva, K./Cunningham, H./Gaizauskas, R./Hamza, O./Maynard, D./Tablan, V./Ursu, C./Jayaram, B./Leisher, M. (2004), Corpus Linguistics and South Asian Languages: Corpus Creation and Tool Development. In: *Literary and Linguistic Computing* 19(4), 509−524.

Barlow, M. (1998), *A Corpus of Spoken Professional American English*. Houston, TX: Athelstan.

Beare, J./Scott, B. (1999), The Spoken Corpus of the Survey of English Dialects: Language Variation and Oral History. In: *Proceedings of ALLC/ACH 1999*. Charlottesville, VA. Available at: http://www.iath.virginia.edu/ach-allc.99/proceedings/scott.html.

Biber, D. (1988), *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D./Finegan, E./Atkinson, D. (1994), ARCHER and its Challenges: Compiling and Exploring a Representative Corpus of Historical English Registers. In: Fries, U./ Tottie, G./Schneider, P. (eds.), *Creating and Using English Language Corpora*. Amsterdam: Rodopi, 1−14.

Böhmová, A./Hajič, J./Hajičova, E./Hladka, B. (2003), The Prague Dependency Treebank: Three-level Annotation Scenario. In: Abeille, A. (ed.), *Treebanks: Building and Using Syntactically Annotated Corpora*. Dordrecht: Kluwer, 103−127.

Burnard, L. (2002), Where did we Go Wrong? A Retrospective Look at the British National Corpus. In: Ketterman, B./Marko, G. (eds.), *Teaching and Learning by Doing Corpus Analysis: Proceedings of the Fourth International TALC*. Amsterdam: Rodopi, 51−70.

Burnard, L. (2003) Reference Guide for BNC-baby. Available at: http://www.natcorp.ox.ac.uk/corpus/baby/.

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

Cavar, D./Geyken, A./Neumann, G. (2000), Digital Dictionary of the 20th Century German Language. In: Erjavec, T./Gros, J. (eds.), *Proceedings of the Language Technologies Conference*. Ljubljana, Slovenia. Available at: http://nl.ijs.si/isjt00/index-en.html.

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

Cheng, W./Greaves, *C.*/ Warren, M. (2005), The Creation of a Prosodically Transcribed Intercultural Corpus: The Hong Kong Corpus of Spoken English (Prosodic). In: *ICAME Journal* 29, 47−68.

Choukri, K. (2003), Brief Overview of Recent Activities in Europe. In: *Proceedings of COCOSDA Workshop 2003*. Geneva, Switzerland. Available at: http://www.cocosda.org/meet/2003/kc-cocosda.pdf.

Coxhead, A. (2000), A New Academic Word List. In: *TESOL Quarterly* 34(2), 213−238.

Crowdy, S. (1993), Spoken Corpus Design. In: *Literary and Linguistic Computing* 8(4), 259−265.

Culpeper, J./Kytö, M. (1997), Towards a Corpus of Dialogues, 1550−1750. In: Ramisch, H./Wynne, K. (eds.), *Language in Time and Space: Studies in Honour of Wolfgang Viereck on the Occasion of his 60th Birthday*. Stuttgart: Franz Steiner Verlag, 60−73.

Culpeper, J./Kytö, M. (2000), Data in Historical Pragmatics: Spoken Interaction (Re)cast as Writing. In: *Journal of Historical Pragmatics* 1(2), 175−199.

Culpeper, J./Kytö, M. (forthcoming), *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: Cambridge University Press.

Dalli, A. (2001), Interoperable Extensible Linguistic Databases. In: *Proceedings of IRCS Workshop on Linguistic Databases*. Philadelphia, PA, 74−81. Available at: http://www.ldc.upenn.edu/annotation/database/proceedings.html.

Denison, D. (1994), A Corpus of Late Modern English Prose. In: Kytö, M./Rissanen, M./Wright, S. (eds.), *Corpora across the Centuries*. Amsterdam: Rodopi, 7−16.

Du Bois, J./Chafe, W./Meyer, C./Thompson, S. (2000−2005), *Santa Barbara Corpus of Spoken American English* Parts 1−4. Philadelphia, PA: Linguistic Data Consortium.

Ellis, N./O'Dochartaigh, C./Hicks, W./Morgan, M./Laporte, N. (2001), *Cronfa Electroneg o Gymraeg (CEG): A 1 Million Word Lexical Database and Frequency Count for Welsh*. Available at: http://www.bangor.ac.uk/development/canolfanbedwyr/ceg.php.en.

English Language Institute (2003), *MICASE Manual: The Michigan Corpus of Academic Spoken English* (version 1.1). University of Michigan. Available at: http://lw.lsa.umich.edu/eli/micase/MICASE_MANUAL.pdf.

Erjavec, T. (2002), The IJS-ELAN Slovene-English Parallel Corpus. In: *International Journal of Corpus Linguistics* 7(1), 1−20.

Erjavec, T. (2004), MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: *LREC 2004 Proceedings*. Paris, France. Available at: http://nl.ijs.si/ME/bib/mte-lrec2004.pdf.

Farr, F./Murphy, B./O'Keeffe, A. (2004), The Limerick Corpus of Irish English: Design, Description and Application. In: *Teanga* 21, 5−29.

Fries, U./Schneider, P. (2000), ZEN: Preparing the Zurich English Newspaper Corpus. In: Ungerer, F. (ed.), *English Media Texts: Past and Present*. Amsterdam: John Benjamins, 1−24.

Garabík, R. (2006), Computer(ized) Linguistic Resources at the Ł. Štúr Institute of Linguistics. In: *Proceedings of the Conference Applied (Computer) Linguistics*. Kiev, Ukraine, 56−59.

Garside, R./Leech, G./Váradi, T. (1992), *Manual of Information for the Lancaster Parsed Corpus*. Lancaster: Lancaster University.

Garside, R./Hutchinson, J./ Leech, G./McEnery, A./Oakes, M. (1994), The Exploitation Of Parallel Corpora in Projects ET10/63 and CRATER. In: *New Methods in Language Processing*. Manchester: UMIST, 108−115.

Garside, R./Leech, G./Sampson, G. (eds.) (1987), *The Computational Analysis of English: A Corpus-based Approach*. Harlow: Longman.

Gautier, G. (1998), Building a Kurdish Language Corpus. Paper presented at ICEMCO 98 6th International Conference and Exhibition on Multilingual Computing. Cambridge, United Kingdom, April 1998. Available at: http://ggautierk.free.fr/e/icem_98.htm.

Gianitsová, L. (2005), Morphological Analysis of the Slovak National Corpus. In: M. Šimková (ed.), *Insight into Slovak and Czech Corpus Linguistics*. Bratislava: Veda, 166−178.

Gillard, P./Gadsby, A. (1998), Using a Learners' Corpus in Compiling ELT Dictionaries. In: Granger, S. (ed.), *Learner English on Computer*. London: Longman, 159−171.

Glover, W. (1998), Toward a Nepali National Corpus. In: Yadava, P./Kansakar, T. (eds.), *Lexicography in Nepal: Proceedings of the Institute on Lexicography, 1995*. Kamaladi, Kathmandu: Royal Nepal Academy, 24−28.

Gómez Guinovart, X./Sacau Fontenla, E. (2004), Parallel Corpora for the Galician Language: Building and Processing of the CLUVI (Linguistic Corpus of the University of Vigo). In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Lisbon, Portugal, 1179−1182.

Grabe, E./Post, B./Nolan, F. (2001), *The IViE Corpus*. Department of Linguistics, University of Cambridge. Available at: http://www.phon.ox.ac.uk/IViE/.

Granger, S. (2003), The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. In: *TESOL Quarterly* 37(3), 538−546.

Granger, S./Tyson, S. (1996), Connector Usage in the English Essay Writing of Native and Non-native EFL Speakers of English. In: *World Englishes* 15(1), 17−27.

Greenbaum, S./Svartvik, J. (1990), The London-Lund Corpus of Spoken English. In: Svartvik, J. (ed.), *The London Lund Corpus of Spoken English: Description and Research*. (Lund Studies in English 82.) Lund: Lund University Press. Available at: http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX. HTM.

Gu, Y. (2002), Sampling Situated Discourse for Spoken Chinese Corpus. Manuscript available at: http://ling.cass.cn/dangdai/gu_papers/sampling%20situated%20discourse.pdf.

Guerra, L. (1998), Research in Language and Literature: Old Problems, New Solutions. Paper presented at the conference of *The Future of Humanities in the Digital Age*. Bergen, Norway, 25−26 September 1998. Available at: http://ultibase.rmit.edu.au/Articles/dec98/guerra1.htm.

Gui, S./Yang, H. (2002), *Chinese Learner English Corpus*. Shanghai: Shanghai Foreign Language Education Press.

Hajič, J. (2004), *Complex Corpus Annotation: The Prague Dependency Treebank*. Bratislava, Slovakia: Jazykovedný ústav Ł. Štúra, SAV.

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

Hatzigeorgiu, N./Gavrilidou, M./Piperidis, S./Carayannis, G./Papakostopoulou, A./Spiliotopoulou, A./Vacalopoulou, A./Labropoulou, P./Mantzari, E./Papageorgiou, H./Demiros, I. (2000), Design and Implementation of the Online ILSP Greek Corpus. In: *Proceedings of LREC 2000*. Athens, Greece, 1737−1742.

Hoffmann, S./Evert, S. (2006), BNCweb (CQP-Edition) − the Marriage of Two Corpus Tools. In: Braun, S./Kohn, K./Mukherjee, J. (eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt am Main: Peter Lang, 177−195.

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

Holmes-Higgin, P./Abidi, S./Ahmad, K. (1994), A Description of Texts in a Corpus: "Virtual" and "Real" Corpora. In: Martin, W./Meijs, W./Moerland, M./ten Pas, E./van Sterkenburg, P./Vossen, P. (eds.), *EURALEX'94 Proceedings*. Amsterdam: Vrije Universiteit, 390−402.

Horváth, J. (1999), Advanced Writing in English as a Foreign Language. A Corpus-based Study of Processes and Products. PhD thesis, Janus Pannonius University.

Huang, C. (2006), Automatic Acquisition of Linguistic Knowledge: From Sinica Corpus to Gigaword Corpus. In: *Language Corpora: Their Compilation and Application. Proceedings of the 13th NIJL International Symposium*. Tokyo, Japan, 41−48.

Huang, C./Chen, K. (1995/1998), *CKIP Technical Report 95-02/98-04*. Taipei: Academia Sinica.

Huang, C./Chen, F./Chen, K./Gao, Z./Chen, K. (2000), Sinica Treebank: Design Criteria, Annotation Guidelines, and Online Interface. In: Bagga, A./Pustejovsky, J./Zadrozny, W. (eds.), *Proceedings of NAACL-ANLP 2000 Workshop: Syntactic and Semantic Complexity in Natural Language Processing Systems*. Seattle, Washington, 29−37.

Hundt, M./Sand, A./Siemund, R. (1998), *Manual of Information to Accompany the Freiburg-LOB Corpus of British English ("FLOB")*. Available at: 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")*. Available at: http://khnt.hit.uib.no/icame/manuals/frown/INDEX.HTM.

Izumi, E./Isahara, H. (2004), Investigation into Language Learners' Acquisition Order Based on the Error Analysis of the Learner Corpus. Paper presented at IWLeL 2004. Tokyo, Japan.

Izumi, E./Uchimoto, K./Isahara, H. (2004), SST Speech Corpus of Japanese Learners' English and Automatic Detection of Learners' Errors. In: *ICAME Journal* 28, 31−48.

Johansson, S./Ebeling, J./Oksefjell, S. (2002), *English-Norwegian Parallel Corpus: Manual*. Oslo: University of Oslo.

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: University of Oslo.

Kang, B./Kim, H. (2004), Sejong Korean Corpora in the Making. In: *Proceedings of LREC 2004*. Lisbon, Portugal, 1747−1750.

Kim, H. (2006), Korean National Corpus in the 21st Century Sejong Project. In: *Language Corpora: Their Compilation and Application. Proceedings of the 13th NIJL International Symposium*. Tokyo, Japan, 49−54.

Kroch, A./Taylor, A. (2000), *The Penn-Helsinki Parsed Corpus of Middle English. Second Edition*. University of Pennsylvania. Available at: http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-2/.

Kruyt, J. (1995), Nationale tekstcorpora in internationaal perspectief. In: *Forum der Letteren* 36(1), 47−58.

Kruyt, J./Dutilh, M. (1997), A 38 Million Words Dutch Text Corpus and its Users. In: *Lexikos 7* (Afrilex-reeks/series 7), 229−244.

Kruyt, J./Raaijmakers, S./van der Kamp, P./van Strien, R. (1996), Language Resources for Language Technology. In: *Proceedings of the First TELRI European Seminar*. Tihany, Hungary, 173−178.

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

Kučera, K. (2002), The Czech National Corpus: Principles, Design, and Results. In: *Literary and Linguistic Computing* 17(2), 245−257.

Kytö, M. (1996), *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts*. Helsinki: University of Helsinki. Available at: http://khnt.hit.uib.no/icame/manuals/HC/INDEX.HTM.

Kytö, M./Walker, T. (2006), *Guide to* A Corpus of English Dialogues 1560−1760. (Studia Anglistica Upsaliensia 130.) Uppsala: Acta Universitatis Upsaliensis.

Laitinen, M. (2002), Extending the Corpus of Early English Correspondence to the 18th Century. In: *Helsinki English Studies* 2002(2). Available at: http://www.eng.helsinki.fi/hes/Corpora/ extending_the_corpus2.htm.

Lee, D. (2001), Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. In: *Language Learning and Technology* 5(3), 37−72.

Leech, G./Smith, N. (2005), Extending the Possibilities of Corpus-based Research on English in the Twentieth Century: A Prequel to LOB and FLOB. In: *ICAME Journal* 29, 83−98.

Lewandowska-Tomaszczyk, B. (2003), The PELCRA Project − State of Art. In: Lewandowska-Tomaszczyk, B. (ed.), *Practical Applications in Language and Computers*. Frankfurt: Peter Lang, 105−121.

Linguistic Data Consortium (1995), SWITCHBOARD: A User's Manual. Philadelphia, PA: LDC, University of Pennsylvania. Available at: http://www.ldc.upenn.edu/Catalog/readme_files/ switchboard.readme.html

MacWhinney, B. (1995), *The CHILDES Project: Tools for Analyzing Talk*. Second Edition. Hillsdale, NJ: Erlbaum.

Malten, T. (1998), Tamil Studies in Germany. Lecture given at Max Mueller Bhavan, Chennai on 17 March 1998. Available at: http://www.tamilnation.org/literature/malten.htm.

Marcus, M. (1999), *Manual of ICAMET (Innsbruck Computer-Archive of Machine-Readable English Texts)*. (Innsbrucker Beiträge zur Kulturwissenschaft, Anglistische Reihe 7.) Innsbruck: Leopold-Franzens-Universität Innsbruck, Institut für Anglistik.

Marcus, M./Santorini, B./Marcinkiewicz, M. (1993), Building a Large Annotated Corpus of English: The Penn Treebank. In: *Computational Linguistics* 19, 313−330.

Marcus, M./Kim, G./Marcinkiewicz, M./MacIntyre, R./Bies, A./Ferguson, M./ Katz, K./Schasberger, B. (1994), The Penn Treebank: Annotating Predicate-argument Structure. In: *Proceedings of the ARPA Human Language Technology Workshop*. Plainsboro, NJ, 114−119.

McEnery, A./Xiao, R./Mo, L. (2003), Aspect Marking in English and Chinese: Using the Lancaster Corpus of Mandarin Chinese for Contrastive Language Study. In: *Literary and Linguistic Computing* 18(4), 361−378.

Milton, J./Chowdhury, N. (1994), Tagging the Interlanguage of Chinese Learners of English. In: Flowerdew, L./Tong, A. (eds.), *Entering Text*. Hong Kong: The Hong Kong University of Science and Technology, 127−143.

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

Nelson, G./Wallis, S./Aarts, B. (ed.) (2002), *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.

Nevalainen, T. (2000), Gender Differences in the Evolution of Standard English. In: *Journal of English Linguistics* 28(1), 38−59.

Rayson, P./Tono, Y./Morita, Y./Hoshino, M./Nakamura, T./Aizawa, H./Watanabe, R. (2005), Building a Corpus of Professional English. Poster presented at the *Corpus Linguistics 2005* conference, July 14−17, Birmingham, UK.

Reppen, R./Ide, N. (2004), The American National Corpus: Overall Goals and the First Release. In: *Journal of English Linguistics* 32(2), 105−113.

Rissanen, M. (2000), The World of English Historical Corpora. In: *Journal of English Linguistics* 8(1), 7−20.

Riza, H. (1999), *The Indonesia National Corpus and Information Extraction Project (INC-IX)*. Jakarta: BPP Teknologi.

Rossini Favretti, R./Tamburini, F./de Santis, C. (2004), A Corpus of Written Italian: A Defined and a Dynamic Model. In: Wilson, A./Rayson, P./McEnery, T. (eds.), *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*. Munich: Lincom-Europa. Available at: http:// corpora.dslo.unibo.it/People/Tamburini/CL2001.pdf.

Sampson, G. (1987), The Grammatical Database and Parsing Scheme. In: Garside, R./Leech, G./Sampson, G. (eds.), *The Computational Analysis of English*. London: Longman, 82−96.

Sampson, G. (1995), *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Oxford: Clarendon Press.

Sampson, G. (2000), *CHRISTINE Corpus: Documentation*. Sussex: University of Sussex. Available at: http://www.grsampson.net/ChrisDoc.html.

Sampson, G. (2005), *The LUCY Corpus: Documentation*. Sussex: University of Sussex. Available at: http://www.grsampson.net/LucyDoc.html.

Sanchez, M. (2002), CREA: Reference Corpora for Current Spanish. In: *Proceedings of Language Corpora: Present and Future*. Donostia, Spain, 24−25.

Santos, D./Rocha, P. (2001), Evaluating CETEMPúblico, a Free Resource for Portuguese. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, ACL'2001*. Toulouse, France, 442−449.

Schmied, J. (1994), The Lampeter Corpus of Early Modern English Tracts. In: Kytö, M./Rissanen, M./Wright, S. (eds.), *Corpora across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora. St Catharine's College Cambridge, 25−27 March 1993*. Amsterdam: Rodopi, 81−89.

Schneider, P. (2002), Computer Assisted Spelling Normalization of 18th Century English. In: Peters, P./Collins, P./Smith, A. (eds.), *New Frontiers of Corpus Research*. Amsterdam: Rodopi, 199−214.

Scott, M. (2004), *WordSmith Tools*. Oxford: Oxford University Press.

Sharoff, S. (2006), Methods and Tools for Development of the Russian Reference Corpus. In: Archer, D./Wilson, A./Rayson, P. (eds.), *Corpus Linguistics Around the World*. Amsterdam: Rodopi, 167−180.

Šimková, M. (2005), Slovak National Corpus − History and Current Situation. In: M. Šimková (ed.), *Insight into Slovak and Czech Corpus Linguistics*. Bratislava: Veda, 152−159.

Souter, C. (1993), Towards a Standard Format for Parsed Corpora. In: Aarts, J./Haan, P./Oostdijk, N. (eds.), *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam: Rodopi, 197−214.

Stern, K. (1997), The Longman Spoken American Corpus: Providing an In-depth Analysis of Everyday English. In: *Longman Language Review* 3. Available at: http://www.longman.com/dictionaries/pdfs/Spoken-American.pdf.

Stibbard, R. (2001), Vocal Expression of Emotions in Non-laboratory Speech: An Investigation of the Reading/Leeds Emotion in Speech Project Annotation Data. PhD thesis, University of Reading.

Tan, M. (2005), Authentic Language or Language Errors? Lessons from a Learner Corpus. In: *ELT Journal* 59(2), 126−134.

Taylor, L./Knowles, G. (1988), *Manual of Information to Accompany the SEC Corpus: The Machine Readable Corpus of Spoken English*. University of Lancaster. Available at: http://khnt.hit.uib.no/icame/manuals/sec/INDEX. HTM.

Thompson, P. (2001), A Pedagogically-motivated Corpus-based Examination of PhD theses. PhD thesis, University of Reading, UK.

Tsou, B./Tsoi, W./Lai, T./Hu, J./Chan, S. (2000), LIVAC, a Chinese Synchronous Corpus, and Some Applications. In: *Proceedings of the ICCLC International Conference on Chinese Language Computing*. Chicago, IL, 233−238.

van Bergen, L./Denison, D. (2004), A Corpus of Late Eighteenth Century Prose. In: Beal, J./Corrigan, K./Mosil, H. (eds.), *Models and Methods in the Handling of Unconventional Digital Corpora*, vol. 2. *Diachronic Corpora*. Basingstoke, UK: Palgrave Macmillan, 228−246.

Váradi, T. (2002), The Hungarian National Corpus. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas, Spain, 385−389.

Wang, J. (2001), Recent Progress in Corpus Linguistics in China. In: *International Journal of Corpus Linguistics* 6(2), 281−304.

Wang, K. (ed.) (2004), *The Development of the Compilation and Application of Parallel Corpora*. Beijing: Foreign Language Education and Research Press.

Wittenburg, P./Brugman, H./Broeder, D. (2000), Summary. In: Broeder, D./Cunningham, H./Ide, N. (eds.), *Proceedings of LREC 2000 Pre-conference Workshop on Meta-descriptions for Multimedia Language Resources*. Athens, Greece. Available at: http://www.mpi.nl/world/ISLE/documents/papers/LREC2000Workshop.pdf.

Xue, N./Xia, F./Chiou, F./Palmer, M. (2005), The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. In: *Natural Language Engineering* 11(2), 207–238.

Zhan, W./Chang, B./Duan H./Zhang H. (2006), Recent Developments in Chinese Corpus Research. In: *Language Corpora: Their Compilation and Application. Proceedings of the 13th NIJL International Symposium*. Tokyo, Japan, 19–30.

Richard Xiao, Ormskirk (UK)

21. Corpora of less studied languages

1. Why less studied?
2. Three main motives for language corpora exemplified
3. Cultural archives
4. Artificial corpora
5. Linguistic documentation
6. Some general sources for less studied languages
7. A survey of global coverage
8. Special problems with corpora of less studied languages
9. Conclusions
10. Literature

1. Why less studied?

1.1. Definitions

This article examines the provision of corpora to represent "less studied languages", sometimes also known as "lesser used languages". This term evidently has vague boundaries. For this volume, the need is to complement the coverage of the world's languages in article 20, "Well-known and influential corpora". That article is devoted predominantly to a variety of corpora for English (often prepared within a foreign context), together with a few corpora for other major national languages (Chinese, Russian, Polish, Korean, German), some major multilingual corpora for Europe and the Indian subcontinent, and (most pertinently to this article) large-scale distributors of language corpora, some of which are in practice global in what they offer.