Linguistica Silesiana 21, 2000
ISSN 0208-4228

EWA MYRCZEK
University of Silesia Katowice

CORPUS – ITS DEFINITION, COMPILATION, TAXONOMY AND FUTURE

The aim of this article is to discuss a few issues related to corpus, mainly, its definition, development, compilation and taxonomy. The author demonstrates a distinction between a corpus and a text archive or a text database (text bank). The development of corpora is divided into the two following stages: pre-electronic and electronic. Corpora are differentiated and classified according to language variables such as monolingual vs. multilingual, plain vs. annotated and data resources such as speech and written language corpora. The author discusses only a few features of a corpus, mainly its representativeness, size [static (closed) corpus vs. dynamic (monitor, open-ended) corpus] and form (machine-readable vs. print). The European and North American centres of corpus linguistics are surveyed. The author argues that the invention of the computer a turning point in the field of corpus linguistics as modern corpora are more precise and flexible than a couple centuries ago. The final conclusion of this article is that the Chomsky's criticism levelled at the practicality of corpus linguistics is no longer valid.

0. Introduction

One of the most important and influential factors in dictionary construction is a corpus. Lexicography, although a part of applied linguistics overlaps with corpus linguistics when it comes to an in-depth analysis and consideration of data collection. In this article we will try to investigate a few issues related to corpus linguistics, mainly corpus taxonomy, and compilation. Before we discuss the very definition of corpus linguistics we need to explain two terms which are frequently used in internet terminology, namely *online* and *machine-readable*.

0.1. The definition of *online*

1) Interactive – accessible via a computer or on the Internet. (Free On-line Dictionary of Computing; 1999)
2) Turned on and connected, that is, ready to send or receive data – of computers, printers, etc... (Webopaedia; 1999)
3) Actively using a computer system – of users when they are connected to a computer service through a modem – they are actually *on the line*.

Thus, *online* data shall be any computer data we can have an access to when we are *online*, i.e. we are connected to a computer service provided our computer and modem are *online* too, i.e. they are turned on and working properly.

0.2. The definition of *machine-readable*

A form that is accepted by a computer. Machine-readable data includes files stored on disks or tapes, or data that comes from a device connected to a computer. Now, that we clarified any doubts as to online services and communication we can proceed to discuss corpus linguistics as such.

1. The Definition of CORPUS LINGUISTICS

Nowadays, corpus linguistics is mainly perceived as a study of language related to processing, usage and analysis of written and spoken corpora. This is a fairly comprehensive definition with the notion of corpus in the foreground. Let us have a look at different definitions of corpus formulated by lexicographers and corpus linguists:

1.1. Definition of a Corpus

CORPUS [from Latin *corpus* body. The plural is usually *corpora*]:

1) Any collection of more than one text (Tony McEnery and Andrew Wilson 1998)
2) A body of texts, utterances, or other specimens considered more or less representative; of a language, and usually stored as an electronic database (*The Oxford Companion to the English Language* ed. McArthur & McArthur; 1992);
3) A collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point for linguistic description or as a means verifying hypotheses about a language (David Crystal, *A Dictionary of Linguistics and Phonetics,* Blackwell, 3rd Edition, 1991);
4) A collection of naturally occurring language texts, chosen to characterise a state or variety of a language. (John Sinclair, *Corpus, Concordance, Collocation,* Oxford University Press, 1991).

The first and second definitions are definitely inadequate for two reasons:

a) We could take electronic text projects such as ***Gutenberg Project*** or books online as corpora.
b) For many centuries corpora were not stored in a machine-readable form.

One must also take into consideration a distinction made between a corpus and a text archive or a text database (text bank). According to Graeme Kennedy (Graeme Kennedy 1998) a corpus is normally a systematic, planned and structured compilation of text, whereas an archive can be any text repository, very often huge, and unstructured. However, many lexicographers can and do make use of text archives as electronic repositories that can be submitted to further processing and analysing. When this happens we may say that a text archive is converted into a corpus. This conversion require specialised tools and software which are already available on the computer market.

At present a corpus may be defined as a collection of written and (or) spoken texts in an electronic database. It may be complete and self-contained. It can be and usually is gathered according to particular principles for some particular purpose. A corpus is always a potential source of linguistic data as its component texts allow for statements to be

made about language as a whole. It represents some specified population or genre.

1.2. The development of Corpora

Before the invention of the computer corpora collection and analysis were also conducted, though in a less adequate and thorough way. Graeme Kennedy (1998) in *An Introduction to Corpus Linguistics* divided corpora into two types:

A. PRE-ELECTRONIC CORPORA which derived from biblical and literary as well as educational resources – they were in form of print,

B. ELECTRONIC CORPORA which can be further subdivided into:

a) first generation corpora (1960-1970) – ICAME corpora: corpora available from ICAME (International Computer Archive of Modern and Medieval English available at http://www.hd.uib.no/icame.html):

The Brown Corpus (The Standard Corpus of Present-Day Edited American English) – approximately 1,000,000 words of American written English printed in the year 1961; the first corpus to be put on computer medium; it consists of 500 American English texts of 2,000 words each representing the genre categories parallel to those of LOB corpus such as newspaper reportage, press editorials, memoirs, religion, science fiction, detective fiction, romance novels. The Brown Corpus is available from the Oxford Text Archive and the ICAME archive. Its tagged version was produced at Brown University during the period 1970-1978 and its parsed version known as Gothenburg Corpus is available from the University of Göteborg, Sweden.

The Lancaster-Oslo/Bergen Corpus (LOB Corpus) – approximately 1,000,000 words of written British English from 1961; it was compiled in the 1970s under the direction of Geoffrey Leech, University of Lancaster and Stig Johansson, University of Oslo. It contains 500 texts of roughly 2000 words apiece and it is made up of 15 different genre categories, available as orthographic text and tagged with the CLAWS1 part-of-speech tagging system.

The London-Lund Corpus (LLC) – is 500,000 words of spoken educated British English, collected in the period 1960-1970 from speakers of various background with a range of discourse types. It contains 100 texts which can be obtained on the ICAME CD-ROM

b) second generation corpora (1980-1990):

the Cobuild project – Bank of English was launched in 1991 by COBUILD (a division of HarperCollins Publishers) and The University of Birmingham. Since 1980 COBUILD, which is based within the School of English at Birmingham University, has been collecting a corpus of texts on computer for dictionary compilation and language study. In 1991 HarperCollins decided on a major initiative to increase the scale of the corpus to 200 million words, to form the basic data resource for a new generation of authoritative language reference publications. On 20 July 1998 the latest release of the corpus amounted to 329 million words and it continues to grow with the constant addition of new material. Written texts come from newspapers, magazines, fiction and non-fiction books, brochures, leaflets, reports, letters, and so on. The spoken word is represented by transcriptions of everyday casual conversation, radio broadcasts, meetings, interviews and discussions, etc. The corpus includes millions of words of transcribed speech from the BBC World Service radio broadcasts, and the American National Public Radio.

The mix and variety of texts represented in the Bank of English is kept under constant review and new sources are introduced to maintain the balance of the material so that it reflects the mainstream of current English today. The Bank of English is available for linguistic study at the University of Birmingham. By arrangement with COBUILD or the School of English, visitors may consult the corpus online and carry out their own analysis.

The Longman Corpus Network is a diverse, far-reaching group of databases consisting of many millions of words. The Network provides Longman lexicographers and course book writers with general knowledge about words, usage, language trends and grammatical patterns using modern technology. Five language databases form the nucleus of the Network (Addison Wesley Longman; 1998):

the Longman/Lancaster Corpus with over 30 million words covers a range of written language taken from literature, magazines, papers and more ephemeral materials such as leaflets and packaging. It began in 1985. Words can be looked at in the corpus via a concordancing programme. One calls up the word one wants to examine and the corpus shows every occurrence of that word in context in its 30 million word databank. Example after example of that word comes up on the screen and from so many examples the lexicographer is able to deduce a great deal of information about how the given word is being used.

the Longman Learners' Corpus – recorded and monitored the written output of students of English in order to pinpoint their specific needs;

the Longman Written American Corpus – comprised of 100 million words including running text from newspapers, journals, magazines, best-selling novels, technical and scientific writing, and coffee-table books

the Longman Spoken American Corpus – a resource of 5 million words of everyday American speech;

the Spoken British Corpus – gives information on what spoken English is like and how it differs from written British English.

The British National Corpus (BNC) – a 100 million word corpus of written and spoken British English from the early 1990s, it contains extracts from 4124 modern British English texts of all kinds, both spoken and written, each text is segmented into orthographic sentence units, and each word automatically assigned a part of speech code, it carries a grammatical tag, that is, a label indicating its part of speech. This process was carried out at **Lancaster University's Unit for Computer Research on the English Language (UCREL),** using the CLAWS4 automatic tagger developed by Roger Garside at Lancaster. There are 6 and a quarter million sentences, and over 100 million words. BNC was produced by collaborative efforts of **Oxford University Press (OUP), Longman, Chambers-Larousse**) and academic research centres: **Oxford University, Lancaster University and the British Library**. BNC is monolingual – as it handles modern English only, synchronic – it covers British English of the late twentieth century, general – it encompasses many different styles and varieties

According to the lexicographers' intentions 75% of the written texts were to be chosen from *informative* writing: of which roughly equal quantities should be chosen from the fields of applied sciences, arts, belief & thought, commerce & finance, leisure, natural & pure science, social science, world affairs. 25% of the written

texts were to be *imaginative*, that is, literary and creative works. 60% of the written texts were to be books, 25% were to be periodicals (newspapers etc.) between 5 and 10% should come from other kinds of miscellaneous published material (brochures, advertising leaflets, etc.) between 5 and 10% should come from unpublished written material such as personal letters and diaries, essays and memoranda, etc. a small amount (less than 5%) should come from material written to be spoken (for example, political speeches, play texts, broadcast scripts, etc.)

Every one of the 100 million words in BNC carries a grammatical tag, that is, a label indicating its part of speech. In addition, each text is divided into sentence-like segments. This process was carried out at **Lancaster University's Unit for Computer Research on the English Language (UCREL,** available at http://www.comp.lancs.ac.uk/computing/research/ucrel), using the CLAWS4 automatic tagger developed by Roger Garside at Lancaster. The basic BNC tagset (known as C5) distinguishes 61 categories found in most „traditional" grammars, such as adjectives, articles, adverbs, conjunctions, determiners, nouns, verbs etc. Tags are also attached to major punctuation marks, indicating their function. This automatic procedure has an error rate of around 1.7%. In addition, about 4.7% words could not be assigned unambiguously to a single category. To overcome these problems, a 2% sample of the corpus was manually post-edited, using an enriched BNC tagset known as C7, in which over 160 categories are distinguished, with a much lower error rate (less than 0.3%).

The International Corpus of English (ICE) – began in 1988 for the purpose of providing comparable data of national varieties of English internationally (discussed in detail later in this article, more information available also at http://www.ucl.ac.uk/english-usage/ice/index.htm)

2. Data Collection

Data collection is a very significant stage of corpus compilation which is fairly easy to carry out nowadays. All linguistic material is converted into an electronic form which in turn will be processed and analysed later on. We will start with a discussion of written language data.

2.1.Written Language

When collecting written data one must take into consideration several factors such as: data availability, speed of written data conversion into electronic data, accuracy, cost.

2.1.1. SCANNING (i.e. optical character reading), i.e. creating digital images of the pages of the text using scanners which resembles using photocopiers in the way they work, instead of producing a paper copy of the image they produce a copy on the computer. In the next stage we need to convert what is a series of pixels, into a text file. The special OCR software (short for object and character recognition software) has been designed to recognise patterns of pixels in digital images that correspond to characters, and so solve the problem of turning an image into a text file. The very conversion process is extremely complicated and the software that does this is very technical. Afterwards, a spell check programme is applied as an aid to spotting scanning errors and other mistakes. Spelling

checkers are usually based on a collection of word forms representing an actual corpus and are used to find spelling errors in a text. They are most probably the number one commercial application! Only high-quality original texts are required to ensure that the error rate is low, however hand-editing is still required in order to correct scanning errors and insert **textual mark-up** (in computerised document preparation, a method of adding information to the text indicating the logical components of a document, or instructions for layout of the text on the page or other information which can be interpreted by some automatic system). Let us take a look at the project carried out at Carnegie Mellon University Libraries. They launched upon a project of converting approximately one million pages of the congressional papers of Pennsylvania Senator John Heinz into digital format. Processed documents were scanned, converted to ASCII form via OCR (Optical Character Recognition) software, verified and annotated, and then indexed using the CLARIT natural language processing software. Scanners are however, not efficient enough at recognising small typefaces, lower-quality typography, or handwriting. Then it is better to choose:

2.1.2 KEYBOARDING – manually typing in texts, which is more time consuming. This is normally the case with magazines and some ephemera.

2.1.3. The fastest and most convenient way of written language collection seems to be REUSE OF EXISTING ELECTRONIC TEXTS – electronic texts defined as all textual information that is stored in data files and can be easily retrieved on computers. Consequently all publishers' and typesetters' versions of newspapers, magazines and some books – converting to the standard format required for the corpus is fairly straightforward.

A very high percentage of all published material exists in electronic online form which is available at Linguistic Data Consortium (LCD, http://www.ldc.upenn.edu), the Oxford Text Archive (http://info.ox.ac.uk/~archive.ota.html), Association for Computational Linguistics Data Collection Initiative (ACL/DCI), the International Computer Archive of Modern English (ICAME), a wide variety of texts in a machine-readable form often out-of-copyright books, CD-ROMs, disks etc. Many universities have already created Electronic Text Centres such as the one at the University of Virginia which has pursued twin missions (David Seaman; 1999, available at http://www.lib.virginia.edu/ecenters.html/) to build and maintain an internet-accessible collection of SGML (Standard Generalised Markup Language) – texts and images; to build and maintain a user community adept at the creation and use of these materials .

2.2. Spoken Language

2.2.1. WIRING FOR SOUND. The structure of spoken language is shaped by many factors such as the phonological, syntactic and prosodic features of the language being spoken, by the acoustic environment and context in which it is produced – e.g., people speak differently in noisy or quiet environments – and the communication channel through which it travels. In creation of speech corpora one has to take into consideration the fact that speech is produced differently by each speaker. Each utterance is produced by a unique vocal tract which assigns its own signature to the signal. Speakers of the same language have different dialects, accents and speaking rates. Their speech patterns are influenced by the physical environment, social context, the perceived social status of the participants, and their emotional and physical state. Large amounts of annotated speech data are needed to model

the affects of these different factors. Let us take an example of spoken British Corpus which was collected by a market research agency which was commissioned to make a selection of British-English speakers in the UK. Each person selected was given a small Walkman and a microphone and asked to record all the speech he or she may hear or take part in over a one week period. The tapes were sent back to lexicographers at Longman, transcribed and entered onto the computer. The total number of participants was over two thousand. Four broad categories of social context were taken into consideration:

Educational and informative events, such as lectures, news broadcasts, classroom discussion, tutorials.

Business events such as sales demonstrations, trades union meetings, consultations, interviews.

Institutional and public events, such as sermons, political speeches, council meetings, parliamentary proceedings.

Leisure events, such as sports commentaries, after-dinner speeches, club meetings, radio phone-ins.

3. Taxonomies of corpora

Corpora can be differentiated and classified according to various criteria. In this article we will focus on language variables such as monolingual vs. multilingual, plain vs. annotated and data resources such as speech and written language corpora. The very corpus design depends on a few principles that are followed by the corpus constructor. However, according to many linguists it is hardly possible to define the rules and procedures which should be applied in corpus compilation.

> *There is no consensus in the community as to the procedures to be followed in corpus design (balanced, opportunistic, statistically sophisticated and defiantly naive approaches all struggle with each other for acceptance) ...*
> (C. M. Sperberg-McQueen, RE: Q: bilingual corpora, TEI-L 1994)

Before we commence to gather linguistic data there are a few questions to be answered:

- Who are the intended users? (e.g. personal research vs. a general resource)
- What is the purpose of the corpus? (e.g. as a basis for a dictionary; to create a word frequency list; to study some linguistic phenomenon; to study the language of a particular author or time period; to train a NLP system; as a teaching resource for non-native speakers; to study language acquisition ...)
- How much data is needed/realistic? What variables should be anticipated?
- Sampling? or exhaustive? (e.g.: the complete OE corpus is available online; a complete Early Middle English corpus is feasible; a complete 20th c. British or American English corpus is not feasible)

3.1. Language Variables: monolingual vs. bilingual (multilingual)

Monolingual corpora contain texts in a single language only, e.g. Bank of English.

As far as bilingual and multilingual corpora are concerned they can be further subdivided into 2 types

TYPE 1 – contains small collections of individual monolingual corpora where the same procedures and categories are used for each language but each contains completely different texts making up the corpora, e.g. the Aarthus corpus of Danish French and Eng-

lish contract law.

TYPE 2 – *parallel (aligned) corpus* (where texts of source language and target language are aligned on the level of sentence) covers a variety of corpora types, but in general it refers to texts that are translations of each other (or are at least on the same topic), that is, texts and their translations are aligned, sentence by sentence. It dates back to the mediaeval times, when grammar-translation method of teaching was very popular and 'polyglot bibles' were produced. There are several options among parallel corpora:

- parallel corpus containing only texts originally written in language A and their parallel translations into languages B (and C...); mono-directional translations from SL into TL
- parallel corpus containing an equal amount of texts originally written in languages A and B and their respective translations; bi-directional translations
- parallel corpus containing only translations of texts into languages A, B and C, whereas the texts were originally written in language Z." (Teubert 1996:245); translations into TLs without the original version

The arguments against using parallel corpora, i.e., corpora based on translations, have generally been that a) translations distort the TL because they give a mirror image of the SL, b) the translated language is different from the original language and c) translators are unreliable and make mistakes.

Below is the list of multilingual corpora:

1. English-Norwegian Parallel Corpus at the University of Oslo
2. Translation Corpus of English and German at the Technical University Of Chemnitz-Zwickau includes EC-material, academic textbooks, modern fiction and tourist brochures (approx. 500000 words altogether). The researchers are currently looking at aspects such as culture-specific problems in translation or translationese.
3. TRIPTIC TRIlingual Parallel Text Information Corpus. A trilingual corpus developed for the analysis of prepositions in English, French and Dutch.
4. LINGUA PROJECT. A project involving the construction of multilingual corpora for English, French, Greek and some others, for use in language pedagogy.
5. MULTEX-PROJECT. Building tools for multilingual corpus access, and also a bunch of sample corpora.
6. MULTEX-EAST. Parallel and comparable corpora in Eastern European languages made up of
   - *Multilingual Parallel Corpus: 1984* (7 x 100k words) is the multilingual parallel corpus, consisting of the English original and the translated data in the six languages of the project. The parallel data chosen for the project was the novel „1984" by George Orwell.
   - *Multilingual Comparable Corpus: Fiction* – first part of the comparable corpus (6 x 100k words) this part is composed either from excerpts of novels or of collections of short stories.
   - *Multilingual Comparable Corpus: News* – the second part of the comparable corpus (6 x 100k words) composed of newspaper articles from the six countries of the project.
   - *Multilingual Speech Corpus: EUROM* – a small parallel speech corpus. For this corpus, a sample (200 sentences) of the English part of the EUROM1 multilingual speech database was selected. This text was translated into the six languages and,

except for Bulgarian and Czech, recorded by one male speaker and digitised in accordance with the EUROM recommendations.

7. ET10-63 CORPUS. A bilingual parallel corpus of English and French, containing EC official documents on telecommunications. The ET10-63 corpus is part-of-speech tagged and lemmatized, approximately 1,250,000 words of each language.
8. The International Communication Union (ITU) or CRATER Corpus. A trilingual corpus of Spanish, French and English made up of texts from telecommunication domain, is part-of-speech tagged, approximately 1,000,000 words.
9. The PIXI Corpora consist of 450 naturally occurring conversations recorded in bookshops in England and Italy, for the purpose of cross-cultural comparisons of discourse structure. They are available in electronic form from the Oxford Text Archive, and in book form in Gavioli & Mansfield (1990), together with careful details of the data gathering, discourse contexts, analytic approach and bibliography of related publications.
10. PEDANT – the parallel texts in Göteborg. PEDANT consists of texts in several languages and aims at providing a wide collection of text types and language pairs in order to facilitate the creation of sub-set corpora for the specific purposes various researchers might have. Developed by Pernilla Danielsson and Daniel Ridings. Searches, resulting in something that could be likened to a parallel concordance, can be done in Swedish, English, French and German.

3.2. Plain vs. annotated

Perfectly plain: produced by scanning; no information about text (usually, not even edition) – e.g. Project Gutenberg texts (available at http://promo.net/pg/), all text archives can be considered plain corpora.

Annotated for part of speech, syntactic structure, discourse information, etc. – British National Corpus, Bank of English.

3.3. Resources of Corpora

The variety of resources can be roughly divided into written and spoken language data. First, we will have a look at speech corpora.

3.3.1. Speech corpora

The Centre for Spoken Language Understanding (CSLU) at the Oregon Graduate Institute is one of the most famous for its collection, annotation and distribution of telephone speech corpora. The Centre's activities are supported by its industrial affiliates, but the corpora are made available to universities world-wide free of charge. Overviews of speech corpora available from the Centre, and current corpus development activities, can be found in: Multi-Language Corpus (also available through the LDC).

Europe is by nature multilingual, with each country having their own language(s), as well as dialectal variations and lesser used languages. Corpora development in Europe is thus the result of both national efforts and efforts sponsored by the European Union (typically under the ESPRIT (European Strategic Programme for Research and Development in Information Technology), LRE (Linguistic Research and Engineering), and TIDE (Technology Initiative for Disabled and Elderly People) programs, and now for Eastern Europe under the PECO (Pays d'Europe Centrale et Orientalle)/Copernicus programs).

Below is the list of speech corpora:

1. SPIDRE Corpus – Recorded Telephone Conversation, The Map Task Corpus is in the form of 8 CD-ROMs containing linked audio and transcriptions of a total of about 18 hours of spontaneous speech that was recorded from 128 two-person conversations. Corpus details 64 different speakers, 32 female, 32 male, all adults, each took part in four conversations in a quiet recording studio. They were all students at the University of Glasgow, 61 of them being native Scots. The conversations were carried out in an experimental setting in which each participant has a schematic map in front of them, not visible to the other. Each map was comprised of an outline and roughly a dozen labelled features (e.g. „a white cottage", „an oak forest", „Green Bay", etc.). The task was for the participant without the route to draw one on the basis of discussion with the participant with the route. In addition to the conversations, each speaker provided a wordlist reading, consisting of the major vocabulary items contained in the conversations. The conditions of the conversations were balanced: In half of them the speakers were strangers, in half friends; in half of them the speakers could see each other's faces, in half they could not. According to the authors the total corpus runs to about 18 hours of speech, with the transcripts consisting of around 150,000 word tokens drawn from just over 2,000 word form types. Transcription is at the orthographic level, and it includes filled pauses, false starts and repetitions, broken words, etc. Transcripts were connected to the acoustic sampled data by sample numbers marked every few turns. Transcriptions are provided for each conversation, marked up with TEI-compliant SGML (Standard Generalised Markup Language – a formal language that describes the relationship between a document's content and its structure).
2. SWITCHBOARD Telephone Speech Corpus – a collection of approximately 2,430 spontaneous conversations among 543 speakers, includes 302 males and 241 females from American English.
3. The TRAINS Spoken Dialogue Corpus – distributed by the Linguistics Data Consortium – a corpus of task-oriented spoken languages made up of 55000 transcribed words, with 98 dialogues collected using 20 different speech tasks and 34 different speakers.
4. ATC Air Traffic Control Corpus .
5. TI-DIGITS corpus (Texas Instruments Speaker-Independent Connected-Digit Corpus), recorded in 1984, has been (and still is) widely used as a test base for isolated and connected digit recognition.

In the United States, the development of speech corpora has been funded mainly by agencies of the Department of Defense (DoD). Such DoD support produced two early corpora:

- Road Rally for studying word spotting,
- King Corpus, for studying speaker recognition, excludes poetry and drama.

6. The Corpus of Spoken American English (CSAE) – is a database of one million words of spoken American English, encompassing a wide range of spoken language types (Chafe, Du Bois, &Thompson, 1992). The corpus is disseminated as widely as possible in several formats, including a printed book and an interactive computer format that will allow simultaneous access to transcription and sound. The creation of the Corpus of Spoken American English will be co-ordinated with the ICE project, of which the CSAE is the officially designated representative for the United States.
7. TIMIT Acoustic-Phonetic Continuous Speech Corpora TIMIT – a phonetically transcribed corpus of read sentences used for modelling phonetic variabilities and for

evaluation of phonetic recognition algorithms, and task related corpora such as Resource Management (RM), funded by the Advanced Research Projects Agency (ARPA) of the DoD

8. British English: WSJCAM0, Bramshill, SCRIBE, and Normal Speech Corpus;
9. Scotish English: HCRC Map Task;
10. Dutch: Groningen;
11. French: BREF;
12. German: PHONDAT1 and PHONDAT2, ERBA and VERBMOBIL;
13. Italian: APASCI;
14. Spanish: ALBAYZIN;
15. Swedish: CAR and Waxholm;
16. The Translanguage English Database (TED) – a corpus of multi-dialect English and non-native English of recordings of oral presentations at Eurospeech'93 in Berlin. TED speeches contain data ranging in style from read to spontaneous, under varying degrees of stress. An associated text corpus TED texts contain written versions of the proceedings articles, which can be used to define vocabulary items and to construct language models. Two auxiliary sets of recordings were made: one consisting of speakers recorded with a laryngograph (TEDlaryngo) in addition to the standard microphone, and the other a set of Polyphone-like recordings (TEDphone) made by the speakers in English and in their mother language. This corpus was partially funded by the LRE project EuroCocosda.

3.3.2. Written corpora

1. The Brown University Corpus – made up of approximately 1,000,000 words of American English dating from 1960.
2. The Kolhapur Corpus of Indian English – made up of approximately 1,000,000 words of Indian English dating from 1978. Its texts were selected from the same text categories as the Brown Corpus and is available from ICAME.
3. Longman-Lancaster Corpus – made up of approximately 14,5 million words of English from various geographical locations but mainly British and American. Begun in 1985, it contains varied stylistic levels and text types, and is intended for lexicographic and academic research.
4. The Corpus of English-Canadian Writing – consists of 3 million words of Canadian English magazines, books and newspapers collected in 1984 and representing a variety of genre categories.
5. The Macquarie (University) Corpus (Peters, 1987) consists of 1 million words of Australian English and is intended to be comparable to the Brown Corpus. It was compiled at Macquarie University, Australia.
6. The American Heritage Intermediate Corpus (Carroll, Davies, &Richman, 1971) consists of over 5 million words of written American English from the most widely used books in grades 3 through 9. It was compiled as a database for the American Heritage School Dictionary.
7. The Birmingham Collection of English Text (BCE) (Runoff, 1984,1987; Sinclair & Kirby, 1990), compiled from 1980-1985 by J. Sinclair,A. Renouf, and J. Clear, contains 20 million words of written (18.5) and spoken (1.5) language (mostly British) used in producing a series of Collins COBUILD reference and teaching works. It also

contains 20 million words of speech from a public inquiry including the complete transcripts of the 18- month-long inquiry into the plan for constructing the Sizewell nuclear power station. It is intended to be representative of modern British English and therefore consists of samples of current and general usage (rather than technical use), from adult speakers without regional dialects.

8. The Bellcore Lexical Research Corpora (Walker, 1987) were compiled to support corpus linguistics and computational lexicography research. They include text bases of 200 million words of newswire text (New York Times, Associated Press), 50 million words of magazine and journal articles, a collection of English machine-readable dictionaries and another machine-readable reference books, electronic-mail digests, and assorted smaller texts.
9. The Melbourne-Surrey Corpus – made up of 100,000 words of Australian newspaper texts, available from ICEMAN.
10. The Nijmegen TOSCA Corpus (Oostdijk, 1988) is a text bank of 75 works (1.5 million words) of educated written British English drawn from a variety of genres meant to be read rather than spoken (i.e., excluding poetry, plays and speeches), compiled for studies of linguistic variation).
11. The Warwick Corpus is approximately 2.5 million words of written British English (letters, fiction and other genres) compiled by J. M Gill for use in research aimed at the automatic generation of Brailleby computer (available from the Oxford Text Archive).

3.4. Language States: synchronic vs. diachronic

Majority of corpora are synchronic. Most of linguistic data in a machine readable form comes from present-day language. Therefore, there are only a few diachronic corpora to be found, for example:

- The Helsinki Corpus of Historical English – made up of 1,500,000 words from law, handbooks, science, trials, sermons, diaries, documents, plays, private and official correspondence, contains samples from texts covering the Old, Middle, and Early Modern English periods. The Helsinki Corpus is available on the ICAME CD-ROM.
- The Lampeter Corpus of Early Modern English Tracks – made up of approximately 1,000,000 words of English pamphlet literature from years 1640-1740, part-of-speech tagged and lemmatised.

3.5. Treebanks

Treebanks are databanks of text containing part of speech tags and labelled constituent structures (e.g., noun phrase, adverbial phrase, co-ordinate clause) (J. Edwards, 1998).

At present there are many treebanks available for public use, for example:

1. Treebanks available from the Lancaster University Centre for Computer Corpus Research on Language :
   - The Lancaster-Leeds Treebank: A manually parsed subsample of the LOB corpus showing the surface phrase structure of each sentence. Approximately 45,000 words taken from all the genre categories of the LOB corpus.
   - The Lancaster Parsed Corpus (LPC): A subsample of the LOB corpus, parsed by computer and manually corrected by several researchers. Approximately 140,000 words with samples from each of the 15 categories in the LOB corpus.

- The American Printing House for the Blind Treebank (APHB): A skeleton-parsed corpus of a wide range of English texts. 200,000 words.
- The Associated Press Treebank (AP): A skeleton-parsed corpus of American newswire reports. 1,000,000 words.
- The Canadian Hansard Treebank: A skeleton-parsed corpus of proceedings in the Canadian Parliament. 750,000 words.
- The IBM Manuals Treebank: A skeleton-parsed corpus of computer manuals. 800,000 words.
- The Anaphoric Treebank: A subsample of the AP corpus, annotated to show the reference of pronouns and lexical cohesion. Approximately 100,000 words.
- The Market Research Corpus: A corpus of approximately 1,500,000 words of in depth market research interview transcripts (from the ACAMRIT project). The data have been tagged for part of speech and word sense, but only about 10% of the corpus has been manually examined.

2. ICAME Treebanks: Treebanks available from International Computer Archive of Modern and Medieval English in Bergen, Norway.
   - LOB Corpus
   - The Lancaster Parsed Corpus (LPC)

3.6. Language types: one type against variety of the same language

Such a corpus is a collection of data necessary for comparative studies of varieties of the same language throughout the world, for example:

1. International Corpus of English which consists of spoken and written material produced after 1989. Each corpus contains one million words half from spoken and half from written language. The components of ICE are:
   - Corpus of Canadian English
   - Corpus of American English
   - Corpus of Caribbean English
   - Corpus of British English
   - Corpus of Irish English
   - Corpus of Nigerian English
   - Corpus of Ghanaian English
   - Corpus of South African English
   - Corpus of Cameroonian English
   - Corpus of Kenyan English
   - Corpus of Tanzanian English
   - Corpus of Malanian English
   - Corpus of Sierra Leone English
   - Corpus of Indian English
   - Corpus of Hong Kong English
   - Corpus of Singapore English
   - Corpus of Philippinian English
   - Corpus of Australian English
   - Corpus of New Zealand English.
2. The Cornell corpus (Hayes 1988; Hayes & Ahrens, 1988) is a 1.6 million word corpus, consisting of 1151 written or spoken British and American English texts, representing

a wide variety of language types. It was compiled in the 1980»s for a study on lexical adaptation of parents to children. The spoken samples range from abortion debates to the Patty Hearst trial to television situation comedies. It is available from the CHILDES archive.

3.7. Specialised corpora

Specialised corpora are based on texts in narrowly defined media or highly specific domains such as: native speaker vs. learner corpora or children corpora.

3.7.1. Native speaker vs. learner (e.g. corpora of learner compositions)

Learner Corpora are of much interest and value in the teaching of English for example:

- the Longman Learner's Corpus (first assembled in the late 1980s);
- the International Corpus of Learner English (ICLE), centralised in Louvain, contains over 2 million words of writing by learners of English from 14 different mother tongue backgrounds and is the result of collaboration between a large number of universities internationally.

3.7.2. Children Corpora

The CHILDES database at Carnegie-Mellon at University of New Mexico The Child Language Data Exchange System (CHILDES) (MacWhinney, 1991;MacWhinney & Snow, 1vs. imaginative prose, or spoken vs. written language) – the main archive for child language data.

A collection of utterances of children of different age groups. The total size of the database is approximately 150 megabytes. The corpora are divided into six major directories: English data, non-English data, story-telling or narrative data, data on language impairments, data from second language acquisition, and data not transcribed. (MacWhinney, Brian, *The CHILDES project*, Lawrence Erlbaum Associates, pp.280, 1995.)

4. Characteristics of a corpus

We will discuss only a few features of a corpus, mainly its representativeness, size, and form. These seem to be most general and applicable to any corpus available.

4.1. Representativeness

According to many linguists the truly representative corpus is a great aid in lexicographic work – the reliable statistics help to make linguistic judgements that support the final entry for a word in the dictionary. The question of representativeness really depends what we need to represent. Moreover, different corpora will provide the lexicographer with different frequency counts. Therefore, it is not very easy to make a corpus representative. Although some corpus linguists may have ready definitions of corpus representativeness it seems to be a bit far-fetched to formulate them as Tony Berber Sardinha did:

A representative corpus should include the majority of the types in the language as recorded in a comprehensive dictionary. Thus:

- *assuming that a dictionary entry is analogous to a type;*
- *dictionary x is comprehensive*

– *dictionary x has 100,000 entries*
– *a majority is 1/2 + 1*

A representative corpus would need to have as many tokens as necessary to include 50,001 types.

Is it really as simple and straightforward a matter as that? What if we want the 50,0001st word to reach some larger threshold frequency — say 5 — because we need to make statements about its typical usage, and one instance may not be enough). I do not question the idea of calculations performed on the basis of collected language data. There is always a need for a certain base to rely on, but calculations performed by Sardinhba seem to be a bit over-generalised and we cannot be confident that the results of such generalisations are the correct ones.

We also ecounter the problem of multi-word units. We need to decide whether to count different forms of the same word as instances of the same type.

According to Michael Klotz (from the internet discussion list):

...the basic type-unit is not the lemma but what Cruse calls the lexical unit, i.e. „a lexical form with a single sense". This is all the more important, since different lexical units that share a lexical form can behave differently e.g. with regards to subcategorisation. For example, there is „be friendly to" (i.e. behave in a friendly way) and „be friendly with" (i.e. be friends with). In a representative corpus you would want to make sure that both senses of „friendly" are covered. Once you take meaning into account, your estimate will be much higher of course.

What Klotz wants is a satisfactory representation of certain multi-word combinations. To represent all the multi-word units in the language clearly requires an infinite corpus. To make progress one would need to specify which multi-word units are of interest, which in most cases would be very hard and arbitrary. Let us a have a look at the way Addison Wesley Longman team tried to handle the issue of representativeness of general corpus. They perceived 'general language' as being reducible into distinct text types which have been conceptualised as amalgams of:

– *subject area in the written corpora (fiction, politics, science, poetry)*
– *medium (books, periodicals, tapes...)*
– *level (high, medium or low in written corpora)*
– *context (Educational, Leisure, Business, Public or Institutional in spoken corpora)*

They do not mention in what proportion the textual and nontextual materials have been compiled. This seems, however, of paramount importance because inappropriate proportions may distort and falsify the ultimate results of lexicographic analysis.

4.2. Finite size

According to Tony McEnery and Andrew Wilson (Tony McEnery & Andrew Wilson; 1998) 'the term 'corpus' also implies a body of finite size'. However, they acknowledged the fact that it is not always so which can be easily illustrated by the COBUILD corpus which is a ***monitor corpus*** that is a growing, non-finite collection of texts- as Sinclair's team call it 'an open-ended entity' since new texts are being constantly added to it. The Word Bank of Australian English is another example of monitor corpus. It contains written and transcribed spoken texts which are representative of the varieties of English used by Australians from all sections of the community.

According to Chris Brew (the internet discussion list), the larger corpus the better. For a long time corpora like LOB and Brown were considered large with 1,000,000 words. The BNC is now large at 100,000,000 words. We can create ad-hoc corpora larger than this from electronically available texts.

Monitor corpus is intended to be not finite or temporally bounded but rather gaining and losing texts over time in parallel with the fluidity of the language itself, one of its functions is to *monitor* linguistic change and innovation whereas *reference* corpora have fixed composition.

(Sinclair, 1982; 1992). Consequently we can differentiate two types of corpora:

- *static (closed) corpus* – of definite size, of finite number of words, e.g. the Brown Corpus
- *dynamic (monitor, open-ended) corpus* – constantly changing in size and updated by addition of new texts

According to Michael Rundell (from the internet discussion list), the size of corpora is related to two different approaches to corpus construction:

- *opportunistic approach* – consists in collecting all data that can be easily got hold of (Bank of English) The ECI disk falls into this category. Also Canadian Hansard and other documents of public record. Old text which is out of copyright (Conan-Doyle, Shakespeare, Bible ...). In short, all texts available free of charge with out of copyright status can make up an opportunistic corpus.
- *principled approach* – only the texts meeting design requirements are collected (Brown Corpus)

It is usually more difficult to get hold of spoken data and majority of corpora contain written language which is more easily available.

4.3. Machine readable vs. print form

Machine readability is characteristic of modern corpora – as it makes corpora searchable and easy to manipulate and update at speed. As far as print is concerned, we can still have a book version of Corpus of English Conversation (Svartvik and Quirk 1980).

5. European Centres of Corpus Linguistics

1. The International Computer Archive of Modern English (ICAME) at the University of Bergen in Norway.
2. Corpus Linguistics at the University of Birmingham
3. The University of Birmingham and Collins Publishers COBUILD Bank of English
4. The European Corpus Initiative at Edinburgh University
5. The HCRC Map Task Corpus at Edinburgh University
6. The International Corpus of English at the Survey of English Usage, University College London
7. The Machine Readable Corpus of Spoken English (Universities of Reading and Leeds)
8. The SUSANNE Corpus – contains annotations of a 130,000-word cross-section of written American English (it is based on a subset of the million-word Brown Corpus). The genesis of the SUSANNE scheme lay in work on statistics-based parsing techniques led by Geoffrey Leech and Roger Garside in the early 1980s at Lancaster University.

The 45,000-word Lancaster-Leeds Treebank which Geoffrey Sampson developed for Geoffrey Leech and Roger Garside's parsing project, though small, was apparently the first in the field, the very term treebank being coined by them

9. Corpus holdings at the University of Lancaster
10. Bergen Corpus of London Teenage Language (COLT)
11. Corpus of Written British Creole
12. The Lingua Parallel Concordancing Project
13. RELATOR (European Linguistic Resources Repository Network)
14. ShATR: A Speech Science Corpus
15. English-Norwegian Parallel Corpus
16. English-Turkish Aligned Parallel Corpora
17. The Standard Corpus of Everyday English Usage. A multimedia project.

6. North American Centres of Corpus Linguistics

1. The Linguistic Data Consortium. The DARPA-funded Linguistic Data Consortium (LDC) was inaugurated in the Spring of 1992. Its formation was stimulated by the establishment of the Data Collection Initiative (DCI) of the Association for Computational Linguistics (ACL), but also strongly influenced by co-operative work in the speech community that led to the development of corpora consisting of digits and of acoustic-phonetic data pronounced by multiple speakers. Based at the University of Pennsylvania, the LDC includes more than 100 companies, universities and government agencies. One of the LDC's goals is to provide the U.S. sponsored databases to foreign researchers and to help to negotiate arrangements for general access by the U.S. researchers to resources from abroad. The LDC is intended to develop and distribute large amounts of linguistic data (e.g., speech, text, lexicons, and grammars) to assist the development of speech- and text-processing systems. The data includes large quantities of raw and annotated (i.e., syntactically and/or semantically tagged) text and speech (billions of words of text and thousands of hours of speech), a large lexicon, and a broad coverage grammar of English. The data also includes whatever additional materials (including foreign language materials) the Consortium can obtain by exchange or on other reasonable terms. Data are to be provided on CD-ROM on a subscription basis to universities and corporations. Although the Consortium does not need exclusive rights to donated data, DARPA does intend to make its growing holdings available exclusively through the Consortium. General membership fees are set at affordable levels, and foreign members are considered if access to foreign data can be assured. The Consortium may be established as a separate legal entity, such as a non-profit corporation or other form of association. Department of Linguistics, University of Pennsylvania, Philadelphia
2. The Penn-Helsinki Parsed Corpus of Middle English
3. The TRAINS Spoken Dialogue Corpus
4. Air Traffic Control Corpus (ATC Corpus)
5. SPIDRE Corpus – Recorded Telephone Conversation
6. ARTFL: American and French Research on the Treasury of the French Language
7. MLTS: Multilanguage Telephone Speech Corpus

7. Final Remarks

The variety of corpora available to lexicographers and other linguists is growing every minute. This is due to the technological development and advanced information systems which we use in data collection, information transmission and processing. This includes optical character recognition, text retrieval and understanding. The invention of the computer is undoubtedly a turning point in the field of corpus linguistics. Nowadays we can store and retrieve huge amounts of texts quickly on the screen. Massive volumes of information are accessible at low cost in a machine readable form – the most popular form of a modern corpus. Corpus linguistics regained its strong position in linguistics. Now, that our corpora are more precise and flexible than a couple centuries ago the Chomsky's criticism levelled at the practicality of corpus linguistics is no longer valid. At present corpus linguistics seems to be flourishing as it is applicable in lexicography, natural language processing and other branches of linguistics which also have substantial commercial value. The results of research carried out by corpus linguists are taken into consideration in foreign language teaching and dictionary compilation. It will never be possible to compile a corpus which will comprise all language senses as such. The infinite nature of the language cannot be pinned down in a form of sloppy disk or CD-ROM. However, we can try to describe and define its senses and meanings with more accuracy than ever before.

References[1]

Ahmad K., Corbett G. (1987). The Melbourne-Surrey Corpus. *ICAME Journal* 11, 39-43.

Altenberg B. (1990). Spoken English and the dictionary. In: J. Svartvik (ed.), *Directions in corpus linguistics: Proceedings of the Nobel Symposium 82*. New York: Mouton de Gruyter.

Altenberg B. (1991). A Bibliography of Publications Relating to English Computer Corpora. In: S. Johansson, A.B. Stenström (Eds.), *English computer corpora: Selected papers and research guide*. New York: Mouton de Gruyter.

Atkins B.H., Clear J., Ostler N. (1992). Corpus Design Criteria. *Literary and Linguistic Computing* 7, 1-16.

Barlow M. (1999). Corpus Linguistics. Available at http: //www.ruf.rice.edu/~barlow/corpus.html

Boguraev B., Briscoe T. (eds.) (1988). *Computational Lexicography for Natural Language Processing*. London: Longman.

Boguraev B., Briscoe T., Calzolari N., Cater A., Meijs W., Zampolli A. (1988). *Acquisition of lexical knowledge for natural language processing systems. Proposal for ESPRIT basic research activities*. Cambridge: Cambridge University Press.

Burnard L. (1991). What is SGML and how does it help? [Document No. TEI EDW 25]. Text Encoding Initiative listserver (available at http: //listserv@uicvm.bitnet).

Burnard L. (1995). *Users' Reference Guide to British National Corpus*. Oxford: Oxford University Computing Services.

[1] As most of the following materials come from the electronic resources (mainly Internet) it was impossible to give the exact pages which particular quotations came from. Instead any details on how and where they are available online have been given. All of the following is valid as through November 1st,1999.

Carroll J.B., Davies P., Richman B. (1971). *The American Heritage word frequency book.* Boston: Houghton Mifflin.

Chafe W. (1992). The Importance of Corpus Linguistics to Understanding the Nature of Language. In: J. Svartvik (ed.), *Directions in corpus linguistics: Proceedings of the Nobel Symposium 82* (pp. 79-97). New York: Mouton de Gruyter.

Chafe W., Du Bois J.W., Thompson S.A. (1992). Corpus of Spoken American English. Unpublished manuscript, Linguistics Department, University of California, Santa Barbara.

Crowdy S. (1993). Spoken Corpus Design. *Literary and Linguistic Computing* 8 (4).

Crystal D. (1991). *A Dictionary of Linguistics and Phonetics*. Blackwell.

Dawson J.L. (1977). Texts in Machine-readable Form and the University of Cambridge Literary and Linguistics Computing Centre. *CAMDAP* 7, 25-30.

Edwards J., Lampert M.D. (1993). *Talking Data: Transcription and Coding in Discourse Research.* Chapter 10*: Survey of Electronic Corpora and Related Resources for Language Researchers* (pp. 263-310). London and Hillsdale, NJ: Erlbaum.

Fisher W.M., Doddington G.R., Goudie-Marshall K.M. (1986). *Proceedings of the Speech Recognition Workshop*. Defense Advanced Research Projects Agency, Information Processing Techniques Office Report No. AD-A165 977.

Francis W.N., Kucera H. (eds.) (1979). *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English for Use with Digital Computers*. Providence, RI: Brown University, Department of Linguistics.

Francis W.N. (1982). *Problems of Assembling and Computerizing large corpora.* In: S. Johansson (ed.), *Computer corpora in English language research* (pp. 7-24). Bergen: Holtz and Schaeder.

Gavioli L., Mansfield G. (1990). *The PIXI Corpora: Bookshop Encounters in English and Italian*. Bologna: CLUEB.

Gellerstam M. (1992). Modern Swedish Corpora. In: J. Svartvik (ed.), *Directions in corpus linguistics* (pp. 149-163). New York: Mouton de Gruyter.

Gellerstam M. (ed.) (1988). *Studies in Computer-aided Lexicology.* Stockholm: Almqvist & Wiksell International.

Greenbaum S. (1992). A New Corpus of English: ICE. In: J. Svartvik (ed.), *Directions in corpus linguistics* (pp. 1761-179). New York: Mouton de Gruyter.

Hayes D.P., Ahrens M.G. (1988). Vocabulary Simplification for Children: A Special Case of Motherese? *Journal of Child Language* 15, 395-410.

Howe D. (1999). *Free On-line Dictionary of Computing*. Available at http: //wombat.doc.ic.ac.uk/foldoc and (or) http://foldoc/doc.ic.ac.uk/

Ihalainen O. (1987). The Helsinki Corpus of English Texts: Diachronic and dialectical-Report on work in progress. *ICAME Journal* 11, 58-60.

Johansson S., Atwell E., Garside R., Leech G. (1986). *The tagged LOB corpus: Users manual*. Bergen: Norwegian Computing Centre for the Humanities.

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: Department of English, University of Oslo.

Kennedy G. (1998). *An Introduction to Corpus Linguistics*. Longman.

Knowles G., Lawrence L. (1987). Automatic Intonation Assignment. In: R. Garside, G. Leech, G. Sampson (eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Kucera H. (1992). Brown Corpus. In: S.C. Shapiro (ed.), *Encyclopedia of Artificial Intelligence. Vol. 1* (pp. 128-130). New York: John Wiley & Sons.

Kucera H., Francis W.N. (1967). *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Linguistic Data Consortium (1996-1999). University of Pennsylvania, available at http://www.ldc.upenn.edu/

MacWhinney B. (1995) *The CHILDES project*. Lawrence Erlbaum Associates, pp.280.

McArthur T. (1992). *The Oxford Companion to the English Language*. Oxford: Oxford University Press.

McEnery T., Wilson A. (1998). *Corpus Linguistics*. Available at http://www.ling.lancs.ac.uk/pbbin/. Also published by Edinburgh University Press.

PC Webopaedia (1999). Internet.com Corp., Ketchum St. Westport, CT. 06880. Available at http://www.webopedia.com/

Project Gutenberg & PROMO.NET (1971-1998). available at http://promo.net/pg/

Rayson P., Needham C. (1993-1999). *University Centre for Computer Corpus Research on Language (UCREL)*. Lancaster University. Available at http://www.comp.lancs.ac.uk/computing/research/ucrel

Rundell M. (1996). The corpus of the future, and the future of the corpus. Talk at Exter special conference on *New Trends in Reference Science*. Abstract 21/4/96.

Rundell M., Stock P. (1994). The Corpus Revolution. *English Today* 8 (4).

Summers D. (1993). Longman Lancaster English Language Corpus Criteria and Design. *International Journal of Lexicography* 6 (3).

Sampson G. (1999). *SUSANNE Corpus and Analytic Scheme*. Available at http://www.cogs.susx.ac.uk/users/geoffs/index.html

Seaman D. (1999). *Goals and Missions*. Available at http://www.lib.virginia.edu/ecenters.html/. University of Virginia Library

Sinclair J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sperberg-McQueen C.M. (1994). *Bilingual corpora*. TEI-L

Summers D., Rundell M., et al. (1993). *Longman Language Activator*. Harlow: Longman.

Summers D., Rundell M., et al. (1995). *Longman Dictionary of Contemporary English*. Harlow: Longman.

The London-Lund Corpus of Spoken English: Description and Research (1980). Lund: Lund University Press.