Linguistic Corpus Analysis (Script for Download)

1. What is a corpus?

A language corpus can be defined as a collection of speech or writing in any given language or language variety. The size of a corpus can vary from a few speech samples to larger collections of written or spoken material, encompassing billions of words. Large corpora are usually put into an electronic format and they are prepared in such a way that an electronic search is possible. Very often, you can do a corpus search on-line. Many large corpora are freely available online, and we will introduce some of them in the course of this tutorial.

2. What does a corpus look like?

Large electronic corpora usually share the following characteristics:

1. They are presented in a written electronic format, i.e. spoken language samples are transcribed in one way or another. For example, you can find an ordinary orthographic transcription or you find phonetic transcriptions or even conversation-analytic transcriptions showing turn-organisation in conversations. In most cases, the transcripts are introduced by a general section giving basic information concerning the participants, location, languages used etc.
2. The language samples are analysed and tagged for various linguistic functions, such as word category, morphemes, syntax.
3. The language samples are usually also annotated. This means that for each utterance or sentence you find additional lines providing a phonological, morphological or syntactic analysis and sometimes even comments about situational aspects or non-verbal aspects of the conversation (e.g. gesture).
4. Electronic corpora provide analysis programs with specific commands that we can use to search the corpus for specific aspects we are interested in.

3. Corpus Types

There are different corpus types, allowing the researcher to analyse different aspects of language. First, there are corpora for different languages. For our purposes, we will mainly concentrate on English language corpora. In the poster, we provide the names of each corpus in the abbreviated version. A list with the full names and links to the corpus websites can be downloaded separately.

4. General Corpora of English

What we refer to as general corpora of English are corpora that are based on a large variety of spoken and written text types or genres. You can either conduct searches for the totality of texts, or narrow the search down for a specific text type if necessary. Examples of general corpora of English are the *Corpus of Contemporary American English* (COCA), the *British National Corpus* (BNC), or the *International Corpus of English* (ICE) with several sub-corpora for different varieties of English. These corpora represent contemporary English as they provide collections of texts from the last ten decades and are regularly updated.

5. Specialised Corpora of English

Specialised corpora can be used for more specific purposes. For example, diachronic corpora, such as the *Corpus of Historical American English* (COHA), enable the user to analyse linguistic changes over time because they comprise texts going back to earlier centuries. You can use them to study English in a particular period or at a particular point in time or to investigate how language has changed from one period to the other. Other specialised corpora provide samples of specific text types, such as the *TIME Magazine Corpus* based on articles from the *TIME* magazine, or the *Michigan Corpus of Academic Spoken English* (MiCASE) based on academic discourse, or even the *Corpus of American Soap Operas* (SOAP) based on scripted conversations from TV shows. Finally, you can find corpora specialised for varieties other than British and American English. The most extensive collection of varieties can be found in sub-corpora of the *International Corpus of English* (ICE). The list of varieties covered is constantly growing. If you are interested in global web-based English, you can consult the *Corpus of Global Web-based English* (GloWbE).
Most of the corpora we have mentioned so far are available online via a website provided by *Brigham Young University* (http://corpus.byu.edu/). All corpora provided via this website operate with the same search interface or analysis window. You can find tutorials for searching these corpora in section 11 in the poster.

6. Learner Corpora of English

Besides corpora of adult and native-speaker English, there are also corpora providing samples from first and second language learners. A large database that can be freely accessed is the *Talkbank* database (http://talkbank.org/), which comprises several sub-corpora such as the *Child Language Data Exchange System (CHILDES)* for first language acquisition, the *SLA Bank* for second language acquisition, and several others. You can find data for English as well as for other first and second languages there. Many data sets can be accessed online and are referred to as 'browsable database'. All data files can also be downloaded and analysed with an analysis programme (CLAN) that is also available for download on the website. In section 11 in the poster we provide tutorials introducing the search interface and frequent commands. The tutorials and materials are based on the CHILDES sub-corpus, but they are applicable to all the transcripts in the *Talkbank* database, since they all use the same transcript format and analysis program.
Another corpus for L2 English is the *International Corpus of Learner English* (ICLE). This corpus is not available online, but it is available in the English department of the *TU Braunschweig* and can be used by our students. A basic tutorial introducing the corpus and how to conduct searches is provided in section 6.3 on the poster.

7. Comparative corpora

Sometimes, we want to be able to compare languages by, for instance, checking whether a term is more or less frequent in one language than in the other. Since individual corpora often use particular tags, annotations, or frequency counts, and since they come in very different sizes, you cannot compare corpora freely with one another. However, there are some corpora available that present comparable data from several languages. For instance, the CELEX corpus offers frequency data from written English, German and Dutch. This corpus is also available online as WebCELEX (http://celex.mpi.nl/). There are also comparable corpora for

spoken language, available in the Subtitle Word Frequencies Corpora (SUBTLEX) that are based on subtitles from films and television series (http://crr.ugent.be/programs-data/subtitle-frequencies). You get data for American and British English, Dutch, Chinese, German, and many other languages.

8. Corpora of German

For some research projects, you might want to consult a corpus for German, for example when studying German learners of English or German-English bilinguals or when working on German and English contrasts. Examples of corpora for German are the DWDS corpus (das *Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts*) or the corpora hosted by the *Institut für Deutsche Sprache* (IDS), which are searchable on-line via the interface COSMAS II (http://www.ids-mannheim.de/cosmas2/). You can also check out the large corpus from the Leipizig-based *Projekt Deutscher Wortschatz* (http://wortschatz.uni-leipzig.de/).

9. Why use corpora?

After this short overview of corpus types, we will now turn to the very relevant question of why linguists use corpora in the first place. Corpora can either be used as the primary database for conducting analyses of language, or they can be used for identifying structures that are typical in language learning. For example, you can choose transcripts of ten English-speaking children at the age of two years from the CHILDES database and analyse their language in terms of the word categories they use. Alternatively, you could check the *British National Corpus* for typical words following, for example, the ditransitive verb *give*.
In other cases, you need a corpus not as a primary data source, but as a way to check and norm linguistic stimuli you use in experiments. Linguistic stimuli are words or sentences you use in experiments in order to elicit reactions from your participants. For most experiments, it is necessary to have stimuli that are similar in frequency, length and other aspects. Corpora allow you to choose words that fulfil the conditions required for your experiment, e.g. by running a frequency search.

10. Choice of corpus for research question

As you might have noticed, a particular research project requires the choice of a corpus that provides suitable data for investigating your research question. For choosing the right corpus, you have to think about various parameters. For example, if you are interested in historical aspects of language, do not choose a contemporary corpus. If you are interested in features of spoken language, choose corpora or sub-corpora that are based on spoken language samples. Choose a corpus of English L2 if your research project is based on L2 learning. For some research projects, a closer check of individual files is necessary. For example, if you want to investigate the speech of ten 2-year-old English children, you will probably choose the CHILDES corpus. However, it is also necessary to check whether the corpus provides transcripts of ten different children at the relevant age.

11. Tutorials: Searching Online Corpora

Once you have chosen a corpus that is suitable for your research project, you must get familiarised with the analysis tools of that corpus. We have linked tutorials for some of the corpora in the poster. These links are represented by the red plus buttons. The tutorials for ICLE and CHILDES are provided by LingTUtor. The tutorials for the online corpora provided by BYU (e.g. BNC, COCA) are from external sources.