2 – Introduction to Corpus Linguistics

Cedric

26-27 March 2012 · Bangor University
Korpora und DaF-Materialien – eine praktische Werkstatt

Translation of the (German-language) Workshop:
Corpora and Materials for German as a Foreign Language – a Practical Workshop

Slides designed by Dr. Cedric Krummes
C.Krummes@gmail.com
What is Corpus Linguistics?

“Corpus linguistics [...] makes use of data to gain insight into how language works.”
(Pérez-Paredes & Díez Bedmar 2009: 7)

Left photo: corpus linguistics can be seen as “pea counting”, a German expression for being very accurate.
Right photo: corpus linguistics is not arm chair linguistics, where people just think and reflect and make up their own examples
What is a corpus?

“We understand a corpus as a digitized collection of texts that have been collected and documented according to pre-established demands and specifications. These texts, whether spoken, written, or signed in one or several language varieties can be accompanied by metadata and linguistic annotation.”

(Ensslin & Krummes forthcoming, 2012)

Signed = sign languages transcribed; also includes any multi-modal language such as speech + movements

<table>
<thead>
<tr>
<th><strong>Breaking down a corpus</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A corpus is a</strong> (das Korpus)</td>
</tr>
<tr>
<td><strong>large</strong></td>
</tr>
<tr>
<td><strong>principled collection</strong></td>
</tr>
<tr>
<td><strong>of texts</strong></td>
</tr>
<tr>
<td><strong>in electronic form</strong></td>
</tr>
<tr>
<td><strong>analysed by software.</strong></td>
</tr>
</tbody>
</table>

In German, “Korpus” is neutral: “das Korpus” – and not “der Korpus”, despite the Latin “us”-ending.
Types of Corpora

1. sample (or reference) corpus
2. monitor corpus
3. parallel and comparable corpus
   (also referred to as a “translation corpus”)
4. specialised corpus
5. spoken corpus
6. learner corpus

(Breyer 2011 : 26)

Reference corpora, like the BNC, are very large corpora. They aim to reflect the language as broadly as possible and are carefully designed to include x% from various genres (e.g. newspaper, sermons, ads, literature) or so many words.
A monitor corpus shows us how words rise and decline over time. During the South African world cup, for instance, “vuvuzuela” became a fashionable word (and object) to use. A monitor corpus would be able to show the rise of that word and the decline too.

The German example, “geil”, used to just mean “lustful”, but its usage has increased since the 1980s for its new (slang) meaning of “great, cool”, etc.
Parallel corpora show you one text in several languages. Very useful for translators.
4. Specialised Corpus

Specialised corpora are smaller corpora with specific aims in terms of what kind of language they want to represent. In this case, the language is Luxembourgish emails.

http://cedrickrummes.org/lewc.php
Spoken corpora represent spoken language, i.e. unprepared, authentic language, usually dialogues. Spoken data is always transcribed and newer spoken corpora, like this German one, also provide a way to listen to the language. Language transcription can be quite loose (like simple fiction) or quite academic, adopting transcription conventions from conversation analysis.
A learner corpus aims to represent the language used by learners of a language. In this case, WHiG is interested in how British learners of German write German essays. Learner corpora can be augmented by a comparison corpus of native speakers performing the same task. In this case, the corpus Falko-L1 has native speakers of German writing the same essays in German.
There are three ways to use corpora in the classroom:

- Using ready-made textbooks (e.g., Touchstone) which contain corpus-informed information
- Getting the students to use corpus software
- Teachers using the corpus software to create corpus-informed paper handouts

Corpora in the Classroom

- **Corpus-Informed Materials**
  - ready-made
  - no input from teacher or learner
- **Data-Driven Learning ("DDL")**
  - using a computer
  - students become language explorers
- **Paper-Based Worksheets**
  - all the goodness of corpus software
  - all the safety of handouts
Corpus Insights

“language looks rather different when you look at a lot of it at once”
(Sinclair 1991: 100)

“[a] corpus does not contain new information about language, but the software offers us a new perspective on the familiar“
(Hunston 2002: 3)


Take the word “workshop”, for instance, all these words attached to it, these “collocations”, have been retrieved using corpus linguistics and corpus software. “Collocations” tell us which words prefer sticking to other words. In English, for instance, meat is “tough” and not hard, and men are “handsome” and not normally “beautiful”.

<table>
<thead>
<tr>
<th>category</th>
<th>verb</th>
<th>category</th>
<th>adjective</th>
</tr>
</thead>
<tbody>
<tr>
<td>verb + workshop</td>
<td>attend, organise, run, hold, lead, organise</td>
<td>adjective + workshop</td>
<td>one-day, two-day, hands-on, interactive, practical, intensive</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>category</th>
<th>... noun</th>
<th>category</th>
<th>noun ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>workshop + noun</td>
<td>facilitator, participant, session</td>
<td>noun + workshop</td>
<td>half-day, drama, craft, participatory, circus</td>
</tr>
</tbody>
</table>
There are three ways to create corpus materials:
- Start with a theory, from a grammar book or academic literature, for instance, then look up the phenomenon in a corpus.
- Work with your corpus, then see how that phenomenon fits in current academic journals or grammar books, for instance.
- Have your corpus examples run through a committee (e.g. editor, publisher, authors) and only keep the essence of it for a “light touch” textbook.

[Note: “bzw.” stands for German “beziehungsweise” and means in English “and/or”.]