A web-based concordance system for spoken language corpora

Knut Hofland
HIT Centre
University of Bergen
Allegt. 27
N-5007 Bergen
Norway

Telephone: +47 5558 9463
Fax: +47 5558 9470
E-mail: Knut.Hofland@hit.uib.no
URL: http://www.hit.uib.no/hit/hofland.htm

Keywords: spoken corpus, transcripts, Web concordance, digitized sound

1. Summary
At the University of Bergen we have developed a web-based environment for the study of spoken language corpora. This system has been used in several projects both with English and Norwegian material. The recordings (done with different kinds of recorders) are digitized and orthograpically transcribed. The transcripts and sound files are aligned. From a web concordance it is possible to play the part of the sound file that corresponds to the text in the concordance line. The concordance lines can then be classified by entering codes on the concordance line. These are saved and can be used as a basis for statistical analysis (cross tabulation). A link to a figure of the system is given in the reference section. The paper/poster will describe this system in more detail.

2. Digitizing and transcription
In some projects the transcription has been done from the original tapes (COLT). In other projects the recordings have first been digitized (and also improved) and the transcription has then been done by playing the sound files on the computer. Most of the time a standard player and word processor have been used. In some cases we have used the special player developed by the University of Santa Barbara (VoiceWalker). In some recent projects we have used the program system Praat, developed at the University of Amsterdam.

Most of the recordings were digitized by using a PC with a sound card and a sound editor like CoolEdit. The recordings were saved in WAV format on CD-R (before the availability of large and cheap hard disks) and later on a large disk system. For some recordings the digital signal was sent directly from the recorder (DAT or MiniDisc) to the sound card.

For each informant a set of demographic parameters, such as gender, age, and occupation were also registered.

3. Alignment of text and sound
To be able to play the sound from the concordance we need to know the time stamp for each word in the text. This was done differently for the different projects. The COLT files were sent to a company in England (SoftSound) for automatic text and sound alignment. Due to the quality of the recordings (they were done by equipping the informants with portable Walkmans) and a lot of concurrent speech, there were a great deal of errors in the time stamps. We have manually gone through the material and corrected many of these errors.

For the Norwegian recordings we have mostly done the alignment manually. In the transcripts we put a time mark for every ten seconds. For the words spoken between these marks we did an interpolation (divided the ten seconds by the number of words). For recordings containing few long pauses this was sufficient, but in some cases we also had to mark the time at other places in the transcript.
In some projects we used Praat. This program keeps track of the time codes for the beginning and end of each segment which is transcribed. The text and the time codes can be read by other programs and are converted to the main format used for searching (we had to interpolate and also re-sort the different "tiers").

Some corpora are distributed with text and sound alignment. An example is the Santa Barbara Corpus of Spoken American English. Each line has a start and end time code. In Santa Barbara they used their own program, SoundWriter, to add these codes to the transcripts in an interactive process (using a suggestion based on the rate of speech).

In general, 10-15 hours of work are required in order to transcribe and manually time align one hour of speech.

4. Tagging

Most of the spoken corpora were also tagged with part of speech tags. The COLT Corpus were tagged with the CLAWS tagger in Lancaster. The Norwegian material was tagged by the new Norwegian tagger which was originally developed in Oslo and has been further developed in Bergen. In Norway we have two written standards for what linguistically is considered as one language. The transcripts were done in the standard most similar to the dialect of the informant. But often the informants used a mixture of the two standards and this is a challenge for the tagger (which originally works on one of the standards at a time). In one project the user manually tagged a set of words with her own SGML-tags.

5. Searching

We used the program Corpus WorkBench from the University of Stuttgart for searching. This program uses a format of structural attributes (as COCOA mark-up) for information on the text and positional attributes (divided by tab characters) for each word. Words and punctuation marks are placed on separate lines. Each line has a column for time code (we only keep the reference in whole seconds) and additional columns for lemma, part of speech and information about each informant, such as gender, age group and home place. We have made CGI scripts for accessing the Corpus Query Program through the web. The user fills in a form in his web browser, giving information about words and additional information about the informant. The scripts return a concordance and each line has a link to an URL which extracts a sound fragment and returns this sound fragment either as a WAV or a MP3 file (the MP3 file is compressed with a factor of at least 10). On the server a special program that takes three arguments (a file, a start time and a duration) is used to extract the sound fragment. The user can save the sound file for processing in special purpose programs like Praat.

The user may filter away some of the concordance lines, classify the rest and then store the lines on the web server. Files can be merged and then basic cross tabulation may be done. The data can also be exported to be used in other programs.

6. Conclusion

The system has been used by researchers and students at the universities of Bergen and Oslo. They have got a valuable tool for the study of spoken language material. The Norwegian texts processed so far will be included in a future national spoken corpus.

7. References

Figure: http://www.hit.uib.no/knut/mons/tale-fig.gif
Example of search form: http://helmer.hit.uib.no/knut/sbcsae-test.html
Example of concordance: http://helmer.hit.uib.no/knut/sbcsae-snd.html
Praat program: http://www.fon.hum.uva.nl/praat/
SoftSound: http://www.softsound.com/
CWB: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/
COLT: http://helmer.hit.uib.no/colt/
Norwegian Spoken Corpus (pilot, in Norwegian):
http://gandalf.hit.uib.no/talemaalskorpus/Hovedside.htm