

Working with language corpora, 7,5 HP
 Spring 2012
 The Humanities Lab, Lund University

Lärare: Victoria Johansson, it-pedagog@humlab.lu.se

Schedule (19 Jan 2012)

There may be some changes regarding the literature (all, or most of it, will be available electronically).

Date	Teacher	Content	Literature
Mon 6 Feb 10-12 B054	VJ	Introduction What is a corpus? Different kind of corpora	[5] [9]
Wed 8 Feb 9-12 B054	SS & SL	Audio and video-recordings	[20]
Mon 13 Feb 10-12 H140	VJ	Ethical aspects Recruiting and anonymizing	Look at [1]; [2]; [3]
Wed 15 feb 9-12 B054	VJ	Transcription I CHAT Choice of transcription	[10]; [15]; [8] [20]
Wed 22 feb 9-12 B054	VJ	Transcription II CHAT	[20]
Mon 27 Feb 10-12 B054	VJ	Oral report on a thesis	[19]
Wed 29 feb 9-12 B054	VJ	Coding and analyses I CLAN	[6] [16] [20]
Mon 5 Mar 10-12 B054	VJ	Discussion of first draft of course paper	
Wed 7 Mar 13-16 time! B054	VJ	Coding and analyses II CLAN	[17] [20]
Wed 14 Mar 9-12 B054	FA	Metadata	[11]; [12]; [13] [7]
Wed 21 Mar 9-12 B054	MG	ELAN: transcription, coding and analyses	[18] [20]
Mon 26 Mar 10-12 H135b	VJ	Discussion of second draft of course paper	
Wed 28 Mar 9-12 B054	JW	Statistical analyses	[21]
10-13 Apr	VJ	Individual supervision of course paper	
Wed 25 Apr 9-12 B054	VJ	Oral report of course paper	

Teachers

Victoria Johansson	(VJ)	it-pedagog@humlab.lu.se
Susanne Schötz	(SS)	it-pedagog@humlab.lu.se
Stefan Lindgren	(SL)	stefan.lindgren@humlab.lu.se
Joost van de Weijer	(JW)	metodolog@humlab.lu.se
Maria Graziano	(MG)	maria.graziano@humlab.lu.se
Felix Ahlner	(FA)	felix.ahlner@ling.lu.se

About the course

The purpose of the course is to describe all the processes of collecting and organizing a corpus with linguistic material. The course provides possibilities for practical training in audio- and video-recording, in transcribing linguistic material, in coding and analyzing the material according to one's purposes. The course also gives theoretical and methodological perspectives on all the choices a researcher must make in order to build a corpus.

During the course we will mostly work with the transcription system CHAT and the coding program CLAN.

Doctoral level or master level?

The course has previously been given with the course code HTXA02, but is now given on the doctoral level. We will accept master students, if there is room in the course. All masterstudents must have explicit approval of their home department or MA-programme in the form of a signed copy of the form "Särskilt tillstånd att delta i kurser inom forskarutbildningen". (The form can be acquired from Victoria Johansson, it-pedagog@humlab.lu.se, if you need it).

Course requirements

To fulfill the course, the student is expected to take part in at least 5 of 7 laborations. The student are also expected to finish the task from the previous lab section before coming to the next, e.g. finishing transcribing an audiofile before the coding session etc.

Further, the student must choose a thesis (or similar) concerned with collecting, coding and/or transcribing data in one way or another, read relevant parts of it, and during an oral report evaluate the method and the studies of the thesis.

Finally, the student should write a course paper of approx. 10 pages. The paper can deal with any part of the course content. Some examples can be: describe a method for coding already transcribed data, or outline the collection of audio-data, or present a scheme for metadata, or discuss the difficulties with a specific analysis method.

Signing up

If you are interested in following in the course, please send an email to Victoria Johansson, at it-pedagog@humlab.lu.se, no later than 27 Jan 2012.

It will also be possible to only take part in the laborations, for those who want to know more about video recording, transcribing or something else.

References

- [1] Vetenskapsrådet. 2011 *CODEX. Rules and guidelines for research* <http://www.codex.uu.se/en/index.shtml>
- [2] Datainspektionen, Socialstyrelsen & Statistiska centralbyrån, 2008. *Personuppgifter i forskningen - vilka regler gäller?* <http://www.datainspektionen.se/Documents/faktabroschyr-pul-forskning.pdf>
- [3] Victoria Johansson, Roger Johansson & Åsa Wengelin. Körschema [opublicerat, 2010]. Fås som kopierat material.
- [4] Martin Wynne (ed.). 2006. *Developing Linguistic Corpora: a Guide to Good Practice*. <http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm>
- [5] John Sinclair. 2006. "Corpus and Text: Basic Principles" in Wynne, M. (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. <http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm>
- [6] Geoffrey Leech. 2006. "Adding Linguistic Annotation" in Wynne, M. (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. <http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter2.htm>
- [7] Lou Burnard. 2006. "Metadata for Corpus Work" in Wynne, M. (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. <http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter3.htm>
- [8] Paul Thompson. 2006. "Spoken Language Corpora" in Wynne, M. (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. <http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter5.htm>
- [9] John Sinclair. 2006. "Appendix to chapter one: How to make a corpus" in Wynne, M. (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. <http://www.ahds.ac.uk/creating/guides/linguistic-corpora/appendix.htm>
- [10] Catrin Norrby. 2004. *Samtalsanalys*, Lund: Studentlitteratur (Kap 6.)
- [11] IMDI Browser http://www.let.ru.nl/sign-lang/ECHO/IMDI/IMDI_intro.html
- [12] CLARIN project <http://www.clarin.eu/>
- [13] DAM-LR <http://www.mpi.nl/DAM-LR/>
- [14] Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates. 2000.
- [15] Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk - Electronic Edition. Part 1: The CHAT Transcription Format*. Carnegie Mellon University. 2007-08-07. <http://childes.psy.cmu.edu/manuals/chat.pdf>
- [16] Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk - Electronic Edition. Part 1: Volume 2: Transcription Format and Programs. Part 2: The CLAN Programs*. Carnegie Mellon University. 2008-04-19. <http://childes.psy.cmu.edu/manuals/clan.pdf>
- [17] Victoria Johansson. *Developmental aspects of text production in writing and speech..* Doctoral Thesis , Lund University, Lund Sweden 2009.

- [18] Hellwig, B. and Uytvanck, D. van. *ELAN - Linguistic Annotator* <http://www.mpi.nl/corpus/manuals/manual-elan.pdf>
- [19] (Parts of) A thesis that deals with any aspect of collecting, analyzing or coding a corpus.
- [20] Various tutorials for laborations.
- [21] Gries, S. T. (2009) *Quantitative Corpus Linguistics with R*. A chapter from this book.