

Introduction to Corpus Linguistics

Anke Lüdeling, Humboldt-Universität zu Berlin
anke.luedeling@rz.hu-berlin.de

Level: Basic

Prerequisites: Basic background in linguistics, no background in computational linguistics or corpus linguistics required

Credits: 3

Requirements for credits:

- Regular attendance.
- Mini-projects (done in groups) which will be presented orally (10 minutes). Additionally a short written summary is required (5 pages).

I will distribute the necessary corpora and search tools on CDs in the class.

Background reading: McEnery, Tony & Wilson, Andrew (2001) *Corpus Linguistics*. Edinburgh University Press, Edinburgh (2nd edition).

Schedule

	topics	supplementary reading
Aug 15, 2006	<ul style="list-style-type: none"> ○ different kinds of linguistic data: introspection, psycholinguistic and neurolinguistic experiments, field data etc.: where does corpus data fit in? for which research question can corpus data be used? ○ brief overview over the history of corpus linguistics 	<p>McEnery, Tony & Wilson, Andrew (2001) <i>Corpus Linguistics</i>. Edinburgh University Press, Edinburgh (2nd edition).</p> <p>Kepser, Stephan & Reis, Marga (eds): <i>Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives</i>. Mouton de Gruyter, Berlin</p>
Aug 17, 2006	<ul style="list-style-type: none"> ○ corpus design ○ pre-processing 1: tokenizing & tagging 	<p>Garside, Roger; Leech, Geoffrey & McEnery, Tony (eds) (1997) <i>Corpus Annotation: Linguistic Information from Computer Text Corpora</i>. Addison Wesley Longman, New York</p>
Aug 19, 2006	<ul style="list-style-type: none"> ○ pre-processing 2: lemmatizing, syntactic annotation, phonological annotation 	<p>Jurafsky, Daniel S. & Martin, James H. (2000) <i>Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition</i>. Prentice Hall, Upper Saddle River, NJ</p>
Aug 22, 2006	<ul style="list-style-type: none"> ○ annotation models: flat models vs. standoff models ○ evaluation: gold standards and inter annotator agreement 	<p>Mitkov, Ruslan (ed, 2003) <i>The Oxford Handbook of Computational Linguistics</i>. Oxford University Press, Oxford</p>
Aug 24, 2006	<ul style="list-style-type: none"> ○ evaluation of corpus data: qualitative and quantitative measures 	<p>Manning, Christopher & Schütze, Hinrich (1999) <i>Foundations of Statistical Natural Language Processing</i>. MIT Press, Cambridge MA</p> <p>Baroni, Marco (to appear) Distributions in text. In Anke Lüdeling and Merja Kytö (eds.) <i>Corpus linguistics: An international handbook</i>, Mouton de Gruyter, Berlin. Available online at: http://sslmit.unibo.it/~baroni/research.html</p>
Aug 26, 2006	<ul style="list-style-type: none"> ○ case study 1: language acquisition, corpora in language teaching, learner corpora 	<p>Granger, Sylviane (2002) A bird's-eye view of learner corpus research. In: Granger, Sylviane; Hung, Joseph; Petch-Tyson, Stephanie (eds, 2002) <i>Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching</i>. John</p>

		Benjamins, Amsterdam, 3-33 Nesselhauf, Nadja (2004) Learner corpora and their potential for language teaching. In: Sinclair, John (ed, 2004) <i>How to Use Corpora in Language Teaching</i> . John Benjamins, Amsterdam, 125-152
Aug 29, 2006	<ul style="list-style-type: none"> ○ case study 2: corpora from the Web 	<p>Lüdeling, Anke; Evert, Stefan & Baroni, Marco (to appear) Using Web data for linguistic purposes. In Marianne Hundt, Caroline Biewer and Nadja Nesselhauf (eds.), <i>Corpus linguistics and the Web</i>. Amsterdam: Rodopi.</p> <p>Baroni, Marco & Bernardini, Silvia (2004) BootCaT: Bootstrapping corpora and terms from the web. <i>Proceedings of LREC 2004</i>, Lisbon: ELDA. 1313-1316. Available online at: http://sslmit.unibo.it/~baroni/research.html</p>
Aug 31, 2006	<ul style="list-style-type: none"> ○ presentation of mini-projects ○ final discussion 	