Corpus Linguistics, class 1

Victoria Kamasa & Katarzyna Klessa

Empirical Linguistics & Language Documentation

Course information

Teachers: Katarzyna Klessa & Victoria Kamasa

Duty hours (Katarzyna Klessa):

- Tuesday, 11:30-12:30, Room 312a (building B Coll. Novum)
- Wednesday 12:45-13:15, Room 312a (building B Coll. Novum)

E-mail: klessa@amu.edu.pl
Website: katarzyna.klessa.pl

Course information - aims

- Introducing basic concepts of corpus linguistics
- Provide the basic knowledge on texts and speech corpora;
- Familiarize with tools to design, annotate and analyse texts and/or speech corpora;
- Demonstrate various uses of corpus data in linguistics.

Course information - expected outcomes

- understanding of basic notions of corpus linguistics;
- ability to design a text or a speech corpus appropriate for answering a given research question;
- ability to answer simple research questions using text and speech corpora;
- understanding of the role of corpus evidence in linguistic research;
- getting familiar with basic software for text and speech corpora analysis;
- knowledge of various possible levels of annotation, their aims, benefits and ways of applying them to text and speech corpora;
- development of teamworking abilities within a research project.

Course information – selected references

- Anthony, L. (2005). AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning, pp. 7-13, see also: http://www.laurenceanthony.net/software.html
- Baker, P. (Ed.). (2009). *Contemporary corpus linguistics*. London, New York: Continuum.
- Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.
- Bird, S., E. Klein, and E. Loper, *Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit*, see: http://www.nltk.org/book/ , http://www.nltk.org/ and (for beginners) Python course at: https://www.codecademy.com/learn/learn-python
- Boersma, P. & Weenink, D. (2013). *Praat: doing phonetics by computer [Computer program]*. Ver. 5.3.51, retrieved 2.06.2013 from www.praat.org
- Dimitriadis, A., & Musgrave, S. (2009). *Designing linguistic databases: A primer for linguists* (p. 13). Berlin: Walter de Gruyter.
- Klessa, K. (2015). *Annotation Pro [Software tool]*. Ver. 2.2.4.0. Retrieved from: annotationpro.org on 2015-05-19.
- McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. *Cambridge textbooks in linguistics*. Cambridge, New York: Cambridge University Press.
- McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction* (2nd ed.). *Edinburgh textbooks in empirical linguistics*. Edinburgh: Edinburgh University Press.
- Sloetjes, H., & Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), on-line: https://tla.mpi.nl/tools/tla-tools/elan/
- Viana, V., Zyngier, S., & Barnbrook, G. (Eds.). (2011). *Studies in corpus linguistics: v. 48*. *Perspectives on corpus linguistics*. Amsterdam, Philadelphia: J. Benjamins Pub.
- Warren Tang. (2011). *A Simple Guide to Using Antconc*. Retrieved from http://www.laurenceanthony.net/software/antconc/resources/help_AntConc321_english.pdf

Basic concepts

- corpus linguistics;
- language documentation;
- computational linguistics;
- ...

Language documentation

- *Language documentation (documentary linguistics), is the subfield of linguistics that deals with creating **multipurpose records of languages** through audio and video recording of speakers and signers and with annotation, translation, **preservation**, and distribution of the resulting materials.* (Oxford Bibliographies)
- *Language documentation is the **process by which a language is documented** from a documentary linguistics perspective. It aims "to provide a **comprehensive record** of the **linguistic practices** characteristic of a given speech community." Language documentation seeks to create as thorough a record as possible of the speech community for both **posterity and language revitalization**. Language documentation also provides a firmer foundation for linguistic analysis in that it creates a citable set of materials in the language on which claims about the structure of the language can be based.* (Wikipedia)

Computational linguistics

- *Computational linguistics is the scientific study of language from a computational perspective. Computational linguists are interested in providing* ***computational models*** *of various kinds of linguistic phenomena (...). Work in computational linguistics is in some cases motivated from a scientific perspective (...) and in other cases the motivation may be more purely technological.* (ACL: Association for Computational Linguistics)
- *… a very broad view of computational linguistics, covering diverse linguistic areas (...). Theoretical foci include* ***models*** *for parsing and learning grammatical structure, models of communication, conversation, and dialogue, computational psycholinguistics, and computational models of social interaction.* (Stanford)
- *Computational linguistics is an interdisciplinary field concerned with the* ***statistical or rule-based modeling*** *of natural language from a computational perspective, as well as the study of appropriate computational approaches to linguistic questions.* (Wikipedia).

Corpus linguistics

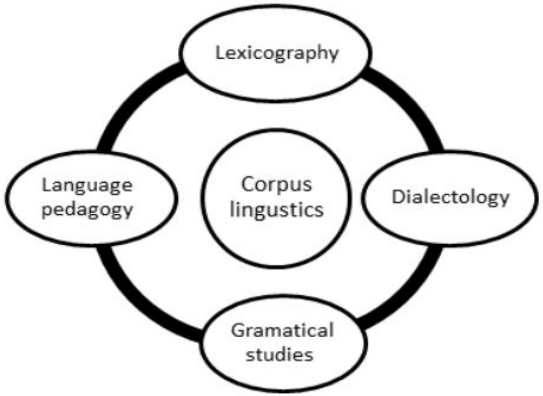
- *Corpus Linguistics is now seen as the **study of linguistic phenomena through large collections** of machine-readable texts: **corpora**.* (Uni. Essex)
- *Corpus linguistics is a **method** of carrying out linguistic analyses. As it can be used for the investigation of many kinds of linguistic questions and as it has been shown to have the potential to yield highly interesting, fundamental, and often surprising new insights about language, it has become one of the most wide-spread methods of linguistic investigation in recent years*. (Uni. Heidelberg)
- *Corpus linguistics is the **study** of language as expressed in **corpora** (samples) of "real world" text.* (Wikipedia)

Corpus linguistics – towards the digital era

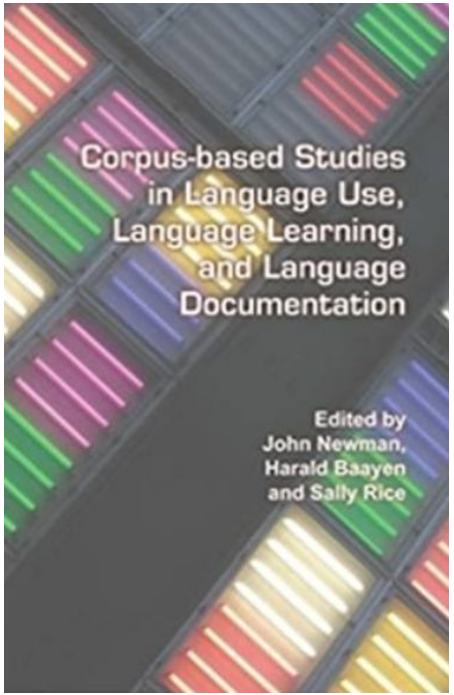
- The earliest corpus-based studies of religious / sacred texts (the Vedas, Quran, Bible);
- Corpus-based background for monolingual dictionaries;
- 1967: *Computational Analysis of Present-Day American English* based on the Brown corpus (Henry Kučera and W. Nelson Francis) as one of the first works in the digital era;
- Increasing interest in the domain as a consequence of the increasing importance of computer software capabilities
  - fundamental research in lexicography, discourse studies, sociology, phonetics...
  - Applications: Machine translation, Translation Studies, Contrastive Analysis, Speech and language technology...

The relationships between domains

- How are the domains / fields related to one another?
- What can be the possible profits of interdisciplinary studies?



https://phonetic-sciences.blogspot.com/2016/12/the-origin-and-history-of-corpus.html



What is a corpus? Why use it?

What is a corpus? Why use it?

https://www.futurelearn.com/courses/corpus-linguistics/5/steps/149215

Why use corpora?

- Possibility to analyze language “as it is”, in use, in practice - and to confront its real-life usage with theories, to verify new or existing hypotheses;
- Support from large amounts of data can make our claims more reliable;
- New discoveries about language in use not possible to be seen based on individual experience even by language specialists;
- Thanks to digitalization of resources, computer processing is possible which enhances (improves & speeds-up) data analysis -> “manual” work is prone to errors

Written vs. spoken corpora

?

Teamworking:

1. Try to define what is a spoken and written corpus.
2. What kind of knowledge can be derived from both types of the corpora?
3. What domains of linguistics can be associated with the two types of corpora?
4. What possible practical applications can be related to corpus studies with spoken / written corpora?

Report your answers in the slides here:

https://docs.google.com/presentation/d/1XxCPY4lLsqR64YFAMFRpOz6qznnx-yolT_1FoJDH0L4/edit?usp=sharing

General purpose / referential vs. specialized corpora

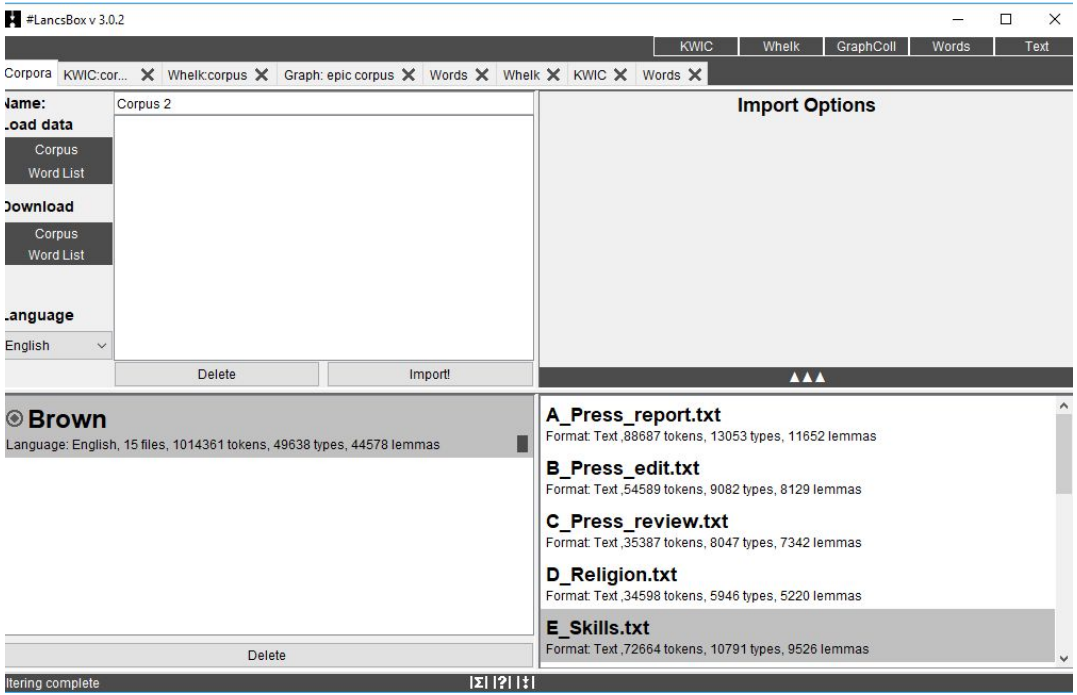
Teamworking (homework):

1. Try to specify differences between general purpose / referential and specialized corpora.
2. Find examples of such corpora.
3. Name the possible user groups for the corpora depending on their type.

Report your answers in the slides here (continued from the previous task):

https://docs.google.com/presentation/d/1XxCPY4lLsqR64YFAMFRpOz6qznnx-yolT_1FoJDH0L4/edit?usp=sharing

A toy :) LancsBox



http://corpora.lancs.ac.uk/lancsbox/help.php