Corpus Linguistics, class 2

Katarzyna Klessa

Empirical Linguistics & Language Documentation

Questions we can answer with various corpora

?

Questions we can answer with various corpora

- What are the most frequent words / phrases in a language?
- What are the differences between spoken and written utterances?
- What are the popular collocations, sequences of words?
- Is a word X more frequently used in speaking or in writing?
- How often do people use proper names in conversations?
- What is the typical number of words used by a native speaker and a non-native speaker of a language?
- How do native and non-native speakers differ in terms of the usage of idiomatic expressions?
- What is the size of vocabulary used by L2 (L3...) learners at particular education levels?
- ….

Other types of questions

- What is the organization of an utterance within conversation?
- What are the features of turn-taking in an interactive setting?
- How do interlocutors implement self-repairs in dialogues?
- How are the interlocutors' cognitive states or abilities displayed through their interactive utterances?
- What might be the reasons of certain quantitative observations (e.g. higher frequencies of certain words over another)?
- ….

Other types of questions: Conversation analysis

- What is the organization of an utterance within conversation?
- What are the features of turn-taking in an interactive setting?
- How do interlocutors implement self-repairs in dialogues?
- How are the interlocutors' cognitive states or abilities displayed through their interactive utterances?
- What might be the reasons of certain quantitative observations (e.g. higher frequencies of certain words over another)?
- ….

These questions require support from other methodologies, such as Conversation analysis.

Specialized corpora - examples

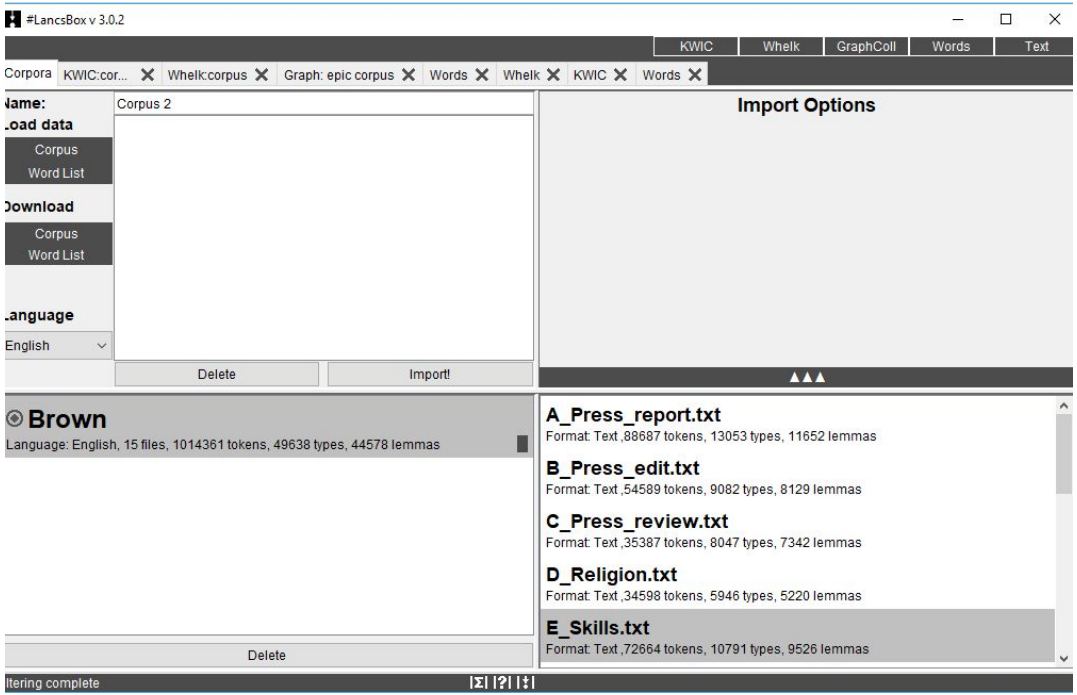
- Corpora used in forensic studies
- Corpora of specific types of texts: academic writing, newspapers, poetry -> potential usability in stylometry, text genre studies, other.
- Language teaching corpora
- Corpora of under-resourced languages can also be treated as specialized to some extent (lack of representativeness, specific age /social groups of speakers, etc.)
- ….

Extended reading tasks*

Please read the following texts (PDF files are available for downloads from the Student's Page here: http://elldo.amu.edu.pl/?page_id=620)

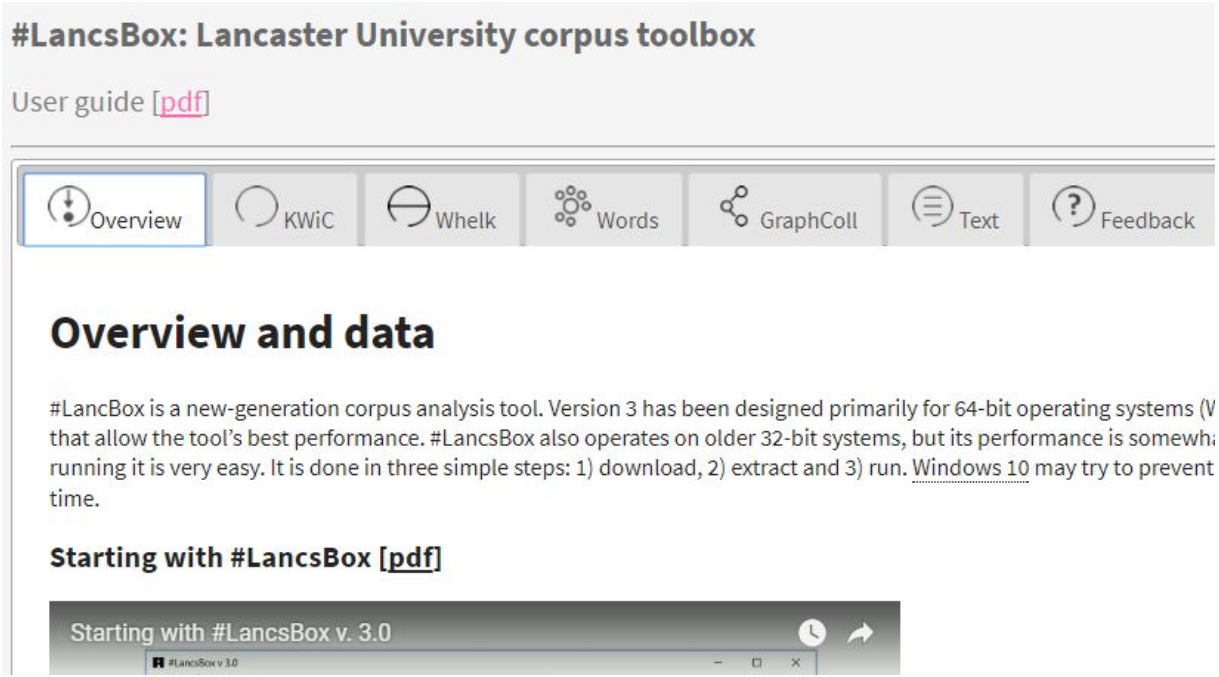
- *Applying corpus linguistics and conversation analysis in the investigation of small group teaching in higher education* (Anne O'Keeffe, Steve Walsh)
- *Why Forensic Linguistics Needs Corpus Linguistics* (Susan Blackwell).
- *How specialized are specialized corpora? Behavioral evaluation of corpus representativeness for Maltese* (Jerid Francom, Amy LaCross, Adam Ussishkin, available here)

LancsBox



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

LancsBox



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

LancsBox – key features

- The *KWIC* tool generates **a list of all instances of a search term in a corpus in the form of a concordance.**
- The *Whelk* tool provides information about **how the search term is distributed across corpus files**.
- The *Words* tool allows in-depth analysis of **frequencies of types, lemmas and POS categories** as well as comparison of corpora.
- The *GraphColl* tool identifies collocations and displays them in a table and as a **collocation graph or network**.

LancsBox - concordances

KWIC

The KWIC tool generates a list of all instances of a search term in a corpus in the form of a concordance.

- Find the frequency of a word or phrase in a corpus.
- Find frequencies of different word classes such as nouns, verbs, adjectives.
- Find complex linguistic structures such as the passives, split infinitives etc. using 'smart searches'.
- Sort, filter and randomise concordance lines.

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

LancsBox – key words

- **Concordance** - all instances of an item appearing in corpus, keywords in context
- **distribution of a search term** - ? http://corpora.lancs.ac.uk/lancsbox/help.php
- **frequencies of types, lemmas and POS categories** - ? http://corpora.lancs.ac.uk/lancsbox/help.php
- **collocation graph or network** - ? http://corpora.lancs.ac.uk/lancsbox/help.php
- insight into the context in which a word / phrase is used - ? http://corpora.lancs.ac.uk/lancsbox/help.php

http://www.laurenceanthony.net/software/antconc/

Laurence Anthony's Website

Home Resume Publications Software Classes Photo Albums Links Contact

AntConc Homepage

Latest Release

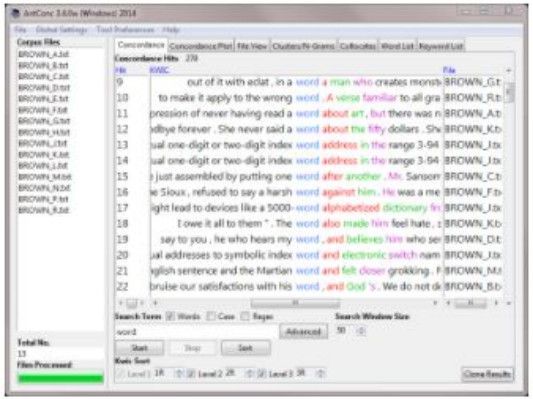
AntConc

A freeware corpus analysis toolkit for concordancing and text analysis.

[AntConc Homepage] [Screenshots] [Help]

Downloads:

- Windows (3.4.4)
- Macintosh OS X 10.7-10.12 (3.4.4)
- Macintosh OS X 10.6 (3.4.1)
- Linux (3.4.3)
- Older versions

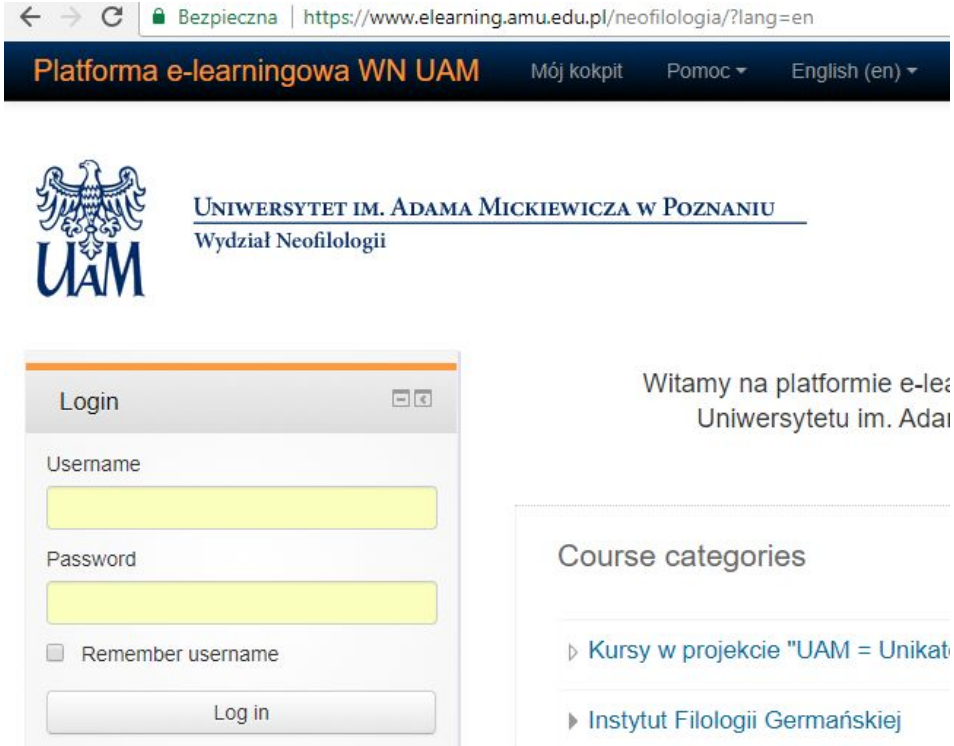


See also...

MOODLE: www.elearning.amu.edu.pl/neofilologia/

- Starting from next week (18.10) - a new task in Moodle
  - Please follow instructions in Moodle and contact us in case of any doubts or questions;
- Meeting in class on 08.11 - corpus annotation & tagging;
- Individual consultations on each Wednesdays.

MOODLE: www.elearning.amu.edu.pl/neofilologia/



MOODLE: www.elearning.amu.edu.pl/neofilologia/