Topics EMLex A6
Computational Lexicography: Corpus exploration for lexicographic purposes

Summer Term 2020, Evert / Heid / Kabashi

Monday, 04.05.2020

▷ 10:00 – 11:00 {*1h*} zoom

- General welcome
- Structure of the course (→ team projects, online teaching)
- *Get to know:* Background and interests of the participants

▷ 11:00 – 13:00 {*2h*}

- Reading: Lexicography and corpora, relevant phenomena moodle
  - readings of (potentially polysemous) items
  - authentic usage examples
  - typical syntactic and lexical contexts of words
  - collocations
  - idioms and their variation
  - terminology and specialized phraseology
- Screencasts: Unit B1.5 *Corpus linguistics & corpus search* moodle
- *Practice:* First steps in SketchEngine, following B1.5 examples
- Technical support via Zoom / Moodle forum

▷ 14:30 – 16:00 {*1.5h*}

- Reading: Corpus design moodle
  - Rundell, Michael and Atkins, B. T. Sue (2013). Criteria for the design of corpora for monolingual lexicography. In R. H. Gouws, U. Heid, W. Schweickard, and H. E. Wiegand (eds.), *Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume: Recent Developments with Focus on Electronic and Computational Lexicography*, volume 5.4 of *HSK*, chapter 96, pages 1336–1343. Mouton de Gruyter, Berlin, New York.
  - Atkins, Sue; Clear, Jeremy; Ostler, Nicholas (1992). Corpus design criteria. *Literary and Linguistic Computing*, **7**(1), 1–16.
- Issues to think about
  - representativity and balance
  - principles of corpus design, sampling frame, selection of texts
  - sources of corpus data
  - practical and legal issues of corpus compilation

▷ 16:00 – 18:00 {*2h*}

- Reading: How to sketch a team project moodle
- *Group work:* Suggest and discuss ideas for team projects among students; form teams of 2–3 students each; sketch topic & goals of your corpus-based dictionary
- Ask general questions in Moodle forum (answers from instructors by Tuesday)
- *Task:* Prepare one slide per team: title, members, important bullet points, possibly illustration → must be uploaded to Moodle in PDF format **by Tuesday 9:00**

Tuesday, 05.05.2020

▷ 10:00 – 11:00 {*1h*} zoom

- *Discussion:* Lexicographic objectives vs. corpus construction
  - students propose different types of interesting corpora for lexicography
  - then discussion on how to design and compile these corpora

▷ 11:00 – 12:30 {*1.5h*} zoom

- *Presentations:* Teams present ideas for class projects
  - general topic and goals, corpus design, expected analysis steps
  - one slide per team: title, members, key bullet points, possibly illustration
  - class project should involve (i) compilation and annotation of specialized corpus and (ii) its lexicographic analysis, typically in combination with larger existing corpora
- Feedback from instructors and discussion

▷ 12:30 – 13:30 {*1h*}

- Screencasts: Getting data from the Web moodle
  - principles and challenges of compiling Web corpora
  - searching vs. crawling vs. scraping
  - boilerplate removal, metadata extraction, normalization
  - searching with BootCaT
  - scraping with WebScraper (or Python/Scrapy)
  - collecting Twitter data (FireAnt, Python)
  - extracting text from PDF documents

▷ 15:00 – 18:00 {*2h*}

- Reading: Recommended software, Web interfaces, overview of corpora moodle
- Screencast: Regular expressions for search & substitution moodle
  - introduction to regular expressions (PCRE standard)
  - plain text format, character encodings, Unicode
  - applied to word lists and full-text search
- *Exercises:* Regular expression practice
- *Group work:* Corpus design & compilation for class projects (BootCaT, Web scraping, Twitter data, other sources)
- 16:00–18:00 Technical support and general questions via Zoom / forum

Wednesday, 06.05.2020

▷ 10:00 – 12:30 {*2.5h*}

- Reading: Metadata for corpora
- Reading: Linguistic annotation and pre-processing
  - tokenization
  - part-of-speech tagging (→ tagset)
  - lemmatization and morphological analysis
  - named entity recognition
  - syntactic analysis
  - WebLicht – an online portal for running corpus annotation tools
- Reading: Corpus search – the CQP query language
- Screencast: Short introduction to CQP user interface
- *Exercises:* Corpus queries in CQP UI
  - also try the same queries in SketchEngine on other corpora

▷ 14:00 – 15:00 {*1h*}

- Screencast: Corpus representation & exchange formats
  - XML annotation, "vertical text" format
  - uploading corpora to SketchEngine
- *Practice:* SketchEngine
  - crawling, uploading and annotating texts
  - corpus queries in SketchEngine
- Questions & support, also for corpus queries, via Zoom / Moodle forum

▷ 15:00 – 17:00 {*2h*}

- *Group work:* Continue work on class projects
- Further crawling, corpus compilation & cleaning, annotation
- First experimental corpus queries
- Questions & support via Zoom / forum

▷ 17:00 – 18:00 {*1h*}

- Online Q&A session with instructors
- Answers to questions on the forum
- Discussion of progress with team projects

Thursday, 07.05.2020

▷ 10:00 – 12:30 {*2.5h*}

- Quantitative analysis – collocations, keywords, term candidate extraction
- Reading: Lexicographic applications
- Reading: Mathematical background & implementation
  - Evert, Stefan (2013). Tools for the acquisition of lexical combinatorics. In R. H. Gouws, U. Heid, W. Schweickard, and H. E. Wiegand (eds.), *Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume: Recent Developments with Focus on Electronic and Computational Lexicography*, volume 5.4 of *HSK*, chapter 104, pages 1415–1432. Mouton de Gruyter, Berlin, New York.
  - [optional: detailed explanation of mathematics, read up to Sec. 4.3] Hardie, Andrew (2014). A single statistical technique for keywords, lockwords, and collocations. Internal CASS working paper no. 1, unpublished.
- Screencast: Collocations & keywords in SketchEngine

▷ 14:00 – 17:00 {*3h*}

- *Group work:* Analyze project corpora with SketchEngine or other software tools
  - corpus queries and reading of the concordances
  - iterative refinement of the search with more complex queries
  - quantitative analysis (collocations, keywords, . . .)
  - both on corpus compiled by team and on large background corpus
- *Task:* Prepare material for the presentation of the projects (e.g. handout, slides, example lists, etc.) → upload in PDF format to Moodle **by Friday 8:00**
- Support available on demand via Zoom / forum

▷ 17:00 – 18:00 {*1h*}

- Online Q&A session on demand

Friday, 08.05.2020

▷ 09:30 – 12:00 {*2.5h*}

- *Presentations:* Final presentation of class projects with preliminary results
  - 10–15 minutes presentation per team
  - 10 minutes feedback, comments, discussion

▷ 12:00 – 13:00 {*1h*}

- *Discussion:* Where to go from here?
  - insights from the course, general questions
  - feedback on the course (esp. virtual classroom and online teaching methods)
  - remaining work on the student projects
  - the future of corpus lexicography