

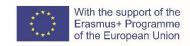


European Master in Lexicography (EMLex)

Advanced module A6: Computational Lexicography

Name of module	Advanced module A6: Computational Lexicography 5 ECTS
Courses	Block seminar: 5-8 days, 24-30 teaching hours
Lecturers	Ulrich Heid / Stefan Evert / Besim Kabashi
Person responsible	Stefan Evert
for module	
Contents	Foundations of corpus linguistics
	Principles and methods of corpus analysis
	Applications of corpus data in lexicography
	Types of corpora, overview of existing corpora
	Corpus design, representativity, data sources, metadata
	Corpus compilation
	Building corpora from online data: Web scraping etc.
	Boilerplate removal, normalization, metadata extraction
	Representation and exchange formats
	Online and stand-alone tools for Web corpus compilation
	• Automatic linguistic annotation (POS, lemma, NER, parsing,)
	Online and stand-alone tools for linguistic annotation
	Searching corpora
	Regular expressions
	Character encodings and the Unicode standard
	CQP query language for lexico-grammatical patterns
	Practical exercises with SketchEngine and CQPweb
	Quantitative analysis
	Frequency lists and metadata distribution
	Collocations and word sketches
	Keyword analysis
	Lexicographic interpretation of results
	 Foundations of statistical inference
	Reproducibility
	Research methodology and documentation
	Data management, sustainability of corpus resources
Learning outcomes	The students should be able
	• to formulate their corpus requirements for a lexicographic project
	and specify the design of a representative corpus;
	• to compile such a corpus from Web pages or other sources;
	• to annotate the corpus with linguistic information using automatic
	natural language processing tools;
	• to search the corpus with regular expressions and more complex
	queries based on lexico-grammatical patterns;
	to apply quantitative techniques such as collocation or keyword
	analysis and interpret the results appropriately;
	• to communicate the results of their work to fellow students;
	to lead academic discussions about technical and methodological
	aspects of corpus-based research; and
	to document and archive corpus data and analysis results.





Requirements for participation	25 ECTS marks from the first semester
Positioning within the schedule	Elective module in the second semester (6 or 7 from 10).
Applicability of the module	For the Master degree programme EMLex
Examination	 The teachers choose one of the following (option b recommended): a) 90-minute final exam on the contents of the seminar or b) presentation of class project plus a short paper (ca. 10 pages) or c) longer paper (15-20 pages)
Calculation of the final mark for the module	100% of the mark obtained in the final exam / assignment
Retakes of the exam	1
Frequency	Annually, during summer term
Workload	Attendance: 5-8 days, 24-30 teaching hours Private study: 120 hours
Duration of module	Block seminar (date and duration to be announced)
Language of teaching	English or German
Selected literature	HSK 5.4, Ch. XVIII + XIX
Last changes	Feb. 2019