

Stefan Evert – Research – **Teaching** – CV – Publications – Software – Private Life

Teaching

Osnabrück – Stuttgart – Summer schools – Other courses – Students

At the University of Osnabrück

Introduction to Computational Linguistics (2005–2010)

Introduction to Computational Linguistics is a compulsory lecture for first-year students in the Cognitive Science programme (2005–2006 with Graham Katz, 2008 with Peter Bosch, 2010 with Maria Cieschinger). [<http://cogsci.uos.de/~CL/teaching/10s/IntroCL/>]

Statistical Natural Language Processing (2006–2010)

This course introduces the fundamental **probabilistic techniques** used in natural language processing. Topics include Markov models, weighted finite-state automata and transducers, probabilistic context-free grammars, the EM algorithm, statistical machine translation, collocations, and maximum entropy models. **Video recordings** of the lectures are available on the course homepage.

[current term: <http://cogsci.uos.de/~CL/teaching/10w/StatNLP/>] [complete handouts and video recordings: <http://cogsci.uos.de/~CL/teaching/08w/StatNLP/>]

Practical NLP (2008–2010)

The seminar **Practical NLP** is taught jointly by the entire Computational Linguistics group. It is a follow-up course to *Introduction to Computational Linguistics* and a useful companion for the more theory-oriented *Statistical NLP*, *Introduction to Syntax* and *Introduction to Semantics*. In this course, participants gain **hands-on experience** of grammar engineering, locating and using resources, the implementation of (statistical or symbolic) NLP algorithms, and the many practical problems involved in building a real-life system that introductory courses tend to gloss over. [<http://cogsci.uos.de/~CL/teaching/10s/PracticalNLP/>]

Analyzing Linguistic Data (2008)

Quantitative linguistic data – whether from a corpus, an eye-tracking study, some other psycholinguistic experiment, or a survey of speaker intuitions – have to be analyzed and explored with statistical tools in order to assess their significance, understand their structure, and reveal the properties and interconnections of the underlying phenomena. This seminar explores the most useful **statistical methods** available for this purpose, including *hypothesis tests* and *correlation* measures, *clustering* and *classification* algorithms, linear and generalized *statistical models*, and *data visualization* techniques. Participants will gain hands-on experience with real-world linguistic data, using the open-source statistical software **R**. [<http://www.cogsci.uni-osnabrueck.de/~CL/teaching/08w/AnalyzingData/>]

Practical Data Analysis (2007)

An interdisciplinary practicum organised together with the Neuroinformatics group (Martin Lauer), in which students gain **hands-on experience** in the application of supervised and unsupervised machine learning techniques to real-life problems (including natural language processing and time series prediction). [<http://www.cogsci.uni-osnabrueck.de/~CL/classes/07s/dataAnalysis/>]

Information Processing in Machine Learning and Computational Linguistics (2006)

Interdisciplinary course held together with the Neuroinformatics group (Martin Lauer), focussing on **vector space representations**, data processing and **dimensionality reduction techniques** (SVD, PCA, LSA), which are used both in machine learning and in statistical natural language processing. [<http://www.cogsci.uni-osnabrueck.de/~CL/classes/06s/informationProcessing/>]

Seminar on Quantifying Linguistic Experience (2005)

Quantifying Linguistic Experience is a hands-on introduction to statistical methods for the quantitative analysis of corpus frequency data, which can be understood as an approximative model for the linguistic experience of a human speaker. In addition to learning the necessary statistical theory, participants are taught how to apply it to real-world data using the statistical programming language **R**

[<<http://www.r-project.org/>>].

Seminar on Lexical Statistics and Computational Morphology (2001)

Seminar on **Word Frequency Distributions** and their application to **Computational Morphology** (with Anke Lüdeling).

At the University of Stuttgart

Proseminar Formale Sprachen (2000–2004)

Introductory class on **Formal Language Theory** for 2nd year students (in German). Handouts are available from <<http://www.ims.uni-stuttgart.de/Lehre/teaching/2004-WS/Formale-Sprachen/>>.

Proseminar Statistische Methoden (2002–2004)

Introductory class on **Statistical Methods** for 2nd year students (in German). Handouts are available from <<http://www.ims.uni-stuttgart.de/Lehre/teaching/2004-SS/Statistische-Methoden/>>.

Miscellaneous Courses (2000–2001)

Software course **Werkzeuge für Computerlinguisten** (IMS Corpus Workbench and R, 2001).

Hauptseminar **Lexikostatistik und Computermorphologie** (2001, with Anke Lüdeling).

Hauptseminar **Terminologie-Extraktion aus Texten** (2001, with Ulrich Heid).

Software course **Perl für Computerlinguisten** (2000, with Arne Fitschen and Wolfgang Lezius).

Hauptseminar **Maschinelle Lexikographie** (2000, with Ulrich Heid).

Summer schools

Computational Lexical Semantics (ESSLLI 2009)

An overview of current research in **computational lexical semantics**, combining theoretical and methodological background with hands-on experience. One-week introductory course at the European Summer School on Logic, Language and Information (ESSLLI 2009), Bordeaux, France (with Gemma Boleda, UPC, Barcelona).

Course homepage: <<http://clselli09.wordpress.com/>>

Distributional Semantic Models – Theory and Empirical Results (ESSLLI 2009)

One-week advanced course on the **mathematical foundations of distributional semantic models** and their **evaluation**, with a particular focus on relating computational models to fundamental issues of semantic theory. At the European Summer School on Logic, Language and Information (ESSLLI 2009), Bordeaux, France (with Alessandro Lenci, U of Pisa).

Course homepage: <<http://wordspace.collocations.de/doku.php/course:start>>

Statistical Analysis of Corpus Data with R (EMA Summer School 2008)

An introduction to **statistical methods** for the **analysis of corpus data** and their practical application with **R** [<<http://www.r-project.org/>>]. One-week course given at the 9th Summer School of the European Masters in Language and Speech Technology Programme, Stuttgart, Germany. Slides and materials are available from <http://purl.org/stefan.evert/SIGIL/sigil_R/>.

Statistical Programming in R for Computational Linguists (DGfS/CL Fall School 2007)

A hands-on introduction to **Statistical Programming** in the **R** language [<<http://www.r-project.org/>>] for **Computational Linguists**, a two-week course at the DGfS/CL Fall School in Computational Linguistics 2007, Potsdam, Germany (with Marco Baroni, CIMEC, U of Trento). Slides and materials can be downloaded from <http://purl.org/stefan.evert/SIGIL/potsdam_2007/>.

Counting Words: An Introduction to Lexical Statistics (ESSLLI 2006)

Introduction to Lexical Statistics and **mathematical models of word frequency distributions** (one-week introductory course) at the European Summer School on Logic, Language and Information (ESSLLI 2006),

Malaga, Spain (with Marco Baroni, U of Bologna, Forli). Slides can be downloaded from <http://purl.org/stefan.evert/zipfR/>.

Computational Approaches to Collocations (ESSLLI 2003)

One-week introductory course on **Computational Approaches to Collocations** at the European Summer School on Logic, Language and Information (ESSLLI 2003), Vienna, Austria (with Brigitte Krenn, OFAI). PowerPoint slides can be downloaded from <http://www.collocations.de/EK/>.

Miscellaneous courses

Introduction to Statistical Data Analysis (Zürich, 2010)

A block course taught for the Doctorate Programme in Linguistics at the University of Zürich. It forms the basis for a restructured version of the SIGIL tutorial slides to be released at <http://sigil.r-forge.r-project.org/>.

Statistics for Linguists with R (Saarbrücken, 2009)

A SIGIL tutorial on **statistical data analysis** for linguistics and translation studies, using the open-source software **R** [<http://www.r-project.org/>]. Slides and materials are available from http://purl.org/stefan.evert/SIGIL/sigil_R/.

Statistics for Corpus Linguists (ICAME 2007)

A short tutorial on the **foundations of statistics for corpus linguists**, given at the 28th ICAME Conference, Stratford-upon-Avon, UK, May 2007 [<http://rdues.uce.ac.uk/icame/>].

Handouts: http://purl.org/stefan.evert/PUB/Handout_ICAME_Statistics.pdf

Linear Algebra in a Nutshell (CIMEC, U of Trento, 2007)

A 6-hour crash course on **linear algebra** and **vector space models** held at the Rovereto campus of the University of Trento, Italy, in March 2007. This course was part of an exchange funded by the Erasmus teacher mobility programme.

Handouts: http://purl.org/stefan.evert/PUB/Handout_LA_Trento_1.pdf (vector spaces),
http://purl.org/stefan.evert/PUB/Handout_LA_Trento_2.pdf (distance, norm, kernel),
http://purl.org/stefan.evert/PUB/Handout_LA_Trento_3.pdf (dimensions & PCA)

Statistical Methods for Corpus Data (EURAC, Bolzano, 2005)

A three-day seminar on **statistical methods** for the analysis of **corpus frequency data**, held at the European Academy (EURAC), Bolzano, Italy, in September 2005 (with Marco Baroni, U of Bologna, Forli).

Handouts: http://purl.org/stefan.evert/PUB/Handout_EURAC_Statistics_1.pdf,
http://purl.org/stefan.evert/PUB/Handout_EURAC_Statistics_2.pdf

Selected students (supervised theses)

PhD – Master – Bachelor

PhD theses

Exploiting Linguistic and Statistical Knowledge for a Text Alignment System (Bettina Schrader, 2007)

A novel architecture for text alignment (ATLAS) performs alignment at multiple levels in parallel (e.g. paragraph, sentence and word alignment) and is designed for easy integration of various linguistic knowledge sources.

Supervisors: Peter Bosch, Helmar Gust, Stefan Evert

Master (MSc) theses

Hybrid Sweeping: Streamlined Perceptual Structured-Text Refinement (Egon Stemle, 2009)

A description of the KrdWrd Project, developed in cooperation with Johannes Steger and other students at the Institute of Cognitive Science. The goals of the project are (i) to provide tools and infrastructure for the

acquisition, visual annotation, merging and storage of Web pages for the purpose of corpus building and content mining; (ii) to develop a classification engine that learns to annotate and clean Web pages automatically based on visual renderings of the pages; and (iii) to provide graphical tools for the manual inspection of annotation/cleaning results.

Coordinated with MSc theses *Web Attention Technology: JAMF and KrdWrd* by Johannes Steger (supervised by Peter König & Stefan Evert). The KrdWrd technology will form the basis of the second CLEANVAL contest on boilerplate removal for the Web as Corpus, to be held in 2010.

KrdWrd project homepage: <<https://krdwrd.org/>>

Supervisors: Stefan Evert, Peter König

An Evaluation of POS Taggers for the Web as Corpus (Eugenie Giesbrecht, 2008)

Part-of-speech (POS) tagging is often considered a “solved task” in computational linguistics, with state-of-the-art taggers reporting accuracies around 97%. However, this performance is achieved only for texts that are sufficiently similar to the training data, and may drop markedly when the tagger is applied to a different genre. This thesis evaluates three widely-used statistical taggers (TreeTagger, Stanford Tagger and Apache UIMA Tagger) on manually annotated samples from a German Web corpus. The results show the expected loss of accuracy, large differences between genres (such as online newspapers vs. discussion forums), and the importance of good probabilistic models for unknown words.

PDF version of the thesis: <http://www.cogsci.uos.de/~CL/download/MSc_Giesbrecht2008.pdf>

Supervisors: Stefan Evert, Marco Baroni

Junk-Email Classification Using NLP Features (Oleksandr Kolomiyets, 2007)

Experiments on the automatic detection of e-mail spam (UBE = unsolicited bulk e-mails) with various machine-learning algorithms (Naive Bayes, Maximum Entropy and Decision Trees). In contrast to the bag-of-words approach of most standard spam filters, these experiments focus on linguistic properties such as part-of-speech tags and phrase patterns, achieving good performance with comparatively low-dimensional feature spaces.

Supervisors: Veit Reuer, Stefan Evert

Bachelor (BSc) theses

Morphology Mining (Thorben Krüger, 2009)

A study on unsupervised learning of German inflectional morphology, using readily available linguistic knowledge about regular inflectional paradigms (as provided by SMOR FST). Stem/class hypotheses are generated by a modified FST (Adolphs 2008) and then ranked and filtered (i) with a MDL algorithm based on corpus frequency data and (ii) with a heuristic method that uses Google queries to find out how many surface forms predicted by a hypothesis are attested on the Web.

PDF version of the thesis, software and data sets can be obtained here: <<http://www-lehre.inf.uos.de/~thkruege/downloads.html>>

The Paradigmatic Structure of Person Marking: A Game Theoretical Analysis (Dominik Hlusiak, 2009)

Design of a game-theoretical model for the typological paradigms of person marking in pronoun systems, as an adaptation of Jäger’s (2007) use of evolutionary game theory to explain case marking. Contains excellent concise summaries of evolutionary game theory and person marking (based on Cysouw 2003).

Supervisors: Stefan Evert, Helmar Gust

Qualitative Enhancements by Quantitative Analysis (Daniel Berndt, 2009)

A thorough reanalysis of surprising findings from a corpus-based study of nation + noun expressions in English (contrasting the Adj-N and N-Prep-N constructions), which was carried out during an internship at the UPC Barcelona. The original observation is explained as a mathematical artefact, and the underlying core phenomenon is revealed (as a basis for linguistic interpretation).

PDF version of the thesis: <http://www.cogsci.uos.de/~CL/download/BSc_Berndt_2009.pdf>

Supervisors: Stefan Evert, Louise McNally

Word Spaces (Alexander Frey, 2009)

Development of a Java GUI toolkit for interactive exploration of “word spaces”, i.e. distributional representations of the usage and meaning of a word. The toolkit is demonstrated and tested in a number of case studies.

Supervisor: Stefan Evert

Subsymbolic String Representations Using Simple Recurrent Neural Nets (Gabriel Pickard, 2008)

Preliminary experiments on the completely unsupervised acquisition of patterns in natural and artificial languages by training recurrent neural networks (rNN) as auto-encoders. Results show that NN in general, and the standard training algorithms for rNN in particular, are not very well suited for the representation of symbolic structures. Better performance is obtained with hand-crafted networks using a special “Cantor encoding”.

Supervisors: Helmar Gust, Stefan Evert

Learning the Semantics of Wikipedia Hyperlinks (Daniel Bauer, 2007)

Experiments on the automatic identification of semantic relations between Wikipedia “concepts” (i.e. articles) based on information from hyperlinks between these articles. An XML-annotated corpus is compiled from the Wikipedia database, and training data are obtained by a semi-automatic mapping of Wikipedia articles to WordNet synsets.

PDF version of the thesis: <http://www.cogsci.uos.de/~CL/download/BSc_Bauer2007.pdf>

Supervisor: Stefan Evert

Prototype-Based Relevance Learning for Genre Classification (Jan Gasthaus, 2007)

Several prototype-based supervised learning algorithms from the Learning Vector Quantization (LVQ) family are implemented in a Java library and evaluated with respect to their suitability for text classification tasks in computational linguistics. They are found to achieve high accuracy (comparable to support vector machines) on the task of genre classification for the British National Corpus.

PDF version of the thesis, software and data sets can be downloaded here: <<http://www.cogsci.uos.de/~CL/study/gasthaus2007/>>

Supervisors: Stefan Evert, Martin Lauer

Using Suffix Trees for Text Categorization in Computational Linguistics (Maria-Hendrike Peetz, 2007)

Suffix trees, an efficient text indexing algorithm used in bioinformatics, are adapted to develop a more fine-grained similarity measure for texts (and other formal strings) than simple n-gram overlap. The new algorithm achieves a remarkable accuracy of 92% for gender classification in the British National Corpus, and marks the first application of suffix trees in computational linguistics to our knowledge.

Supervisors: Stefan Evert, Volker Sperschneider

Automatic Suggestion of Wikipedia Categories (Thomas Göbel, 2007)

Wikipedia, the free encyclopedia, uses a category system to establish a hierarchical order for its articles. This thesis explores the possibility of automatically assigning such categories to new articles, using a k-Nearest Neighbour algorithm based on textual features that are unique to Wikipedia articles.

Supervisors: Stefan Evert, Peter Geibel

Requirements for and design of a flexible tokenization system (Meike Aulbach, 2006)

Tokenization is a first and essential step in most natural language processing (NLP) pipelines for written text. While often considered a “trivial” problem, the accuracy of tokenizers is often unsatisfactory, especially on less formal genres such as personal Web pages. This study specifies requirements for a high-quality tokenization system, analyzes different types of problematic cases, evaluates a number of commonly used tokenizers, and outlines the architecture of a more flexible tokenization system.

PDF version of the thesis: <http://www.cogsci.uos.de/~CL/download/BSc_Aulbach_2006.pdf>

Supervisors: Bettina Schrader, Stefan Evert

Using Statistical and Linguistic Spam Filters to Improve the Quality of Web Corpora (Florian Groß, 2006)

Experiments on using spam filters and support vector machines (SVM) for the automatic detection of unwanted pages in Web corpora (termed “WaC spam”). Results on an existing collection of WaC spam and additional manually classified Web pages are encouraging, especially for the SVM classifiers.

PDF version of the thesis, software and data sets can be downloaded here:
<http://wacky.sslmit.unibo.it/old_wiki/doku.php?id=cite:gross_2006>

Supervisors: Stefan Evert, Peter Geibel

POD ERRORS

Hey! **The above document had some coding errors, which are explained below:**

Around line 295:

Unterminated I< ... > sequence