

# Language Processing II

Fridays, 10:15-12:00, 24.4.07

Lecturers:

Manex Agirrezabal and Patrizia Paggio

**Course Description:** Have you ever wondered how to build a system that can process text automatically, for example reveal syntactic structure, extract semantic information or capture stylistic features of different authors?

In this class we will explore advanced algorithms and models for natural language processing, and discuss the application of these algorithms and models to key problems in natural language processing, such as part of speech tagging, named entity recognition and syntactic parsing, as well as more general applications such as author verification and hate speech profiling.

**Credit Hours:** 2 hours of class

## **Course Objectives:**

At the completion of this course, students will:

1. Understand advanced concepts of natural language processing.
2. Demonstrate theoretical insight into natural language processing by identifying complex problems and solutions in the context of practical applications
3. Apply feature extraction methods and modelling techniques in natural language processing.
4. Deal with specific challenges arising from processing user-generated content, evaluate systems or system components.

**Assessment:** Take-home assignment, set subject

**Grade:** 7-point grading scale

**Scope:** 6-10 standard pages. 5 days are given for assignment completion.

**Regulations for group exams:** The exam can be taken individually or as a group exam by 2-3 students but with individual assessment. For group exams, each individual participant's contribution to the assignment must be readily identifiable, and the joint part must not exceed 50% of the total work. The scope for the take-home assignment for groups is: 9-15 standard pages (2 students) or 12-20 standard pages (3 students).

**Exam language(s):** English.

**Make-up exam/resit:** Take-home assignment, set subject. 6-10 standard pages. 5 days are given for completion of the assignment. It can only be taken individually.

The topic of the assignment will be agreed on together with the students, and will take the form of a project topic preferably based on a research challenge from the NLP research field.

**Textbook:** Daniel Jurafsky & James H. Martin (2023). Speech and Language Processing. Draft of Jan, 2023 available at <https://web.stanford.edu/~jurafsky/slp3/> (J&M in what follows).

**Tentative Course Plan:**

The weekly coverage might change as it partly depends on the progress and the interests of the class.

Rsp.	Sess no.	Day	Title	Readings/Homework
PP	1	Feb 9	Introduction to study unit and challenge topics Practical on style analysis	One of the following challenges will be chosen for the exam:  1) Multi-Author Writing Style Analysis from <a href="https://pan.webis.de/clef24/pan24-web/">https://pan.webis.de/clef24/pan24-web/</a> 2) Human value detection, subtask 1 from <a href="https://touche.webis.de/clef24/touche24-web/human-value-detection.html">https://touche.webis.de/clef24/touche24-web/human-value-detection.html</a>

PP	2	Feb 16	HMM (pos tags)	J&M: Chapter 8, Sections 1, 2 and 4
MA	3	Feb 23	HMM+CRF / POS CRF / NER + chunking	<p>J&amp;M: Chapter 8, Section 5</p> <p>Additional reading:            McCallum, Andrew, Dayne Freitag, and Fernando CN Pereira. "Maximum entropy Markov models for information extraction and segmentation." <i>Icml</i>. Vol. 17. No. 2000. 2000.</p> <p>Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001). (The article will be uploaded to Absalon)</p>
PP	4	March 1	Constituency parsing	J&M, Chapter 17.1-5 and 17.8.
PP	5	March 8	Dependency parsing <b>Deciding project topic</b>	<p>J&amp;M: Chapter 18.1-2 and 4.</p> <p>Python Notebook on constituency parsing</p>
PP	6	March 15	Dependency parsing, finish. Lexicons for Affect, Sentiment, Emotions and Human Values.	<p>Python Notebook on dependency parsing.</p> <p>J&amp;M: Chapter 25.1-7.</p>
PP	7	March 22	Intro to chosen project topic, and State of the Art for it. Writing + assessment criteria. Midterm evaluation?	To be announced

EASTER				
MA	8	April 5	Recurrent Neural Network Language Models 1 (character features)	J&M: Chapter 7, Section 7.5 Chapter 9, Section 9.1–9.2 (Optional 9.3–9.4)
MA	9	April 12	Recurrent Neural Network Language Models 2 (LSTM and attention)	J&M: Chapter 9, Section 9.5, 9.7, 9.8
MA	10	April 19	Brainstorming of reasoning behind methods (Stylometric/textual features)	Stamatatos, Efstathios. "A survey of modern authorship attribution methods." <i>Journal of the American Society for information Science and Technology</i> 60.3 (2009): 538-556. <a href="#">Link to article</a>
MA	11	April 26	Extracting features from models	TBA
MA	12	May 3	Finetuning models and prompting	TBA
MA and PP	13-14	May 10	Project workshop (4 hours)	