

# Corpus Linguistics (PS)—Summer/Winter Term 2019/20

Wed 12-2 p.m., Seminarhaus - SH 4.105

Course Web Page: <http://english-linguistics.de/nschenk/teaching.html>

## Dr. Niko Schenk

**E-mail:** [n.schenk@em.uni-frankfurt.de](mailto:n.schenk@em.uni-frankfurt.de)

**Website:** <http://www.uni-frankfurt.de/46551536/Schenk>

**Offices:** Norbert-Wollheim-Platz 1, IG-Farben Haus, Raum: 3.216 (Campus Westend)  
Robert-Mayer Straße 11-15, Raum: 110 (Campus Bockenheim)

**Office hours (Westend):** Wed. 2-3 by appt.

**Tel:** 069-798/32530

This syllabus may be subject to unconstrained revisions.

**Course Description:** This course provides theoretical and practical training in the use of corpus linguistics. The goal is to show how linguistic hypotheses can be derived from patterns in corpus data. Participants will learn ways to search for linguistic elements of interest and evaluate them by means of simple statistical tests.

The first part of this course will give a brief overview of currently used corpora and methodologies for text digitization, and introduce basic terminology. In the second part, participants will learn how to apply particular techniques to real corpus data, for instance to find syntactic phrases, semantically related words or collocations. Students will study differences in word meaning based to the topological co-occurrence of words and their contexts, or detect long repetitions within corpus data (plagiarism).

Finally, the course will provide a general overview of currently used methods in information retrieval, and provide hand-on exercised for computing various statistics on text data.

**Prerequisite(s):** A notebook / tablet computer for in-class exercises.

**Note(s):** Registration will be done in the first course meeting.

## Course Objectives:

At the completion of this course, students will be able to:

1. use state-of-the art corpora for linguistic research.
2. understand corpus-specific terminology.
3. know how linguistic corpora are created and digitized.
4. work with diachronic corpus data.
5. apply basic concepts of statistical information retrieval.
6. manually compute statistics for word frequencies and word-co-occurrences.
7. test the outcome of an experiment for statistical significance.
8. read and understand current scientific literature in the domain of CL.

## Course Requirements:

- **Teilnahmeschein**

- Active **in-class participation** & presentations possible (Referat), **weekly homework assignments**.
- Homework assignments in poor quality will not be sufficient for an ungraded Teilnahmeschein.

- **Leistungsschein**

- Active **in-class participation** (presentations possible), **weekly homework assignments**.
- **Course project** at the end of the seminar.  
short paper/kleine Hausarbeit (große Hausarbeit possible for BA students)

## Course Policies:

- **In-class Presentations**

- TBD: Each student is supposed to give a short in-class presentations (10-30 minutes) depending on a pre-selected topic by the instructor.

- **Assignments**

- Students are expected to work independently on homework assignments. Offering and accepting solutions from others is an act of plagiarism. Discussion amongst students is encouraged, but when in doubt, direct your questions to the lecturer.
- There will be regular (weekly) assignments that you are expected to do in preparation for the next session. The assignments are part of the participation requirements. If you miss one, talk to the instructors about making up the work. **Assignments are due at 6 p.m. on Monday before the class** to the instructor. Note, that assignments are to be sent **via email attachment in a separate PDF or Word file**.
- **No late assignments will be accepted under any circumstances!**
- Possible topics for students needing a Leistungsschein of any sort will be announced by the end of the seminar.

- **Attendance and Absences**

- Attendance is expected and will be taken each class. You are not allowed to miss more than two class meetings during the semester without a proper excuse (doctor's note). Any further absences will result in point and/or grade deductions. Please let me know in advance if you are going to miss class.
- Students are responsible for all missed work, regardless of the reason for absence. It is also the absentee's responsibility to get all missing notes or materials. Please, ask your fellow students for help.

## Data for Research Disclosure:

Any and all results of in-class and out-of-class assignments and examinations are data sources for research and may be used in published research. All such use will always be anonymous.

### Tentative Course Outline:

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

Week	Content
Week 1 (Oct 16)	• <b>Organizational, A Simple Introduction to CL</b>
Week 2 (Oct 23)	• <b>Introduction to Corpus Linguistics (I)</b> (Use Case, Corpora, Creation, Applications) • Homework assignment: Google NGram Viewer
Week 3 (Oct 30)	• No class
Week 4 (Nov 6)	• <b>Introduction to Corpus Linguistics (II)</b> (Use Case, <b>Corpora, Creation</b> , Applications)
Week 5 (Nov 13)	• <b>Introduction to Corpus Linguistics (III)</b> (Use Case, Corpora, Creation, <b>Applications</b> ) • Homework assignment: <a href="http://corpus.byu.edu/coca/">http://corpus.byu.edu/coca/</a>
Week 6	• <b>TBA</b>
Week 7	• <b>TBA</b>
Week 8	• <b>TBA</b>
Week 9	• <b>TBA</b>
Week 10	• <b>TBA</b>
Week 11	• <b>TBA</b>
Week 12	• <b>TBA</b>
Week 13	• <b>TBA</b>