

## Course Description

*Empirical methods* have revolutionized nearly all subfields of linguistics in the recent years. Since the 1990s, language resources, in particular the corpus, have played a vital role in virtually all empirically-oriented linguistic studies.

This course offers an introduction of corpus linguistics for linguistic graduate students, including the necessary tools, techniques, and analyzing methodologies for doing corpus-based studies and corpus annotation projects. Existing major corpora will be scrutinized for a better understanding of their linguistic uses. To keep students up to date with the latest developments in corpus linguistics, rationals and methods for using web as corpus will be the main focus as well. Goals of this course include: [1] passing on essential knowledge and skills in building, annotating and searching corpora, and [2] familiarizing students with the methods to perform a corpus-based quantitative analysis of some linguistic phenomena. Students will also gain hands-on experience in the lab session, and learn how to formulate research questions by working on a specific topic of their own interest.

**Teaching assistants** Chan-Chia (Mike) [chanchiah@gmail.com]  
and Yu-Yun (Taco) [yuyun.unita@gmail.com].

**Course webpage** Course web page at NTU CEIBA  
<https://ceiba.ntu.edu.tw/1021corpus>, and lecture  
scripts will be made publicly available at SlideShare  
[<http://www.slideshare.net/shukaihsieh/>]

## Syllabus (Schedule subject to change)

---

Week	Date	Topic	Lab
1	09/12	Orientation: What is corpus linguistics?	
2	09/19	{National Holiday}	
3	09/26	Corpus building, processing, and evaluation	BNC/BNY-web
4	10/03	Pre-processing	Corpus tools
5	10/10	{National Holiday}	
6	10/17	Corpus annotation (I)	Annotation exercise
7	10/24	Corpus annotation (II)	Annotation exercise
8	10/31	Corpus annotation (III)	Annotation exercise
9	11/07	Corpus-based analysis (I) Neo-Firthian	LexicoR package
10	11/14	Corpus-based analysis (II) Synchronic and diachronic variation	LexicoR package LexicoR package
11	11/21	Corpus-based Analysis (III) Corpus methods in cognitive-functional linguistics	LexicoR package
12	11/28	Basic corpus statistics (I)	LexicoR package
13	12/05	Basic corpus statistics (II) and term project proposal	LexicoR package
14	12/12	Web as Corpus / Brain as Corpus	Bootcat
15	12/19	[Paris: official leave] (Guest lecture)	
16	12/26	Methodological issues in Corpus Linguistics	LexicoR package
17	01/02	Class project workshop	
18	01/09	Term paper due	

## Class Activities and Requirements

- You are expected to complete **weekly readings and assignments**, and actively participate in classroom activities and discussions. There will be no required paper textbooks for this course, instead, a *Course Reader* compiled by me will be distributed in next class. Lecture notes and slides will be made available in the course web page and SlideShare.
- Each class will be divided into **two sessions**: (i) lecture, presentation and discussion (two hours), and (ii) lab (one hour). Each one will be assigned at least one paper for seminar presentation. For some homework assignments, we will be using the server of LOPE lab. Instruction about how to use the lab server will be given before the first assignment. In the Lab session, you will learn how to construct a corpus from scratch, search corpus using existing scripts and other corpus tools, as well as perform quantitative analysis of corpus data.
- You are required to submit a **term project** involving original corpus-based research. It can be either a local corpus construction project (e.g., emotional corpus), a program package with documentation (e.g., R or Python) to perform some substantial corpus processing task, or a research paper on a corpus-based topic. You will be asked to turn in a proposal early in the semester so that I can help you design and execute your project. Proposed topics will also be discussed in class later.
- There will be (in principle) no exams. **Grading** is based on: Homework and Lab exercise (30%); Seminar Discussion and Presentation (30%); Term paper (40%).

# Bibliography

- [1] McEnery, Tony and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.
- [2] Baker, Paul. (ed). 2009. *Contemporary Corpus Linguistics*. Continuum Publisher.
- [3] Garside, Roger et al. 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Addison Wesley.
- [4] Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. Longman.
- [5] McEnery, Tony and Andrew Wilson. 1996. *Corpus Linguistics*. Edinburgh University Press.
- [6] Sampson, Geoffrey and Diana McCarthy. 2004. *Corpus Linguistics: Readings in a Widening Discipline*. Continuum.
- [7] Scott, Mike and Christopher Tribble. 2006. *Textual Patterns: Key words and corpus analysis in language education*. John Benjamins.
- [8] Gries, Stefan. 2008. *Quantitative Corpus Linguistics with R: A Practical Introduction*. Routledge.
- [9] Teubert, Wolfgang and Anna Čermáková. 2007. *Corpus Linguistics: A Short Introduction*. Continuum Press.
- [10] Teubert, Wolfgang and Anna Čermáková. 2008. *Corpus Semantics: An Introduction*. Continuum Press.
- [11] Johnson, Mark. 2008. *Essential Python for corpus linguistics*. Blackwell Publishers.

- [12] Martin Wynne (ed). *Developing Linguistic Corpora: a Guide to Good Practice*. available at <http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm>
- [13] O'Keeffe and McCarthy (eds). 2010. *The Routledge Handbook of Corpus Linguistics*. Routledge.
- [14] Baker et al (eds). 2006. *A Glossary of Corpus Linguistics*. Edinburgh University Press.
- [15] Lüdeling and Kytö (eds). 2008. *Corpus Linguistics: An International Handbook*. Volumn1-2. Walter de Gruyter.