Shu-Kai Hsieh
shukaihsieh@ntu.edu.tw

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. In the lab session, this course will cover the fundamentals of computer programming skills in Python/ NLTK, a very popular programming language and modules for many corpus linguistic applications. This course will be taught in English and assumes only minimal background with computers, no programming skills or knowledge are required.

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 Hsu (Mike) [chanchiah@gmail.com] and Yu-Yun Chang (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 http://lope.linguistics.ntu.edu.tw/courses/corpusling2015

Shu-Kai Hsieh
shukaihsieh@ntu.edu.tw

Syllabus (Schedule subject to change)

Week	Date	Topic	Lab
1	09/18	Orientation	
2	09/25	Introduction to Corpus Linguistics: building, processing, and evaluation	BYU
3	10/02	Corpus-based analytical tools	Antconc Family
~~4~~	~~10/09~~	~~National Holiday~~	
5	10/16	Corpus-based analytical tools	Word Sketch Engine
6	10/23	Corpus annotation (I)	Annotation exercise
7	10/30	Corpus annotation (II)	Annotation exercise
8	11/06	Corpus-based analysis	Beginning Python
9	11/13	Corpus Data Collection	WAC.py
10	11/20	Corpus Data Preprocessing	WAC.py
11	11/27	Corpus Data Preprocessing	WAC.py
12	12/04	Basic corpus statistics (I)	WAC.py
13	12/11	Basic corpus statistics (II) and term project proposal	WAC.py
14	12/18	Corpus data and Brain data	WAC.py
15	12/25	(Guest lecture)	
16	01/01	Methodological issues in Corpus Linguistics Evaluation and Comparison	WAC.py
17	01/08	Class project workshop	
18	01/14	Term paper due	

Shu-Kai Hsieh
shukaihsieh@ntu.edu.tw

Class Activities and Requirements

- You are expected to complete **weekly readings and assignments**, and actively participate in classroom activities and discussions. The *Course Reader* will be mainly based on two textbooks ([1] and [2]), and will be distributed in next class. Lecture notes and slides will be made available in the course web page.

- 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/chapter for seminar presentation. For the completion of your in-class exercise, a laptop is required to bring to the class. For some homework assignments, we might use the server of LOPE lab. Instruction about how to use the lab server will be given before the 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., Taiwan Parliament Corpus; NTU-GIL thesis 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%).

Resources

Software Tools for Corpus Linguistics: `http://www.uow.edu.au/~dlee/software.htm`

Bibliography

[1] McEnery, Tony and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

[2] Schäfer, Roland and Felix Bildhauer. 2013. *Web Corpus Construction*. Morgan & Claypool Publishers.

[3] Baker, Paul. (ed). 2009. *Contemporary Corpus Linguistics*. Continuum Publisher.

[4] Garside, Roger et al. 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora.* Addison Wesley.

[5] Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. Longman.

[6] McEnery, Tony and Andrew Wilson. 1996. *Corpus Linguistics*. Edinburgh University Press.

[7] Sampson, Geoffrey and Diana McCartthy. 2004. *Corpus Linguistics: Readings in a Widening Discipline.* Continuum.

[8] Scott, Mike and Christopher Tribble. 2006. *Textual Patterns: Key words and corpus analysis in language education*. John Benjamins.

[9] Gries, Stefan. 2008. *Quantitative Corpus Linguistics with* R*: A Practical Introduction*. Routledge.

[10] Teubert, Wolfgang and Anna Čermáková. 2007. Corpus Linguistics: A Short Introduction. Continuum Press.

Shu-Kai Hsieh
shukaihsieh@ntu.edu.tw

[11] Teubert, Wolfgang and Anna Čermáková. 2008. Corpus Semantics: An Introductrion. Continuum Press.

[12] Johnson, Mark. 2008. *Essential Python for corpus linguistics*. Blackwell Publishers.

[13] Martin Wynne (ed). *Developing Linguistic Corpora: a Guide to Good Practice*. available at `http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm`

[14] O'Keeffe and McCarthy (eds). 2010. The Routledge Handbook of Corpus Linguistics. Routledge.

[15] Baker et al (eds). 2006. A Glossary of Corpus Linguistics. Edinburgh University Press.

[16] Lüdeling and Kytö (eds). 2008. Corpus Linguistics: An International Handbook. Volumn1-2. Walter de Gruyter.