Computational Corpus Linguistics

LING-367 (Fall 2020)
Mon+Wed, 2:00-3:15

Instructor:
Amir Zeldes (Assoc. Prof. of Computational Linguistics)
E-Mail: amir.zeldes@georgetown.edu
Website: https://corpling.uis.georgetown.edu/amir
Office: online only due to the pandemic (office hours Fri 9:30-11:30, via Zoom)

Teaching assistant: Luke Gessler (lg876@georgetown.edu)

Summary:
Digital linguistic corpora, i.e. electronic collections of written, spoken or multimodal language data, have become an increasingly important source of empirical information for computational, theoretical and applied linguistics in recent years. This course is meant as a theoretically founded, practical introduction to corpus work with a broad selection of data, including non-standardized varieties such as language on the Internet, learner corpora and spoken corpora. We will discuss issues of corpus design, annotation and evaluation using quantitative methods and both manual and automatic annotation tools for different levels of linguistic analysis, from parts-of-speech, through aspects of syntax, semantics and discourse annotation. As part of the course, students will participate in the creation of a multilayer corpus that will be built up as the course progresses.

Course requirements:

Attendance	
Final project	35%
Graded assignments	50%
Participation	15%

Assignments and final project:
Assignments will include reading assignments, possibly including a brief writing assignment about the text (reviewing an article, discussing some question), corpus search assignments, and annotation assignments, in which we will produce annotated corpus data using a variety of coding schemes for parts of speech, syntactic annotation and discourse level annotations.

Each student will be responsible for one short text that they will be annotating throughout the semester. At the end of the course, students will be given the opportunity to make their corpus materials available to the public under a Creative Commons license over the corpus linguistics server (optional). Results from previous semesters can be found here for reference:

http://corpling.uis.georgetown.edu/gum/

The final project will be in a conference short paper format and should discuss a linguistic phenomenon using either the corpus created in this course, or another available corpus depending on the phenomenon or language in question.

Asynchronous participation, absences and timely assignment submission:
Students are expected to attend all classes and to complete all assignments on time. Due to pandemic constraints, it is understood that students will participate virtually, and that some students will attend from remote location. Students who are joining from a time zone which makes synchronous participation unfeasible are expected to review all lectures asynchronously before the next scheduled session; every effort will be made to make face to face contact possible for all students, including via individual or group office hours outside of normal course times with the instructor and the TA, as well as any special sessions required to make course assignments work remotely.

Beyond these measures, students may excuse themselves via e-mail from up to three meetings at their discretion, provided that they make up for lost course work and submit the assignments. Excused absences can be made up asynchronously and it is understood that absent students may not be able to review materials in time for the next session. That said, absences (not including asynchronous but timely participation) make following the course difficult, and repeated unexcused absences of this kind may have an adverse effect on grades, up to and including failure. Additional absences for special reasons (religious observances, prolonged illness etc.) may be coordinated on a case by case basis with the instructor (documentation may be required as applicable).

Course plan: (approximate)

Week	Topics	Readings and workshops	Assignments
Week 1	Introduction	Fillmore (1992)	1 Page write up
Week 2	Corpus Design	Biber (1993) or Hunston (2008) (assigned by group)	
Week 3	Preprocessing	Workshop - preprocessing	Preprocessing assignment
Week 4	Part of speech tagging	Workshop - tagging	Tagging assignment
Week 5	Corpus query		
Week 6	Lexicography and collocations	Gries (2015), Gablasova et al. (2017)	Collocation assignment
Week 7	Treebanks and dependency grammar		Association assignment
Week 8		Workshop - syntax annotation	Syntax assignment
Week 9	Multilayer corpora		
	Information structure and referentiality	Krifka (2008)	Review assignment (mock reviewing of a short conference paper)
Week 10			
	Coreference		
Week 11		Workshop - entities and coreference	Coreference assignment
	Rhetorical Structure Theory	Mann & Thompson (1988)	
Week 12		Workshop - Rhetorical Structure	Rhetorical structure assignment
Week 13			
Week 14	Conclusion		

Reading list: (all readings will be available over Canvas)

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243–257.

Evert, Stefan. 2009. Corpora and collocations. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics. An International Handbook*. Vol. 2, 1212–1248. Berlin: Mouton de Gruyter.

Fillmore, Charles J. 1992. 'Corpus linguistics' or 'computer-aided armchair linguistics'. In Jan Svartvik (ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, 35–60. Berlin and New York: Mouton de Gruyter.

Gablasova, Dana, Vaclav Brezina & Tony McEnery. 2017. Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. Language Learning 67(1), 155–179.

Gries, S. T. 2015. 50-something years of work on collocations: what is or should be next... Sebastian Hoffmann, Bettina Fischer-Starcke, & Andrea Sand (eds.), *Current issues in phraseology*. Amsterdam & Philadelphia: John Benjamins, 135-164.

Hunston, Susan. 2008. Collection strategies and design decisions. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics. An International Handbook*, 154–168. Berlin: De Gruyter.

Krifka, Manfred. 2008. Basic notions of information structure. *Acta Linguistica Hungarica* 55. 243–276.

Mann, William C. & Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8(3). 243–281.

Optional readings:

Biber, Douglas & James K. Jones. 2009. Quantitative methods in corpus linguistics. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics. An International Handbook*. Vol. 2, 1286–1304. Berlin: Mouton de Gruyter.

Kilgarriff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2). 263–275.

Kübler, Sandra & Heike Zinsmeister. 2015. *Corpus Linguistics and Linguistically Annotated Corpora*. London: Bloomsbury.

van Langendonck, Willy. 2003. The dependency concept and its foundations. In Vilmos Ágel, Ludwig M. Eichinger, Hans Werner Eroms & Peter Hellwig (eds.), *Dependency and Valency. An International Handbook of Contemporary Research*. Vol. 1, 170–188. Berlin: Walter de Gruyter.

Manning, Christopher D. 2003. Probabilistic syntax. In Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probabilistic Linguistics*, 289–341. Cambridge, MA: MIT Press.

McEnery, Tony, Richard Xiao & Yukio Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. (Routledge Applied Linguistics.) London and New York: Routledge.

Sampson, Geoffrey. 2013. The Empirical Trend. Ten Years on. *International Journal of Corpus Linguistics* 18(2), 281–289.

Santorini, Beatrice. 1990. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)*. University of Pennsylvania, Technical Report.

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.

Weisser, Martin. 2016. *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis*. Oxford: Wiley Blackwell.

Zeldes, Amir. 2018. *Multilayer Corpus Studies*. (Routledge Advances in Corpus Linguistics 22.) London: Routledge.

Notice regarding sexual misconduct:
Please know that as a faculty member I am committed to supporting survivors of sexual misconduct, including relationship violence, sexual harassment and sexual assault. University policy also requires me to report any disclosures about sexual misconduct to the Title IX Coordinator, whose role is to coordinate the University's response to sexual misconduct.

Georgetown has a number of fully confidential professional resources who can provide support and assistance to survivors of sexual assault and other forms of sexual misconduct. These resources include:

Jen Schweer, MA, LPC
Associate Director of Health Education Services for Sexual Assault Response and Prevention
(202) 687-0323
jls242@georgetown.edu

Erica Shirley, Trauma Specialist
Counseling and Psychiatric Services (CAPS)
(202) 687-6985
els54@georgetown.edu

More information about campus resources and reporting sexual misconduct can be found at http://sexualassault.georgetown.edu.