Computational Corpus Linguistics

LING-4427 (Fall 2023)
Mon+Wed, 11:00-12:15

Instructor:
Amir Zeldes (Assoc. Prof. of Computational Linguistics)
E-Mail: amir.zeldes@georgetown.edu
Website: https://gucorpling.org/amir
Office: Wed 3:30-5:00 (Poulton Hall 243)

Teaching assistant: Wes Scivetti (wss37@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.

Learning goals:

By the end of this course students will know:

- How to create a new annotated corpus for computational linguistics research
- Practical and ethical considerations in data selection for building and using corpora
- Detailed knowledge of some of the most important frameworks for annotation today
- How to use query languages for corpus search and visualization systems
- Quantitative computational methods including collocation analysis and keyword analysis
- A survey of some of the most frequently used corpora in the field

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. These assignments will support the learning goals of mastering contemporary annotation frameworks and provide case studies to get to know prominent corpora and methods.

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:

https://gucorpling.org/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. This paper will challenge students to use the methods they have learned while taking care to use data in an informed, responsible and ethical way to tackle a real and novel research question.

Absences and timely assignment submission:
Students are expected to attend all classes and to complete all assignments on time. Absences may have an adverse effect on grades in a course, up to and including failure. That said, 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. Any additional absences for special reasons (religious observances, athletic travel, prolonged illness etc.) may be coordinated on a case by case basis with the instructor (documentation may be required as applicable).

Use of AI and conversational agents:
Although it is unlikely to be useful for most of the assignments in this class (specifically for annotation assignments), using AI agents such as ChatGPT, Bing AI or Google Bard is not forbidden, unless explicitly specified (especially for some writing assignments). AI assistants are becoming a part of our world, and are based on models trained on linguistically annotated corpora themselves, so we should learn about them, and realistically, students will use them if they remain in the field.

However, I recommend using AI models for homework assignments mainly to assist in formulations or grammar and spelling corrections, for example for non-native speakers. Using AI models for writing effectively requires sufficient experience to understand when their suggestions are right, wrong, or something in between. Overly relying on AI is likely to diminish your learning experience. Nevertheless, using AI models where appropriate is allowed.

Approximate course plan

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

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:

Aoyama, Tatsuya, Behzad, Shabnam, Gessler, Luke, Levine, Lauren, Lin, Jessica, Liu, Yang Janet, Peng, Siyao, Zhu, Yilun and Zeldes, Amir (2023) GENTLE: A Genre-Diverse Multilayer Challenge Set for English NLP and Linguistic Evaluation. In: *Proceedings of the Seventeenth Linguistic Annotation Workshop (LAW-XVII 2023)*, 166–178. Toronto, Canada.

Batinić, Josip, Elena Frick and Thomas Schmidt. 2021. Accessing spoken language corpora: an overview of current approaches. *Corpora* 16(3), 417-445.

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.

Crible, Ludivine. 2022. The syntax and semantics of coherence relations: From relative configurations to predictive signals. *International Journal of Corpus Linguistics* 27(1), 59-92.

Egbert, Jesse & Paul Baker. 2021. *Using Corpus Methods to Triangulate Linguistic Analysis*. London: Routledge.

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.

McConnell, Kyla & Alice Blumenthal-Dramé. 2022. Effects of task and corpus-derived association scores on the online processing of collocations. *Corpus Linguistics and Linguistic Theory* 18(1), 33-76.

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 Title IX/Sexual Misconduct:

Georgetown University and its faculty are committed to supporting survivors and those impacted by sexual misconduct, which includes sexual assault, sexual harassment, relationship violence, and stalking. Georgetown requires faculty members, unless otherwise designated as confidential, to report all disclosures of sexual misconduct to the University Title IX Coordinator or a Deputy Title IX Coordinator. If you disclose an incident of sexual misconduct to a professor in or outside of the classroom (with the exception of disclosures in papers), that faculty member must report the incident to the Title IX Coordinator, or Deputy Title IX Coordinator. The coordinator will, in turn, reach out to the student to provide support, resources, and the option to meet. [Please note that the student is not required to meet with the Title IX coordinator.]. More information about reporting options and resources can be found on the Sexual Misconduct Website: https://sexualassault.georgetown.edu/resourcecenter.

If you would prefer to speak to someone confidentially, Georgetown has a number of fully confidential professional resources that can provide support and assistance. These resources include:
Health Education Services for Sexual Assault Response and Prevention: confidential email sarp@georgetown.edu

Counseling and Psychiatric Services (CAPS): 202.687.6985 or after hours, call (833) 960-3006 to reach Fonemed, a telehealth service; individuals may ask for the on-call CAPS clinician

More information about reporting options and resources can be found on the Sexual Misconduct Website.