Introduction to Quantitative Textual Analysis

DH 0120 / CLS 181
Tufts University
Fall 2024
M 6–9 p.m.
Miner Hall 112
Instructor: Charles Pletcher (@pletcher)
TA: Zoë Spriggs (@zspriggs)
Office hours: By appointment (https://cal.com/pletcher)

Course Description

Despite the recent proliferation in digital sources for a variety of literary fields and subfields, quantitative textual analysis has often been viewed as anathema to pursuits in the humanities. Rather than set computational and literary methods at odds, this course seeks to reconcile them through careful application of statistical methodologies alongside literary modes of inquiry. Far from a positivistic approach to literary texts, this course will guide students towards enriching their understanding of the texts that they study by taking a "distant reading" approach (cf. Moretti 2000) that complements, rather than supplants, close reading and critical analysis.

As a class, we will be focusing on the text of Pausanias's *Hellados Periēgēsis* (*Description of Greece*), but you should feel free to start working on your own corpora as time and interest allow.

This course introduces humanists to the tools and methodologies of quantitative textual analysis through corpus linguistics. Using the Python programming language, students will learn how to build and evaluate corpora related to their areas of expertise, and they will gain experience with basic statistics and probability theory, hypothesis testing and experiment design, and methods in sociolinguists, stylistics, and diachronic textual analysis. Students will gain substantial practical experience through weekly labs and homework assignments, culminating in a final presentation and paper aimed at publication.

NB: Although programming is a major part of this course, the focus is on corpus linguistics and statistical methods, not on programming per se.

Prerequisites

None. Prior programming experience is welcome but not required.

Outcomes

This course will prepare students to:

- Describe the process of creating useful corpora

- Analyze linguistic corpora using statistical methods and the Python programming language
- Test and evaluate hypotheses based on these analyses
- Apply these techniques to literary works within their fields of expertise

Textbooks

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. 1st ed. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316410899.

Stefanowitsch, Anatol. 2020. *Corpus Linguistics: A Guide to the Methodology*. Berlin: Language Science Press. https://doi.org/10.5281/ZENODO.3735822.

Wynne, Martin, ed. 2004. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books. https://users.ox.ac.uk/~martinw/dlc/index.htm.

We will follow Brezina's textbook for most of the course, supplementing his focus on statistics with articles on techniques and case studies along the way. A PDF copy of Brezina's textbook is available through the library.

Requirements

1. Attendance and participation (20% of final grade). This course is fast-paced and multi-disciplinary, and it involves a substantial amount of teamwork. It is therefore critical that you come to class prepared, both to support your own work in the course and to support the work of your teams. Remember, your participation in class is not graded for correctness. If you have a question, please speak up, and don't be afraid to make mistakes.

Starting with the second class, we will have a brief "journal club" at the beginning of the session, during which a volunteer will present and lead a short discussion of a recent and/or important article relevant to the class. This is an opportunity not only to discuss methods and case studies but also to build a shared bibliography with the class, which will be helpful as you perform increasingly independent research and analyses over the course of the semester.

2. Homework/labs (40% of final grade). Each week's assignments involve a non-trivial lab exercise. Often, especially after the first couple of weeks, you will be required to collect, clean, and analyze data on your own or in pairs. Your submissions from the previous week should be submitted before the next class.

3. Final presentation (20% of final grade). Over the course of the semester, you will tackle various problems in corpus linguistics related to your area of expertise. This presentation is meant to give you space to present preliminary findings, get feedback from the class, and refine your research questions in preparation for the final paper.

4. Final project (20% of final grade). The final project will be the culmination of your work in this class. Ideally, you will have identified a journal to which you might like to submit this project, and your project will conform to the structure and length requirements of that journal. We will start planning for and working towards these projects as early in the semester as possible. Do not leave your project until the last minute.

For the final project and final presentation, you may optionally work with a coauthor/classmate. In that case, you will both be expected to submit the same work, and you will both receive the same grade. Please make sure to distribute your efforts accordingly.

Course Schedule

With a course as fast-paced as this one, class-time will be focused on addressing any questions and working collaboratively on labs in which you implement analyses similar to those discussed in the readings for that day.

Readings are to be done *before* class. Labs will be done *in* class, and must be finished before the next class.

x	Week	Reading Assignment	Journal Club	Homework
x	1	None	Syllabus Overview	Finish lab exercises
x	2	Brezina Ch. 2, pp. 38–41	Ehrett et al. 2024. "Shakespeare Machine."	Reading quiz
x	3	Brezina Ch. 2, pp. 42–65	Burns. 2019. "Text Pipeline"	Finish lab
x	4	Review of Ch. 2, pp. 42–65	Sanchez-Marco et al. 2010.	Dispersion lab
x	5	Brezina Ch. 3, pp. 66–75	TBD	Reading quiz
x	6	Brezina Ch. 3, pp. 76–101	TBD	Finish lab
	7	Brezina Ch. 4, pp. 102–117	TBD	Reading quiz
	8	Brezina Ch. 4, pp. 117–138	TBD	Finish lab
	9	Brezina Ch. 5, pp. 139–151	TBD	Reading quiz
	10	Brezina Ch. 5, pp. 151–182	TBD	Finish lab
	11	Brezina Ch. 6, pp. 183–199	TBD	Reading quiz
	12	Brezina Ch. 6, pp. 199–218	TBD	Finish lab

x	Week	Reading Assignment	Journal Club	Homework
	13	Brezina Ch. 7, pp. 219–235	TBD	Work on final project
	??	Brezina Ch. 7, pp. 235–256	TBD	Work on final project

Week 7 (November 4, 2024)

November 4: Building a corpus from TEI XML; word embeddings (https://www.tensorflow.org/text/guide/word_embeddings#create_a_classification_model) and Word2Vec (https://www.tensorflow.org/text/tutorials/word2vec)

HW: Finish word2vec lab, build and evaluate initial corpus

Week 8 (November 12, 2024)

November 18: Brezina ch. 5 (correlation measures); text classification (https://www.tensorflow.org/tutorials/keras/text_classification, https://www.tensorflow.org/text/tutorials/cl

HW: Finish text classification with BERT lab; fine-tuning lab (https://www.tensorflow.org/text/tutorials/bert_ or for Greek: https://huggingface.co/bowphs/GreBerta); finalize corpora for final projects

Week 9 (November 18, 2024)

November 25: Brezina ch. 7 (change over time), semantic similarity (https://www.tensorflow.org/hub/tutorials/semantic_similarity_with_tf_hub_universal_encoder)

HW: Finish in-class lab, work on final projects

Week 10 (November 25, 2024)

Independent work on final projects; additional labs/catch-up as needed

Week 11 (December 2, 2024)

Independent work on final projects; additional labs/catch-up as needed

Week 12 (December 9, 2024)

Final project presentations

Labs

Dispersion Lab

For this lab, you'll want to use a Colab notebook or a new repository on GitHub.

Your assignment is to calculate and visualize dispersion measures for a small corpus, consisting of Pausanias and one or more other texts.

You might start by identifying interesting lemmata, such as most frequent words (after removing stop words), hapaxes, etc. You can even identify clusters of words based on your own semantic understanding, if you wish.

> Remember, you'll want to report absolute and relative frequencies for the tokens, types, and lemmata in the works in your corpus.

You'll then want to calculate different dispersion measures. Start with Range2 to get a sense of what you're up against.

You should also consider reporting Juilland's *D*, Deviation of Proportions (DP) and Average Reduced Frequency (ARF) or other members of dispersion, depending on your hypotheses and the information that you are hoping to find.

> Remember to consider the tradeoffs associated with each statistic.

Finally, report on the basic lexical diversity of your corpus — and the parts of your corpus — using TTR, STTR, and/or MATTR.

If you need help designing experiments, the Exercises in @Brezina2018 ch. 2 (§2.8, p. 62–64) might provide helpful guidance.

Submission instructions Please upload your lab notebook or include a link to it on Canvas.

You should also write a short (4–6-page) report of your findings. Detail your methods and hypotheses, and try to include at least one visualization. (You can use Excel or matplotlib or something similar.)

Bibliography

Additional Bibliography

This bibliography will grow over the course of the semester as we add articles from journal club and other related readings.

- Berti, Monica, ed. 2019. *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution.* De Gruyter. https://doi.org/10.1515/9783110599572-201.
- Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide.* 1st ed. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316410899.
- Dobson, James E. 2022. "Vector Hermeneutics: On the Interpretation of Vector Space Models of Text." Digital Scholarship in the Humanities 37 (1): 81–93. https://doi.org/10.1093/llc/fqab079.
- Ehrett, Carl, Lucian Ghita, Dillon Ranwala, and Alison Menezes. 2024. "Shakespeare Machine: New AI-Based Technologies for Tex-

tual Analysis.” Digital Scholarship in the Humanities 39 (2): 522–31. https://doi.org/10.1093/llc/fqae021.

- Fleming, Paul. 2017. “Tragedy, for Example: Distant Reading and Exemplary Reading (Moretti).” New Literary History 48 (3): 437–55. https://doi.org/10.1353/nlh.2017.0021.
- Hanks, Patrick. 2013. *Lexical Analysis: Norms and Exploitations.* Cambridge: The MIT Press.
- Herrmann, J Berenike. 2023. “Tool Criticism in Practice. On Methods, Tools and Aims of Computational Literary Studies.” Digital Humanities Quarterly 17 (2).
- Honkapohja, Alpo, and Jukka Suomela. 2022. “Lexical and Function Words or Language and Text Type? Abbreviation Consistency in an Aligned Corpus of Latin and Middle English Plague Tracts.” Digital Scholarship in the Humanities 37 (3): 765–87. https://doi.org/10.1093/llc/fqab007.
- Hoover, David L. 2016. “Argument, Evidence, and the Limits of Digital Literary Studies.” In Debates in the Digital Humanities 2016, edited by Matthew K. Gold and Lauren F. Klein: 230–50. University of Minnesota Press. https://doi.org/10.5749/j.ctt1cn6thb.
- Moretti, Franco. 2000. “Conjectures on World Literature.” New Left Review 1: 54–68.
- Real Academia Española. n.d. “Banco de Datos (CORDE) [En Línea]. Corpus Diacrónico Del Español.” Accessed April 11, 2024. https://corpus.rae.es/cordenet.html.
- Säily, Tanja. 2014. “Sociolinguistic Variation in English Derivational Productivity’: Studies and Methods in Diachronic Corpus Linguistics.” Helsinki: University of Helsinki. https://helda.helsinki.fi/items/c320c8fe-255c-43c6-9403-a4a898212ed7.
- Stefanowitsch, Anatol. 2020. *Corpus Linguistics: A Guide to the Methodology.* Berlin: Language Science Press. https://doi.org/10.5281/ZENODO.3735822.
- Vatri, Alessandro, and Barbara McGillivray. 2020. “Lemmatization for Ancient Greek: An Experimental Assessment of the State of the Art.” Journal of Greek Linguistics 20 (2): 179–96. https://doi.org/10.1163/15699846-02002001.
- Wynne, Martin, ed. 2004. *Developing Linguistic Corpora: A Guide to Good Practice.* Oxford: Oxbow Books. https://users.ox.ac.uk/~martinw/dlc/index.htm.

Course and University Policies

Academic Integrity

Tufts holds its students strictly accountable for adherence to academic integrity. The consequences for violations can be severe. It is critical that you understand the requirements of ethical behavior and academic work as described in Tufts’ Academic Integrity handbook. If you ever have a question about the expectations concerning a particular assignment or project in this course, be sure to ask me for clarification. The Faculty of the School of Arts and Sciences and the School of Engineering are required to report suspected cases of academic integrity violations to the Dean of Student Affairs Office. If I suspect that you have cheated or plagiarized in this class, I must report the situation to the dean.

Religious Accommodations

Tufts University faculty, staff, and administration highly value and acknowledge the religious diversity of its student body. Students seeking religious accommodations related to their holy days are encouraged to collaborate with faculty to make arrangements during the first week of each semester. Consult the Multifaith Calendar for upcoming holidays, links to the University Religious Accommodations Policy, and members of the University Chaplaincy who are available to respond to questions on religious observances.

Accommodations for Students with Disabilities

Tufts is committed to providing equal access and support to all qualified students through the provision of reasonable accommodations. If you have a disability that requires reasonable accommodations, contact the StAAR Center at StaarCenter@tufts.edu or 617-627-4539. Please be aware that accommodations cannot be enacted retroactively, making timeliness a critical aspect for their provision.

Academic Support at the StAAR Center

The StAAR Center offers a variety of FREE resources to all students. Students may make an appointment to work on any writing-related project or assignment, attend subject tutoring in a variety of disciplines, or meet with an academic coach to hone skills like time management and navigating procrastination. Students can make an appointment for any of these services by visiting https://students.tufts.edu/staar-center.

Student Support, including Mental Health

As a student, there may be times when personal stressors or difficulties interfere with your academic performance or well-being. The Dean of Student Affairs Office offers support and care to undergraduates and graduate students who are experiencing difficulties, and can also aid faculty in their work with students. In addition, through Tufts' Counseling and Mental Health Service (CMHS) students can access mental health support 24/7, and they can provide information on additional resources. CMHS also provides confidential consultation, brief counseling, and urgent care at no cost for all Tufts undergraduates as well as for graduate students who have paid the student health fee. To make an appointment, call 617-627-3360. Please visit the CMHS website: http://go.tufts.edu/Counseling to learn more about their services and resources.

In-Person Classroom Health and Safety Policy

Tufts is not currently mandating mask-wearing in the classroom. I might wear a mask at times, and others should feel free to do so as well. We cannot know everyone's personal or familial health situations, so please take the steps you need to feel comfortable in an in-person setting. Should the university's masking requirements change, we will update our practices as well. Here is the link to the current Tufts COVID policy: Healthy@Tufts | Coronavirus (COVID-19).

Policy on sharing

This course is designed for everyone to feel comfortable participating in discussion, asking questions, learning, and facilitating the learning of others. In order for that atmosphere to be maintained, the recordings of our conversations will only be shared with the enrolled students in the class (not posted publicly) and it is prohibited for any of us who have access to the video to share it outside the course. Similarly, I have specifically designed the syllabus, exams, handouts, and lectures for the people who are enrolled in the course this term and those may not be shared outside this course. It is against Tufts policy for anyone to share any content made available in this course including course syllabi, reading materials, problems sets, videos, handouts, and exams, with anyone outside of the course without the express permission of the instructor. This especially includes any posting or sharing of videos or other recordings on publicly accessible websites or forums. Any such sharing or posting could violate copyright law or law that protects the privacy of student educational records.