Computational linguistics

Linguistics 290L
Fall 2020

MW 11am-12:30pm, online

Description: This course provides a graduate-level introduction to computational linguistics. We will explore computational principles and methods that cross-cut different branches of linguistics, and will apply those principles to replicate and extend computational analyses in a selection of published papers.

Instructor: Terry Regier (email: firstname dot lastname at berkeley dot edu).

Prerequisites and expectations: The course is open to graduate students in linguistics or related disciplines. Access for other students is by permission of instructor. Some basic prior experience with programming is necessary, but no prior experience with computational linguistics is required. Starter code for homework assignments will be provided, giving students a basis on which to build further. Programming will be in Python.

Learning goals, or what I hope you will gain from this class: Familiarity with computational principles and methods in linguistics, and experience in conducting computational analyses.

Recording: The class will be delivered remotely, by Zoom, and all class sessions will be recorded, to accommodate those who may benefit from being able to view these recordings asynchronously. Please note that all course materials, including all video recordings, are to remain internal to this class. Reposting to third party sites or any other form of redistribution is prohibited.

Grading:

- Attend office hours at least once during the semester (10%)
- In-class discussion of readings (45%)
- Homeworks, replicating and extending computational analyses in papers (45%)

Format:

- In general, Mondays are devoted to “thinking”, and Wednesdays to “doing”. Specifically, each week will proceed as follows.
    - Monday: lecture and discussion of readings. Please come to class prepared to ask questions about the readings, and to suggest comments or critiques that engage and go beyond the readings.
    - Wednesday: training in hands-on practical skills, review of the previous week’s homework assignment, and presentation and discussion of the next homework assignment.

Course policies

- All readings will be made available electronically.
- Please contact the instructor for accommodation of religious beliefs, disabilities, or other circumstances.
- Intellectual dishonesty of any sort will not be tolerated. The course will follow the University Policy on Cheating and Plagiarism.

Readings and schedule

Wed Aug 26: **Introduction and orientation**
Overview of course, and infrastructure for homeworks.

Mon Aug 31: **Languages, automata, and the Chomsky hierarchy**
[Optional, for background] Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, volume 2, issue 3, 113-124.
Sipser, M. (2013). Excerpt from Chapter 1, Regular languages, in *Introduction to the theory of computation, third edition*. **Read pp. 31-38**.
Chomsky hierarchy

Wed Sep 2: **Datasets: Multilingual pronunciation data from Wiktionary (WikiPron)**

- Wikipron
- Regular expressions in Python:
  - Background
  - Syntax
- Homework: Use regular expressions to manipulate WikiPron data.

Mon Sep 7: ***No class meeting – Labor day***

Wed Sep 9: **Datasets: US given names, Social Security Administration**

- Pandas tutorial
- US given names dataset
- Homework: Use Pandas to explore given names in the US over time.

Mon Sep 14: **Introduction to probability theory**

- Bishop, C.M. (2006). Probability theory (pp. 12-17). Excerpt from *Pattern Recognition and Machine Learning*. New York: Springer.
- Hayes, B. (2013). First links in the Markov chain. *American Scientist, 101*.
- Marr, D. (1982). Vision. New York: Freeman. Pp. 24-29.

Wed Sep 16: **Datasets: The CMU pronouncing dictionary (CMUDICT)**

- CMU pronouncing dictionary
- Homework: Phoneme transition probabilities in CMUDICT.

Mon Sep 21: **Introduction to information theory**

- Stone, J.V. (2018). Information theory: A tutorial introduction. Pay special attention to sections 1-6.

Wed Sep 23: **Datasets: CHILDES, in XML**

- CHILDES, the Child Language Data Exchange System
- XML parsing in Python
- Homework: Basic information-theoretic quantities in CHILDES.

Mon Sep 28: **Universals and variation in color naming**

- [Optional, for background] Roberson, D., et al. (2000). Color categories are not universal. *Journal of Experimental Psychology: General, 129*, 369-398.
- Regier, T. et al. (2005). Focal colors are universal after all. *PNAS, 102*, 8386-8391.

Wed Sep 30: **Datasets: The World Color Survey (WCS)**

- The World Color Survey
- Homework: Replicate results of Regier et al. (2005).

Mon Oct 5: **Word segmentation in speech**

- Hockema, S. (2006). Finding words in speech: An investigation of American English. *Language Learning and Development, 2*, 119-146.

Wed Oct 7: **Implementation**

- Homework: Replicate results of Hockema (2006).

Mon Oct 12: **Naive Bayes**

- [Optional] Mitchell, T. (2017). Generative and discriminative classifiers: Naive Bayes and logistic regression. Chapter 3 of *Machine Learning*. **Read pp. 1-7**.
- Jufafsky, D., & Martin, J.H. (2019). Naive Bayes and sentiment classification. Chapter 4 of *Speech and Language Processing, Third edition draft*.
- Michael, L., et al. (2014). Exploring phonological areality in the circum-Andean region using a naïve Bayes classifier. *Language Dynamics and Change, 4*, 27-86.

Wed Oct 14: **Datasets: The South American Phonological Inventory Database (SAPhon)**

- The South American Phonological Inventory Database (SAPhon)
- Naive Bayes in scikit-learn
- Homework: Replicate results of Michael et al. (2014).

Mon Oct 19: **Maximum entropy / logistic regression**

- [Optional] Mitchell, T. (2017). Generative and discriminative classifiers: Naive Bayes and logistic regression. Chapter 3 of *Machine Learning*. **Read pp. 7-16**.
- Jufafsky, D., & Martin, J.H. (2019). Logistic regression. Chapter 5 of *Speech and Language Processing, Third edition draft*.

Goldwater, S. & Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Workshop on Variation within Optimality Theory*, pp. 111-120.

Wed Oct 21: **Implementation**

Homework: Replicate results of Goldwater & Johnson (2003).

Mon Oct 26: **Clustering**

Hierarchical clustering in scikit-learn (section 2.3.6)

Hierarchical clustering

White, A.S., et al. (2014) Discovering classes of attitude verbs using subcategorization frame distributions. *Proceedings of the Northeast Linguistics Society 43*.

Wed Oct 28: **Implementation**

Homework: Replicate results of White et al. (2014).

Mon Nov 2: **Neural networks**

Jufafsky, D., & Martin, J.H. (2019). Neural networks and neural language models. Chapter 7 of *Speech and Language Processing, Third edition draft*. Excerpts.

Wed Nov 4: **Application**

Lewis, J.D., & Elman, J.E. (2001). A connectionist investigation of linguistic arguments from the poverty of the stimulus: Learning the unlearnable. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*.

No homework.

Mon Nov 9: **TBD** - swing space for now.

Wed Nov 11: ***No class meeting – Veterans' day***

Mon Nov 16: **Categorical perception**

[Optional, for background] Huttenlocher, J., et al. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review, 98*, 352-376.

Feldman, N.H., et al. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review, 116*, 752–782.

Wed Nov 18: **Implementation**

Homework: Replicate results of Feldman et al. (2009).

Mon Nov 23: **"Guest" lecture**

Research talk by Terry Regier

Wed Nov 25: ***No class meeting – Thanksgiving break***

Mon Nov 30: **Bayesian phylogenetics**

Huelsenbeck, J.P. et al. (n.d.). An introduction to Bayesian inference of phylogeny.

Sicoli, M.A. & Holton, G. (2014). Linguistic phylogenies support back-migration from Beringia to Asia. *PLOS ONE, 9*, e91722.

Wed Dec 2: **Summary and conclusions**