Measuring and Modeling Language Change

Jacob Eisenstein
Georgia Institute of Technology and Facebook Artificial Intelligence Research
me@jacob-eisenstein.com

1 Description

This tutorial is designed to help researchers answer the following sorts of questions about how language usage varies over time:

- Are people happier on the weekend?
- What was 1861's word of the year?
- Are Democrats and Republicans more different than ever?
- When did "gay" stop meaning "happy"?
- Are gender stereotypes getting weaker, stronger, or just different?
- Who is a leader, and who is a follower?
- How can we get internet users to be more polite and objective?

Such questions are fundamental to the social sciences and humanities, and scholars in these disciplines are turning to computational techniques for answers (e.g., Evans and Aceves, 2016; Underwood et al., 2018; Barron et al., 2018). Meanwhile, the ACL community is increasingly engaged with data that varies across time (e.g., Rayson et al., 2007; Yang and Eisenstein, 2016), and with the social insights that can be offered by analyzing temporal patterns and trends (e.g., Tsur et al., 2015). The purpose of this tutorial is to facilitate this convergence in two main ways.

First, by synthesizing recent computational techniques for handling and modeling temporal data, such as dynamic word embeddings, the tutorial will provide a starting point for future computational research. It will also identify useful text analytic tools for social scientists and digital humanities scholars, such as dynamic topic models and dynamic word embeddings.

Second, the tutorial will provide an overview of techniques and datasets from the quantitative social sciences and the digital humanities, which are not well-known in the computational linguistics community. These techniques include hypothesis testing, survival analysis, Hawkes processes, and causal inference. Datasets include historical newspaper archives, social media, and corpora of contemporary political speech.

1.1 Format

The format of this three-hour tutorial will combine lecture-style surveys of various research areas with interactive coding demonstrations. The coding demonstrations will use Jupyter notebook and the numpy, scipy, and pandas libraries. These notebooks will be shared along with publicly available data in a github repository for the tutorial.[1]

1.2 Scope

This tutorial is focused on **corpus-based** methods for measuring and modeling changes in language usage from time-stamped documents. Another body of research is built on **type-level** resources, such as lists of aligned words across languages, which can support phylogenetic analysis of language history (e.g. Gray and Atkinson, 2003; Bouchard-Côté et al., 2013). Other researchers use **simulation** to test the consequences of theoretical models of language change (e.g. Niyogi and Berwick, 1997; Cotterell et al., 2018). Finally, sociolinguists make use of **apparent time**, a technique for measuring language change by comparing the speech of individuals of various ages (e.g., Tagliamonte and D'Arcy, 2009). These three methods all contribute to our overall understanding of language change, but in the interest of a compact and coherent presentation, this tutorial will focus exclusively on corpus-based techniques.

[1] https://github.com/jacobeisenstein/language-change-tutorial

Proceedings of NAACL HLT 2019: Tutorial Abstracts, pages 9–14
Minneapolis, Minnesota, June 2, 2019.

The tutorial will engage with statistical analysis (e.g., hypothesis testing, causal inference) to a greater extent than most NAACL papers. Every effort will be made to make this material accessible to the typical NAACL attendee.

2 Topics

The bulk of the tutorial consists of hands-on exploration of time-stamped textual data, which will be conducted in the form of Jupyter notebooks. These practical sessions will be book-ended by an introduction to theoretical and methodological perspectives on language change, and a brief discussion of open questions for future work.

2.1 How and why to measure language change?

The tutorial begins with a survey of theoretical questions and associated methodological approaches. Sociolinguists and historical linguists are interested in changes to the linguistic system (Weinreich et al., 1968; Pierrehumbert, 2010); digital humanists model changes in text over time to track the evolution of cultural and literary practices (Michel et al., 2011); computational social scientists use time-stamped corpora to understand the transmission and evolution of social practices (Kooti et al., 2012; Garg et al., 2018) and to identify causes and effects in social systems (Bernal et al., 2017; Chandrasekharan et al., 2018). We will survey some of the ways in which various disciplines approach language change, and briefly discuss alternatives to the corpus-based perspective taken in this tutorial.

2.2 Tracking changes in word frequency

Question: Are people happier on the weekend?

Data: Twitter sentiment (Golder and Macy, 2011)

Methods: hypothesis testing, regression, python dataframes

In a seminal paper in social media analysis, Golder and Macy (2011) use Twitter data to quantify sentiment by time-of-day and day-of-the-week. This provides an opportunity to apply fundamental methods in quantitative social science to a time-stamped corpus of text, while gaining familiarity with the python data science stack. We will replicate the results of Golder and Macy, and then extend them, exploring Simpson's paradox and questions of representativeness (Biber, 1993; Pechenick et al., 2015).

2.3 Quantifying differences over time

Question: Are Democrats and Republicans more polarized than ever?

Data: Legislative floor speeches (Gentzkow et al., 2016)

Methods: topic models, information theory, randomization

Many observers have concluded that American politicians are increasingly polarized. Voting records are the main empirical foundation for this claim (e.g., Bateman et al., 2016), but legislative votes may be taken for non-ideological reasons, such as party discipline (Peterson and Spirling, 2018). Text analysis has therefore been proposed as a technique for quantifying ideological differences across groups, via either individual word frequencies (Monroe et al., 2008; Gentzkow et al., 2016) or latent topics (Tsur et al., 2015; Barron et al., 2018). Similar techniques can be used to track similarity and difference across literary genres (Underwood et al., 2018), academic conferences (Hall et al., 2008), and social media communities (Danescu-Niculescu-Mizil et al., 2013). In this section, we will apply language models, topic models, and information theory to a dataset of legislative speech, quantifying the textual distance between U.S. political parties over time.

2.4 Detecting changes in meaning

Question: When did money become something you can launder?

Data: Legal opinions from `courtlistener.com`

Methods: dynamic word embeddings

Word embeddings capture lexical semantics in vector form, but word meaning can change over time through a variety of linguistic mechanisms (Tahmasebi et al., 2018). This section will survey methods for computing *diachronic* word embeddings, which are parameterized by time (Wijaya and Yeniterzi, 2011; Kulkarni et al., 2015; Hamilton et al., 2016; Garg et al., 2018; Rudolph and Blei, 2018; Rosenfeld and Erk, 2018). We will investigate the application of one such method to a corpus of historical texts, identifying words with particularly fluid semantics, and teasing apart these different meanings.

2.5 Distinguishing leaders and followers

Question: Who is setting the terms of the debate?

Data: 2012 Republican primary debates (Nguyen et al., 2014)

Methods: Granger causation, Hawkes Process

Language changes have leaders and followers, and there is considerable interest in identifying the specific individuals and types of individuals who drive change (Dietz et al., 2007; Gerrish and Blei, 2010; Kooti et al., 2012; Eisenstein et al., 2014; Goel et al., 2016; Gerow et al., 2018; Del Tredici and Fernández, 2018). We will explore data from the 2012 Republican primary debates (Nguyen et al., 2014), applying a Hawkes process model to try to identify individuals whose language most shaped the terms of the debate. This section will also cover epidemiological models that attempt to predict *who* will be affected next in a cascade, and to quantify the factors that make an individual more or less susceptible (Soni et al., 2018).

2.6 Predicting the future

Question: Which innovations will persist?

Data: Reddit neologisms (Stewart and Eisenstein, 2018)

Methods: survival analysis

Some changes pass the test of time, but others are ephemera (Dury and Drouin, 2009). Is it possible to predict what will happen in advance? By attacking this problem, we hope to better understand the social and linguistic mechanisms that underlie language change (Chesley and Baayen, 2010; Del Tredici and Fernández, 2018; Stewart and Eisenstein, 2018). The dataset for this evaluation will consist of a set of lexical innovations from Reddit. We will build models to predict not only which will survive, but for how long.

2.7 Causation and the arrow of time

Question: Can internet policies make people be nicer?

Data: Counts of hate speech lexicons on Reddit (Chandrasekharan et al., 2018)

Methods: interrupted time series

Because causes precede effects, it is natural to ask whether temporal data can support causal inferences. This section will begin by reviewing the potential outcomes framework, which is the classical approach to causal inference from observational data (Rosenbaum, 2017). This framework is based on three main concepts: *treatment* (the manipulation of the environment whose effect we want to test), *outcome* (the quantity to measure), and *confounds* (additional variables that are probabilistically associated with both the treatment and effect). We will discuss how the potential outcomes framework can apply to temporal data through the interrupted time series model (Bernal et al., 2017), and we will experiment with the impact of a discrete policy treatment on textual outcomes in social media (Chandrasekharan et al., 2018; Pavalanathan et al., 2018). This section will also briefly survey approaches to modeling text as a treatment (Fong and Grimmer, 2016; Egami et al., 2018).

2.8 What's next?

We will conclude with a discussion of open research questions for the analysis of language change and diachronic textual corpora (Nerbonne, 2010; Eisenstein, 2013; Maurits and Griffiths, 2014; Perek, 2014).

3 Presenter

Jacob Eisenstein is Associate Professor in the School of Interactive Computing at the Georgia Institute of Technology, which he joined in 2012. He is on sabbatical at Facebook Artificial Intelligence Research in Seattle. His research on computational sociolinguistics is supported by an NSF CAREER award and by a young investigator award from the Air Force Office of Scientific Research (AFOSR). Results from this research have been published in traditional natural language processing venues, in sociolinguistics journals, and in more general venues. Jacob's Georgia Tech course on Computational Social Science covers some of the same themes as this tutorial, and includes some additional material.[2] He recently completed an introductory textbook on natural language processing.

References

Alexander TJ Barron, Jenny Huang, Rebecca L Spang, and Simon DeDeo. 2018. Individuals, institutions, and innovation in the debates of the french revolution. *Proceedings of the National Academy of Sciences*, 115(18):4607–4612.

David A. Bateman, Joshua D. Clinton, and John S. Lapinski. 2016. A house divided? roll calls, po-

[2] https://github.com/jacobeisenstein/gt-css-class

larization, and policy differences in the u.s. house, 1877–2011. 61:698–714.

James Lopez Bernal, Steven Cummins, and Antonio Gasparrini. 2017. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International journal of epidemiology*, 46(1):348–355.

Douglas Biber. 1993. Representativeness in corpus design. *Literary and linguistic computing*, 8(4):243–257.

Alexandre Bouchard-Côté, David Hall, Thomas L Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, page 201204678.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2018. You can't stay here: The effectiveness of Reddit's 2015 ban through the lens of hate speech. In *Proceedings of Computer-Supported Cooperative Work (CSCW)*.

Paula Chesley and R Harald Baayen. 2010. Predicting new words from newer words: Lexical borrowings in french. *Linguistics*, 48(6).

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2018. On the diachronic stability of irregularity in inflectional morphology. *arXiv preprint arXiv:1804.08262*.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 307–318.

Marco Del Tredici and Raquel Fernández. 2018. The road to success: Assessing the fate of linguistic innovations in online communities. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1591–1603.

Laura Dietz, Steffen Bickel, and Tobias Scheffer. 2007. Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning*, pages 233–240. ACM.

Pascaline Dury and Patrick Drouin. 2009. When terms disappear from a specialized lexicon: A semi-automatic investigation into necrology. In *Actes de la conférence internationale" Language for Special Purposes"(LSP 2009)*.

Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2018. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 359–369.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE*, 9.

James A Evans and Pedro Aceves. 2016. Machine translation: mining text for social theory. *Annual Review of Sociology*, 42.

Christian Fong and Justin Grimmer. 2016. Discovery of treatments from text corpora. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1600–1609.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Matthew Gentzkow, Jesse Shapiro, and Matt Taddy. 2016. Measuring polarization in high-dimensional data: Method and application to congressional speech. Technical Report 22423, NBER Working Papers.

Aaron Gerow, Yuening Hu, Jordan Boyd-Graber, David M Blei, and James A Evans. 2018. Measuring discursive influence across scholarship. *Proceedings of the National Academy of Sciences*, page 201719792.

Sean Gerrish and David M Blei. 2010. A language-based approach to measuring scholarly impact. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 375–382.

Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. 2016. The social dynamics of language change in online networks. In *The International Conference on Social Informatics (SocInfo)*.

Scott A Golder and Michael W Macy. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881.

Russell D Gray and Quentin D Atkinson. 2003. Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, 426(6965):435.

David Hall, Daniel Jurafsky, and Christopher D Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 363–371.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Farshad Kooti, Haeryun Yang, Meeyoung Cha, P. Krishna Gummadi, and Winter A Mason. 2012. The emergence of conventions in online social networks. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 194–201.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 625–635.

Luke Maurits and Thomas L Griffiths. 2014. Tracing the roots of syntax with bayesian phylogenetics. *Proceedings of the National Academy of Sciences*, 111(37):13576–13581.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

John Nerbonne. 2010. Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1559):3821–3828.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, Deborah A Cai, Jennifer E Midberry, and Yuanxin Wang. 2014. Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, 95(3):381–421.

Partha Niyogi and Robert C. Berwick. 1997. A dynamical systems model for language change. *Complex Systems*, 11(3):161–204.

Umashanthi Pavalanathan, Xiaochuang Han, and Jacob Eisenstein. 2018. Mind your POV: Convergence of articles and editors towards wikipedia's neutrality norm. In *Proceedings of Computer-Supported Cooperative Work (CSCW)*.

Eitan Adam Pechenick, Christopher M Danforth, and Peter Sheridan Dodds. 2015. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, 10(10):e0137041.

Florent Perek. 2014. Vector spaces for historical linguistics: Using distributional semantics to study syntactic productivity in diachrony. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 309–314.

Andrew Peterson and Arthur Spirling. 2018. Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems. *Political Analysis*, 26(1):120–128.

Janet B. Pierrehumbert. 2010. The dynamic lexicon. In A. Cohn, M. Huffman, and C. Fougeron, editors, *Handbook of Laboratory Phonology*, pages 173–183. Oxford University Press.

Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. 2007. Tagging the bard: Evaluating the accuracy of a modern pos tagger on early modern english corpora. In *Corpus Linguistics Conference*.

Paul R Rosenbaum. 2017. *Observation and experiment: an introduction to causal inference*. Harvard University Press.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 474–484.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1003–1011. International World Wide Web Conferences Steering Committee.

Sandeep Soni, Shawn Ling Ramirez, and Jacob Eisenstein. 2018. Discriminative modeling of social influence for prediction and explanation in event cascades. *arXiv preprint arXiv:1802.06138*.

Ian Stewart and Jacob Eisenstein. 2018. Making "fetch" happen: The influence of social and linguistic context on the success of lexical innovations. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Sali A. Tagliamonte and Alexandra D'Arcy. 2009. Peaks beyond phonology: Adolescence, incrementation, and language change. *Language*, 85(1):58–108.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *arXiv preprint arXiv:1811.06278*.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1629–1638.

Ted Underwood, David Bamman, and Sabrina Lee. 2018. The transformation of gender in english-language fiction. *Cultural Analytics*.

Uriel Weinreich, William Labov, and Marvin Herzog. 1968. Empirical foundations for a theory of language change. *Directions for historical linguistics*, pages 97–188.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 international*

workshop on DETecting and Exploiting Cultural diversiTy on the social web, pages 35–40. ACM.

Yi Yang and Jacob Eisenstein. 2016. Part-of-speech tagging for historical English. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.