Research Statement — Brendan O’Connor
Carnegie Mellon University, Nov. 2013

I study society by computationally analyzing text. Corpora of news, books, and social media encode human beliefs and culture. But it is impossible for a person to read all of today’s rapidly growing text archives. Automated text analysis scales to large data sets, and can assist in discovering patterns and themes, in areas from political science to literature to sociolinguistics (O’Connor et al., 2011). Driven by questions in these social science fields, I develop new **statistical machine learning** and **natural language processing** methods to help answer them—for example, to predict international conflict by analyzing millions of news articles. This is part of the emerging area of **computational social science** (Lazer et al., 2009).

In my work, I often interact and collaborate with colleagues in the social sciences. I advance methods in three deeply intertwined areas:

- Statistical methods: Textual data is noisy and high-dimensional. How do we reliably infer the latent structures behind the data, and tie them to important questions in the domain?
- Computational methods: These datasets can include millions of documents, analyzed with Bayesian inference over as many as tens of millions of parameters. How do we efficiently calculate quantities of interest, with abstractions that generalize to multiple problems?
- Linguistic methods: Text encodes relational meaning, a key element of social interactions. How do we extract rich belief and event structures from text?

These are recurring themes through many projects, including in social measurement from text (§1), advancing deeper text meaning analysis to support it (§2), and analyzing social media and online behavior (§3). In future work, I look forward to advancing the science of computational social analysis (§4).

1 Social measurement from text

In a social measurement problem, the task is to infer the aggregate aspects or activities of a population. In *text-based* social measurement, we analyze textual records to help conduct this inference. Here are two examples.

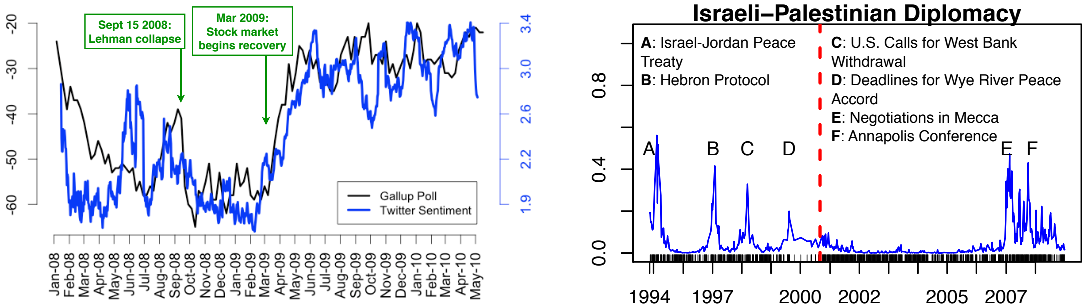

Figure 1: Text-derived time series from two different projects, in public opinion (O’Connor et al., 2010a) and international relations (O’Connor et al., 2013). (a) Aggregate sentiment analysis of jobs-related messages on Twitter, compared to Gallup consumer confidence polling. (b) Posterior frequency of diplomatic Israeli-Palestinian actions from newswire articles, compared to selected major events.

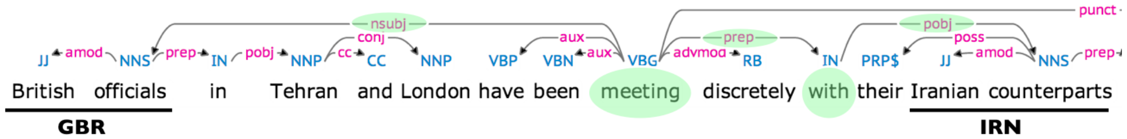

Figure 2: NLP event extraction for international relations, using a syntactic dependency graph.

The measurement of public opinion through surveys informs policy and politics; for example, surveys of consumer confidence aid business and government decision making. Is it possible to **measure public opinion from social media data**? In O'Connor et al. (2010a) I examined whether sentiment on Twitter might correlate, over time, with public opinion polls on several economic and political topics. Analyzing the time series of lexicon-based frequency measures, we found high variability in social media data, though for certain topics there were relatively strong correlations against tracking polls (Figure 1). Our approach was to carefully validate a proposed NLP method by comparing it to a well-understood traditional measurement instrument, randomly sampled telephone-based surveys. This work has been influential in highlighting the potential of social measurement via social media as a complement to existing methods—for example, it's discussed in the presidential address of the American Association for Public Opinion Research (AAPOR), when examining the future of their field (Newport, 2011), and I was invited to present this work at an AAPOR conference panel (2012).

In another project, I seek to measure **cooperation and conflict between countries**. Many analysts and political scientists are interested data of events reported in the news: who did what to whom and when. But this information is scattered among millions of news articles; it has to be extracted, and then can be used to study international relations, forecast future conflict, and understand biases in reporting.

In O'Connor et al. (2013), I leveraged syntactic parsing technology to develop a news event extraction system, and developed an unsupervised Bayesian latent variable model of the events. This model learns a useful ontology of the types of events that occur between countries (see §2). Simultaneously, it also learns how the relationships of every pair of countries has changed through time (Figure 1), and we validate it by comparing its historical inferences to records of real-world conflicts.

2 Deeper analysis of text meaning

In social text analysis, we often use "bag-of-words" methods based on word counts, ignoring the relationships between words in a document. While useful, these methods cannot deal with linguistic structure. For example, countries X and Y and the word "attack" might be mentioned in the same article, but is X attacking Y, or the other way around? In several projects, I advance deeper analysis of *text meaning* to help uncover the deeper belief and event structures of society.

Schema/Frame-semantic learning. My international relations work builds on a theoretical literature in linguistics and artificial intelligence of *schemas* or *frame semantics* (Minsky, 1974; Schank and Abelson, 1977; Fillmore, 1982), which are latent knowledge structures that describe typical event types and their participants. The types of actions between countries constitute event types which our model learns from text: for example, it learns different clusters of verbs that correspond to diplomacy (*meet with*, *sign with*, *praise*) versus military action (*kill*, *fire at*). I take a **machine learning approach** to frame semantics (building on earlier work (O'Connor, 2013)), in which I give the model absolutely no knowledge about the category system. Instead, it learns the ontology com-

ikr	smh	he	asked	fir	yo	last	name	so	he	can	add	u	on	fb	lololol
Int	Gen	Pro	Verb	Prp	Det	Adj	Noun	Prp	Pro	Verb	Verb	Pro	Prp	Proper	Int

Figure 3: Why Twitter NLP is hard: output from our POS tagger on a tweet with non-standard words, spellings, and capitalization. Our tagger correctly identifies *fir* as a preposition ("for"), *fb* as a proper noun ("Facebook"), and *ikr* as an interjection ("I know, right?"). Traditional NLP tools get these wrong, because they are designed for newspaper text. For example, the Stanford POS Tagger makes 7/16 errors here.

pletely automatically from the raw text. This contrasts to the knowledge engineering approach previously used in this area, which spent more than a decade to manually build an ontology containing tens of thousands of textual patterns. (This previous work is enormously valuable: for example, we also evaluate our system by seeing how well it reconstructs this expert-designed ontology, and are exploring how to incorporate it into a joint model.) And unlike traditional machine learning with bag-of-words representations, this requires modeling sophisticated **linguistic structures** from the text, such as syntactic parses (Figure 2) and coreference (O'Connor and Heilman, 2013), derived from NLP tools.

In related work, I used another version of my frame model to learn the latent types of **character personas** that appear in movie plot summaries (Bamman et al., 2013). Character types are learned to correspond to clusters of actions and attributes for each character. The model learns recurring character personas, like "Action Hero" or "Villain," from the text, which we again compare to pre-existing ontologies. This demonstrates the powerful flexibility of Bayesian language modeling: the same core parsers and models can be reused to answer questions in another area, digital humanities.

Social media NLP. Ideally, frame-semantic analysis could also be performed on social media; for example, to support rich analysis of sentiment or ideology. Unfortunately, most current NLP tools, designed for newspaper-like text, are extremely inaccurate when applied to the more informal and creative language in social media. In early work (O'Connor et al., 2010c) I found that even tokenization is challenging. Since part-of-speech (POS) tagging is a further prerequisite for complex linguistic analysis, at CMU I helped lead a large collaboration that developed a novel part-of-speech annotated dataset for Twitter, with a grammatical typology customized to its unique phenomena (Figure 3). Using semi-supervised learning, we developed a state-of-the-art, open source **part-of-speech tagger for Twitter** that has been downloaded thousands of times and used in both industry and academia (Gimpel et al., 2011; Owoputi et al., 2013). For example, it has been used to extract disaster-relevant information from social media (Imran et al., 2013), or to build state-of-the-art sentiment analysis (Mohammad et al., 2013). In ongoing work we seek to develop dependency parsing for Twitter; as a first step, I helped design our syntactic dependency formalism for informal text (Schneider et al., 2013).

3 Social science from online data

I believe the primary goal of computational social science is to develop tools and methodologies to better study society. I often use traditional information sources like mainstream news media or the U.S. Census in my work, which helps connect to long-standing research traditions in the social sciences. At the same time, **online data** is a powerful new resource for social science research, as more and more social behavior is digitized. In several other projects I explore its potential.

Sociolinguistics in social media. People's everyday language is now recorded in the form of online conversational text, such as SMS and social media. This new medium is filled with novel words and expressions, from which we can track language evolution by observing the

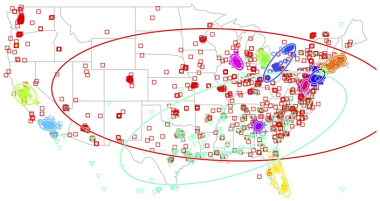

Figure 4: Geographical linguistic communities, as identified by the Geographic Topic Model.

linguistic behavior of millions of people. In a series of projects, we have found that variation of language in social media seems to be reproducing the fault lines of society. From the geographic location of tweets from mobile phones, we find surprisingly region-specific word usage patterns, and can even predict a person's location from the text of their messages (Eisenstein et al., 2010). By cross-referencing locations against U.S. Census data, we find demographic-linguistic relationships (O'Connor et al., 2010b). Given several years of Twitter data, we can track the invention and spread of new slang terms among regions and communities, and statistically test sociolinguistic theories about the determinants of linguistic diffusion (Eisenstein et al. (2012), and journal version under submission).

Media analysis. In my examples of text-based social measurement (§1), an important question is whether we are measuring the true state of society, or are measuring aspects of the media system that produced the text: biases in who expresses opinions on Twitter, or biases in international news reporting, etc. These are crucial topics to address in future work. In one project I've directly analyzed a media system: In (Bamman et al., 2012), we analyzed messages that tend to become deleted on the Chinese microblog service Sina Weibo, and were able to detect patterns of **political content censorship** in action.

Crowdsourcing. When I worked at startups before grad school, we stumbled upon the then-little-known **Mechanical Turk** service, and found it to be a very powerful tool for NLP and information retrieval annotations, as well as various business uses and even human behavioral studies. We had hundreds of thousands of visitors to our blog that informally illustrated such experiments, and our paper on NLP applications (Snow et al., 2008) is a key early paper on crowdsourcing, having inspired workshops[1] and startups.[2] Google Scholar currently lists our work as the most-cited computational linguistics paper in the last five years[3].

4 The future of computational social *science*

My research develops methods for *computational social science with text analysis* (reviewed in O'Connor et al. (2011)), with a variety of applications. Some of these applications areas are very interesting not just to us, but outside academia as well: our various projects on social media research (including sentiment/polls, geography/slang, and censorship) have been covered in hundreds of news articles, including in the *New York Times*, the *Wall Street Journal*, etc. Large-scale analysis of novel, socially relevant online data is of great interest. While it is exciting and important to engage a larger audience, the scientific understanding of these areas is still underdeveloped.

This needs to change. In my work and teaching, I seek to advance the **scientific validity** of computational social science methods, putting them in context with rich research traditions in the social sciences, linguistics, and statistics.

The first important aspect is **statistical and computational methodology**. I chose to go to a Machine Learning Department for graduate school because I think *statistics and machine learning are crucial tools for the future of social science*. I use a wide variety of such techniques in my research—generalized linear models, (ad)mixture models, Bayesian and frequentist estimation, distributed

[1] NAACL 2010, "Creating Speech and Language Data With Amazons Mechanical Turk" (Callison-Burch, personal communication)

[2] Coursera's peer grading system (Andrew Ng, EMNLP 2013 keynote), and the crowdsourcing company Crowdflower, Inc.

[3] http://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computationallinguistics

computation, large-scale optimization, false discovery rate control, dynamic programming, geographic data tools, etc. All these computational and statistical techniques will be a key part of my future work. I also want to teach them to future students for use in both research and industry. The huge amount of interest in **"data science" and "big data"** is driven in part by the necessity of a unified approach to using and teaching these types of tools.

Furthermore, when dealing with text, **linguistics** forms a second methodological pillar. My work shows that *a little bit of NLP can go a long way*: by using computational linguistic techniques founded in syntactic and semantic theory, we uncover socially relevant belief structures. I am very interested in further pursuing rich linguistic analysis methods; the social analysis questions will drive better definition and understanding of the relevant NLP problems.

Finally, besides linguistic, statistical, and computational methodology, it is critical to ground research in substantive **domain questions** from the social sciences and humanities. For example, our international relations project is in the context of a long-standing political science literature on extracting and analyzing event time series; we seek address a specific methodological problem they have (the time-consuming manual development of event ontologies and their dictionaries), which affects how to answer political science questions, such as understanding the causes of conflict.

In general, I have been fortunate to work with researchers not just in computer science but also with backgrounds in linguistics, political science, philosophy, classics, economics, and biology, and this diversity of academic training has brought unique perspectives to the table. While **collaborations** are always hard, they're also the most fun and effective way to perform research, and I plan to continue this way in the future.

References

David Bamman, **Brendan O'Connor**, and Noah A. Smith. Censorship and deletion practices in Chinese social media. *First Monday*, 17(3), 2012.

David Bamman, **Brendan O'Connor**, and Noah A. Smith. Learning latent personas of film characters. In *Proceedings of ACL*, 2013.

Jacob Eisenstein, **Brendan O'Connor**, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277—1287, 2010.

Jacob Eisenstein, **Brendan O'Connor**, Noah A. Smith, and Eric P. Xing. Mapping the geographical diffusion of new words. In *NIPS Workshop on Social Network and Social Media Analysis*, 2012. URL http://arxiv.org/abs/1210.5268.

Charles Fillmore. Frame semantics. *Linguistics in the morning calm*, pages 111–137, 1982. URL http://brenocon.com/Fillmore%201982_2up.pdf.

Kevin Gimpel, Nathan Schneider, **Brendan O'Connor**, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeff Flanigan, and Noah A. Smith. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47. Association for Computational Linguistics, 2011.

Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1021–1024. International World Wide Web Conferences Steering Committee, 2013.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Lszl Barabsi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323(5915):721–723, February 2009. doi: 10.1126/science.1167742. URL http://www.sciencemag.org/content/323/5915/721.short.

Marvin Minsky. A framework for representing knowledge. *MIT-AI Laboratory Memo 306*, 1974. URL http://web.media.mit.edu/~minsky/papers/Frames/frames.html.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June 2013. URL http://www.umiacs.umd.edu/~saif/WebPages/Abstracts/NRC-SentimentAnalysis.htm.

Frank Newport. Presidential address: Taking AAPOR's mission to heart. *Public Opinion Quarterly*, 75(3):593–604, 2011. URL http://poq.oxfordjournals.org/content/75/3/593.

Brendan O'Connor. Learning frames from text with an unsupervised latent variable model. 2013. URL http://arxiv.org/abs/1307.7382.

Brendan O'Connor and Michael Heilman. ARKref: a rule-based coreference resolution system. 2013. URL http://arxiv.org/abs/1310.1975.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *International AAAI Conference on Weblogs and Social Media, Washington, DC*, 2010a.

Brendan O'Connor, Jacob Eisenstein, Eric P. Xing, and Noah A. Smith. A mixture model of demographic lexical variation. In *NIPS Workshop on Machine Learning for Social Computing*, 2010b.

Brendan O'Connor, Michel Krieger, and David Ahn. TweetMotif: Exploratory search and topic summarization for Twitter. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010c.

Brendan O'Connor, David Bamman, and Noah A. Smith. Computational text analysis for social science: Model assumptions and complexity. In *Second Workshop on Comptuational Social Science and the Wisdom of Crowds (NIPS 2011)*, 2011.

Brendan O'Connor, Brandon Stewart, and Noah A. Smith. Learning to extract international relations from political context. In *Proceedings of ACL*, 2013.

Olutobi Owoputi, **Brendan O'Connor**, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL*, 2013.

Roger C. Schank and Robert P. Abelson. *Scripts, Plans, Goals, Understanding: An Inquiry Into Human Knowledge Structures*. Psychology Press, 1977.

Nathan Schneider, **Brendan O’Connor**, Naomi Saphra, David Bamman, Manaal Faruqui, Noah A. Smith, Chris Dyer, and Jason Baldridge. A framework for (under)specifying dependency syntax without overloading annotators. In *Linguistic Annotation Workshop, Proceedings of ACL*, 2013. Extended version, arXiv preprint arXiv:1306.2091.

Rion Snow, **Brendan O’Connor**, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics, 2008.