Modeling Language Games

Chris Proctor

Stanford University, cs224w

Abstract. A central belief of the learning sciences is that learning is situated within communities of practice. This paper demonstrates a method for modeling learning within an online discourse community as two intersecting phenomena: members' trajectories of participation and change in the community's language norms. Using the word2vec algorithm, a language model is trained on 12 million comments over a decade, capturing a snapshot every month. This paper replicates prior findings showing that users' individual language stabilizes while community norms progress, until users depart the community. We also show how the discourse community's word meanings shift over time with respect to relational axes such as gender and morality.

1 Introduction

The last five years have seen great advances in natural language processing. Word2Vec (Mikholov et al, 2013) uses an embedding which maps words to high-dimensional vectors such that nearby words have similar meaning. Word vectors are trained on a prediction task which iteratively passes over a large corpus, taking each word and predicting words that will appear in a surrounding context window. GloVe (Pennington, Socher, & Manning, 2014) takes a different approach, considering global word covariance statistics. GloVe is excellent at identifying word synonyms and also completing analogy tasks via linear substructures. Both models are effective in practice for a wide range of NLP tasks (Baroni, Dinu, & Kruszewski, 2014).

Training these models relies on the circular fact that word meanings are based on how they are used. However, the results frame word meaning as fixed and global, which is a barrier to their use in many social science applications, which view word meaning as dynamically shaped through use in local "language games" (Wittgenstein, 1953). Talk is an important practice through which identities are performed and learning takes place. Within the learning sciences, learning is described as increasingly central participation in the practices of a community (Lave & Wenger, 1991); as "the process of identity formation in figured worlds" (Boaler & Greeno, 2000).

> Communities of practice and other "figured worlds" provide examples of the types of people that it is possible for participants in those communities and worlds to become–street vendor, rap artist, lover of books, chess master, public speaker, basketball player, technology whiz, team

> leader. And they provide opportunities for participants to begin to enact new identities, to take on and to adapt sanctioned ways of behaving, interacting, valuing, and believing. (Hull & Greeno, 2006)

NLP methods could be profoundly useful in analyzing discourse in education and related fields if they could model contextualized linguistic meaning, and how it affects participation. This project proposes steps toward that goal with two research questions:

1. Can we use the speech of participants in a discourse community to predict their trajectories of participation?
2. Can we characterize how word meanings change between everyday language and within the discourse community?

2 Related Work

My approach to the first research question is inspired by Danescu-Niculescu-Mizil et al (2013), who explore the relationship between change in online linguistic communities and relative change in individual users' language with respect to community norms. The central finding is that Labov's adult language stability assumption largely holds in online communities: after a period of linguistic adolescence, users' linguistic practices tend to stabilize. As community norms keep changing, users gradually drift toward the periphery of the community before abandoning the community altogether. Using this phenomenon, the authors were able to predict a user's community lifespan based on the pace of her linguistic maturation within the community.

Danescu-Niculescu-Mizil et al (2013) take monthly snapshots of community practices in the form of bigram probabilities, and then compute the cross-entropy of particular posts to determine the extent to which the post's language was aligned with community norms. This approach effectively gives a distance metric between a user's language use and community norms, but it offers no insight into how a user's linguistic practices are similar to or different from the community norms. The authors switch to particular examples (such as the use of particular adjectives to describe beer flavors) to give an account of how users' practices change. I replicate their approach as a baseline, and then extend it using word vector language models which are able to capture more semantic information.

There have been numerous approaches to identifying bias and stereotyping in speech. Voight (2017) used hand-selected features (such as polite and impolite forms of address) from dialogue from police body cameras to show that officers consistently speak less respectfully to black people than white. Wu (2017) used distributions of words and topics around males and females to show sexism in an online Economics forum. These are compelling examples of how fine-grained linguistic analysis can contribute to higher-level social dynamics. In this paper, my goal is to go beyond analyzing particular features to considering deformation of the word vector space (the embedding) over time as the discourse community develops.

Characterizing changes in high-dimensional spaces is difficult, so I adopt a strategy of analyzing changes within subspaces. Arora et al (2015) argue that the linear structures often observed within word embeddings, such as the well-known example "king - man + woman = queen," should be understood as semantic relations. In this framing, projecting changes in word positions onto a subspace (for example, the one-dimensional line between the words "man" and "woman") may be interpreted as capturing the way a word's meaning changes over time with respect to the man-woman gender relation. Kusner et al (2015) compare documents using 'word mover distance', a measure of document similarity using pointwise distance between their words. In my approach, I use the same document but different embeddings, to characterize differences in the embeddings. Drawing on similar ideas, Bolukbasi et al (2016) used a subspace based on gender to identify bias in word vectors. My approach is methodically similar, but theoretically quite opposed to Bolukbasi et al (2016): while they seek to identify and remove bias from language, my goal is to describe particular forms of bias/sense-making, and how they change over time within a discourse community.

3 Methods

3.1 Data Collection

Hacker News is a discussion forum affiliated with the Bay Area startup incubator Y Combinator. The site is organized as a list of posts, each of which has an attached comments thread. I used Hacker News's public API to retrieve all comments from the site's inception in 2007 through August 2017. Table 3.1 provides summary statistics from the dataset.

Table 1. Hacker News Comments dataset statistics

Posts	12,166,758
Users	315,634
Users with more than 50 posts	31,274
Median sentences per post	2.0 (std: 3.0)
Median words per post	48.0 (std: 85.4)

After retrieving the dataset, I stripped out HTML, and then used Python's NLTK to split each comment into sentences and tokenize it, keeping only lowercase ascii and treating punctuation as separate tokens. I then used Snap (Leskovec & Sosi, 2016) to construct a user graph by creating an edge between users every time they commented in the same thread. (Each comment records its ID and parent ID, so I created comment trees and took each disconnected component as a thread.) The charts below show that the user degrees follow a power law distribution, with a roughly-linearly-declining average clustering coefficient by

degree. These charts show typical structure for a social network, confirming the validity of my graph-building procedure.

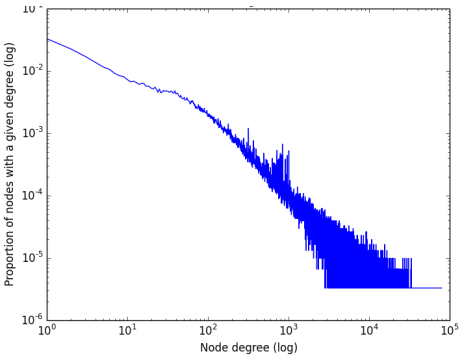

Fig. 1. User degree distribution

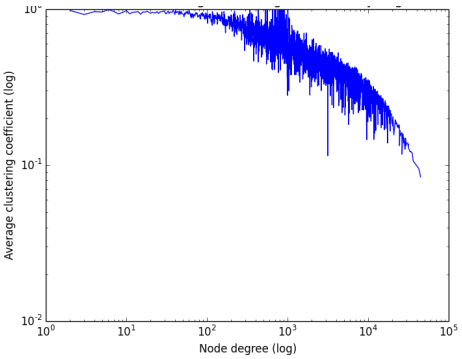

Fig. 2. Clustering coefficient distribution

3.2 Predicting User Trajectories with Linguistic Features

My first research question is addressed by Danescu-Niculescu-Mizil et al (2013), so I partially replicate their findings. Given a user's first w posts, the task is to predict whether a user will soon depart the community (writing fewer than m subsequent posts) or will remain in the community, with at least n total posts. Following one case presented by the prior paper, I use $w = 20$, $m = 30$, and $n = 200$. The distribution of users joining, leaving, staying, and bouncing (joining and leaving in the same year) follows similar trends as the datasets analyzed in the earlier paper.

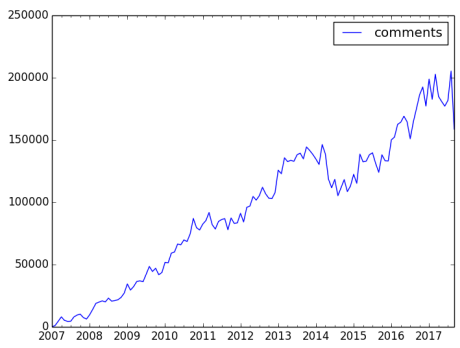

Fig. 3. Monthly comments

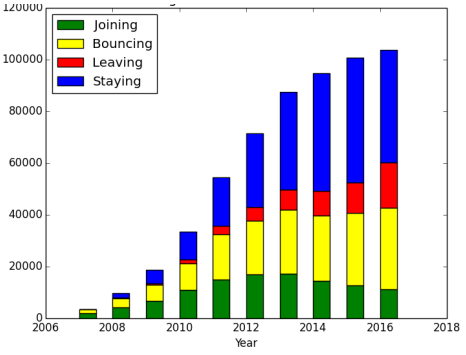

Fig. 4. Changes in user base

I used several approaches to model the community's changing word meanings over time. First, I partitioned the comments by month and used kenlm (Heathfield, 2011) to construct a bigram language model with modified Kneser-Ney (Heathfield et al, 2013) smoothing, for each month of discourse. This model calculates the perplexity of a word or phrase. For consistency with Danescu-Niculescu-Mizil et al (2013), I instead report cross-entropy.

I created a second set of monthly snapshot language models using Mikholov et al's (2013) word2vec (negative sampling skipgram) algorithm. Word2Vec represents word meanings as points in high-dimensional space (I used vectors in $\Re^{300}$), and iteratively optimizes their positions so that frequently-coincident words are closer to one another. During training, the algorithm passes over the corpus once per epoch. For each word, the model predicts the likelihood of seeing each word in a surrounding window as well as the likelihood of seeing several negative samples drawn from the vocabulary. A low loss is achieved when words appearing in the window have a high predicted likelihood and when the negative samples have a low predicted likelihood.

To prepare the corpus for word2vec training, I tokenized comment text as above, and then removed all comments having fewer than 30 tokens. Because word2vec determines word meanings using their immediate neighbors, I used Snap to organize comments by thread rather than in chronological order across the whole site. I initialized my model with an embedding pretrained on 100bn words from Google News, assuming this would be a good representation of everyday word meanings. I then iteratively trained these word vectors over each month's comments using GenSim's implementation of word2vec (Rehurek & Sojka,2010), with constant learning rate of 0.1, and 10 epochs for each month. Each month's embedding is a representation of the development of linguistic community up to that point.

Before considering how linguistic features might predict users' trajectories, I repeated the analysis in Danescu-Niculescu-Mizil et al (2013) showing how users' linguistic flexibility and distance from community language norms changes over the course of the user's lifespan in the community. Using a sample of 10000 users having at least 50 posts (with at least 30 tokens in each post), I sampled posts from each decile of the user's lifespan in the community. Linguistic flexibility was calculated as the Jaccard coefficient of a post with respect to the previous ten posts. A post's distance from the community was calculated using each snapshot language model for the month in which the post was written. For the bigram language model, I computed the cross-entropy of the post. For the word vector model, I computed the log-likelihood of the post.

Finally, I used both sets of snapshot language models to generate features to for the prediction task described above. Each feature is computed for each of a user's first w=20 posts, and then the posts are grouped into bins of size 5 and the feature values are averaged. Additionally, for each feature I compute fmax and fmin, the index of the maximum and minimum value of the feature across posts. I generated the following features:

1. Frequency, the average time between comments

2. Month, the month in which the comment was posted
3. BigramCE, the cross-entropy of the post with respect to the monthly bigram model
4. WordVectorLL, the log-likelihood of the post with respect to the monthly word vector model
5. DiffLL, the difference between the log-likelihood of the post with respect to the monthly word vector model and with respect to the initial model, representing everyday language.

The first two features are known to be highly predictive of user retention (Yang et al, 2010). The prior paper found BigramCE to give a 1-2% improvement in F_1 score for the prediction task. Using these features, I trained a logistic regression model on 60% of the users, reserving 20% for a development set and 20% for a test set. I report precision, recall, and F_1 for each combination of features.

3.3 Modeling Word Meaning Change Over Time

Having trained word vectors models for each month in the discourse community, I next use these to investigate how the space of linguistic meaning changes for Hacker News. Whereas previous studies observed changes in community norms by tracking features over time, I instead measure changes in the space of word meanings. Using two words (or two clusters of words) as anchors, I consider the line between the anchors as a relational axis, and can then project the points corresponding to other words onto this axis. For example, if the anchor clusters are ['man', 'male', 'masculine'] and ['woman', 'female', 'feminine'], projecting other words onto the relational axis shows the extent to which they are associated with masculinity or femininity. If E is the embedding matrix, E_{word} is the word vector for a given word, and the anchor word vectors are E_0 and E_1, then the magnitude of the projection of a word onto a normalized relational axis is given by:

$$|Projection| = \frac{(E_{word} - E_0) \cdot (E_1 - E_0)}{(E_1 - E_0) \cdot (E_1 - E_0)}$$

In this formula, E_0 is positioned at the origin and E_1 is a unit distance away from the origin. Values closer to 0 are more similar to E_0, while values closer to 1 are more similar to E_1. When the same anchor words are used for each monthly language model, we can compute how words' meanings change over time with respect to a relational axis, even though all the words are moving through $\Re^{300}$ space from month to month. In initial experiments, I found that occasionally all the sampled words' projections appear to move in unison; this is best explained by movement in one anchor or the other. Using clusters of words instead of individual words helps to stabilize jitter caused by the movement of individual anchor words over time.

Computing projections is computationally inexpensive when vectorized, so it is possible to select two anchor words and to compute projections for the entire

vocabulary using several language models, resulting in a list of the words whose meanings change the most over time, and as the community's word meanings diverge from everyday meanings. Initial experiments show that often the words whose meanings change most are tokens which are very-little used in everyday language, such as "trackback," "spambots," and hexadecimal codes (all in the top 10). The low frequency of such words in everyday language means they incur very little loss during training regardless of where they are placed. To observe more interesting shifts, I found that it helps to restrict searches to the 1000 or 5000 most common words, according to the Google News model's vocabulary.

4 Results

4.1 Predicting User Trajectories with Linguistic Features

I did not find the clean U-shaped curve of users' distance from the community over lifespan reported by Danescu-Niculescu-Mizil et al (2013). Instead, as shown in Figures 5 and 6, both the bigram language model and the word vectors model show a steady distancing from the community.

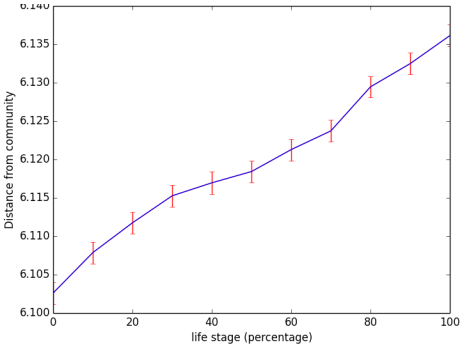

Fig. 5. Distance from community (bigram cross-entropy)

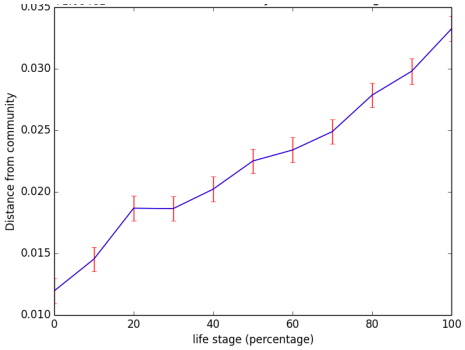

Fig. 6. Distance from community (word vector log-likelihood

Additionally, Danescu-Niculescu-Mizil et al (2013) found new users have high language flexibility, and that after linguistic adolescence, users' language flexibility declines. The same method of calculating linguistic flexibility yields a less clear trend, as shown in Figure 7.

These two findings tell a coherent story: while there is no clear trend in users' linguistic flexibility, the community does appear to have continuously-shifting norms. Because users do not appear to do much adaptation to the community when they join (Figure 7), we do not see users' distance from the community decline rapidly in the early stage of their lifecycles (Figures 5 and 6). Instead, users appear to start the process of marginalization the moment they join.

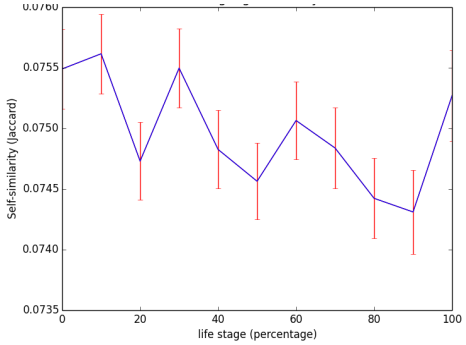

Fig. 7. Linguistic flexibility over user lifecycle

Why should this be? There are several potentially-relevant differences between the Hacker News community and the RateBeer community (analyzed in Danescu-Niculescu-Mizil et al (2013)). Hacker News is an order of magnitude larger. It may be that users offer less less "onboarding" to new members. Additionally, there may be something different in the interests bringing each community together. RateBeer is populated by people who like beer, while Hacker News broadly attracts software developers, startup enthusiasts, and data scientists. It may be that there exists a better-defined cultural niche for hackers and people interested in startups, such that they arrive at the community already part of the discourse community. Hacker news started as an online extension of a face-to-face community, so some of the language adaptation hypothesized by Danescu-Niculescu-Mizil et al (2013) may have happened in the offline space.

Table 2. User trajectory prediction task

Data set	Features	Precision	Recall	F_1	Departed	Living
RateBeer	Activity	0.737	0.193	0.305	261	465
RateBeer	Activity + BigramCE			0.374	261	465
HackerNews	Activity	0.769	0.803	0.786	1977	1602
HackerNews	Activity + BigramCE	0.770	0.805	0.787	1977	1602
HackerNews	Activity + WordVectorLL	0.768	0.804	0.785	1977	1602
HackerNews	Activity + DiffLL	0.771	0.807	0.788	1977	1602
HackerNews	Activity + WordVectorLL + DiffLL	0.769	0.805	0.787	1977	1602

Finally, Table 2 shows the results of the prediction task using various feature sets. (The first two rows are the results reported by Danescu-Niculescu-Mizil et al (2013)). Implementing the same algorithms, the Hacker News dataset achieved a much higher recall than RateBeer from the prior paper. This is at least partly due to the higher percentage of 'living' users sampled in the prior paper. More important for our research question is that none of the linguistic features de-

rived from the language models contributed significantly to the F_1 score. This is consistent with the earlier finding that Hacker News users do not appear to have an early period of linguistic adolescence. The fact that activity is much more predictive in the Hacker News dataset may also be a factor, as there may be less signal remaining for the linguistic features to pick up.

4.2 Modeling Word Meaning Change Over Time

Extending relational axes over time offered many insights into how word meanings shifted. These are illustrated here using only the relational axis of man-woman due to space limitations, but I plan to use this method in future work to analyze other relations as well. Table 3 shows the words whose position on the man-woman relational axis moved the most between January 2008 and January 2017. Several of these words' change over time is plotted in Figure 8.

Table 3. Words with greatest change on man-woman relational axis between January 2008 and January 2017

Word	Change (positive is toward woman)
ability	-0.300
knowledge	-0.281
wikipedia	-0.278
gives	-0.271
security	-0.268
theory	-0.263
education	-0.258
sense	-0.255
wiki	-0.253
available	-0.251

While we cannot draw any firm conclusions about this data here, some of the often-discussed stereotypes about women in technology are clearly visible. Margolis & Fisher (2003) and many others have shown that such stereotypes are pervasive and that they have a strong effect on womens' participation in computer science learning environments. The potential contribution of this method of modeling is to show that sexism is more than a problem of sexists; sexism is woven into the meanings of a discourse community's words. In future work, I hope to more rigorously quantify these findings and show that the orientation of users' language on relational axes of gender and other identity categories is predictive of their future participation trajectories.

5 Conclusions

This paper offers the first practical means of modeling the semantic content of participation trajectories within evolving linguistic communities. While the

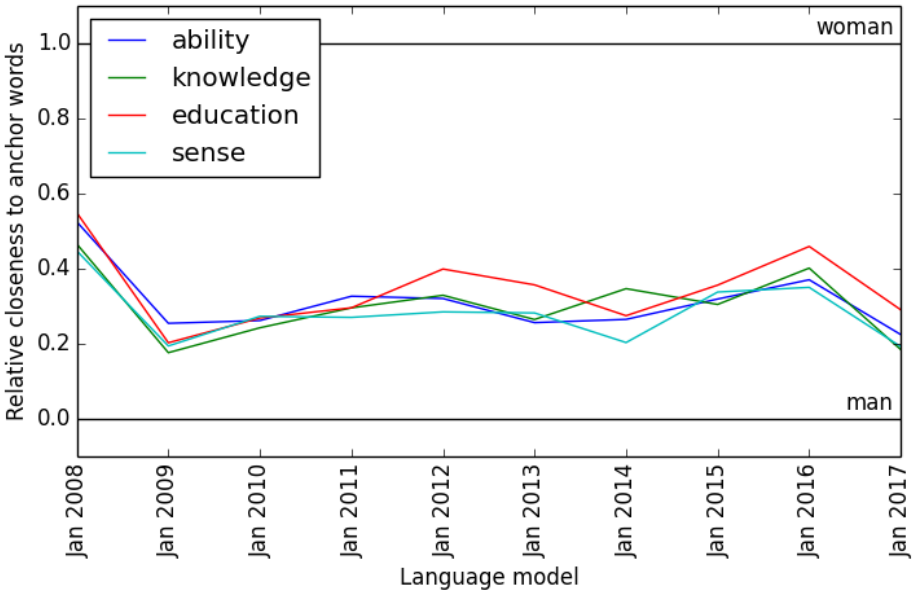

Fig. 8. Change in position on man-woman relational axis over time

word2vec-based linguistic features did not improve performance on the prediction task, the model's characteristics are very similar to the bigram model, which Danescu-Niculescu-Mizil et al (2013) found to significantly improve prediction of users' future participation trajectories. My interpretation of why Hacker News may function differently from RateBeer is consistent with research on how learners move from peripheral to more central participation (Munter & Ma, 2014). Finally, using relational axes to analyze changes in the space of word meanings is a promising avenue for future research.

In addition to methodological improvements discussed above, my future work will involve further analysis of the Hacker News community, supported by a more substantial theoretical framework. Additionally, I plan to use these methods to model in-person discourse communities.

6 References

Arora, S., Li, Y., Liang, Y., Ma, T., & Risteski, A. (2015). Rand-walk: A latent variable model approach to word embeddings. arXiv preprint arXiv:1502.03520.

Baroni, M., Dinu, G., & Kruszewski, G. (2014, June). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In ACL (1) (pp. 238-247).

Boaler, J., & Greeno, J. G. (2000). Identity, agency, and knowing in mathematics worlds. Multiple perspectives on mathematics teaching and learning, 171-200.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in Neural Information Processing Systems (pp. 4349-4357).

Dash, A. (2013). Learn to code switch before you learn to code. http://anildash.com/2013/12/learn-to-code-switch-before-you-learn-to-code.html

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013, May). No country for old members: User lifecycle and linguistic change in online communities. In Proceedings of the 22nd international conference on World Wide Web (pp. 307-318). ACM.

Heafield, K. (2011, July). KenLM: Faster and smaller language model queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation (pp. 187-197). Association for Computational Linguistics.

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013, August). Scalable Modified Kneser-Ney Language Model Estimation. In ACL (2) (pp. 690-696).

Hull, G. A., & Greeno, J. G. (2006). Identity and agency in nonschool and school worlds. Counterpoints, 249, 77-97.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015, June). From word embeddings to document distances. In International Conference on Machine Learning (pp. 957-966).

Lave, J., & Wenger, E. (1991). Situated learning: Legitimate peripheral participation. Cambridge university press.

Leskovec, J., & Sosi, R. (2016). Snap: A general-purpose network analysis and graph-mining library. ACM Transactions on Intelligent Systems and Technology (TIST), 8(1), 1.

Ma, J. Y., & Munter, C. (2014). The spatial production of learning opportunities in skateboard parks. Mind, Culture, and Activity, 21(3), 238-258.

Margolis, J., & Fisher, A. (2003). Unlocking the clubhouse: Women in computing. MIT press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.

Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., ... & Eberhardt, J. L. (2017). Language from police body camera footage shows racial disparities in officer respect. Proceedings of the National Academy of Sciences, 201702413.

Wittgenstein, L. (1953). Philosophical investigations (GEM Anscombe, trans.).

Wu, A. (2017). Gender stereotyping in academia: Evidence from Economics Job Market Rumors Forum.

Yang, J., Wei, X., Ackerman, M. S., & Adamic, L. A. (2010). Activity Lifespan: An Analysis of User Survival Patterns in Online Knowledge Sharing Communities. ICWSM, 10, 186-193.