Diachronic degradation of language models: Insights from social media

Kokil Jaidka
Computer & Information Science
University of Pennsylvania
jaidka@sas.upenn.edu

Niyati Chhaya
Big Data Experience Lab
Adobe Research
nchhaya@adobe.com

Lyle H. Ungar
Computer & Information Science
University of Pennsylvania
ungar@cis.upenn.edu

Abstract

Natural languages change over time because they evolve to the needs of their users and the socio-technological environment. This study investigates the diachronic accuracy of pre-trained language models for downstream tasks in machine learning and user profiling. It asks the question: given that the social media platform and its users remain the same, how is language changing over time? How can these differences be used to track the changes in the affect around a particular topic? To our knowledge, this is the first study to show that it is possible to measure diachronic semantic drifts *within* social media and *within* the span of a few years.

1 Introduction

Natural languages are dynamic–they are constantly evolving and adapting to the needs of their users and the environment of their use (Frermann and Lapata, 2016). The arrival of large-scale collections of historic texts and online libraries and Google Books have greatly facilitated computational investigations of language change over the span of decades. Diachronic differences measure semantic drift specifically for languages over time. For instance, the meaning of the word 'follow' has changed from a reference, then to surveillance, and finally to the act of subscribing to a social media user's feed. In a quantitative analysis, diachronic differences may explain why predictive models go 'stale'. For instance, a sentiment model trained on Victorian-era language would label 'aweful' as positive sentiment; however, in contemporary usage, 'awful' is considered a negative word (Wijaya and Yeniterzi, 2011). Thus motivated, we raise the following research questions:

- How do language models trained at one point in time, perform at predicting age and gender on language from a subsequent time?
- What is the practical benefit of measuring diachronic differences on Twitter?

To our knowledge, there is no existing work which has investigated whether, and how, language models degrade over time, i.e. why predictive models trained on an older sample of language, may fail to work on contemporary language. While previous studies have explored the change in word meanings spanning decades or hundreds of years, we address a research gap by exploring finer temporal granularity and using a more accessible language corpus. Twitter's[1] discourse is rather different from traditional English writing. So far, word embeddings trained on Twitter (Kulkarni et al., 2015; Mikolov et al., 2013) have considered it a static corpus, and have not used it to study short term changes in word connotations. It contributes with the following observations:

- Diachronic differences are greater (hence, language change is faster) for younger social media users than older social media users.
- Diachronic language differences enable the measurement of the change in social attitudes (captured by word embeddings).

In order to study this phenomenon, we define the notion of *temporal cohorts* as a set of social media users who have posted on Twitter during the same time period, e.g., in the year 2011. In this study, we evaluate the linguistic differences between temporal cohorts, e.g. 20-year-olds in 2011 vs. 20-year-olds in 2015.

[1] https://twitter.com/

2 Related work

A number of studies have built language models to predict users' age and gender (Sap et al., 2014), personality (Schwartz et al., 2013) and other traits (Jaidka et al., 2018a) with high accuracy from a sample of their social media posts. We offer the explanation that these language models may have 'degraded' due to the diachronic changes in language over the past few years, as compared to the predictions on their posts in 2011, which is closer to the time period for which their model was actually trained (see Figure 1).

The work by Frerman and Lapata (2016) quantified meaning change in terms of emerging meanings over many time periods, on a corpus collating documents spanning the years 1700-2010. Studies measuring semantic drift using word embedding models trained on Twitter corpora, such as Twitter GloVe and Word2Vec (Mikolov et al., 2013; Kulkarni et al., 2015), have considered microblog posts a static resource, reflective of modern language usage at a single point in time. Szymanski (2017) highlights the need to explore it in contemporary language e.g. social media. We illustrate that it is possible to measure diachronic semantic drifts *within* social media and *within* the span of a few years. Furthermore, we are arguing that year-related change affects different cohorts differently.

We use part-of-speech information about word embeddings to better understand semantic drift in terms of the adjective and affective words used to connotate everyday concepts. In doing so, we follow the approach outlined in previous work by Gart et al. (2017) and Hamilton et al. (2016a) to consider different kinds of contexts (e.g., adjectives, verbs and emotion words) to learn and compare distributional representations of target words.

3 Method

We first establish the diachronic validity of language-based models through predictive evaluations. We then use topic models and word embeddings as the quantitative lens through which to study the diachronic differences in the language of social media users, and linear methods to easily interpret the differences between standardized coefficients as diachronic differences in user trait prediction from language.

3.1 Predictive validity

We test the predictive performance of language models trained on a year's worth of social media posts from a subset of users who have provided their age and gender information.

- We train language models on the age- and gender-labeled primary dataset and evaluate their diachronic validity (Hamilton et al., 2016b), i.e. their predictive performance on subsequently collected language samples.
- We identify age groups which drift faster than others by reporting predictive performance on users, stratified by year of birth.

3.2 Language insights about diachronic differences

We use language for the following insights into diachronic drift:

- **Important changes**: For each temporal cohort, we identify the language features which have the most drift in terms of recalibrated coefficients, in regression models trained in 2011 vs. 2015. In doing so, we follow the approach described by Rieman et al. (2017) to compare the standardized coefficients of the age and gender models trained on the language samples from 2011 and 2015.
- **High drift concepts**: We use word embeddings to identify the semantic differences in the connotations around common concepts, for two sets of users who are a generation apart. Following the framework proposed by Garg et al. (2017), we calculate semantic drift as the changes in the relative normalized distance for the context words describing a set of target concepts.

Target concepts comprise a group of words representing a single idea (Garg et al., 2017), for instance, *positive emotion*, derived from the LIWC psycholinguistic dictionary (Pennebaker et al., 2007), and sexual orientation and gender expression (*LGBTQ issues*), based on a glossary on LGBTQ terms provided by the Human Rights Campaign [2]. We use the relative norm distance to identify the contextual words (mainly adjectives, adverbs and sentiment words) that are *most different* across the two word embedding models. Regular expression matching for part-of-speech, sentiment and emotion words was based on LIWC and the NRC emotion lexicon (Mohammad et al.,

[2] https://www.hrc.org/resources/glossary-of-terms

2013). Among a variety of distance metrics, euclidean distances provided the most interpretable results.

4 Datasets and Pre-processing

Primary data: Our primary dataset consists of the Twitter posts of adults in the United States of America who were recruited by Qualtrics (a crowdsourcing platform similar to Amazon Mechanical Turk) for an online survey, and consented to share access to their Twitter posts. This data was collected in a previous study by Preoţiuc-Pietro et al. (2017) and is available online [3]. We restrict our analysis to tweets posted between January 2011 and December 2015, by those users who indicated English as a primary language, have written at least 10 posts in their posts in each year, and have reported age and binary gender as a part of the survey. This resulted in a dataset of $N = 554$ users, who wrote a mean of 265 and a median of 156 posts per year and over 13.5 million words collectively. The mean age of the population was 33.54 years. 59% of them self-identified as male.

Decahose (Twitter 10%) dataset: For insights based on word embeddings, we used the decahose samples for the years 2011 and 2014 collected by the TrendMiner project (Preotiuc-Pietro et al., 2012), which comprises a 10% random sample of the real-time Twitter firehose using a realtime sampling algorithm. To match with our primary data, we used bounding boxes to consider only those tweets with geolocation information which were posted in the United States. In this manner, we obtained 130 and 179 million Twitter posts for 2011 and 2014 respectively.

Pre-processing: In a dataset of 554 users, the absolute vocabulary overlap may be low. By converting each users string of words into their probabilistic usage of 2000 topics, we expected to get more stable estimates than using word-based language models. We represent the language of each user as a probabilistic distribution of 2000 topics derived using Latent Dirichlet Allocation (LDA) with α set to 0.30 to favor fewer topics per document. These topics are modeled as open-ended clusters of words from actual distributions in social media over approximately 18 million Facebook updates, and are provided as an open-sourced resource in the DLATK python toolkit (Schwartz et al., 2017).

[3] https://web.sas.upenn.edu/danielpr/resources/

Predictive evaluation: We use Python's sklearn library to conduct a ten-fold cross-validation and train weighted linear regression models for age, and binary logistic regression models for gender, on the LDA-derived features for users in nine folds, and test on the users in the held out fold. We use feature selection, elastic-net regularization, and randomized PCA to avoid over-fitting. Although we tested other linguistic features such as n-grams, the best predictive performance was for models trained on the topic features.

Word embeddings: We separately train word embeddings on the language of the Twitter 10% sample from 2011, and the sample from 2014. We use Google's Tensorflow framework (Abadi et al., 2016) to optimize the prediction of co-occurrence relationships using an approximate objective known as skip-gram with negative sampling (Mikolov et al., 2013) with incremental initialization and optimizing our embeddings with a stochastic gradient descent. Embeddings were trained using the top-50000 words by their average frequency over the entire time period. A similar threshold has also been applied in previous papers (Hamilton et al., 2016a,b). We experimented with different window sizes and parameter settings, finally choosing a window size of 4, embeddings with 1000 dimensions, and the negative sample prior α set to $log(5)$ and the number of negative samples set to 500.[4]

5 Results

5.1 Predictive performance

In Figure 1, we report performance error in predicting age as the mean of $(actual - predicted)$ in order to better understand the model bias towards predicting younger or older ages.

We observe that the age- and gender- predictive models by Sap et al. (2014) **degrade in performance on language samples from more recent years**. They have a lower mean error in age prediction and higher accuracy in gender prediction on the a language sample from 2011 as compared to 2015; yet, our test sets are ostensibly drawn from the same corpus as the original training data. This shows that even models trained on large datasets show performance degradation if they are tested against newer language samples for the same set of users. We observe the same

[4] The trained word embedding models can be downloaded from http://www.wwbp.org/publications.html

Age (Mean Error)						
Test set	Sap et. al	2011	2012	2013	2014	2015
2011	2.2	**0.0**	0.2	1.1	1.6	1.8
2012	3.1	0.2	**-0.1**	1.1	1.8	2.1
2013	3.9	-0.2	-0.3	**0.4**	0.9	1.2
2014	4.4	-1.2	-1.2	-1.1	**0.0**	0.7
2015	5.0	-1.3	-1.5	-1.3	-0.4	**0.0**

Gender (Accuracy %)						
Sap et. al	2011	2012	2013	2014	2015
83	**.86**	.79	.75	.75	.75
82	.78	**.87**	.78	.77	.74
80	.78	.78	**.87**	.77	.77
77	.78	.78	.84	**.84**	.72
75	.77	.78	.77	.77	**.87**

Figure 1: Cross-year performance for predicting (a) age (reported as $MeanError = Age_{actual} - Age_{predicted}$) and (b) gender (reported as Accuracy). The columns depict the training set for regression models: language samples posted in a particular year. The rows depict the test sets. Deeper shades of blue reflect higher underestimation errors; deeper shades of red reflect higher overestimation errors. Deeper shades of green depict higher accuracy.

trends on in-sample models trained and tested on our primary data (see Figure 1). Age and gender models perform the best when tested on a sample from the same year, but age models degrade in performance over time, with older models tending to **over-predict** the age on subsequent samples. Newer models tend to **under-predict** the age on older samples of language. Taken together, these insights suggest that the rate of change in age (1 unit per year) is less than the rate of change in language use. Gender models demonstrate an approximately 7-12% drop in accuracy for subsequent or older years.

In the next step, we attempt to understand the rate of change of language for social media users of different ages, which show larger variance as compared to gender. Figure 2 provides the results for a language model trained on the language sample of 2011 and tested to predict age from a language sample from 2011 and the subsequent years. The columns depict the birth year for users in the test set. This figure provides some important insights:

1. The first row has the smallest mean errors, which is expected since it is an in-sample prediction of a 2011-trained language model on itself.
2. The largest mean errors are seen at the bottom left, where the model is consistently **underpredicting** the age of older social media users whose language usage has little change over 5 years.
3. In the bottom-right, social media users born between 1992-1997 observe the highest **overestimation** errors as they 'sound' older than they are. Despite their annual increment in age, the model still **overpredicts** their age by approximately twice as many years.

5.2 Insights about Diachronic Differences

We want to understand performance degradation in terms of the change in the associations of linguistic features associated with higher age or with one of the genders. The age range in 2011 were $[14, 41]$ years and in 2015 were $[18, 45]$ years respectively. To explore diachronic differences in topic usage across two age-matched populations, we subset the population to the subjects to the age range of $[18, 41]$ years during 2011-2015 (N=429).

Important changes: In Table 1, we compare the standardized coefficients ($p < .001$) of the predictors in models trained on the language of the year 2011 against those of 2015. We observe that a lot of the topics typically associated with older social media users in the 2011 model, [5] such as swearing, tiredness and sleep, changed their age bias. On the other hand, topics popular among younger social media users – for instance, topics mentioning employable skills and meetings, percolated upward as the early adopters of social media grew older. In the case of gender, topics related to business meetings, the government, computers, and money were no longer predictive of males, while topics associated with proms, relationships, and hairstyles were no longer predictive of females. [6]

Age	β_{2011}	β_{2015}
Email communication: *(send, email, message, contact)*	168.5	-53.7
Accommodation *(place, stay, found, move)*	162.2	-101.8
Sleep *(bed, lay, sleep, head, tired)*	59.6	-88.5
Swear *(wtf, damn, sh**, wth, wrong, pissed)*	38.1	-46.0
Tiredness *(i'm, sick, tired, feeling, hearing)*	33.6	-98.3
Hacking *(virus, called, open, steal, worm, system)*	-99.6	253.5
Software *(computer, error, photoshop, server, website)*	-87.2	80.7
Feeling *(feeling, weird, awkward, strange, dunno)*	-70.5.0	23.2
Meetings *(meeting, conference, student, council, board)*	-44.0	38.6
Skills *(management, business, learning, research)*	-26.4	158.0
Gender	$\beta_{2011}*$	$\beta_{2015}*$
Apple products *(iphone, apple, ipad, mac, download)*	4.1	(0)
Sports *(win, lose, game, betting, streak, change)*	3.2	(0)
Bills *(pay, money, paid, job, rent)*	2.8	(0)
Government *(government, freedom, country, democracy)*	2.8	(0)
Prom *(dress, prom, shopping, formal, homecoming)*	1.8	(0)
Hairstyles *(hair, blonde, dye, color, highlights)*	1.7	(0)
Relationships *(amazing, boyfriend, wonderful, absolutely)*	1.7	(0)
Negative emotions *(inside, deep, feel, heart, pain, empty)*	1.6	(0)

Table 1: The features whose coefficients had the biggest change and flipped sign when comparing the age and gender prediction models trained on 2011 language against those trained on 2015 language. (0) depicts that the feature was no longer significant in the 2015 model. *:($X10^{-}4$)

High drift concepts: We now illustrate diachronic differences in terms of the changing context around the *same* concept, in Table 2. We have

[5] See Schwartz et al. (2013)

[6] We also estimated feature importance through an ablation analysis, according to the difference made to the overall prediction ($\sum w_i x \beta_i$), which also yielded similar results.

Test set year	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
2011	**4.4**	**4.6**	**4.7**	**5.1**	**3.4**	**1.7**	**1.8**	**1.4**	**0.7**	**0.9**	**-0.1**	**-1.0**	**-0.9**	**0.0**	**-3.5**	**-2.7**	**-2.0**	**-4.3**	**-4.9**	**-2.5**	**-4.0**	**-4.3**	**-3.3**	**-2.7**	**-4.8**	**-5.6**	**-3.1**	**-2.7**
2012	6.5	5.0	6.8	6.2	4.4	3.1	3.2	3.1	1.2	2.2	-1.1	-0.8	-1.2	-2.1	-3.7	-2.9	-2.9	-5.2	-6.0	-2.9	-5.6	-2.9	-4.3	-4.7	-4.0	-7.2	-3.9	-6.1
2013	8.4	6.0	8.2	5.6	6.5	4.2	4.1	3.0	0.5	1.3	-0.4	0.2	-0.8	-1.4	-3.7	-0.8	-2.9	-4.0	-4.5	-3.7	-4.9	-3.5	-4.4	-4.8	-4.3	-8.2	-5.0	-7.0
2014	9.1	8.7	8.1	7.0	7.2	6.1	5.7	3.4	1.1	2.7	0.5	1.9	-0.1	-3.4	-2.4	-1.8	-2.3	-3.2	-4.1	-2.9	-3.8	-4.2	-4.4	-4.1	-4.7	-6.6	-4.8	-7.1
2015	10.9	8.7	10.8	6.7	8.1	6.6	6.1	4.3	3.4	2.7	2.7	2.0	0.1	-2.9	-2.9	-1.2	-3.2	-4.0	-4.5	-2.7	-4.3	-3.3	-3.8	-6.7	-5.9	-6.2	-6.2	-4.9
N	**10**	**9**	**10**	**9**	**17**	**7**	**23**	**19**	**16**	**13**	**28**	**26**	**22**	**21**	**23**	**24**	**19**	**27**	**20**	**25**	**34**	**15**	**27**	**23**	**16**	**14**	**11**	**15**

Figure 2: Cross-year performance for predicting (a) age (reported as Mean Error = $actualage - predictedage$). The rows reflect the test sets: language samples posted in the same or different year. The columns reflect users stratified according to their year of birth. Deeper shades of blue reflect higher underestimation errors; deeper shades of red reflect higher overestimation errors.

identified the words which show the largest drift between 2011-2014, in terms of their association with LGBTQ issues and positive emotion. We observe that this method of comparing the relative differences in distances, proposed by Garg et al. (2017), is able to capture social attitudes towards gender issues, as well as the emerging trends in netspeak. Specifically, in the discussions around LGBTQ issues in 2014, the words that emerge are closer to the actual experiences of the group, with words referring to 'passing' (a reference to transsexuals) and 'coping', as well as more positive emotion words ('yayy', 'harmony').

Concept	Year	Context words
LGBTQ issues	2011	strippers, conservative, pedophile, subjective, shocking
	2014	coping, passed, balance, yayy, finally, harmony
Positive emotion	2011	fagazy, bomb, totally, awesomeness, tight, fly
	2014	kickback, swag, winning, dontgiveafuck, bi*ch, thicka**

Table 2: Context words for concepts in the language of Twitter 2011 vs. 2014, selected among the words with the highest relative norm difference in distances from the concepts in the first column, between the two sets of Twitter embeddings.

6 Discussion

To summarize, our findings show that diachronic differences in language can be observed on social media and their effect differs for social media users of varying ages. In Figure 2, consider again the users of the same age at different points of time. For instance, compare the errors for 22-year old users in 2011 (born in 1989) against those for 22-year-old users in 2012 (born in 1990), and so on. The variance in error along these diagonals, are high in the right side of the table. This suggests that in every subsequent year, the language of late-teens and early-twenties is more different from the language of their contemporaries from the year before. On the other hand, compare the errors for 37-year-olds in 2011 (born in 1974) against those for 37-year-olds in 2012 (born in 1975). The errors have a low variance along the diagonals in the left of the Figure. Among social media users in their late thirties, the language of each cohort of 35-year-olds changes little over the previous year.

Next, consider the quantitative insights from Table 1. The results suggest that over time, young users from 2011 continued to use certain topics, while older users adopted newer trends. We found that if we do not use age-matched samples in this experiment, the coefficients for other topics are also flipped, but this effect is noticeably diminished with an age-matched subset. This suggests that indeed, a part of the language drift appeared because $1/5$th of the population was shifting along the temporal axis, which also associates their distinct topical preferences with an older age group.

7 Conclusion & Future Work

This study offers an empirical study of how gender and age classifiers degrade over time, a qualitative study of the features whose coefficients change the most, and concepts that drift in meaning over time. The language of social media posts can be used to study semantic drift over short periods of time, even from a dataset of 554 social media users. These methods can also find application in the study of other linguistic phenomena such as polysemy (Hamilton et al., 2016b; Szymanski, 2017). However, there is a need to disentangle which differences are due to the changing use of language from the ones due to changes in topics and trends on social media.

Language models degrade over time, but it not always feasible to retrain models with new data. In future work, we plan to explore whether domain adaptation techniques can resolve diachronic performance differences, in addition to generalizing language models to other platforms (Jaidka et al., 2018b) or scaling to measure communities (Rieman et al., 2017).

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *OSDI*. volume 16, pages 265–283.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics* 4:31–45.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2017. Word embeddings quantify 100 years of gender and ethnic stereotypes. *arXiv preprint arXiv:1711.08412* .

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*. NIH Public Access, volume 2016, page 2116.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 1489–1501.

Kokil Jaidka, Anneke Buffone, Salvatore Giorgi, Johannes Eichstaedt, Masoud Rouhizadeh, and Lyle Ungar. 2018a. Modeling and visualizing locus of control with facebook language. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Kokil Jaidka, Sharath Chandra Guntuku, Anneke Buffone, H. Andrew Schwartz, and Lyle Ungar. 2018b. Facebook vs. twitter: Differences in self-disclosure and trait prediction. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pages 625–635.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242* .

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net* .

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 729–740.

Daniel Preotiuc-Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins, and Mahesan Niranjan. 2012. Trendminer: An architecture for real time analysis of social media text .

Daniel Rieman, Kokil Jaidka, H Andrew Schwartz, and Lyle Ungar. 2017. Domain adaptation from user-level facebook models to county-level twitter predictions. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. volume 1, pages 764–773.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 1146–1151.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8(9):e73791.

H Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. Dlatk: Differential language analysis toolkit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. pages 55–60.

Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. volume 2, pages 448–453.

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*. ACM, pages 35–40.