Temporally Distant Trained Language Models: A Computational Approach to the Analysis of Language Change

Krisya Louie Masaki Makitani Federico Tondolo

1 Introduction

How do languages evolve? While it is exceedingly hard for a language speaker to concretely pinpoint at a moment in time and declare “here my language became what it is today!” it takes little to realize that language changes drastically over time, a fact painfully well known to students forced to read classical English literature. However, it remains difficult to quantify the difference between these archaic forms of language and their present-day counterparts. In English, for example, while words may fade out of vocabulary and sentence construction gradually simpler it remains relatively easy for a fluent reader to understand a text from the late 16th or early 17th century such as Shakespeare. Table 1 shows some example phrases found in Shakespeare’s works that have different present-day meanings.

Phrase	Meaning	Example
brave (adj.)	*fine, impressive*	“He has brave utensils”
issue (n.)	*child(ren), family*	“for Banquo's issue have I filed my mind”
merely (adv.)	*completely, entirely*	“Love is merely a madness”
want (v.)	*lack, need, be without*	“such a worthy leader, wanting aid”

Table 1: Shakespearean phrases and meaning.

To better study this phenomenon and attempt to empirically quantify the nature of the relative similarity between Elizabethan English and contemporary parlance we attempted to apply temporally distant corpora to train and test language models and measure the impact on their predictive performance. Specifically, we first created a variety of n-grams, ranging from bigrams to five-grams, alongside a neural network language model (NNLM) as described by Bengio et al. (2003). We then trained two identical instantiations of these models, one on a corpus consisting of a number of Shakespearean plays, and the other on the Brown University Standard Corpus of Present-Day American English. Once training was complete the models were then tested first on their own respective training datasets for reference and then on a third contemporary dataset (to eliminate any possible bias in the Brown-trained models): the Reuters-21578 ‘ApteMod’ corpus.

In an effort to arrive at the nature of the dissimilarities between Early Modern English and Present-Day English, a variety of n-gram and NNLM models were run to attempt to establish the weight that context has on model predictive accuracy (as measured by the average percentage probability assigned to the correct word). The results clearly show that the average accuracy of all the models was severely hampered by the small training datasets and the linguistic dichotomy between the corpora. However, when the probabilities were adjusted to include only word sequences which had been previously observed during training, the results showed that the Shakespeare trained models performed comparably to those trained on the Brown corpus. Moreover, while the Shakespeare bigram and trigram models performed marginally worse than the equivalent Brown-trained models, the four-gram and five-gram algorithms actually outperformed their contemporary counterparts, by a margin of up to circa 10% in the case of the five-grams.

2 Related Work

While the language evolution and language change literature is rich with sociolinguistic theories, the use of computational approaches across historical and modern corpora to capture semantic shifts and language style changes is comparatively less studied. Some quantitative methods previously used include detecting changes in word semantics using part-of-speech tags (Mihalcea and Nastase, 2012) and entropy (Tang et al., 2016). Latent semantic analysis (Salgi et al., 2011), temporal semantic indexing (Basile et al., 2014) and dynamic word embeddings (Rudolph and Blei, 2018) have also been used to analyze changes in word meaning over time. Notably, to our knowledge, there is no published study comparing Elizabethan English to present-day English using computational language models. The current study aims to extend these findings by comparing models trained on corpora more than 400 years apart to provide an interesting alternative exploratory window into how language changes across time.

3 Methods

3.1 Data

We focused on two main datasets, Shakespeare's plays as the Early Modern English corpus and the Brown corpus as the Present-Day English corpus. We also tested model performance against a present-day English Reuters corpus.

Shakespeare's plays: This dataset contains all 38 of Shakespeare's plays obtainable from the Folger Shakespeare Library Online Texts database (The Folger Shakespeare. n.d.). The character names, stage directions and metadata were removed to form a 651,777-word long corpus. Preprocessing consequently brought the length of the corpus down to 551,960 tokens.

Brown corpus: This dataset contains text from 500 sources with a variety of genres, including reviews, news, fiction etc. The original corpus we obtained from the nltk package was 1,191,192 words long (Bird et al., 2009), which is about twice the size of Shakespeare's. In order to reduce any potential biases attributed to the difference in size, we adjusted its length by selectively extracting relevant texts. Particularly, since Shakespeare's plays are inherently based on dialogues, among the various genres which constitute the Brown corpus, we limited our corpus to genres that involve dialogues: 'lore', 'belles_lettres', 'fiction', 'mystery', 'adventure', 'science_fiction', 'romance', 'humor'. (as opposed to genres like 'news' and 'editorial') The length of the extracted Brown corpus after preprocessing is 513,627 tokens, which is fairly close to that of Shakespeare's.

Reuters corpus: This dataset contains text from 10,788 Reuters news documents. We obtained the 172,0901-word long corpus from the nltk package (Bird et al., 2009).

3.2 Pre-processing

After removing all punctuations defined by the Python string library (https://docs.python.org/3/library/string.html) and lower-casing, we tokenized the words from all three corpora. In addition, we partitioned the corpora into training and test sets by the ratio of 8:2. For the Shakespeare corpus, this was simply achieved by taking the first 80% of the data. However, for the Brown and Reuters corpora, since the texts were arranged in the sequential order of genres and categories, respectively, taking the first 80% would have led to a substantial dissimilarity between the training and test set. In order to account for this, we constructed the training dataset by taking the first 80% of each genre and appending them to one another.

3.3 Model

N-grams: We employed a number of n-gram models, consisting of the traditional bigram and trigram combinations, in addition to a four-gram and five-gram to better extrapolate possible trends which might appear in the data as greater context is taken into account. These models have been amply studied and essentially consist of selecting the most likely target word by examining the pool of word sequences which contain the same context terms as those preceding the target term and then measuring which term in the same position as the target is the most common in the set.

NNLM: The NNLM we employed is effectively an approximation of the neural probabilistic language model introduced in Bengio et al. (2003). As opposed to traditional n-gram models in which word frequencies are encoded as discrete variables, the model utilizes continuous representation of words in semantic vector space. For this purpose, our model consists of 12 embedding layers through which the input words are first converted into learned distributed feature vectors. These vectors are then fed into 15 hidden layers and a softmax output layer to learn and output probability functions provided a sequence of context words.

While NNLMs have been studied to generally yield better accuracy than their non-neural counterparts, NNLMs face the issue of computational expensiveness to which we were no exception. Training a model for a single epoch took at least 2 hours, and since we wanted to run for at least 10 epochs to allow the parameters to converge as described by Bengio et al. (2003), the total time of training a model on a corpus spanned over 20 hours. Given such extensive running time and our limited computational resources, we were only able to perform comparative experiments on trigrams. For fair comparison, performance of our trigram neural network was assessed with the same methods as n-grams.

4 Results

4.1 N-grams

In the results to follow, these models were asked, rather than to predict the target word, to return the probability they would have retrospectively assigned the correct term once it was provided to them. These probabilities were then averaged out to arrive at an approximation of the average accuracy of the model (Table 2).

Corpora	2gram	3gram	4gram	5gram
Shakespeare (v. Own)	2.665	3.101	1.067	0.202
Shakespeare (v. Reuters)	1.120	0.533	0.064	0.002
Brown (v. Own)	3.807	3.903	1.459	0.280
Brown (v. Reuters)	3.428	2.856	0.932	0.168

Table 2: Average probability (%) assigned to correct target term.

We also recorded what subset of the tested word sequences had been previously observed by the models, that is, what percentage of cases these models included the correct target word among the possibilities they analyzed (Table 3).

Corpora	2gram	3gram	4gram	5gram
Shakespeare (v. Own)	62.994	17.050	2.228	0.267
Shakespeare (v. Reuters)	21.606	2.348	0.113	0.002
Brown (v. Own)	60.997	18.584	3.005	0.407
Brown (v. Reuters)	42.221	10.605	1.774	0.248

Table 3: Proportion (%) of previously observed sequences in training corpus.

Finally, in these cases where the models did indeed contain among the target term possibilities the correct term, a further accuracy measurement was taken in the method described above (though distinct from those values). These latter measurements are henceforward referred to as Adjusted Accuracy, to indicate that they exclude those cases where the correct target word had a 0% of being selected as it had never been observed during training (Table 4).

Corpora	2gram	3gram	4gram	5gram
Shakespeare (v. Own)	4.230	18.186	47.865	75.504
Shakespeare (v. Reuters)	5.601	22.726	56.666	85.667
Brown (v. Own)	6.242	21.004	48.541	68.665
Brown (v. Reuters)	8.120	26.931	52.545	67.915

Table 4: Average probability (%) of correct target term in previously observed sequences (Adjusted Accuracy).

4.2 NNLM

For fair comparison, performance of our trigram neural network was assessed with the same methods as the n-grams (Table 5).

Corpora	Accuracy	Adjusted Accuracy	Previously Observed
Shakespeare (v. Own)	7.635	15.654	12.849
Shakespeare (v. Reuters)	6.962	29.019	1.176
Brown (v. Own)	10.667	22.273	12.431
Brown (v. Reuters)	7.519	30.633	3.674

Table 5: NNLM 3gram accuracy and adjusted accuracy, with previously observed sequences.

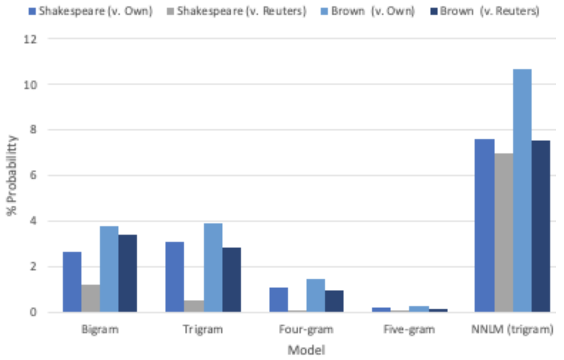

Figure 1: Average probability assigned to correct target term for n-grams and NNLM.

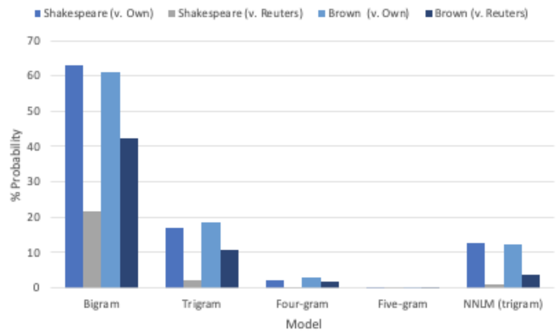

Figure 2: Percentage of previously observed sequences in training corpus for n-grams and NNLM.

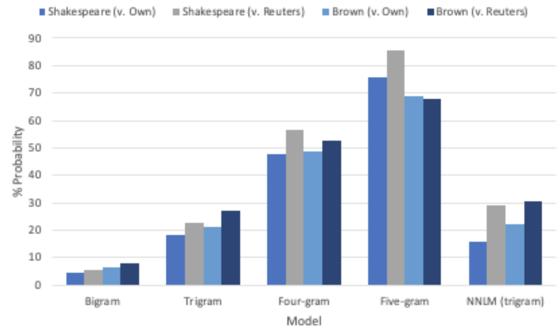

Figure 3: Average probability assigned to correct target term in previously observed sequences for n-grams and NNLM.

5 Additional Analysis

5.1 Preliminary Analysis

Frequency distribution of words can inform much about the semantic properties of the corpus. To this end, we visualized the frequency distribution of the 30 most common words in each corpus in Figures 1 and 2. Two interesting observations can be made from visually comparing the two plots. First, the top four words, “and,” “the,” “to” and “of” are shared between the two corpora. Second, the usage of “the” is significantly pronounced in the Brown corpus (approximately 1 in 12 words is “the”).

Set Difference	Top 30 Words
Shakespeare/Brown	*thou, what, him your, have, my, me, thy, so*
Brown/Shakespeare	*had, at, from, was, she, on, by, her, they*

Table 6: Set difference of the two lists of top 30 words.

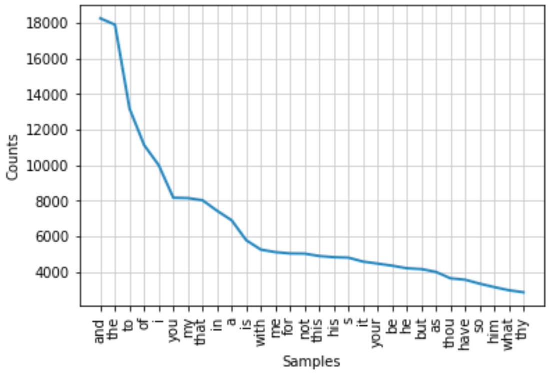

Figure 4: Frequency distribution of the most common words in the Shakespeare corpus.

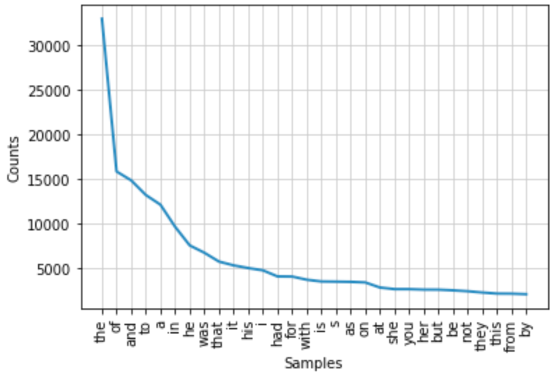

Figure 5: Frequency distribution of the most common words in the Brown corpus.

In addition, we took the set difference of the two lists of top 30 words (Table 6). As expected, archaic pronouns such as “thou” and “thy” were only found in the Shakespeare corpus. More notably, we found that third person pronouns listed from the Brown corpus encompass both genders while those from the Shakespeare corpus were exclusive to men. This might be indicative of the relative dominance of male characters in Shakespeare’s plays.

5.2 <UNK> Token Implementation

We attempted to replace words which occur less than 10 times in a given corpus with <UNK> tokens to attempt to compensate for the aforementioned exceedingly specific vocabulary of the chosen corpora. These results do not appear in this paper however because after days of computation it was determined that the only model which contained any vocabulary meeting these parameters was the Shakespearean one, rendering comparative analysis moot. Furthermore, the <UNK> substitution resulted in word sequences in the resulting Shakespeare corpus such as the

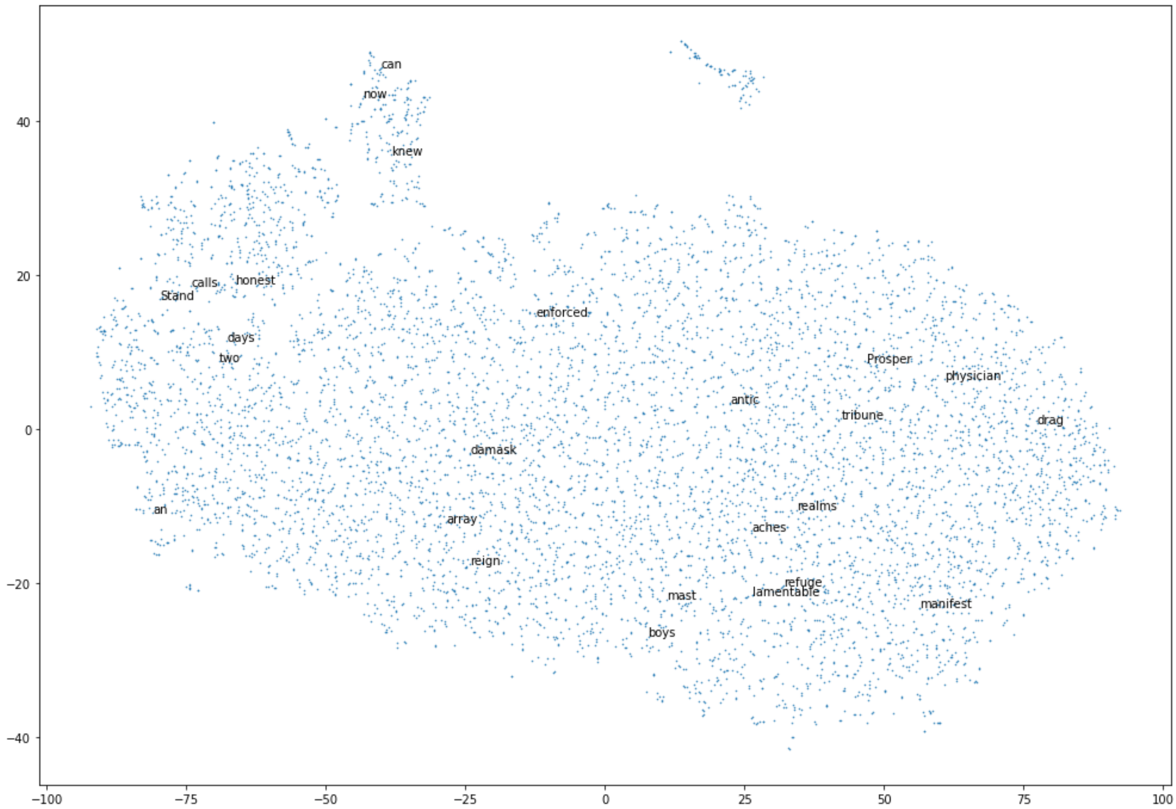

Figure 6: Two-dimensional t-SNE projection of the Shakespeare Word2Vec model.

following: “keep us all in <UNK> <UNK> <UNK> <UNK>”. Such a destructive substitution meant that even if one were to increase the minimum parameters for the operation as to result in token substitutions in the Brown and Reuters corpora the results would be unrecognizable to the point of defeating their linguistic comparison.

5.3 Word2Vec Word Embeddings

In order to better understand the underlying semantic differences between the Shakespeare corpus and the Brown corpus that led to such differences in model performance, we trained two Word2Vec models to obtain cosine similarities of the word embeddings. Qualitative inspection of the differences in highest word vector cosine similarities between the two corpora revealed interesting semantic differences between words (Table 7).

Word	Shakes Match	Shakes Similarity	Brown Match	Brown Similarity
man	*woman*	0.928	*woman*	0.873
brave	*vision*	0.975	*eager*	0.927
want	*naught*	0.960	*ask*	0.937
issue	*throne*	0.960	*method*	0.971

Table 7: Selected words with different cosine similarities in the two corpora.

The word with the highest cosine similarity with *man* in both corpora is *woman*, but the magnitude is slightly lower for the Brown corpus. Words that have an Elizabethan meaning different from present-day meaning (Figure 1) share the highest cosine similarity with words related to their respective meanings, e.g., *want* and *naught* in Elizabethan English both mean a lack of something, while *want* and *ask* are verbs used in similar contexts in present-day English.

Examination of the t-distributed stochastic neighbor embedding (t-SNE) projection of the Shakespeare Word2Vec model (Figure 6) shows a distributed representation of word vectors. There appears to be a general grouping of verbs and nouns common in both Elizabethan and present-day English (*can, now, knew; days, two, calls)*, as well as a grouping of common Elizabethan words that are less common in present-day English (*lamentable, mast, refuge, realms, aches)*.

6 Discussion

The n-gram model results paint a clear picture of models hampered by limited corpora expounded by a lack of substantial vocabulary overlap across the training and testing corpora. Though neither the Brown nor Shakespeare trained models never broke an average ~4% cumulative probability, whether when being tested against their original training corpora or against the Reuters corpus, the precipitous drop in accuracy in the latter tests of the Shakespeare-trained models as compared to the Brown-trained ones indicates a clear lack of learning (comparatively speaking). These lower predictive values are explained when one turns to look at the percent of previously observed sequences in each training corpus n-gram variation.

The percentages of previously observed word-sequences in the testing set demonstrate that although all of them decrease as the word sequences examined by the models become longer, and consequently rarer, the Shakespeare v. Reuters figures are far weaker than the Brown v. Reuters ones. This means that the vocabulary used by the Shakespeare corpus has little overlap with that of the Reuters corpus relative to the Brown corpus, making fewer of the already limited training word sequences apply to any one case, resulting in feebler predictions based on less datapoints. Moreover, as the percentage of pre-observed word sequences in the training corpus of models tested against the same corpora as they were trained on are also consistently lower in the Shakespeare models compared to the Brown models this also indicates a larger linguistic variability within the classical corpus, resulting in less deeply instilled statistic connections within the algorithms.

While the adjusted accuracy values seem to indicate that the Shakespeare-trained models are more precise than the Brown-trained models, this is illusory, as the former is based on an exponentially smaller sample size because of its aforementioned much narrower application. These models all seem to manage to catch on to moderately common English linguistic expressions, and perform well upon them, but begin to flounder almost immediately when the texts break from that small set of specific phrasing.

Further evidence for this is the results, or lack thereof, of implementing <UNK> tokens in the corpora. The fact that only the Shakespeare corpus contained words which occurred demonstrates the wild variety and uniqueness in terminology of the classic corpus, exemplifying the abyssal difference between its language and writing style and that of the two modern corpora. The exploratory word2vec model analyses also support such a distinction between linguistic expressions common to both Elizabethan and present-day English compared to expressions that are unique to a particular time.

NNLMs exhibited similar behavior to n-grams but with less drama. The baseline accuracies of NNLMs, that is, the predictive accuracies of NNLMs on the entire test set (not limited to previously observed texts), were significantly higher than those of n-grams, usually by a couple of factors. However, performance on previously observed texts were comparable to those of n-grams, which is why we did not observe as drastic an increase in performance as n-grams.

Lastly, contrary to our expectations, we observed the NNLM trained on the Shakespeare corpus performing as good as the NNLM trained on the Brown corpus in predicting contemporary English. This was particularly surprising for us because we had been under the impression from the n-gram results that Elizabethan and contemporary English were virtually incompatible entities to be ever amalgamated, even with the assistance of sophisticated computational algorithms including neural networks. Our studies suggest the underlying preservation of syntax across time, and future research may be pursued towards this direction.

7 Conclusion

The fundamental lack of linguistic overlap regarding the terminology in the examined corpora, not just between the Shakespearean and contemporary texts but also to a lesser extent between these modern corpora is not particularly surprising. For one, the moderately technical topic-oriented nature of the Brown and Reuters corpora would inevitably result in distinct vocabulary, and secondly, it is no small secret that most publications of Shakespeare include translations for outdated vocabulary and dialectic mannerisms on the opposite page of the original text. As such, these models which attempt to arrive at their results through the direct analysis of vocabulary are stumped. In future, one could attempt to train similar models on translations of Shakespeare to attempt to eliminate the Bard's peculiar vocabulary and focus on the overall grammatical sentence structure of Early Modern English.

References

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. Analysing word meaning over time by exploiting temporal random indexing. In *First Italian Conference on Computational Linguistics CLiC-it*. 38-42.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research,* 3:1137-1155.

Steven Bird, Edward Loper and Ewan Klein. 2009. *Natural Language Processing with Python.* O'Reilly Media Inc.

Folger Shakespeare Library. n.d. *Shakespeare's Plays, Sonnets and Poems* from The Folger Shakespeare. https://shakespeare.folger.edu.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 259–263.

Maja Rudolph and David Blei. 2018. Dynamic Embeddings for Language Evolution. In *WWW '18: Proceedings of the 2018 World Wide Web Conference*. 1003-1011.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2011. Tracing semantic change with latent semantic analysis. *Current Methods in Historical Semantics*. 161–183.

Xuri Tang, Weiguang Qu, and Xiaohe Chen. 2016. Semantic Change Computation: A successive approach. *World Wide Web*, 19:375–415.