Semantic Change and Emerging Tropes In a Large Corpus of New High German Poetry

Thomas Nikolaus Haider
MPI for Empirical Aesthetics, Frankfurt
IMS, University of Stuttgart
thomas.haider@ae.mpg.de

Steffen Eger
Natural Language Learning Group
Technical University of Darmstadt
eger@aiphes.tu-darmstadt.de

Abstract

Due to its semantic succinctness and novelty of expression, poetry is a great test bed for semantic change analysis. However, so far there is a scarcity of large diachronic corpora. Here, we provide a large corpus of German poetry which consists of about 75k poems with more than 11 million tokens, with poems ranging from the 16th to early 20th century. We then track semantic change in this corpus by investigating the rise of tropes ('love is magic') over time and detecting change points of meaning, which we find to occur particularly within the German Romantic period. Additionally, through self-similarity, we reconstruct literary periods and find evidence that the law of linear semantic change also applies to poetry.

1 Introduction

Following in the footsteps of traditional poetry analysis, Natural Language Understanding (NLU) research has largely explored *stylistic variation* (Kaplan and Blei, 2007; Kao and Jurafsky, 2015), (over time) (Voigt and Jurafsky, 2013), with a focus on *sound devices* (McCurdy et al., 2015; Hench, 2017) and broadly canonised form features such as *meter* (Greene et al., 2010; Agirrezabal et al., 2016; Estes and Hench, 2016) and *rhyme* (Reddy and Knight, 2011; Haider and Kuhn, 2018), as well as *enjambement* (Ruiz et al., 2017) and *noun+noun metaphor* (Kesarwani et al., 2017).

However, poetry also lends itself well to semantic change analysis, as linguistic invention (Underwood and Sellers, 2012; Herbelot, 2014) and succinctness (Roberts, 2000) are at the core of poetic production. Poetic language is generally very dense, where concepts / ideas cannot be easily paraphrased. With a distributional semantics model, Herbelot (2014) finds that the coherence of poetry significantly differs from Wikipedia and random text, allowing the conclusion that poetry is – compared to ordinary language – unusual in its word choice, but still generally regarded comprehensible language. Recently, there has been research with topic models on poetry with Latent Dirichlet Allocation. Navarro-Colorado (2018) explores the overarching topical motifs in a corpus of Spanish sonnetts, while Haider (2019) sketches the evolution of topics over time in a German poetry corpus, identifying salient topics for certain literature periods and applying these for downstream learning how to date a poem.

eher Topic Modelling

Following in this vein, we offer a method to explore poetic tropes, i.e. word pairs such as 'love (is) magic' that gain association strength (cosine similarity) over time, finding that most are gaining traction in the Romantic period. Further, we track the self-similarity of words, both with a change point analysis and by evaluating 'total self-similarity' of words over time. The former helps us to reconstruct literary periods, while the latter provides us with further evidence for the law of linearity of semantic change (Eger and Mehler, 2016) using our new method.

We do this with a model that learns diachronic word2vec embeddings jointly over all our time slots (Bamman et al., 2014), avoiding the need to compute the cosine similarity of two word vector representations on second order to align the embeddings.

Our contributions are: we (1) provide a large corpus of German poetry which consists of about 75k poems, ranging from the 16th to early 20th century with more than 11 million tokens.[1] We then track semantic change in this corpus with (2) two self-similarity experiments and finally (3) by investigating the rise of tropes (e.g. 'love is magic') over time.

[1] http://github.com/thomasnikolaushaider

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 216–222
Florence, Italy, August 2, 2019.

2 Related Work

Semantic change has been explored in various works in recent years. One focus has been on studying laws of semantic change. Xu and Kemp (2015) explore two earlier proposed laws quantitatively: the law of differentiation (near-synonyms tend to differentiate over time) and the law of parallel change (related words have analogous meaning changes), finding that the latter applies more broadly. Hamilton et al. (2016) find that frequent words have a lower chance of undergoing semantic change and more polysemous words are more likely to change semantically. Eger and Mehler (2016) find that semantic change is linear in two senses: semantic self-similarity of words tends to decrease linearly in time and word vectors at time t can be written as linear combinations of words vectors at time $t-1$, which allows to forecast meaning change. Regarding methods, Xu and Kemp (2015) work with simple distributional count vectors, while Hamilton et al. (2016) and Eger and Mehler (2016) use low-dimensional dense vector representations. Both works use different approaches to map independently induced word vectors (across time) in a common space: Hamilton et al. (2016) learn to align word vectors using a projection matrix while Eger and Mehler (2016) induce second-order embeddings by computing the similarity of words, in each time slot, to a reference vocabulary. Kutuzov et al. (2018) survey and compare models of semantic change based on diachronic word embeddings. Dubossarsky et al. (2017) caution against confounds in semantic change models.

An interesting approach besides computing independent word embeddings in each time period has been outlined by Bamman et al. (2014) who *jointly* compute embeddings across different linguistic variables: each word w has an embedding

$$\mathbf{w} = \mathbf{e}_w \mathbf{W}_{\text{main}} + \mathbf{e}_w \mathbf{W}_C,$$

where $\mathbf{W}_{\text{main}} \in \mathbb{R}^{|V| \times d}$ is a main embedding matrix and $\mathbf{W}_C \in \mathbb{R}^{|V| \times d}$ is an embedding matrix for linguistic variable C, and $\mathbf{e}_w$ is a 1-hot vector (index) of word w. In their original work, C ranges over geographic locations (US states). A joint model has several advantages: it better addresses data sparsity (for specific variables) and it directly learns to map words in a joint vector space without necessity of ex-post projection. In our work, we use this latter model for temporal embeddings in that each linguistic variable C corresponds to a time epoch t:

$$\mathbf{w}(t) = \mathbf{e}_w \mathbf{W}_{\text{main}} + \mathbf{e}_w \mathbf{W}_t$$

This dispenses the need to align independently trained embeddings for every time slot. Instead, a joint (MAIN) model is learned that is then re-weighted for every time epoch. While this is convenient, it does not necessarily mean that embeddings of a certain low-frequency word in a given time slot are stable. If there is not enough context for a given word in a certain time period t, the model just learns the MAIN embedding with little to no re-weighting, i.e., the matrix $\mathbf{W}_t$ may not be well estimated (at certain rows).

Corpus

We compile the largest corpus of poetry to date, the **German Poetry Corpus v1**, or Deutsches Lyrik Korpus version 1, **DLK** for short. See table 1 for a size overview. We know of no larger poetry collections in any language. Only the collection from the English Project Gutenberg offers a similar size, but due to a lawsuit, as of 2018 it is not available in Germany anymore.[2]

Tokens	11,849,112
Lines	1,784,613
Stanzas	280,234
Poems	74,155
Authors	269

Table 1: Corpus Size, Deutsches Lyrik Korpus v1

DLK covers the full range of the New High German language (of public domain literature), ranging from 1575 AD (Barock period) up to 1936 AD (Modern period). It is collected from three resources: (1) Textgrid[3] (TGRID), (2) The German Text Archive[4] (DTA), and (3) Antikoerperchen (ANTI-K). The latter two were first described by Haider and Kuhn (2018). All three corpora are set in TEI P5 XML.

TGRID offers around 51k poems with the label 'verse' (TGRID-V). Many of these texts have a unique timestamp. Where this is not the case, we take the average year between the author's birth and death.

[2] http://block.pglaf.org/germany.shtml
[3] textgrid.de
[4] textarchiv.de

DTA offers around 28k poems with the label 'lyrik' (DTA-L). The poetry in DTA is organized by editions (whole books), rather than by single poems. The timestamps are therefore guided by these few books, but give very accurate stamps.

ANTI-K is a collection of only 156 poems of school canon that was mined from antikoerperchen.de/lyrik. It has very accurate annotation, including literary periods, that allow us to gauge the distribution of poems according to canonic periods.

For training our model, we organize the corpus by stanzas, where every stanza represents a document. The reasoning behind this is that for poetic tropes, words are likely to stand in local context. We merge our collections and remove duplicate stanzas that match on their first line. This removes 9600 duplicates. Filtering Dutch and French material further eliminates 3200 stanzas. Since the earliest time slot 1575–1625 is too small, we merge it with the adjacent slot.

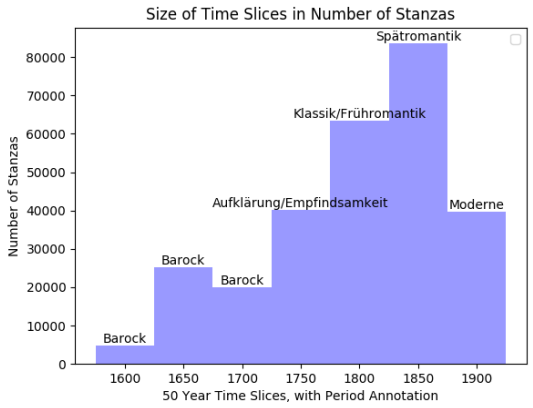

Figure 1: Distribution of stanzas in 50 year slots, 1575–1925 AD. Period labels: Barock (baroque), Aufklärung (enlightenment), Empfindsamkeit (sentimentalism), Klassik (Weimar classicism), Frühromantik (early romantic), Spätromantik (late romantic), Moderne (modernity). First slot (1600) is merged into the adjacent slot.

See figure 1 for the distribution of stanzas in 50 year time slots. The slots are labelled with approximate literature period information based on the clustered annotation in ANTI-K. We can see that the Romantic period (approx. 1750–1875) is overly heavy, while the Barock period is somewhat underrepresented.

We lemmatize based on a gold token:lemma mapping that was extracted from DTA-L in tcf format. Where this does not cover a token, we pos-tag the line with *pattern.de* to feed into germalemma.[5] We publish our corpus in json format.[6]

Experiments

Self-similarity

We investigate semantic self-similarity of words over time in two ways: (1) How does poetic diction change over successive time steps (change point detection), and (2) how does contextual word meaning change in total over the whole time frame with respect to the word's frequency (laws of conformity and linearity)? We use a model with a 25+50 sliding window, where time steps increase by 25 years, with a window size of 50 years. This doubles the data and allows a more fine grained analysis.

Pairwise Self-Similarity

We compute how the contextual use of words changes over successive time steps. We do this by determining the self-similarity of a word w over time by calculating the cosine similarity of the embedding vectors $\mathbf{w}(t)$ for w at time periods $t = t_i$ and $t = t_{i+1}$ as in equation (1):

$$\text{cossim}(\mathbf{w}(t_i), \mathbf{w}(t_{i+1})) \tag{1}$$

where $\text{cossim}(\mathbf{a}, \mathbf{b})$ is defined as $\mathbf{a}^\intercal\mathbf{b}$ for two normalized vectors $\mathbf{a}$ and $\mathbf{b}$.

Thus, we can aggregate the self-similarity for the most frequent words at every time step and plot the change for all these words combined. See figure 2 for a boxplot of this pairwise self-similarity for the 3000 most frequent words.

Results

Our interpretation is that rising similarity signifies a homogenization of overall word use (diction), while a falling similarity signifies semantic diversification. In particular, we see a steady falling trajectory in the period between 1600 and 1675, with a dip at 1700. This period is generally regarded as the 'Barock' period. Then, word use slowly homogenizes, until we see a sharp dip around 1750, the onset of the Romantic period. Then it homogenizes during the Romantic period, until a dip at 1850, the end of the Romantic period, and then a homogenization into the the onset of modernity.

[5] https://github.com/WZBSocialScienceCenter/germalemma

[6] http://github.com/thomasnikolaushaider

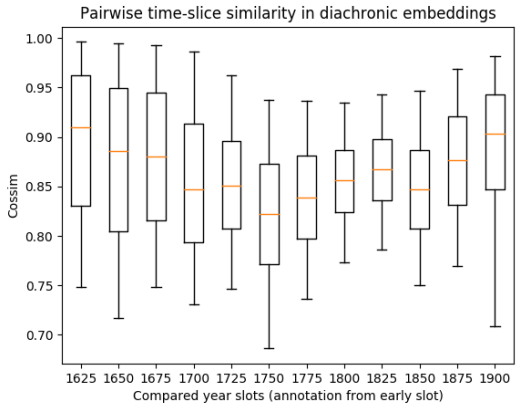

Figure 2: **Pairwise Self-Similarity. Top-3000 most frequent words. Cossine similarities of word w with itself in adjacent time slots** $cossim(w(t_i), w(t_{i+1}))$

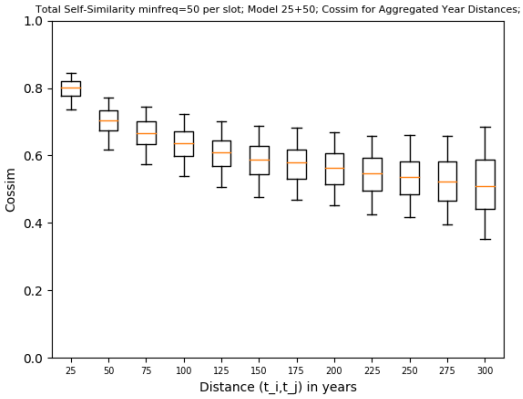

Figure 3: **Total Self-Similarity of words that occur at least 50 times in every time slot. Cossine similarities aggregated by the distance of compared time slots** (t_i, t_j) **averaged for every time slot given a word. Removed stopwords. Whiskers: [5,95] percentiles.**

Total Self-Similarity

We determine change of word meaning across any possible time distances as a probing for the linearity of semantic change in our corpus.

For this, we calculate the semantic self-similarity of a word across all time periods t_i and t_j with $t_i < t_j$. We then aggregate all pairwise distances in years

$$\text{dist}(t_i, t_j) = |t_i - t_j|$$

for all words w.[7] To obtain robust estimates of embeddings, we only allow words that occur at least 50 times in every time slot and remove stopwords, leaving us with 472 words.

The x-axis in Figure 3 gives the distances $\text{dist}(t_i, t_j)$ while the y-axis shows the distribution of cossims over all words w within each distance.

We find that there is approximately a linear relation between the distance of timeslots for an average word, where close slots are more similar, and far apart slots are increasingly dissimilar. However, the variance also increases with distance.

Additionally, we equally divide our words into a low-frequency and a high-frequency band. We find that the low-frequency band shows a generally higher self-similarity than the high-frequency band over all distances. This would mean that, overall, high frequency words tend to be more semantically diverse over time, i.e. stand in more diverse contexts. In contrast, low-frequency words stand in fewer contexts, therefore undergo less change. However, this could also come from the tendency of the model to revert to MAIN.

[7] For all 25,50,...,300 year distances, cossims per word in these distances are averaged, so we are left with one value per distance and word.

Emerging Tropes: Collocations & Metaphors

Method

To detect emerging tropes, we calculate the cosine similarity of word pairs over time. For the sake of visualization we use a 50+50 model with 6 time slots. We then perform Principal Component Analysis (PCA) over the resulting trajectories (Eger, 2010). The resulting principal components show that similar trajectories are co-variant. Component 1 aggregates stable high/low trajectories, while component 2 aggregates rising/falling trajectories. We illustrate our finding with the tropes for the concept 'love' ('Liebe' in German) and determine the most salient word pairs over the whole dataset. 'Love' is a very frequent word in poetry. Nevertheless, this approach works equally well for any word, except very low frequency words that show idiosyncratic behavior as they are not well distributed.

Results

We calculate the distance of 'love' against every other word w, where w has to occur at least 30 times in the corpus, and it needs to be represented in every time slot at least twice. We allow one slot to be empty.

The first 4 components of PCA explain >.95 variance, where component 1 explains 73%, component 2 13%, and component 3 5%. We retrieve the top-25 word pairs at every component extreme.

rising traj.	falling traj.	high traj.	low traj.
frische	aufrechen	liebe	brummen
veilchen	alsbald	freundschaft	krähen
niedersinken	billigkeit	lust	rasseln
duftig	erzeigen	treue	rum
jenseits	unterstehen	trieb	bock
zauber	betragen	seligkeit	dum
entgleiten	stracks	hoffnung	prasseln
künden	zuerkennen	glaube	trommel
hoffend	hierin	keusch	säbel
efeu	schmeissen	treu	traben
enthüllen	anlaß	erkalten	bellen
erfüllung	jederzeit	wahr	block
heimat	muhen	immerdar	bügel
trübe	schimpfen	regung	gaul
gloria	stecken	gegenliebe	grasen

Table 2: Top 15 words per dimension for 'love' tropes from PCA extremes, plotted in figures 4, 5, 6 and 7.

We find that component 1 orders trajectories based on high/low semantic similarity, while component 2 orders based on rising/falling trajectories. See figures 4 (high trajectory), 5 (rising trajectory), 6 (low trajectory) and 7 (falling trajectory). See table 2 for the respective word pairs (collocations) with 'love' as they are plotted.

Stable High Trajectories Trajectories in figure 4 (table 2 column 3) have a consistently high cosine, meaning that these collocations have remained unchanged since the Baroque period: 'love is fidelity',[8] 'love is friendship',[9] or 'love is lust'. These are conventional near-synonyms, just as ('apple', 'tree')[10] or idioms ('apples', 'pears').[11] A k-nearest neighbor (KNN) analysis retrieves these collocations. Performing this analysis for multiple words, we find that the idiom ('apple', 'pear') is a special case, as it strongly loads into both rising and stable high PCA dimensions (both top 20).

Rising Trajectories Figure 5 (table 2 column 1) shows rising collocations that emerge during the Romantic period, i.e. 'fresh love',[12] 'love is magic / enchantment'[13] and 'love is violets'.[14] A metaphorical (trope) interpretation is most likely here.

Falling Trajectories As illustrated in figure 7 (column 2), these collocations fall into obscurity.

[8]('Treue', 'Liebe')
[9]('Freundschaft', 'Liebe')
[10]('Apfel', 'Baum')
[11]('Äpfel', 'Birnen'), 'compare apples and oranges'.
[12]('Frische', 'Liebe')
[13]('Zauber', 'Liebe')
[14]('Veilchen', 'Liebe')

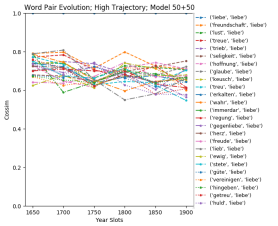

Figure 4: Love: High

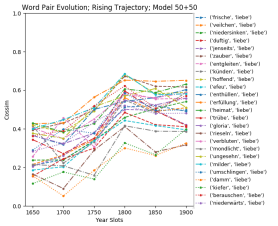

Figure 5: Love: Rising

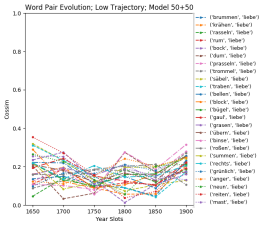

Figure 6: Love: Low

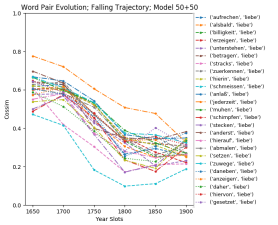

Figure 7: Love: Falling

We find 'cheap love'[15] or things like 'raking'[16] or 'manners / accounting'.[17]

Stable Low Trajectories The lines in figure 6 (column 4) signify word pairs that are always far apart. We find things that make noise, like drums.[18] The 'drums of love' seems to be an oxymoron.

Conclusion

We constructed the largest poetry corpus to date and investigated distributional semantic change with different methods. With self-similarity, we can reconstruct literature period transitions and find that the the law of linear semantic change also applies to poetry. However, for confident analysis of other laws more data and a more robust model is still called for. Finally, we extract emerging and vanishing poetic tropes based on the co-variance of time trajectories of word pairs. This method is applicable more broadly to cluster similar trajectories for any given word pairs. We found trajectories of word similarities that are beyond simple nearest-neighbor analysis, and illustrated findings for reasonable tropes with 'love'. While large, our dataset is still somewhat sparse in the distribution of words over all time slots, partially because many word forms simply emerge / vanish at a certain point ('excitement' is not in Baroque).

[15]billigkeit
[16]aufrechen
[17]betragen
[18]trommel

References

Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. 2016. Machine learning for metrical analysis of English poetry. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 772–781, Osaka, Japan. The COLING 2016 Organizing Committee.

David Bamman, Chris Dyer, and Noah A Smith. 2014. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 828–834.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1147–1156, Copenhagen, Denmark.

Steffen Eger. 2010. Investigating lexical competition - an empirical case study of the german spelling reform of 1996/2004/2006. *JLCL*, 25(1):3–21.

Steffen Eger and Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 52–58, Berlin, Germany. Association for Computational Linguistics.

Alex Estes and Christopher Hench. 2016. Supervised machine learning for hybrid meter. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 1–8.

Erica Greene, Tugba Bodrumlu, and Kevin Knight. 2010. Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 524–533.

Thomas Haider and Jonas Kuhn. 2018. Supervised rhyme detection with siamese recurrent networks. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. At COLING 2018, Santa Fe, New Mexico*, pages 81–86.

Thomas Nikolaus Haider. 2019. Diachronic topics in new high german poetry. In *In Proceedings of the International Digital Humanities Conference DH2019, Utrecht*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. 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)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Christopher Hench. 2017. Phonological soundscapes in medieval poetry. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 46–56.

Aurélie Herbelot. 2014. The semantics of poetry: a distributional reading. *Digital Scholarship in the Humanities*, 30(4):516–531.

Justine T Kao and Dan Jurafsky. 2015. A computational analysis of poetic style. *LiLT (Linguistic Issues in Language Technology)*, 12.

David M Kaplan and David M Blei. 2007. A computational approach to style in american poetry. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 553–558. IEEE.

Vaibhav Kesarwani, Diana Inkpen, Stan Szpakowicz, and Chris Tanasescu. 2017. Metaphor detection in a poetry corpus. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–9.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nina McCurdy, Julie Lein, Katharine Coles, and Miriah Meyer. 2015. Poemage: Visualizing the sonic topology of a poem. *IEEE transactions on visualization and computer graphics*, 22(1):439–448.

Borja Navarro-Colorado. 2018. On poetic topic modeling: extracting themes and motifs from a corpus of spanish poetry. *Frontiers in Digital Humanities*, 5:15.

Sravana Reddy and Kevin Knight. 2011. Unsupervised discovery of rhyme schemes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 77–82.

Phil Roberts. 2000. *How Poetry Works*. Penguin UK.

Pablo Ruiz, Clara Martínez Cantón, Thierry Poibeau, and Elena González-Blanco. 2017. Enjambment detection in a large diachronic corpus of spanish sonnets. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 27–32.

Ted Underwood and Jordan Sellers. 2012. The emergence of literary diction. *The Journal of Digital Humanities, 1(2)*, pages http://journalofdigitalhumanities.org/1–2/theemergence–of–literary–diction–by–ted–underwoodand–jordan–sellers/.

Rob Voigt and Dan Jurafsky. 2013. Tradition and modernity in 20th century chinese poetry. In *Proceedings of the Workshop on Computational Linguistics for Literature*, pages 17–22.

Yang Xu and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *CogSci*. cognitivesciencesociety.org.