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

TECHNISCHE UNIVERSITÄT DARMSTADT

University of Stuttgart Germany

Max Planck Institute for Empirical Aesthetics

Thomas Nikolaus Haider[1,2], Steffen Eger[3]

[1] Max Planck Institute for Empirical Aesthetics, Frankfurt | Department Language and Literature
[2] University of Stuttgart | Institut für Maschinelle Sprachverarbeitung (IMS)
[3] Technical University of Darmstadt | Natural Language Learning Group

Introduction

Poetry lends itself well to semantic change analysis, as **novelty of expression** (Underwood, 2012; Herbelot, 2014) and **succinctness** (Roberts, 2000) are at the core of poetic production.

Self-Similarity can track **literary periods** and show **linearity of semantic change**.

Previous work (Haider, 2019) showed **salient topics of literary periods.** Then how are topics correlated to form metaphors / tropes? We compute **cosine similarity of word vectors over time to see the rise of tropes** (`love is magic'). We find change mainly within the German Romantic period, where tropes are picked up and permeate into Modernity.

We compile a large corpus of German poetry with **75k poems** and **11 million tokens**, ranging from **1575 – 1936 A.D.**, from the Baroque period into Modernity.

Model

Jointly compute word2vec embeddings for **MAIN** corpus and add **each time period** (Bamman et al., 2014)

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

No need to align independently trained embeddings for every time slot. Instead, a joint (MAIN) model is learned that is then reweighted for every time epoch (originally regional variables: US states). This is convenient, but it does not necessarily mean that embeddings of a certain low-frequency word in a given time slot are stable. Also, this concatenation does not allow to look at certain semantic laws (conformity, innovation), because it always reverts to MAIN.

Experiments

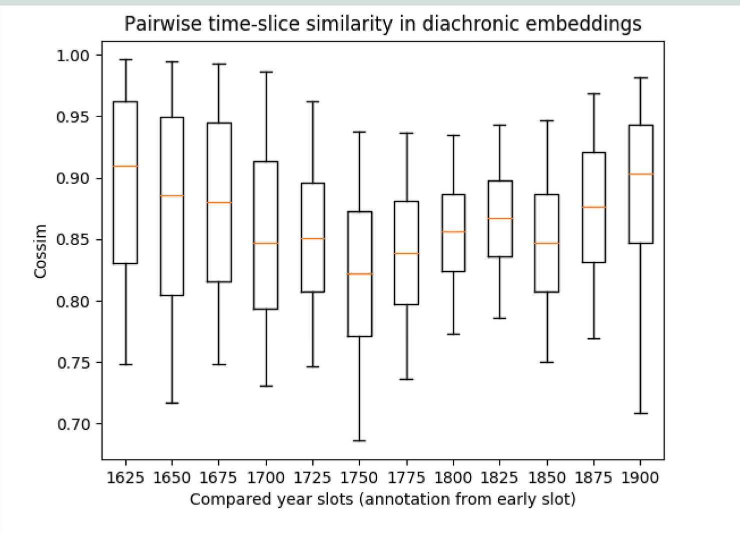


Figure 2: **Pairwise Self-Similarity. Top-3000 most frequent words. Cossine similarities of word w with itself in adjacent time slots** $cossim(w(t_i), w(t_{i+1}))$

Pairwise similarity of a given word over **successive time steps** (13 slots 25+50) tracks **literature periods**. Upward traj. show **homogenization**, downward traj. **diversification** of vocabulary. Dips show onsets of lit. period (1750: Onset of Romantic period).

Self-Similarity

Total similarity of a given word over **all possible time distances** shows an approx. **linear relation** b/w change and time.

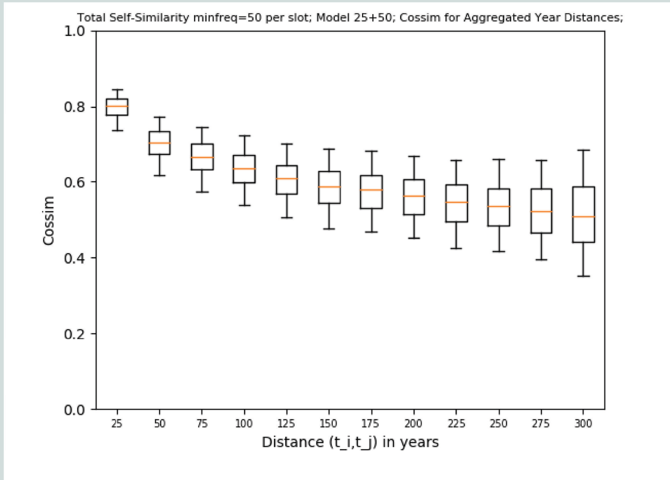
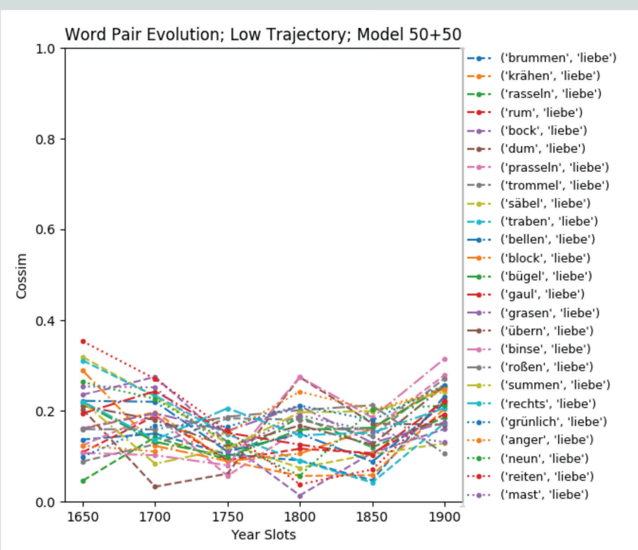


Figure 3: **Total Self-Similarity of words that occur at least 50 times in every time slot. Cossine similarities aggregated by the distance of compared time slots** (t_i, t_j) **averaged for every time slot given a word. Removed stopwords. Whiskers: [5,95] percentiles.**

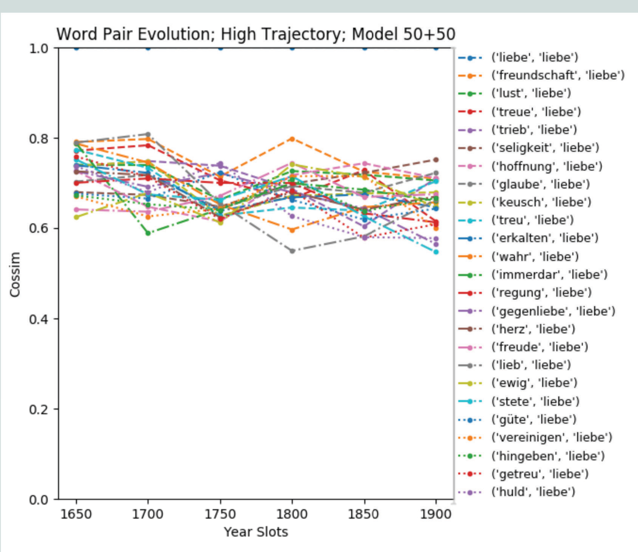


Stable Low trajectories: Always **far apart**. Things that make noise e.g. **'drums of love'**.

To **discover emerging tropes**, we calculate **cosine similarity of 'love' against every other word** over time.

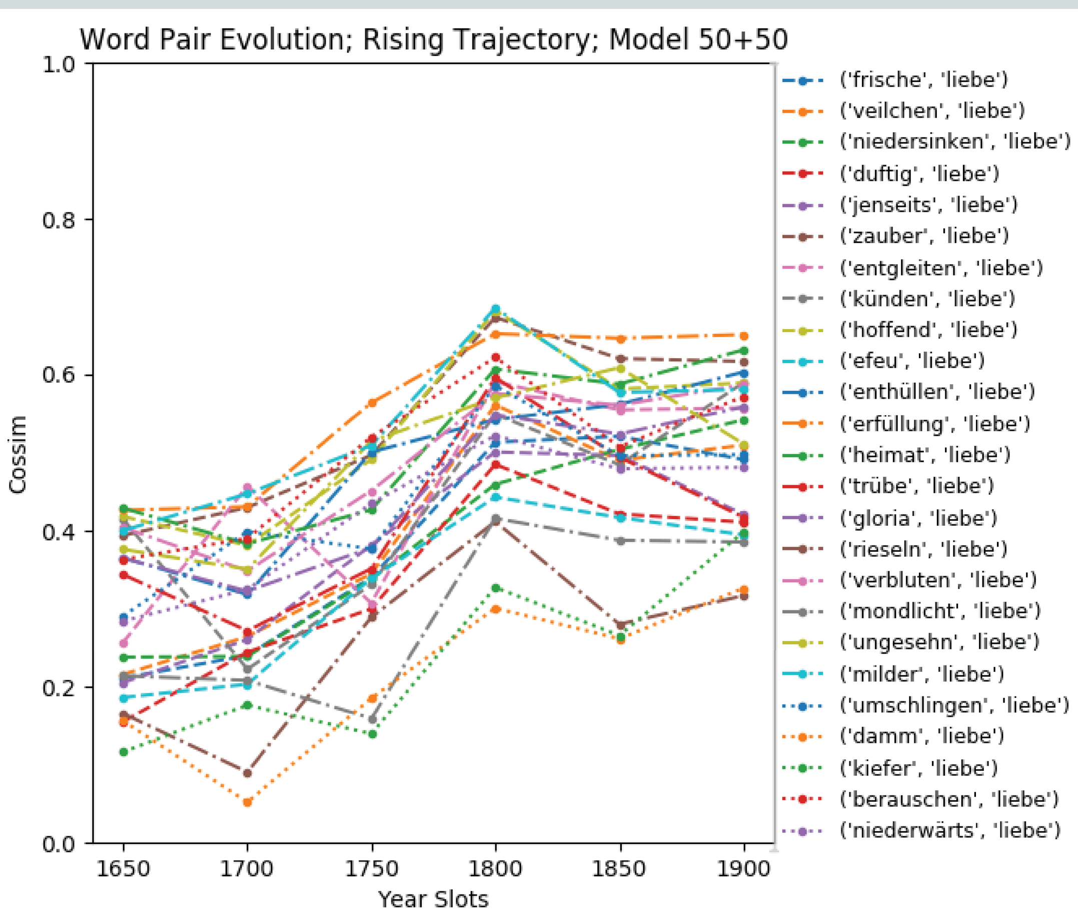
Principal Component Analysis (PCA) over the resulting trajectories show: similar trajectories are co-variant. Component 1 (73%) aggregates **stable high/low trajectories,** while component 2 (13%) aggregates **rising/falling trajectories**. Plotted are top 25 word pairs per dimension (two per component).

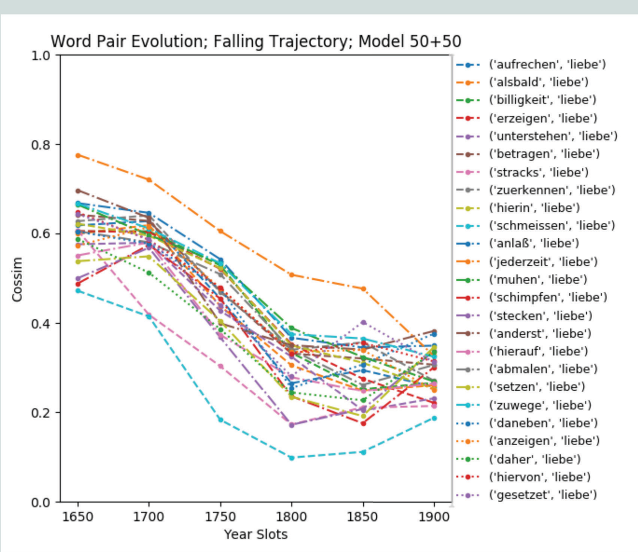
Emerging Tropes



Stable High Trajectories have a **consistently high cossim**. These collocations have remained unchanged since the Baroque period: **'love is fidelity'**, **'love is friendship'**, or **'love is lust'**. These are **conventional near-synonyms**. A k-nearestneighbor (KNN) analysis retrieves these collocations.

Rising trajectories emerge during the Romantic period, i.e. **'fresh love'**, **'love is magic / enchantment'** and **'love is violets'**. A **metaphorical (trope) interpretation** is most likely here.





Falling trajectories fall into **obscurity**: We find **'cheap love'**, **'raking' or 'manners / accounting'**.

Corpus

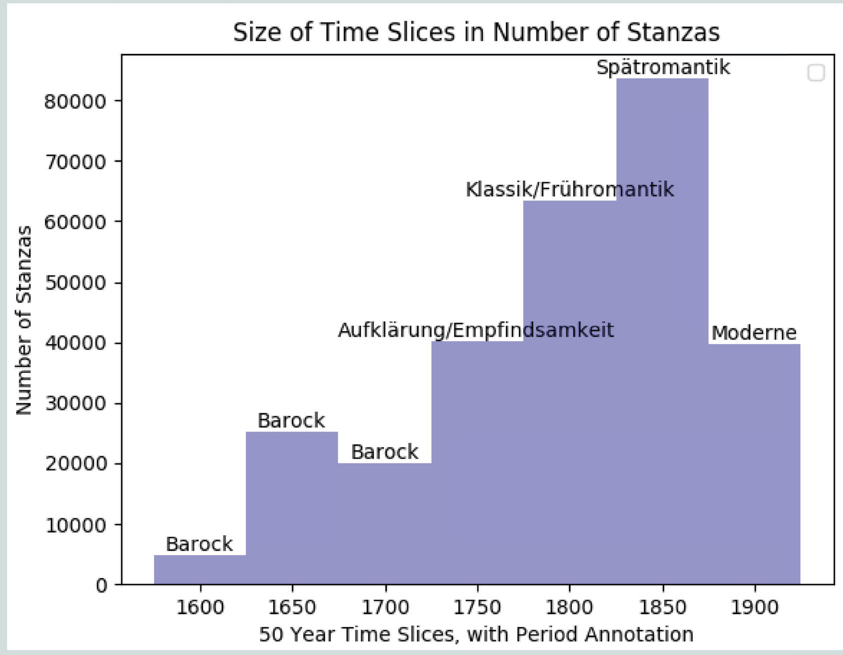


Figure 1: **Distribution of stanzas in 50 year slots, 1575–1925 AD. Period labels: Barock (baroque), Aufklärung (enlightenment), Empfindsamkeit (sentimentalism), Klassik (Weimar classicism), Frühromantik (early romantic), Spätromantik (late romantic), Moderne (modernity).**

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

Table 1: Corpus Size, Deutsches Lyrik Korpus v1

- Largest dataset of New High German poetry to date (consistency from Baroque to Modernity)
- 75k poems (texts), 11M words, 1575 – 1936 A.D.
- Time stamps mostly accurate. If not: average year b/w author birth \& death
- Documents are stanzas (for poetic tropes, words are likely to stand in local context)
- Includes most of the literary canon But far from complete: Half of Rilke's work is missing
- Includes other languages than New High German (Middle German, Dutch, French, Latin) that need to be filtered
- Lemmatization based on a gold token: lemma mapping from DTA + germalemma
- Compiled from (1) Textgrid (51k poems), (2) The German Text Archive DTA (28k poems), and (3) Antikoerperchen (ANTI-K, 150 poems, school canon).

Contact Thomas Haider
thomas.haider@ae.mpg.de |
github.com/thomasnikolaushaider

Acknowledgement
Thanks to Sarah Brendecke for helping with the poster.
Thanks to the MPIEA for funding this work.

References

+ Thomas N. Haider. 2019. Diachronic Topics in New High German Poetry, DH2019 Utrecht
+ David Bamman, Chris Dyer, and Noah A Smith. 2014. Distributed representations of geographically situated language. ACL
+ Steffen Eger and Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. ACL.
+ Aurélie Herbelot. 2014. The semantics of poetry: a distributional reading. Digital Scholarship in the Humanities (DSH).
+ Ted Underwood and Jordan Sellers. 2012. The emergence of literary diction. The Journal of Digital Humanities
+ Phil Roberts. 2000. How Poetry Works. Penguin UK.