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

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.

Experiments

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 approximate linear relation b/w change and time.

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

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.

Corpus

Jointly compute word2vec embeddings for MAIN corpus and add each time period (Bamman et al., 2014) $w(t) = e_p W_{main} + e_t W_t$

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.

• 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: lemmata from DTA + germanlemma

References

+ Aurélie Herbelot. 2014. The semantics of poetry: a distributional reading. Digital Scholarship in the Humanities (DSH)

Acknowledgement

Thanks to Sarah Brandenke for helping with the poster. Thanks to the MPIfG for funding this work.