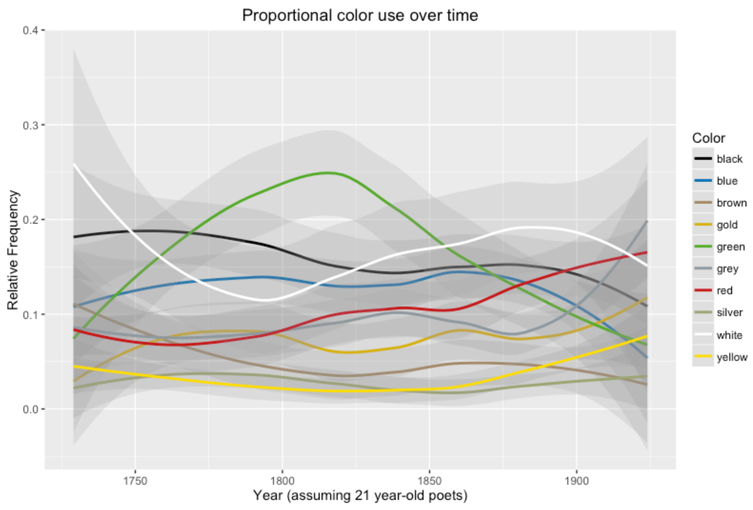
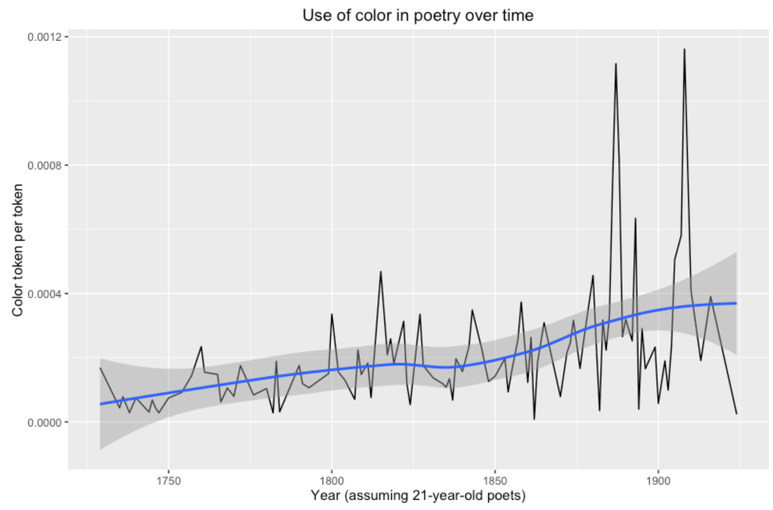
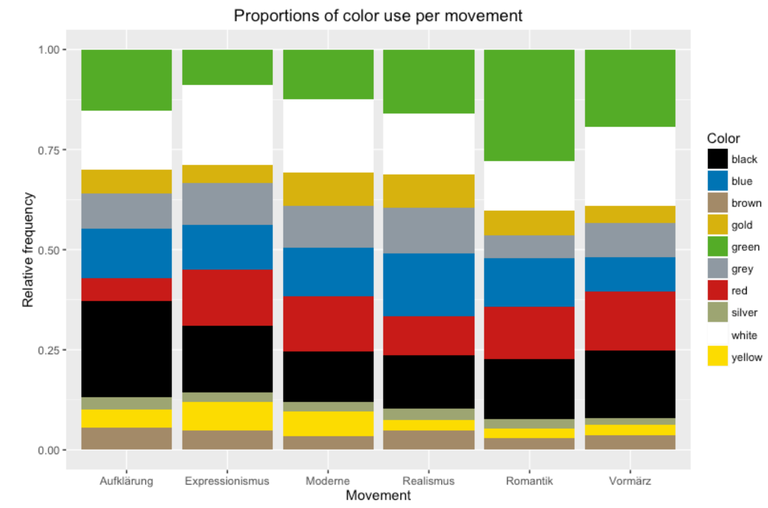
Color Distribution in German Poetry

Laura Bostan, Jonathan Oberländer
Saarland University, Trento University, Charles University







Introduction

Inspired by reading a collection of German poems titled "Blaue Gedichte" we investigated the usefulness of color frequencies in identifying the period/movement in which a poetic work was created.
Color is an obvious feature for image classification, yet, to the best of our knowledge, a feature that was not considered for text classification.

Data

We crawled all poems available on Project Gutenberg-DE using a combination of the Python libraries Requests and Beautiful Soup. The number of different authors was 299, most of which are German, Austrian or Swiss. The rest are of different nationalities (with their works translated to German).

authors	299
authors (German)	266
pages crawled	27523
words (tokens in corpus)	10481697
occurrences of colors (total)	31388

After the crawling was completed, we counted for each author how often they used any individual color name. This was done automatically using a set of regular expressions, which brings a set of problems related to ambiguity, most prominently regarding the color white, which in German is a homonym of several forms of the verb "to know". To solve this we POS-tagged the corpus and took only tokens that were tagged as nouns, adjectives or adverbs.

black	gray	white	red	green	blue	yellow	brown	silver	gold	violet	orange
3797	2185	3974	3066	3859	2967	999	971	575	1733	51	3

The list of authors was annotated with:

- frequency of individual color word usage
- literary movement
- years of birth and death
- total number of tokens in the corpus
- polarity information (from Chen and Skiena, 2014)
- frequency of color words overall
- GND identifier

Berlin Kay

Based on the data collected by Pratt (1898) (17 poets, color counts done manually), McManus (1983) showed that the distribution of color words in English and Chinese poetry correlates with the order of evolution of color words across languages (Berlin and Kay, 1963). Later, McManus (1997) continued the experiment computationally on a larger English dataset. We repeated this study for German poetry and we found no relationship between the order of colors as derived by Berlin and Kay and the frequency of usage of color terms in German poetry.

Classifying

We investigated whether color usage was enough to predict the main literary movement of an author. For this, we chose only the six most frequent movements and discarded all authors that couldn't be assigned to any of them, as well as any translated authors. This left us with 146 authors. We split the data randomly in training and testing data and trained two standard classification models (Naive Bayes and Random Forest) on the task, comparing it with a "stratified" dummy classifier as a baseline. Since our dataset was so small, we repeated the test 200 times and report average accuracy values here.

Classifier	Accuracy
Dummy classifier	16.7
Naive Bayes	26.0
Random Forest	27.0

We also tested whether adding color counts as features to a classification system based on keyword extraction (Mihalcea 2004) would improve the accuracy over the system without the colors as features, but that doesn't seem to be the case.

Color by sentiment

We also investigated how the polarity values related to colors. Since higher frequencies of colors lead to them being in the context of negative and positive polarity higher, we devise a simple positivity ratio of any color c:

$$p(c) = \frac{\sum_{a \in \text{Authors}} \text{freq}(a, c) * \text{freq}(a, +)}{\sum_{a \in \text{Authors}} \text{freq}(a, c) * \text{freq}(a, -)}$$

This leads to a color ranking that is different from their frequencies with green, black, and silver being most used by authors with higher positive polarity, and orange, yellow, and violet linked stronger to negative polarity. Overall, all texts were annotated much more frequently as having more negative polarity tokens.

Conclusion

There are two major drawbacks to our work: The amount of data, and the annotation quality regarding the literary movements. We don't have enough data to say for sure if there is an actual correlation between time and the usage of colors.
The created dataset is uploaded on Github together with the scripts we created to explore the data: `https://github.com/gastrovec/poetry-colors`

References

Berlin, Brent, and Paul Kay. Basic color terms: Their universality and evolution. Univ of California Press, 1991.
Chen, Yanqing, and Steven Skiena. "Building Sentiment Lexicons for All Major Languages." ACL (2). 2014.
McManus, Ian C. "Basic colour terms in literature." Language and Speech 26.3 (1983): 247-252.
McManus, Ian C. "Note: Half-a-million basic colour words: Berlin and Kay and the usage of colour words in literature and science." Perception 26.3 (1997): 367-370.
Pratt, Alice Edwards. The use of color in the verse of the English romantic poets. Haskell House, 1898.