DH 2023, Graz

Modeling Prototypicality and Uncertainty in Genre Concepts

Julian Schröter
Digital Literary Studies, Ludwig-Maximilians-Universität München

Introduction

- The overall project focuses on the difference between the ›Novelle‹ and the ›Erzählung‹ as two novella genres in German literature.
- Some scholars have claimed that both genres are clearly separated. However, another group of scholars maintains there is no difference (Meyer 1987).
- As the PCA in Fig. 1 shows, this project provides initial evidence for the second and skeptical position:

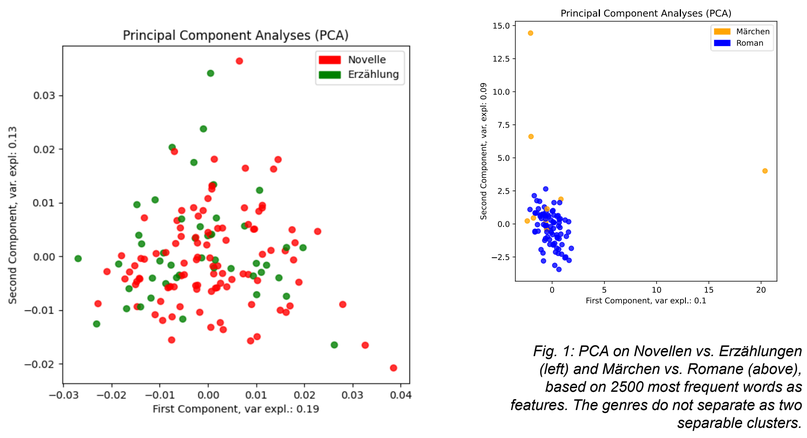

Fig. 1: PCA on Novellen vs. Erzählungen (left) and Märchen vs. Romane (above), based on 2500 most frequent words as features. The genres do not separate as two separable clusters.

Aim

- A model of the prototypicality of genre concepts (Rosch 1973; 1975; 1978; Taylor 2007; Hempfer 2010).
- Estimating the degree of vagueness and looseness of conceptual boundaries and thus the distinctiveness between pairs of genres.
- A better understanding of the change of literary semantics.

Method

Integration of the c@1-accuracy score into perspectival modeling (Underwood 2019).

1. Implementation of the c@1 score (Peñas/Rodrigo 2011), a variant of the accuracy score accounting for undecidability.

$$c@1 = \frac{1}{n}\left(n_{ac} + \frac{n_{ac}}{n} * n_u\right)$$

- n: number of all predictions
- n_{ac}: number of correct predictions (corr. answer)
- n_u: number of undecided predictions (unanswered)

2. Calculate the boundaries of undecidability that lead to optimal c@1-accuracy (= **Optimal Boundary of Undecidability / OBoU**).

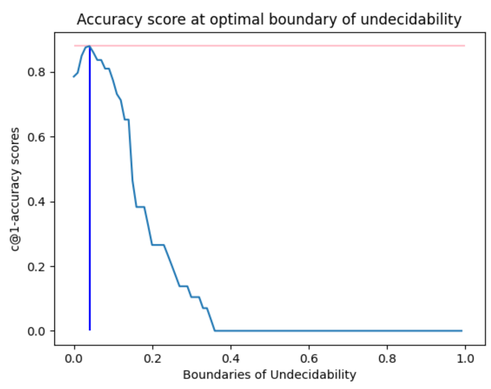

Fig. 4: Grid Search for the OBoU with a symmetric expansion of possible boundaries (horizontal) and accuracy (vertical axis) for fairy tales (M) versus novels (R).

3. Integrate step (2) into the extraction of predictive probabilities in logistic regression models for genre classification within perspectival modeling.

4. Calculate predictive probabilities averaged over 1000 iterated predictions after resampling.

5. Plot inverse predictive probability in a circular scatter plot (Fig. 2, upper and lower left) indicating semantic structure according to prototype theory.

References and Further Information

www.github.com/julianschroeter/Prototypicality

Results

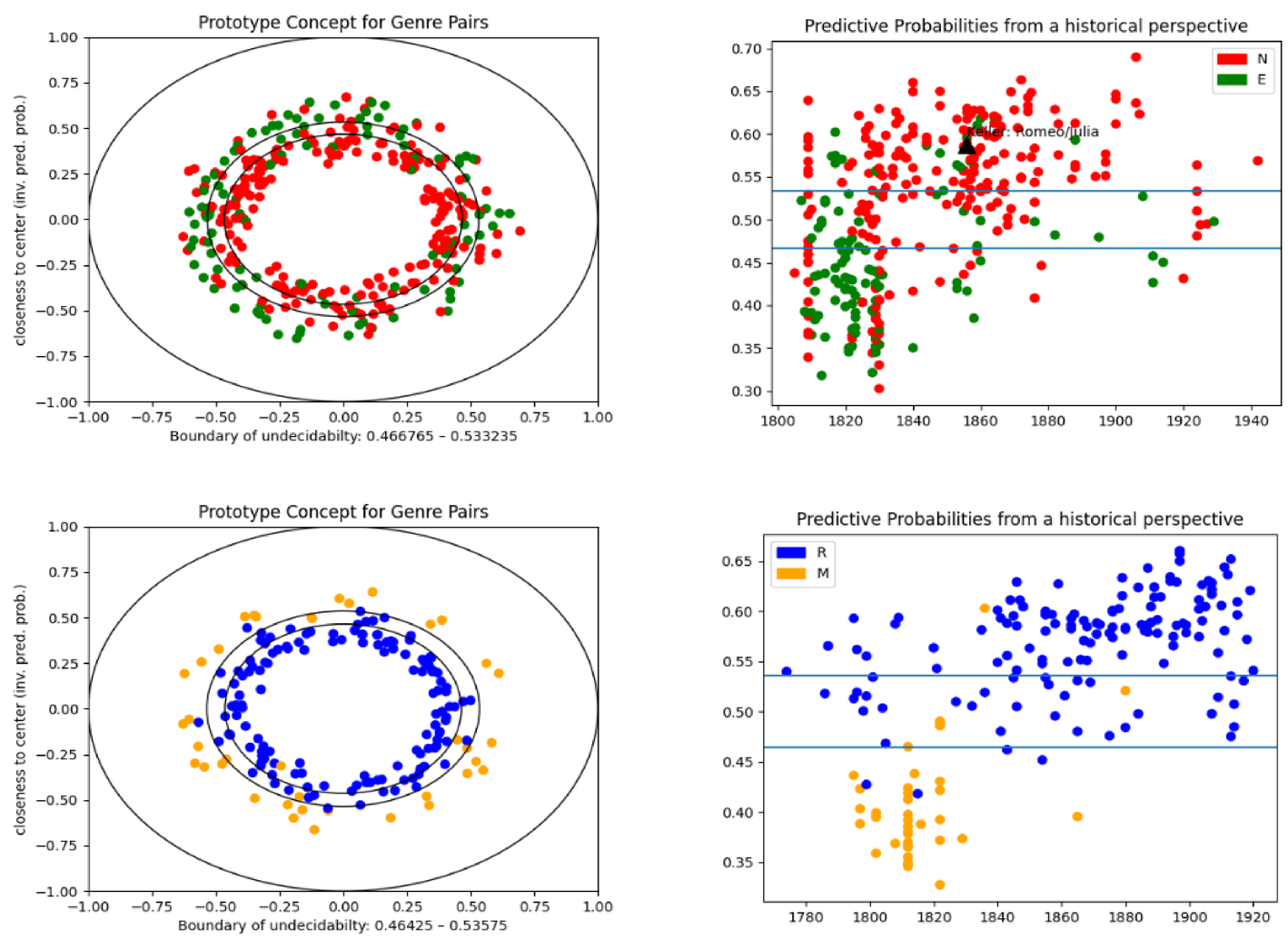

Fig. 3: For Novellen (N) versus Erzählungen (E): Prototypicality based on inverse predictive probability including the boundary of undecidability at the optimal c@1-accuracy score (oBoU) (upper left). Predictive probabilities including temporal index of the year of first publication, showing also the position of Gottfried Keller's novella „Romeo und Julia auf dem Dorfe", published in 1856 as a ‚Erzählung' but canonized later in 1871 as a ‚Novelle' (upper right). Lower level: Novels (R: ‚Romane') versus fairy tales (M: ‚Märchen') as two clearly distinct genres: Prototypicality (lower left) and predictive probabilities including temporal index (lower right).

Genre pairs		Optimal Boundary of Undecidabilty (OBoU)	c@1-accuracy score at OBoU	standard accuracy score
N. Novellen	Erzählungen	0.47–0.53	0.70	0.69
M: Märchen (fairy tales)	R: Romane (novels)	0.46–0.54	0.89	0.87

Table 1: c@1-accuracy scores at Optimal Boundaries of Undecidability (OBoU) in contrast to standard accuracy scores.

- The plots on the right side of Fig. 3 include a temporal index and provide additional information on decreasing or increasing consolidation of genre semantics over time.
- The width of the **OBoU** provides, together with the accuracy score, an approximation to the degree of conceptual looseness under the perspective of genre comparison. Put together, the c@1-accuracy score and the **OBoU** are better interpretable in terms of prototypicality than the standard accuracy score.
- Although both genre pairs (N/E and M/R) have a similar **OBoU**, the OBoU for the M/R-pair generates more clearcut zones of unambiguous classification and of undecidability, which is reflected by a higher c@1-accuracy score.

→ The **OBoU** is not significant in itself as a single measure. The c@1-accuracy score and the **OBoU** have to be combined.

Corpus

The corpus covers 740 German mid-length prose fiction texts from the 19th century.

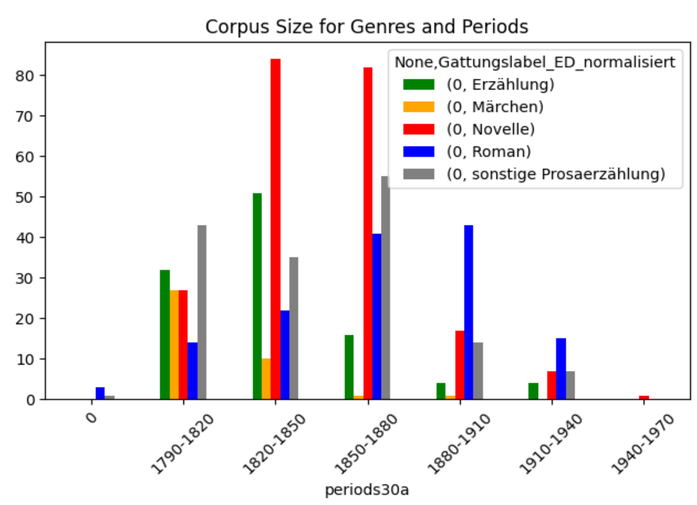

Fig. 2: Composition of the corpus covering the distribution of genres over 30-year-periods in absolute numbers.

Conclusions

- The Optimal Boundary of Undecidability (OBoU) is a useful complement for modeling the prototypicality of literary genre concepts.
- While it uses two existing methods with perspectival modeling and a c@1-accuracy score, the model presented in the poster is new to computational literary studies.
- The OBoU needs additional information on predictive accuracy to be significant and interpretable.
- In future research, numerical results of OBoU and c@1-accuracy score can be used for principled ways of expressing conceptual vagueness in quantitative terms and hypothesis-driven research designs.
- To-do: modeling based on artificial data.