Authorship attribution of Mediaeval German Text: style and contents in *Apollonius von Tyrland*

Sarah Schulz, Jonas Kuhn, and Nils Reiter
Institute for Natural Language Processing (IMS)
University of Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart
firstname.lastname@ims.uni-stuttgart.de

February 26, 2016

1. Introduction

In this paper, we describe computer-aided authorship testing on the Middle High German (MHG) text *Apollonius von Tyrland* written by *Heinrich von Neustadt* (HvN) in the late 13th century. Being based on a Latin original, HvN is suspected to incorporate other sources into the translation. We investigate assumptions regarding a segmentation of this text into parts supposedly tracking back to different sources. Our objective is it to provide a) clarification on the validity of this segmentation and b) on features that show the difference in origin of the segments. In particular, we distinguish between features related to content and to style.

2. Contents and Style

Comparing frequency distributions over frequent words has been established as a state of the art method for contrasting style across different literary texts (cf. Eder et al. (2013)). Quite recently, Herrmann et al. (2015) proposed to define style as a property constituted by "formal features which can be observed quantitatively or qualitatively" (p. 44). An important aspect of it is that style has to be based on *observable* features.

We propose cluster analysis establishing a clear cut between content and style: To measure stylistic differences, we restrict the selection to words appearing in every text of the corpus, thus are observable in each text, assuming that this is a simple way to exclude words that are markers of content. Content words (that presumably only appear in a subset of the texts) do not contribute to this understanding of style. They, in contrast, are extracted by filtering the MFW with a stop word list containing all the function words in a language. We refer to the sets of feature words extracted for a text with **content words** and **style words**.

To validate this idea, we analyse five MHG texts by three authors with the R stylo package (Eder, 2013). Figure 1 shows results for the content (a) and style (b) words. The higher similarity of *Erec* and *Tristan* in (a) compared to *Der arme Heinrich* reflects that both narratives feature knighthood as a main theme. In contrast, the narrative in *Der Arme Heinrich* involves more religious themes (faith, god), which is also reflected in the frequency tables. This distinction is clearly based on content. If we focus our analysis on style words, as in (b), we see the clustering according to authors. Thus, distinguishing frequent words of a corpus in style and content words can give us better insights into the results.

Nr.	Verses	Origin
1	1-2,905	Latin original
2	2,906-15,118	Insertion
3	15,119-17,382	Latin original
4	17,383-20,589	Insertion

Table 1: Partition of *Apollonius* according to Bockhoff and Singer (1911)

Nr.	Verses	Title
1	2920-4126	The fight with Gog and Magog
2	4126-6068	The adventures in Galacites
3	6069-7186	The duel in Syria and the Robinson Island
4	7187-10594	Bulgare war and imprisonment in Nemrot
5	10595-13512	The adventures in Chrysa
6	13517-14929	The return to Tarsus
7	17282-20639	Closing

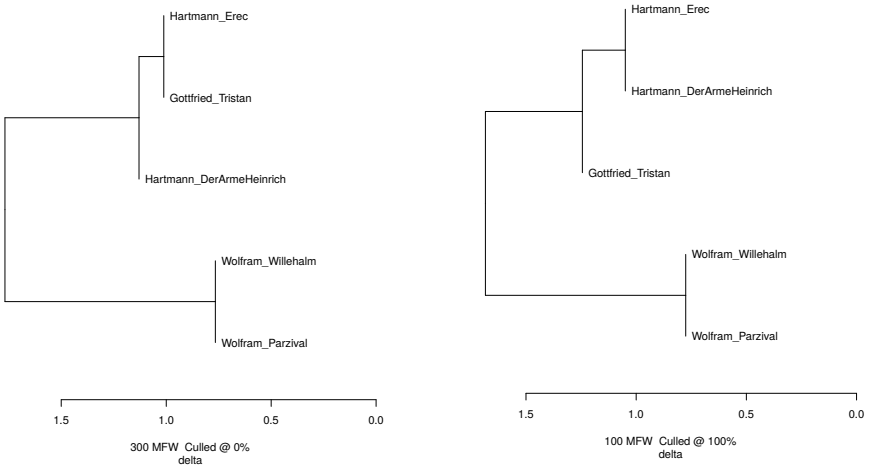
Table 2: Sub parts identified in the third section of *Apollonius*, identified by Bockhoff and Singer (1911)

3. Dissecting Heinrich von Neustadt: Apollonius

Bockhoff and Singer (1911) formulated two hypotheses regarding the internal structure and origin of the ca. 21k verses long text, regarding both the overall structure (Table 1) and the internal structure of one segment (Table 2).

To get an impression of which paragraphs can indeed be found as a distinctive group using content words and style words respectively, we split the text into 71 segments of equal length. These segments are then clustered with stylo, using delta as a similarity measure.

Our baseline consists of randomly assigning distances be-



(a) Clustering analysis using content words.

(b) Clustering analysis using style words.

Figure 1: Comparison of different groups of high frequent words and their performance on a clustering task on MHG text by three different authors. Due to largely uniform editing of MHG text in the 19th century, normalisation can be neglected (Florian, 2015).

Features	**Part**	**Recall**		**Precision**		**F-score**	
		BL	CA	BL	CA	BL	CA
Content words	Historia Apollonii 1	0.46	0.56	0.19	0.4	0.19	0.47
	Adventure	0.33	0.23	0.66	1.0	0.44	0.37
	Historia Apollonii 2	0.30	0.88	0.16	0.44	0.21	0.59
	Final apotheosis	0.19	0.45	0.17	0.25	0.18	0.32
Style words	Historia Apollonii 1	0.46	0.56	0.19	0.21	0.19	0.31
	Adventure	0.33	0.28	0.66	0.92	0.44	0.43
	Historia Apollonii 2	0.30	0.75	0.16	0.33	0.21	0.46
	Final apotheosis	0.19	0.72	0.17	0.5	0.18	0.67

Table 3: Results of the clustering analysis for style and content words respectively regarding the overall structure hypothesis of Apollonius. Since clustering methods do not provide class labels for an evaluation of the performance with respect to precision, recall and F-score, we need to map the clusters onto the parts of the hypothesis. This is done manually in such a way that F-score is maximised. BL: Baseline, CA: Cluster Analysis.

tween the segments, drawn from a uniform distribution. We sample baseline results 1000 times. The baseline results give an impression on how well an uninformed method covers with the hypothesis classification introduced in Section 3. Comparing these to the results of our methods can inform us on which method goes in line with the hypothesis as opposed to a random classification.

Regarding the **first hypothesis** (Table 3), we observe that for both feature sets the F-score lies above the baseline for all parts except the third. This seems reasonable since this part is suspected to be based on different sources and therefore might be more heterogenius both in content and in style. Style seems to be more homogeneous (F-Score above baseline) throughout the entire text whereas content seems to be heterogeneous especially in the adventure part introduced by HvN (F-Score below baseline). This is in line with the hypothesis, considering that HvN's insertions report on different adventures.

Analysing these heterogeneous parts further (**second hypothesis**, Table 4), we see heterogeneity in terms of content for all but one part, *The duel in Syria*. It seems homogeneous in style whereas *The adventures in Galacites* shows tendencies towards an heterogeneous style.

4. Conclusion

Both feature sets show similar tendencies and support a major part of the hypotheses by Bockhoff and Singer (1911) regarding parts suspected as insertions. Nevertheless, differences in content cannot clearly confirm the suspicion that HvN incorporated other sources. He might have created additional adventures by himself. Bockhoff and Singer (1911) do not cite sources from which HvN copied narratives, making it difficult to tackle exactly. Overall differences in style are much less significant than differences in content, which is inline with the hypothesis.

5. References

A. Bockhoff and S. Singer. 1911. *Heinrichs von Neustadt Apollonius von Tyrland und seine Quellen. Ein Beitrag zur mittelhochdeutschen und byzantinischen Literaturgeschichte von A. Bockhoff und S. Singer*. Sprache

Features	Part	Recall		Precision		F-score	
		BL	CA	BL	CA	BL	CA
content words	The fight with Gog and Magog	0.43	0.75	0.26	0.6	0.32	0.67
	The adventures in Galacites	0.36	0.50	0.32	0.5	0.34	0.5
	The duel in Syria	0.39	0.25	0.25	0.11	0.31	0.15
	Bulgare war	0.26	0.33	0.48	0.8	0.33	0.46
	The adventures in Chrysa	0.24	0.5	0.43	0.71	0.31	0.59
	The return to Tarsus	0.25	0.83	0.3	0.71	0.29	0.77
style words	The fight with Gog and Magog	0.43	0.75	0.26	0.375	0.32	0.5
	The adventures in Galacites	0.36	0.25	0.32	0.25	0.34	0.25
	The duel in Syria	0.39	0.5	0.25	0.33	0.31	0.4
	Bulgare war	0.26	0.5	0.48	0.46	0.33	0.48
	The adventures in Chrysa	0.24	0.4	0.43	0.67	0.31	0.5
	The return to Tarsus	0.25	0.33	0.3	0.4	0.29	0.36

Table 4: Results of the clustering analysis for style and content words respectively regarding structure of the parts of Apollonius attributed to Heinrich von Neustadt. Final part has been removed from the discussion due to its short length.

und Dichtung : Forschungen zur Linguistik und Literaturwissenschaft [dann] zur Sprach- und Literaturwissenschaft. J. C. B. Mohr.

Maciej Eder, Mike Kestemont, and Jan Rybicki. 2013. Stylometry with r: a suite of tools. In *Digital Humanities 2013: Conference Abstracts*, pages 487–89, Lincoln, NE. University of Nebraska–Lincoln.

Maciej Eder. 2013. Mind your corpus: systematic errors in authorship attribution. *LLC*, 28(4):603–614.

Kragl Florian. 2015. Normalmittelhochdeutsch. theorieentwurf einer gelebten praxis. *Zeitschrift fur Deutsches Altertum und Deutsche Literatur*, 144:1–27.

J. Berenike Herrmann, Karina van Dalen-Oskam, and Christof Schöch. 2015. Revisiting style, a key concept in literary studies. *Journal of Literary Theory*, 9(1):25–52.