

# Vorläufige Erkenntnisse

Pia Henning

December 16, 2020

### **Abstract**

Vorläufige Erkenntnisse im Zusammenhang mit der Arbeit und Erforschung der Lyrik des 19. Jahrhunderts.

# Chapter 1

## Datenexploration

### 1.1 Korpusaufbau

Korpus	Anzahl Balladen/Autoren	Anzahl Lyrik/Autoren	Gesamt	Anzahl Autoren	Tokens	Anzahl Segmente
ersterKorpus	163/47	63/50 + 9 NA	224	89	67.019	118/15
Autorenkorpus	163/47	302/38	465	47	79.895	41
angepasster Korpus	163/47	206/50 + 9 NA	369	89	78.584	38
angepasster Korpus (groß)	163/47	206/50 + 9 NA	369	89	78.584	118/38

Figure 1.1: Zusammenstellung der Korpora.

Korpora	Textlängen Balladen	Textlängen Lyrik
ersterKorpus	ca. 100-1550	ca. 25-420
Autorenkorpus	ca. 100-1550	ca. 20-270
angepasster Korpus	ca. 100-1550	ca. 20-2000

Figure 1.2: Textlängen.



Figure 1.4: Type-Token-Ratio der Balladen

Figure 1.5: Type-Token-Ratio der Lyrik



Figure 1.6: Type-Token-Ratio der Lyrik im Autorenkorpus

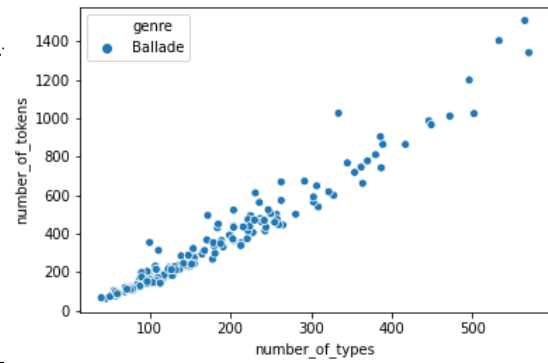


Figure 1.7: Type-Token-Ratio der Lyrik im der Größe nach angepassten Korpus.

### **Korpus**

gilt für alle Korpora (siehe Abbildung 1.4)

### **Balladen**

- bis ca. 600
- cluster im ersten Drittel
- lineare Verteilung

### **Korpus**

ersterKorpus (siehe Abbildung 1.5)

### **Lyrik**

- bis ca. 200
- cluster im ersten Drittel
- lineare Verteilung

Autorenkorpus (siehe Abbildung 1.6)

- bis ca. 160
- cluster in der ersten Hälfte
- lineare Verteilung

Angepasster Korpus (siehe Abbildung 1.7)

- bis ca. 900
- cluster im ersten Viertel
- lineare Verteilung

Figure 1.3: Type-Token-Ratio

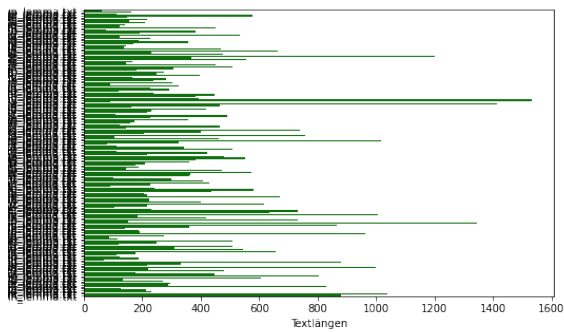


Figure 1.8: Textlängen des Balladenkorpus

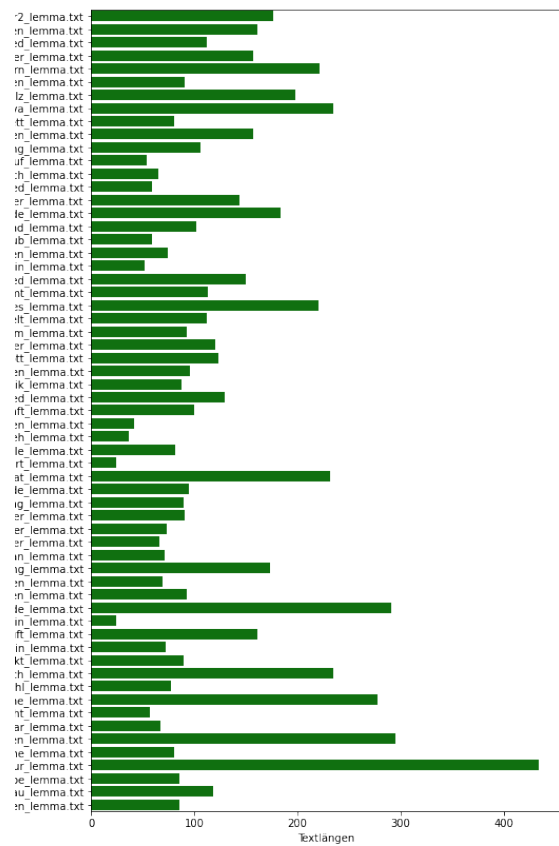


Figure 1.9: Textlängen des Lyrikkorpus

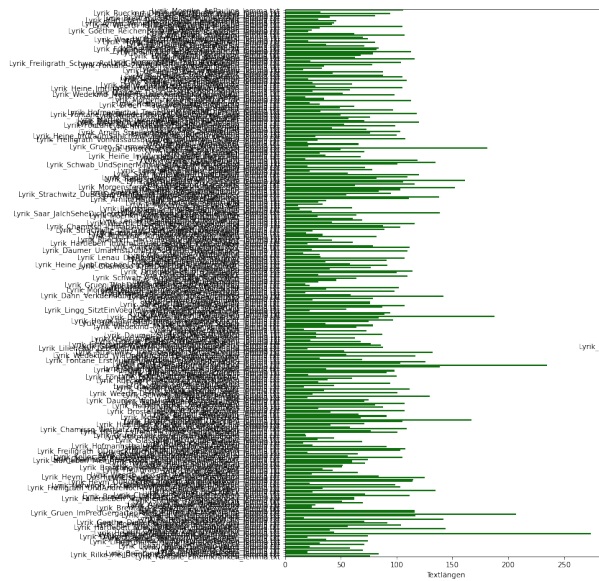


Figure 1.10: Textlängen des Lyrikkorpus im Autorenkorpus

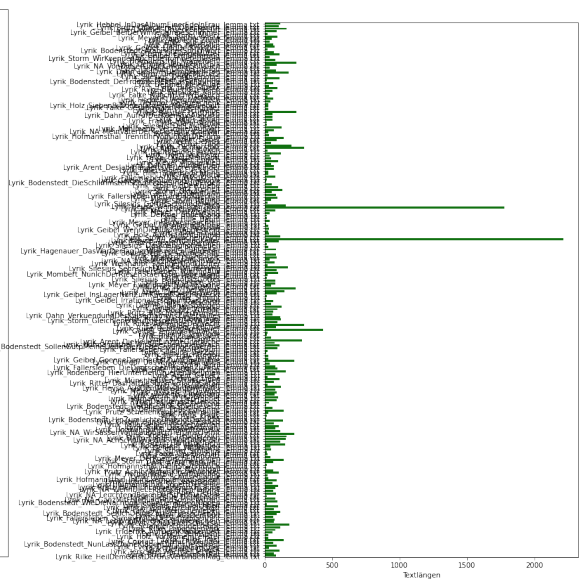


Figure 1.11: Textlängen des Lyrikkorpus im angepassten Korpus



Figure 1.12: Wordcloud des Balladenkorpus

Figure 1.13: Wordcloud des Lyrikkorpus







Figure 1.16: MFW des Balladenkorpus, ohne Stopwords

Figure 1.17: MFW des Lyrikkorpus, ohne Stopwords

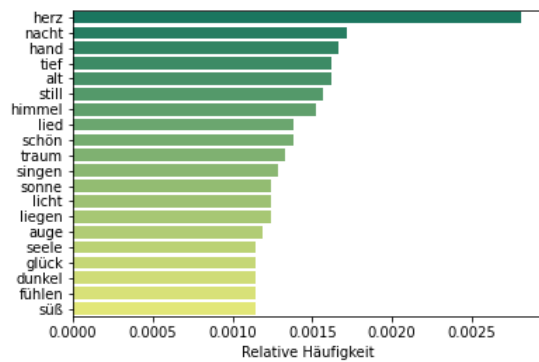


Figure 1.18: MFW des Lyrikkorpus, ohne Stopwords im Autorenkorpus

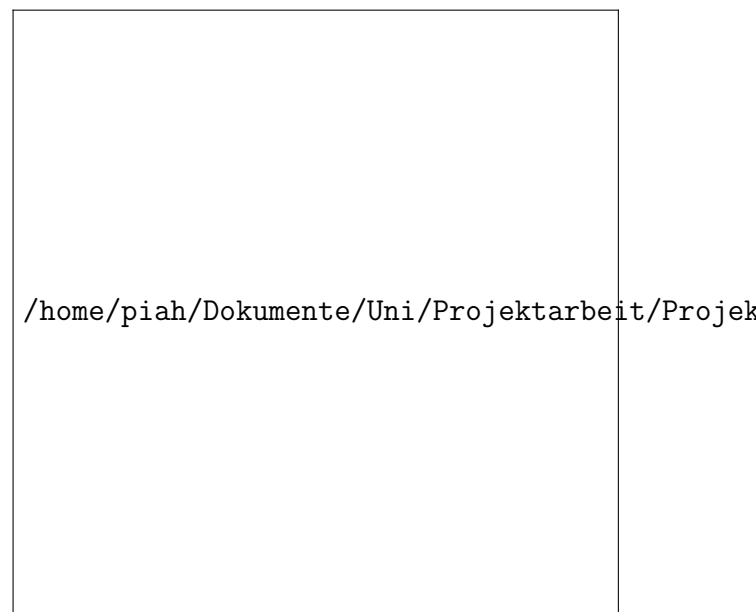


Figure 1.19: MFW des Lyrikkorpus, ohne Stopwords im angepassten Korpus.

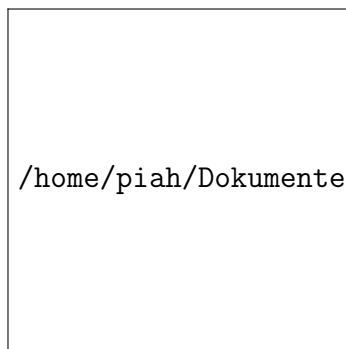


Figure 1.20: Textlängen

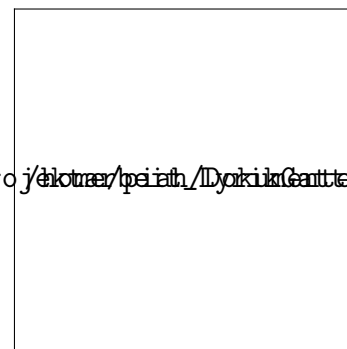


Figure 1.21: Textlängen  
des Autorenkorpus

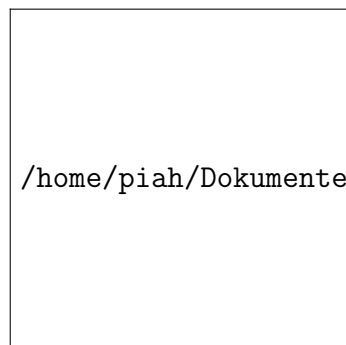


Figure 1.22: Textlängen  
des angepassten Korpus

# Chapter 2

## Stilometrische Analyse

### 2.1 Zeta

Auf Grund der Zetaergebnisse, in Tabellen 2.1 und 2.2 aufgelistet, wird für folgende Analysen eine Segmentgröße von 500 Token pro Segment gewählt. Auch eine Anpassung der Größe erscheint auf Grund der stark abweichenden und gegen -1 laufenden Ergebnisse (starke Tendenz in Richtung Lyrik, auf Grund der geringen Korpusgröße) sinnvoll.

### 2.2 Delta

ersterKorpus

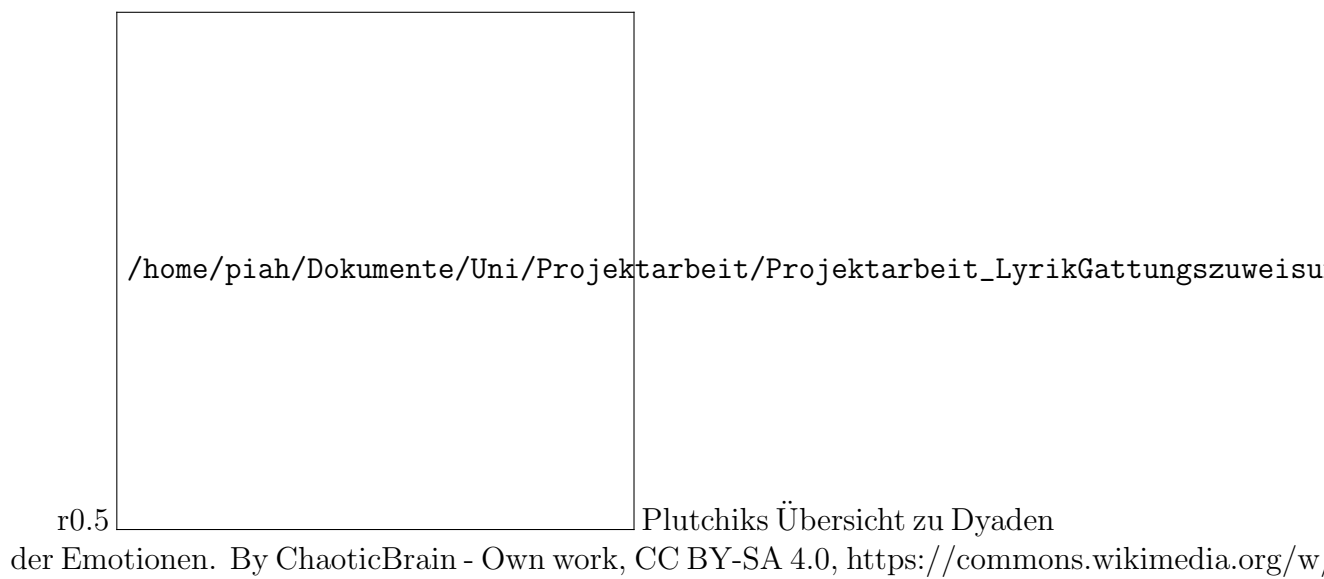
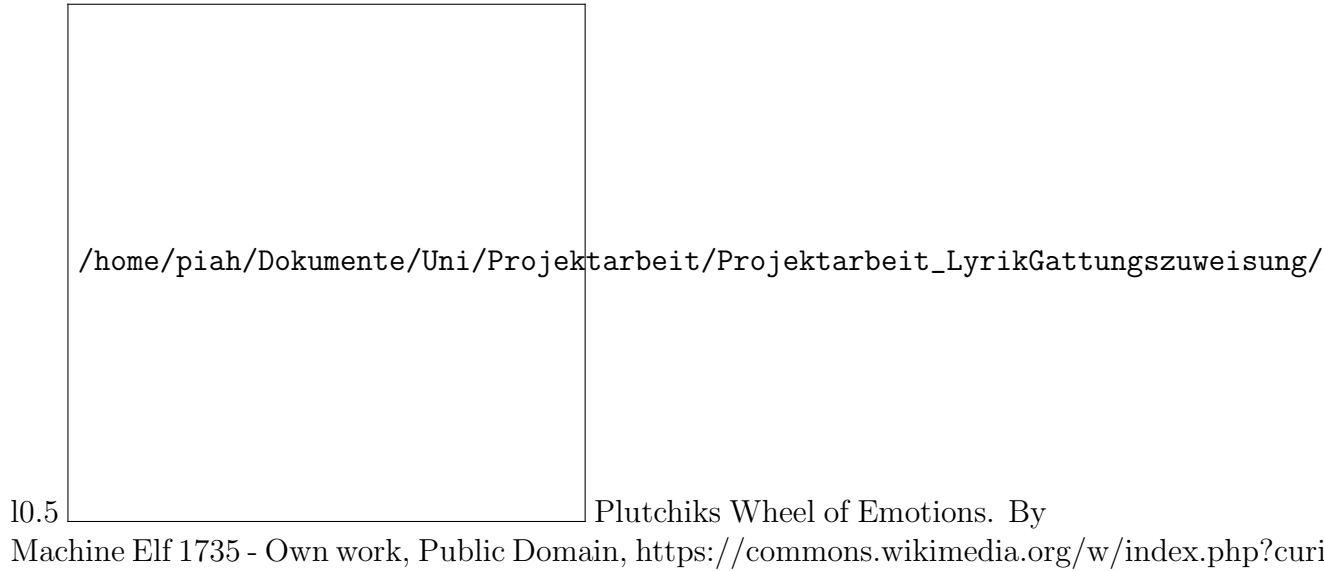
- unbereinigter Korpus
  - kaum Delta Varianz zu erkennen siehe Abbildung 2.27

### 2.3 Wordembedding

- PCA siehe Abbildung 2.30
  - deutliches Cluster
- t-SNE siehe Abbildung 2.31
  - deutliches Cluster

Wie in Abbildung 2.32 und 2.33 zu erkennen ist, sind die Ergebnisse des angepassten Autorenkorpus noch aussagekräftiger.

## 2.4 Scikit-learn



Korpora	100	500	1000
<b>Substantive</b>			
ersterKorpus		5 über 0,2	7 über 0,3 2.3
Autorenkorpus	0	2 über 0,3	4 über 0,3
angepasster Korpus		7 über 0,2	3 über 0,3
<b>Adjektive</b>			
ersterKorpus	0	3 über 0,2	8 über 0,3 2.5
Autorenkorpus	0	3 über 0,2	4 über 0,3
angepasster Korpus		2 über 0,2	5 über 0,3
<b>Verben</b>			
ersterKorpus		14 über 0,2; 3 über 0,3	10 über 0,3 2.9
Autorenkorpus		10 über 0,2	13 über 0,3
angepasster Korpus		10 über 0,2	17 über 0,3
<b>Sprechmarker</b>			
ersterKorpus		4 über 0,2; 1 über 0,4	4 über 0,3 2.7
Autorenkorpus	0	2 über 0,2	4 über 0,2; 2 über 0,5
angepasster Korpus		2 über 0,2	4 über 0,2; 1 über 0,5
<b>Emotionen + SentiWS</b>			
ersterKorpus		4 über 0,2	7 über 0,3 ??
Autorenkorpus	0	2 über 0,2	6 über 0,3
angepasster Korpus		1 über 0,2	6 über 0,3
<b>Emotionen</b>			
ersterKorpus		0 über 0,2	3 über 0,2 ??
Autorenkorpus	0	0	1 über 0,2
angepasster Korpus	0	0	2 über 0,2

Figure 2.1: Zetawerte der Balladen

Korpora	100	500	1000
<b>Substantive</b>			
ersterKorpus		8 über -0,4	7 über -0,6
Autorenkorpus	0	3 über -0,3	4 über -0,4; 14 über -0,3
angepasster Korpus		4 über -0,3	7 über -0,4
<b>Adjektive</b>			
ersterKorpus	0	11 über -0,3	11 über -0,4
Autorenkorpus	0	11 über -0,2	9 über -0,3
angepasster Korpus		19 über -0,2	18 über -0,3
<b>Verben</b>			
ersterKorpus		11 über -0,3	8 über -0,5
Autorenkorpus		7 über -0,2	9 über -0,3
angepasster Korpus		10 über -0,2	20 über -0,3
<b>Sprechmarker</b>			
ersterKorpus		2 über -0,2	5 über -0,2
Autorenkorpus	0	1 über -0,2	3 über -0,2
angepasster Korpus		1 über -0,2	3 über -0,2
<b>Emotionen + SentiWS</b>			
ersterKorpus		8 über -0,4	9 über -0,5
Autorenkorpus	0	8 über -0,25	8 über -0,4
angepasster Korpus		10 über -0,25	15 über -0,4
<b>Emotionen</b>			
ersterKorpus		10 über -0,2	11 über -0,3
Autorenkorpus	0	3 über -0,2	6 über -0,3
angepasster Korpus	0	7 über -0,2	9 über -0,3 (fühlen bei ca. -0,83

Figure 2.2: Zetawerte der übrigen Gedichte.



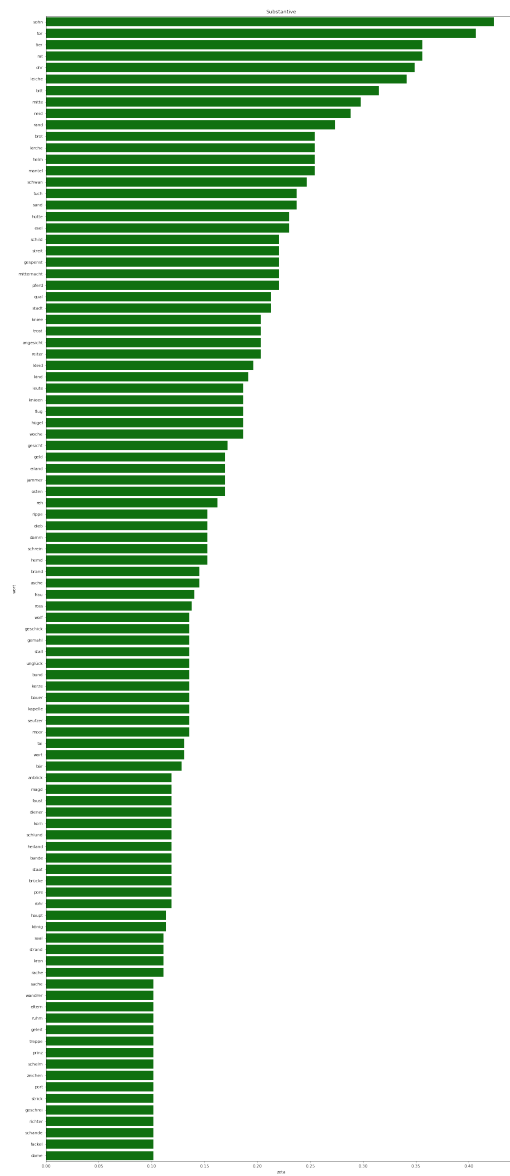


Figure 2.3: Zetawerte der Substantive des Balladenkorpus 15

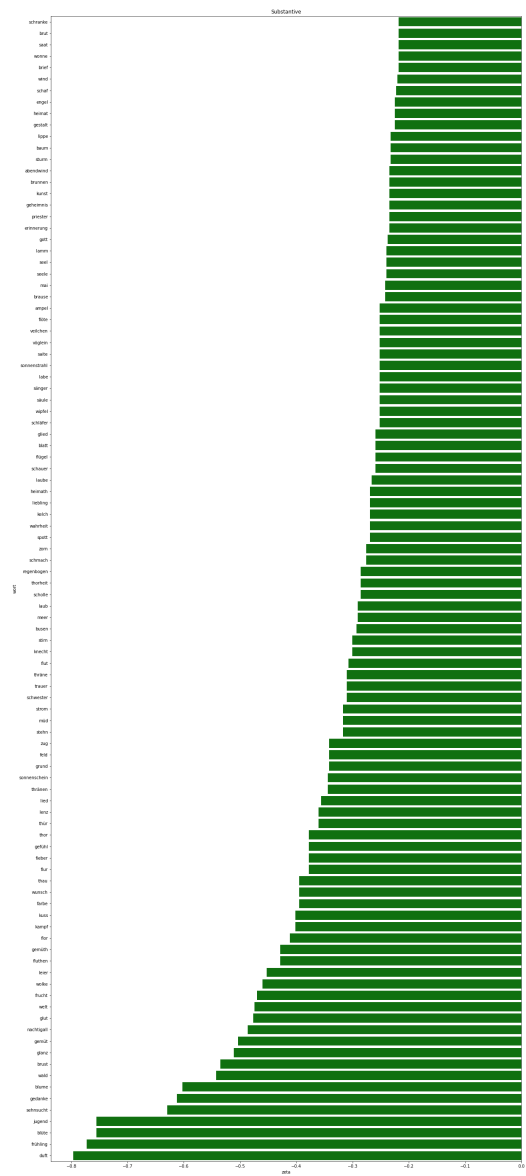
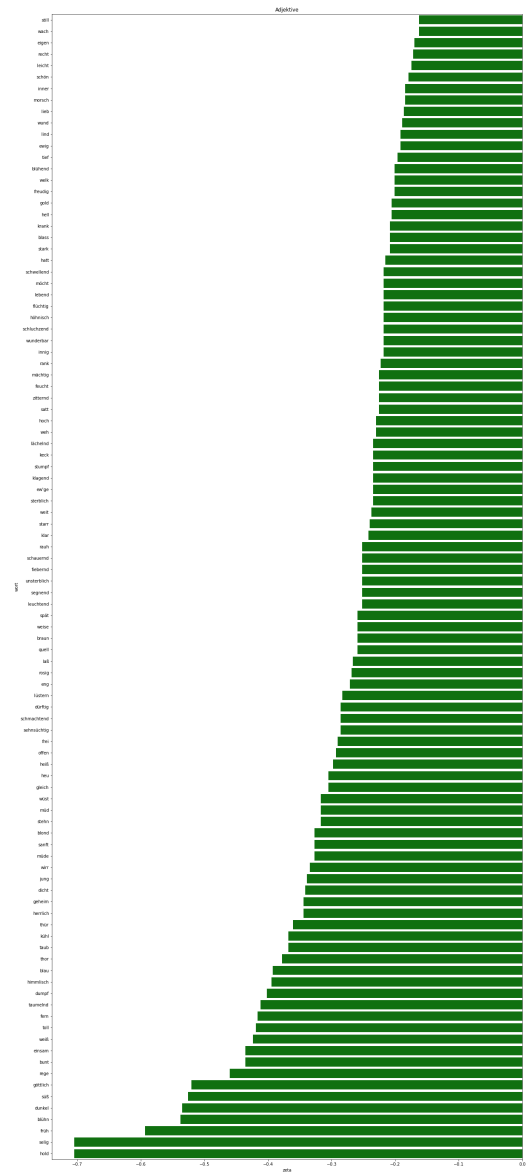
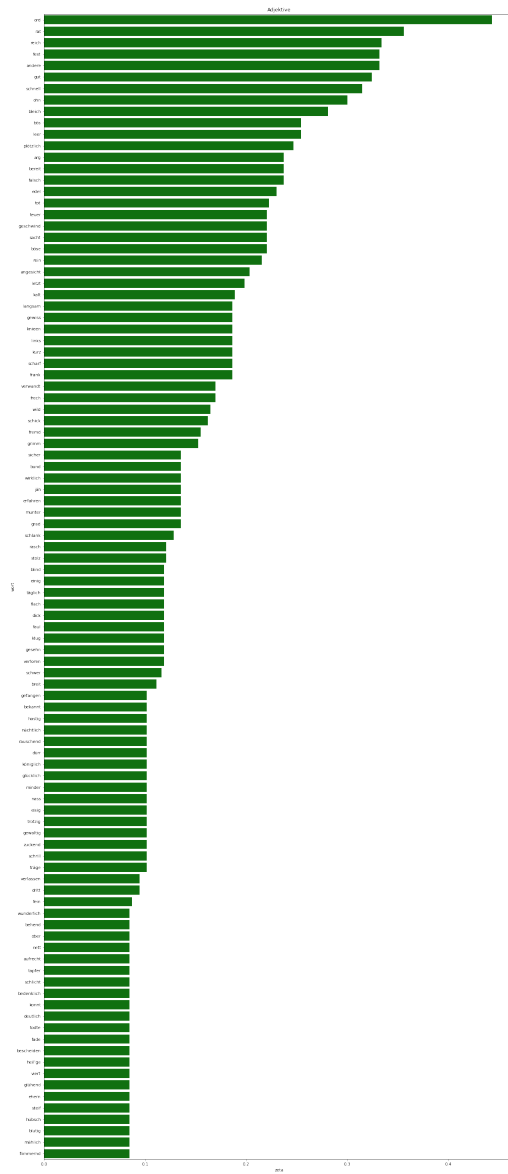
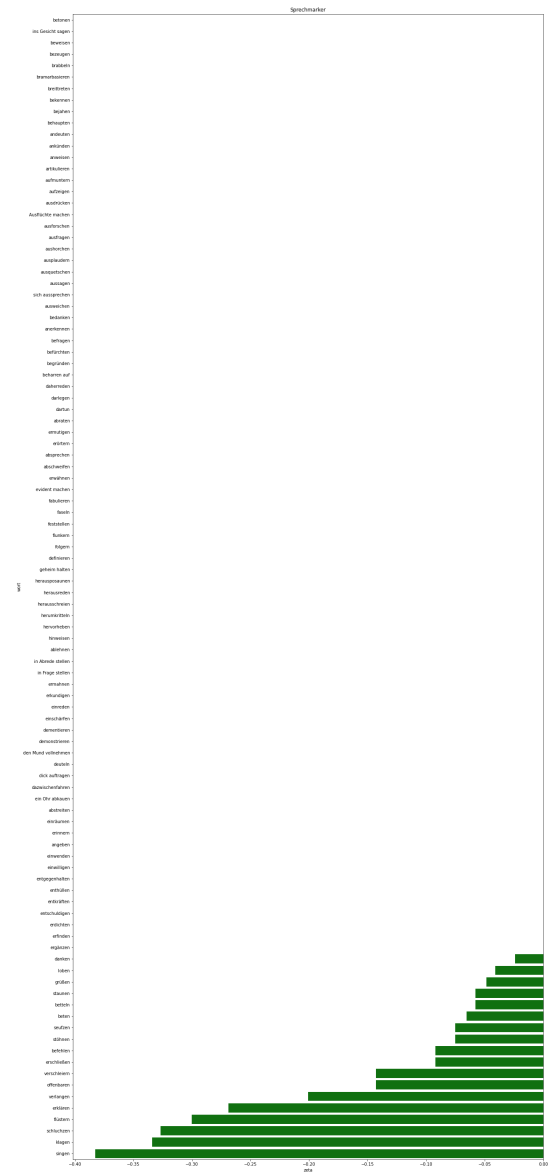
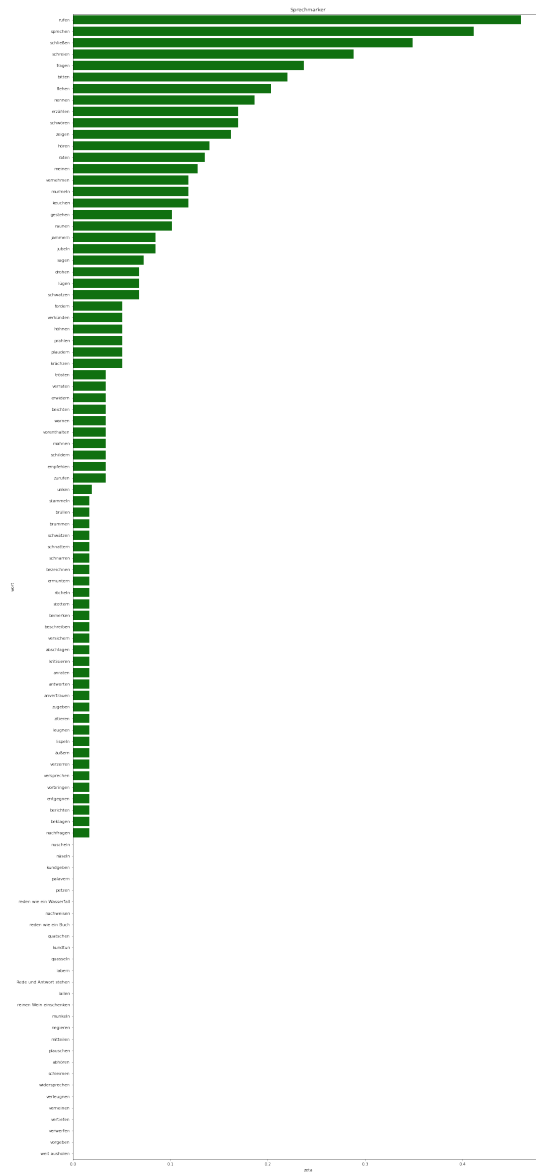
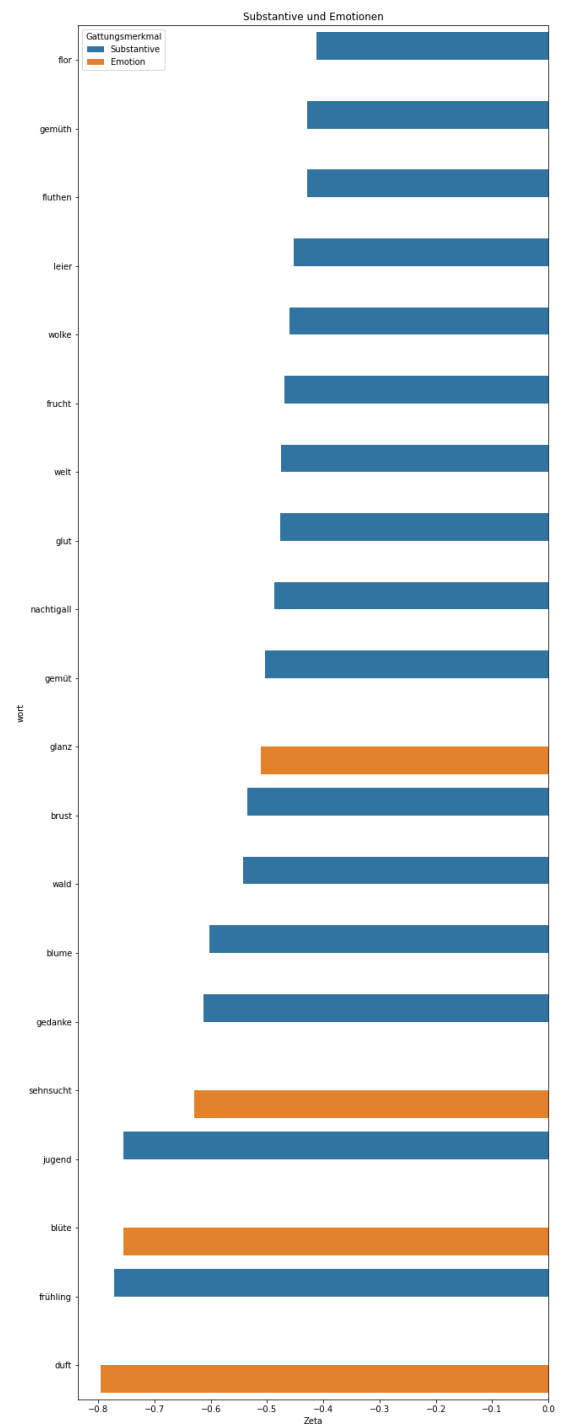
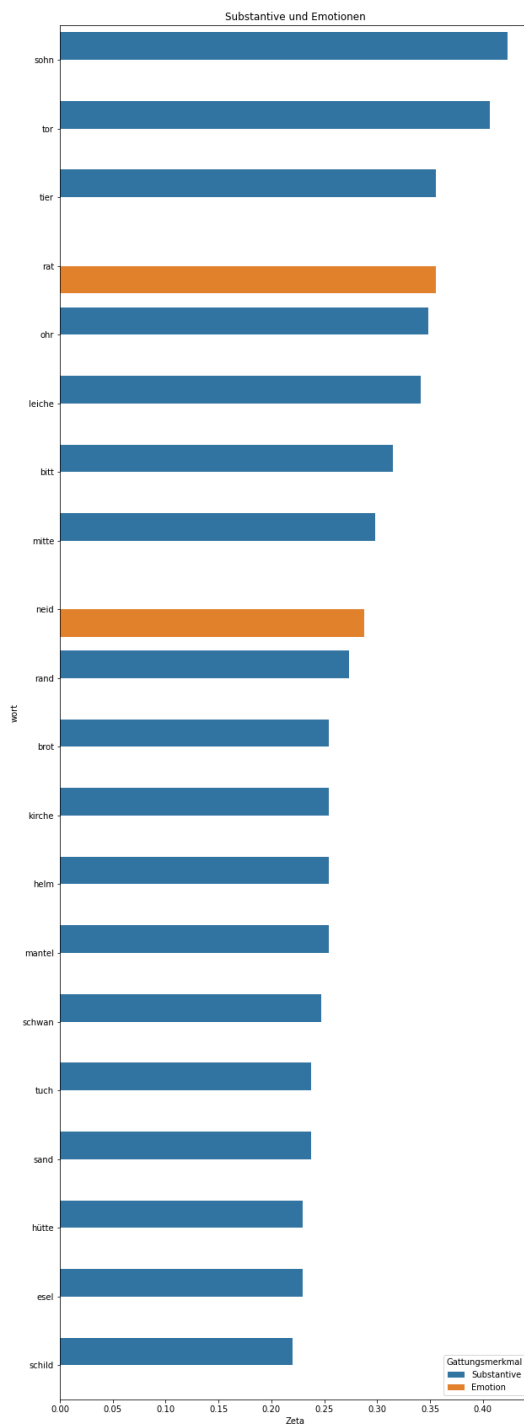


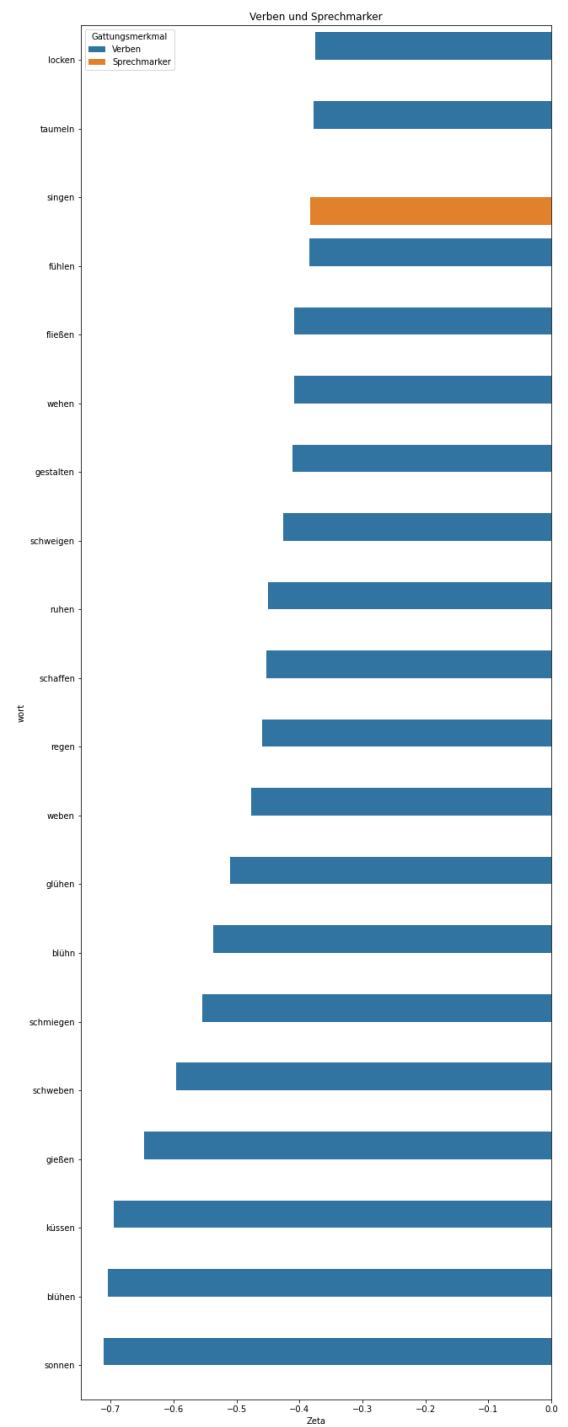
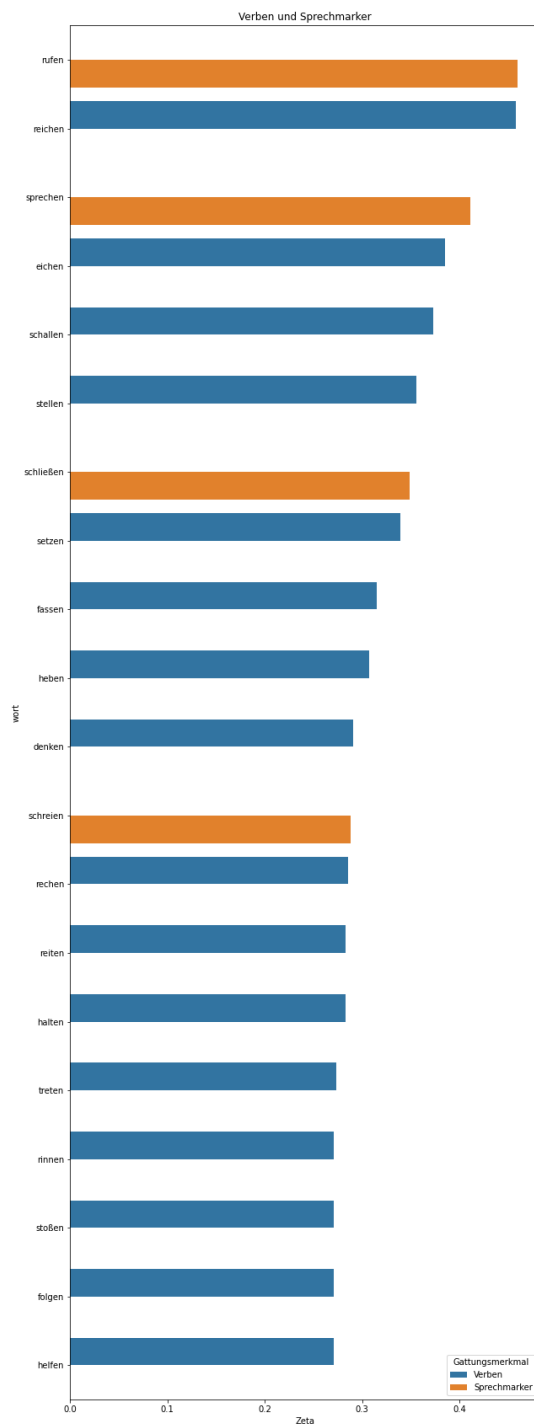
Figure 2.4: Zetawerte der Substantive des Lyrikkorpus











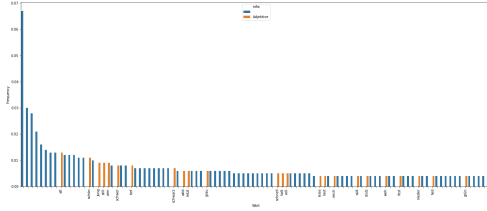


Figure 2.15: Adjektive innerhalb der MFW im Balladenkorpus

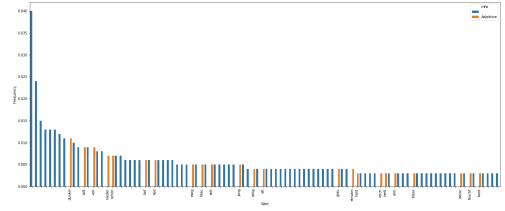


Figure 2.16: Adjektive innerhalb der MFW im Lyrikkorpus

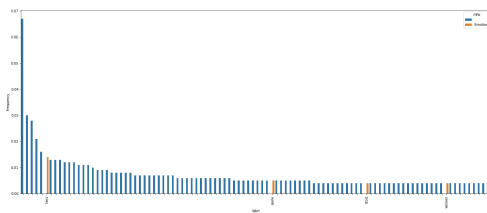


Figure 2.17: Emotionswörter innerhalb der MFW im Balladenkorpus

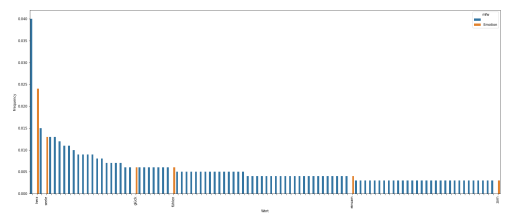


Figure 2.18: Emotionswörter innerhalb der MFW im Lyrikkorpus

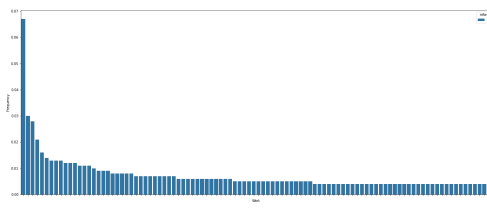


Figure 2.19: Plutchik-Emotionswörter innerhalb der MFW im Balladenkorpus

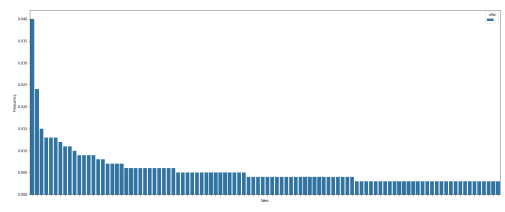


Figure 2.20: Plutchik-Emotionswörter innerhalb der MFW im Lyrikkorpus

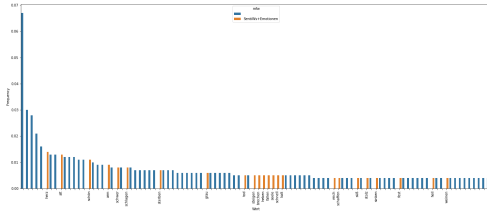


Figure 2.21: SentiWS-Emotionswörter innerhalb der MFW im Balladenkorpus

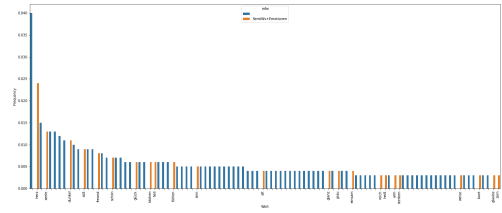


Figure 2.22: SentiWS-Emotionswörter innerhalb der MFW im Lyrikkorpus

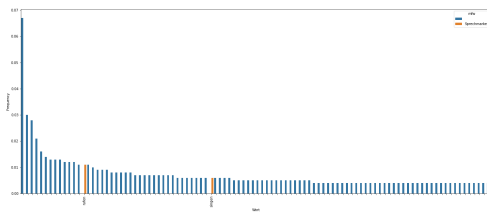


Figure 2.23: Sprechmarker innerhalb der MFW im Balladenkorpus

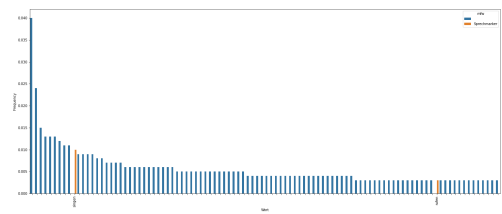


Figure 2.24: Sprechmarker innerhalb der MFW im Lyrikkorpus

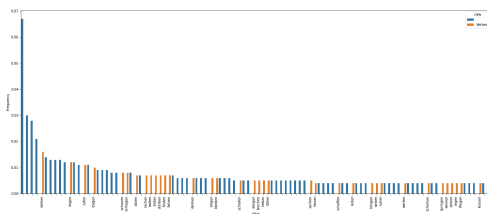


Figure 2.25: Verben innerhalb der MFW im Balladenkorpus

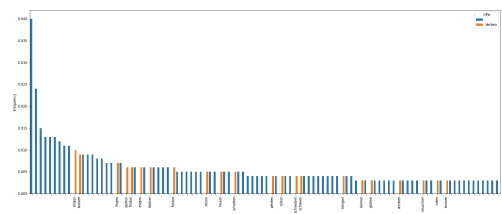


Figure 2.26: Verben innerhalb der MFW im Lyrikkorpus



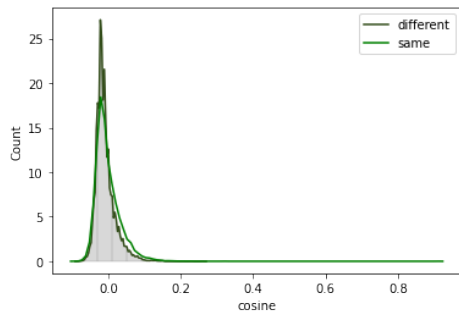


Figure 2.27: Unterscheidung von Balladen und Lyrik allgemein auf Grund des Deltas

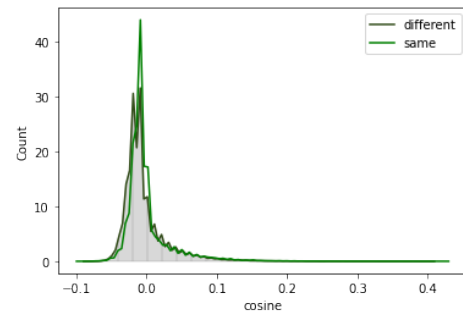


Figure 2.28: Unterscheidung von Balladen und Lyrik allgemein auf Grund des Deltas (Autorenkorpus)

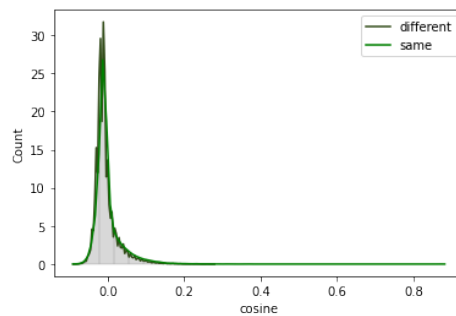


Figure 2.29: Unterscheidung von Balladen und Lyrik allgemein auf Grund des Deltas (angepasster Korpus)

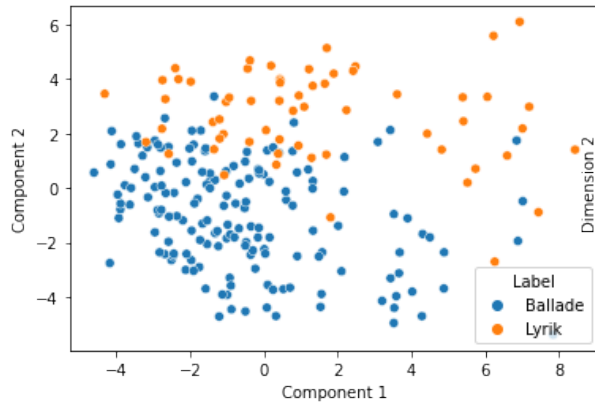


Figure 2.30: Wordembedding Cluster nach PCA

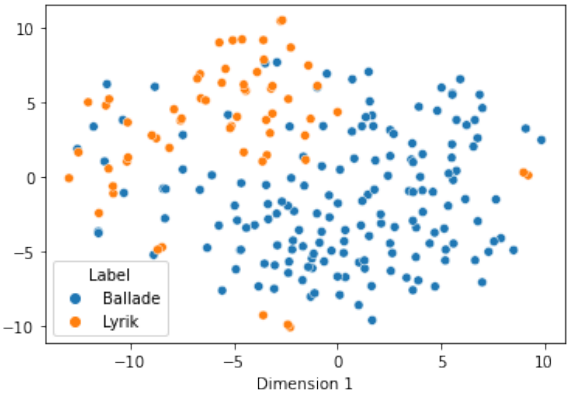


Figure 2.31: Wordembedding Cluster nach t-SNE

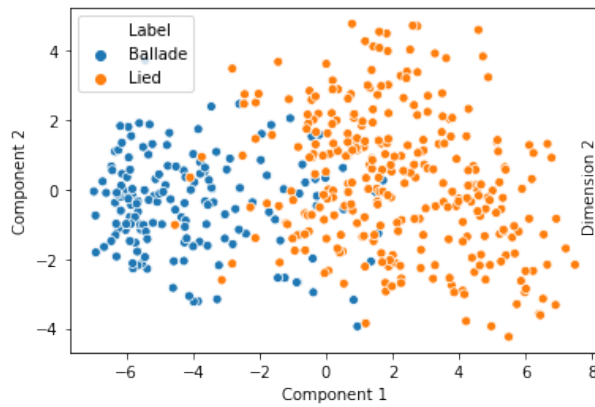


Figure 2.32: Wordembedding Cluster nach PCA (Autorenkorpus)

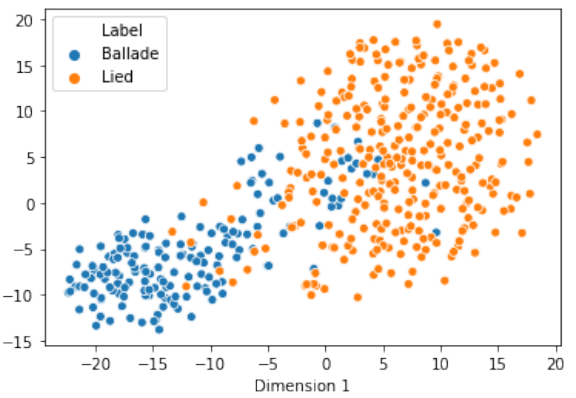


Figure 2.33: Wordembedding Cluster nach t-SNE (Autorenkorpus)