

Bedeutungsanalyse und *distant reading*

Fotis Jannidis (Würzburg)

Ringvorlesung ‚Digital Humanities‘ (BBAW), 16.1.2018

Was ist ‚distant reading‘?

‚distant reading‘

‚macro analysis‘

‚scalable reading‘

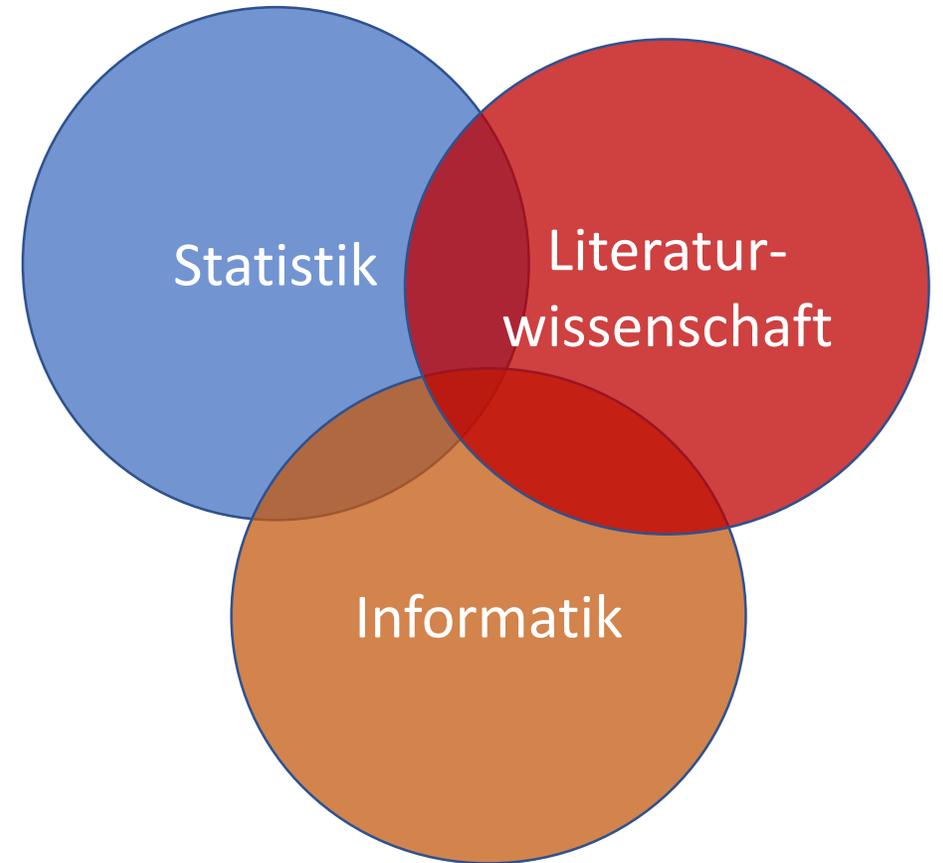
‚quantitative analysis of literary texts‘

‚computational literary studies‘

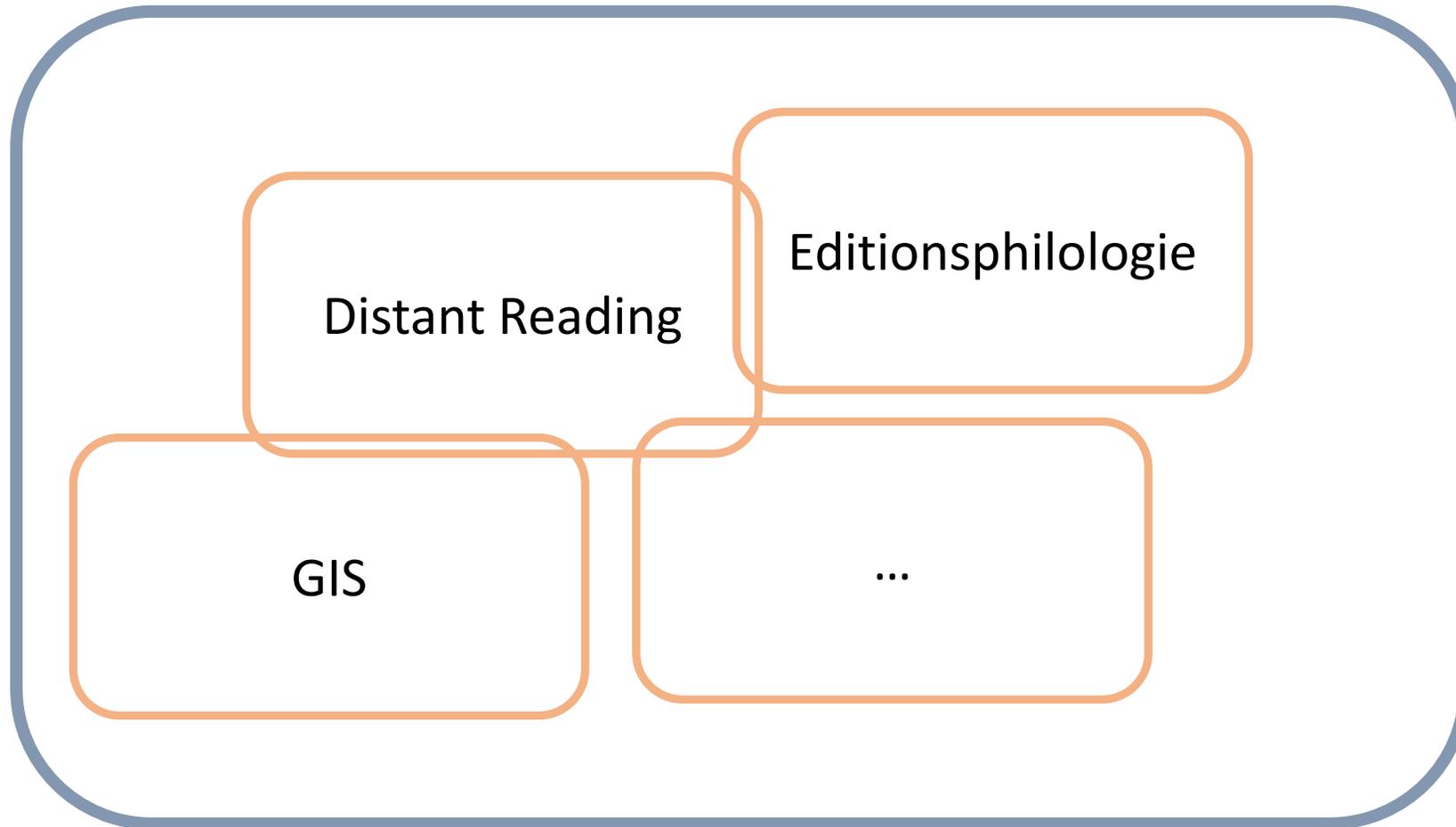
‚korpusbasierte Analyse literarischer Texte‘

‚cultural analytics‘

‚Stilometrie‘



Verhältnis zu DH



Arbeitsfelder des ‚distant reading‘

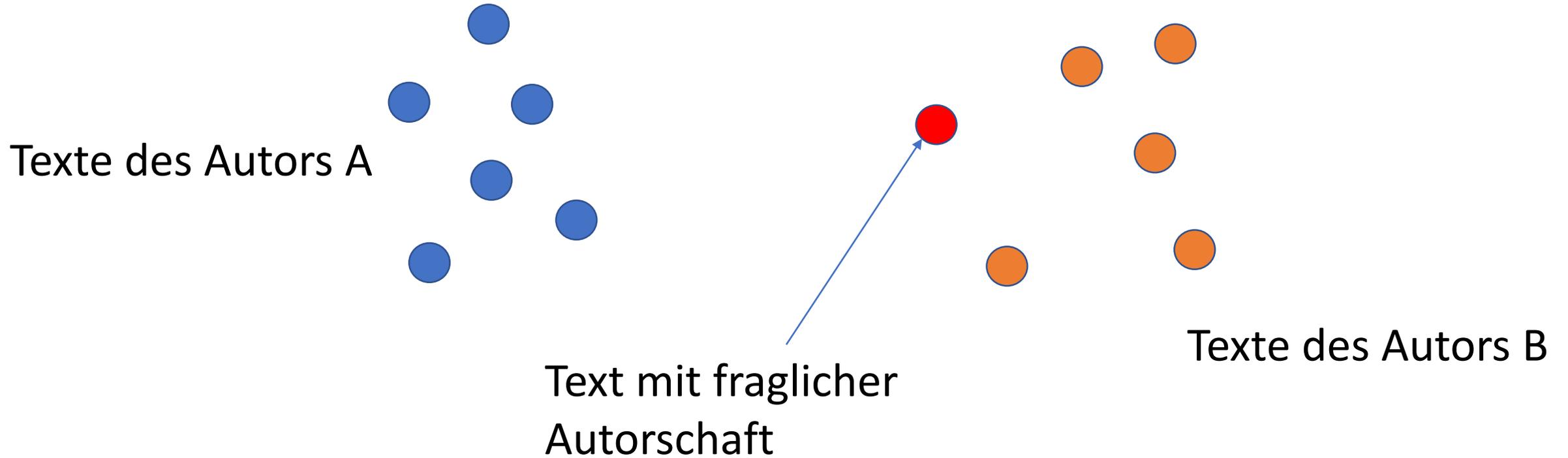
- Stilometrie
- Figurenkonstellationen
- Sentiment Analyse
- Gattungsanalyse
- ...
- Individualisierte Untersuchungsanordnungen

Stilometrie

Stilometrie: Burrow's Delta

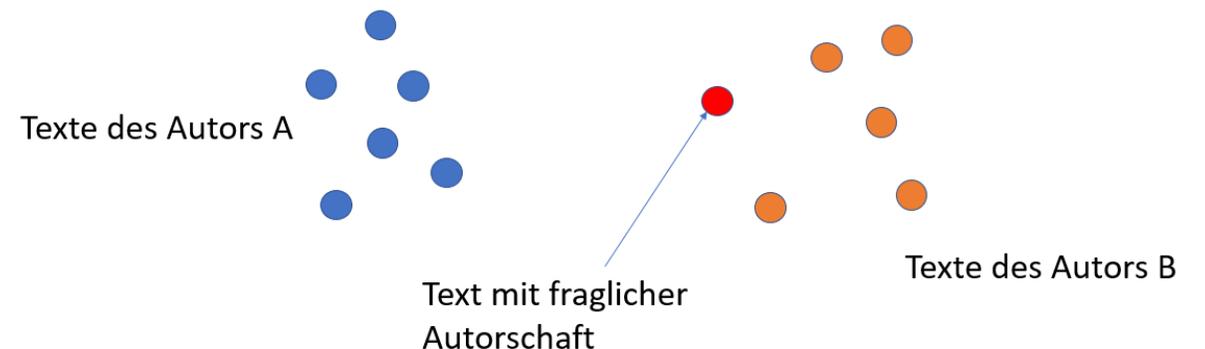
- Geschichte: John Burrows (Newcastle, Australien) seit den 1980ern
 - Busa-Preisrede 2001
- Heute: Erfolgreichstes und besterforschtes Verfahren der Stilometrie
- Grundlage der Analyse sind die n häufigsten Wörter eines Textes, wobei n zwischen 2.000 und 3.000 liegen sollte.
 - Bag of words – Modell des Textes
- Ursprüngliche Verwendungsweise: Autorschaftsattribuion
 - Funktion: Reduktion der Anzahl möglicher Kandidaten
 - Burrows Kritik an Foucault

Prinzipielle Funktionsweise



Funktionsweise

- Delta ist ein Clustering-Verfahren
- Grundlage ist die Berechnung des Abstands zwischen den Texten
- Kleinerer Abstand wird als größere Ähnlichkeit gedeutet
- Zuschreibung zum ähnlichsten Kandidaten
 - Stimmt natürlich nur, wenn der richtige Autor Teil des Untersuchungssets ist



Burrow's Delta - Algorithmus

- Nimm ein Korpus von Texten und zähle alle Wörter in allen Texten
- Ermittle die n häufigsten Wörter bezogen auf alle Texte
- Berechne den Mittelwert μ und die Standardabweichung σ für jedes der n Wörter über alle Texte
- Für jeden Text des Korpus:

Berechne für jedes der n häufigsten Wörter (x_i = Häufigkeit des Worts i im Text):

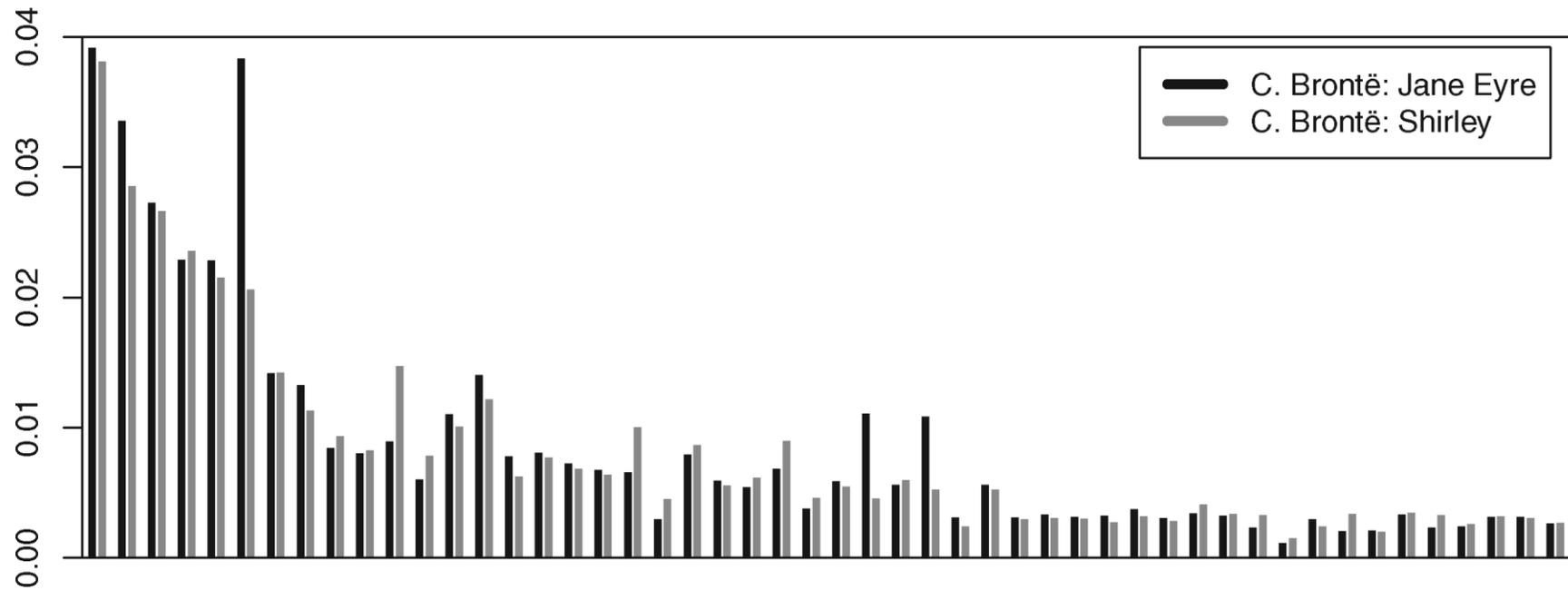
$$u_i = \frac{x_i - \mu_i}{\sigma_i} \quad (\text{z-score})$$

- Berechne den Abstand jedes Text zu jedem Text, durch:

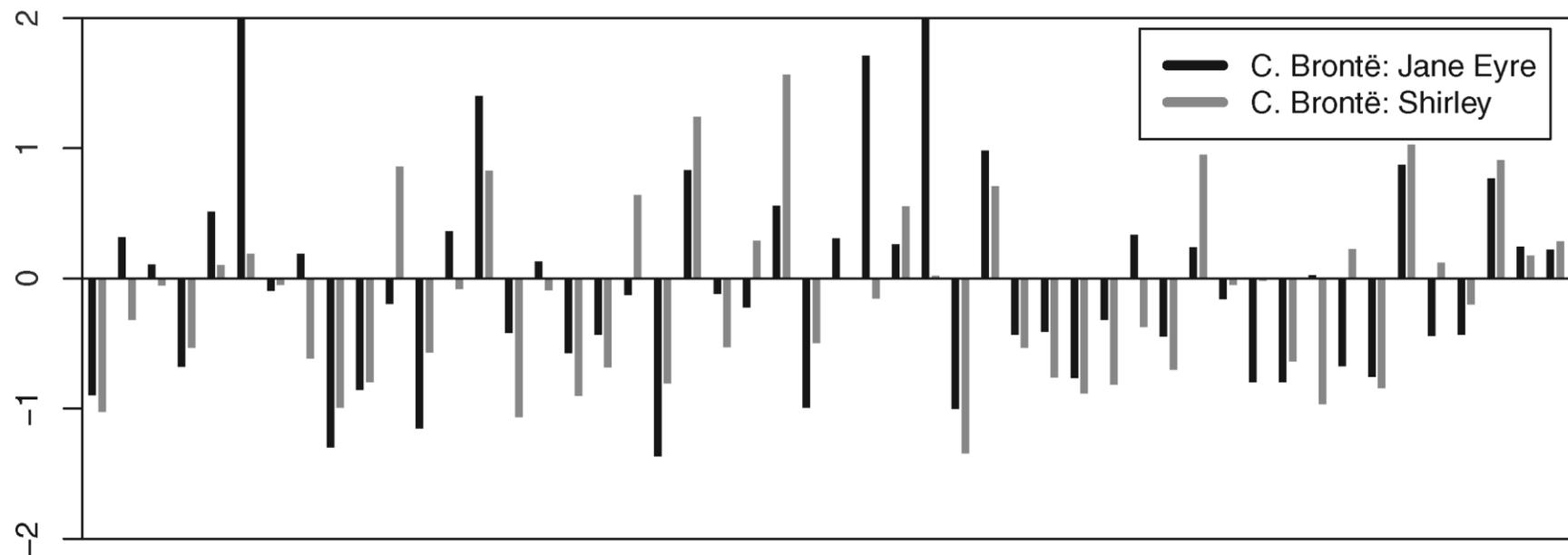
$$\sum_{i=1}^n |u_i - v_i|$$

u = z-score für Wort i in Text 1, v = z-score für Wort i in Text 2

relative frequencies



standardized z-scores



	anfang	aufstand	der	die	frißt	erhebung	hat
T1	0	0	0	1	1	0	0
T2	0	0	0	0	0	0	1
T3	1	1	3	0	0	0	0
T4	0	0	0	0	0	0	0
T5	0	0	0	0	0	0	0
Abfrage	0	0	0	0	0	1	0

...

Dokumentvektor



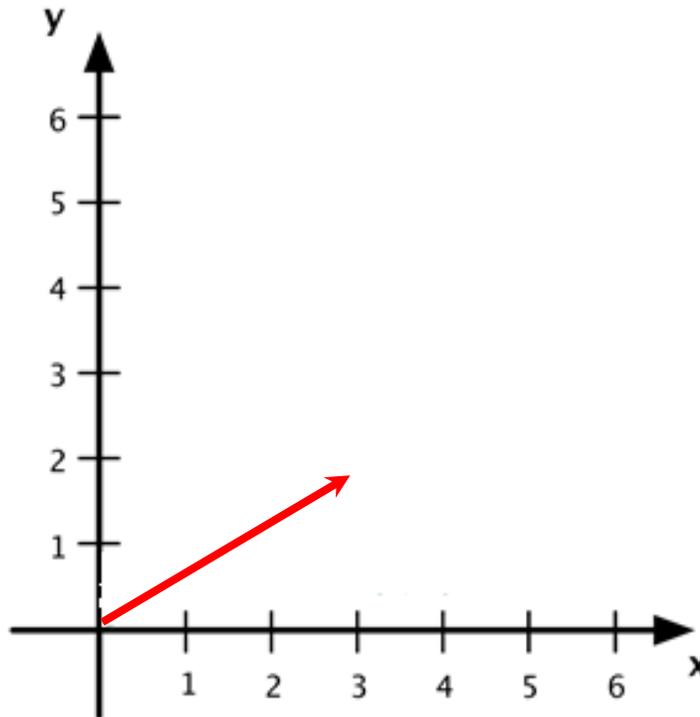
Bag of words (hier 26 Texte x 5000 Wörter)

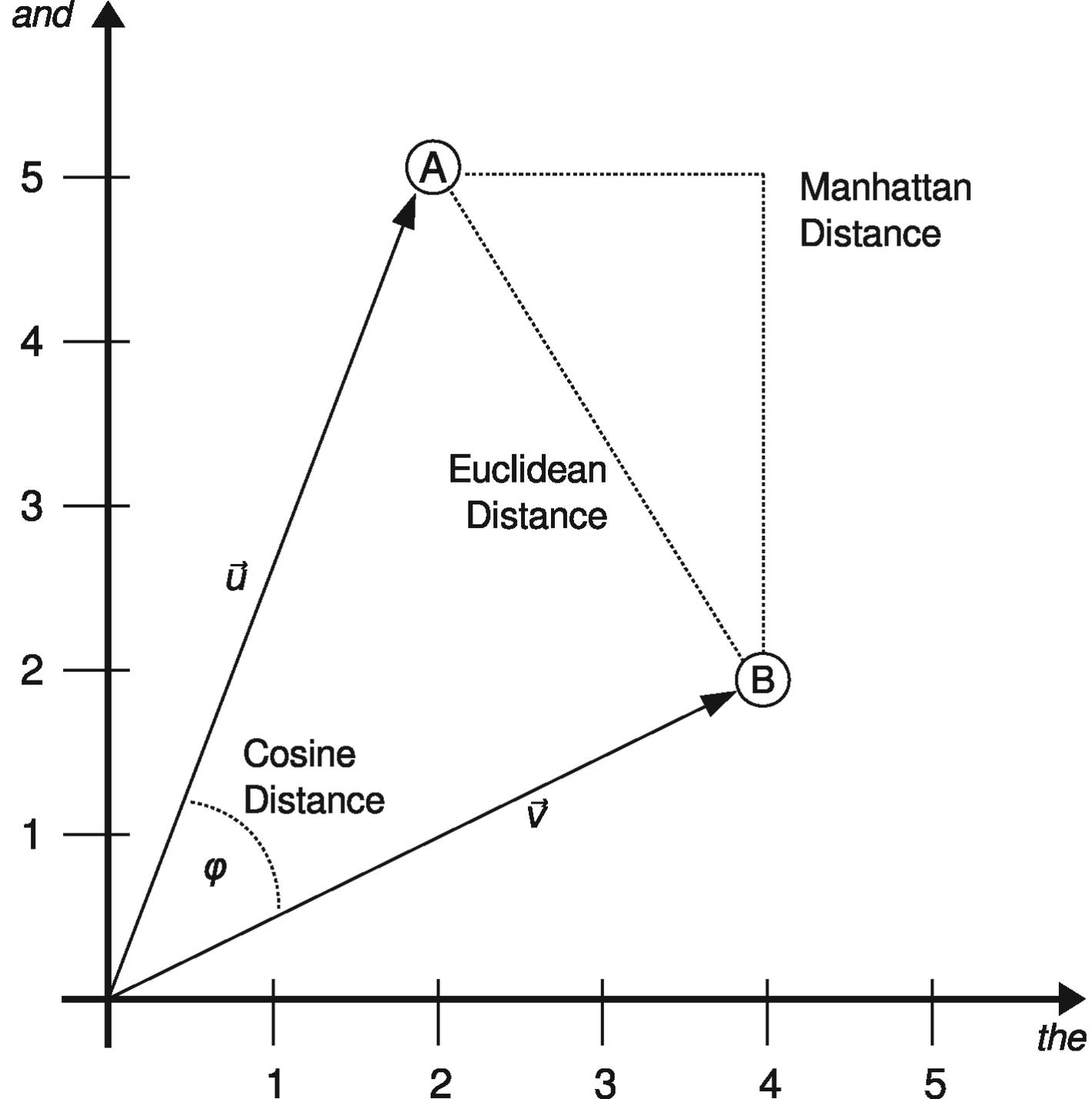
	und	die	der	zu	sie	ich
Ahlefeld-Ch_Erna	3.1348972032852	2.75377100220087	2.61957163562188	2.24381340920071	2.12750729149892	0.49206434412296207
Anonym_Schwester	4.57764782326186	2.62833008447044	2.31968810916179	1.41650422352177	1.35152696556205	1.74463937621832
Arnim-Lu_Armut	2.70232639792725	2.8506395626880403	2.76044094411923	1.33421312299096	2.0612503102469297	1.29849689147714
Bernhardi-So_Evremont	3.1855540492776604	2.9341100358374796	3.0079325664738303	1.8554062699936	1.36996720490003	1.43081486651544
Bruckbräu-Fr_Mitteilungen	2.5580601821245	2.8512934251775404	3.0986756520441	1.71034953402572	1.49922159902753	0.944744193981788
Eichendorff-Jo_Ahnung	3.88322549453262	2.6539660333494504	2.14786088280374	1.16342464485203	1.62529702817524	0.9741495478999721
Fischer-Ca_Margarethe	3.49278263618165	1.77536613633969	1.40396164787694	1.34337793699294	2.53661363396902	3.6350226530397203
Fouqué-Ca_Frau	3.2744119102235403	2.3613547429496697	1.96404743699306	1.4078172086538001	2.87110751285626	1.21441101065983
Fouqué-Ca_Magie	3.4197607359039104	2.7691721080978002	2.40932272055669	1.43463133311091	2.46413421667223	0.610075782851151
Fouqué-Ca_Resignation	2.36501884635697	2.34693213141084	2.0300528855545	1.5858431664773602	2.66670525165855	2.83382649776086
Fouqué-Ca_Spanier	3.44406661435173	2.49658499723893	1.86589938093992	1.41540965501206	2.37742319876769	1.1073328101839701
Frölich-He_Virginia	3.4959263941564798	2.46172215198764	2.2123893805309702	1.42927377440652	1.09741536732687	1.7418176710212099
Goethe-Jo_Wilhelm	3.53130342750142	2.39423050829059	2.0373687209965	2.02515672067085	1.57874026432196	1.28022470080599
Hauff-Wi_Lichtenstein	3.1337947217796005	2.48372852601454	2.67161351186371	1.34531991977232	1.23354024464686	0.978270348261073
Hauff-Wi_Mitteilungen	2.7998625018018903	2.39734761540423	2.58141778383953	1.58566470399077	1.55794329308184	2.1833383231873
Hoffmann-E_Elixiere	3.0569318500514195	2.52033280358979	3.05412732541834	1.4060016827147799	0.9385809105356641	2.58764139478358
Hoffmann-E_Lebensansichten	2.7351702024532996	2.5666455389433698	2.9305395026285397	1.3206069870623798	0.93657954587823	1.9618955296223102
Huber-Th_Ellen	2.23087339201083	2.23200180546152	1.8077183480027101	2.41141954412097	1.88332204919883	2.7725118483412303
Keller-Go_Martin	3.63371325354657	2.96718038639893	2.2605076635435	1.46827366931097	1.80734981884249	1.02356632054844
Manzoni-Al_Verlobten	3.2101767420916403	2.9044241810199303	2.5527411910390603	2.00833051896882	1.6323314195654601	0.9312169312169309
Pichler-Ka_Agathocles	3.38429031297418	2.39800331834877	2.0314250464427	1.85274472822156	1.05790093169023	2.4781258419955496
Schopenhauer-Jo_Gabriele	2.6479030924748304	1.99373824293274	2.1161093789664798	2.15711672774375	2.4975428136248503	1.1742421776854899
Schopenhauer-Jo_Tante	2.59040716891511	2.11704984126207	1.98810077614279	2.10072717479127	1.65838291343274	1.4706722490186
Schröder-Devrient-Wi_Aus	3.09275874079679	2.44538128010185	1.8928884565032	1.5746045472561498	1.95414309564131	2.95943982031973
Spindler-Ca_Jude	3.66738000144956	2.52947751088122	2.96892225413791	1.63952561329626	0.739274229541215	1.5403453761029
Zschokke-He_Goldmachedorf	5.524901129609019	2.86072891751723	2.3042129443247203	1.2148625287138601	1.3332701825846	0.45231723778625

Zur Erinnerung

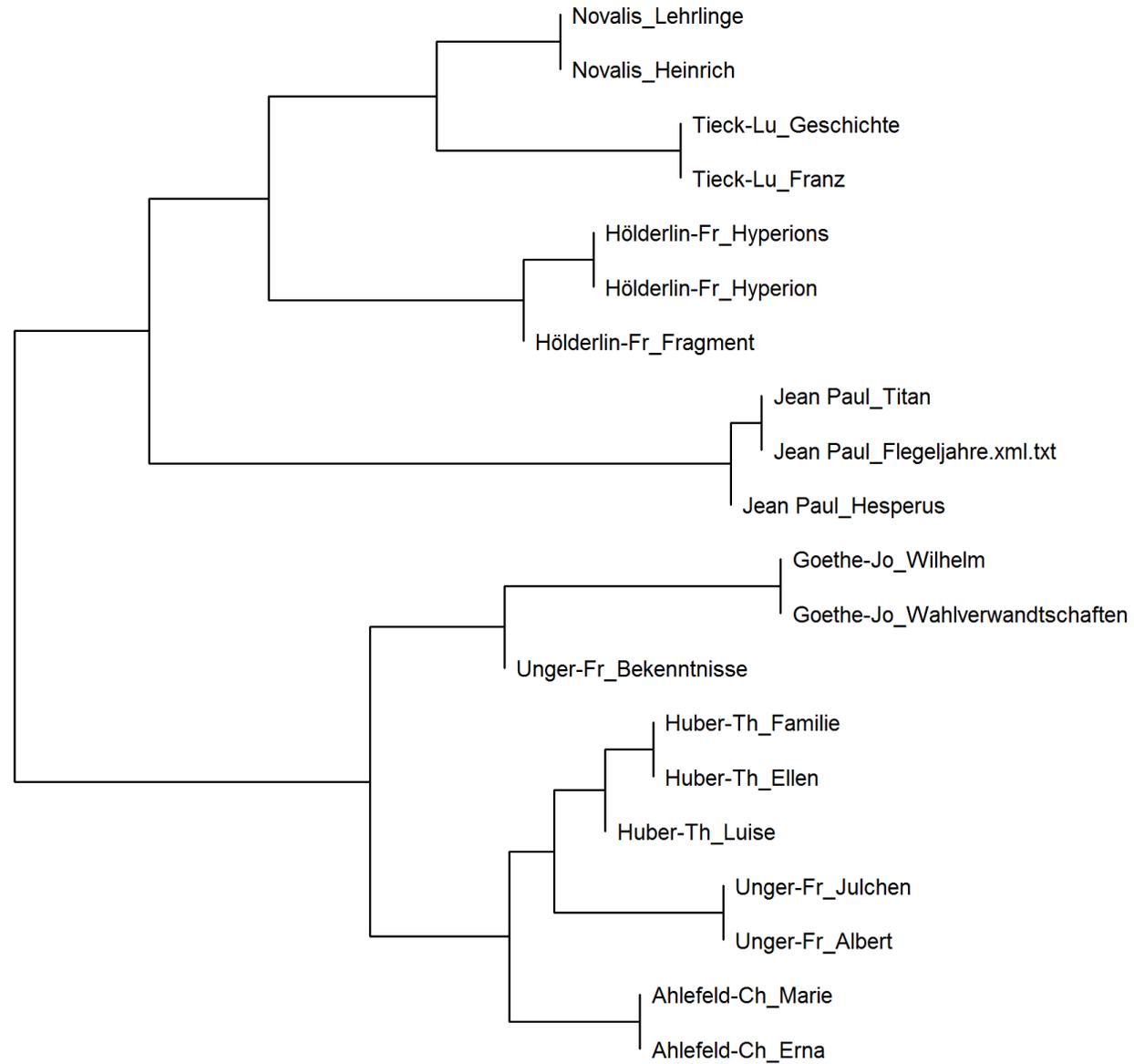
- Ein Vektor, also eine Folge von n Zahlen, lässt sich geometrisch interpretieren, nämlich als die Punktkoordinaten in einem n -dimensionalen Raum.

- Z.B. $(3, 2)$

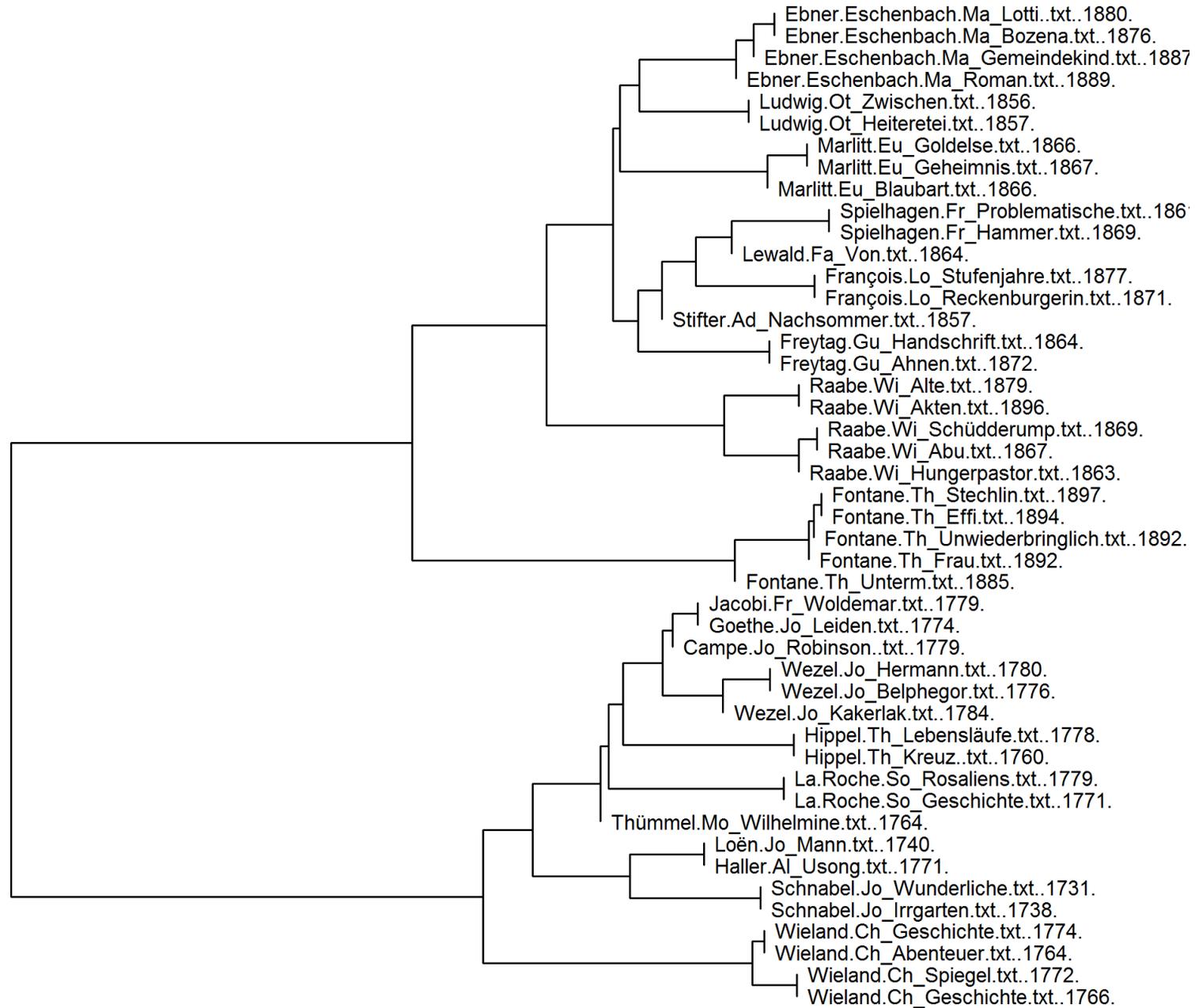




**Authors around 1800
Cluster Analysis**



AufklaerungundRealismus Cluster Analysis



Figurenkonstellationen

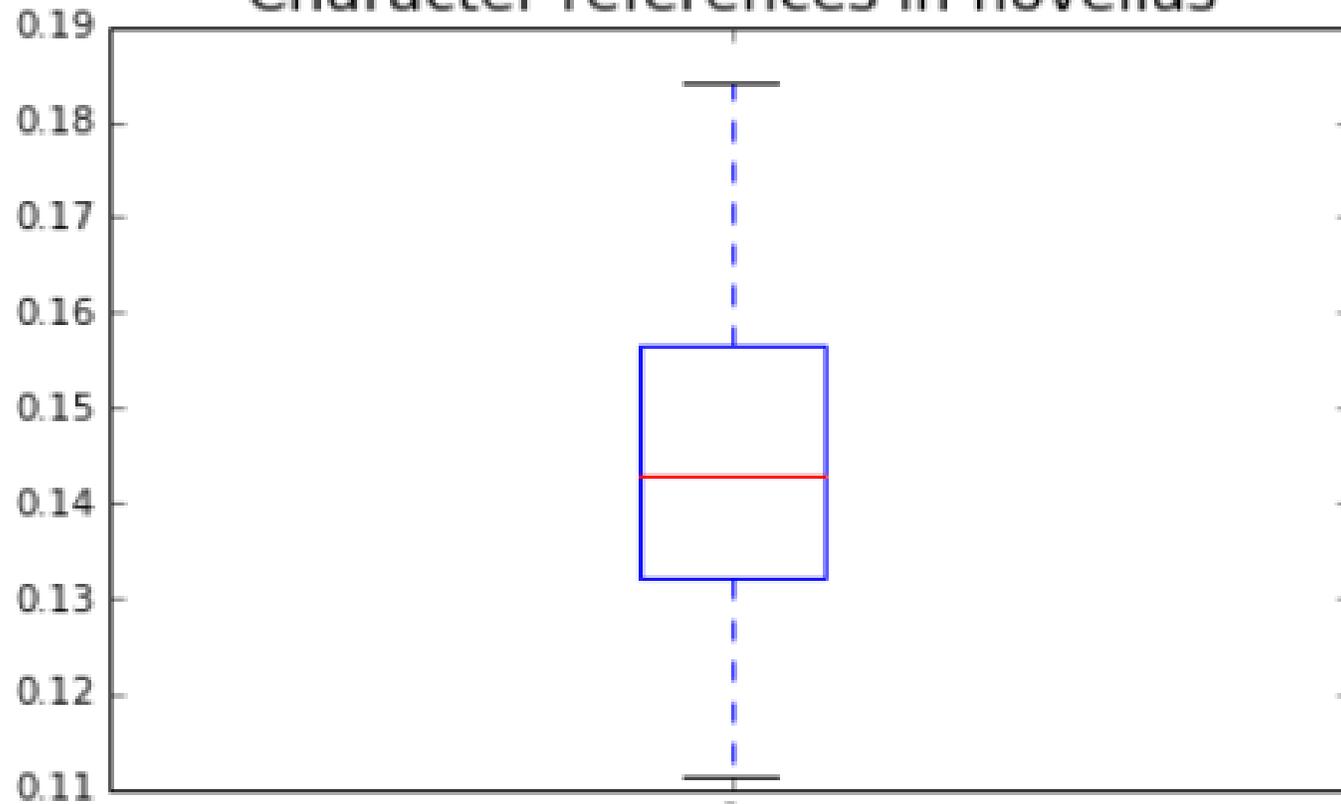
In der Einleitung zum *Novellenschatz* schreiben Paul Heyse und Hermann Kurz, dass sich Novellen von Romanen dadurch unterscheiden:

„Wenn der Roman ein Kultur- und Gesellschaftsbild im Großen, ein Weltbild im Kleinen entfaltet, bei dem es auf ein gruppenweises Ineinandergreifen oder ein concentrisches Sichumschlingen verschiedener Lebenskreise recht eigentlich abgesehen ist, so hat die Novelle in einem einzigen Kreise einen einzelnen Conflict, eine sittliche oder Schicksals-Idee oder ein entschieden abgegrenztes Charakterbild darzustellen und die Beziehungen der darin handelnden Menschen zu dem großen Ganzen des Weltlebens nur in andeutender Abbrüviatur durchschimmern zu lassen.“

Figurenreferenzen im ‚Novellenschatz‘

ParagraphId	SentencelId	TokenId	Begin	End	Token	Lemma	CPOS	POS	Morphology	DependencyHead	DependencyRelation	NamedEntity	CorefId
0	0	0	0	4	?Der	?der	ADJA	ADJA		Stern	NK	0	-
0	0	1	5	10	Stern	Stern	N	NN	masc sg	war	SB	0	-
0	0	2	11	14	der	der	ART	ART	masc sg	Schönheit	NK	0	-
0	0	3	15	24	Schönheit	Schönheit	N	NN	masc sg	Stern	AG	B-PER_CORE	39
1	0	4	26	32	August	August	N	NE	masc sg	Wolf	PNC	I-PER_CORE	39
1	0	5	33	37	Wolf	Wolf	N	NE	masc sg	Schönheit	NK	I-PER_CORE	39

Character references in novellas



war	MO	0	-
ROOT	--	0	-
Gesellschaft	NK	0	-
war	SB	0	-
Gesellschaft	MNR	0	-
bei	NK	0	-
Madame	UC	0	-
ROOT	NK	0	-
gewesen	--	0	-
waren	SB	0	-
ROOT	--	0	-
viele	MO	0	-
zusammengekommer	MO	0	-
viele	AG	B-PER_APP	40
zusammengekommer	MO	0	-
worden	OC	0	-
zusammengekommer	--	0	-
viel	MO	0	-
gesprochen	MO	0	-
zusammengekommer	CJ	0	-
gesprochen	--	0	-
gesprochen	CJ	0	-
gelacht	--	0	-
gelacht	CJ	0	-
getanzt	CD	0	-
viel	MO	0	-

Interaktion modellieren

- Figuren werden im gleichen Absatz erwähnt

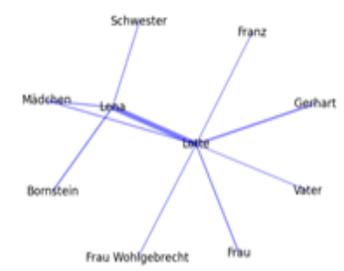
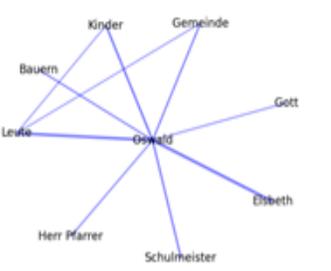
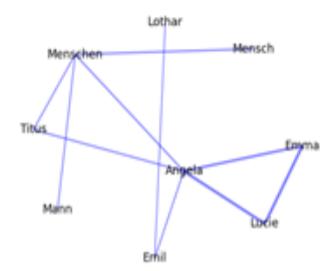
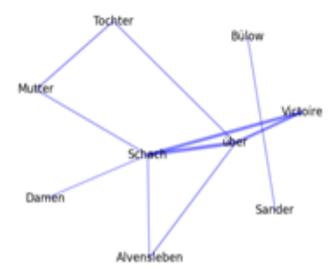
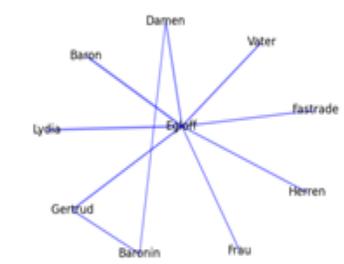
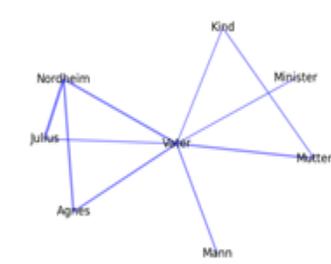
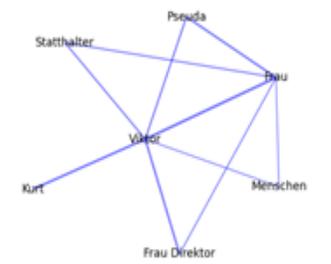
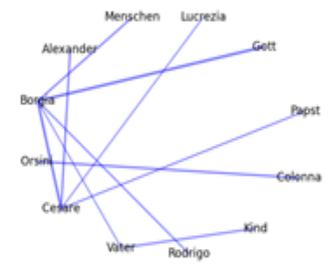
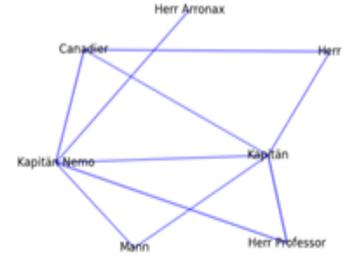
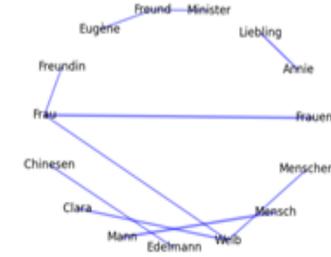
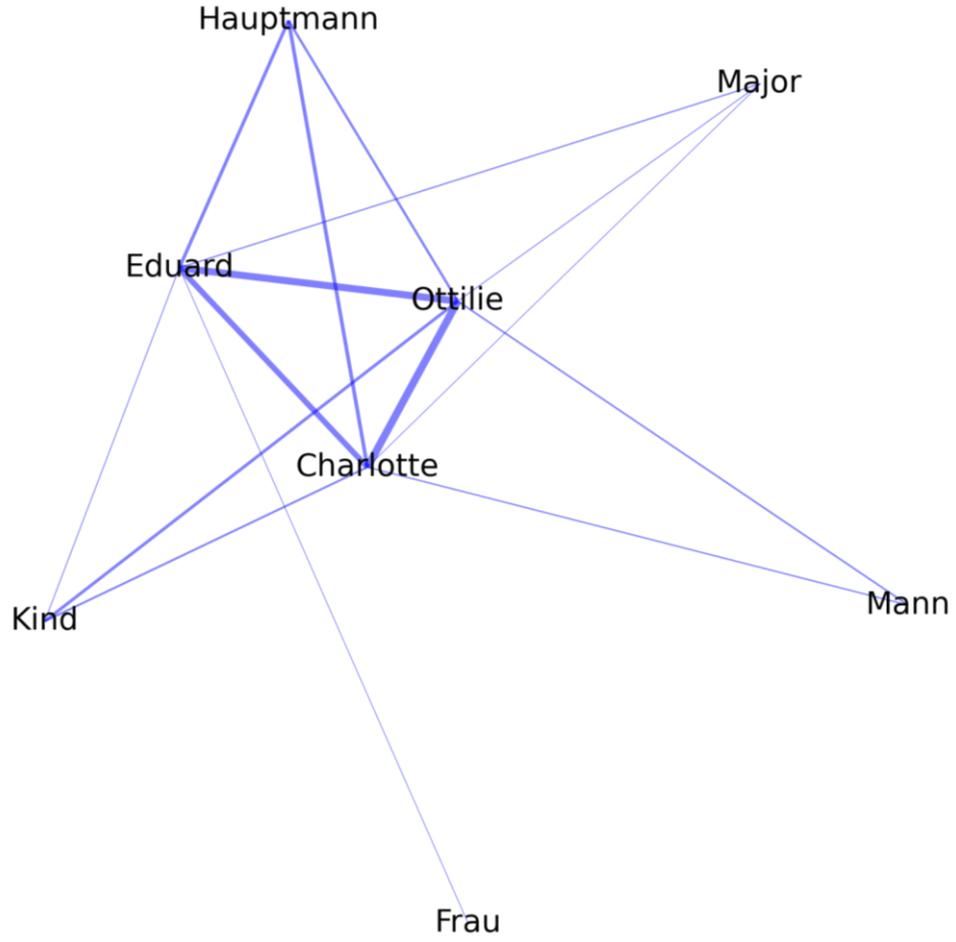
„Die Baronin winkte ihrem Mann zum Abschied zu und wandte sich zu ihrer Freundin Wanda.“

Folgende Interaktionen:

Baronin - Ehemann

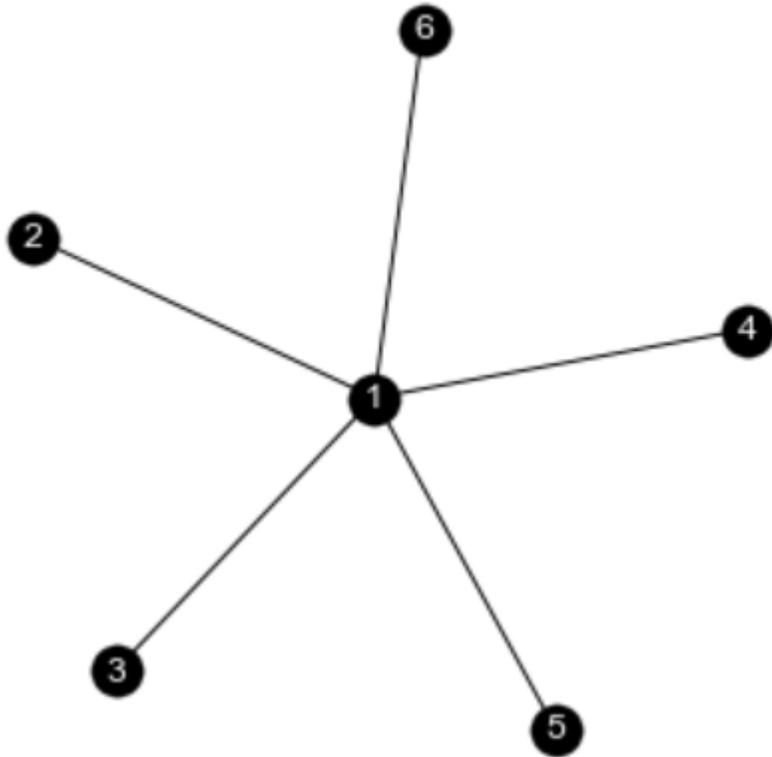
Baronin – Wanda

Ehemann – Wanda

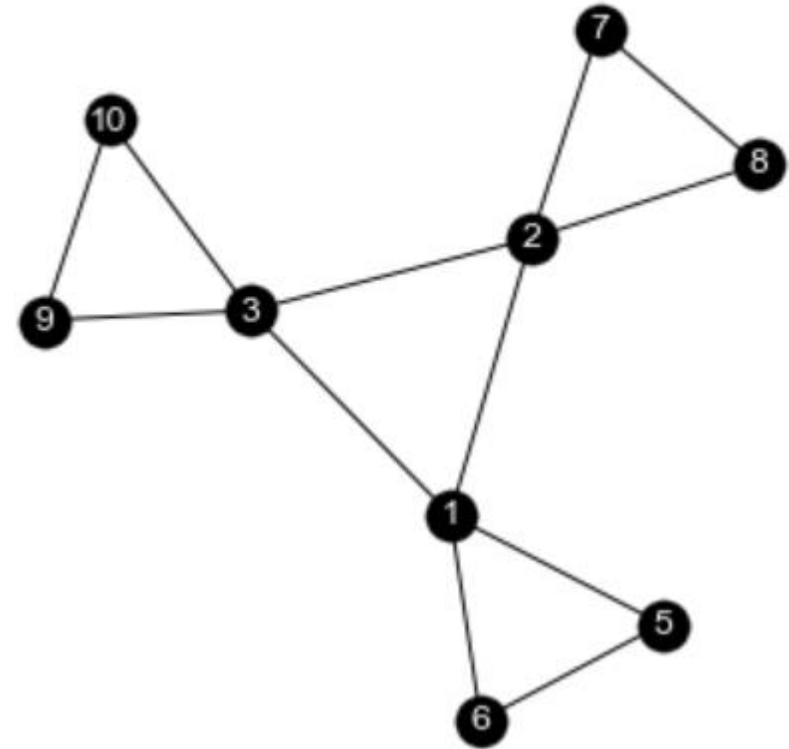


Netzwerktypen und Netzwerkmaße

Star network (GDC: 1.0)

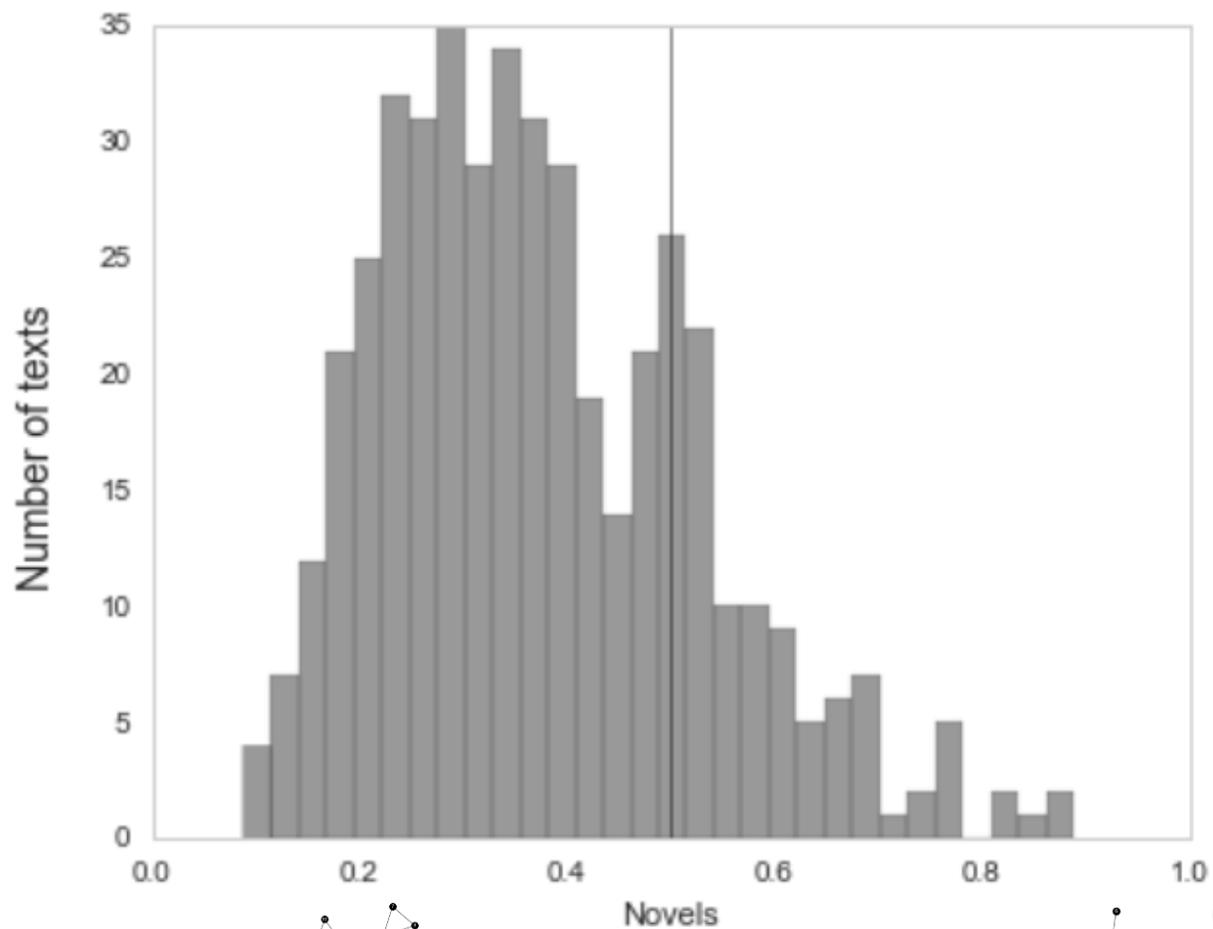


Network of groups (GDC: 0.21)

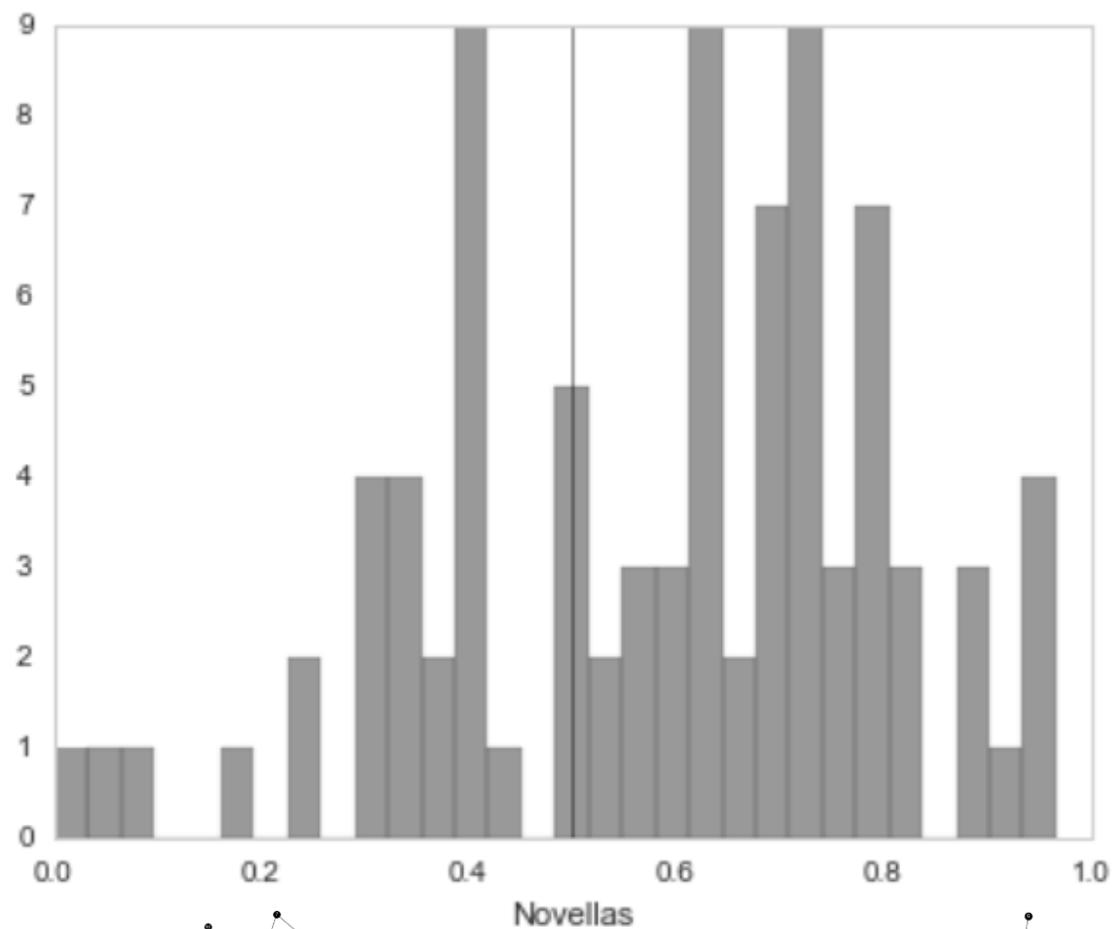


GDC: Group degree centrality

GDC of novels and novellas



GDC

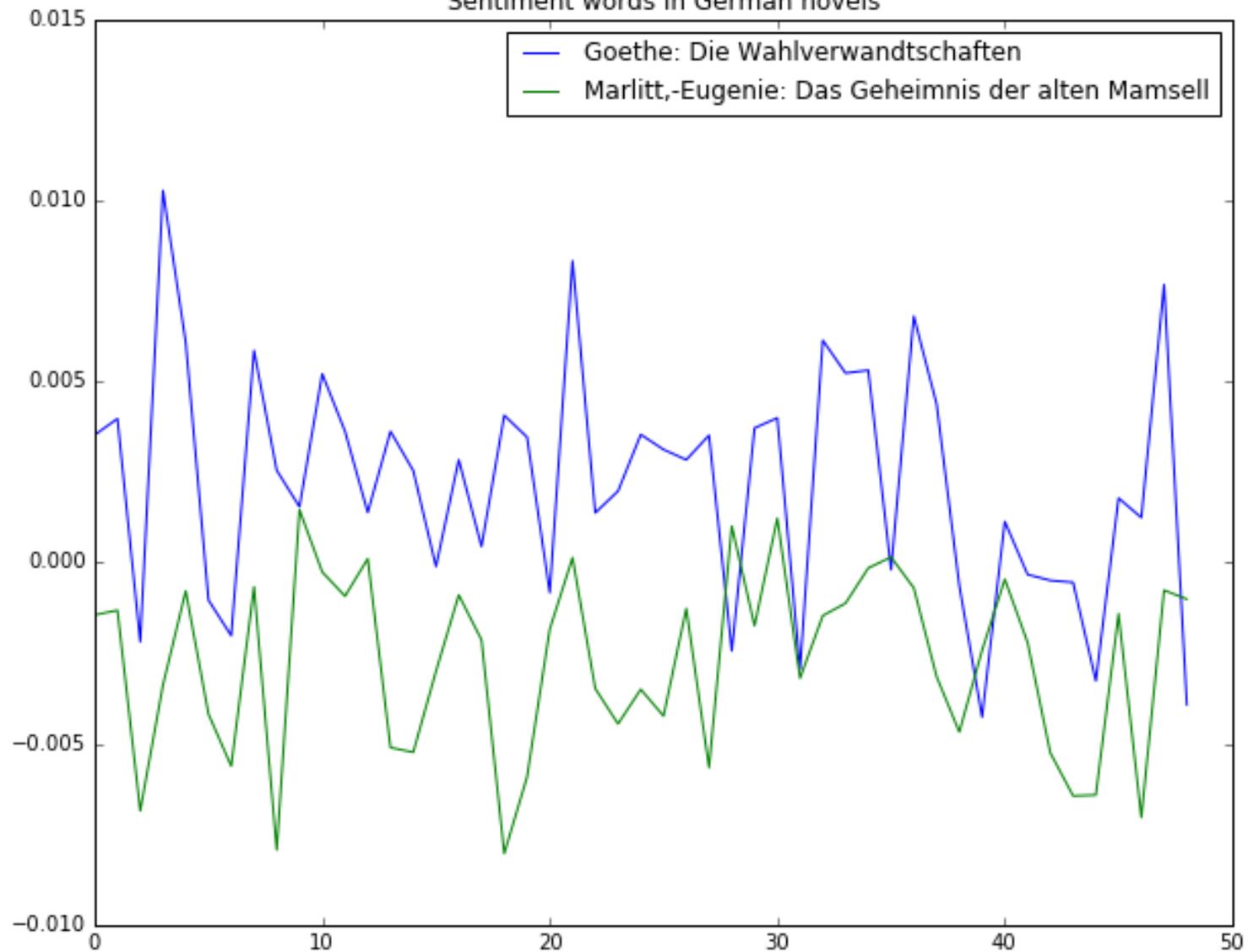


Resultat

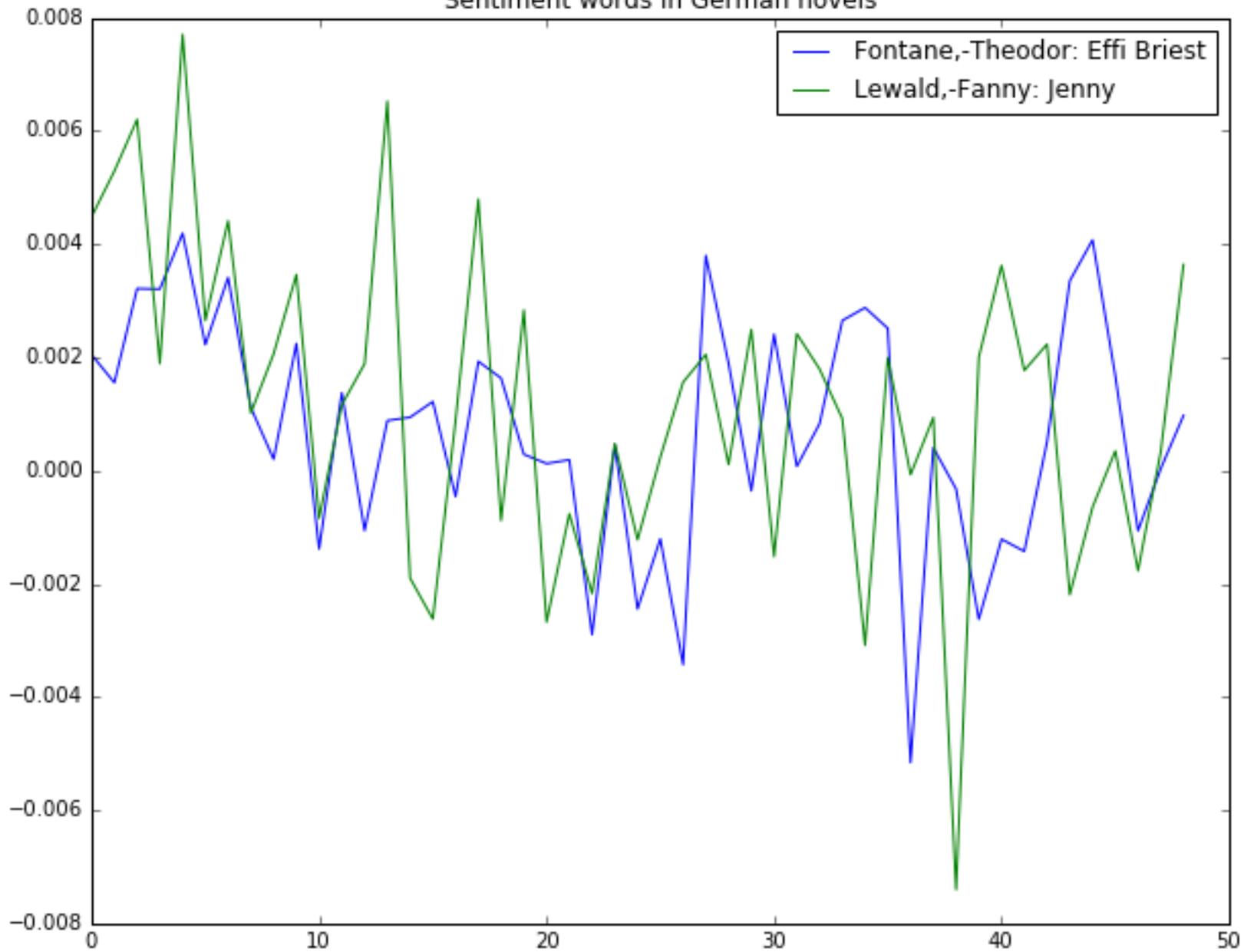
- Die Figurenkonstellation in den Novellen unterscheidet sich von der F. in den Romanen (hochsignifikant). Das bestätigt die Beschreibung der Herausgeber.
- Noch unklar: Wie sieht die Figurenkonstellation in anderen kürzeren Erzähltexten der Zeit aus? Ist das beobachtete Phänomen
 - ein Effekt der Gattung Novelle oder
 - ein Effekt der Auswahl der Herausgeber oder
 - ein Effekt der Länge der Erzählungen

Sentiment Analysis

Sentiment words in German novels



Sentiment words in German novels



Happy ends?

Topic Modeling

Topic Models

Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

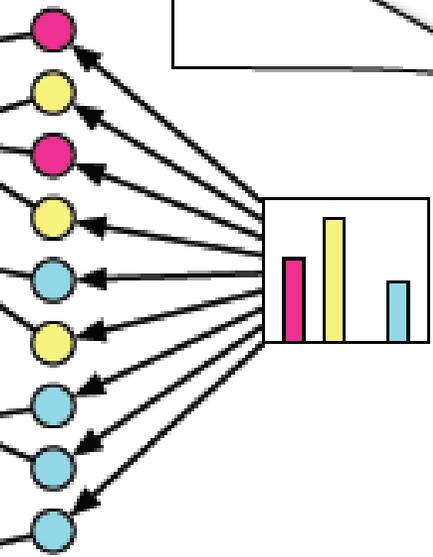
"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson, a research University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic** numbers game, particularly as more and more **genomes** are sequenced, mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



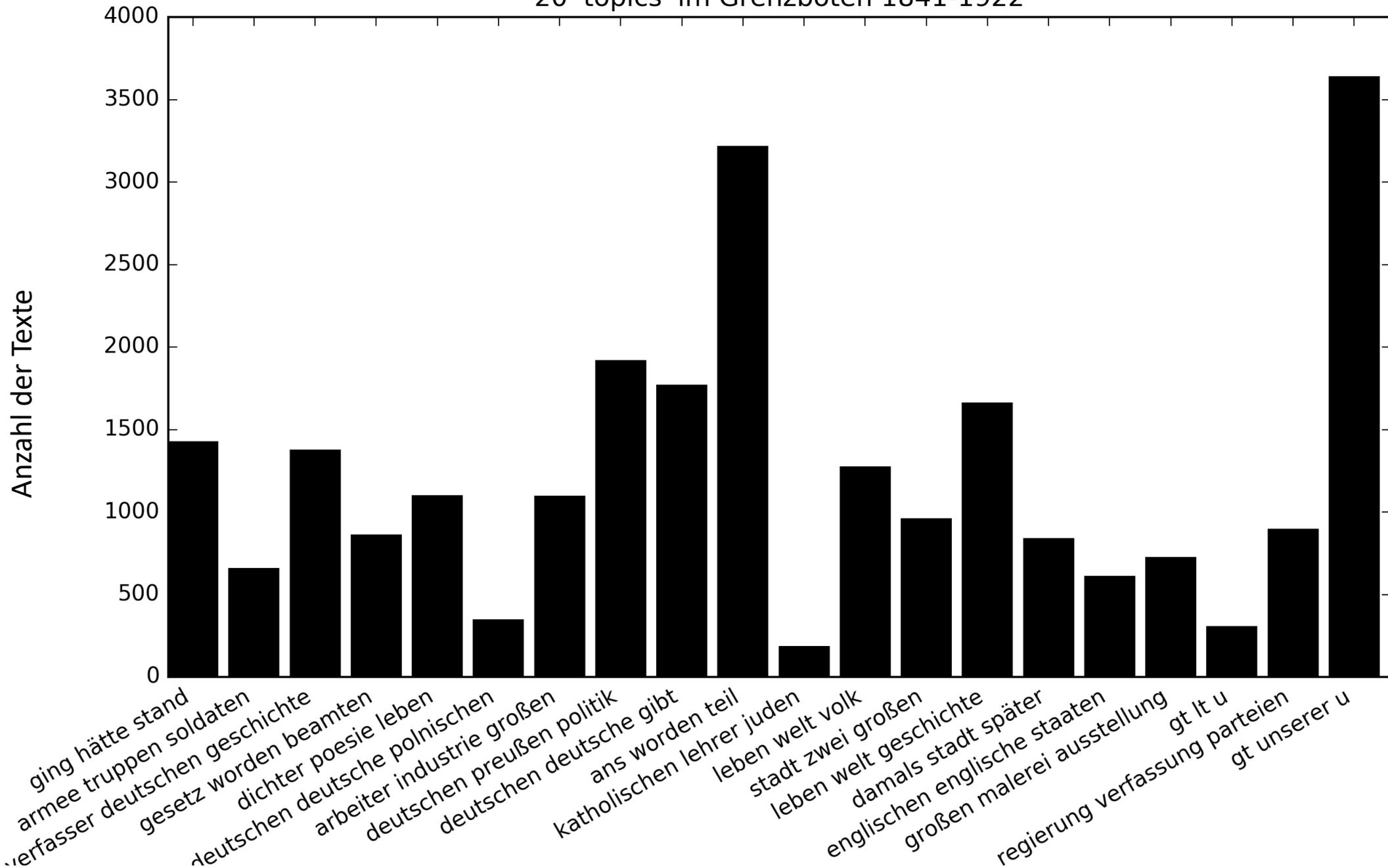
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

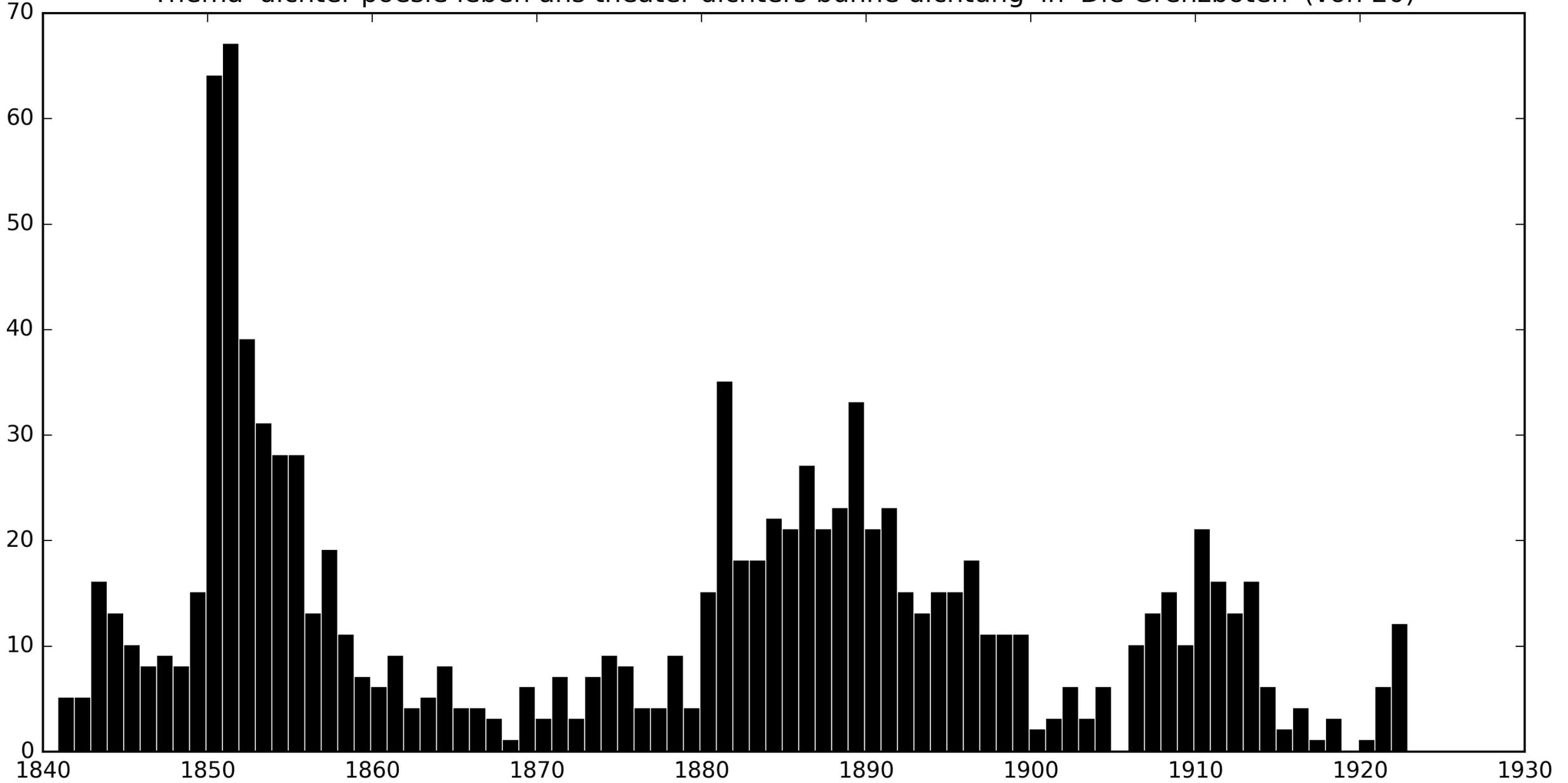
Topic proportions and assignments



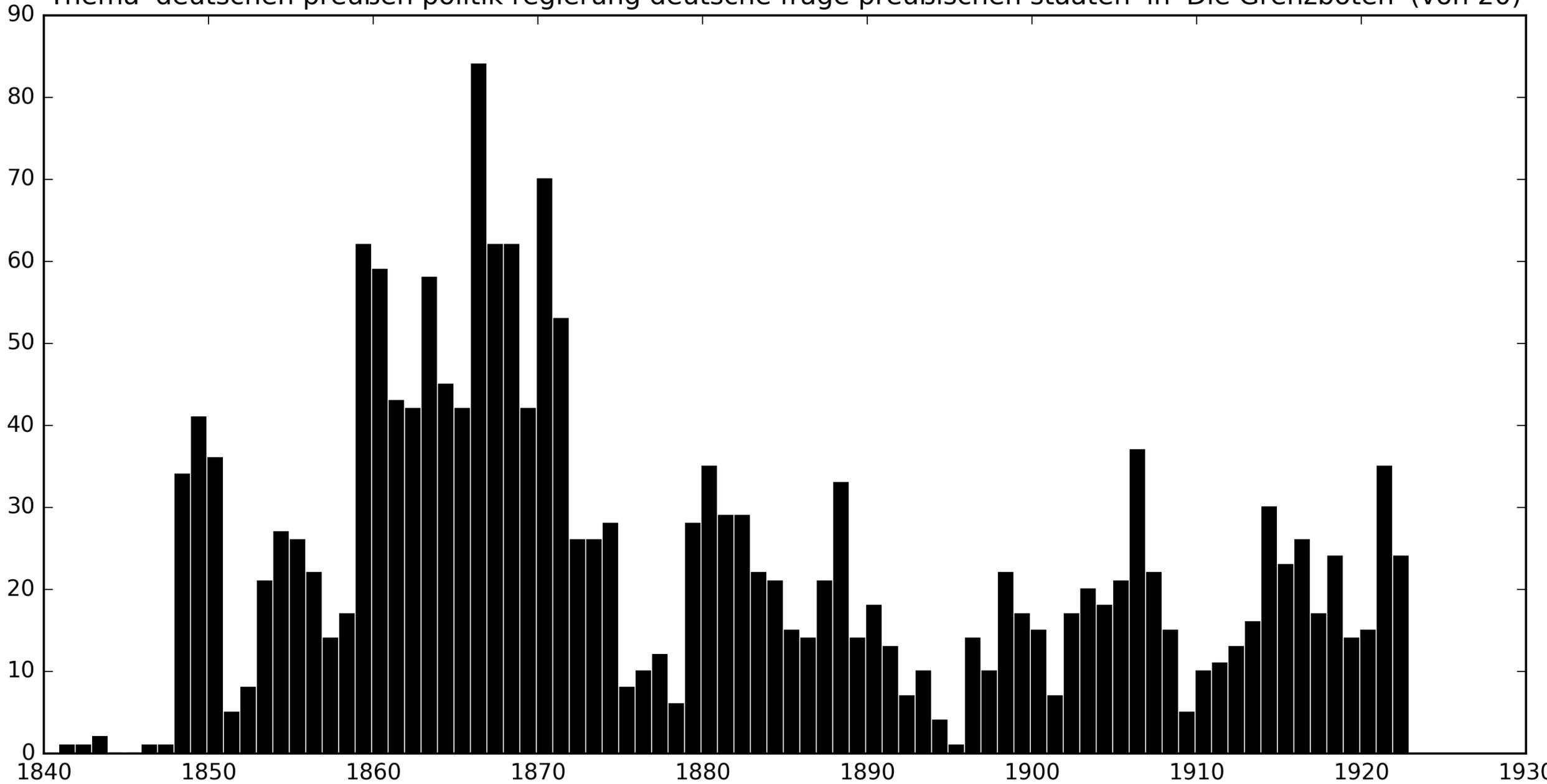
20 'topics' im Grenzboten 1841-1922



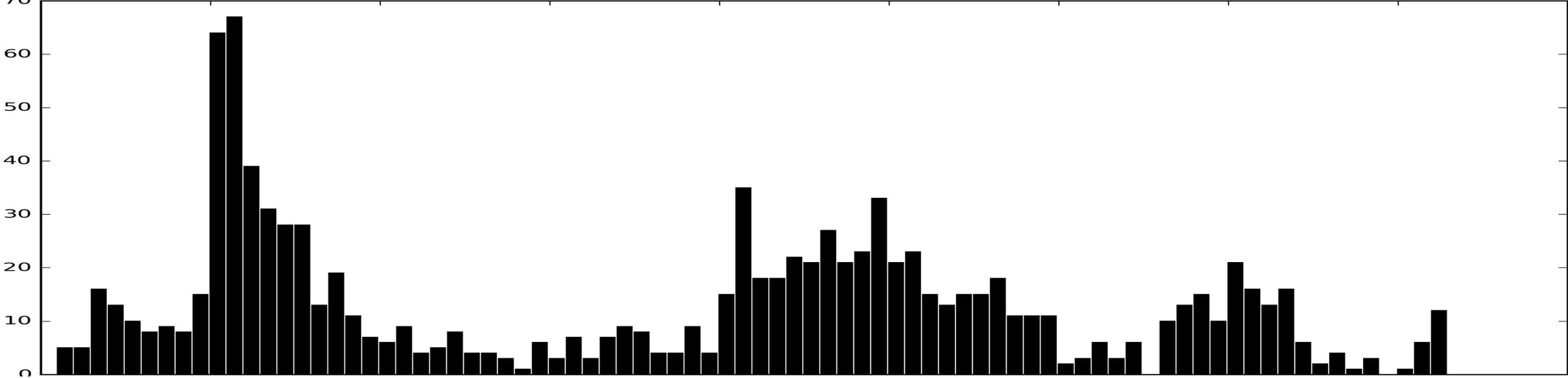
Thema 'dichter poesie leben ans theater dichters bühne dichtung' in 'Die Grenzboten' (von 20)



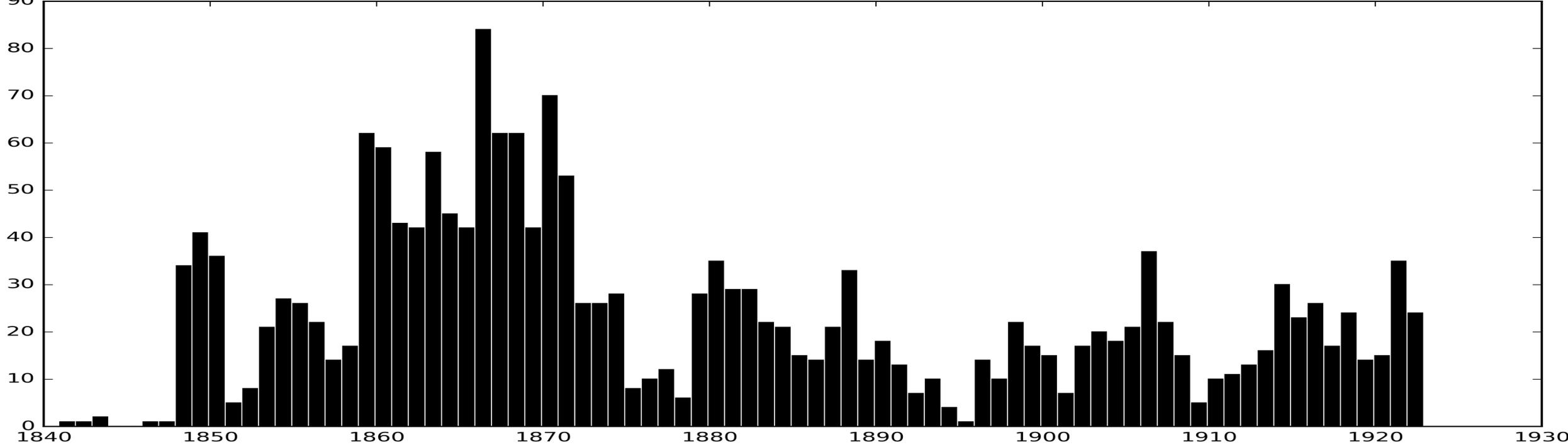
Thema 'deutschen preußen politik regierung deutsche frage preußischen staaten' in 'Die Grenzboten' (von 20)



Thema 'dichter poesie leben ans theater dichters bühne dichtung' in 'Die Grenzboten' (von 20)

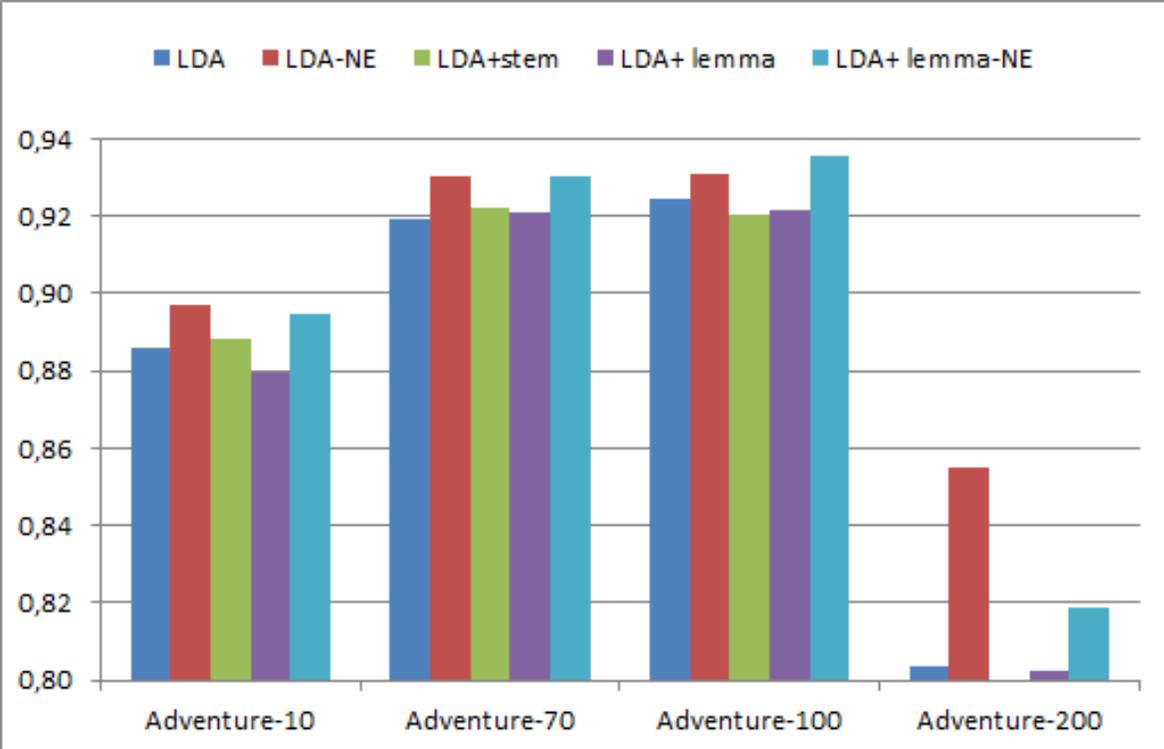


Thema 'deutschen preußen politik regierung deutsche frage preußischen staaten' in 'Die Grenzboten' (von 20)

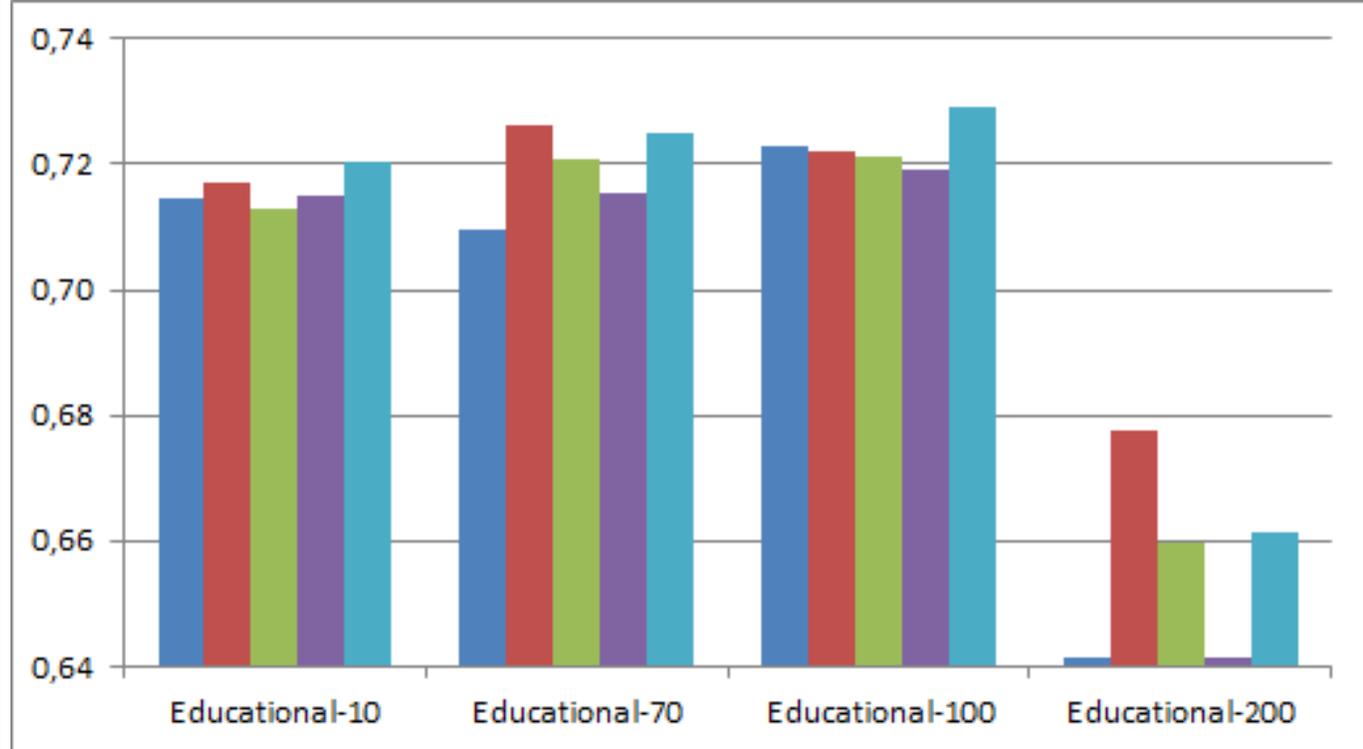


Topic models for novels

- wasser felsen ufer berg berge wiesen
- pferd pferde ritt reiter reiten sattel ritter
- papier hand lesen buch las brief schreiben blatt geschrieben feder
- liebe lieben herz hätte glück freundschaft leidenschaft herzen
- trug haar schwarzen rock weißen kleid hut kleider mantel
- augen gesicht haar stirn lippen mund gestalt sah mann nase wangen
- schiff see meer schiffe wasser ufer insel boot land hafen bord wind
- gott menschen himmel heiligen welt willen seele glauben sünde
- leben welt lebens liebe glück seele herz zukunft allein nie jugend
sehnsucht tage
- gefahr genug allein retten sicher mittel bringen unglück sicherheit
flucht furcht suchen
- zeit arbeit arbeiten lernen kennen art mann kleinen lehrer erziehung
jahre großen
- garten baum blumen bäume bäumen grünen erde kleinen natur
schönen früchte

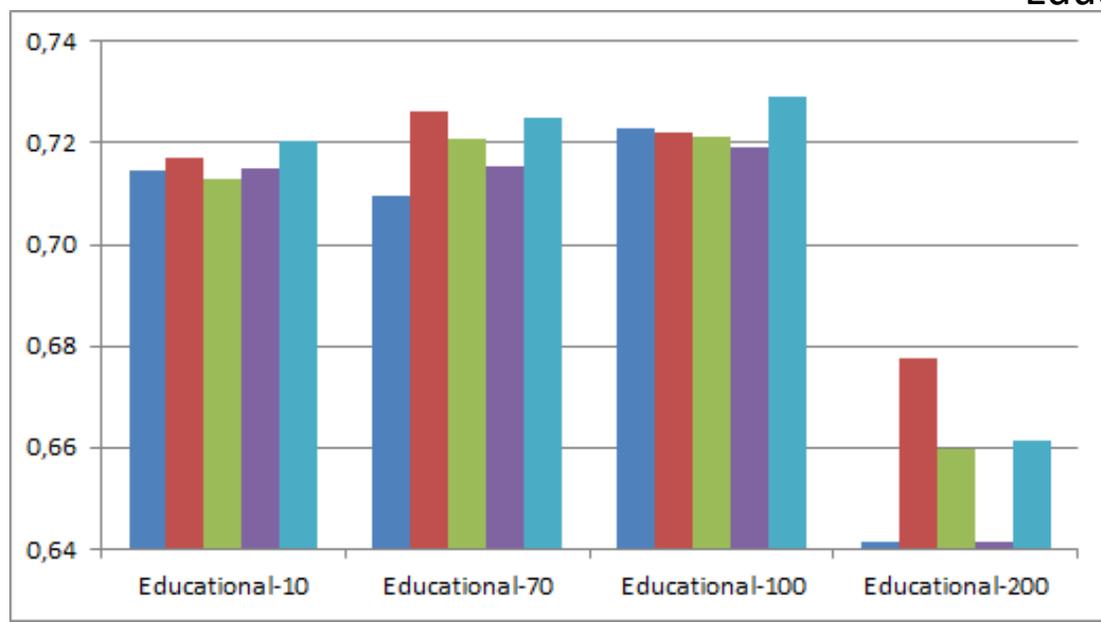


Adventure / non-adventure

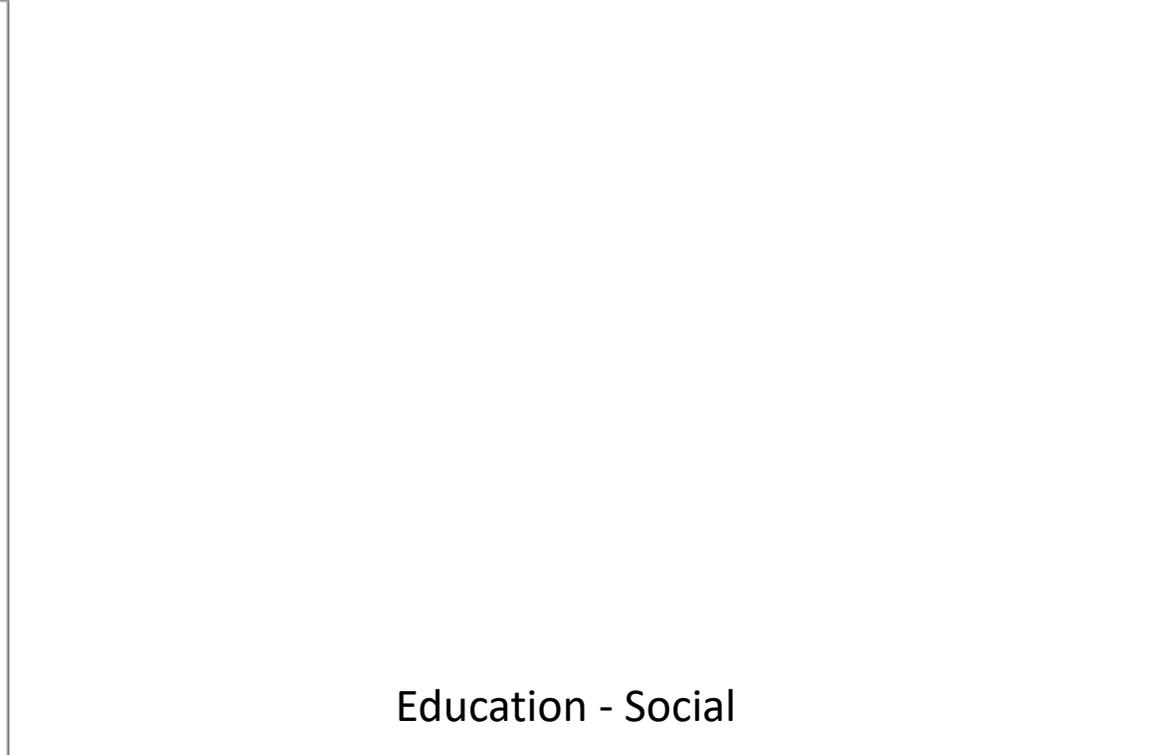
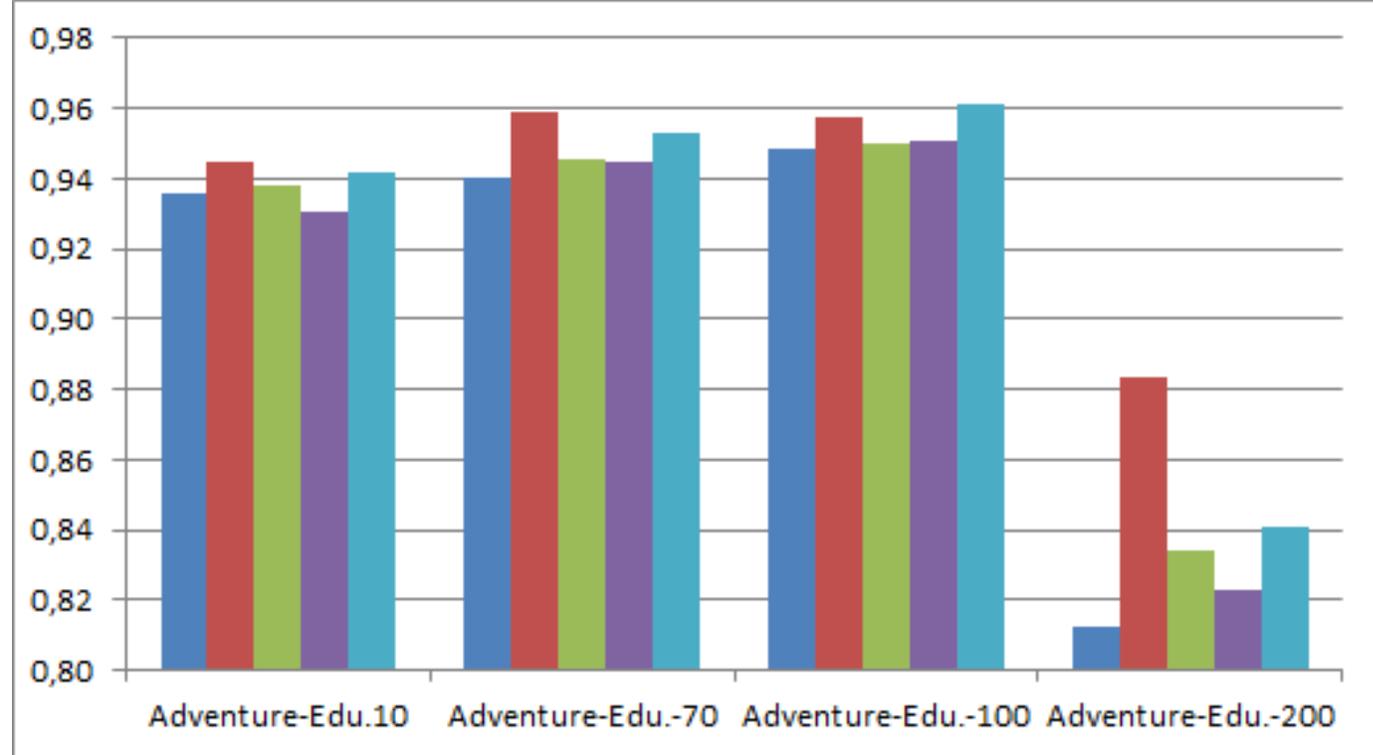


Educational / non-educational

10, 70, 100, 200 =
n topics

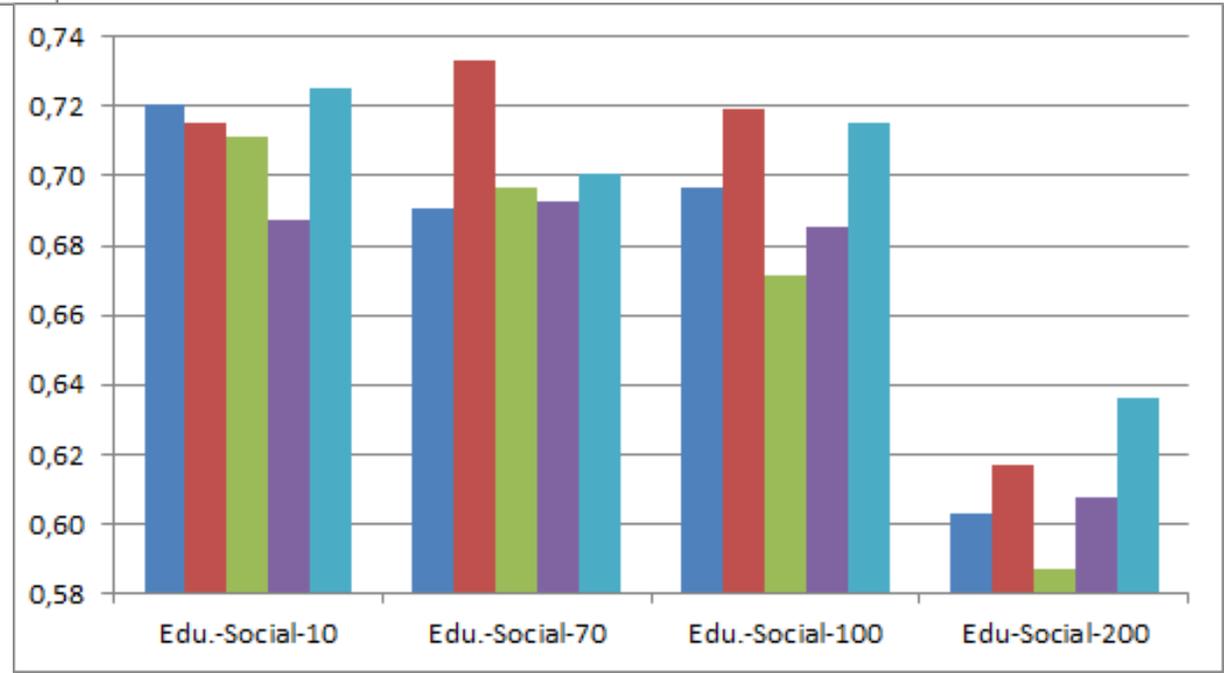


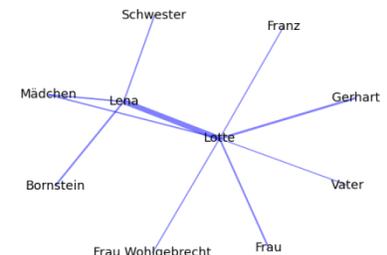
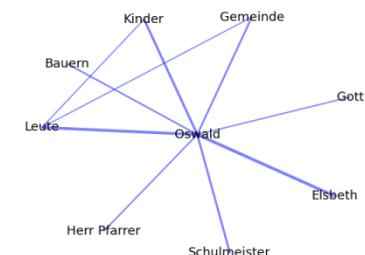
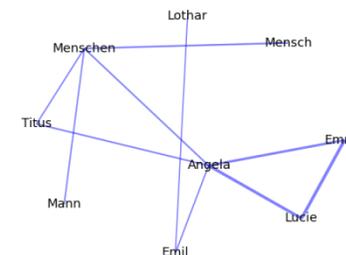
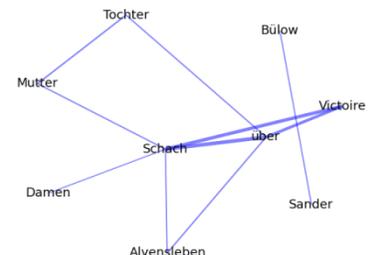
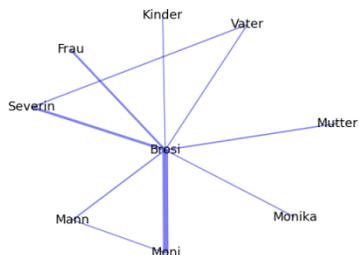
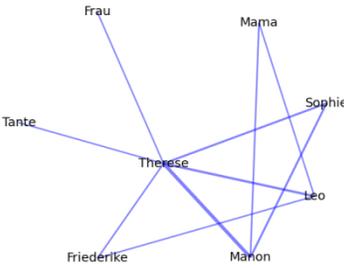
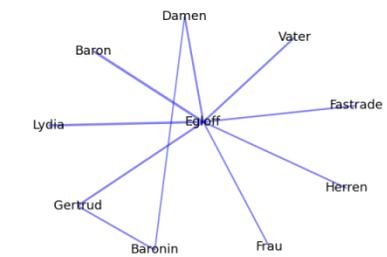
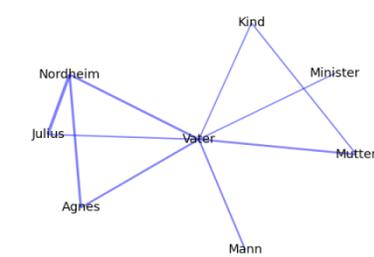
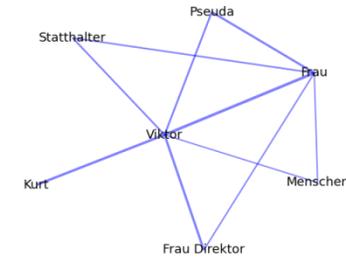
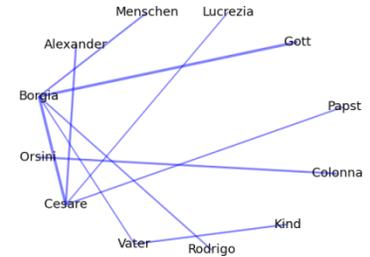
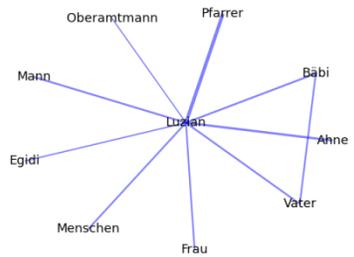
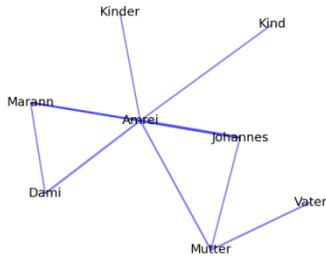
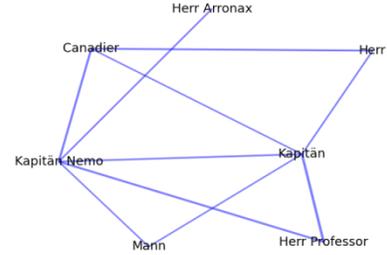
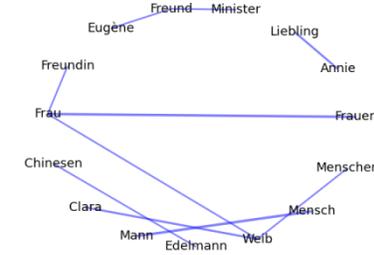
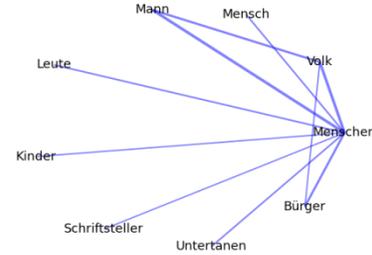
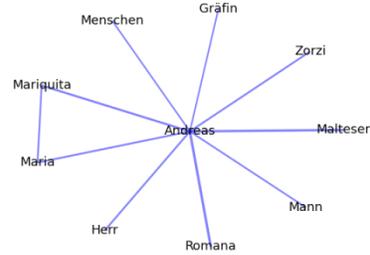
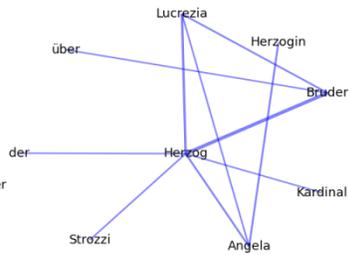
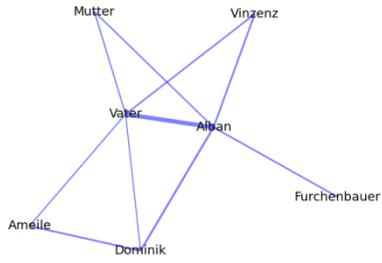
Social / non-social



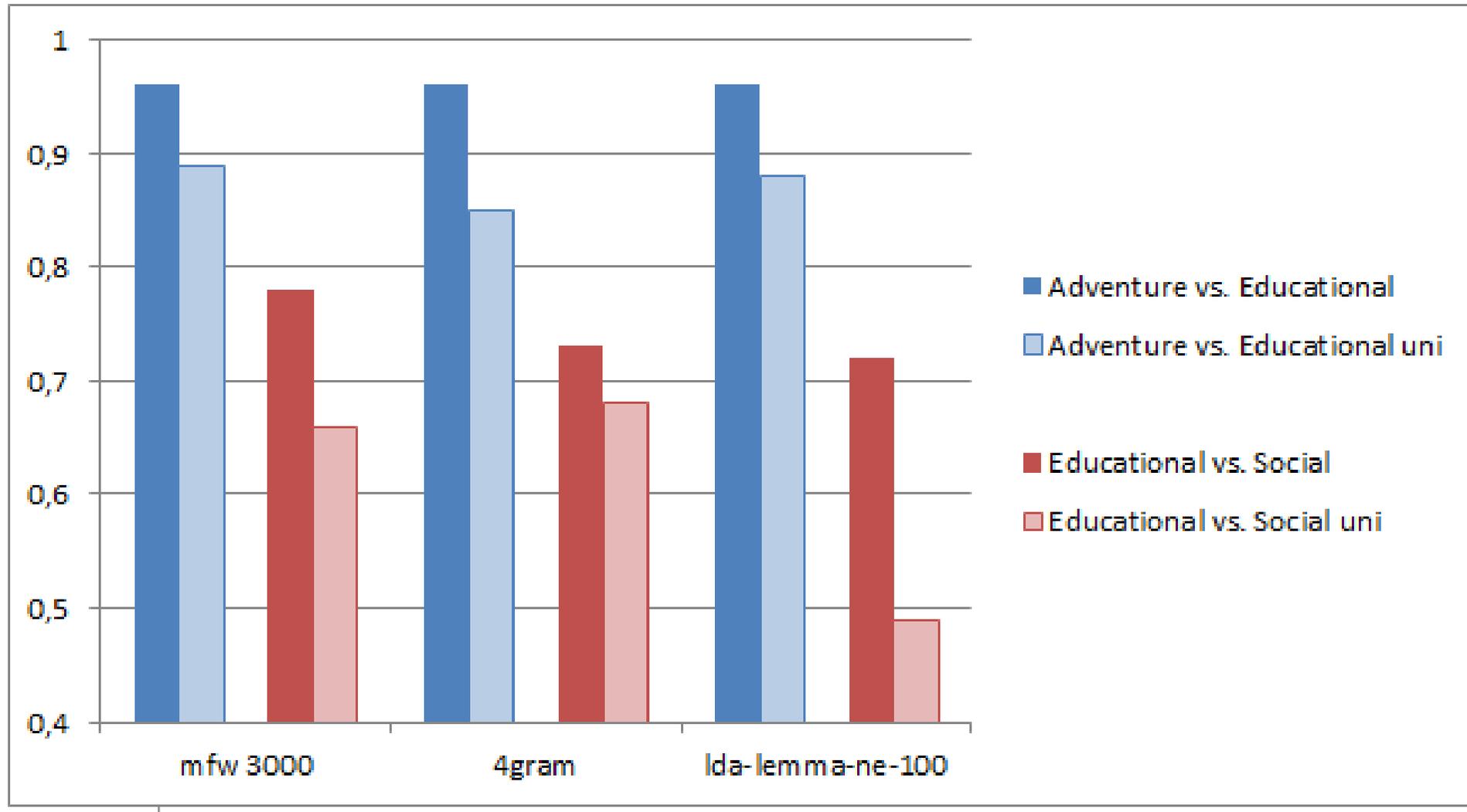
Adventure - Education

Education - Social





How well-defined is a genre? And what are its features?



Das Problem

Die meisten Methoden setzen längere Texte voraus

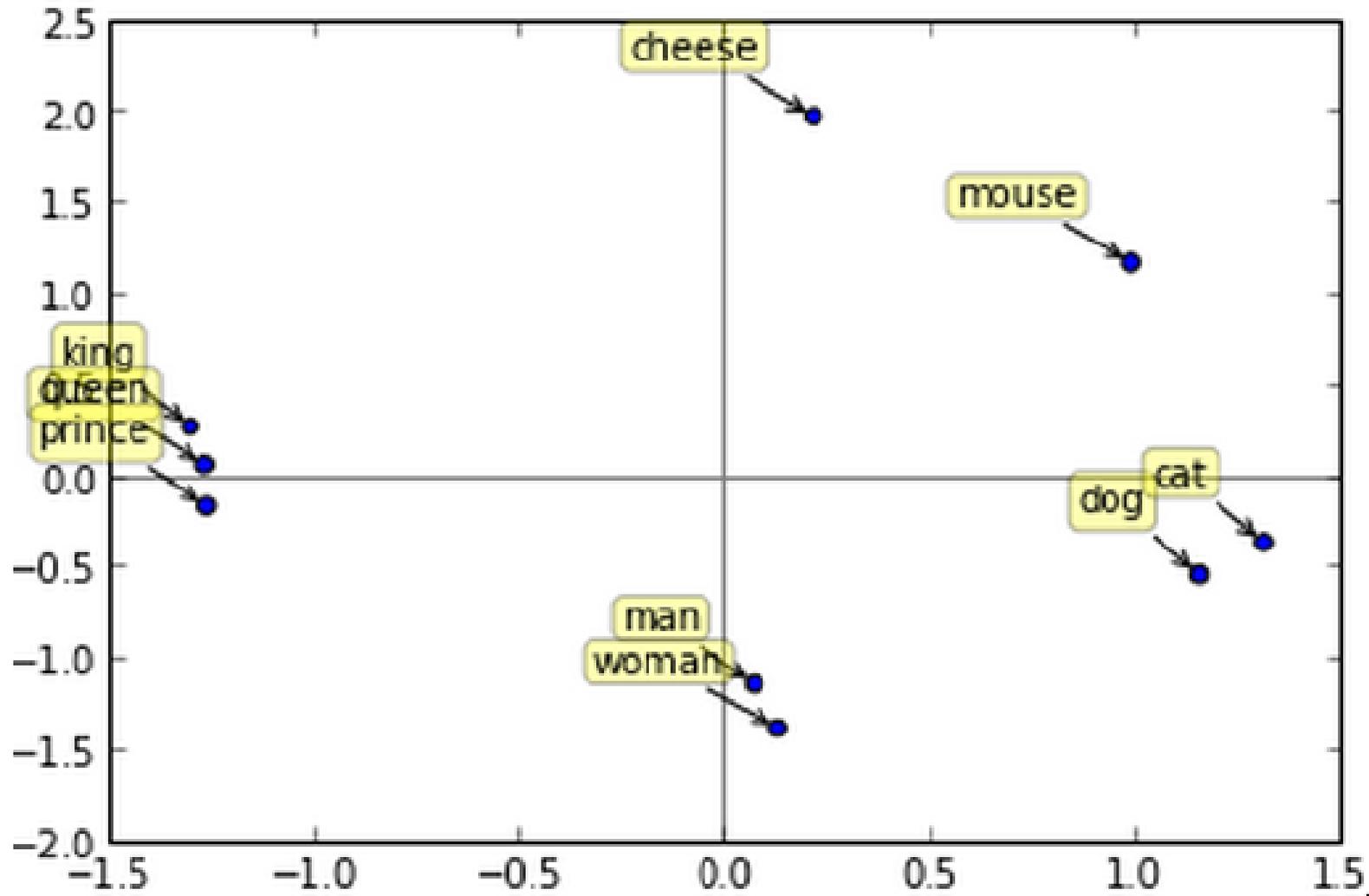
Verteilung der Einzelwörter über die Texte

Kookurrenz der Wörter

Distributionelle Semantik

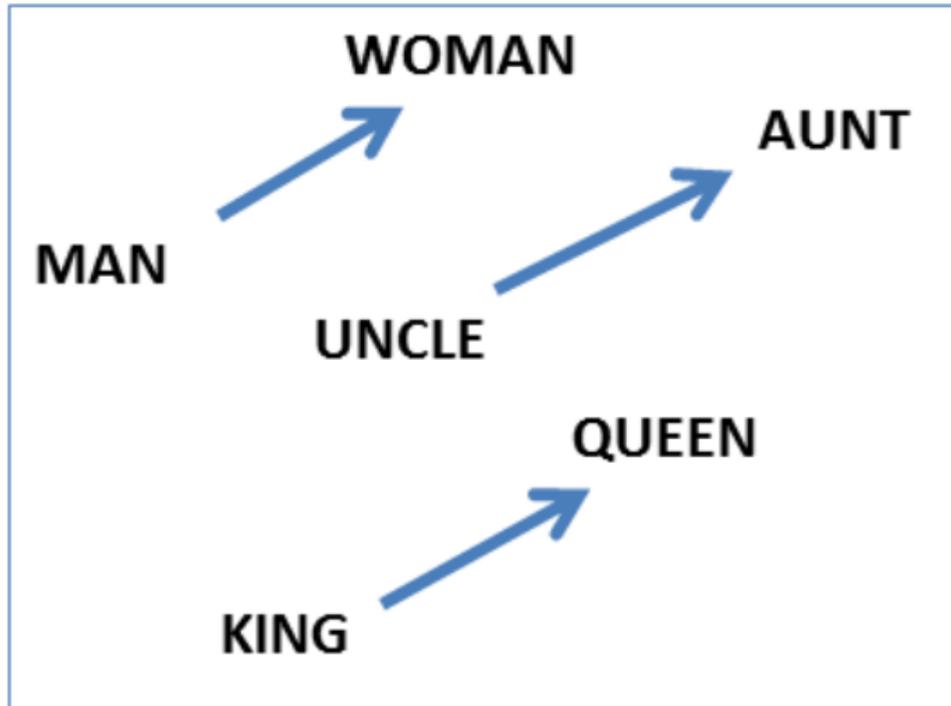
Word Embeddings

Erläuterung



Räumliche Nähe =
Semantische Ähnlichkeit

Beziehungen zwischen den Worten durch Raumbeziehungen kodiert

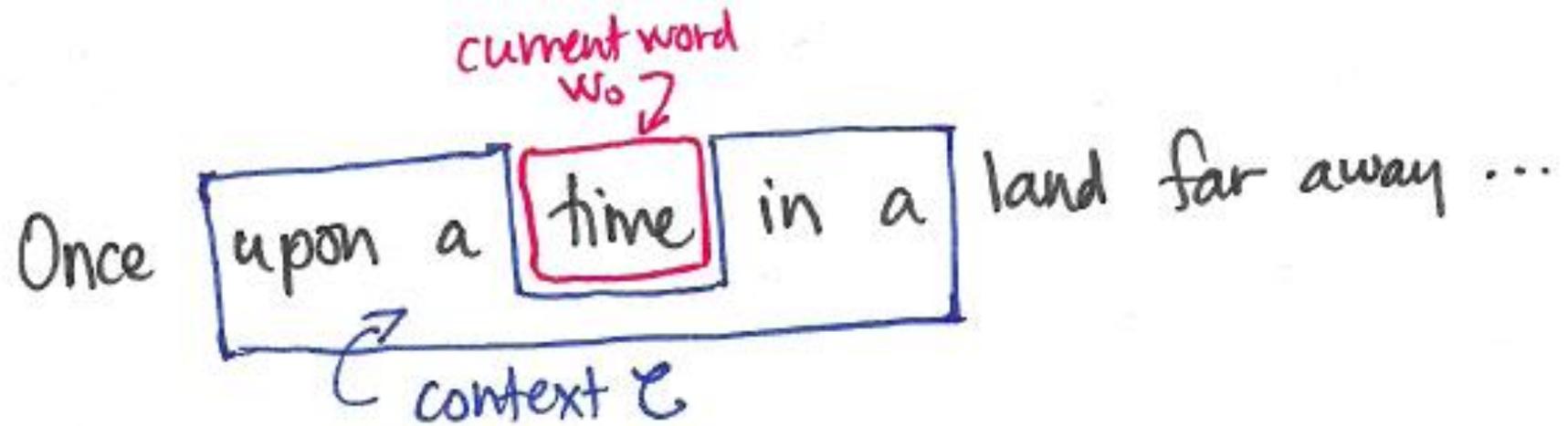


Ermöglicht Vektoralgebra:

$$V[\text{King}] - V[\text{Queen}] = V[\text{Man}] - V[\text{Woman}]$$

Vorhersage von w_0 bei gegebenem C (CBOW)

INPUT TEXT:



Modell

- Eine Wortliste und zu jedem Wort ein Vektor, häufig 200-300.
- Dichter Vektor (verglichen mit Dokument-Wort-Matrix)
- Enthält Information über Ähnlichkeit als Abstand
- Enthält weitere Informationen als Richtungsvektoren

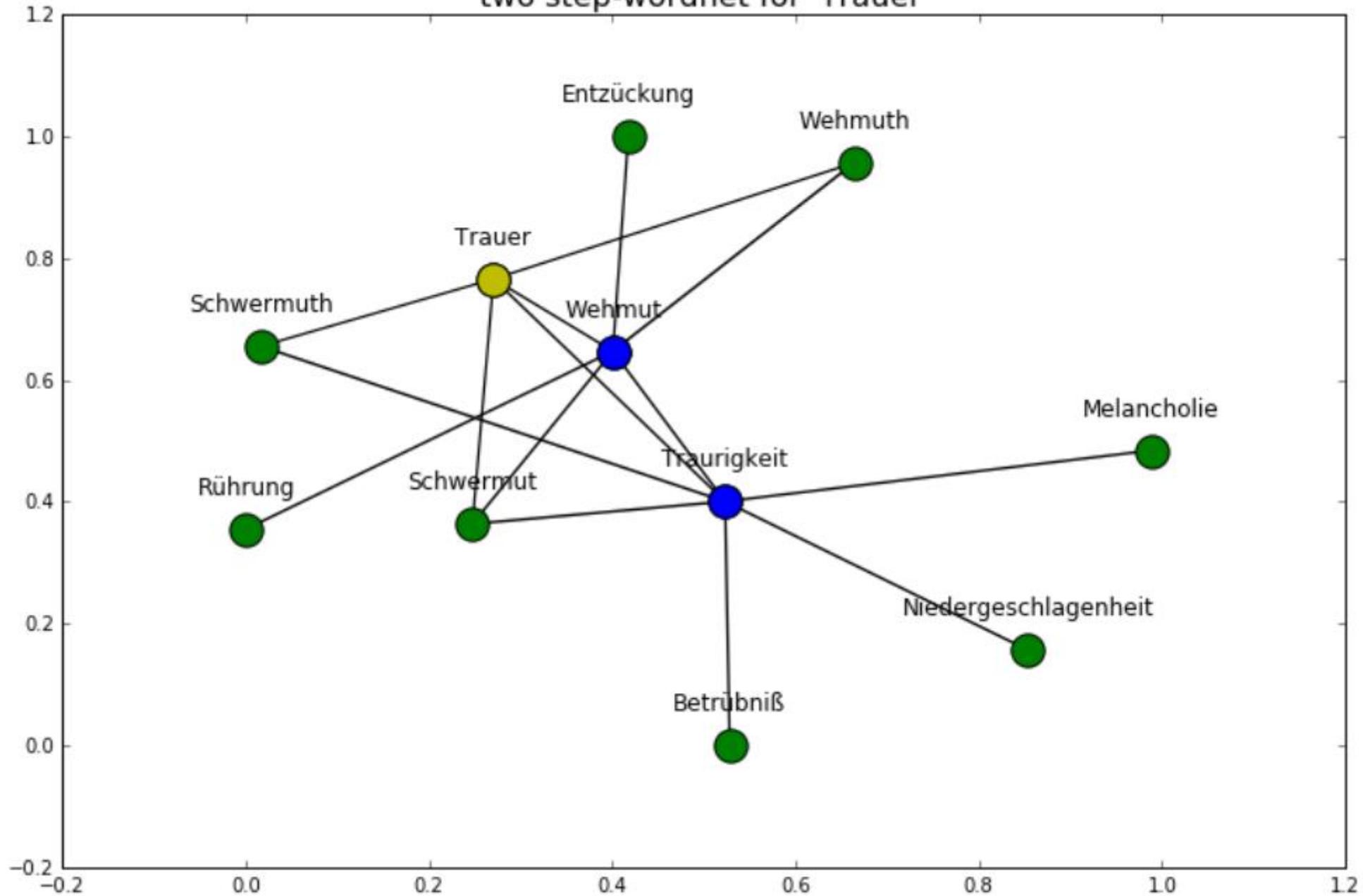
```
1 model.similar_by_word("Strukturalismus", topn=20)
```

```
[('strukturalistischen', 0.7546783089637756),  
 ('Strukturfunktionalismus', 0.7180768251419067),  
 ('Poststrukturalismus', 0.7170661687850952),  
 ('strukturalistische', 0.7160986661911011),  
 ('poststrukturalistischen', 0.7148711681365967),  
 ('poststrukturalistische', 0.6963427066802979),  
 ('analytischen_Philosophie', 0.6950880289077759),  
 ('Gegenwartsphilosophie', 0.6882011890411377),  
 ('Diskurstheorie', 0.6843668818473816),  
 ('Gilles_Deleuze', 0.6843351125717163),  
 ('Positivismus', 0.6828920841217041),  
 ('Strukturalisten', 0.6828206777572632),  
 ('kritischen_Realismus', 0.6826651096343994),  
 ('Dekonstruktivismus', 0.6809293031692505),  
 ('Roland_Barthes', 0.6795011758804321),  
 ('Holismus', 0.6781435012817383),  
 ('Evolutionismus', 0.6779667139053345),  
 ('Essentialismus', 0.6773746609687805),  
 ('systemtheoretischen', 0.6760599613189697),  
 ('konstruktivistischen', 0.675438404083252)]
```

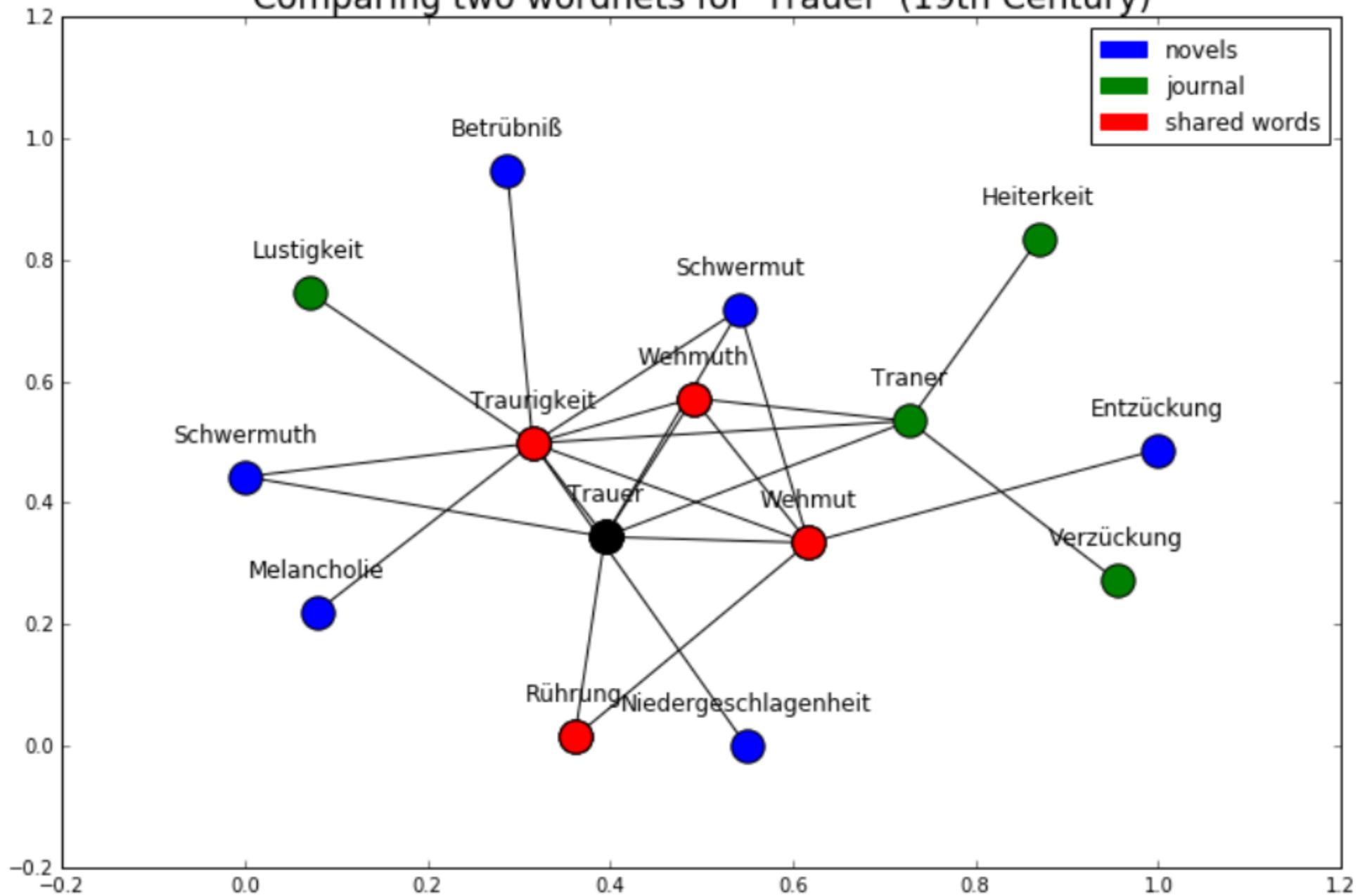
```
1 model.similar_by_word("Barthes", topn=10)
```

```
[('Derrida', 0.6827796697616577),  
 ('Roland_Barthes', 0.6703132390975952),  
 ('Deleuze', 0.6700422763824463),  
 ('Althusser', 0.660750150680542),  
 ('Foucault', 0.6597222089767456),  
 ('Althusser', 0.6522998809814453),  
 ('Lyotard', 0.6496005654335022),  
 ('Jean-François_Lyotard', 0.6457963585853577),  
 ('Baudrillard', 0.6423356533050537),  
 ('Bourdieu', 0.6418837308883667)]
```

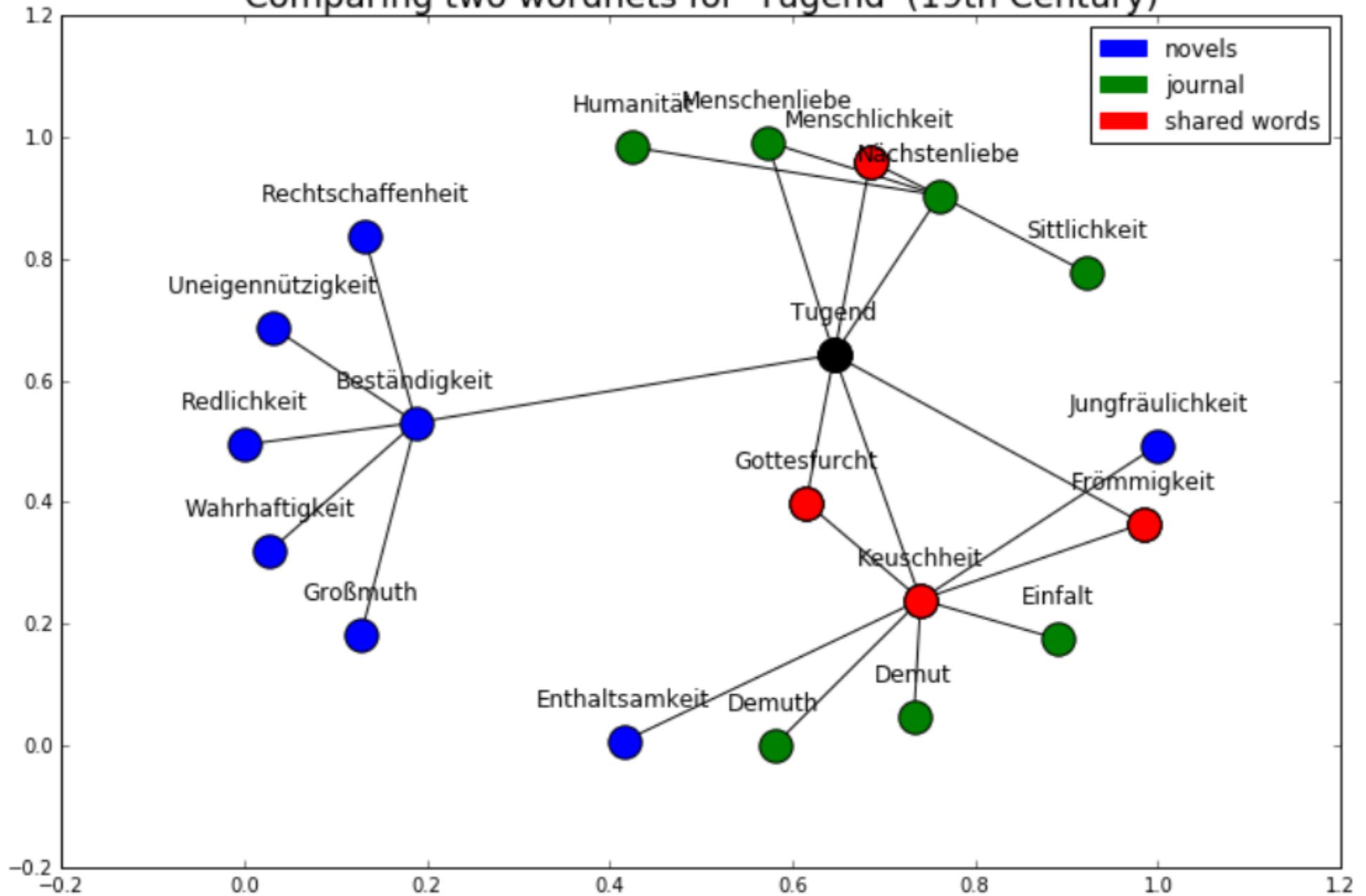
two step-wordnet for 'Trauer'



Comparing two wordnets for 'Trauer' (19th Century)



Comparing two wordnets for 'Tugend' (19th Century)



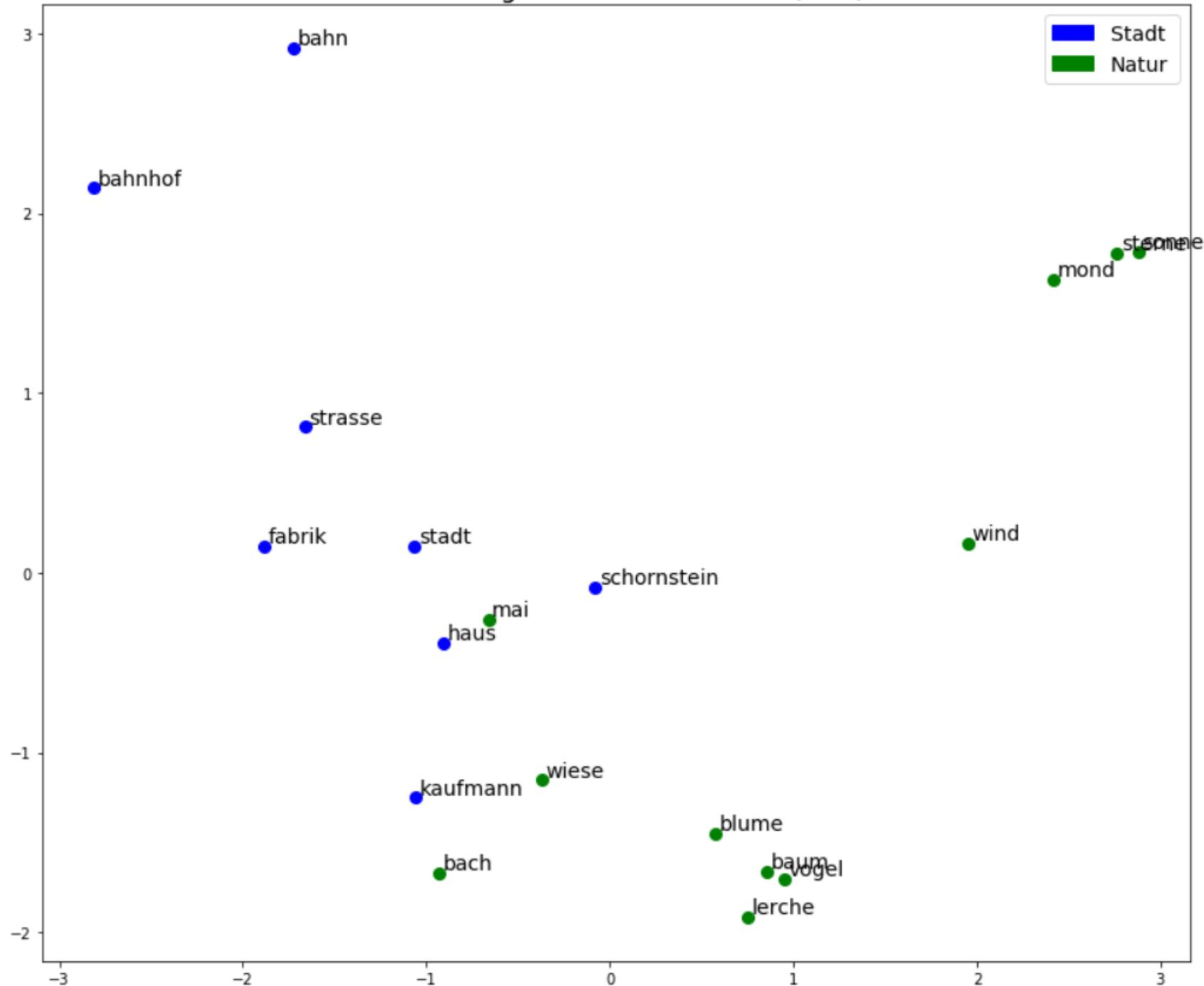
Spieldaten

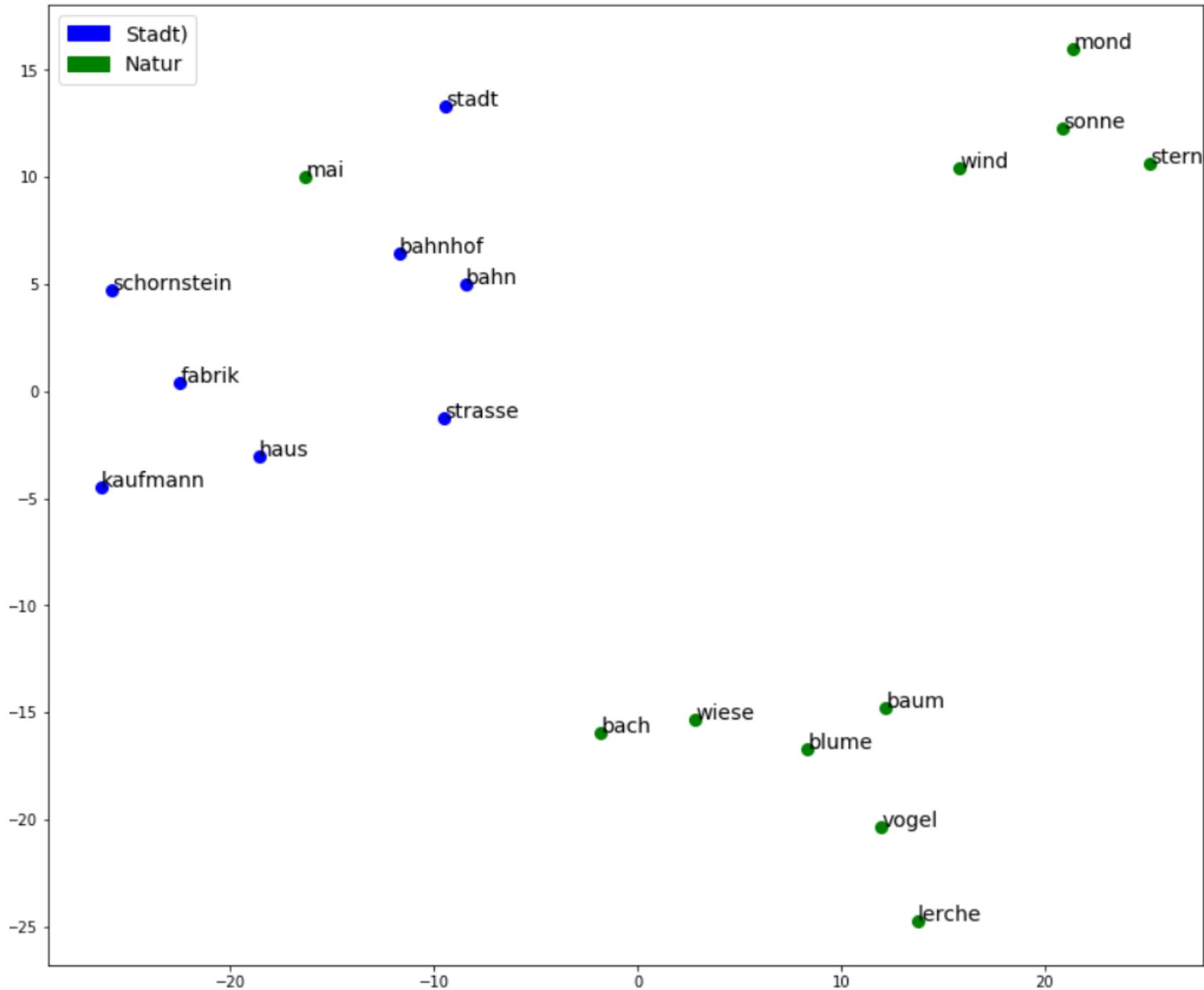
- text_nat = "Sonne Mond Sterne Blume Wiese Bach Wind Baum Vogel
Lerche Mai"
- text_cit = "Stadt Strasse Haus Schornstein Bahn Fabrik Bahnhof
Kaufmann"

Embedding Modell

- Fasttext
- 296 Sprachen
- Buchstaben-NGramme

Embeddings für Natur vs. Stadt (PCA)





t-sne

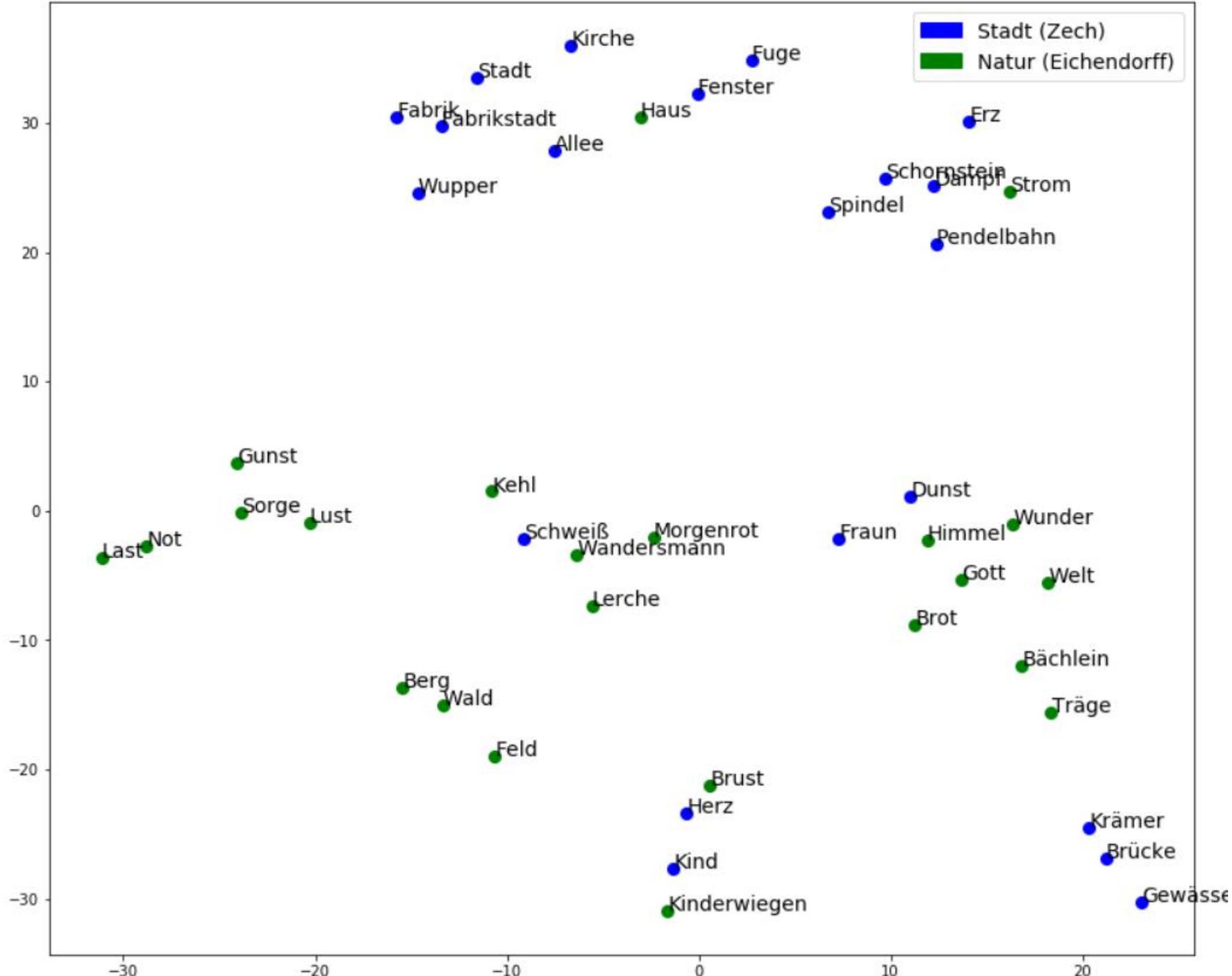
Problem 1:

Distanz zwischen Texten und Modell

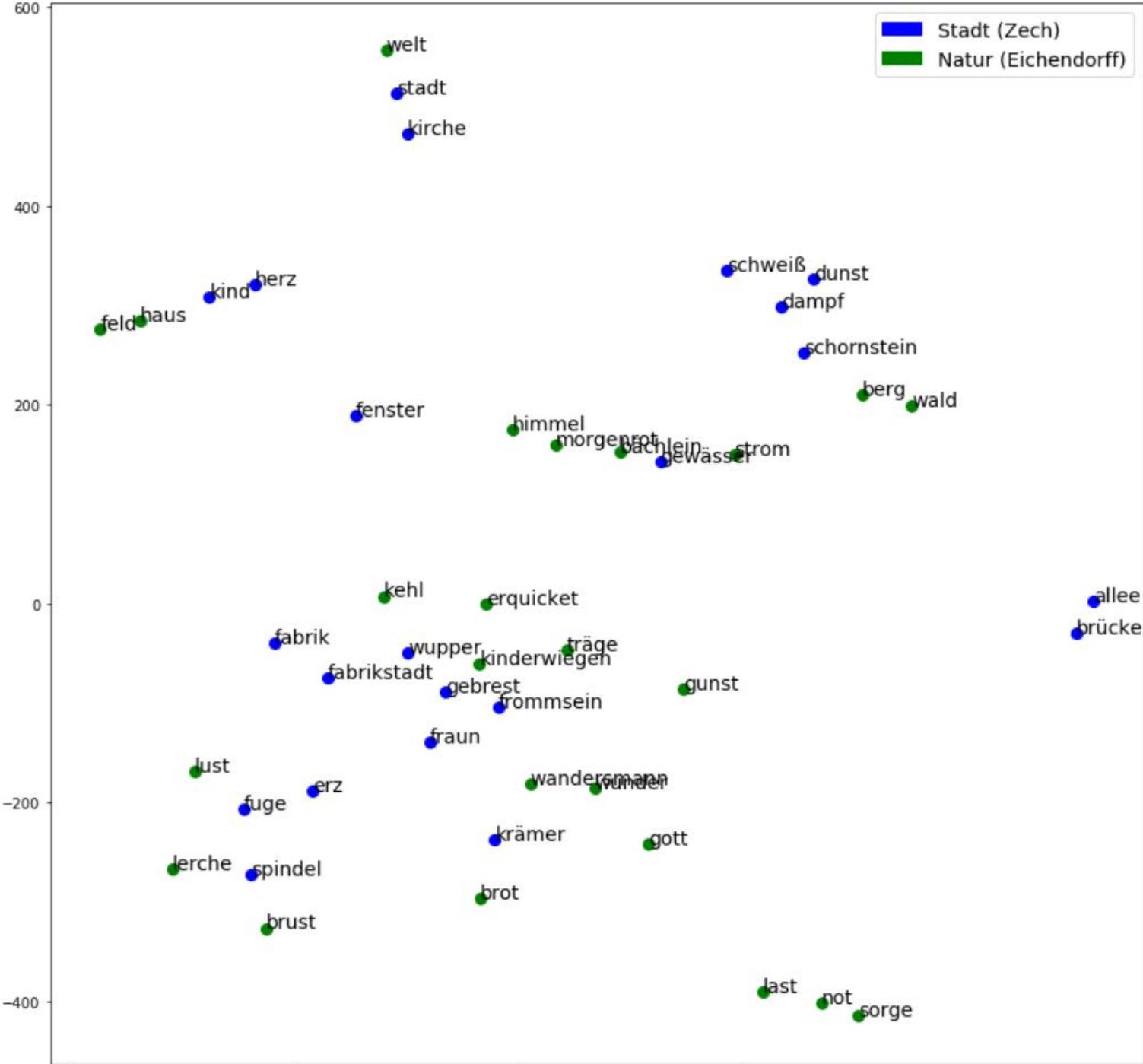
- Das Modell von Fasttext ist auf der deutschsprachigen Wikipedia von 2016 trainiert. Die Gedichte stammen aus dem langen 19. Jh.
- Differenz der Textsorten
- Historische Differenz

Texte des 19. Jahrhunderts

- Auswahl an Stadt- und Naturgedichten: (nur Worte, keine Satzzeichen)
 - Tokens: 7785
 - Types: 3713
- Sammlung 19. Jahrhundert (rd. 20.000 Texte)
 - 294.754.982 Tokens
 - 3.441.196 Types



Fasttext Modell
t-sne



Modell 19. Jahrhundert
t-sne

Problem 2: Wörter, die im Modell fehlen

- 196 Types, ca 5%, nicht in der Liste der Word Embeddings:
- z.B.: heißrot
- entbröckelt
- sprühregenflitter
- spiegelschein
- vorübersterbe menschengraus binsenkraut unregbar tigerhaut
heidewelt bedornte seegewürm huppenklänge riesensteinmeer
enteitert puderhände seelenburgen schattengewühl ...

Lösungsansätze

- Ignorieren
- Durchschnittswert aus zufällig anderen seltenen Worten bilden
- Character ngramme im Modell verwenden (Fasttext)
- Überbegriff aus Thesaurus nehmen
- Viele der unbekanntes Wörter sind Komposita: Komposita zerlegen und dann über Durchschnitt der beteiligten Wörter mitteln
- Mischung aus diesen Verfahren

Problem 3: Vom Wort zum Satz/Text

- Par2vec, Doc2vec (braucht sehr große Korpora)
- Durchschnitt über Wortvektoren berechnen
- Word Mover Distance

Durchschnitt berechnen

Word₁ = [i₁, i₂, i₃, ..., i_n]

Word₁ = [i₁, i₂, i₃, ..., i_n]

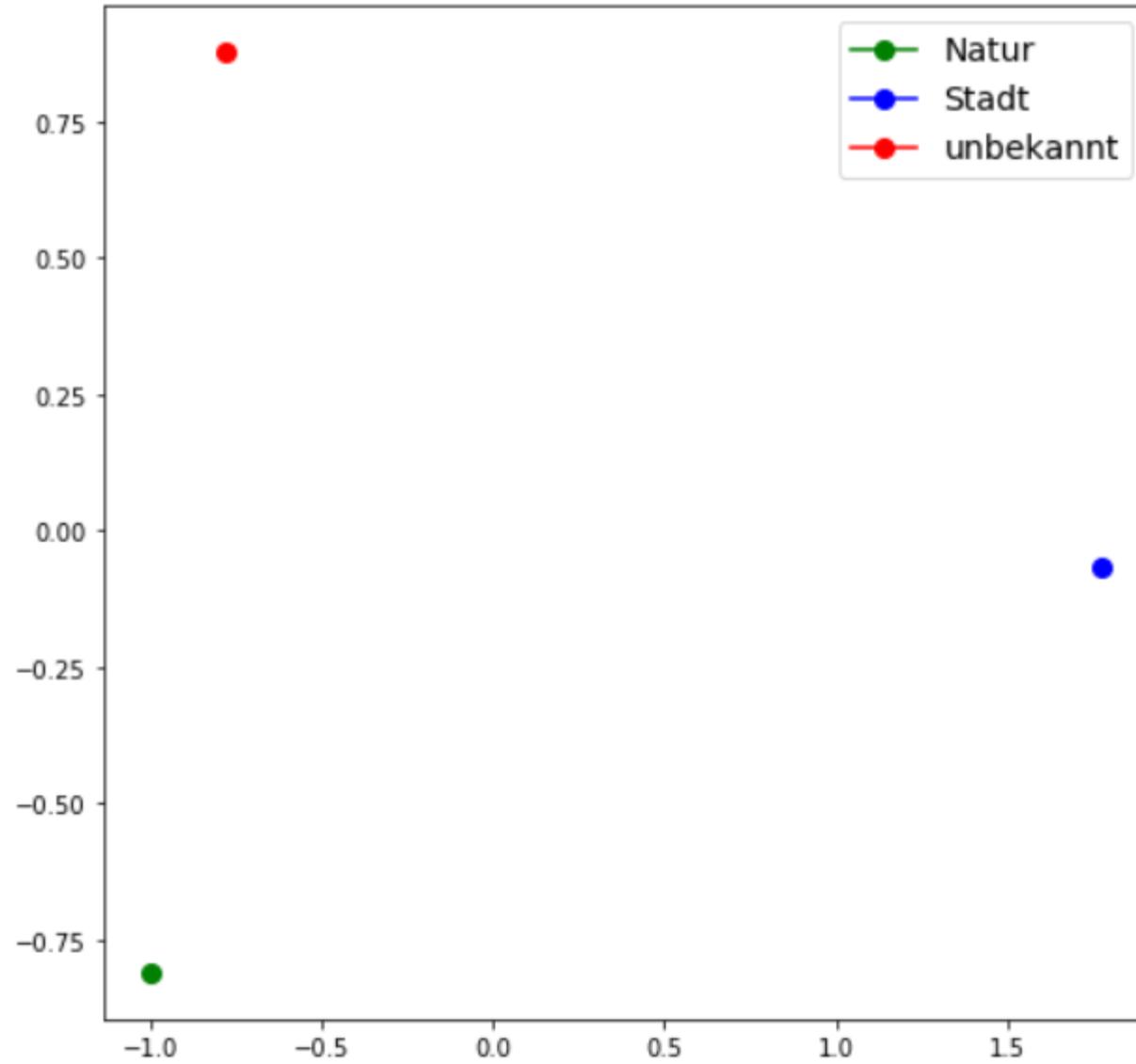
...

Word_n = [i₁, i₂, i₃, ..., i_n]

Text = [i_{1-n}]

Spielmaterial

- `text_nat = "Sonne Mond Sterne Blume Wiese Bach Wind Baum Vogel Lerche Mai".split()`
- `text_cit = "Stadt Strasse Haus Schornstein Bahn Fabrik Bahnhof Kaufmann".split()`
- `text_X = "Sonnenschein Frühling Rose Veilchen Gras Busch Tal Reh Hase Ameise Luft".split()`



Textvektoren: Durchschnitt

Word Mover Distanz

- $d(\text{nat-cit})$: 10.51
 - $d(\text{nat-X})$: 11.12
 - $d(\text{cit-X})$: 10.99
- D.h. der Abstand des Naturgedichts zum Naturgedicht X ist größer als als zwischen den Naturgedichten und dem Stadtgedicht. Passt nicht zu unserem Textverständnis.

Bedeutung im ‚distant reading‘

- Nicht integral, sondern selektiv
- Nicht wahrnehmungsbasiert, sondern rein relational
- Perspektivenspezifisch
- Schnell bewegliches Ziel

Literatur

- David Blei, Andrew Ng, and Michael Jordan. "Latent dirichlet allocation." In: Journal of machine Learning research 3 (2003): 993-1022.
- David Blei: Probabilistic Topic Models. Communications of the ACM 55,4 (2012), p. 77-84.
- John Burrows (2002). 'Delta'—A measure of stylistic difference and a guide to likely authorship . Literary and Linguistic Computing, 17 (3): 267–87.
- Stefan Evert, Thomas Proisl, Isabella Reger, Steffen Pielström, Christof Schöch, Thorsten Vitt: Understanding and explaining Delta measures for authorship attribution. In: Digital Scholarship Humanities 2017. DOI: <https://doi.org/10.1093/lc/fqx023> 
- Gunter E. Grimm: Deutsche Naturlyrik. Vom Barock bis zur Gegenwart. Stuttgart 1995.
- Fritz Hofmann u.a. (Hg.): Über die großen Städte. Gedichte 1885-1967. Berlin, Weimar 1968.
- Fotis Jannidis: Quantitative Analyse literarischer Texte am Beispiel des Topic Modeling. In: Der Deutschunterricht 68,5 (2016), 24-35.
- Fotis Jannidis: Netzwerke. In: F.J./ Hubertus Kohle / Malte Rehbein (Hg.): Digital Humanities. Eine Einführung. Stuttgart: Metzler 2017, 147-161.
- Matt J. Kusner, Yu Sun, Nicholas I Kolkin, Kilian Q. Weinberger: From Word Embeddings To Document Distances. *Proceedings of the 32 nd International Conference on Machine Learning, Lille, France, 2015.*
- Wolfgang Rothe (Hg.): Deutsche Großstadtlyrik vom Naturalismus bis zur Gegenwart. Stuttgart 1975.

Werkzeuge, Bibliotheken, Ressourcen

- Jupyter Notebook
 - Python: matplotlib, pandas, gensim, pyemd, scikit-learn
 - Dkpro und Dkpro Wrapper
 - Modelle: Fasttext
-
- Meinen Dank an all die Kolleginnen und Kollegen: Andreas, Frank, Isabella, Markus, Albin, Leonard, Lukas ...