Text- und Data Mining in der Literaturwissenschaft

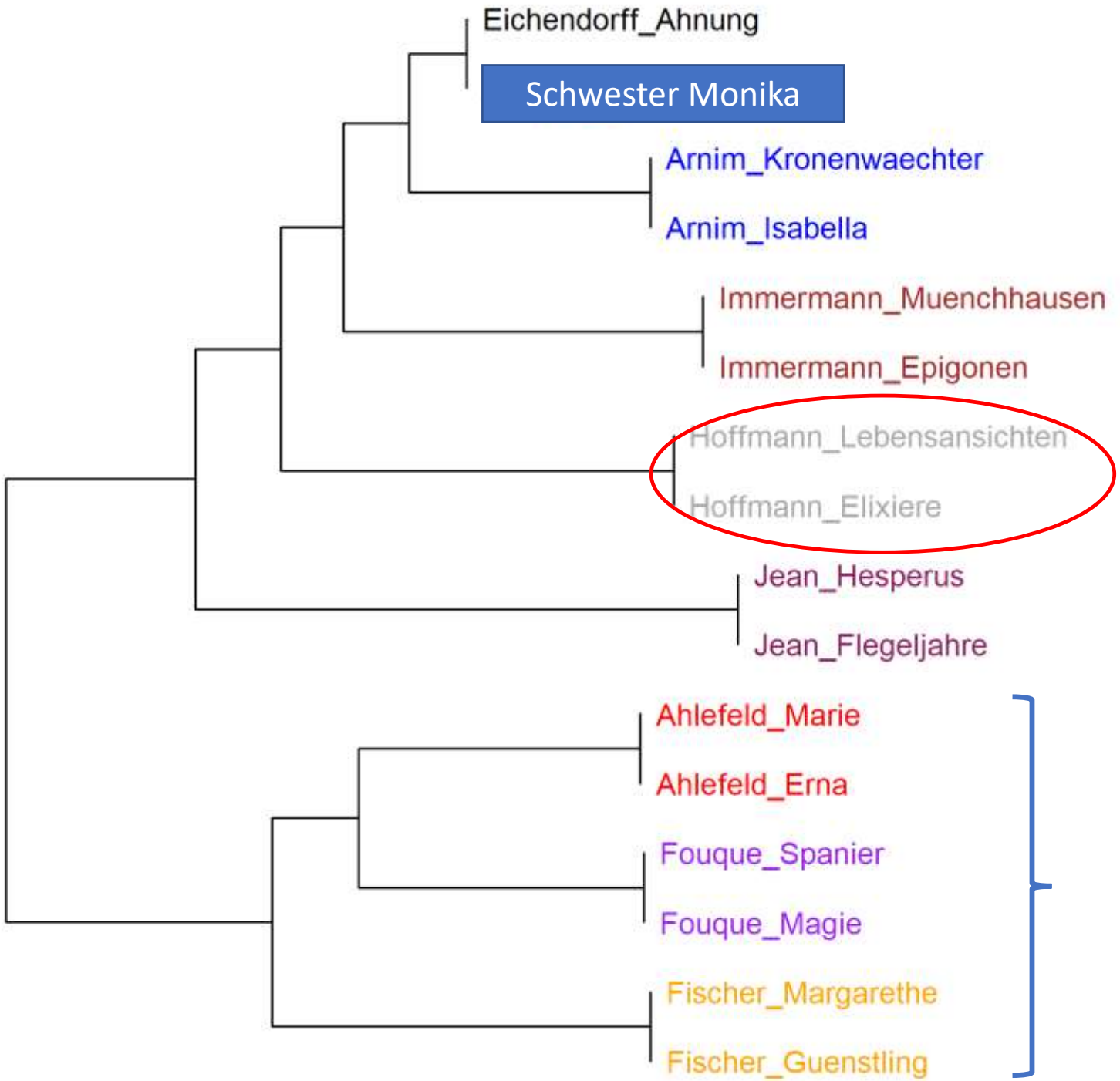
Fotis Jannidis (Würzburg)

Typische Verwendungen digitaler Daten in den Literaturwissenschaften

- Lesen
- Suchen (insbesondere Google Books)
- Edition, also Annotation von textueller Varianz und Genese
- Weitere Annotationen, z.B. Kommentar, Intertextualität usw.

- Optical Character Recognition: Goldstandards
- Textsammlungen für Textmining

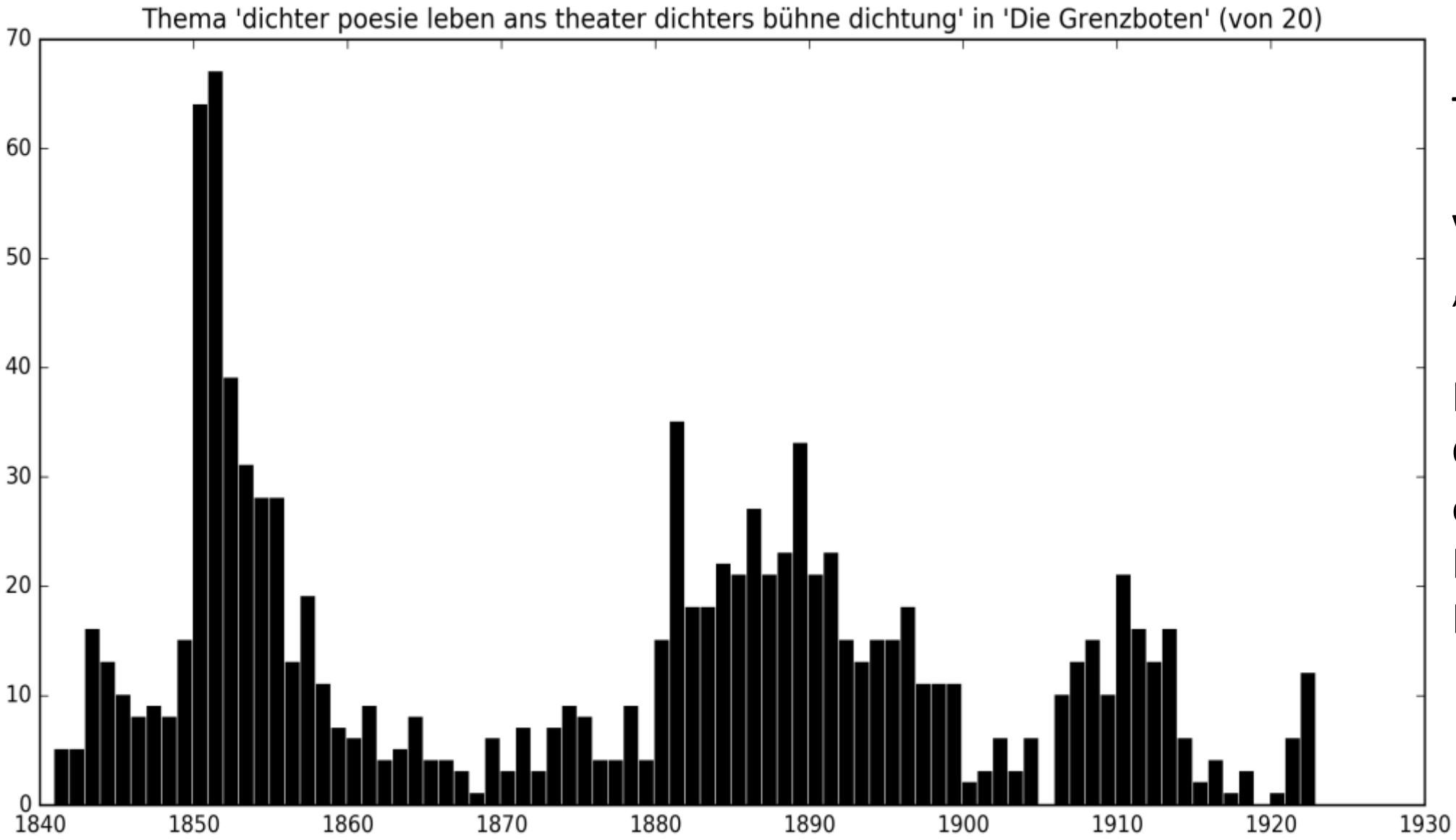
Dokument-Term-Matrix
Bag of words



Stilometrie

Fragestellung:
Ist E.T.A Hoffmann der Autor der ‚Schwester Monika‘?

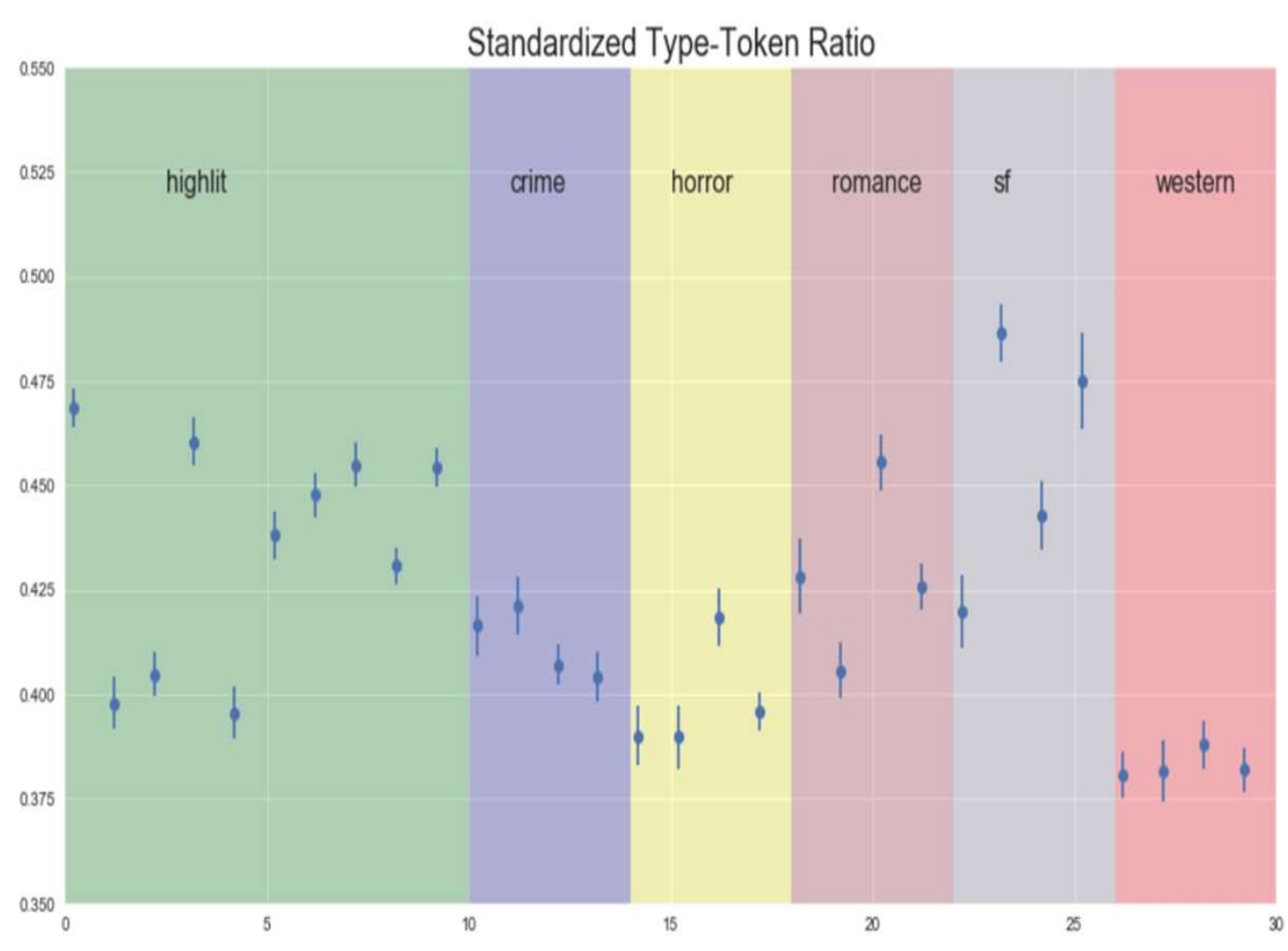
Weibliche Autoren



Topic Modeling

Verteilung von ‚topics‘ über Texte

Fragestellung
Gibt es nach 1848 ein verstärktes Interesse an Literatur?



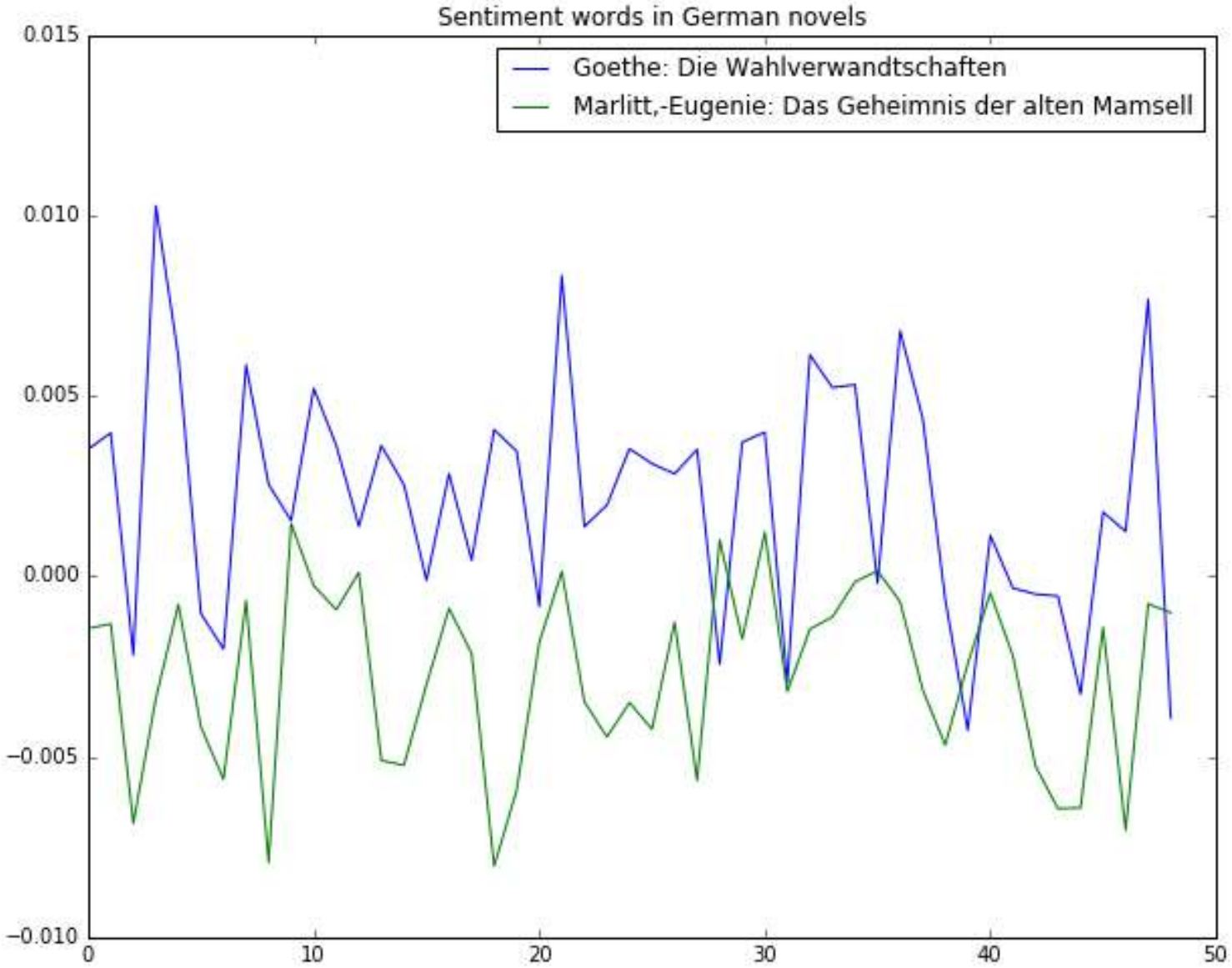
Komplexität des Vokabulars

Fragestellung:
Haben Texte der Hochliteratur ein komplexeres Vokalbular

	Ahlefeld_Erna	Ahlefeld_Marie	Anonym_Schwester	Arnim_Isabella
und	3.12795700844477	3.50684290560299	4.56459764157056	2.93920011945945
der	2.61377229472782	2.09849105158498	2.31307502915641	2.76001095044922
die	2.74767456392494	2.92899754357235	2.62083711286769	2.54349070456186
er	1.628251593437	1.25160837524857	0.424387715433459	1.86655384385655
in	1.6657442288122	1.49257223067025	1.3152779577556	1.62016873646748
sie	2.12279730767171	3.2635395952743	1.34767396656732	2.45389612005674
zu	2.23884594097588	1.83647210200023	1.41246598419075	1.3862273213708
ich	0.490974987056114	1.39197566966897	1.74290527407023	0.674448122246833
den	1.14441806073807	1.08082816703708	1.23752753660749	1.2543241830716
das	1.04622306332685	0.89133231956954	1.16301671634055	0.997984121848635
sich	1.45507132527539	1.12527781027021	0.761306207075288	1.2344142754038
mit	0.974808519755048	1.403672944204	0.939484255539717	0.958164306513029
nicht	0.926603702844084	0.88431395484852	0.732149799144745	0.978074214180832
dem	0.68379425469997	0.65972628377588	0.660878579758974	0.910878275801996
ein	0.624877256253236	0.577845362030647	0.751587404431774	0.861103506632488
so	0.992662155647998	0.844543221429407	0.748347803550603	0.851148552798586
es	0.748067343914588	0.706515381916014	0.49565893481923	0.744132799084144
auf	0.635589437789006	0.556790267867587	0.907088246728003	0.579876060824768
von	0.810555069539912	0.608258275821734	0.864973435272774	0.756576491376521
daß	0.592740711645927	0.479588255936367	0.515096540106259	0.846171075881636
wie	0.596311438824517	0.4912855304714	0.647920176234288	0.711779199123964
als	0.835550159790041	0.67376301321792	0.667357781521317	0.676936860705308
war	0.555248076270733	0.797754123289274	0.330439289879487	0.694358029914636
an	0.530252986020603	0.61059773072874	0.534534145393288	0.477837784027277
des	0.578457802931567	0.3158264124459	0.573409355967345	0.460416614817949

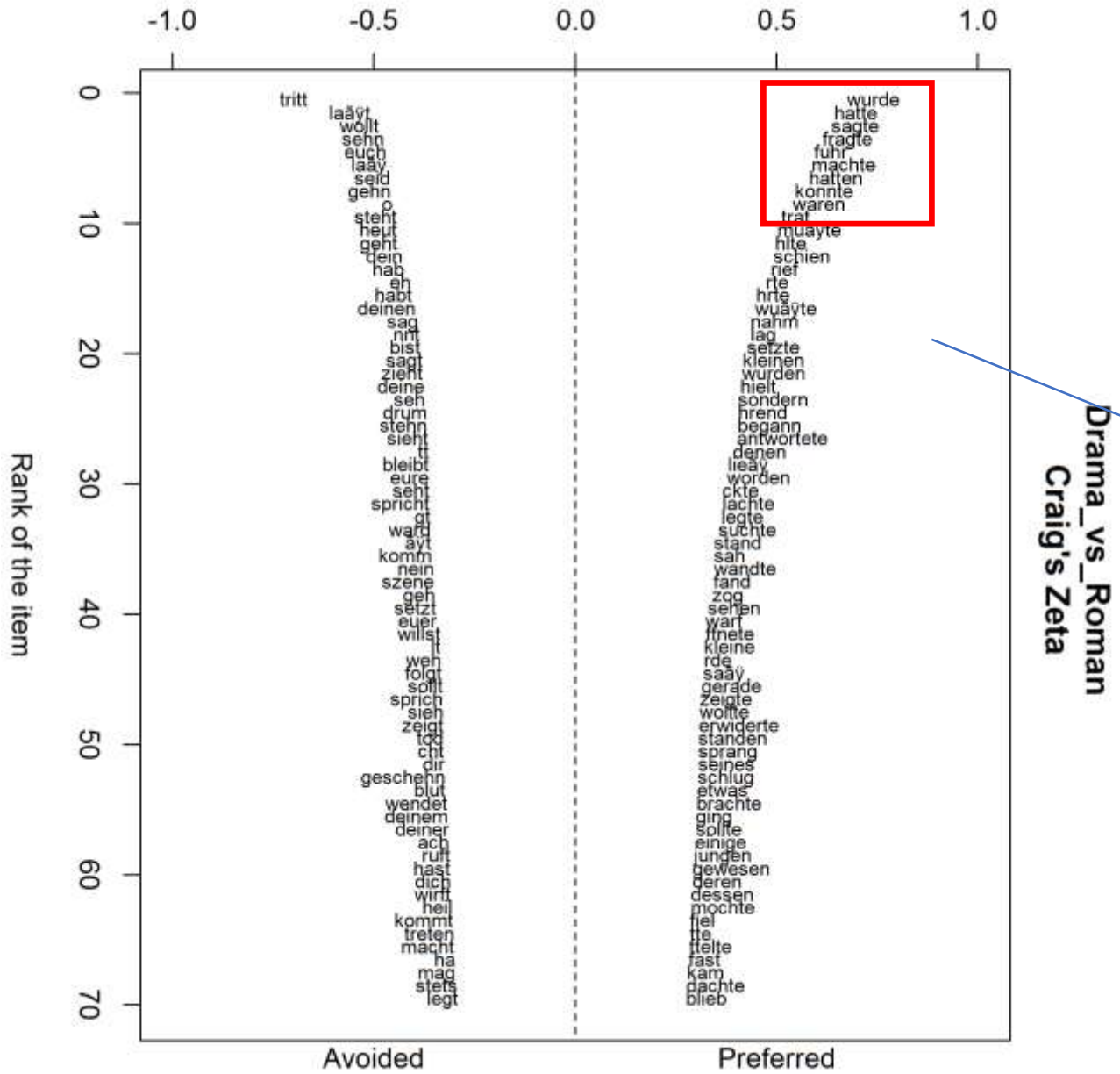
Term-Dokument-Matrix

Segmentiertes BOW-Modell



Sentiment Analyse

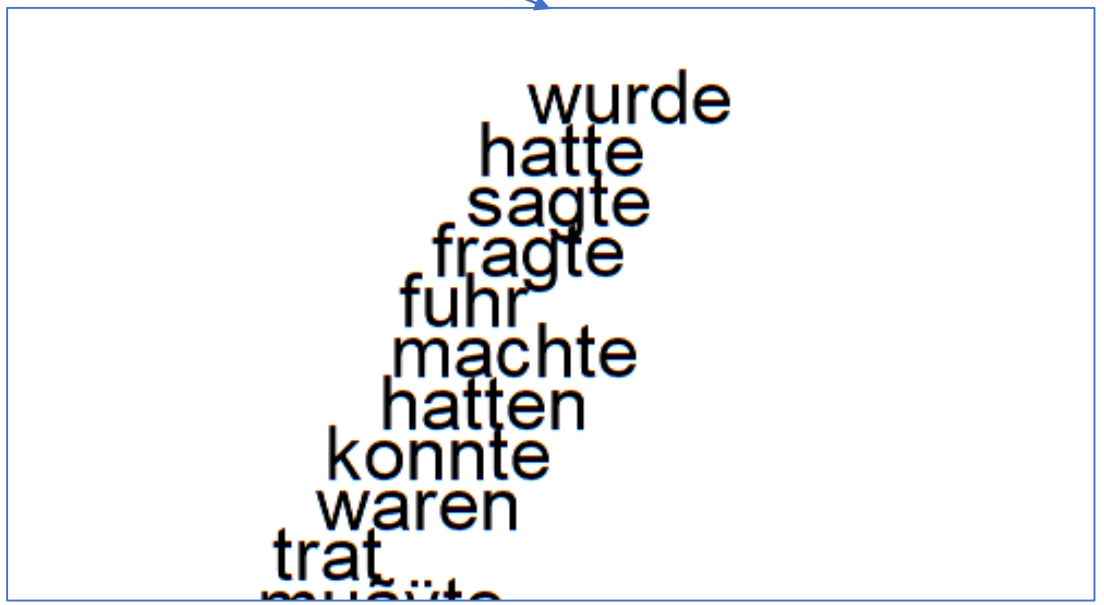
Fragestellung:
Kann man ‚Happy Ends‘ automatisch erkennen?
Und:
Ist Trivialliteratur positiver als Hochliteratur?

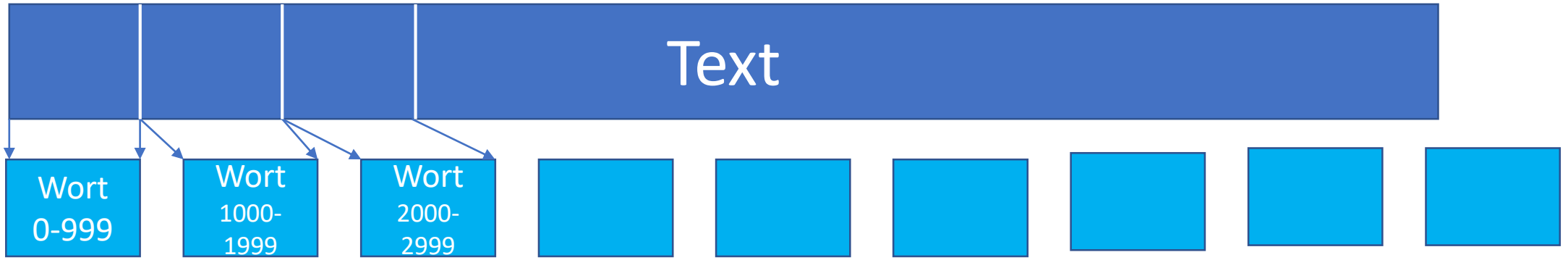


Zeta:

Vergleich von Textgruppen aufgrund von bevorzugten / vermiedenen Worten

Fragestellung:
Wodurch unterscheidet sich Prosa vom Drama





und	3.12795700844477
der	2.61377229472782
die	2.74767456392494
er	1.628251593437
in	1.6657442288122
sie	2.12279730767171
zu	2.23884594097588
ich	0.490974987056114
den	1.14441806073807
das	1.04622306332685
sich	1.45507132527539
mit	0.974808519755048
nicht	0.926603702844084
dem	0.68379425469997

Wort-Häufigkeitsliste pro Segment

Ngramme

Ngramm-Suche

Fragestellung:

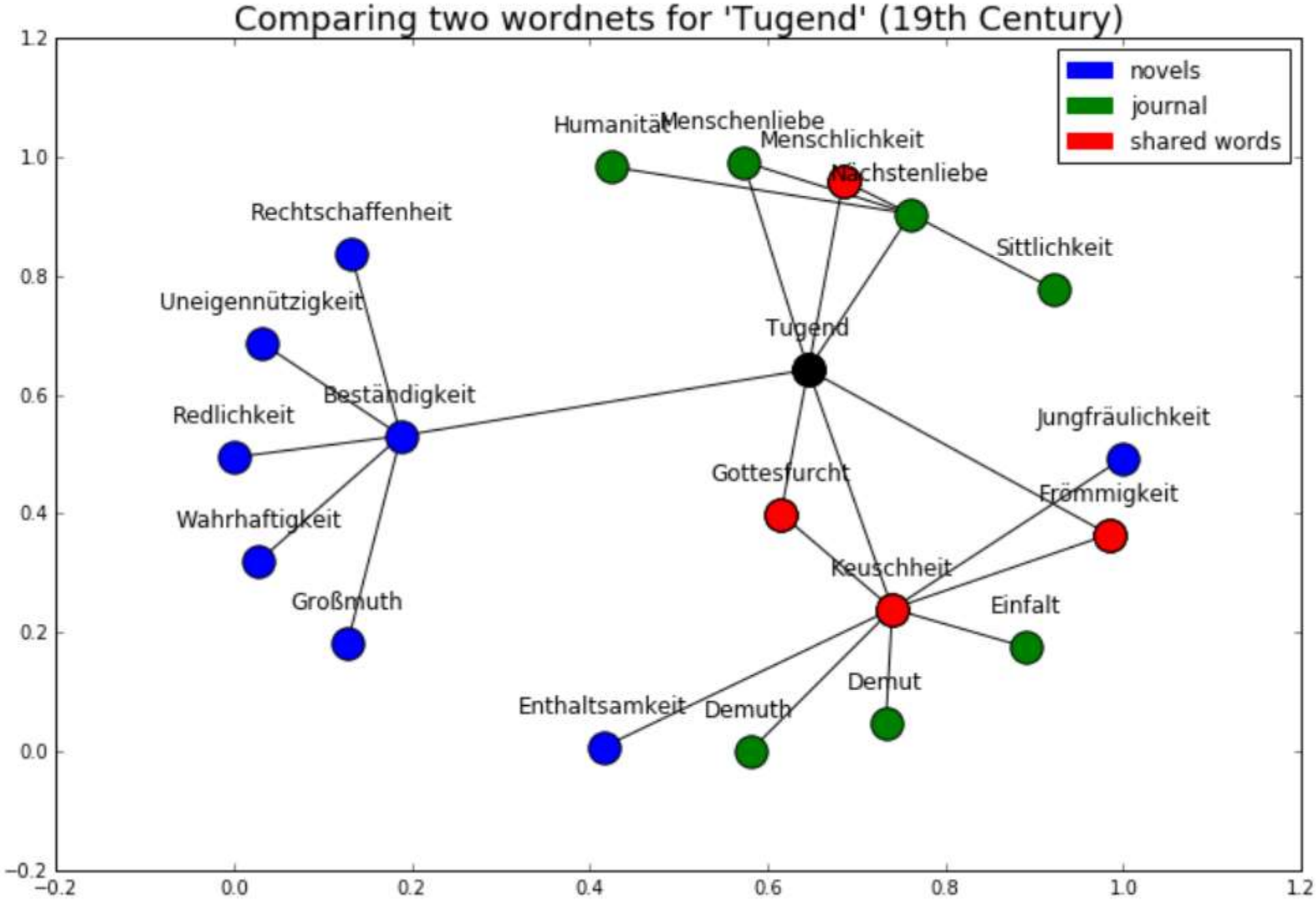
Ist Alter oder Aussehen die wichtigere Wahrnehmungsdimension?

Ngramm-Format Google 2

Bigramm	Jahr	Häufigkeit	Bücher
schöne Frau	1978	335	91
Schöne Frau	1978	261	46
schöne Frau	1980	348	95
Schöne Frau	1980	233	43

1-Gramme bis 5-Gramme, Anzahl < 10 werden nicht aufgenommen

Wortvektoren (*word embeddings*)



Word Embeddings

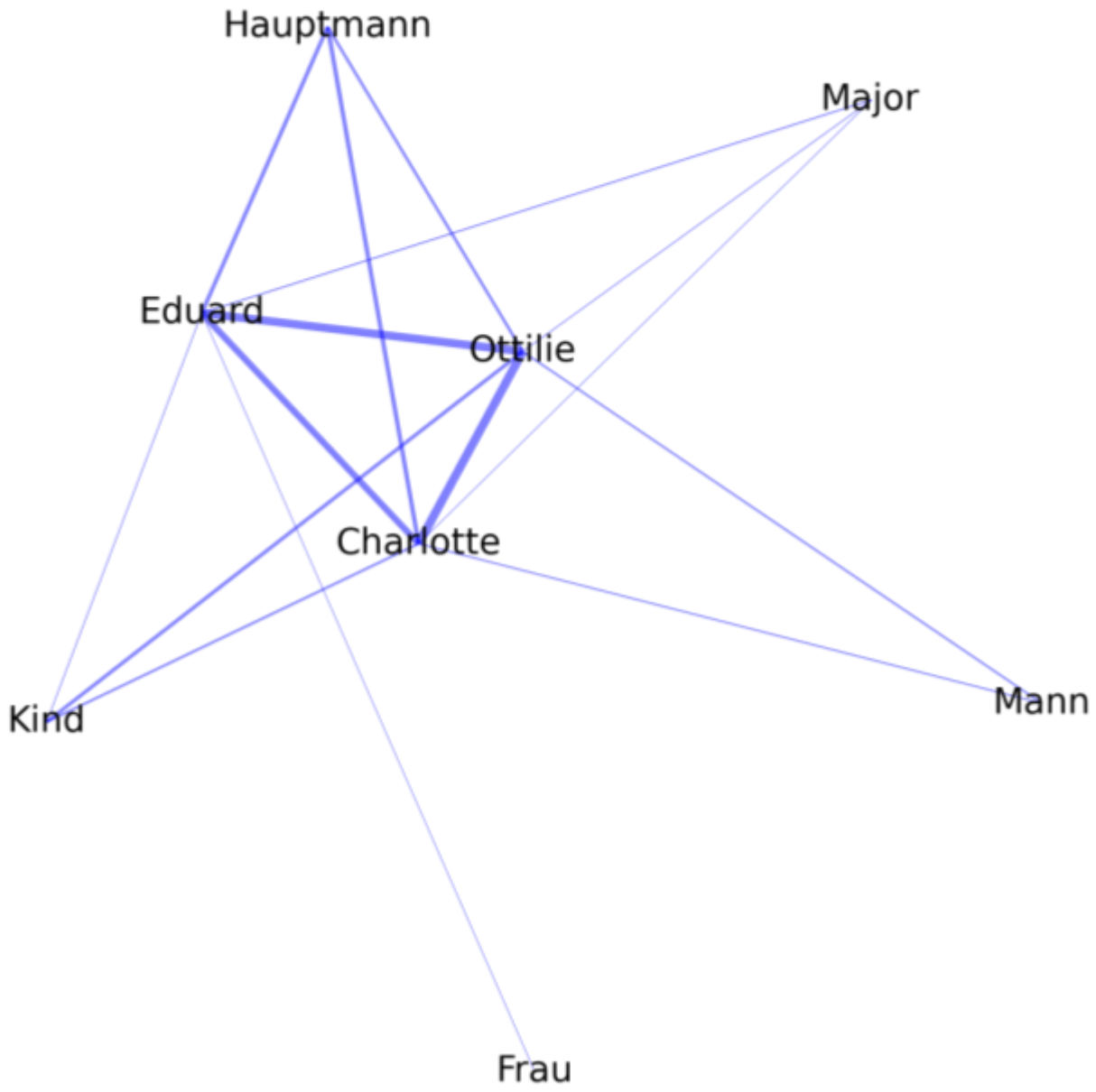
Fragestellung:
Wie verhalten sich Wortbedeutungen in unterschiedlichen Textsorten?

dritten -0.3235 0.34409 -0.60493 -0.27091 -0.14214 0.091305 …
ehe -0.71767 0.19702 -0.35163 -0.4692 0.22042 0.338 0.016256 …
koronarsport -0.32817 0.10931 0.038537 0.28134 0.91661 0.177 …
austrobaileya -0.28194 0.74169 0.094848 0.18052 0.19486 0.50 …
längerdauernde -0.062823 0.022873 0.12147 -0.2547 -0.19937 …
hausvaters -0.30787 0.55782 -0.21791 -0.062345 -0.38914 -0.1 …
pfeifkonzert -0.028291 0.97464 -0.3016 -0.65689 0.57015 0.11 …
beitrittes -0.30852 -0.041101 -0.27951 -0.76973 0.53353 0.53 …
förderbeitrag 0.29313 -0.12741 -0.18391 -0.057058 0.53828 0. …
pernthaler -0.14985 0.48605 -0.26858 0.19827 0.24308 0.22706 …
dachknauf -0.10577 -0.056043 -0.093128 -0.21443 -0.052583 -0 …
ausstattungsniveau 0.048696 -0.37609 -0.30301 -0.27865 -0.28 …
birkebeinern -0.61793 1.1085 -0.88521 -0.88392 -0.22911 0.31 …
malepartus -0.2116 0.3991 0.1918 -0.21442 0.15148 -0.065504 …
gravenberg 0.039838 0.49038 -0.39001 0.16243 -0.028735 0.402 …
norfolks -0.34032 0.12549 -0.015844 -0.2381 0.078124 0.33239 …
frankoprovenzalische -0.32825 0.63102 -0.20729 -0.51876 0.21 …
parschin -0.19682 0.47458 -0.51535 -0.43258 0.047059 0.4212 …
grabmalkunst 0.039913 0.48895 -0.41024 -0.063175 -0.23528 0. …

Wort und n-dimensionaler Vektor

zumeist 100-300

Netzwerkdaten

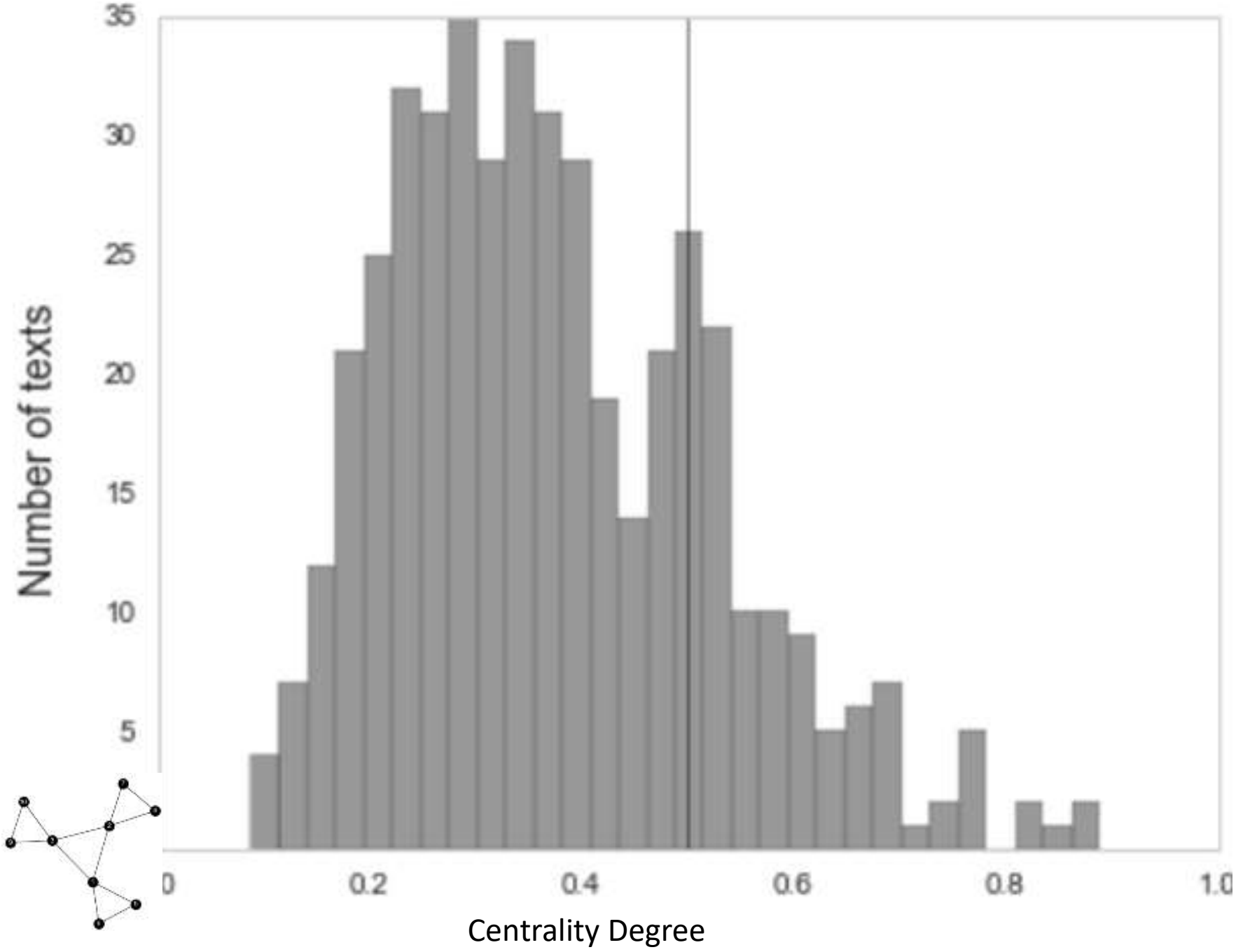


Netzwerkanalyse

Fragestellung:
Was sind Hauptfiguren des Textes
und wie interagieren sie

Verteilungen von Netzwerkmaßen

Fragestellung:
Konzentrieren sich Romane zumeist auf eine Hauptfigur?



ParagraphId	SentenceId	TokenId	Token	Lemma	POS	Morphology	NamedEntity	CorefId
0	0	7	Annerl	Annerl	NE	masc|sg|	B-PER_CORE	180
4	6	111	Bäuerin	Bäuerin	NN	fem|sg|	B-PER_CORE	4
4	8	194	Schürze	Schürze	NN	|pl|	B-PER_CORE	182
12	29	718	Straße	Straße	NN	fem|sg|	B-PER_CORE	46
12	30	741	Bäuerin	Bäuerin	NN	masc|sg|	B-PER_CORE	4
16	49	1157	Thaler	Thaler	NN	masc|sg|	B-PER_CORE	65
30	69	1652	Herrn	Herr	NN	masc|sg|	B-PER_CORE	35
37	77	1780	Fähnrich	Fähnrich	NN	masc|sg|	B-PER_CORE	202
42	85	1998	Mütterchen	Mütterche	NN	fem|pl|	B-PER_CORE	206
52	111	2593	Uhlane	Uhlane	NN	masc|sg|	B-PER_CORE	40
52	114	2678	Finkel	Finkel	NN	masc|sg|	B-PER_CORE	41
52	117	2768	Straßburg	Straßburg	NE	neut|sg|	B-PER_CORE	46
52	128	2989	Gieb	Gieb	ADJA	masc|sg|	B-PER_CORE	214
53	137	3180	Meister	Meister	NN	masc|sg|	B-PER_CORE	121
54	141	3269	Pharisäer	Pharisäer	NN	masc|pl|	B-PER_CORE	218
58	167	4021	Gottes	Gott	NN	masc|sg|	B-PER_CORE	15
64	188	4432	Annerl	Annerl	NE	masc|sg|	B-PER_CORE	69
64	192	4549	Annerl	Annerl	NE	neut|sg|	B-PER_CORE	69
64	197	4678	Gottes	Gott	NN	masc|sg|	B-PER_CORE	15
65	199	4755	Vater	Vater	NN	masc|sg|	B-PER_CORE	155
65	199	4771	Annerl	Annerl	NE	masc|sg|	B-PER_CORE	69
65	200	4808	Herr	Herr	NN	masc|sg|	B-PER_CORE	72
65	205	4925	Kasper	Kasper	NN	masc|pl|	B-PER_CORE	74
65	206	4951	Kasper	Kasper	NN	masc|pl|	B-PER_CORE	74

CONLL reduziert auf NER

Angereicherte Volltextdaten

Grundlage für Filterprozesse

SectionId	ParagraphId	SentenceId	TokenId	Begin	End	Token	Lemma	CPOS	POS	Morphology	DependencyHead	DependencyRelation	NamedEntity	CorefId
null	0	0	0	0	4	?Der	?der	ADJA	ADJA	||	Invalide	NK	O	-
null	0	0	1	5	10	tolle	toll	ADJA	ADJA	||	Invalide	NK	O	-
null	0	0	2	11	19	Invalide	Invalide	N	NN	||	saß	SB	B-PER_APP	0
null	0	0	3	20	23	auf	auf	APPR	APPR	||	Invalide	MNR	O	-
null	0	0	4	24	27	dem	der	ART	ART	masc|sg|	Ratonneau	NK	O	-
null	0	0	5	28	32	Fort	Fort	N	NN	masc|sg|	Ratonneau	PNC	O	-
null	0	0	6	33	42	Ratonneau	Ratonneau	N	NE	masc|sg|	auf	NK	B-PER_CORE	81
null	0	0	7	43	46	von	von	N	APPR	||	Ratonneau	PG	I-PER_CORE	81
null	1	0	8	49	54	Achim	--	N	NE	||	von	NK	I-PER_CORE	81
null	1	0	9	55	58	von	von	APPR	APPR	||	Achim	PG	I-PER_CORE	81
null	1	0	10	59	64	Arnim	Arnim	N	NE	||	Dürande	PNC	I-PER_CORE	81
null	2	0	11	68	72	Graf	Graf	N	NE	||	Dürande	PNC	I-PER_CORE	81
null	2	0	12	73	80	Dürande	Dürande	N	NE	||	von	NK	I-PER_CORE	81
null	2	0	13	80	81	.	--	SYM	$.	||	Dürande	--	O	-
null	2	0	14	82	85	der	der	ART	ART	masc|sg|	Commandant	NK	O	-
null	2	0	15	86	90	gute	gut	ADJA	ADJA	||	Commandant	NK	O	-
null	2	0	16	91	95	alte	alt	ADJA	ADJA	masc|sg|	Commandant	NK	O	-
null	2	0	17	96	106	Commandant	Commandant	N	NN	masc|sg|	Invalide	CJ	B-PER_CORE	3
null	2	0	18	107	110	von	von	APPR	APPR	||	Commandant	PG	O	-
null	2	0	19	111	120	Marseille	Marseille	N	NE	masc|sg|	von	NK	O	-

Besten Dank!