



TECHNISCHE UNIVERSITÄT DARMSTADT
DIGITAL HUMANITIES COOPERATION
UNIVERSITÄT KONSTANZ

Thomas Weitin, Katharina Herget
Falkentopics

Pamphlet #4
November 2016

LitLab

Thomas Weitin, Katharina Herget

Falkentopics

Abstract

Topic Modeling gehört zu den vielversprechenden Verfahren für die quantitative Exploration semantischer Strukturen. Die Schöpfer der entsprechenden Algorithmen beschreiben die Anwendung im Hinblick auf große Textmengen, in denen versteckte thematische Zusammenhänge ermittelt werden, die das bloße Auge nicht wahrnehmen kann.

Wir haben Topic Modeling dagegen an einem mittelgroßen Korpus von Novellen erprobt, das sich wohl durch individuelle Lektüre als auch durch statistische Verfahren erschließen lässt. Angetrieben hat uns dabei eine bislang wenig beachtete Beobachtung: Diejenigen Wissenschaftler, denen es gelungen ist, statistische Verfahren bei der literarischen Korpusanalyse so angemessen einzusetzen, dass sie einschlägig werden konnten, kannten die verwendeten Korpora jeweils ausgezeichnet.

Weil wir über unser Novellenkorpus ebenfalls eine Menge wussten, war es uns möglich, die ermittelten Topics textrelevanten Themen zuzuordnen. Unter dem Stichwort ‚Falkentopics‘ beschreiben wir darüber hinaus eine andere Art von Topics, die den speziellen Plot einzelner Texte wiederzugeben scheinen.

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie. Sie ist in der Zeitschriftendatenbank (ZDB) und im internationalen ISSN-Portal erfasst. Detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar. Alle Rechte, auch die des auszugsweisen Nachdrucks, der fotomechanischen Wiedergabe und der Übersetzung, vorbehalten. Dies betrifft auch die Vervielfältigung und Übertragung einzelner Textabschnitte, Zeichnungen oder Bilder durch alle Verfahren wie Speicherung und Übertragung auf Papier, Transparente, Filme, Bänder, Platten und andere Medien, soweit es nicht §§ 53 und 54 UrhG ausdrücklich gestatten.

©2016 Thomas Weitin, weitin@linglit.tu-darmstadt.de
Katharina Herget, herget@linglit.tu-darmstadt.de

ISSN: 2364-5342

Falkentopics

I.

Die Semantik stellt für die digitale Literaturanalyse nach wie vor eine der größten Herausforderungen dar. Wenn der Mensch liest, wird er in den weitaus meisten Fällen intuitiv bedeutungsorientiert verfahren. Er erwartet einen Sinn und versucht, die Bedeutung des Gelesenen zu verstehen. Professioneller Literaturunterricht an der Schule oder an der Universität muss nicht selten diese hermeneutische Intuition zunächst einmal ausbremsen, um vor der Interpretation eine Analyse überhaupt stattfinden lassen zu können. Gerade vorschnelle Lesarten können als ein Indiz dafür gelten, wie wohl sich unser Gehirn mit Bedeutung fühlt. Für den Computer ist der Umgang mit Bedeutung dagegen ein großes Problem, weil sie sich regelmäßig als zu komplex erweist, um operationalisiert und berechnet werden zu können. Topic Modeling zählt zu den distant reading-Verfahren, die auf eine Berechnung von Bedeutung angelegt sind. Es handelt sich dabei um statistische Modelle, deren Algorithmen in großen Textkorpora latente semantische Strukturen sichtbar machen. Ermittelt werden Gruppen von Wörtern, deren gemeinsames Auftreten wahrscheinlich ist. Anhand dieser ‚Topics‘ erfahren wir etwas über die Themen der analysierten Texte, über Figuren, die darin vorkommen, über Handlungen und Handlungsorte. Ohne selbst zu lesen erhalten wir Einblick in inhaltliche Zusammenhänge. David Blei, dessen Beiträge das Topic Modeling in den Digital Humanities entscheidend popularisiert haben, bemüht das klassische *too big to read*-Argument, um das Einsatzgebiet des Verfahrens zu markieren: Große Textmengen, die das menschliche Lesevermögen bei weitem übersteigen, werden auf versteckte thematische Informationen hin untersucht, die für das Auge des einzelnen Lesers unsichtbar bleiben.¹ Nur mit gehörigem Abstand, bei der statistischen Analyse sehr vieler Texte, zeichnen sich diese „hidden structures“² ab.

Breit praktiziert wird das Topic Modeling bislang in den quantitativ arbeitenden Sozialwissenschaften, vor allem in der Politikwissenschaft. Hier finden sich Textsorten und Korpora, die speziell geeignet scheinen, Bundestagsreden etwa oder Protokolle der Beratungen von Exekutive und Legislative.³ Ein zweiter Schwerpunkt ist die Analyse digitaler Kommunikation, etwa von E-Mails oder in

¹„While more and more texts are available online, we simply do not have the human power to read and study them [...]. To this end, machine learning researchers have developed probabilistic topic modeling, a suite of algorithms that aim to discover and annotate large archives of documents with thematic information.” (David Blei: Probabilistic Topic Modeling. In: Communications of the AMC 55.4, S. 77)

²Ebd., S. 79.

³Das Schweizerische Bundesarchiv in Bern etwa wertet derzeit in einem groß angelegten Projekt die Topics der Beratungen des Bundesrates, der Schweizerischen Regierung, aus. Und auch die politischen Agenden

sozialen Medien wie Facebook und Twitter. Auch hier scheinen sich die Textkorpora für die Topic Modeling Algorithmen besonders anzubieten. Zwei der drei populärsten Topic Modeling packages für die statistische Software R gehen auf Forscher zurück, die im Kontext von Facebook arbeiten.⁴ Für die digitale Literaturwissenschaft ist das Topic Modeling mit einem großen Versprechen verbunden. Zu Beginn der Renaissance quantitativer Verfahren in den Philologien ist es vor allem stilometrischen Analysen gelungen, auch die Aufmerksamkeit traditionell arbeitender Kolleginnen und Kollegen zu erregen. Da diese Analysen meist auf Worthäufigkeiten beruhen, entsteht bei Kritikern regelmäßig der Eindruck, digitale Literaturwissenschaft erschöpfe sich im Zählen von Wörtern und dringe zu dem, was Leserinnen und Leser am meisten interessiert, gar nicht vor, zum Inhalt und zur Bedeutung von Texten. Matthew Jockers beginnt daher in seinem Buch *Macroanalysis* das Kapitel zum Topic Modeling, indem er suggeriert, die semantisch orientierte Methode könne am Ende „plot, character, and theme“⁵ erreichen. Allerdings konzentriert sich das Kapitel ausschließlich auf die „themes“, die klassischerweise mit Topic Modeling ermittelt werden. Es wird demonstriert, dass sich das Verfahren sowohl für die Einzeltext- als auch für die Korpusanalyse eignet. Und in vergleichenden Korpusanalysen gelingt es Jockers, Topic Models zur Klassifikation von Texten nach den Variablen ‚Gender‘ und ‚Nationalität‘ einzusetzen. Die Ergebnisse zeigen, was man von „macro trends“ vielleicht befürchtet hat, es treten ausschließlich Klischees zutage: Männer schreiben am liebsten über „pistols and other guns“, Frauen hingegen über „female fashion“, die Schotten sind mit ihrem „dialect“ beschäftigt, die Amerikaner mit „us dollars and us cities“ und die Briten selbstredend mit „hounds and shooting sport“.⁶ Jockers weist auf die Bedeutung individueller Outlier (immerhin 33 Prozent bei Nationality und 14 Prozent bei Gender), betrachtet sie allerdings nicht im Kontext der verwendeten Methode. In neueren Arbeiten konzentriert er seine thematischen Analysen sehr stark auf den Bereich der Emotion, wobei das dafür in *Macroanalysis* noch verwendete Topic Modeling⁷ zugunsten einer wörterbuchbasierten Sentiment Analysis aufgegeben worden ist.⁸

des Europäischen Rates wurden schon mithilfe des Topic Modelings betrachtet (vgl. Derek Greene, James P. Cross: Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach).

⁴Nämlich auf Johnathan Chang und David Blei selbst (vgl. Matthew L. Jockers: Text Analysis with R for Students of Literature. Cham u.a. 2014, S. 136).

⁵Matthew L. Jockers: *Macroanalysis. Digital Methods & Literary History*. Urbana u.a. 2013, S. 118.

⁶Ebd., S. 149-153.

⁷Vgl. ebd., S. 136.

⁸Mit seinem „syuzhet“ package für R hat Jockers eine heftige Netz-Diskussion ausgelöst (vgl. hierzu: Anni Swafford: Problems with the Syuzhet Package, Blogpost vom 2. März 2015 <https://annieswafford.wordpress.com/2015/03/02/syuzhet/> und Matthew L. Jockers: Some thoughts on Annie's thoughts... About Syuzhet, Blogpost vom 4. März 2015 <http://www.matthewjockers.net/2015/03/04/some-thoughts-on-annies-thoughts-about-syuzhet/>).

Was genau wollen und können wir mit Topic Models herausfinden? Auf der Ebene des Einzeltex-tes scheint die Antwort ebenso klar wie bei großen Korpora. Im Einzeltext können wir uns zum Beispiel vergegenwärtigen, welche Themen in welchen Kapiteln eines Romans dominant sind. Und in großen Korpora können wir verborgene thematische Strukturen sichtbar machen. Im ersten Anwendungsgebiet leistet die Methode etwas, das wir auch lesend leisten können. Im zweiten greift sie dort, wo die individuelle Lektüre nicht hinreicht, und tritt an deren Stelle. Natürlich kann man ein distant reading Ergebnis aus dem Topic Modeling durch gezieltes close reading evaluieren und dann zum Beispiel feststellen, dass ein bestimmter männlicher Autor aus dem Korpus von Jockers ganz dem Makrotrend entsprechend Räuberpistolen erzählt, während eine Autorin vom Trend abweichend völlig andere Themen bedient als „female fashion“. Beim Einsatz statistischer Methoden in der Literaturwissenschaft treten häufig zwei Denkgesten gepaart auf: Man freut sich, wenn man zeigen kann, dass eine Methode funktioniert, was in der Regel über die Bestätigung einer bereits vorhandenen Erkenntnis oder durch ein vom Textkorpus her erwartbares Ergebnis gelingt: Amerikaner haben tatsächlich andere Themen als Engländer! Und man erklärt gleichzeitig, gerade die Abweichungen vom Makrotrend seien doch interessant.

Statistisch gehören die Outlier so zuverlässig zum Makrotrend wie die Standardabweichung zum normalverteilten Datensatz. Literaturwissenschaftlich erinnert das Denkgestenpaar an den alten hermeneutischen Wunsch, das Verstehen von Einzelem und Ganzem möge sich wechselseitig erhellen. Dazu reicht es aber nicht aus, die Outlier interessant zu finden. Und wenn ich in der Lektüre eines einzelnen Textes, der statistisch abweicht, meine Gründe für diese Abweichung zu entdecken, stellt sich immer die Frage, ob der hermeneutische Einzelbefund und das Ergebnis der statistischen Korpusanalyse überhaupt etwas miteinander zu tun haben. Stellen wir uns vor, dass ich mir aus dem Jockers-Korpus den Text einer Autorin greife, der den gemessenen topic trends für female writers sehr stark widerspricht. Und ich finde heraus, dass diese Autorin im bürgerlichen Beruf eine Physikerin war, die mit „female fashion“ zufällig nichts am Hut hatte. Dann wäre in diesem Fall die interpretative Verbindung von historischem Einzeltextverständnis und Topic Modeling des Korpus korrekt. Es spricht überhaupt nichts dagegen, eine einzeltextbezogene These derart mit quantitativer Evidenz zu erhärten, schon weil man den Umgang mit allen quantitativen Verfahren sehr viel praktizieren und üben muss, wozu konkrete Fragen am stärksten motivieren. Gerade wer aber dann Übung darin hat, wird nach Wegen suchen, Methoden wie das Topic Modeling nicht mehr nur deskriptiv, sondern auch explorativ einzusetzen. Wir wollen uns nicht nur Topic-Anteile und Topic-Verteilungen anschauen, wir wollen dadurch etwas herausfinden über Beziehungen und Relationen zwischen Texten und Autoren mit ih-

ren jeweils mannigfaltigen Eigenschaften. Es geht uns um Zusammenhänge innerhalb des modellierten Korpus und darüber hinaus. Bei der Ermittlung solcher Zusammenhänge hilft es ungemein, wenn wir etwas über die Texte und Autoren wissen, auch wenn wir die Texte nicht einzeln durchlesen können. Insofern ist das auf den ersten Blick so einleuchtende *too big to read*-Argument zum Teil irreführend. Aus der Perspektive von Facebook-Analitikern mag die Emergenz von *hidden structures* aus Datenmengen, die nur statistisch auszuwerten sind, das richtige erkenntnistheoretische Leitbild sein. Wir Literaturwissenschaftlerinnen und Literaturwissenschaftler aber haben eine andere Erkenntnispraxis. Wir untersuchen in den meisten Fällen Korpora, über die wir eine ganze Menge wissen, sei es durch die Kenntnis bestimmter Texte, durch historisches Kontextwissen oder durch systematische Kenntnisse über Gattungen, Formen, Intertextualität usw. Diejenigen Wissenschaftler, denen es gelungen ist, statistische Verfahren bei der literarischen Korpusanalyse so angemessen einzusetzen, dass sie einschlägig werden konnten, kannten die verwendeten Korpora jeweils ausgezeichnet. Das gilt zum Beispiel für John Burrows, der das nach ihm benannte *Delta*-Verfahren zur Bestimmung der stilometrischen Distanz zwischen Texten an einem Korpus mit 25 englischen Autoren des 17. Jahrhunderts erprobte,⁹ oder für Matthew Jockers, der sein Korpus irisch-amerikanischer Literatur dank überragender Kenntnisse in dieser speziellen Literaturgeschichte zum idealen Studienobjekt von *Macroanalysis* machen konnte. Im Darmstädter LitLab haben wir *Topic Modeling* am Korpus des *Deutschen Novellenschatzes* erprobt, einer von Paul Heyse und Hermann Kurz herausgegebenen Sammlung von insgesamt 86 Novellen. Die 24 Bände dieser Sammlung sind zwischen 1871 und 1876 erschienen. Wir haben sie digitalisiert, als Korpus im *TXT* und *TEI-XML*-Standard aufbereitet (*corrected OCR*) und mit Metadaten zu den einzelnen Texten und deren Autorinnen und Autoren angereichert. Dabei haben wir viel über unser Korpus gelernt. Es handelt sich um eine Gattungsanthologie, die sich als „Mustersammlung“¹⁰ mit klar realistischem Epochenbewusstsein versteht. 49 der 86 Novellen stammen aus der Zeit nach 1848, die beiden ältesten von 1811, die jüngste von 1875. Versammelt sind fast ausschließlich Einzeltexte, nur vier Autoren sind mit zwei Texten vertreten. 12 weiblichen stehen 70 männliche Autoren gegenüber.

Ein Vorteil, den wir bei der Arbeit mit diesem Korpus schätzen gelernt haben, liegt in seiner mittleren Größe. Es ist noch klein genug für unsere individuelle Lesekompetenz, aber doch schon so groß, dass auch statistische Analysen möglich und sinnvoll sind. In einer großen *Excel*-Tabelle sammeln wir kontinuierlich neue Daten und Metadaten, wobei die Impulse dafür häufig aus der Lektüre einzelner Texte kommen. Neben den Novellen spielen dabei auch literaturgeschichtlich relevante Texte aus dem

⁹Vgl. John F. Burrows: ›Delta‹: a Measure of Stylistic Difference and a Guide to Likely Authorship, in: *LLC* 17 (2002), S. 267–287.

¹⁰Paul Heyse, Hermann Kurz: „Einleitung“. In: Dies. (Hrsg.): *Deutscher Novellenschatz*. München 1871, S. 24.

Kontext des Novellenschatzes eine wichtige Rolle, wenn es darum geht, die Funktion der Sammlung als Instrument der Literaturgeschichtsschreibung zu verstehen. Wir konsultieren zum Beispiel die Einleitung, die Paul Heyse zum ersten Band des *Deutschen Novellenschatzes* verfasst hat. Heyse skizziert darin eine realistische Gattungspoetik der Novelle, die das Verhältnis von Thema und Form als Alleinstellungsmerkmal zu bestimmen versucht. Mit Blick auf die geläufige Abgrenzung zum Roman verwirft Heyse das „Längenmaß“ als untaugliches Kriterium, vielmehr müsse „im Thema“ etwas liegen, „das mit Nothwendigkeit zu der einen oder andern Form hindrängt“.¹¹ Heyse glaubt, dass „eine so einfache Form“ wie die Novelle „nicht für jedes Thema unseres vielbrüchigen modernen Culturlebens“ geeignet ist.¹² Anstatt jedoch den Zuständigkeitsbereich der Gattung inhaltlich einzugrenzen, wird der Unterschied zum Roman über die Heuristik der Novelle definiert, das heißt über die speziellen Verfahren zur Reduktion von Komplexität, die nur der Novelle eigen sind.¹³ Während der Roman als Reflexionsgattung seine Themen „erschöpfend von allen Seiten“ beleuchtet, „verdichtet“ die Novelle alles „auf einen Punkt“.¹⁴ Die Themenbearbeitung der Novelle beruht auf der „Isolierung des Experiments“, was poetologisch bedeutet, sich auf ein „Grundmotiv“ zu konzentrieren und „eine starke Silhouette“ auszubilden. Ob das gelingt, kann Heyse zufolge in einem „Versuch“ überprüft werden, indem man nämlich testet, ob es gelingt, „den Inhalt in wenige Zeilen zusammenzufassen“.¹⁵

Diese empirisch ausgerichtete Gattungspoetik ist als ‚Falkentheorie‘ in die Literaturgeschichte eingegangen, weil Heyse seinen Test anhand jener Novelle Boccaccios vorführt, in der ein aus Liebe geopferter Falke die Verdichtungsqualität eines Grundmotivs besitzt, über das sich die Novellenhandlung tatsächlich in kurzen Sequenzen zusammenfassen lässt. Der ‚Falkentest‘ wird als zentrales Auswahlkriterium angegeben, nach dem die Texte für den *Deutschen Novellenschatz* ausgewählt worden sind. Schriftstellern, die Novellen produzieren wollen, wird ausdrücklich empfohlen, ihren Stoff zuallererst diesem Test zu unterziehen und zu fragen, „wo ‚der Falke‘ sei“.¹⁶

Wir haben für das Topic Modeling des Novellenschatzes das R package MALLETT innerhalb von R-Studio verwendet.¹⁷ Da wir beim Topic Modeling systematisch nach inhaltlichen Zusammenhängen suchen, muss das Korpus zunächst von Funktionswörtern und denjenigen Information tragenden

¹¹Ebd., S. 17.

¹²Ebd., S. 20.

¹³Vgl. Thomas Weitin: Heuristik der Novelle. In: Albrecht Koschorke u.a. (Hg.): *Komplexität und Einfachheit*. Stuttgart 2017 (im Erscheinen).

¹⁴Paul Heyse, Hermann Kurz: „Einleitung“, S. 18 (erstes Zitat), 17.

¹⁵Ebd., S. 18 (erstes Zitat), 19.

¹⁶Ebd., S. 20.

¹⁷Die im Folgenden verwendeten Analysen gehen zurück auf die Masterarbeit von Katharina Herget: *Die Literaturgeschichtsschreibung des Deutschen Novellenschatzes Paul Heyses: Qualitative und quantitative Perspektiven*. Typoskript. Konstanz 2015.

Wörtern gereinigt werden, deren Häufigkeit die Ergebnisse dominieren würde (Beispiel: ‚haben‘). Wir haben eine vorhandene Stoppwortliste für das Deutsche für unsere Zwecke angepasst, wobei vor allem eine große Zahl der Namen berücksichtigt worden ist, die bei ersten Analyseversuchen mit dem Korpus sehr stark in den Topics präsent gewesen waren. Grundsätzlich geht Topic Modeling von zwei Annahmen aus:¹⁸ Es gibt eine bestimmte Zahl von gemeinsam verwendeten Wörtern, deren gemeinsames Auftreten sich in Texten wie ein Muster regelmäßig wiederholt. Das sind die Topics. (1) Jeder einzelne Text in einem Korpus lässt sich danach beschreiben, wie stark jedes dieser Topics in ihm präsent ist und welche Wörter jeweils dazu gehören. (2) Unter dieser Voraussetzung arbeitet der Algorithmus mit einem bag of words aller im Korpus verbliebenen Wörter und verteilt sie so lange zufällig auf die vorab definierte Zahl von Topics, bis sich ein gemeinsames Auftreten stabilisiert. Die beobachteten Muster werden mit dem tatsächlichen Auftreten der Topics in den Texten abgeglichen. Gemessen werden also die Topics, die Topic-Anteile im Text und die Zugehörigkeit von Wörtern in einem Text zu den Topics.

Dem entsprechend sind die Korpusprämissen gestaltet, die bei der Arbeit mit dem MALLET package festgelegt werden müssen und deren Festlegung das jeweilige Modell darstellt, mit dem das Experiment arbeitet. Ist die Stoppwortliste bestimmt, kommt es darauf an, drei Parameter zu optimieren: die Anzahl der Topics, die Anzahl der Wörter pro Topic und die Anzahl der Iterationen bis zur erhofften Stabilisierung der Ergebnisse. Stabil ist ein Modell dann, wenn die kookkurrenten Wortgruppen, die es als Topics produziert, sich wiederholen lassen. Es geht um die Zuverlässigkeit (reliability) des Modells. In unseren Versuchen mit den 86 Texten des Novellenschatz-Korpus hat ein Modell mit 100 Topics, 10 Wörtern pro Topic und 10000 Iterationen relativ zuverlässige Ergebnisse gebracht. 57% der Topics erwiesen sich als reproduzierbar — allerdings nicht sofort. Um das zu erreichen, mussten wir eine wesentliche Veränderung im Experiment vornehmen. Auf der Basis der 86 Einzeltexte unserer Sammlung war trotz erheblicher Mühe mit ganz verschiedenen Einstellungen der drei Parameter Topicanzahl, Wörter pro Topic und Iteration kein reproduzierbares Ergebnis zu erreichen. Dafür mussten wir das Korpus erst so präparieren, dass der Algorithmus nicht mehr den Einzeltext als Grundeinheit, sondern automatisch generierte chunks von je 300 Wörtern verarbeitet, insgesamt 5380 einzelne Dokumente. Nach diesem preprocessing funktionierte unser Modell dann relativ stabil.

Bekanntermaßen lautet eine Grundformel der Novelle $2 + 1$: Zwei sind einander bestimmt, ein Dritter stört, es kommt zum Konflikt mit gutem oder schlechtem Ende. Hannelore Schläffer hat das in Anlehnung an Gottfried Keller die „Novellenmathematik“¹⁹ genannt. Aus der Einzeltextlektüre des

¹⁸Vgl. Blei, S. 77f.

¹⁹Vgl. Hannelore Schläffer: Poetik der Novelle, Stuttgart 1993.

Novellenschatzes wussten wir, dass es sich bei einer überwältigenden Mehrheit der enthaltenen Texte um solche Heiratsnovellen handelt. Zählt man Ehenovellen, bei denen nicht der Weg zur Hochzeit, sondern das nachherige Eheleben den Plot bildet, mit dazu, liegt der Anteil bei über 90 Prozent. Dementsprechend wunderten wir uns nicht, dass unter den 100 Topics unseres Modells²⁰ entsprechende semantische Muster auftraten:

Heirat

- 12 müllerin heirathen jurancon verwandtschaft base müller sich's stiefvater verlobung schwager
- 13 onkel großmutter julie oberförster brink traueung liebes berger brigitten vollends
- 17 wirthin baron gegend thurn freifrau braut lich fräulein lothar schlosse
- 22 herzog reise advocat durchlaucht baron studenten polen kurland bursch hochzeit
- 40 mädchen signora urballa burschen signor heirathen villa padrona signoria baldo

Von unserer Lektüre her hatten wir allerdings erwartet, dass sich dieses absolut dominante Thema noch deutlicher ausprägen würde.²¹ Immerhin ist in den meisten Novellen das Heiraten ein großes Thema. Wir sehen natürlich, dass unsere Topics Schönheitsfehler haben: „sich's“ hätte unter die stop-words gehört, „lich“ ist offensichtlich ein Fehler, und es sind auch einige Namen stehengeblieben. Ein Vergleich untereinander zeigt, dass die Topics semantisch und von den versammelten Wortarten her unterschiedlich einschlägig für das Thema Heiraten sind. Einschlägig ist sicher Topic 12 mit „verlobung“, „verwandtschaft“, „schwager“ und „heirathen“. Ähnlich Topic 40 mit „mädchen“, „burschen“ und „heirathen“. Die anderen drei Topics enthalten jeweils nur ein einschlägiges Substantiv (13: „traueung“, 17: „braut“, 22: „hochzeit“), 17 und 22 sind überhaupt reine Substantivtopics. Entscheiden wir uns, diese fünf Topics als Heiratstopics zu verstehen, können wir sie mit anderen thematischen Clustern vergleichen, die sich aus unserem Modell mit 100 Topics bilden lassen. Es handelt sich dabei um die Themenbereiche Militär, Religion, Justiz, Künste, Vermögen und Bauern, denen sich jeweils mindestens drei Topics zuordnen lassen.

²⁰Vgl. Anhang, S. 16f.

²¹ Die absoluten Häufigkeiten einschlägiger Hochzeitsbegriffe im Korpus haben die vom Leseindruck herrührende Erwartung gestärkt. Von den rund 1,6 Mio Tokens unseres Gesamtkorpus entfallen immerhin mindestens 1309 auf einen ganz eng definierten Hochzeitsthesaurus: hochzeit* (260), braut* (333), heirat* (515), traueung* (37), bräutigam* (125), vermählt* (39).

Militär

20 offizier graf nacht pferde offiziere pferd husar junge husarenoffizier hauptquartier
66 chef lieutenant freund ehre freunde gefährten ring flagge offiziere freundes
76 major haide pferde ritten herum uwar maroshely gitter ritt lande
78 teufel general soldaten könig hauptmann commandant offizier dienst oberst wache
98 franzosen krieg frankreich könig männer vaterland volk kaiser feind französischen

Religion

15 gott ehre grab alte schwert herzog gnade begraben schönen letzten
60 heiligen nonne bild kloster antlitz kunst kirche malen bildes entgegnete
76 major haide pferde ritten herum uwar maroshely gitter ritt lande
91 vermittlung gemeinde klöster gesunden gottlosen zustand dünnen gebilde gesunde eingebüßt

Justiz

10 silberburg rothenburg stadt scharfrichter wolf heyliger thal römerhöhe garten rath
21 müller mühle schmied justizrath actuar gerichtsdienner knabe meister schmiede müllers
70 frei schuldig staatsanwalt rothmann geschworenen verhandlung schwiegersohn
angeklagten männer vorsitzende
92 richter stadtschreiber galgen gericht urtheil herren rath strafe freiheit zuletzt

Künste

43 brief las schrieb lesen briefe schreiben geschrieben buch papier gelesen
45 novelle geschichte namen form roman dichter stoff meister erzählungen krieges
47 alte bild maler kunst künstler freund bilder gemälde fremde junger
54 kennen natur lernen nennen bildung kunst geist meisten höheren erscheint
83 lied musik singen spielen sang talent spielte geige klavier sängerin
100 tanz tanzen tanze vicomte tanzte jungen musik mädchen musikanten altenkreuz

Vermögen

50 schatten gold herrn grauen sonne tasche goldes seckel pracht vieles
61 geld gulden thaler handel summe wasser schulden littauer wirth verkauft
67 ring finger musik eltern palast schöne niemals kästchen stehen gewaltigen
86 fräulein sprach schmuck meister könig arbeit that kästchen geheimniß paris
96 kästchen gesellschaft wagen zimmer schöne beutel geld wodurch leben vergnügen

Bauern

8 bauer gott großmutter alte magd kind heiligen knabe worte seele
68 jäger bauern mühle förster bauer wald thal hütte jüngling felsen
74 acker kinder bauern puppe nase steine jahren wilde aecker loch

Wir sehen im Überblick, dass die Heiratstopics in etwa gleich stark sind wie die anderen Themenbereiche. Der Eindruck der absoluten Dominanz des Themas, den wir in der Lektüre gewonnen haben, bestätigt sich also nicht. Durchaus aber entspricht die Gesamtschau der Topics dem, was wir über un-

ser Korpus und dessen einzelne Texte wissen. Heiraten und Familiengründung sind in den realistischen Novellen meist mit Vermögensfragen verbunden. Es ist das Leben einfacher Menschen (Bauern), das in den im Korpus gut vertretenen Dorfgeschichten geschildert wird. Religion spielt eine große Rolle, es gibt ein starkes Spannungsfeld von Aberglaube und Säkularisierung. Viele Konflikte werden aber auch vor der Justiz behandelt. Gleichzeitig geht es um den Abstieg des Adels und den Aufstieg des Bürgertums. Militärische Karrieren gibt es auf allen sozialen Ebenen, sie haben auch mit sozialer Mobilität zu tun. Es gibt einzelne Künstlernovellen, die starke Präsenz dieses Topics kommt aber für Leserin und Leser etwas überraschend.

Diese Zusammenschau von Topic Modeling, Lektüre und literaturgeschichtlichem Wissen könnte uns schlussfolgern lassen, dass unser Modell funktioniert. Die statistischen Topics, die der Algorithmus produziert hat, sind relativ konsistent und wir können sie relevanten Themen zuordnen. Unsere historisch informierte Inhaltsdeutung und die statistische Korpusanalyse scheinen zu korrespondieren.

Aber ist das tatsächlich so? Oder besser gefragt: Wie angemessen ist es, die Ergebnisse eines statistischen Topic Models derart hermeneutisch zu ordnen und zu deuten? Und wohin führt das? Betrachten wir ein Beispiel. Dem Themenbereich Vermögen haben wir diese beiden Topics zugeordnet:

67 ring finger musik eltern palast schöne niemals kästchen stehen gewaltigen
86 fräulein sprach schmuck meister könig arbeit that kästchen geheimniß paris

Wenngleich „ring“ und „palast“ in Topic 67 und „schmuck“ mit „könig“ in Topic 86 jeweils in Kombination mit „kästchen“ als Vermögensindikatoren verstanden werden können, wurde unsere Zuordnungsentscheidung womöglich durch ein anderes Topic beeinflusst. In Topic 96 tritt „kästchen“ gemeinsam mit dem einschlägigen Nomen „geld“ auf:

96 kästchen gesellschaft wagen zimmer schöne beutel geld wodurch leben vergnügen

Wie schon bei den Heiratstopics könnte man sagen, dass einzelne Topics unterschiedlich einschlägig für ein Thema sind. Wie wir solche Zuordnungen vornehmen, das ist letztlich eine Frage der Interpretation.

Als wir die Ordnung unserer Topics diskutierten, hatten wir neben der Literaturgeschichte des Realismus zwei Dinge im Hinterkopf. Wir dachten daran, was der Topic Modeling Algorithmus berechnet: Topics im Korpus, Topic-Anteile im Text und die Zugehörigkeit von Wörtern in einem Text zu den Topics. Und wir dachten an die Herausgeber des Novellenschatzes, die in die Sammlung nur Texte hatten aufnehmen wollen, die den ‚Falkentest‘ bestehen, also über ein starkes Grundmotiv verfügen, anhand dessen sich der Inhalt leicht zusammenfassen lässt. Da wir die Texte unseres Korpus

kannten und bestimmte Namen und Ortsbezeichnungen stehen geblieben waren, fiel uns zu einigen Topics sofort eine passende Novelle ein. In diesen Fällen fingen wir an, den Test zu machen und probierten aus, ob sich tatsächlich alle Wörter des betreffenden Topics in der Novelle finden würden. Das Ergebnis erstaunte uns: Bei 32 der 100 Topics stammten tatsächlich alle 10 Topic-Wörter aus ein und derselben Novelle. Goethes *Neue Melusine*, die Novelle, die die Herausgeber an den Anfang des Novellenschatzes gesetzt haben, weil Heyse sie als Stilvorlage unentbehrlich fand, bildet als einzige sogar zwei vollständige Topics aus.

Wir waren irritiert. Um verlässliche Ergebnisse zu erzielen, hatten wir die Einheit der Novellen aufgelöst und stattdessen 5380 chunks von je 300 Wörtern zur Ermittlung der Topics verwendet. War es da nicht sehr unwahrscheinlich, dass in fast einem Drittel der Fälle die Wörter eines Topics auch die eines bestimmten Textes waren? Einen Namen hatten wir natürlich sofort für unsere Beobachtung. Paul Heyse zu Ehren fingen wir an, von Falkentopics zu sprechen. Bildeten sie eine semantische ‚Silhouette‘ aus, so wie Heyse sich das vorgestellt hatte? Konnten wir den charakteristischen Handlungsgang der Novellen diesen Topics entnehmen? Schauen wir uns ein paar Beispiele an:

Falkentopics

Tieck: Des Lebens Überfluss

65 treppe polizei holz ofen stufen lachen emmerich sieh niemals feuer

Goethe: Die neue Melusine

67 ring finger musik eltern palast schöne niemals kästchen stehen gewaltigen

96 kästchen gesellschaft wagen zimmer schöne beutel geld wodurch leben vergnügen

E.T.A. Hoffmann: Das Fräulein von Scudery

86 fräulein sprach schmuck meister könig arbeit that kästchen geheimniß paris

Kleist: Die Verlobung in St. Domingo

72 fremde neger fremden alte familie knaben herrn weißen bette babekan

Schmid: Mohrenfranzel

89 königin spielen burschen stück saba bursche schiffer schwarze bühne zunftmeister

Roquette: Die Schlangenkönigin

48 kahn wasser bild schlangenkönigin dorfe schlangen ufer entgegen gegend leipe

In Topic 65 erkennt man sofort den ‚Falken‘ der Tieckschen Novelle *Des Lebens Überfluss*, in der ein Liebespaar ökonomischer Not trotzend die „Treppe“ zur Außenwelt verbrennt, um sich ganz in die romantische Zweisamkeit zurückzuziehen. Die Topics 67 und 96 enthalten beide das „Kästchen“, in dem die Miniaturwelt der schönen Zwergenprinzessin in Goethes *Neuer Melusine* samt Palast transportiert wird. Auch in Topic 86 ist ein „Kästchen“ das Dingsymbol. Bei E.T.A. Hoffmann transportiert es den für die adligen Empfänger höchst unheilvollen Schmuck des Goldschmieds René Cardillac (*Das Fräulein von Scudery*). Kleists Haiti-Novelle *Die Verlobung in St. Domingo* profiliert ihre historisch-dramatische ‚Silhouette‘ ebenso im Konflikt um *race* und „Fremde“ wie Hermann Schmidts Novelle *Mohrenfranzel* über eine von Schiffern betriebene Wanderbühne in der bayerischen Provinz, deren einzige schwarze Darstellerin die Königin von Saba spielt. Schließlich bringen die „Schlangen“ in Otto Roquettes Spreewaldnovelle *Die Schlangenkönigin* die misogynen Sexualitätsmetaphorik des Textes auf den Punkt.

Wer einmal daran glaubt, dass die Novellen, die solche Falkentopics ausbilden, innerhalb des Korpus tatsächlich herausstechen und auf besondere Weise unterscheidbar sind, ist versucht, dem statistischen Phänomen weitere Erklärungslasten aufzubürden. Etwa in Verbindung mit der Beobachtung, dass vier von fünf Topics im Themenbereich Vermögen Falkentopics bilden, während sich unter den Themen Religion und Künste überhaupt keine Falkentopics finden. Schaut man sich die entsprechenden Topics genauer an, scheint der hermeneutische Eindruck dem zu entsprechen. Gegenüber den starken Falken im Vermögensbereich rufen etwa die folgenden beiden Beispiele aus Religion und Künste auch beim kundigen Leser keinen bestimmten Text auf:

81 kirche pfarrer sonntag heilige messe gemeinde priester geistliche heiligen predigt

45 novelle geschichte namen form roman dichter stoff meister erzählungen krieges

Wir wollen indes der Spekulation an dieser Stelle Einhalt gebieten und lieber versuchen, herauszufinden, ob der vermeintliche Fund nicht eher ein systematischer Fehler oder zumindest Ergebnis der Parametermanipulation unseres Modells ist. Eine Überlegung wäre, dass 100 Topics einfach zu viel sind für die Größe unseres Korpus, so dass sich die gemeinsam auftretenden Wörter deshalb innerhalb ein und desselben Textes finden.²² Verlangen wir dem Algorithmus weniger Topics ab, so müsste die Überlegung weitergehen, verteilen sich die mit Wahrscheinlichkeit auftretenden Wortmuster über meh-

²²Dank an Sabine Bartsch (Darmstadt), die uns auf diese Idee gebracht hat.

rere Texte und die Falkentopics verschwinden. Wir haben die Anzahl der Topics schrittweise reduziert, dabei aber diese Vermutung nicht bestätigt gefunden. Die Falkentopics gehen zwar mit sinkender Topiczahl zurück, sie verschwinden aber nicht ganz. Waren es bei 100 Topics noch 32, finden sich unter 20 Topics immerhin noch zwei Falken. Das 20er-Modell gibt folgende Topics aus:

1	frau vetter munde conrectorin vetterchen that schäferle stadt general buchenberg
2	mynheer see chef schiff boot lieutenant meister zorgenhof freimeister amsterdam
3	hunde bloß major herren pferde dachte notar junker koppigen wilde
4	könig königin herzog stadt ehre soldaten franzosen officier königs könige
5	schatten ring kästchen fräulein schmuck arbeit könig meister paris arzt
6	wagen pferde pferd kutscher dame klarinett reiter graf burschen signora
7	pfarrer kirche berg bauer pfarrerin großmutter burg magd schwester knabe
8	sonne wasser erde himmel luft wald straße garten schatten land
9	zeit ließ hause tage jungen tag wußte fand that tochter
10	müller küster mühle schmied stadt justizrath silberburg herrn rothenburg hedeper
11	landrichter jäger lappen förster felsen fjord kaufmann missionär thal hütte
12	herr marquis doctor herrn hahn stadt professor todten gast fräulein
13	rief vater alte herr mann frau fort sagen sehen mädchen
14	hand augen stand schien nacht gesicht hielt trat lag weg
15	dorf arbeit bauern hof juden jude littauer feld acker wasser
16	wein sprach tisch glas trinken gäste trank keller herren rose
17	buch prinzessin geschichte london novelle werth bildung wirklichkeit deutsche gelesen
18	bild kunst maler nonne künstler musik heiligen spielen bilder bilde
19	leben liebe herz welt menschen seele glück herzen fühlte thränen
20	graf frau baron kind vater fräulein gräfin grafen marquis frug

Vergleichen wir das 20er- mit dem 100er-Modell,²³ fallen vor allem systematische Gemeinsamkeiten auf. Wir finden in beiden sowohl Falkentopics (rot) als auch thematische Topics (grün). Der Anteil der Falkentopics reduziert sich wie gesagt von rund einem Drittel (32/100) auf 10 Prozent (2/20), während der Anteil der thematischen Topics in beiden Modellen bei rund 40 Prozent liegt. Wir finden in dem stark reduzierten Modell nicht alle unsere Themen wieder, aber doch ausschließlich solche, die sich im viel größeren Modell auch finden, z.B. Topics zu den oben diskutierten Themen Künste (17 und 18), Religion (7) und Adel/Militär (4, 20) sowie zu Natur (8), Essen (16) und Liebe (19). Wir hatten gesehen, dass die entsprechenden thematischen Zuordnungen Interpretationssache sind. Systematisch wichtiger ist die Feststellung, dass sich semantische Kohärenz innerhalb der Topics durch die Reduktion der Topicanzahl nicht verbessert hat. Sie bleibt gleichwohl gut, und wir sehen erneut, dass thematische Topics unterschiedlich einschlägig für ihr Thema sein können oder, wie wir besser sagen, es mehr oder weniger kohärent ausdrücken. Das von der Systematik der Ergebnisse her sehr ähnliche, aber kompaktere Modell macht uns Vergleiche leichter. Gehen wir also dieser Beobachtung noch ein

²³Siehe Anhang, S. 17ff.

Stück weiter nach und vergleichen die thematischen Topics des 20er-Modells untereinander. Mit Blick auf die Topics Topics 16 und 19 könnte man versucht sein zu glauben, dass Kohärenz entsteht, wenn ein Topic neben Substantiven auch einschlägige Verben enthält.

- 16 wein sprach tisch glas trinken gäste trank keller herren rose
- 19 leben liebe herz welt menschen seele glück herzen fühlte thränen

Hier scheint es so, als sei die semantische Kohärenz auch ein Effekt möglicher syntaktischer Verbindungen, die durch die Kookkurrenz unterschiedlicher Wortarten ‚getriggert‘ werden. Man kann sich zu den beiden Topics ganz leicht Szenen vorstellen, die durch Sätze beschrieben werden können, die sich aus den Wörtern der Topics bilden lassen. Ein Vergleich mit Topic 8 belehrt uns indes darüber, dass auch rein substantivische Topics einschlägige Szenerien aufrufen können. Wir können uns zu den Begriffen von Topic 8 leicht einen Spaziergang durch die Natur vorstellen:

- 8 sonne wasser erde himmel luft wald straße garten schatten land

Wir haben im 20er-Modell neben unserer Unterscheidung zwischen Falkentopics und thematischen Topics noch zwei weitere Gruppen von Topics grau bzw. gelb markiert. Diese Topics stehen jeweils einer der Topicarten systematisch nah, ohne dazu gezählt werden zu können. Die gelb markierten gaben uns durch bestimmte Wörter, insbesondere durch Namen, Eigennamen und Ortsbezeichnungen, Anlass, sie für Falken zu halten. Ein Test ergab indes, dass sie nicht nur aus Wörtern ein und desselben Textes bestehen, auch wenn der Anteil von Wörtern, für die das gilt, jeweils beträchtlich ist. Die grau markierten Topics erscheinen semantisch sehr kohärent, ohne wirklich thematisch zu sein. Schauen wir uns diese Topicgruppe näher an:

- 9 zeit ließ hause tage jungen tag wußte fand that tochter
- 13 rief vater alte herr mann frau fort sagen sehen mädchen
- 14 hand augen stand schien nacht gesicht hielt trat lag weg

Ohne weiteres können wir aus diesen aus Verben, Substantiven und Adjektiven bestehenden Topics bedeutungsvolle Sätze bilden, und sie rufen auch bestimmte Szenen in uns wach. Die Zuordnung zu einem Thema aber bedürfte jeweils einer starken Interpretation, die zumindest bei den Topics 9 und 13 nicht allein aus dem Topic heraus zu rechtfertigen wäre. Topic 14 indes könnte mit seiner aufgerufenen nächtlichen Szenerie unter Umständen zu den thematischen Topics gezählt werden. Es ist ein Grenzfall.

Schauen wir uns jetzt unsere beiden verbliebenen Falkentopics an. Bei Topic 5 handelt es sich um einen alten Bekannten. Alle seine Wörter stammen aus E.T.A. Hoffmanns *Das Fräulein von Scudery*, das schon im 100er-Modell als Falke aufgefallen war.

E.T.A. Hoffmann: Das Fräulein von Scudery

Modell mit 100 Topics

fräulein sprach schmuck meister könig arbeit that kästchen geheimniß paris

Modell mit 20 Topics

schatten ring kästchen fräulein schmuck arbeit könig meister paris arzt

7 Topic-Wörter sind im Vergleich beider Modelle stabil. Und beide Varianten zeigen uns das Falkentopic als ein typisches seiner Art: Es enthält Hinweise auf die Figuren (René Cardillac wird in der Novelle beständig der „Meister“ genannt, die Scudery nicht nur im Titel „das Fräulein“), auf den Handlungsort („Paris“) und auf Gegenstände, die für den Plot charakteristisch sind („Schmuck“, „Kästchen“). Topic 15 ist der zweite Falke des 20er-Modells. Alle Wörter dieses Topics stammen aus Ernst Wicherts Novelle *Ansas und Grita*, die anders als die Hoffmann-Novelle im 100er-Modell noch kein Falkentopic ausgebildet hatte.

Ernst Wichert: Ansas und Grita

Modell mit 20 Topics

dorf arbeit bauern hof juden jude littauer feld acker wasser

Während die geografische Bezeichnung „Littauer“ für ein Falkentopic typisch ist und bei entsprechender Lektürekennntnis sofort an die betreffende Novelle denken lässt, ist die kohärente Ausprägung des Bauerntemas im Vergleich zu den anderen Falkentopics durchaus ungewöhnlich. Im Vergleich mit den oben im 100er-Modell dem Thema ‚Bauern‘ zugeordneten Topics enthält das Topic der Wichert-Novelle sogar deutlich mehr einschlägige Inhaltswörter. Dieses Topic ist also zugleich Falke und thematisch. Zur Veranschaulichung noch einmal die oben angeführten Falkentopics aus dem 100er-Modell:

Falkentopics (Modell mit 100 Topics)

Tieck: Des Lebens Überfluss

65 treppe polizei holz ofen stufen lachen emmerich sieh niemals feuer

Falkentopics bieten immer auch thematische Anhaltspunkte, die aber in der Regel in dem Maße schwächer und weniger kohärent sind, wie sie durch Namen, Eigennamen und Ortsbezeichnungen

Goethe: Die neue Melusine

67 ring finger musik eltern palast schöne niemals kästchen stehen gewaltigen
96 kästchen gesellschaft wagen zimmer schöne beutel geld wodurch leben vergnügen

E.T.A. Hoffmann: Das Fräulein von Scudery

86 fräulein sprach schmuck meister könig arbeit that kästchen geheimniß paris

Kleist: Die Verlobung in St. Domingo

72 fremde neger fremden alte familie knaben herrn weißen bette babekan

Schmid: Mohrenfranzel

89 königin spielen burschen stück saba bursche schiffer schwarze bühne zunftmeister

Roquette: Die Schlangenkönigin

48 kahn wasser bild schlangenkönigin dorfe schlangen ufer entgegen gegend leipe

unterbrochen werden. Dadurch verweisen sie eher auf den betreffenden Text und seine Handlung als (rein) auf ein Thema. Beim Topic Modeling als reinem distant reading fallen Falkentopics deshalb, so vermuten wir, einfach aus der Betrachtung heraus. Man braucht einen dezidiert auch am Einzeltext interessierten Ansatz in der Korpusanalyse, um sie überhaupt interessant zu finden. Matthew Jockers hat für namen- und eigennamenlastige Topics einen eigenen Term geprägt und spricht von ‚topical topics‘,²⁴ die er allerdings als bloße Verzerrung des erhofften Ergebnisses in Gestalt thematischer Topics begreift. Unsere Falkentopics sind topisch in ihrer doppelten Verortungsfunktion. Betrachten wir statistisch errechnete Topics hermeneutisch auf der Basis von Textkenntnis, sind es vor allem Namen, Eigennamen und Hinweise auf Orte, die uns helfen, den entsprechenden Text auch ohne Volltextsuche im Gesamtkorpus aufzufinden. Gleichzeitig verorten Falkentopics die Handlung der betreffenden Novellen, ihr Setting und ihren Plot.

Nimmt man die Idee topischer Topics ernst, führt das möglicherweise ins Zentrum aktueller methodischer Diskussionen um das Topic Modeling. Wir haben das große Glück, mit unserem Novellenschatz-Korpus Teil einer mit Topic Modeling befassten Forschergruppe von DARIAH zu sein. Mit den Teams von Gerhard Lauer und Simone Winko (Göttingen), Fotis Jannidis (Würzburg) und Peer Trilcke (Potsdam) diskutieren wir immer wieder kritisch die Frage, wie geeignet Topic Modeling für Literatur eigentlich ist. Die Würzburger Gruppe mit Christof Schöch und Steffen Pielström, die an der Verbesserung des Algorithmus arbeitet, denkt intensiv darüber nach, warum im sozial- und politikwissenschaftli-

²⁴Matthew Jockers: Text Analysis with R for Students of Literature. Cham 2014, S. 152.

chen Bereich, Topic Modeling Analysen regelmäßig glattgehen, während bei der Analyse literarischer Korpora ebenso häufig Probleme auftreten. Eine interessante Hypothese geht davon aus, dass die semantischen Felder literarischer Texte im Unterschied etwa zu Bundestagsreden oder zur Social Media Kommunikation viel stärker vom Setting der Texte überlagert werden, von Wörtern, die für die Handlung charakteristisch, aber nicht unbedingt thematisch sind. Nach unseren Experimenten wäre unsere vorläufige Position dazu: In topischen Topics tritt diese Überlagerung zutage. Was wir vom historischen Kontext unseres Novellenschatz-Korpus ausgehend als Falkentopics beschrieben haben, das kann man gleichermaßen als Folge unzureichender Parameteroptimierung und als statistische Spur von Literarizität verstehen.

Anhang

1	mäuse stephan hölzerne ratten höfen habicht ähnlichen topf paulus wahrhaftigkeit
2	weg nacht luft himmel sonne standen hinein schritte erde kamen
3	bertram süden spitze gipfel land osten stiefel gebüsch gänge inseln
4	pfarrer schnee hofrichter meßner förster köchin kirche warmen kellermeister stunden
5	blumen erde blume goldenen blätter gesegnet natur brunnen früchte zweige
6	nacht tag schlief bette bett uhr schlaf saß schlafen erwachte
7	schiff ufer see schiffer boot schiffe wasser wellen segel hamburg
8	bauer gott großmutter alte magd kind heiligen knabe worte seele
9	bäcker pastor baumann moor löffel schwieg busch kirche fritzens stube
10	silberburg rothenburg stadt scharfrichter wolf heyliger thal römerhöhe garten rath
11	frau graf kind grafen gräfin bruder dame verwalter frug schlosse
12	müllerin heirathen jurancon verwandtschaft base müller sich's stiefvater verlobung schwager
13	onkel großmutter julie oberförster brink traueung liebes berger brigitten vollends
14	zeit ließ schien hause alten tage wußte blieb fand that
15	gott ehre grab alte schwert herzog gnade begraben schönen letzten
16	hunde beute bloß koppigen junker herren desto dachte wild alten
17	wirthin baron gegend thurn freifrau braut lich fräulein lothar schlosse
18	wirth stiefeln netze kalender pantoffeln zusammenhalten dreifach schlingen verstecken kann's
19	rose wein sprach keller judas kellermeister apostel namen jungfer römer
20	offizier graf nacht pferde offiziere pferd husar junge husarenoffizier hauptquartier
21	müller mühle schmied justizrath actuar gerichtsdienner knabe meister schmiede müllers
22	herzog reise advocat durchlaucht baron studenten polen kurland bursch hochzeit
23	brunken maler schulmeister doctor ungeheuer kunst assessor bild freundes aufgewachsen
24	bäuerin särke dunklen blitzstrahl blitze gericht vorübergehen bärbele knopf damm
25	herr mann fort sagen frau lassen alte sehen fragte weiß
26	london grauen woche deutschen bettler england ohnedies selbstpeinigung ländlichen künftigen
27	marquis schloß paris herrn anstalt drachen zwerglein schlosse geschichte gott
28	marquis doctor professor alten jungen rom ghetto corso italien valencia
29	gewissenhaft agram damaligen ertönen heiligkeit violine romantischen ueber hauptstadt mühsam
30	landrichter erwiderte haus lappen kaufmann fjord missionär probst malanger henrik
31	küster hedeper frau küsterei wittwe meierin küsters herrn jahr schwägerin
32	leben liebe herz welt menschen seele glück herzen freude armen

33	tante notar spendvögtin glücklich freundin bloß julie marei jungfer städtchen
34	schreiber base schreibers ofen registranten keller syndicus züge erschöpfung untersten
35	dorf kessel milch schönsten dirnen schultheiß schultheißen pfanne brunnen streit
36	muhme muhme-lieutnanten striethast hanepich lene lebte erbin muhme-lieutnant-saloppel schwestern professor
37	vater kind jahre tochter haus jungen jahren kinder mädchen sohn
38	versetzte arzt wasser gräfin köln fremden ring schatulle dame felsen
39	könig königin herzog frau weib könige königs frug land volke
40	mädchen signora urballa burschen signor heirathen villa padrona signoria baldo
41	droben schwester rosine tante bruder meran bursch zehnuhrmesser hochwürden mal
42	wesen schönen gesellschaft gefühl verstand niemals gewissen schönheit leidenschaft natur
43	brief las schrieb lesen briefe schreiben geschrieben buch papier gelesen
44	vater hab laß weißt ist's komm geh kind ich's lieb
45	novelle geschichte namen form roman dichter stoff meister erzählungen krieges
46	prinzessin wiederum hauses magd rom größe römischen prinz damen gatten
47	alte bild maler kunst künstler freund bilder gemälde fremde junger
48	kahn wasser bild schlangenkönigin dorfe schlangen ufer entgegen gegend leipe
49	schnee tante mühle Kindes doktor nordwind blankenheim margrets tropfen wölfe
50	schatten gold herrn grauen sonne tasche goldes seckel pracht vieles
51	kranken kranke krankheit arzt klagte heilung gehoben uebel ausdauer stock
52	ruft büchse sieht gesellen wolf bär juden fällt thiere schuß
53	mynheer meister zorghof amsterdam freimeister buchhalter see jüngling insel familie
54	kennen natur lernen nennen bildung kunst geist meisten höheren erscheint
55	alte schneider mädchen versetzte sohn pfarrerin gesicht burschen entgegnete bursche
56	paris deutschen frankreich töchter damen schwestern monseigneur allmählich bordeaux hamburg
57	sprach eremit freunde überall traum richter fest riefen schatz engel
58	hand augen rief stand gesicht hielt kopf trat lag mädchen
59	amtman baron fräulein versetzte halden dame amtmanns referendar weis barons
60	heiligen nonne bild kloster antlitz kunst kirche malen bildes entgegnete
61	geld gulden thaler handel summe wasser schulden littauer wirth verkauft
62	erfreute korb betraten verleiten feine bescheidenheit fressen zahl geschirr verschiedene
63	frau conrectorin vetter vetterchen general eigentlich tone lachen erzählen weile
64	herr hahn herrn papa gast todten mama herbesheim fräulein stadt
65	treppe polizei holz ofen stufen lachen emmerich sieh niemals feuer
66	chef lieutenant freund ehre freunde gefährten ring flagge offiziere freundes
67	ring finger musik eltern palast schöne niemals kästchen stehen gewaltigen
68	jäger bauern mühle förster bauer wald thal hütte jüngling felsen

69	scheint neffe entgegnete würdest verbreiten ansicht eva ludwig paradiese streben
70	frei schuldig staatsanwalt rothmann geschworenen verhandlung schwiegersohn angeklag- ten männer vorsitzende
71	sprach rhein trug nebi fischer klausur klotz manch berg kniggemann
72	fremde neger fremden alte familie knaben herrn weißen bette babekan
73	klarinetten schloß dame wald bursch garten student fenster draußen droben
74	acker kinder bauern puppe nase steine jahren wilde aecker loch
75	wein tisch glas trinken tische essen trank gäste flasche tafel
76	major haide pferde ritten herum uwar maroshely gitter ritt lande
77	salonichi diener juden paradise griechen brittischen arbeiten athen konstantinopel alex- ander
78	teufel general soldaten könig hauptmann commandant officier dienst oberst wache
79	weißen trug schwarzen haar langen schön schönen schwarze rothen blauen
80	pfarrer burg berg tubus pfarrerin gesellschaft pfarrers gleichfalls nämlich examen
81	kirche pfarrer sonntag heilige messe gemeinde priester geistliche heiligen predigt
82	theils laden wobei korb kunden fände georg's landenberger kanzlei wittwe
83	lied musik singen spielen sang talent spielte geige klavier sängerin
84	todt tod blut hülfe leiche wunde wuth todten mörder kampf
85	used total openjdk-i usr/lib/jvm/java dff acf thread xae heap jre/lib/i
86	fräulein sprach schmuck meister könig arbeit that kästchen geheimniß paris
87	doctor bauer drachen pirna erbschaft kappbauer doctors flurschütz stube pastor
88	wagen pferde fahren pferd kutscher straße reiter fremde pferden knecht
89	königin spielen burschen stück saba bursche schiffer schwarze bühne zunftmeister
90	gutsherr ohm dorf knaben förster dennoch holz uhr juden knabe
91	vermittlung gemeinde klöster gesunden gottlosen zustand dünnen gebilde gesunde ein- gebüßt
92	richter stadtschreiber galgen gericht urtheil herren rath strafe freiheit zuletzt
93	fenster zimmer haus thür thüre hinaus nacht treppe öffnete licht
94	vater graf mann frau marquis sehen dachte frug lies baron
95	munde vetter frau that schäferle buchenberg haus stube kapitel schäfer
96	kästchen gesellschaft wagen zimmer schöne beutel geld wodurch leben vergnügen
97	weib sag schuster lustig pfand lachen pfeife bauern frauen liebe
98	franzosen krieg frankreich könig männer vaterland volk kaiser feind französischen
99	oberst meister fräulein erwiderte leise stadt gnädiger obersten tochter gefangene
100	tanz tanzen tanze vicomte tanzte jungen musik mädchen musikanten altenkreuz