Versuch einer literarischen Topologie

Der Mythos der literaturzentristischen russischen Kultur scheint auch vor dem Runet nicht haltzumachen: In zahlreichen Webauftritten wird über Literatur diskutiert, in ebenso zahlreichen Webauftritten werden eigene literarische Texte einer manchmal breiten, manchmal weniger breiten Öffentlichkeit vorgestellt. Dieser Fokus des Runet auf das literarische Schreiben lässt sich unter anderem daran ablesen, dass sich ein eigener institutionalisierter Literaturbetrieb im Netz herausgebildet hat, der Wettbewerbe, Literaturzeitschriften und Bibliotheken umfasst (Schmidt 2011: 78-191). Viele außerhalb des Netzes bekannte Autorinnen und Autoren bloggen; für die vorliegende Untersuchung schriftstellerischer (Selbst-)Inszenierungen liegt damit umfangreiches Material vor. Die dominante Blog-Plattform im Runet ist das auf Seite 76 vorgestellte ŽŽ. Ende der 2000er-Jahre steigt dann die Bedeutung sozialer Netzwerke stetig an, literarische Texte werden aber nach wie vor eher im ŽŽ veröffentlicht (Roesen/Zvereva 2014: 78f.). Als Grundlage der vorliegenden quantitativen Untersuchung dienen deshalb ŽŽ-Blogs, die in einzelnen Fällen um Profile in sozialen Netzwerken ergänzt werden, um ein umfassenderes Bild von Strategien der (Selbst-)Inszenierung zeichnen zu können.

Die literarische ›Schlagseite‹ des ŽŽ zeigt sich immer wieder; so sind einzelne Blogs allein der Frage gewidmet, wer sich zu den russischen Schriftstellerinnen und Schriftstellern im Netz zählen darf, sie versuchen also, kanonbildend zu wirken. Die wohl umfassendste Bestandsaufnahme findet sich in der »Biblioteka Živogo Žurnala« [»ŽŽ-Bibliothek«], dem Blog ⟨biblioteka-2013⟩. Dieser setzt sich zum Ziel, »найти авторов ›бумажных‹ книг, которые ведут блоги в Живом Журнале« [»Autoren von ›Papier‹-Büchern zu finden, die einen ŽŽ-Blog führen«] und wird vom ŽŽ finanziell unterstützt (Biblioteka Živogo Žurnala 2015).[1] Jede Leserin und jeder Le-

1 | An dieser Stelle sei daran erinnert, dass Nicknamen auf Internetplattformen in dieser Arbeit mit spitzen Klammern gekennzeichnet werden: ⟨…⟩.

ser kann Autorinnen bzw. Autoren zur Aufnahme in die ŽŽ-Bibliothek vorschlagen, das Publikum wird also aktiv an den Kanonisierungsbestrebungen beteiligt. Das Resultat dieser Bemühungen ist eine Liste, die 2015 insgesamt 677 Namen umfasste. Zusätzlich werden bevorzugte Gattungen und gegenwärtige Wohnorte angeführt (Biblioteka Živogo Žurnala 2013a, Biblioteka Živogo Žurnala 2013b).

Andere Bloggerinnen und Blogger wiederum versuchen, eine qualitative Auswahl zu treffen. Ekaterina Pachomčik, ihres Zeichens PR-Direktorin des ŽŽ, hat unter dem vielsagenden Pseudonym ⟨lytdybr⟩[2] eine Liste von 29 Blogs veröffentlicht, »[к]ого уж точно стоит почитать« [»(w)en es sich wirklich zu lesen lohnt«] (Pachomčik 2013).[3] Eine Zeit lang führte das ŽŽ selbst Buch über ›seine‹ Schriftstellerinnen und Schriftsteller. Die Seite »Dajdžest Živogo Žurnala« [»Übersicht über das ŽŽ«] ist seit 2014 nicht mehr verfügbar, kann aber noch über *archive.org* aufgerufen werden. Insgesamt 123 »celebrities« aus dem Bereich Literatur werden hier gelistet, als einziger nicht-russischsprachiger Schriftsteller kommt der amerikanische Fantasyautor George R. R. Martin vor, der unter dem Nick ⟨grrm⟩ bloggt (LiveJournal 2013).

Diese Versuche der Kanonisierung erleichtern die Auswahl einzelner Autorinnen und Autoren für die vorliegende Analyse von (Selbst-)Inszenierungen. Grundsätzlich soll ein möglichst breites Spektrum an inszenatorischen Strategien abgedeckt werden. Dafür kommen quantitative Methoden zum Einsatz, die es erlauben, ein großes Textkorpus kursorisch zu erfassen und in repräsentative Gruppen zu gliedern. Das vorliegende Kapitel zur literarischen Topologie vermittelt zunächst einen Überblick über den Einsatz quantitativer Methoden in der Literaturwissenschaft, für den Franco Moretti den Begriff »distant reading« geprägt hat. Anschließend stellt es das für die vorliegende Untersuchung verwendete Verfahren des »topic modeling« vor und geht auf weitere mathematische Ansätze ein, die genutzt werden können, um mehr über das Korpus herauszufinden. Zusätzlich werden die Ergebnisse des »topic modeling« und der nachgeschalteten Verfahren präsentiert. Diese erlauben es, die Blogs der Schriftstellerinnen und Schriftsteller in drei grundlegende Gruppen zu unterteilen, nämlich in solche mit Politik-, Alltags- und Literatur-Schwerpunkt. Abschließend können technisch Interessierte einen Blick hinter die Kulissen wagen. Dabei werden die Ergebnisse aus quantitativer Perspektive diskutiert und Details der Implementierung erläutert.

2 | Die Buchstabenfolge ›lytdybr‹ entspricht auf der kyrillischen Tastatur dem Wort ›дневник‹ [›Tagebuch‹] (Schmidt 2011: 273).

3 | Der Eintrag verspricht eigentlich 30 Blogs, allerdings fehlt Nummer 18.

Tabelle 1: Die in der vorliegenden Arbeit analysierten 37 Webauftritte, Ergänzungen zu Pachomčik (2013) sind markiert (*). Gezählt wurden alle öffentlichen Einträge bis Ende 2014. Für Profile in den sozialen Netzwerken Facebook *(*FB*) und* Vkontakte *(*VK*) ist nicht das Datum der Registrierung angegeben, sondern das des ersten Eintrages.*

	Name	*Blog*	*Einträge*	*Online seit*
1.	Akunin, Boris	⟨borisakunin⟩	311	07. 11. 2010
2.	Akunin, Boris*	⟨borisakunin⟩ (*FB*)	349	12. 08. 2012
3.	Bagirov, Ėduard	⟨bagirov⟩	92	17. 10. 2006
4.	Berezin, Aleksej	⟨alex-aka-jj⟩	604	08. 02. 2009
5.	Bormor, Petr	⟨bormor⟩	2.404	05. 08. 2002
6.	Bykov, Dmitrij	⟨ru-bykov⟩	7.698	28. 03. 2005
7.	Divov, Oleg	⟨divov⟩	993	01. 11. 2003
8.	Ėksler, Aleksej	⟨exler⟩	15.795	13. 06. 2001
9.	Fraj, Maks	⟨chingizid⟩	5.323	24. 05. 2002
10.	Galkovskij, Dmitrij	⟨galkovsky⟩	934	10. 10. 2003
11.	Gluchovskij, Dmitrij	⟨dglu⟩	605	31. 05. 2005
12.	Goralik, Linor	⟨snorapp⟩	3.510	16. 04. 2002
13.	Goralik, Linor*	⟨snorapp⟩ (*FB*)	3.408	16. 09. 2009
14.	Grišковec, Evgenij*	⟨e-grishkovets⟩	685	24. 04. 2007
15.	Gromov, Aleksandr	⟨lemming-drover⟩	698	12. 04. 2005
16.	Ketro, Marta	⟨marta-ketro⟩	1.504	21. 02. 2005
17.	Klimova, Marusja	⟨marussia⟩	902	22. 12. 2001
18.	Kudrjaševa, Alja	⟨izubr⟩	948	03. 06. 2003
19.	Kudrjaševa, Alja*	⟨khaitlina⟩ (*VK*)	570	16. 10. 2007
20.	Kudrjaševa, Alja*	⟨kudryasheva⟩ (*FB*)	281	06. 09. 2012
21.	Kudrjaševa, Alja*	⟨xelbot⟩	330	11. 09. 2005
22.	Limonov, Ėduard*	⟨limonov-eduard⟩	2.616	11. 03. 2009
23.	Loginov, Svjatoslav	⟨sv-loginow⟩	472	13. 03. 2008
24.	Luk'janenko, Sergej*	⟨doctor-livsy⟩	823	16. 07. 2003
25.	Luk'janenko, Sergej	⟨dr-piliulkin⟩	2.373	13. 07. 2008
26.	Mart'janov, Andrej	⟨gunter-spb⟩	9.332	06. 03. 2007
27.	Mel'nikova, Julija	⟨avit-al⟩	405	13. 06. 2008
28.	Minaev, Sergej	⟨amigo095⟩	1.742	01. 05. 2003
29.	Morozova, Tat'jana	⟨maroosya⟩	259	21. 03. 2013
30.	Perumov, Nik	⟨captain-urthang⟩	51	29. 08. 2003
31.	Polozkova, Vera	⟨mantrabox⟩	2.606	29. 12. 2002
32.	Prilepin, Zachar	⟨prilepin⟩	841	06. 03. 2007
33.	Rubinštejn, Lev	⟨levrub⟩	732	19. 01. 2005
34.	Sabitova, Dina	⟨feruza⟩	10.175	06. 09. 2002
35.	Sė, Slava	⟨pesen-net⟩	240	22. 04. 2007
36.	Tolstaja, Tat'jana	⟨tanyant⟩	441	14. 12. 2007
37.	Truskinovskaja, Dalija	⟨_runcis⟩	1.899	10. 08. 2005

Quelle: G. H.

Distant Reading

Um ein möglichst umfassendes Bild der (Selbst-)Inszenierung russischer Autorinnen und Autoren zeichnen zu können, ist es wünschenswert, viele repräsentative Webauftritte in die Überlegungen miteinzubeziehen. Die vorliegende Untersuchung konzentriert sich auf die 29 Autorinnen und Autoren aus Pachomčiks Liste, weil diese als prototypische ŽŽ-Schriftstellerinnen und -Schriftsteller präsentiert werden. Eine Übersicht über die bearbeiteten Blogs ist in Tabelle 1 auf Seite 83 dargestellt. An Pachomčiks ursprünglicher Liste wurden kleinere Modifikationen vorgenommen, so wurde Sergej Luk'janenkos aktueller Blog ⟨dr-piliulkin⟩ um den älteren Blog ⟨doctorlivsy⟩ ergänzt. Bei Boris Akunin, Linor Goralik und Alja Kudrjaševa wurden auch Auftritte in sozialen Netzwerken in das Korpus aufgenommen, weil diesen Autorinnen und Autoren jeweils ein eigenes Kapitel zur Detailanalyse gewidmet ist.

Hinzugefügt wurden weiters zwei bekannte Blogger, die nicht in der Liste aufscheinen, weil sie nicht (mehr) im ŽŽ posten: Evgenij Griškovec und Aleksandr Ėksler. Griškovec hat mehrere Jahrgänge seines Blogs als Buch veröffentlicht (Griškovec 2009, Griškovec 2010a, Griškovec 2011a, Griškovec 2012a, Griškovec 2013a, Griškovec 2014). Ėksler wiederum ist als graphomanischer Blogger bekannt, weshalb ihm Henrike Schmidt (2011: 460-470) in ihrer Monographie ein eigenes Kapitel widmet. Hier erscheint es besonders spannend, im Zuge der vorliegenden Untersuchung Schmidts qualitative Interpretation mit den Ergebnissen quantitativer Verfahren abzugleichen. Als letzte Ergänzung ist Ėduard Limonovs Blog ⟨limonov-eduard⟩ zu nennen. Limonov ist mittlerweile ausschließlich politisch tätig. Sein Blog markiert einen inhaltlichen Grenzfall, der als Vergleichsfolie für die (Selbst-)Inszenierung der anderen Schriftstellerinnen und Schriftsteller dienen soll. Entfernt wurde hingegen ⟨denis-balin⟩. Der Grund dafür ist technischer Natur. Balin veröffentlicht in seinem Blog häufig pornographische Bilder, die mit einer Altersabfrage geschützt sind. Diese Abfrage erschwert eine automatisierte Bearbeitung des Blogs.

Trotz der Beschränkung auf 37 Webauftritte ist die Gesamtanzahl der Texte immer noch groß. Einzelne Blogs bestehen immerhin seit 15 Jahren, manche Schriftstellerinnen und Schriftsteller veröffentlichen mehrere Einträge pro Tag. Für die vorliegende Untersuchung wurden alle Einträge, die bis zum 31. Dezember 2014 um 23:59 Uhr veröffentlicht worden sind, erfasst. Es handelt sich um 78.268 Texte mit insgesamt 17.816.195 Wörtern. Zum Vergleich:[4] Lev Tolstojs *Vojna i mir* [*Krieg*

4 | Als Grundlage dieses Vergleichs dient das Digitalisat von Tolstojs *Sobranie sočinenij v 22 tomach* [*Gesammelte Werke in 22 Bänden*] (Tolstoj 1979, Tolstoj 1980a, Tolstoj 1980b, Tolstoj 1981).

und Frieden] besteht aus 459.188 Wörtern, das für die vorliegende Studie verwendete Textmaterial ist damit fast vierzig Mal so umfangreich. Zwar kann die thematische Ausrichtung einzelner Webauftritte durch Querlesen grob bestimmt werden, die Beziehungen zwischen den Webauftritten und feinere thematische Einschätzungen bleiben aber verborgen; die schiere Textmasse widersetzt sich einer detaillierten Lektüre. Im Kontext der Digital Humanities mag der Umfang des Korpus wiederum bescheiden wirken, Michel et al. (2011) bearbeiten etwa im Zuge ihrer vielzitierten und -kritisierten »culturomics« Millionen von Büchern. Dazu ist festzuhalten, dass diese großangelegten Vorhaben in der Regel sehr allgemeine Fragestellungen bearbeiten, Michel et al. möchten etwa Genaueres über die Dynamiken menschlicher Kultur in ihrer Gesamtheit (sic!) herausfinden. Das hier präsentierte Forschungsvorhaben hat hingegen einen klar umrissenen Fokus auf russische Autorinnen und Autoren, deshalb erscheint es sinnvoll, das Korpus entsprechend einzugrenzen.

Da sämtliche Webauftritte in ihrer Gesamtheit als integraler Bestandteil der (Selbst-)Inszenierung verstanden werden, muss diese mediale Masse in irgendeiner Form handhabbar gemacht werden. Im Jahr 2000 schlägt der Literaturwissenschafter Franco Moretti, als er mit einem ähnlichen Problem konfrontiert wird, den Begriff des »distant reading« vor. Um Weltliteratur als solche untersuchen zu können, reiche es nicht, Einzeltexte einer genauen Lektüre zu unterziehen, es sei vielmehr notwendig, auf quantitative Methoden zurückzugreifen, um die Gesamtheit der Texte erfassen zu können:

> [L]iterary history will […] become ›second hand‹: a patchwork of other people's research, *without a single direct textual reading*. […] [T]he ambition is now directly proportional *to the distance from the text*: the more ambitious the project, the greater must the distance be. […] we know how to read texts, now let's learn how *not* to read them. Distant reading: where distance […] *is a condition of knowledge*: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems. And if, between the very small and the very large, the text itself disappears, well, […] one can justifiably say, Less is more. If we want to understand the system in its entirety, we must accept losing something. (Moretti 2000: 57, Hervorh. i. O.)

Dieser Ansatz ist radikal und wird entsprechend kritisiert (u.a. Prendergast 2001, Arac 2002, Kristal 2002, Parla 2004). Moretti reagiert auf diese Kritik, indem er zwar zwischen den Zeilen andeutet, an einem Nebeneinander von literarischen Mikro- und Makrostrukturen interessiert zu sein, sodass »the ›details‹ […] will be highlighted, not erased by models and ›schemas‹« (Moretti 2003: 81). Trotzdem bekräftigt er seine Ablehnung des »close reading« erneut: »This is of course the old question

of whether the proper object of historical disciplines are individual cases or abstract models; and […] I will argue […] for the latter« (Moretti 2003: 80). Wie »distant reading« einzelne Details in den Vordergrund rücken könnte, führt Moretti allerdings nicht aus.

Einen Ausweg aus dieser polemischen Sackgasse weist N. Katherine Hayles, die drei Arten des Lesens skizziert: Neben dem traditionellen »close reading« führt sie das »hyperreading« an, also das überblicksmäßige Querlesen von (Bildschirm-)Texten (Hayles 2010: 66), sowie das algorithmische »machine reading«, das Morettis »distant reading« entspricht (ebd.: 72f.). Vom Menschen durchgeführtes »hyperreading« und »machine reading« seien sich demnach insofern ähnlich, als beide Varianten genutzt werden könnten, Texte für das »close reading« auszuwählen (ebd.: 74). Zur Kombination von »machine« und »close reading« schreibt Hayles:

> Relatively context-poor, machine reading is enriched by context-rich close reading when close reading provides guidance for the construction of algorithms; […]. On the other hand, machine reading may reveal patterns overlooked in close reading[.] (Ebd.: 75)

Erst aus der Wechselwirkung dieser beiden Ebenen ergibt sich dann ein vollständiges Bild; analog zu diesem Ansatz kommt auch in der vorliegenden Arbeit eine Kombination aus qualitativen und quantitativen Verfahren zum Einsatz.

Franco Moretti ist zweifellos einer der bekanntesten Proponenten eines quantitativen Zugangs zur Literatur, er ist damit allerdings keineswegs der erste. Quantitative Verfahren für die Literaturwissenschaft werden bereits im 19. Jahrhundert angedacht.[5] Augustus de Morgan schlägt 1851 vor, lexikalische Statistiken für die Autorschaftsbestimmung der Paulusbriefe zu nutzen (Hockey 2004: 5). Parallel dazu äußert der russische Mathematiker Viktor Bunjakovskij 1847 die Vermutung, die Wahrscheinlichkeitstheorie ließe sich gewinnbringend auch in der Literaturwissenschaft anwenden (Kelih 2008: 31f.). Da Naturwissenschaft und Geisteswissenschaft in Russland nicht so strikt voneinander getrennt werden wie im Westen, beschäftigt sich eine Reihe von russischen Wissenschaftern in der ersten Hälfte des 20. Jahrhunderts mit quantitativen Verfahren in einem literarischen Kontext (ebd.: 271). Darunter sind so bekannte Namen wie der Literaturkritiker Nikolaj Černyševskij, der Mathematiker Andrej Markov, der Symbolist Andrej Belyj und der Formalist Boris

5 | Einen Überblick über die Geschichte der Digital Humanities unter besonderer Berücksichtigung quantitativer Verfahren in der Literatur- und Sprachwissenschaft findet sich bei Hockey (2004). Kelih (2008) wiederum hat eine Darstellung quantitativer Ansätze in der russischen Literatur- und Sprachwissenschaft verfasst.

Tomaševskij. Überlegungen zum russischen Vers stehen dabei im Vordergrund (ebd.: 39-41; 47f.; 83-88; 272-278).

All diese Versuche eint, dass sie sozusagen noch ›von Hand‹ berechnet werden. Erst 1949 wird ein Computer für literaturwissenschaftliche Zwecke eingesetzt. In diesem Jahr beginnt der Theologe Roberto Busa gemeinsam mit IBM, eine Konkordanz aller 11 Millionen Wörter in Thomas von Aquins Texten zu erstellen. Dreißig Jahre später ist diese Arbeit abgeschlossen (Hockey 2004: 4). Bis Ende der 1980er-Jahre werden quantitative Verfahren in der Literaturwissenschaft fast ausschließlich dafür eingesetzt, Konkordanzen zu erstellen und Fragen nach Autorschaft und Stil zu klären. Die 1990er-Jahre werden dann von Digitalisierungsbestrebungen dominiert (ebd.: 10-13). Franco Moretti und seinem »distant reading« gebührt also das Verdienst, eine Neuausrichtung der quantitativen Literaturwissenschaft hin zu bislang nur qualitativ bearbeiteten Fragestellungen angestoßen zu haben; und das zu einer Zeit, als vom kommenden Boom quantitativer Methoden in den Geisteswissenschaften noch wenig zu spüren war. Wie Matthew Kirschenbaum ausführt, hat John Unsworth den spätestens seit der 2010er-Jahre inflationären Begriff *Digital Humanities* etwa erst im Jahr 2004 geprägt (Kirschenbaum 2010: 56f.).

Im Folgenden wird das »distant reading« der ausgewählten russischen Webauftritte im Detail beschrieben. Dabei werden quantitative Verfahren eingesetzt, um eine ›topologische‹ Übersichtskarte, die Topic-Karte, zu zeichnen, die thematische Beziehungen zwischen einzelnen Webauftritten visualisiert. Die erhobenen Daten und die zur Verarbeitung verwendeten Skripte können mitsamt einer kurzen Anleitung auf der Online-Plattform *Github* eingesehen werden.[6] Dies soll die Nachvollziehbarkeit der im Rahmen dieser Untersuchung verwendeten quantitativen Verfahren gewährleisten und es interessierten Wissenschafterinnen und Wissenschaftern ermöglichen, die verwendeten Techniken für eigene Forschungsfragen anzupassen.

Topic Modeling

Unter dem Begriff »topic modeling« werden verschiedene Verfahren zusammengefasst, deren Ziel es ist, die Themen eines großen Textkorpus automatisiert zu bestimmen.[7] Diese Algorithmen operieren in abstrahierter Form auf der Inhaltsebene und bieten sich deshalb für die geplante quantitative Analyse russischer Webauftritte an.

6 | https://github.com/ghowa/russian-blogs, letzter Aufruf 10. September 2019.

7 | Für eine Einführung in das »topic modeling« siehe Blei (2012b).

Hier zeigt sich die von David Blei (2012a) konstatierte durchlässige Grenze zwischen Information Retrieval (Informationsrückgewinnung), einem Teilgebiet der Informatik, und Digital Humanities.

Die bekannteste Variante des »topic modeling« ist die von David Blei, Andrew Ng und Michael Jordan 2003 vorgestellte *Latent Dirichlet Allocation* (LDA), benannt nach dem deutschen Mathematiker Johann Peter Gustav Lejeune Dirichlet. Diese ist allerdings nicht, wie auf den ersten Blick vermutet werden könnte, ein Beispiel für künstliche Intelligenz, die Texte lesen, verstehen und zusammenfassen kann. Bei LDA handelt es sich vielmehr um ein statistisches Modellierungsverfahren, das Wortgruppen zu abstrakten Themenkreisen zusammenschließt. Im Folgenden soll deshalb nicht von ›Themen‹ im klassischen literaturwissenschaftlichen Sinn gesprochen werden, sondern von ›Topics‹.

Das »topic modeling« geht von einer grundlegenden Prämisse aus: Texte werden nicht von Menschen geschrieben, sondern entstehen durch Zufallsprozesse, die von Wahrscheinlichkeitsverteilungen auf zwei Ebenen gesteuert werden. Zunächst bestimmt eine Wahrscheinlichkeitsverteilung, aus welchen Topics ein Text zusammengesetzt werden soll. Weitere Wahrscheinlichkeitsverteilungen ordnen bestimmte Wörter einzelnen Topics zu. Ihnen entsprechend wird das Wortmaterial bestimmt, aus dem der Text zusammengefügt wird (Blei et al. 2003: 996). Ein Topic ist für LDA also nichts anderes als eine konkrete Wahrscheinlichkeitsverteilung über Wörter, die im Vorfeld festgelegt wurde. Ein Text wiederum entspricht einer bestimmten Wahrscheinlichkeitsverteilung über gleichfalls vorher festgesetzte Topics.

Ziel der LDA ist nicht, neue Texte zu schreiben; ganz im Gegenteil sollen auf Basis bereits bestehender Texte die im Hintergrund vermuteten (latenten) Wahrscheinlichkeitsverteilungen abgeleitet werden (ebd.: 1003). Als Resultat liefert ein LDA-Lauf basierend auf den Worthäufigkeiten in den Texten die Wortverteilungen pro Topic und die Topicverteilungen pro Text. Obwohl der mathematische Kontext mit seinen scheinbar genauen Prozentzahlen eine exakte Methode suggeriert, darf nicht vergessen werden, dass es sich um eine Modellrechnung handelt; die Topicverteilungen der Texte kann nicht genau bestimmt werden, sondern wird näherungsweise ermittelt. Die Ergebnisse einzelner LDA-Läufe, die mit den gleichen Texten operieren, können sich daher in Details unterscheiden.

Der LDA-Algorithmus modelliert eine vom Menschen vorgegebene Anzahl von Topics. Jedes Topic ist im Grunde eine Liste von Wörtern; die menschliche Interpretationsleistung verschiebt sich also weg von der Textlektüre zur Verbindung der Topics mit tatsächlichen Themenkomplexen. Damit »topic modeling« sinnvoll eingesetzt werden kann, müssen die modellierten Topics für den Menschen deshalb einfach zu interpretieren sein. Für die Qualität der Topics spielen Wortstellung, Satzein-

heit und andere syntaktische Konstellationen keine Rolle, nur das Wortmaterial ist ausschlaggebend. Jeder Text wird von der LDA als »bag of words« verstanden, das heißt, es wird gezählt, wie häufig jedes Wort vorkommt. Für die vorliegende Untersuchung werden deshalb zunächst Interpunktionszeichen, Sonderzeichen, Zahlen und lateinische Buchstaben entfernt. Letzteres sorgt dafür, dass die LDA nur mit kyrillischen Wörtern operiert. Sämtliche Großbuchstaben werden durch Kleinbuchstaben ersetzt. Wie eine empirische Testreihe mit dem Korpus gezeigt hat, ist es dann hinsichtlich der subjektiven Topicqualität zielführend, nur Nomina zu verwenden; einen ähnlichen Ansatz verfolgen etwa Chang et al. (2009: 292). Um die Unschärfen, die das Russische als flektierende Sprache mit sich bringt, etwas abzumildern, werden die Nomina in den Nominativ Singular gesetzt. Solcherart aufbereitet, reduziert sich das Wortmaterial dabei von ursprünglich 17.816.195 Wörter auf 4.960.044 Nomina. Weitere Versuche mit anderen Wortarten, mit einer Reduktion auf synthetische Stammformen (Stemming) sowie mit dem Filtern von Eigennamen haben zu keiner Verbesserung der subjektiven Topicqualität geführt.

Neben der Aufbereitung der Ursprungstexte nimmt die Anzahl der zu suchenden Topics maßgeblichen Einfluss auf das Ergebnis des »topic modeling«. Eine Versuchsreihe mit jeweils 30, 50, 75 und 100 Topics hat zu folgender Einschätzung geführt: Bei einer größeren Topicanzahl, also 75 und 100, entstehen viele sehr kleinteilige Topics, die wenig Aussagekraft haben. Eine Anzahl von 30 mündet wiederum in wenigen großen, nichtssagenden Topics, die mehrere Themenkomplexe in sich vereinen. 50 Topics stellen einen guten Kompromiss zwischen den beiden Extremen dar, deshalb beruhen die nachfolgenden Analysen auf dieser Topicanzahl.

Wie einfach ist es, die von der LDA modellierten Topics zu interpretieren? Um diese Frage zu beantworten, werden im Folgenden einige exemplarische Resultate des für die folgende Analyse verwendeten LDA-Laufes mit 50 Topics angeführt. Auf Grundlage aller in den 37 Blogs verwendeten Nomina wird etwa ›Topic Nr. 40‹ vom Algorithmus mit folgenden Wörtern modelliert: *власть, россия, путин, гражданин, президент, страна, выбор, партия, политика, оппозиция, митинг, площадь, владимир, москва, свобода...* [*Macht, Russland, Putin, Bürger, Präsident, Land, Wahl, Partei, Politik, Opposition, Demo, Platz, Vladimir, Moskau, Freiheit...*]. Hier kommt die (Staats-)Macht gleich an erster Stelle, gefolgt von den Signalwörtern *Russland* und *Putin*.

Ergänzt wird dieses innenpolitische Topic durch das außenpolitische ›Topic Nr. 37‹: *россия, война, украина, страна, европа, народ, государство, ссср, сша, территория, германия, крым, запад, америка, киев...* [*Russland, Krieg, Ukraine, Land, Europa, Volk, Staat, UdSSR, USA, Territorium, Deutschland, Krim, Westen, Amerika, Kiew...*]. Dass sich diese erkennbar unterschiedlichen Topics trotz einiger gemeinsa-

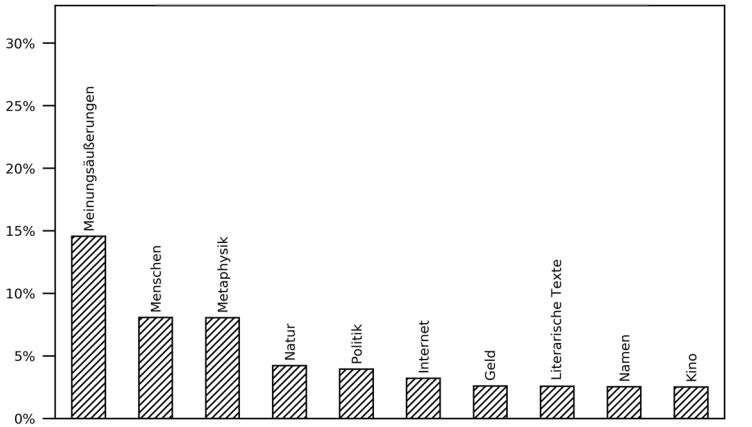
mer Wörter nicht mischen, spricht für die feine Granularität der Methode bei 50 Topics. Leicht einem konkreten Topic zuzuordnen ist auch die Wortkonstellation von ›Topic Nr. 10‹: *книга, автор, роман, литература, писатель, читатель, текст, книжка, рассказ, издательство, проза, герой, произведение, чтение, фантастика...* [*Buch, Autor, Roman, Literatur, Schriftsteller, Leser, Text, Büchlein, Erzählung, Verlag, Prosa, Held, Werk, Lesung, Phantastik...*]. Hier handelt es sich vorwiegend um Wörter aus dem Feld der Literatur, entsprechend ausgerichtet sind dann auch Texte, die dieses Topic beinhalten. Für manche Topics lässt sich ein konkreter Bezug allerdings nicht auf den ersten Blick herleiten, so wirken die Wörter von ›Topic Nr. 34‹ zunächst etwas wahllos: *свет, голос, музыка, ночь, рука, звук, огонь, старик, глаз, воздух, звезда, голова, шаг, сон, луна...* [*Licht, Stimme, Musik, Nacht, Hand, Geräusch, Feuer, alter Mann, Auge, Luft, Stern, Kopf, Schritt, Schlaf, Mond...*].

Ein Querlesen von Texten, in denen dieses ›Topic Nr. 34‹ sehr präsent ist, zeigt, dass es sich in den meisten Fällen um *literarische* Texte handelt: Gedichte, Erzählungen oder Auszüge aus Romanen; hier scheint ein fehlender offensichtlicher innerer Zusammenhang und die auf Wortebene eher lose Verbindung eine Konstante darzustellen. Da die Topics auf Wörtern basieren, die häufig gemeinsam auftreten, sind sie manchmal weniger thematisch oder inhaltlich zu verstehen, sondern zeigen unter Umständen verschiedene Genres oder andere Gemeinsamkeiten der Texte an. Die ermittelten Topics lassen sich demnach zwei grundlegenden Klassen zuordnen: Die meisten Topics beschreiben, *was* ein Text vermittelt und erinnern damit an Themen im ›klassischen‹ Sinn: *Literatur*, *Politik* oder *Ukraine*. Das Topic *Literarische Texte* skizziert jedoch, *wie* ein Text seine Inhalte vermittelt, stilistische Eigenschaften des jeweiligen Textes stehen also im Zentrum. Diese Beispiele zeigen auch, dass Topics sozusagen auf unterschiedlichen Ebenen ansetzen können; neben breit gefassten Topics gibt es auch sehr spezielle. Einziges Kriterium für den Algorithmus ist, dass die Wörter eines Topics in möglichst vielen Texten gemeinsam vorkommen.

Wie bereits erwähnt, ist die LDA als Modell zu verstehen, das eine gewisse Fehlerquote aufweist. Texte, die typisch für das ›Topic Nr. 34‹, also das Topic *Literarische Texte* sind, müssen nicht automatisch literarische Texte sein; umgekehrt gibt es literarische Texte, die sich nicht in diesem Topic finden. Wie gut die automatisiert ermittelten Topics tatsächlich die ihnen zugeordneten Texte beschreiben, wird im Kapitel »Hinter den Kulissen« ab Seite 99 erörtert. Dort finden sich auch weitere Beispiele für modellierte Topics.

Nach der Modellierung der Topics durch die LDA kann diese dafür genutzt werden, die Topicverteilungen in einzelnen Texten, in einzelnen Blogs sowie im Gesamtkorpus zu bestimmen. Diese vermitteln im Sinne des »distant reading« einen ersten Einblick in die thematischen Schwerpunkte einzelner Autorinnen und Auto-

Abbildung 1: Die zehn häufigsten Topics im Gesamtkorpus



Quelle: G. H.

ren. Abbildung 1 zeigt die zehn bestimmenden Topics, die der oben erwähnte LDA-Lauf identifiziert hat. Es dominiert das Topic *Meinungsäußerungen*, auf dem zweiten Platz landet das Topic *Menschen*, das häufig literarische Texte enthält. Im folgenden Topic *Metaphysik* finden sich Einträge über die ›großen Themen‹ wie Liebe, Tod, und Gott. Zum Topic *Natur* auf Platz vier zählen vor allem Reiseberichte, aber auch Landschaftsbeschreibungen in literarischen Texten. Als erstes ›handfestes‹ Topic taucht *Politik* (Platz fünf) auf, was aufgrund der Verwerfungen in der russischen Innenpolitik, sprich: Putins Wiederwahl 2012, wenig verwundert. Die selbstreflexiven Einträge im Topic *Internet* auf Platz sechs, die beispielsweise auf andere Blogs verweisen, kommen ebenfalls relativ häufig vor, viele Autorinnen und Autoren vernetzen sich offenbar ausgiebig mit der russischen Blogosphäre und in sozialen Netzwerken. Diese Vernetzung mit anderen kehrt im Topic *Namen* auf Platz neun wieder. Wie das Topic *Literarische Texte* auf Platz acht zeigt, dienen die Webauftritte auch dazu, Schaffensproben zu veröffentlichen, um Rückmeldungen der Leserinnen und Leser zu erhalten. Das Topic *Geld* auf Platz sieben deckt den kommerziellen Aspekt des Literaturbetriebs ab, enthält aber auch Aufrufe zu wohltätigen Zwecken. Das Topic *Kino* auf Platz zehn schließlich besteht vorwiegend aus Filmrezensionen, die in unglaublicher Zahl von Aleksej Ėksler ins Netz gestellt werden.

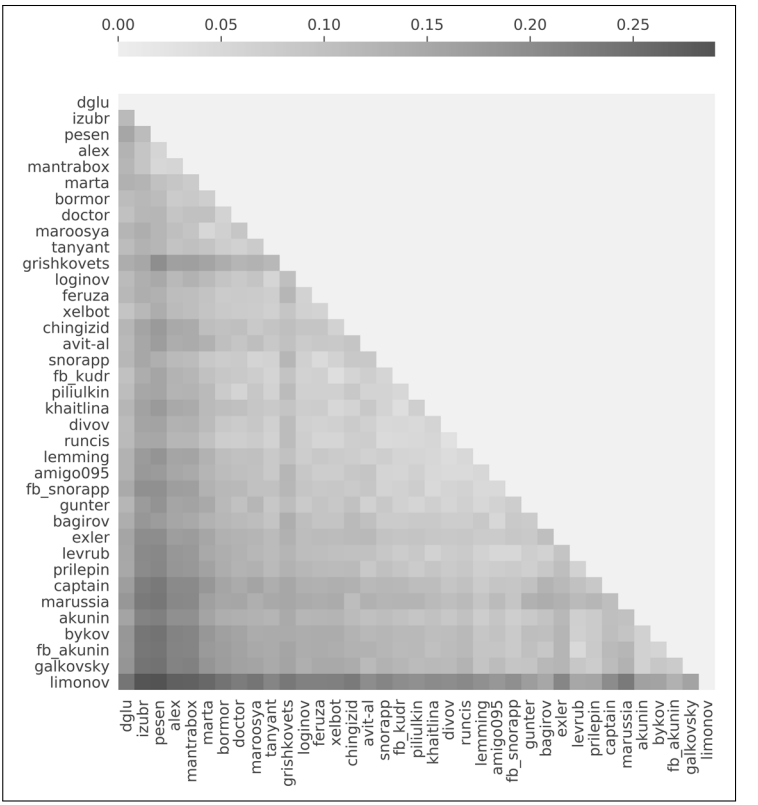
Das Korpus zwischen Politik, Alltag und Literatur

Das durch das »topic modeling« offengelegte Topicspektrum gewährt zwar erste Einblicke in das Gesamtkorpus, allerdings geht daraus nicht hervor, wo einzelne Autorinnen bzw. Autoren thematische Schwerpunkte setzen. Natürlich ließen sich die Topicverteilungen für jeden Webauftritt einzeln anzeigen, diese miteinander zu vergleichen wäre allerdings mit einigem Aufwand verbunden. Deshalb werden weitere quantitative Verfahren hinzugezogen, um mehr über das Gesamtkorpus, vorherrschende Topics bzw. die Beziehungen der Webauftritte untereinander zu erfahren. Ergänzend zur grundlegenden thematischen Einschätzung des Korpus liefert das »topic modeling« auch eine Möglichkeit, Unterschiede und Gemeinsamkeiten der Webauftritte zu berechnen. Dabei wird jeder Webauftritt durch seine Verteilung der fünfzig Topics repräsentiert, also durch einen Zahlenvektor mit fünfzig Dimensionen.

Diese Kodierung erlaubt es, mathematische Abstandsmaße zu verwenden, um Unterschiede bzw. Ähnlichkeiten zwischen zwei Topicverteilungen zu berechnen. Zwei gebräuchliche Abstandsmaße sind euklidischer Abstand und Cosinusähnlichkeit. Ersterer misst die Unterschiede, zweitere die Ähnlichkeiten zwischen zwei Punkten (Manning et al. 2009: 292). Für den vorliegenden Fall wurde empirisch festgestellt, dass die Unterschiede zwischen diesen beiden Maßen kaum ins Gewicht fallen. Verwendet wird deshalb der euklidische Abstand. Je ähnlicher zwei Webauftritte thematisch sind, desto näher sind sich ihre Topics bzw. ihre Topicvektoren in geometrischer Hinsicht. Die berechneten Abstände zwischen den Webauftritten lassen sich in Form einer Abstandsmatrix darstellen, wie in Abbildung 2 gezeigt. Der SPIN-Algorithmus (»Sorting Points into Neighborhoods«) von Tsafrir et al. (2005) wird dabei verwendet, um die Webauftritte so gut wie möglich nach Ähnlichkeit zu sortieren. Diese Sortierung erlaubt es dann, Gruppen ähnlicher Webauftritte über die hellen Flächen in der Abstandsmatrix zu identifizieren.

Die entsprechend beschrifteten Zeilen und Spalten in Abbildung 2 repräsentieren die einzelnen Webauftritte, anhand der Farben lassen sich die Abstände zu den anderen Blogs ablesen. Dementsprechend können die Abstände von Boris Akunins Blog ⟨borisakunin⟩ zu den anderen in der fünften Zeile von unten und in der fünften Spalte von rechts abgelesen werden. Das helle Rechteck rechts unten in der Graphik deutet eine Gruppe ähnlicher Blogs an: ⟨galkovsky⟩, ⟨borisakunin⟩ auf *Facebook*, ⟨rubykov⟩, ⟨borisakunin⟩ im ŽŽ, ⟨levrub⟩ und ⟨prilepin⟩. Thematisch sehr unterschiedlich von allen anderen Blogs ist, wie die dunkle Färbung andeutet, ⟨limonov-eduard⟩ in der untersten Zeile. Auch die Blogs ⟨izubr⟩, ⟨pesen-net⟩, ⟨grishkovets⟩, ⟨exler⟩,

Abbildung 2: Unterschiede zwischen Webauftritten, dargestellt als Abstandsmatrix und durch SPIN nach Ähnlichkeit sortiert



Quelle: G. H.

⟨captain-urthang⟩, ⟨marussia⟩ heben sich von den anderen Blogs ab. Neben einzelnen Gruppen sind auch die ›Eckpunkte‹ von Relevanz, also Ėduard Limonovs Blog ⟨limonov-eduard⟩ einerseits sowie Slava Sės Blog ⟨pesen-net⟩ und Alja Kudrjaševas Blog ⟨izubr⟩ andererseits; Dmitrij Gluchovskijs Blog ⟨dglu⟩ steht zwar ganz links, die hellere Färbung weist seine Position aber als weniger extrem aus. Der Topic-Raum scheint sich also zwischen den Polen ⟨limonov-eduard⟩ einerseits und ⟨pesen-net⟩ sowie ⟨izubr⟩ andererseits aufzuspannen.

Die Abstände für sich genommen gewähren bereits interessante Einblicke, trotzdem ist es wünschenswert, eine Art der Visualisierung zu finden, die Konstellationen zwischen den einzelnen Webauftritten noch deutlicher sichtbar macht. Neben der oben beschriebenen Abstandsmatrix gibt es eine Reihe von elaborierten Ver-

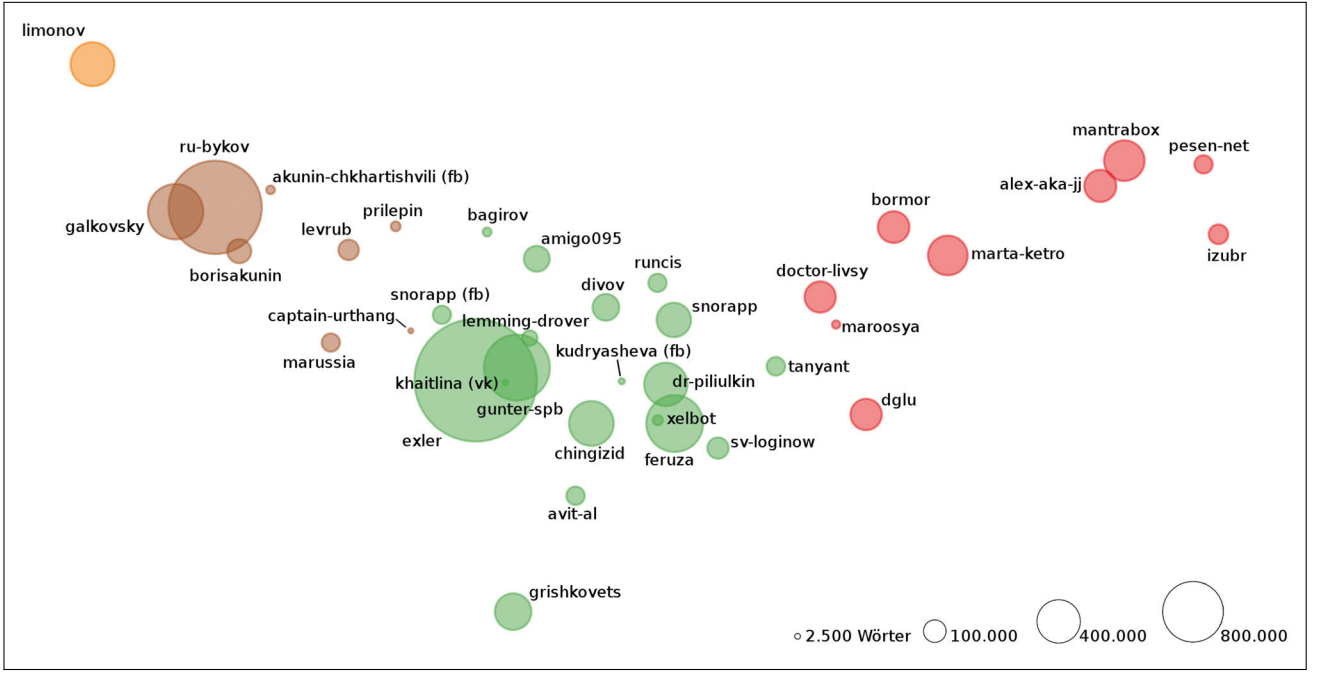
fahren, die hochdimensionale Probleme auf einige wenige Dimensionen projizieren und damit darstellbar machen. Eine gerade im Kontext der Digital Humanities weit verbreitete Variante ist die Hauptkomponentenanalyse (Rehbein 2017: 338). In der vorliegenden Arbeit wird allerdings die von Tenenbaum et al. (2000) entwickelte Isomap verwendet, die die ursprünglichen Abstände zwischen den Webauftritten besser erhält, als dies bei der Hauptkomponentenanalyse möglich ist.

Was zeigt nun die Topic-Karte in Abbildung 3 auf Seite 95? Die Ausnahmepositionen von ⟨limonov-eduard⟩ (links) und ⟨pesen-net⟩ sowie ⟨izubr⟩ (rechts) bestätigen sich in dieser Darstellung. Auch die Abstände zwischen einzelnen Webauftritten bleiben im Großen und Ganzen erhalten. Zu kleinen Verzerrungen kommt es trotzdem, so sind Alja Kudrjaševas Online-Auftritte ⟨khaitlina⟩, ⟨kudryasheva⟩ und ⟨xelbot⟩ zueinander am ähnlichsten, auf der Topic-Karte kommt allerdings Maks Frajs Blog ⟨chingizid⟩ zwischen ihnen zu liegen. Ein Blick in die Abstandsmatrix zeigt, dass dieser Blog Kudrjaševas Auftritten aber zumindest sehr nahe ist. Ein großer Vorteil der Abstandsmatrix ist, Relationen zwischen einzelnen Webauftritten ohne Verzerrung darstellen zu können. Diese Art der Visualisierung ist aber abstrakt und wenig intuitiv. So schlägt sich die Ausnahmeposition von Evgenij Griškovec' Blog ⟨grishkovets⟩ auch in der Matrix nieder. Wo sich dieser Blog in Relation zu den anderen Autorinnen und Autoren befinden, geht aber klarer aus der Topic-Karte hervor. Dafür muss allerdings eine leichte Verzerrung in Kauf genommen werden.

Abschließend erfolgt noch der Versuch, die Webauftritte automatisiert nach Ähnlichkeit in verschiedene Gruppen einzuteilen, also zu »clustern«. Die Aufteilung des Korpus in Untergruppen soll helfen, Ordnung in die Masse der Webauftritte zu bringen. Dementsprechend werden Cluster-Algorithmen auch im Information Retrieval dafür eingesetzt, das Durchsuchen großer Datenmengen zu erleichtern (Manning et al. 2009: 351). Im konkreten Fall der Blogs ist eine klare Aufteilung in Untergruppen nicht zu erwarten, denn die Topicspektren der Autorinnen und Autoren überschneiden sich regelmäßig, was den Cluster-Algorithmen die Arbeit erschwert.

Für das Clustern der Webauftritte wird der K-Means-Algorithmus (MacQueen 1967) verwendet, der eine einfache Methode darstellt, nicht-hierarchische Daten zu gruppieren (Manning et al. 2009: 350). Bei diesem Algorithmus muss vorgegeben werden, in wie viele Gruppen die Daten gegliedert werden sollen (ebd.: 355). Diese Zahl ist für die Blogs vorderhand nicht bekannt; unter Zuhilfenahme etablierter Metriken haben sich vier Gruppen als ein guter Kompromiss erwiesen. Kapitel »Hinter den Kulissen« ab Seite 99 geht genauer auf die verwendeten Clustermetriken und die Herleitung einer optimalen Anzahl von Gruppen ein. Die Aufteilung in vier Gruppen wird in Abbildung 3 auf Seite 95 durch unterschiedliche Farben wiedergegeben. ⟨limonov-eduard⟩ bildet allein eine Gruppe, die zweite Gruppe reicht von ⟨ru-bykov⟩

Abbildung 3: Topic-Karte der 37 Webauftritte; die Farben kennzeichnen die automatisch erstellten Gruppierungen, die Größe der Kreise entsprechen dem jeweiligen Umfang der Webauftritte.



Quelle: G. H.

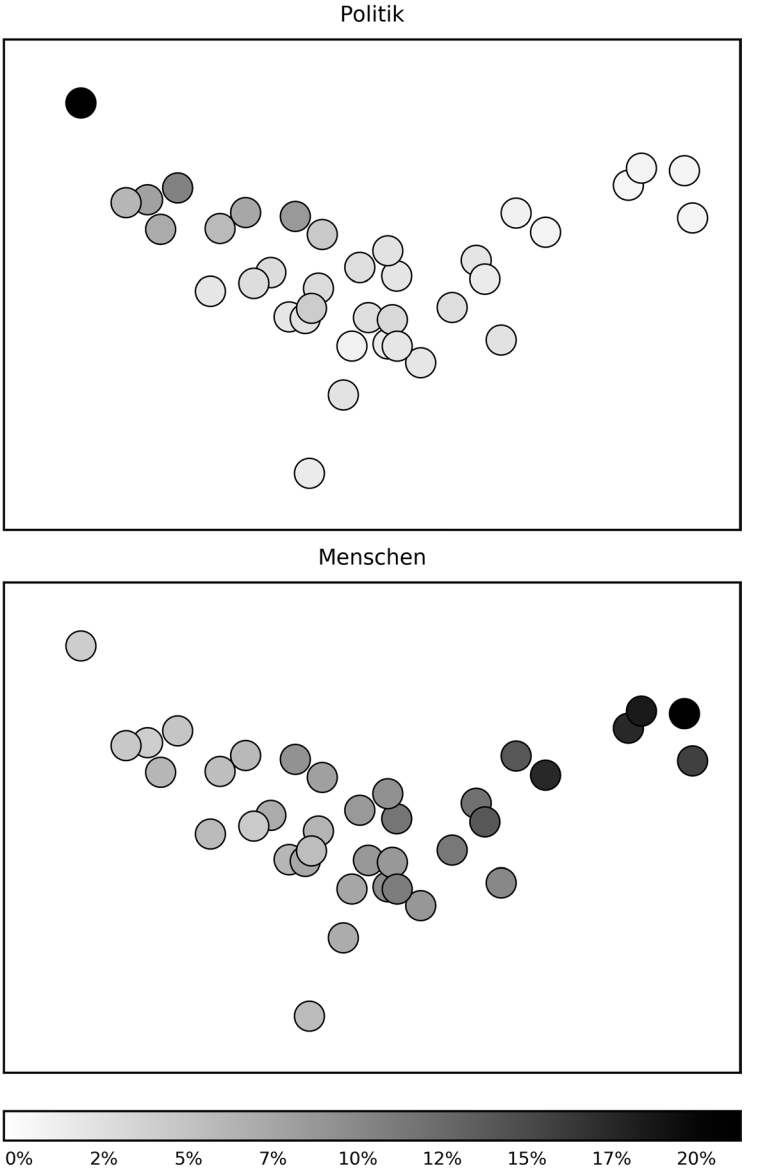
und ⟨galkovsky⟩ bis ⟨prilepin⟩ bzw. ⟨captain-urthang⟩. Im Bereich zwischen ⟨exler⟩ und ⟨bargirov⟩ auf der einen und ⟨tanyant⟩ sowie ⟨sv-loginow⟩ auf der anderen Seite liegt die dritte Gruppe, rechts davon Gruppe vier, die bei ⟨dglu⟩ und ⟨doctor-livsy⟩ beginnt und bis ⟨izubr⟩ bzw. ⟨pesen-net⟩ reicht.

Nun ist zu klären, worin die Eigenheiten dieser vier algorithmisch erstellten Gruppen bestehen. Dafür hat es sich als hilfreich erwiesen, die Intensität jedes Topics in den einzelnen Webauftritten zu visualisieren. In dieser Darstellung können Themen identifiziert werden, die typisch für eine einzige Gruppe sind. Wie Abbildung 4 auf Seite 97 zeigt, ist das Topic *Politik* in den Gruppen eins und zwei dominant. Das Topic *Menschen* ist hingegen in Gruppe vier sehr präsent; wie bereits erwähnt, versammelt dieses Topic zahlreiche literarische Texte. Diese ›tragenden‹ Topics geben bereits eine gewisse Richtung für die Interpretation der Gruppen vor: Die Topic-Karte wird nach links politischer und nach rechts literarischer. Anders formuliert finden sich in den Webauftritten am linken Rand mehr politische Einträge als in jenen am rechten Rand; mit literarischen Texten verhält es sich genau umgekehrt.

Diese erste Einschätzung wird durch die durchschnittlichen Topicverteilungen der einzelnen Gruppen unterstützt. In der ersten Gruppe, die nur aus dem Blog ⟨limonov-eduard⟩ besteht, ist das Topic *Politik* an erster Stelle; weitere politisch ausgerichtete Topics finden sich ebenfalls, beispielsweise *Ukraine* auf Platz sechs. In der zweiten Gruppe kommt *Politik* auf Platz drei, *Ukraine* auf Platz fünf; auch hier sind diese Topics präsenter als im Gesamtkorpus. Beide Gruppen weisen damit eine relative Häufung politischer Einträge auf, es liegt also nahe, sie zu einer einzigen Gruppe zusammenzufassen. Limonovs Blog wird demnach nicht mehr als eigene Gruppe geführt, sondern als Extrembeispiel des politischen Teilkorpus betrachtet.

Auch der vermutete literarische Fokus der vierten Gruppe wird durch die Topicverteilung unterstützt. Neben der verstärkten Präsenz des Topics *Menschen* ist auch das Topic *Literarische Texte* zu nennen, das dort fast doppelt so präsent ist wie im Gesamtkorpus. In den entsprechenden Webauftritten kommen also häufiger literarische Texte vor als im Durchschnitt des Korpus. Schwieriger ist es, den inhaltlichen Schwerpunkt der letzten Gruppe, die sich zwischen den eher literarischen und den eher politischen Webauftritten befindet, zu bestimmen. So lassen sich keine eindeutigen ›tragenden‹ Topics benennen. Die betreffenden Webauftritte scheinen sich eher durch die Abwesenheit eines klaren Überhangs politischer oder literarischer Einträge auszuzeichnen. Ein »close reading« ausgewählter Beiträge zeigt dann, dass diese Webauftritte häufig Privat- oder Berufsleben der jeweiligen Schriftstellerin bzw. des jeweiligen Schriftstellers beleuchten.

Abbildung 4: Die ›tragenden‹ Topics in den politischen und literarischen Webauftritten



Quelle: G. H.

Aufbauend auf diesen Erkenntnissen wird für die nächsten Kapitel folgende Grobeinteilung der Webauftritte in drei Teilkorpora vorgenommen: ein politisches (Gruppe eins und zwei), ein alltägliches (Gruppe drei) und ein literarisches (Gruppe vier). Zu dieser Einteilung ist zu sagen, dass die Grenzbereiche zwischen den einzelnen Gruppen nicht klar definiert sind. Sowohl zwischen politischen und alltäglichen Webauftritten als auch zwischen alltäglichen und literarischen kommt es zu Überlappungen.[8] Selbst unter den beiden Extrema, dem politischen Blog ⟨limonov-eduard⟩ und dem literarischen Blog ⟨izubr⟩, sind unter den häufigsten zehn Topics sechs gleiche, die sich nur in der Gewichtung unterscheiden: *Meinung*, *Metaphysik*, *Zeitangaben/Flüge*, *Menschen*, *Stadt* und *Internet*. Aufgrund dieser Überlappungen kann es durchaus vorkommen, dass in einem Webauftritt des literarischen Teilkorpus Einträge beispielsweise zur russischen Innenpolitik auftreten. Solche politisch ausgerichteten Posts sind aber im politischen Teilkorpus anteilsmäßig stärker vertreten.

Trotz dieser Grauzonen liefert die durch die quantitativen Verfahren aufgeworfene Dreiteilung wertvolle Hinweise zum Korpus in seiner Gesamtheit und macht dieses leichter handhabbar. Insbesondere die Ausrichtung der Webauftritte entlang der thematischen Achse *Politik – Alltag – Literatur*, die die Topic-Karte und das Clustering offenbaren, ist hier von Bedeutung, weil diese aus der simplen Reihung der häufigsten Topics, die das »topic modeling« vornimmt, nicht hervorgeht.

Zusammenfassend lässt sich sagen, dass die im vorliegenden Kapitel vorgeschlagenen Methoden des »distant reading« helfen können, einen umfassenden Überblick über die 37 Webauftritte zu vermitteln. Dabei wird das »topic modeling« eingesetzt, um von der Wortoberfläche abstrakte Topics, also letztlich Gemeinsamkeiten mehrerer Texte, abzulesen. Für jeden Webauftritt wird dann berechnet, welche Topics vorhanden sind; diese Topicverteilungen können genutzt werden, um Ähnlichkeiten zwischen einzelnen Webauftritten zu finden und diese entsprechend dieser Ähnlichkeiten in Gruppen zu unterteilen. Bei jedem dieser Schritte werden sowohl durch genaues Lesen von Einzeltexten als auch durch Querlesen einzelner Webauftritte die Parameter für das »distant reading« optimiert und dessen vorläufige Ergebnisse evaluiert. In dieser Vorgehensweise greifen die von Hayles beschriebenen Vorteile der drei unterschiedlichen Arten des Lesens ineinander und ermöglichen Einblicke, die sich einem reinen »close reading«, einem reinen »hyperreading« oder einem reinen »machine reading« verschließen.

8 | Da die Felder Politik – Alltag – Literatur auch thematisch ineinandergreifen können, verwundert diese Überlappung nicht. Alltägliche Einträge weisen unter Umständen einen impliziten politischen Hintergrund auf, genauso wie Gedichte eine politische Aussage tätigen können.

An dieser Stelle der Arbeit eröffnen sich zwei unterschiedliche Lektürepfade. Leserinnen und Leser, die gleich mit den Detailanalysen der 37 Webauftritte fortfahren möchten, finden diese ab Seite 109. Diejenigen, die sich für die technischen Hintergründe und eine quantitative Evaluation des verwendeten »topic modeling« interessieren, können gleich im Anschluss weiterlesen.

Hinter den Kulissen

Um Transparenz und Nachvollziehbarkeit der vorliegenden Studie zu sichern, werden in diesem Unterkapitel Hintergrundinformationen vermittelt, die es ermöglichen sollen, den quantitativen Teil der Untersuchung zu evaluieren bzw. nachzuprogrammieren.[9] Dementsprechend richtet sich dieser Abschnitt an ein technisch interessiertes Publikum. Die vorliegende Untersuchung von 37 Blogs im Runet basiert auf einer in Python implementierten Skriptsammlung. Einzelne Pakete werden für spezifische Funktionalität herangezogen, beispielsweise Radim Řehůřeks *Gensim*, eine quelloffene Sammlung verschiedener »topic modeling«-Algorithmen (Řehůřek/Sojka 2010). Weiters kommen folgende Pakete zum Einsatz: *scipy* für das Clustering, *scikit-learn* für PCA, Isomap sowie die Berechnung der Silhouettenkoeffizienten, *scrapy* zum Sammeln des Rohmaterials, *pymorphy2* für Part-of-Speech-Tagging russischer Texte, sowie *nltk* für den Snowball-Stemmer. Für SPIN wurde Jonatas Cesars Implementierung auf *Github* verwendet (Cesar 2013).

Um die Ergebnisse des »topic modeling« quantitativ erfassen und vergleichen zu können, muss zunächst definiert werden, was ein ›gutes‹ Ergebnis ist. Das hier vorgestellte System zur Topicmodellierung hat drei unmittelbare Ziele. Vorrangig sollen die modellierten Topics einfach zu interpretieren sein. Zudem soll die Zuordnung von Topics zu einzelnen Texten treffend sein, und die Topicverteilungen sollen es erlauben, einzelne Blogs nach thematischen Gruppen zusammenzufassen. Nachfolgend werden für diese drei Ziele passende Metriken vorgestellt und ausgewertet.

Wie gut wird das erste Ziel, leicht verständliche Topics zu modellieren, erreicht? Die automatisierte Bewertung von automatisch erstellten Topics stellt ein bekanntes Problem dar. Ein Ansatz ist, das Modell in seiner Gesamtheit zu evaluieren, indem dessen Perplexität berechnet wird. Dabei handelt es sich um eine Metrik, die beschreibt, wie gut ein (trainiertes) statistisches Modell auf (unbekannte) Testdaten

9 | Die verwendeten Skripte und Daten sind auf *Github* verfügbar: https://github.com/ghowa/russian-blogs, letzter Aufruf 10. September 2019.

passt. Dieser Zugang hat sich zwar etabliert, misst aber nicht, wie gut sich die Topics von Menschen interpretieren lassen. Wie Chang et al. (2009: 288f.) experimentell belegen, geht eine hohe Perplexität häufig sogar mit besonders unverständlichen Topics einher. Deshalb postulieren sie eine qualitative Bewertung der Topics durch den Menschen, die nach bestimmten Kriterien formalisiert wird (ebd.: 291f.). Mimno et al. (2011: 265) wiederum stellen fest, dass die Anzahl der Kookkurrenzen von Wörtern in einzelnen Dokumenten Aussagen bezüglich der Topicqualität erlaubt, und entwickeln darauf aufbauend das Kohärenzmaß C:

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})}$$

Dabei entspricht $V^{(t)} = (v_1^{(t)}, \ldots, v_M^{(t)})$ der Liste der M wahrscheinlichsten Wörter in Topic t, $D(v)$ der Anzahl der Dokumente, in denen Wort v mindestens einmal vorkommt, und $D(v, v')$ der Anzahl der Dokumente, in denen mindestens einmal sowohl Wort v als auch Wort v' vorkommen (ebd.). Anzumerken ist, dass Mimno et al. zur Überprüfung ihrer Metrik wieder auf menschliches Urteilsvermögen zurückgreifen (ebd.: 263f.).

Um die Aussagekraft dieses Kohärenzmaßes zu prüfen, werden zunächst versuchsweise die Kohärenzen jenes LDA-Laufes mit 50 Topics berechnet, der in der folgenden Analyse zum Einsatz kommt. Verwendet werden dafür die jeweils 20 häufigsten Wörter jedes Themas. In Tabelle 2 auf Seite 101 sind die sieben kohärentesten Topics gelistet, weiters drei Topics mit mittlerer Kohärenz und die drei inkohärentesten. Tatsächlich stimmt die berechnete Kohärenz im Wesentlichen mit der subjektiven Einschätzung überein. Zwar sind ausgerechnet die drei ›kohärentesten‹ Topics etwas kryptisch, die folgenden vier sind dafür einfach zu interpretieren. Nicht immer müssen Kohärenz und offensichtliche Interpretation also Hand in Hand gehen.

Dies zeigt sich auch anhand des Topics *Literarische Texte*, dessen Wörter wie in Abschnitt »Topic Modeling« ab Seite 87 ausgeführt auf den ersten Blick zusammenhanglos zu sein scheinen. Trotzdem weist dieses Topic mit einem Wert von -854 eine hohe Kohärenz auf und landet auf Platz 15. Das Kohärenzmaß ›identifiziert‹ damit ein gutes Thema, während der Mensch auf den ersten Blick zu einer anderen Bewertung kommt. Erst der Blick in für dieses Topic typische Texte zeigt dem menschlichen Auge, dass hier tatsächlich ein roter Faden zu finden ist.

Die Topics mit mittlerer Kohärenz lassen sich zwar noch interpretieren, allerdings offenbaren sich immer wieder Anzeichen problematischer Topicmodellierung. Dazu zählen »intruders«, also Wörter, die nicht zu den anderen passen (ebd.: 264); ein Beispiel ist das Wort »Tunnel« im Tier-Topic (Platz 26). Die vier Topics auf den

Tabelle 2: Kohärenzen von 13 ausgewählten Topics mit den jeweils häufigsten Wörtern; gereiht wird nach fallender Kohärenz

1.	-525	дело вопрос история случай работа человек место проблема отношение. . .
		Sache, Frage, Geschichte, Fall, Arbeit, Mensch, Platz, Problem, Beziehung
2.	-547	человек жизнь друг мир слово дело бог любовь сила смысл
		Mensch, Leben, Freund, Welt, Wort, Sache, Gott, Liebe, Kraft, Sinn
3.	-602	рука глаз женщина лицо мужчина голова нога девушка муж девочка
		Hand, Auge, Frau, Gesicht, Mann, Kopf, Bein, junge Frau, Ehemann, Mädchen
4.	-667	фильм кино режиссёр актёр герой роль картина сценарий сериал сцена
		Film, Kino, Regisseur, Schauspieler, Held, Rolle, Bild, Szenario, Serie, Szene
5.	-689	власть россия путин гражданин президент страна выбор партия политика. . .
		Macht, Russland, Putin, Bürger, Präsident, Land, Wahl, Partei, Politik
6.	-702	россия война украина страна европа народ государство ссср сша территория
		Russland, Krieg, Ukraine, Seite, Europa, Volk, Staat, UdSSR, USA, Territorium
7.	-734	книга автор роман литература писатель читатель текст книжка рассказ...
		Buch, Autor, Roman, Literatur, Schriftsteller, Leser, Text, Büchlein, Erzählung...
24.	-1143	иван король дракон господин рыцарь принцесса царь ефрем принц меч
		Ivan, König, Drache, Herr, Ritter, Prinzessin, Zar, Ephraim, Prinz, Schwert
25.	-1257	камент кнопка колонка устройство просмотр тролль журналистика. . .
		Comment, Knopf, Kolumne, Organisation, Durchsicht, Troll, Journalistik
26.	-1298	собака кошка даша собачка котик котёнок туннель птичка щенок животное
		Hund, Katze, Daša, Hündchen, Katerchen, Kätzchen, Tunnel, Vögelchen, Welpe, Tier
48.	-1562	сказка обзор плата дерьмо властелин алёна доклад рыбка гаджет каток
		Märchen, Übersicht, Lohn, Scheiße, Herr, Alena, Vortrag, Fisch, Gadget, Eislaufplatz
49.	-1572	диск форум прилепин паспорт захар папка роммель пират тандем документ
		Scheibe, Forum, Prilepin, Pass, Zachar, Ordner, Rommel, Pirat, Tandem, Dokument
50.	-1617	картинка метка вильнюс удальцов испания комиссар консервы травля. . .
		Bild, Zeichen, Vilna, Udal'cov, Spanien, Kommissar, Konserven, Hetze

Quelle: G. H.

Plätzen 48 bis 50 weisen schließlich keinen sichtbaren inneren Zusammenhang mehr auf, die Zusammenstellung der Wörter wirkt großteils zufällig. Tatsächlich lässt sich hier im Unterschied zum Topic *Literarische Texte* auch nach Durchsicht typischer Texte kein innerer Zusammenhang feststellen. Für das am schlechtesten bewertete Topic wird allerdings nach Begutachtung der dazugehörigen Texte deutlich, dass auch ein fehlender Zusammenhang einen Zusammenhang bilden kann. Konkret gehören zu diesem Topic Texte, die zwar auf kyrillisch verfasst worden sind, nicht aber auf russisch; beispielsweise tauchen hier serbische Texte auf, oder solche, die einer Phantasiesprache entspringen.

Die von Mimno et al. postulierte Kohärenz stimmt damit im vorliegenden Fall mit der subjektiven Einschätzung der Topicqualität großteils überein. Zu erwähnen ist, dass die zehn häufigsten Topics des Korpus (Abbildung 1) allesamt zu den kohärentesten vierzehn Topics gehören. Damit bieten die gezeigten Topicverteilungen eine solide Grundlage für eingehendere Interpretationen. Trotz dieses positiven Eindruckes ist es schwierig, die Ergebnisse verschiedener Experimente über deren Kohärenzwerte zu vergleichen. Die Ergebnisse in Tabelle 3 auf Seite 102 zeigen, dass mit steigender Topicanzahl die Kohärenz immer schlechter wird. Dies scheint dem subjektiven Empfinden entgegengesetzt zu sein, das 50 Topics den Vorzug über 30

Tabelle 3: Ergebnisse der LDA im Vergleich. Oben Kohärenz, unten Silhouettenkoeffizient für eine Gruppierung in vier Gruppen. Hervorgehoben ist der nachfolgend verwendete Lauf mit 50 Topics

	Topics	*1. Lauf*	*2. Lauf*	*3. Lauf*	*4. Lauf*	*5. Lauf*	∅
Kohärenz	30	-1097	-1064	-1108	-1119	-1064	-1090
	50	-1162	-1178	-1159	-1157	-1163	-1163
	75	-1254	-1243	-1267	-1233	-1266	-1252
	100	-1288	-1288	-1269	-1283	-1292	-1284
Silhouetten	30	0.27	0.16	0.30	0.22	0.26	0.24
	50	0.20	0.27	0.25	0.27	0.29	0.26
	75	0.19	0.23	0.22	0.24	0.17	0.21
	100	0.22	0.29	0.25	0.25	0.26	0.25

Quelle: G. H.

gibt. Ein Beispiel aus einem LDA-Lauf mit 30 Topics sei hier angeführt, das eine (im Vergleich sehr gute) Kohärenz von -585 aufweist: *рука, глаз, дом, голова, нога, лицо, ночь, минута, дверь, свет, голос, друг, стол, сон...* [*Hand, Auge, Haus, Kopf, Bein, Gesicht, Nacht, Minute, Tür, Licht, Stimme, Freund, Tisch, Schlaf*]. Subjektiv kann die hohe Kohärenz nicht ganz nachvollzogen werden, vermischen sich doch in diesem Topic drei Wortfelder, nämlich *Menschen*, *Wohnen* und *Zeit*.

Der Grund für dieses Auseinanderklaffen von subjektiver und objektiver Einschätzung ist schnell gefunden. Das Kohärenzmaß untersucht, wie viele Wörter eines Themas wie häufig gemeinsam in Texten auftreten. Mit steigender Topicanzahl werden die Topics immer kleiner. Während 30 Topics durchschnittlich aus 346 Wörtern bestehen, schrumpft diese Zahl bei 100 Topics um ein Drittel auf 224 Wörter. Je weniger Wörter in einem Topic sind, desto unwahrscheinlicher werden Kookkurrenzen dieser Wörter. Aus diesem Grund ist die Kohärenz nicht für einen Vergleich zwischen den LDA-Läufen mit unterschiedlicher Anzahl von Topics geeignet. Auf Tabelle 3 umgelegt, bedeutet dies, dass die Kohärenz nur für einen zeilenweisen Vergleich taugt. Dieser belegt, dass die Unterschiede zwischen verschiedenen Läufen mit gleicher Topicanzahl sehr gering sind.

Ähnlich schwierig ist es, das zweite Ziel zu quantifizieren, also die Zuordnung der Topics zu gewissen Texten. Indem die typischsten Texte für jedes Topic angezeigt werden, kann dieses Problem als eine klassische Fragestellung aus dem Information Retrieval aufgefasst werden. Die dafür gängigen Maßzahlen heißen Präzision (»precision«) und Vollständigkeit (»recall«), wobei erstere beschreibt, wie viele der gefundenen Texte relevant sind, und zweitere, wie viele der relevanten Texte gefunden worden sind (Manning et al. 2009: 155f.). Beide Zahlen werden häufig kombiniert,

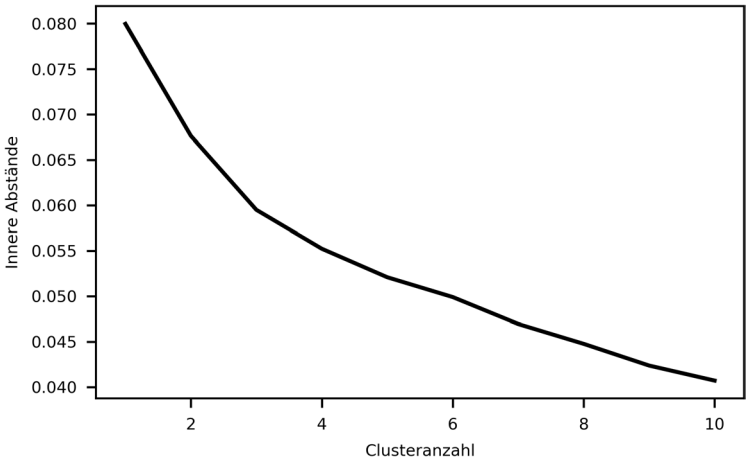
indem ihr harmonisches Mittel als F-Maß (»F measure«) bezeichnet wird (ebd.: 156). Das Problem im vorliegenden Fall der Blogs ist, dass nur die Präzision berechnet werden kann; wieder basierend auf menschlicher Einschätzung.

Für die Ermittlung der Vollständigkeit muss hingegen die Anzahl der relevanten Texte bekannt sein, ein Mensch müsste also zunächst manuell die Topics aller Einträge bestimmen. Dies ist aufgrund des Korpusumfangs nicht möglich. Eine Alternative würde darin bestehen, Präzision und Vollständigkeit nur für einen Webauftritt zu berechnen, doch selbst dann wäre der manuelle Aufwand noch sehr hoch. Für aussagekräftige Ergebnisse raten Buckley/Voorhees (2000: 39), mindestens 50 Topics zu überprüfen, bei 50 modellierten kommen damit sämtliche Topics zum Einsatz. Allerdings sind einige Topics, wie eben gezeigt worden ist, sehr inkohärent, weshalb es auch sinnlos ist, für diese relevante Texte zu suchen. Selbst wenn diese Probleme überwunden werden könnten, wäre die Aussagekraft eines einzelnen F-Maßes gering. Diese käme erst zum Tragen, wenn verschiedene LDA-Läufe miteinander verglichen werden. Für jeden LDA-Lauf müssten dann aber erneut die Topics interpretiert und dann die Einträge eines Webauftritts entsprechend zugeordnet werden.

Die Zuordnung der Topics zu einzelnen Texten kann aufgrund des großen Aufwandes nur kursorisch überprüft werden. Dafür wird die Präzision des LDA-Laufes berechnet, auf dem die Detailanalysen der folgenden Kapitel beruhen. Dies soll ein Gefühl dafür vermitteln, ob das System offensichtliche Fehler aufweist. Die daraus gewonnene subjektive Einschätzung wird durch die Einzelanalysen der folgenden Kapitel ergänzt, in denen über die Resultate des »topic modeling« relevante Texte gesucht und einem prüfenden »close reading« unterzogen werden. Als Test erfolgt die Berechnung der jeweils 30 typischsten Texte für die 15 kohärentesten Topics; darunter sind gleichzeitig die zehn häufigsten. Dabei fällt auf, dass die Topiczuordnung für Texte mit weniger als fünf Nomina nicht mehr funktioniert. Zu beachten ist, dass dieses Problem auch Auswirkungen auf die Berechnung der Topicverteilung eines Webauftritts in seiner Gesamtheit hat, die ermittelten Topicverteilungen der Einzeltexte müssen nämlich dementsprechend gewichtet werden. Je länger ein Text ist, desto mehr soll er sich auf die Gesamtstruktur der Topics auswirken.

Wie hoch ist nun die Präzision des LDA-Laufes für die 15 kohärentesten Topics? Da pro Topic die 30 Texte überprüft werden, die laut »topic modeling« am relevantesten sind, muss die Relevanz von insgesamt 450 Texten subjektiv bestimmt werden. 420 Texte passen thematisch tatsächlich, 30 Texte werden falsch zugeordnet, was einer Relevanz von 93% entspricht. Bei fünf Topics erreicht diese Quote sogar 100%: *Politik*, *Ukraine*, *Natur*, *Künstlerinnen und Künstler* sowie *Geld*. Mit 70% am schlechtesten schneidet das Topic *Meinungsäußerungen* ab, wohl nicht zuletzt deshalb, weil es etwas schwammig formuliert ist. Einige der falschen Zuordnungen lassen sich auf

Abbildung 5: Durchschnittlicher innerer Abstand für verschiedene Clusteranzahlen



Quelle: G. H.

die Tatsache zurückführen, dass Wörter mehrere Bedeutungen haben können, was durch die LDA nicht immer erfasst wird. Insgesamt ist die thematische Zuordnung jedoch treffend.

Um das letzte Ziel, die Gruppierung der Webauftritte, bewerten zu können, müssen zwei Dinge voneinander getrennt betrachtet werden. Zunächst gilt es, die optimale Clusteranzahl zu ermitteln, anschließend kann die Clusterqualität bestimmt werden. Für die Anzahl der Cluster gibt es eine einfache Faustregel, laut Mardia et al. (1979: 365) seien für n Datenpunkte $\sqrt{n/2}$ Cluster sinnvoll. Im konkreten Fall der 37 Webauftritte ergibt diese Gleichung vier Cluster. Eine weitere einfache Variante, um die optimale Clusteranzahl abzuschätzen, ist die sogenannte Ellenbogenmethode. Grundgedanke ist dabei, die Anzahl an Clustern zu identifizieren, ab der sich der durchschnittliche innere Abstand der Punkte in den Clustern nicht mehr wesentlich verringert, wo der Funktionsgraph der Abstände also einen Knick aufweist (Thorndike 1953: 274-276). Abbildung 5 zeigt den durchschnittlichen inneren Abstand in Abhängigkeit von der Clusteranzahl; der ›Ellbogen‹ lässt sich insbesondere aufgrund des Knicks bei sechs Clustern nicht eindeutig bestimmen; eine Möglichkeit wäre, ihn bei vier Clustern anzusetzen. Die von der Faustregel ins Spiel gebrachten vier Cluster erscheinen damit auch im Kontext der Blogs als optimal. Diese Aufteilung entspricht den drei Gruppen Politik – Alltag – Literatur, wobei der Ausnahmefall ⟨limonov-eduard⟩ als eigene Gruppe geführt wird.

Um die Qualität der Cluster zu ermitteln, kommt schließlich der Silhouettenkoeffizient zum Einsatz. Dieser misst, wie kompakt die gefundenen Cluster sind und wie gut sie sich voneinander abgrenzen lassen; die ›wahren‹ Cluster müssen als Vergleichswert nicht bekannt sein. Rousseeuw (1987: 55f.) definiert den Silhouettenkoeffizienten wie folgt:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Wenn i ein Datenpunkt ist, der Cluster A zugerechnet wird, dann bezeichnet $a(i)$ den durchschnittlichen Abstand von i zu allen anderen Punkten in A. $b(i)$ ist der durchschnittliche Abstand zu den Punkten des nächstgelegenen Clusters B. Die Werte von $s(i)$ reichen von -1 bis 1. -1 heißt, i liegt im Zentrum des nächstgelegenen anderen – also falschen – Clusters, bei 0 liegt i genau zwischen dem eigenen und dem nächstgelegenen anderen Cluster, und bei 1 ist i das Zentrum des eigenen Clusters. Nach Kaufman/Rousseeuw (2005: 88) kann der Silhouettenkoeffizient folgendermaßen interpretiert werden: Unter einem Wert von 0,25 existiert keine Struktur, von 0,26 bis 0,5 eine schwache, von 0,51 bis 0,7 eine mäßige, und ab 0,71 eine starke. Diese Einschätzung ist jedoch rein subjektiv und sollte dementsprechend *cum grano salis* genommen werden.

Für 20 LDA-Läufe, davon jeweils fünf mit 30, 50, 75 und 100 Topics, wird ein K-Means-Clustering in vier Gruppen durchgeführt. Im Anschluss daran kann der Silhouettenkoeffizient berechnet werden. Die Ergebnisse dieser Berechnung sind in Tabelle 3 auf Seite 102 zusammengefasst. Insgesamt liegen die Werte eng beieinander und implizieren, es existiere keine Struktur. Einzelne Läufe erreichen eine schwache Struktur, darunter auch der für die Analyse ausgewählte 3. Lauf mit 50 Topics. Eine Normalisierung der Topicvektoren führt zu keiner Verbesserung des Silhouettenkoeffizienten. Dies erscheint nachvollziehbar, weil durch die Normalisierung zwar die Relationen zwischen den Topics erhalten bleiben, aber verloren geht, wie stark die einzelnen Topics vertreten sind. Auffällig ist, dass erneut bei 30 Topics die besten Ergebnisse erzielt werden. Hier zeigt sich der »curse of dimensionality«, häufig geht mit mehr Datendimensionen eine Verschlechterung des Klassifizierungsergebnisses einher (Beyer et al. 1999: 231f.).

Die im vorigen Kapitel geäußerte Vermutung, die thematischen Überschneidungen verhinderten eine klare Gruppierung der Blogs, erweist sich als zutreffend. Nur die Aufteilung in zwei Gruppen, sprich: ⟨limonov-eduard⟩ und die restlichen Blogs, erreicht Silhouettenkoeffizienten um 0,5 und kann damit als ›starke‹ Struktur verstanden werden. Das Clustering hat also Probleme, die subtilen Unterschiede zwischen den schriftstellerischen Blogs zu erfassen. Große Differenzen, wie sie zwi-

schen Limonovs fast ausschließlich politischen Themen gewidmetem Blog und allen anderen Blogs auftreten, werden aber problemlos erkannt und markiert.

Wie aus Tabelle 3 auf Seite 102 hervorgeht, ergeben sich prinzipbedingte Unterschiede zwischen verschiedenen LDA-Läufen. Allerdings überwiegen die Kontinuitäten, wie ein Vergleich der Abstandsmatrizen von vier LDA-Läufen mit 50 Topics gzeigt hat. Die in Abbildung 2 auf Seite 93 ersichtlichen dunklen Linien, die die Webauftritte von Ėduard Limonov, Alja Kudrjaševa, Slava Sė, Evgenij Griškovec, Aleksej Ėksler, Nik Perumov und Marusja Klimova als jeweils relativ weit entfernt von anderen Webauftritten markieren, waren auch in den anderen drei Testläufen erkennbar. Ebenso bleiben die Regionen sehr ähnlicher Webauftritte erhalten, beispielsweise die im vorherigen Unterkapitel erwähnte politische Gruppe rechts unten.

Häufig kommt es dazu, dass ähnliche Topics in verschiedenen LDA-Läufen Variationen in der Zusammensetzung ihrer Wörter aufweisen oder dass zwei sehr ähnliche Topics in einem anderen Lauf als ein einziges gewertet werden. Die Reihenfolge der Topics insgesamt verändert sich ebenfalls von Lauf zu Lauf, die thematischen Abstände zwischen den Webauftritten und damit auch die Beziehungen der Webauftritte zueinander bleiben aber größtenteils konstant. Zudem haben stichprobenartige Einblicke während der Durchführung von über 60 LDA-Läufen gezeigt, dass das Topicspektrum tatsächlich relativ konstant bleibt. Dies gilt besonders für auffällige Positionierungen, wie der von Politik dominierte Blog ⟨limonov-eduard⟩ oder die Kochrezepte bei ⟨chingizid⟩. Es ist daher im Kontext der vorliegenden Untersuchung ausreichend, einen LDA-Lauf als beispielhaft auszuwählen und diesen für die weitere Interpretation zu verwenden.

Zusammenfassend lässt sich bemerken, dass keines der drei Ziele Topicqualität, Relevanz der Topiczuordnung und Clustering rein objektiv bewertet werden kann. Wie in diesem Kapitel herausgearbeitet, spielen für alle drei Beispiele auch in informatischer Fachliteratur subjektive Einschätzungen eine große Rolle, sei es als Rückfallebene für den Kohärenzwert, bei der Feststellung der Relevanz von Suchergebnissen oder bei der Ellbogen-Methode. Eine Kombination subjektiver und objektiver Kriterien ist deshalb anzustreben. Klar definierte Metriken können helfen, subjektive Einschätzungen zu formalisieren, vergleich- und wiederholbar zu machen. Dabei dürfen die inhärenten Probleme verschiedener Berechnungsmethoden, etwa die Abhängigkeit der Kohärenz von der Anzahl der Topics, nicht übersehen werden.

Nicht verschwiegen werden soll, dass die Cluster-Struktur der ermittelten Gruppen vom Silhouettenkoeffizienten immer als schwach bewertet wird. Dies ist auf die starken thematischen Überschneidungen der einzelnen Webauftritte zurückzuführen. Aus subjektiver Perspektive erscheint die automatisch erstellte Aufteilung

jedenfalls nachvollziehbar zu sein, weshalb sie für die Gliederung der folgenden Kapitel herangezogen wird.