Erschienen in: Konopka, Marek/Wöllstein, Angelika (Hrsg.): Grammatische Variation. Empirische Zugänge und theoretische Modellierung. - Berlin/Boston: de Gruyter, 2017. S. 331-343. (Jahrbuch des Instituts für deutsche Sprache 2016)

Felix Bildhauer (Mannheim)/Roland Schäfer (Berlin)

Induktive Topikmodellierung und extrinsische Topikdomänen[1]

Abstract: Eine reichhaltige Auszeichnung mit Metadaten ist für alle Arten von Korpora für die linguistische Forschung wünschenswert. Für große Korpora (insbesondere Webkorpora) müssen Metadaten automatisch erzeugt werden, wobei die Genauigkeit der Auszeichnung besonders kritisch ist. Wir stellen einen Ansatz zur automatischen Klassifikation nach Themengebiet (*Topikdomäne*) vor, die auf dem lexikalischen Material in Texten basiert. Dazu überführen wir weniger gut interpretierbare Ergebnisse aus einer so genannten *Topikmodellierung* mittels eines überwachten Lernverfahrens in eine besser interpretierbare Kategorisierung nach 13 Themengebieten. Gegenüber (automatisch erzeugten) Klassifikationen nach Genre, Textsorte oder Register, die zumeist auf Verteilungen grammatischer Merkmale basieren, erscheint eine solche thematische Klassifikation geeigneter, um zusätzliche Kontrollvariablen für grammatische Variationsstudien bereitzustellen. Wir evaluieren das Verfahren auf Webtexten aus DECOW14 und Zeitungstexten aus DeReKo, für die jeweils getrennte Goldstandard-Datensätze manuell annotiert wurden.

1 Motivation

Dank der Verfügbarkeit von viele Milliarden Wörter großen Korpora ist die Bedeutung von Korpusdaten in vielen Bereichen der Linguistik weiter gestiegen. Solche Korpora eröffnen zum Beispiel die Möglichkeit, seltene Phänomene zu untersuchen, ohne dafür Daten experimentell erheben oder auf Introspektion zurückgreifen zu müssen. Bei der korpuslinguistischen Untersuchung eines sprachlichen Phänomens werden aber oft auch Informationen *über* die Texte, aus denen die Belege stammen, einbezogen – sei es, weil das Hauptinteresse auf sprachlichen Varietäten liegt, sei es, weil man bei der Erforschung innersprachlicher Bedingungen für Varianz bestimmte Texteigenschaften kontrollieren möchte. Solche *Metadaten* können ganz unterschiedliche Aspekte beschreiben: neben sozio-

1 Roland Schäfers Beitrag zu dieser Arbeit wurde finanziert durch die Deutsche Forschungsgemeinschaft (DFG, SCHA1916/1-1).

demographischen Angaben zum Verfasser oder zur Verfasserin des Texts auch Information über die Kommunikationssituation, die beabsichtigte Wirkung des Texts, das Textthema usw.

Das Fehlen solcher Metadaten in sehr großen gecrawlten Webkorpora wird gelegentlich kritisch gesehen (z.B. Leech 2007). Dabei sollte man allerdings nicht außer Acht lassen, dass die gewünschten Metadaten auch für traditionelle Korpora oft nicht vorliegen. Dies gilt insbesondere für abstrakte Kategorien wie *Register*, *Genre*, *Textsorte* usw.[2] Obwohl die Varianz sprachlicher Phänomene in Abhängigkeit von solchen Kategorien seit Jahrzenten Gegenstand der korpuslinguistischen Diskussion ist, gibt es bis heute keine allgemein akzeptierten Definitionen für diese Begriffe. Der fehlende Konsens macht sich wie zu erwarten auch bei der Erstellung von Taxonomien bemerkbar. Schon für klassische Medien konnte kein einvernehmliches Inventar von Genres gefunden werden, und das Aufkommen neuer Textformen im WWW macht die Situation nur komplexer (vgl. die Beiträge in Mehler/Sharoff/Santini (Hg.) 2010). Dies führt dazu, dass bei einer manuellen Klassifikation von Texten die Übereinstimmung zwischen Annotatorinnen oft unbefriedigend ist. Daher verwundert es nicht, dass auch die automatische Klassifikation von Genres – zumal für Webdaten – selbst in rezenten Experimenten nur unbefriedigende Ergebnisse liefert. So berichten Biber/Egbert (2016), dass ihr automatischer Klassifizierer eine Genauigkeit von 42,1% auf 32 Kategorien aufweist.

Darüber hinaus verwenden Methoden zur automatischen Klassifikation von Genres meist sprachliche (oft grammatische) Merkmale als Grundlage. Während dies für manche praktischen (typischerweise nicht-linguistischen) Anwendungen kein Problem darstellt, ist es konzeptuell fragwürdig, Korpora für die linguistische Forschung mit Metadaten auszuzeichnen, die über das (gemeinsame) Auftreten grammatischer Phänomene definiert wurden. Sobald eines der zur Klassifikation verwendeten Phänomene (oder ein anderes, mit ihm korrelierendes) anhand dieses Korpus linguistisch untersucht wird, droht Zirkularität: Ein Phänomen P tritt häufiger in Genre G auf, aber die Dokumente wurden unter anderem als zu Genre G zugehörig klassifiziert, weil Phänomen P in ihnen häufig auftritt.

Eine andere Möglichkeit ist die Klassifikation von Dokumenten nach ihrem Thema. Dies ist im Prinzip orthogonal zur Klassifikation nach Genres, obwohl es plausibel ist, von deutlichen Abhängigkeiten zwischen Textthema und Genre

2 Im Folgenden verwenden wir „Genre“ stellvertretend für Kategorien, die in der Literatur oft auch als „Register“ oder „Textsorte“ o.Ä. behandelt werden.

auszugehen. Informationen über das Textthema[3] können auch für Untersuchungen linguistischer Phänomene relevant sein. Gleichzeitig ist bei einer Klassifikation auf der Basis von Inhaltswörtern das Problem der Zirkularität geringer, zumindest wenn anhand des thematisch klassifizierten Korpus vorwiegend grammatische Phänomene untersucht werden.[4] Die thematische Zusammensetzung eines Korpus ist zudem ein gutes Kriterium, anhand dessen Korpora miteinander verglichen werden können.

Die wichtigste Frage bei der thematischen Erschließung von Korpora ist die nach der verwendeten Taxonomie. Auch hier gibt es keinen Konsens, sondern verschiedene nicht miteinander kompatible Klassifikationssysteme, wie schon Sinclair/Ball (1996) anmerken. Eine weitere Schwierigkeit ist, dass oft auch innerhalb eines Klassifikationssystems eine eindeutige Zuordnung eines Textes zu genau einem Thema nicht möglich ist, weil Texte häufig verschiedene Themen einer gegebenen Taxonomie kombinieren. Erschwerend kommen Themenwechsel innerhalb eines Textes hinzu.

Um der Beliebigkeit bei der Erstellung einer Thementaxonomie entgegenzuwirken, bietet sich eine Kombination von *externen* und *internen* Klassifikationskriterien an (Sinclair/Ball 1996). Ein internes Kriterium ist das im Text auftretende lexikalische Material. Ein datengetriebenes Aufdecken von so genannten *Topiks* (relativ speziellen Einzelthemen) ist auf Basis dieses lexikalischen Materials objektiv, aber die resultierenden Kategorien kommen ohne aussagekräftige Bezeichnungen und müssen erst einmal inhaltlich interpretiert werden (vgl. die Ausführungen zur Topikmodellierung unten sowie Tabelle 1), was jedoch in vielen Fällen schwer fallen dürfte. Nimmt man an, dass für die linguistische Forschung die Interpretierbarkeit von thematischen Kategorien oft wichtig ist, sollte ein Kompromiss zwischen objektiver datengetriebener Klassifikation und Interpretierbarkeit gefunden werden. Eine Möglichkeit besteht darin, eine externe Taxonomie so auszurichten, dass ihre Kategorien möglichst gut mit lexikalischem Material korrespondieren, und sie sich damit auch möglichst gut für eine automatische Klassifikation eignen.

3 Wir verwenden „Textthema“ und „Topik“ gleichbedeutend, d.h. „Topik“ hat hier nichts mit dem Begriff „(Satz-)topik“ zu tun, der in der Literatur zur Informationsstruktur eine Rolle spielt. Auch mit „Topikdomäne“ ist hier nicht der informationstrukturelle Begriff gemeint.

4 Thematische Kategorien können bei der statistischen Modellierung von Variationsphänomenen in verschiedener Weise einbezogen werden. Einführend zu gemischten Modellen und ihren Verwendungsmöglichkeiten in der Linguistik siehe z.B. Gries (2015).

Dieser Artikel geht daher der Frage nach, in welchem Maße extern definierte grobe Themenbereiche (*Topikdomänen*) mithilfe von unüberwacht generierten Topiks automatisch erschlossen werden können. Das Ziel ist eine extern motivierte, für die linguistische Forschung attraktive Taxonomie, die gleichzeitig objektiv in lexikalischen Verteilungen verankert ist und eine geeignete Basis für eine akkurate automatische Klassifikation darstellt. Dieser Ansatz ist als solcher nicht neu. Selbst für einen Teil der von uns verwendeten Daten wurde ein ähnlicher Ansatz in Weiß (2005) bereits verfolgt. Unsere Methode unterscheidet sich von Weiß (ebd.) jedoch in mehreren Punkten. So verwenden wir ein Clusteringverfahren, das dezidiert für die Verarbeitung von Sprache entwickelt wurde. Wir kombinieren darüber hinaus Korpora sehr unterschiedlicher Art, und wir führen eine Evaluation durch, die den üblichen Standards der Computerlinguistik genügt. Es geht hier ausdrücklich nicht darum, neue Algorithmen zu entwickeln. Vielmehr sollen etablierte Verfahren kombiniert und für den Aufbau von Korpora für die linguistische Forschung nutzbar gemacht werden.

2 Vorgehen

Unser Experiment gliedert sich in drei Schritte. Zunächst wird eine Stichprobe von 1756 Dokumenten manuell nach Topikdomänen (thematischen Großbereichen) als Goldstandard-Datensatz annotiert. Im zweiten Schritt werden unabhängig von diesen Annotationen mithilfe eines unüberwachten Verfahrens (*Topikmodellierung*) individuelle *Topiks* (nicht Topik*domänen*) aufgedeckt. Die Charakterisierungen der einzelnen Dokumente in Bezug auf diese Topiks dienen im dritten Schritt als Trainingsdaten für ein überwachtes Lernverfahren, bei dem Dokumente den zuvor manuell annotierten Topikdomänen zugeordnet werden. Dabei kombinieren wir verschiedene Varianten der Korpusvorverarbeitung mit verschiedenen Parametern bei der Topikmodellierung und beim überwachten Lernen.

3 Daten und Goldstandard

Die Daten für unsere Untersuchung wurden zwei verschiedenen Korpora entnommen: 870 Dokumente stammen aus DECOW14A, einem Korpus aus gecrawlten HTML-Dokumenten aus dem WWW (ca. 17 Mio. Wörter; Schäfer/Bildhauer 2012; Schäfer 2015). Weitere 886 Dokumente stammen aus DeReKo 2014-II, das überwiegend Zeitungstexte enthält (ca. 28 Mio. Wörter; Kupietz et al. 2010). Die Auswahl der Korpora ist naheliegend, weil neben Gemeinsamkeiten auch deutliche

Unterschiede zwischen den Korpora in der Verteilung von Topiks zu erwarten sind. Dass diese Annahme zutreffend ist, zeigt Abbildung 1. Eine wichtige Frage ist, ob ein heterogener Datensatz für unser Verfahren geeignet ist oder ob Daten aus derart verschiedenen Korpora besser getrennt verarbeitet werden.

Die Dokumente wurden manuell nach dem COWCat-Klassifikationsschema[5] (Schäfer/Bildhauer 2012) annotiert, das wiederum auf Arbeiten von Sharoff (2006) aufbaut. Zielgröße beim Aufbau des Schemas war eine moderate Anzahl von ca. 10–20 Topikdomänen. Die Taxonomie wurde in mehreren Annotationsdurchläufen entwickelt und angepasst. Die von uns hier verwendete Version umfasst die folgenden 13 Kategorien:

- Science
- History
- Philosophy
- Public-Life-and-Infrastructure
- Individuals
- Law
- Life-and-Leisure
- Technology
- Business
- Beliefs
- Politics-and-Society
- Medical
- Fine-Arts

Abbildung 1 zeigt die Verteilung dieser Kategorien in den manuell annotierten Stichproben. Die Schriftgröße spiegelt den Anteil der jeweiligen Kategorie wider. In beiden Korpora dominieren Texte über Freizeitthemen (*Life and Leisure*). In DECOW14A (links) sind darüber hinaus Texte über Kunst oder Wirtschaft häufig, während in DeReKo-Texten Themen aus den Bereichen Politik, Gesellschaft und öffentliche Einrichtungen bzw. Infrastruktur vorherrschen.

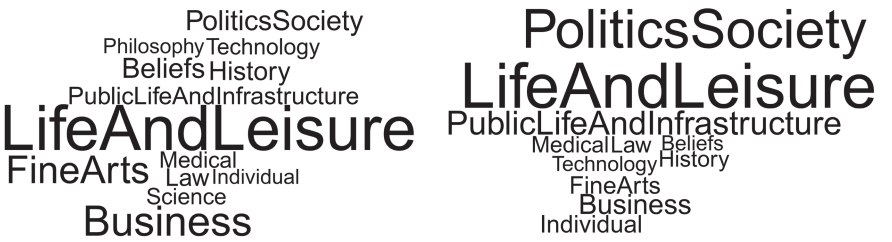

Abb. 1: Verteilung der COWCat-Topikdomänen in den Stichproben aus DECOW14A (links) und DeReKo-2014-II (rechts)

5 Eine weiterentwickelte Version findet sich unter http://corporafromtheweb.org/cowcat/.

4 Topikmodellierung

Der Begriff *Topikmodellierung* bezeichnet eine Reihe von Verfahren zur automatischen thematischen Erschließung größerer Textmengen, die in Teilen der digitalen Geisteswissenschaften seit Jahren etabliert sind (Hall/Jurafsky/Manning 2008; Jockers 2014; Jockers/Mimno 2012; Nelson 2016; Rhody 2012). Die meisten Verfahren dieser Art stützen sich ausschließlich auf die Vorkommenshäufigkeit von lexikalischen Wörtern in Texten und quantifizieren mit verschiedenen mathematischen Verfahren die thematischen Ähnlichkeiten der Texte untereinander. Sehr bekannt sind *Latent Semantic Indexing* (LSI; Landauer/Dumais 1997) und *Latent Dirichlet Allocation* (LDA; Blei/Ng/Jordan 2003). Beide decken ohne vorgegebene Kategorien semantische Strukturen in Textsammlungen auf. Lediglich die gewünschte Anzahl der Topiks – also letztlich die Feinheit der Klassifikation – wird vorgegeben. Weder LSI noch LDA liefern aber *Bezeichnungen* für die induzierten Topiks. Die Ausgabe beschränkt sich vielmehr auf eine gewichtete Liste besonders charakteristischer Wörter für ein Topik sowie eine gewichtete Zuordnung einzelner Texte zu den Topiks. Diese namenlosen Topiks versuchen wir im nächsten Schritt vorgebenen Topikdomänen zuzuordnen.

Wir verwenden die in der Software *gensim* (Řehůřek/Sojka 2010) implementierte Variante von LSI. Die zu induzierende Anzahl von Topiks wurde experimentell von 20 bis 90 variiert. Als Eingabe wurden die Häufigkeiten von Substantiven, Adjektiven, Verben und Adverbien in den einzelnen Dokumenten verwendet. Weil unsere Stichproben für Topikmodellierung eher klein sind, wurden zudem schrittweise weitere Dokumente aus den beiden Korpora verwendet, jedoch nicht beim anschließenden überwachten Lernen (da für diese zusätzlichen Texte eben keine manuelle Annotation der Topikdomäne vorlag). Solche und ähnliche Methoden werden in der Topikmodellierung öfter zur Stabilisierung des Verfahrens eingesetzt.

Tab. 1: Charakteristische Wörter für einige ausgewählte Topiks

Topik1	Topik2	...	Topik29	Topik30
spiel	*hotel*		*diabetes*	*album*
mannschaft	*ferienhaus*		*kirche*	*kind*
sieg	*unternehmen*		*stellungnahme*	*band*
punkt	*markt*		*turnier*	*polizei*
team	*deutsch*		*patient*	*song*
minute	*deutschland*		*album*	*prozent*

Topik1	Topik2	...	Topik29	Topik30
platz	*kunde*		*euro*	*konzert*
trainer	*fahren*		*platz*	*music*
spielen	*service*		*haut*	*lied*
gewinnen	*bieten*		*schule*	*diabetes*

Das Ergebnis ist für jedes Topik eine Liste charakteristischer Wörter. Tabelle 1 illustriert für einige Topiks die zehn typischsten Wörter, in absteigender Gewichtung. Während *Topik1* und *Topik2* eindeutig mit Fußball und Tourismus in Zusammenhang stehen, ist die Interpretation von *Topik30* weniger eindeutig. Offenbar geht es um Musik und Konzerte, doch passen nicht alle Wörter dazu. *Topik29* ist noch schwieriger zu interpretieren. Hier scheinen Schlüsselbegriffe aus unterschiedlichen Themen wie Gesundheit, Glaube und Sport vermischt zu sein. Dies illustriert, dass induzierte Topiks ohne weitere Verarbeitung eher nicht der linguistischen Vorstellung von relevanten Metadaten entsprechen. Darüber hinaus erzeugt das Verfahren eine Dokument-Topik-Matrix, in der jedes der induzierten Topiks für jedes Dokument gewichtet ist. Man erhält damit ein Maß für die Zugehörigkeit eines Dokuments zu den einzelnen Topiks. Tabelle 2 illustriert eine solche Matrix.

Tab. 2: Beispiel einer Dokument-Topik-Matrix

	Topik1	Topik2	...	Topik29	Topik30
Dokument_1	.067	.045		-.002	-.040
Dokument_2	.149	.123		-.007	.008
Dokument_3	.093	.171		.026	.083
...					
Dokument_1756	.219	-.062		.157	.066

Abbildung 2 zeigt einen Vergleich der beiden verwendeten Teilkorpora hinsichtlich der Verteilung ausgewählter Topiks. Dargestellt ist das Verhältnis der Anteile der Dokumente in jedem der Korpora, für die das jeweilige Topik unter den drei am höchsten gewichteten Topiks ist.

Abb. 2: (Logarithmiertes) Verhältnis relativer Häufigkeiten ausgewählter LSI-Topiks in den beiden Korpora. Je weiter ein Topik von x=0 entfernt ist, umso charakteristischer ist es im direkten Vergleich für DeReKo (links) oder DECOW (rechts)

5 Automatische Klassifikation

Im nächsten Schritt sollen in einem überwachten Verfahren die im Goldstandard annotierten Topikdomänen aus den gewichteten Zuordnungen von Dokumenten zu Topiks gelernt werden. Als Eingabe dient dabei also eine Dokument-Topik-Matrix wie in Tabelle 2. Zur Klassifizierung verwenden wir eine Support Vector Machine, weil wir damit im Vergleich zu anderen Klassifizierern die besten Ergebnisse erzielen konnten. Da in den Trainingsdaten einige Topikdomänen nur sehr selten vorkommen und in diesen Fällen keine verlässlichen Generalisierungen durch den Klassifizierer zu erwarten sind, wurde auf zwei Versionen des Daten-

satzes trainiert: dem vollen Datensatz und einem reduzierten Datensatz, bei dem schwach repräsentierte Kategorien ausgefiltert wurden. Wie oben erwähnt, wurde bei den Topikmodellen, die als Eingabe für den Klassifizierer dienen, sowohl die Anzahl der induzierten Topiks als auch die Anzahl der zusätzlich beigemischten Dokumente variiert. Tabelle 3 zeigt die jeweils besten Kombinationen dieser Parameter.

Tab. 3: Beste erreichbare Genauigkeit für die reduzierten Datensätze bei 10-facher Kreuzvalidierung. *Precision*, *Recall* und *F1-Score* sind gewichtete Mittelwerte über alle Kategorien

Korpus	zusätzl. Dok.	Topiks	Genauigkeit	Prec.	Rec.	F1
DECOW	3200	20	68,77%	0,69	0,69	0,67
DeReKo	3600	40	73,00%	0,73	0,73	0,70
DECOW + DeReKo	0	30	51,87%	0,43	0,52	0,42

Aus den Ergebnissen ist ersichtlich, dass Topikmodelle, die aus den DECOW- und DeReKo-Daten jeweils für sich genommen erzeugt werden, einen besseren Input für den Klassifizierer liefern als Topikmodelle, die aus beiden Korpora gemeinsam erzeugt werden. Dies ist bemerkenswert, da eine größere Menge an Trainingsdaten typischerweise mit einer höheren Genauigkeit bei der Klassifikation einhergeht. Wir gehen davon aus, dass dies durch die starke Ungleichverteilung der Kategorien bedingt ist, die entsteht, wenn die Daten aus beiden Korpora kombiniert werden: Mit *Life and Leisure* sowie *Politics and Society* gibt es dann in den Trainingsdaten zwei sehr stark ausgeprägte Kategorien, die den Klassifizierer dazu verleiten, einen Großteil der Dokumente diesen beiden Klassen zuzuordnen. Ein Blick auf die Konfusionsmatrizen[6] (Tabellen 4–6) legt außerdem die Vermutung nahe, dass die Kategorie *Life and Leisure* möglicherweise zu weit gefasst ist, da bei allen drei Datensätzen eine erhebliche Anzahl Dokumente fälschlich als *Life and Leisure* klassifiziert wird. Gleiches gilt für *Politics and Society*. Es muss allerdings beachtet werden, dass die zum Training verwendeten Goldstandard-

6 Die Konfusionsmatrizen vergleichen für jedes Dokument die manuelle Annotation mit dem Ergebnis der automatischen Klassifikation. Bei einem perfekten Ergebnis wären alle Werte jenseits der Diagonalen null. Die unterschiedliche Anzahl der Kategorien bei den einzelnen Korpora ergibt sich aus den jeweils ausgeschlossenen, schwach repräsentierten Kategorien.

korpora relativ klein sind und damit – gerade bei einer sehr unausgewogenen Verteilung über die Topikdomänen – schlicht nicht genug Daten vorhanden sind, um schwach repräsentierte Kategorien zu lernen.

Tab. 4: Konfusionsmatrix für die DECOW-Daten

	DECOW	**klassifiziert**							
		PolSoc	**Business**	**Life**	**Arts**	**Public**	**Law**	**Beliefs**	**History**
annotiert	PolSoc	**26**	12	10	1	1	0	1	0
	Business	5	**105**	40	7	1	2	1	1
	Life	3	14	**286**	6	4	1	1	1
	Arts	3	2	36	**78**	1	0	2	6
	Public	0	3	11	0	**9**	1	0	0
	Law	3	9	8	0	1	**8**	0	0
	Beliefs	4	3	11	6	1	0	**30**	1
	History	9	0	9	7	1	1	2	**15**

Tab. 5: Konfusionsmatrix für die DeReKo-Daten

	DeReKo	**klassifiziert**					
		PolSoc	**Business**	**Life**	**Indiv**	**Arts**	**Public**
annotiert	PolSoc	**223**	6	39	0	0	8
	Business	20	**24**	9	0	0	0
	Life	24	1	**324**	0	0	1
	Indiv	5	0	17	**0**	0	1
	Arts	2	0	28	0	**6**	0
	Public	35	0	30	0	0	**34**

Tab. 6: Konfusionsmatrix für kombinierte DECOW- und DeReKo-Daten

DECOW + DeReKo		klassifiziert								
		PolSoc	Business	Medical	Life	Arts	Public	Law	Beliefs	History
annotiert	PolSoc	**199**	7	0	109	0	12	0	0	0
	Business	18	**23**	0	172	0	2	0	0	0
	Medical	6	0	**0**	29	0	1	0	0	0
	Life	25	4	0	**632**	0	5	0	0	0
	Arts	2	2	0	160	**0**	0	0	0	0
	Public	46	2	0	56	0	**19**	0	0	0
	Law	8	0	0	31	0	0	**0**	0	0
	Beliefs	0	0	0	0	59	0	0	**0**	0
	History	4	0	0	50	0	0	0	0	**0**

6 Zusammenfassung und Ausblick

Die Ergebnisse zeigen deutlich, dass zwischen datengetrieben aufgedeckten Topiks und extern definierten Topikdomänen eine Verbindung besteht. Bei der Verteilung solcher Topiks und Topikdomänen bestehen ausgeprägte Unterschiede zwischen Zeitungs- und Webkorpora. Innerhalb beider Korpora sind einige Topikdomänen stark unterrepräsentiert, während andere die Verteilung klar dominieren. Weitere Experimente werden zeigen, ob sich das Klassifikationsergebnis durch größere Trainingskorpora verbessern lässt.

Die relativ häufig auftretenden Fehlklassifikationen in den Kategorien *Life and Leisure* sowie *Politics and Society* legen nahe, dass diese zu weit gefasst und somit nicht distinktiv genug sind. Auf Grundlage dieser Erkenntnisse kann das verwendete Annotationschema für Topikdomänen (COWCat) dahingehend angepasst werden, dass die postulierten Kategorien eine bessere empirische Fundierung (nämlich in lexikalischen Verteilungen) haben. Rückmeldungen der beteiligten Annotatorinnen, die auf Probleme mit diesen beiden Kategorien hingewiesen und eine Teilung vorgeschlagen haben, konvergieren mit den Ergebnissen des Klassifikationsexperiments. Wir sehen daher das Ergebnis unseres Experiments als wichtigen Schritt in Richtung eines empirisch fundierten Annotationsschemas für die automatische Auszeichnung großer Korpora mit Metadaten.

Literatur

Biber, Douglas/Egbert, Jesse (2016): Using grammatical features for automatic register identification in an unrestricted corpus of documents from the open web. In: Journal of Research Design and Statistics in Linguistics and Communication Science 2, S. 3–36.

Blei, David M./Ng, Andrew Y./Jordan, Michael I. (2003): Latent dirichlet allocation. In: Journal of Machine Learning Research 3, S. 993–1022.

Gries, Stefan Th. (2015): The most underused statistical method in corpus linguistics: multi-level (and mixed-effects) models. In: Corpora 10, S. 95–126.

Hall, David/Jurafsky, Daniel/Manning, Christopher D. (2008): Studying the history of ideas using topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08). Stroudsburg, S. 363–371.

Jockers, Matthew L. (2014): Text analysis with R for students of literature. (= Quantitative Methods in the Humanities and Social Sciences). Cham u.a.

Jockers, Matthew L./Mimno, David (2012): Significant themes in 19th-century literature. (= DigitalCommons@University of Nebraska – Lincoln). Lincoln, NE. Internet: http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1105&context=englishfacpubs (Stand: 20.9.2016).

Kupietz, Marc et al. (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta et al. (Hg.): Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC '10). Valletta, S. 1848–1854.

Landauer, Thomas K./Dumais, Susan T. (1997): A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. In: Psychological Review 104, S. 211–240.

Leech, Geoffrey (2007): New resources or just better old ones? The Holy Grail of representativeness. In: Hundt, Marianne et al. (Hg.): Corpus linguistics and the web. Amsterdam/New York, S. 133–149.

Mehler, Alexander/Sharoff, Serge/Santini, Marina (Hg.) (2010): Genres on the web: Computational models and empirical studies. (= Text, Speech and Language Technology 42). New York.

Nelson, Robert K. (2016): Mining the dispatch. Internet: http://dsl.richmond.edu/dispatch (Stand: 20.9.2016).

Řehůřek, Radim/Sojka, Petr (2010): Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, S. 45–50.

Rhody, Lisa M. (2012): Topic modeling and figurative language. In: Journal of Digital Humanities 2, 1.

Schäfer, Roland (2015): Processing and querying large web corpora with the COW14 architecture. In: Bański, Piotr et al. (Hg.): Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3). Mannheim, S. 28–34.

Schäfer, Roland/Bildhauer, Felix (2012): Building large corpora from the web using a new efficient tool chain. In: Calzolari, Nicoletta et al. (Hg.): Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC '12). Istanbul, S. 486–493.

Sharoff, Serge (2006): Creating general-purpose corpora using automated search engine queries. In: Baroni, Marco/Bernardini, Silvia (Hg.): WaCky! Working papers on the web as corpus. Bologna, S. 63–98.

Sinclair, John McH./Ball, J. (1996): Preliminary recommendations on text typology. Technical report EAG-TCWG-TTYP/P. Internet: www.ilc.cnr.it/EAGLES/texttyp/texttyp.html (Stand: 20.9.2016).

Weiß, Christian (2005): Die thematische Erschließung von Sprachkorpora. (= OPAL – Online publizierte Arbeiten zur Linguistik 1/2005). Mannheim.