Induktive Topikmodellierung und extrinsische Topikdomänen

# Motivation

Mit der Verfügbarkeit sehr großer Korpora (im Bereich von Milliarden Wörtern) haben Korpusdaten für viele Bereiche der Linguistik an Bedeutung gewonnen. Einer der Gründe dafür ist, dass große Korpora die Möglichkeit eröffnen, auch seltene Phänomene zu untersuchen, ohne dafür Daten experimentell zu erheben oder auf bloße Introspektion zurückzugreifen. Bei der korpuslinguistischen Untersuchung eines sprachlichen Phänomens werden üblicherweise Informationen *über* die Texte, aus denen die Belege stammen, miteinbezogen: Sei es, weil das Hauptinteresse bei 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 soziodemographischen Angaben zum/zur Verfasser/in des Texts auch Information über die Kommunikationssituation, die beabsichtigte Wirkung des Texts, das Textthema usw.

Das Nichtvorhandensein solcher und anderer Metadaten bei sehr großen, gecrawlten Webkorpora ist ein gelegentlich geäußerter Kritikpunkt (z.B. Leech 2007), doch sollte man dabei nicht außer Acht lassen, dass auch für „traditionelle“ Korpora oft nicht in gewünschtem Maße Metadaten vorliegen. Dies gilt insbesondere für so abstrakte Kategorien wie „Register“, „Genre“, „Textsorte“ etc.[[1]](#footnote-1) Obwohl die Varianz sprachlicher Phänomene in Abhängigkeit von solchen Kategorien seit Jahrzenten Gegenstand der korpuslinguistischen Diskussion ist, gibt es bis heute keine allgemein akzeptierte Definitionen für diese Begriffe (meistens macht man Gebrauch von einer oder mehreren anderer Metadadaten-Kategorien wie Kommunkationssituation oder Intention; Lee 2001 gibt einen Überblick). Der fehlende Konsens macht sich erwartbarerweise auch bei der Erstellung von Taxonomien bemerkbar. Schon für „klassische“ Medien konnte kein halbwegs einvernehmliches Inventar von Genres angegeben werden, und das Aufkommen neuer Textformen im WWW macht die Situation nur komplexer (vgl. die Beiträge in Mehler et al. (ed.) 2007). Vorgeschlagene Kategorien sind meist prototypisch, was eine Klassifikation von Texten enorm erschwert. Zudem lassen sich Ober- und Unterkategorien schwer gegeneinander abgrenzen. Dies führt dazu, dass bei einer manuellen Klassifikation von Texten die Übereinstimmung zwischen voneinander unabhängigen Annotatoren oft unbefriedigend ist, und die automatische Klassifikation, zumal für Webdaten, auch in sehr rezenten Experimenten äußerst unbefriedigende Ergebnisse liefert. So berichten Egbert et al. (2016) von 42,1% korrekt klassifizierter Texte über 32 Kategorien. Darüber hinaus verwenden Methoden zur automatischen Klassifikation von Genres meist sprachliche Merkmale (oft grammatischer Natur) als Grundlage. Während dies für manche praktischen Anwendungen kein Problem darstellt (z.B. für ein Ranking von Suchergebnissen, das neben dem Inhalt eines Dokuments auch dessen Genre mit einbezieht), ist es konzeptuell höchst fragwürdig, Korpora für die linguistische Forschung mit Metadaten auszuzeichnen, die aus dem (gemeinsamen) Auftreten grammatischer Phänomene gewonnen wurden. Sobald nämlich eines der zur Klassifikation verwendeten Phänomene (oder ein anderes, mit ihm korrelierendes) anhand dieses Korpus linguistisch untersucht wird, droht Zirkularität: Phänomen X tritt signifikant häufiger in Genre Y auf; aber Dokumente wurden ursprünglich als Genre Y klassifiziert, weil (unter anderem) Phänomen X dort häufig auftritt.

Eine andere, weitaus geläufigere Klassifikation von Dokumenten ist die thematische. Sie ist im Prinzip orthogonal zur Klassifikation nach Genres, es ist jedoch plausibel, von deutlichen Abhängigkeiten zwischen Textthema und Genre auszugehen. Die thematische Zusammensetzung eines Korpus ist sicher ein sinnvolles Kriterium für den Vergleich von Korpora (wenn auch keineswegs das einzige; die Vergleichbarkeit von Korpora ist wichtiges und häufig diskutiertes Thema in der Korpuslinguistik, siehe u.a. Kilgarriff 2001; Baroni & who????; Schäfer & Bildhauer 2013). Auf der anderen Seite kann Information über das Textthema[[2]](#footnote-2) auch für Untersuchungen sprachlicher Phänomene eine Rolle spielen. Gleichzeitig ist bei einer Klassifikation auf der Basis von Inhaltswörtern das Problem der Zirkularität wahrscheinlich weniger gravierend, zumindest dann, wenn anhand des thematisch klassifizierten Korpus vorwiegend grammatische Phänomene untersucht werden.

Die wichtigste Frage, die sich bei der thematischen Erschließung von Korpora stellt, ist die nach der verwendeten Taxonomie. Auch hier gibt es keinen Konsens, sondern verschiedene, parallel existierende, nicht miteinander kompatible Klassifikationssysteme, wie Sinclair und Ball (1996) anmerken. Eine weitere Schwierigkeit ist, dass eine eindeutige Zuordnung von Texten zu Themen oft nicht einmal innerhalb eines Systems vorgenommen werden kann, sofern nur ein einziges Etikett vergeben werden darf. Der Grund dafür ist, dass Texte häufig verschiedene Themen einer gegebenen Taxonomie (mit einer überschaubaren Anzahl von Kategorien) kombinieren, etwa bei Themen wie „Wirtschaftskriminalität“, „Umweltpolitik“, „Sponsoring in der Bundesliga“ etc.; erschwerend hinzu kommen auch Themenwechsel innerhalb eines Textes.

Um der Beliebigkeit bei der Erstellung einer Thementaxonomie entgegenzuwirken, bietet sich eine Kombination von „externen“ und „internen“ Klassifikationskriterien an (Sinclair & Ball 1996). Als internes Kriterium kommt eine induktive Klassenbildung aufgrund des auftretenden lexikalischen Materials in Frage. Ein solches datengetriebenes Aufdecken von Topiks ist möglich und objektiver, die resultierenden Kategorien sind jedoch nicht immer intuitiv interpretierbar (vgl. Abschnitt „

Topikmodellierung“). Nimmt man an, dass für die linguistische Forschung (oder zumindest für die meisten Linguisten) die intuitive Interpretierbarkeit von thematischen Kategorien wichtig ist, muss hier ein Kompromiss gefunden werden, und eine Möglichkeit besteht darin, eine „externe“ Taxonomie so auszurichten, dass die Kategorien möglichst gut mit lexikalischem Material korrespondieren („in den Daten wurzeln“) und sie sich damit auch möglichst gut für eine automatische Klassifikation eignen.

Dieses Papier beschäftigt sich mit der Frage, in welchem Maße extern definierte thematische „Groß“-Bereiche („Topikdomänen“) mithilfe von unüberwacht generierten thematischen Charaktersierungen von Dokumenten automatisch erschlossen werden können. Ziel ist es einerseits, eine extern motivierte, für Linguisten und Linguistinnen intuitive und attraktive Taxonomie zu bekommen, die aber andererseits objektiv in lexikalischen Verteilungen verankert ist und einen geeigneten Input für eine aussichtsreiche Klassifikation darstellt. Dieser Ansatz ist als solcher nicht neu. Selbst für einen Teil von uns verwendeten Daten wurde ein ähnlicher Ansatz bereits verfolgt (Weiß 2005). Von letzterem unterscheidet sich unser Experiment in mehreren Punkten. So verwenden wir als Clusteringverfahren Methoden, die dezidiert für die Verarbeitung von Sprache entwickelt wurden. Wir kombinieren darüber hinaus Korpora sehr unterschiedlicher Art und führen schließlich eine Evaluation durch, die dem gängigen Standard in der Computerlinguistik entspricht. Es soll im Folgenden ausdrücklich nicht darum gehen, neue Algorithmen für die Textklassifikation vorzustellen. Vielmehr sollen etablierte Verfahren in sinnvoller Weise kombiniert und für die Konstruktion von Korpora nutzbar gemacht werden, die in der linguistischen Forschung verwendbar sind.

# Übersicht

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

# Daten & Goldstandard

Die Daten für unsere Untersuchung wurden als Stichprobe aus zwei verschiedenen Korpora gezogen: 870 Dokumente aus DECOW14, einem Korpus aus gecrawlten HTML-Dokumenten aus dem World Wide Web (ca. 17 Mio. Wörter; Schäfer und Bildhauer 2012) sowie 886 Dokumente aus DeReKo 2014-II, das überwiegend Zeitungstexte enthält (ca. 28 Mio. Wörter; Kupietz et al. 2010). Die Auswahl der Korpora ist durch die Vermutung motiviert, dass neben einigen Gemeinsamkeiten auch deutliche Unterschiede in der Verteilung von Topiks auftreten werden. Eine wichtige Frage ist dann, ob der so entstandene, heterogene Datensatz für das Auffinden von Topiks und das Trainieren eines Klassifizierers geeignet ist, oder Daten aus derart verschiedenen Korpora besser getrennt verarbeitet werden sollten. Für das spätere topic modelling und die automatische Klassifizierung wurden alle Dokumente identisch aufbereitet (dies beeinhaltet Tokenisierung, Wortarten-Tagging, Lemmatisierung, Eigennamenerkennung).

Die Dokumente wurden manuell nach dem CowCat-Klassifikationsschema[[3]](#footnote-3) annotiert, das wiederum auf Arbeiten von Sharoff (2006) aufbaut. Zielgröße beim Aufbau des Schemas war eine moderate Anzahl (ca.10–20) von Topik*domänen*, d.h. thematische Großbereiche, von denen jeder eine Anzahl einzelner Topiks subsumiert. Die Taxonmie wurde in mehreren Annotationsdurchläufen (auf anderen Datensätzen als den in diesem Experiment verwendeten) entwickelt und angepasst, dabei wurden jeweils die Rückmeldungen der Annotatoren/Annotatorinnen berücksichtigt. Die von uns verwendete Version umfasst die in Tabelle 1 dargestellten 13 Kategorien:

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

Tabelle : Kategorien für Topikdomänen aus CowCat 2013

Abbildung 1 zeigt die manuell ermittelte Verteilung dieser Kategorien in den Stichproben aus DECOW14 und DeReKo-2014-II. (Die Schriftgröße spiegelt den Anteil der jeweiligen Kategorie wider.) In beiden Korpora dominieren Texte über „Freizeitthemen“ (Kategorie *Life and Leisure*). Bei DECOW14 (links) sind darüber hinaus Texte über Kunst oder Wirtschaft/Unternehmen auffällig, während in DeReKo-Texten die Themen aus den Bereichen Politik, Gesellschaft und öffentliche Einrichtungen/Infrastruktur vorherrschen.



[Abbildung 1 hier (decow-dereko-cattle-dist.pdf)]

Abbildung : Verteilung der CowCat-Topikdomänen in den Stichproben aus DECOW2014 (links) und DeReKo-2014-II

# Topikmodellierung

Topikmodellierung bezeichnet eine Reihe von Verfahren zur automatischen thematischen Erschließung größerer Textmengen, die in Teilen der digitalen Geisteswissenschaften seit Jahren etabliert sind und deren Nutzen als heuristisches Werkzeug in zahlreichen Fallstudien dokumentiert wurde (paar Beispiele). Die am meisten verbreiteten Verfahren dieser Art stützen sich ausschließlich auf (Ko-) Okkurrenzen von Wörtern in Texten und erfassen durch verschiedene numerische Techniken thematische Ähnlichkeiten zwischen den Texten. Ein sehr bekanntes Verfahren ist dabei *Latent Semantic Indexing* (LSI, auch bekannt als *Latent Semantic Analysis*; Landauer und Dumais 1997). Ein anderes, etwas rezenteres Verfahren ist *Latent Dirichlet Allocation* (LDA, Blei et al. 2003). Bei beiden handelt es sich um unüberwachte , bottom-up Verfahren, bei denen nicht nach vorgegebenen Kategorien klassifiziert wird, sondern die heuristisch semantische Strukturen in Textsammlungen aufdecken. Im Zusammenhang dieses Papiers ist es wichtig zu betonen, dass weder LSI noch LDA *Bezeichnungen* für die induzierten Topiks liefern. Die Ausgabe beschränkt sich vielmehr auf eine Liste besonders charakteristischer Wörter für ein Topik, sowie eine gewichtete Zuordnung einzelner Texte zu den einzelnen Topiks. Inwieweit auf diese Weise induzierte Topiks-ohne-Namen automatisch zu vorgebenen Topikdomänen zusammengefasst werden können, ist Gegenstand dieser Studie.

Wir verwenden die im gensym-Toolkit (Řehůřek und Sojka 2010) implementierte Variante vom *Latent Semantic Indexing*. Die zu induzierende Anzahl von Topiks wurde zwischen 20 und 90 variiert; für COW- und DeReKo-Daten zusammen genommen ergab sich dabei N=30 als optimale Anzahl an Topiks (mit Blick auf das nachfolgende Klassifikationsexperiment). Als Eingabe wurden TF-IDF-gewichtete Häufigkeiten von Substantiven, Adjektiven, Verben und Adverbien berücksichtigt, in Form von Tupeln bestehend aus dem Lemma und der Wortart (haus\_nn, franzkafka\_ne etc.). Weil die von uns verwendeten Stichproben für Zwecke der Topikmodellierung als klein gelten müssen, wurden zudem schrittweise weitere Dokumente aus den beiden Korpora für die Topikmodellierung hinzugefügt, jedoch nicht beim anschließenden überwachten Lernen verwendet (da für diese zusätzlichen Texte eben keine manuelle Annotation der Topikdomäne vorlag).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Topik1 | Topik2 | ... | Topik29 | Topik30 |
| *spiel*  *mannschaft*  *sieg*  *punkt*  *team*  *minute*  *platz*  *trainer*  *spielen*  *gewinnen*  *ball*  *spieler*  *turnier*  *saison* | *hotel*  *ferienhaus*  *unternehmen*  *markt*  *deutsch*  *deutschland*  *kunde*  *fahren*  *service*  *bieten*  *zimmer*  *preis*  *haus*  *international* |  | *diabetes*  *kirche*  *stellungnahme*  *turnier*  *patient*  *album*  *euro*  *platz*  *haut*  *schule*  *behandlung*  *halbfinale*  *gott*  *gewinnen* | *album*  *kind*  *band*  *polizei\_*  *song*  *prozent*  *konzert*  *music*  *lied*  *diabetes*  *jugendliche*  *telefon*  *auto*  *beamter* |

Tabelle : Charakteristische Wörter für einige ausgewählte Topiks

Im Ergebnis kann für jedes der 30 Topiks eine Liste von Wörtern angegeben werden, die für das jeweilige Topik charakteristisch sind. Tabelle 2 illustriert für einige Topiks die 14 typischsten Wörter, in absteigender Gewichtung. Während *Topik1* und *Topik2* wohl unkontrovers mit „Fußball” und „Tourismus” betitelt werden können, ist die Interpretation von *Topik30* weniger eindeutig, offenbar geht es um Musik und Konzerte, doch passen nicht alle dargestellten Wörter zu diesem Etikett. *Topik29* ist noch schwieriger zu interpretieren, hier scheinen Schlüsselbegriffe aus unterschiedlichen Themen (etwa „Gesundheit“, „Glauben“, „Sport“) vermischt zu sein. Dies illustriert den Punkt, dass induzierte Topiks ohne weitere Verarbeitung eher nicht der Vorstellung vom dem entsprechen, was Linguisten normalerweise unter relevanten Metadaten verstehen.

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 3 zeigt einen Auschnitt aus einer solchen Matrix.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Topik1 | Topik2 | Topik3 | | ... | | Topik30 |
| Dokument\_1 | .067 | .045 | ­-.002 |  | | -.040 | |
| Dokument\_2 | .148 | .123 | -.007 |  | | .008 | |
| Dokument\_3 | .093 | .171 | .026 |  | | .083 | |
| ... |  |  |  |  | |  | |
| Dokument\_1756 | .219 | -.062 | .157 |  | | .066 | |

Tabelle : Beispiel einer Dokument-Topik-Matrix

Abbildung 2 zeigt einen Vergleich der beiden verwendeten Teilkorpora hinsichtlich der Verteilung ausgewählter LSI-Topiks (nur solche, die auf Basis ihrer Schlüsselbegriffe einigermaßen interpretierbar sind). 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, gemäß der Dokument-Topik-Matrix. Beispiel: 30% der DeReKo-Dokumente haben *Topik-X* unter ihren drei am höchsten gewichteten Topiks, aber nur 15% der DECOW-Dokumente. Das Verhältnis ist dann 0.15/0.3 = 0.5, und von diesem Wert wird der Logarithmus genommen (≈-0.69).



[Abbildung 2 hier (topics-logratios2.pdf)]

Abbildung : Verhältnis relativer Häufigkeiten (log) ausgewählter LSI-Topiks zwischen den beiden Korpora (= Anteil der der Dokumente, für die das jeweilige Topik unter den drei am gewichteten ist). Je weiter ein Topik von der y-Achse entfernt ist, umso charakteristischer ist es im direkten Vergleich für DeReKo (links) oder DECOW (rechts).

# Automatische Klassifikation

Im nächsten Schritt sollen in einem überwachten Verfahren die im Goldstandard annotierten Topikdomänen aus den gewichteten Zuordnung von Dokumenten zu Topiks gelernt werden. Als Eingabe dient dabei also eine Dokument-Topik-Matrix wie in Tabelle 3. Als Klassifizierer verwenden wir eine SVM mit einem Pearson VII Kernel (Üstün et al.2006), weil damit im Vergleich zu praktisch allen anderen Klassifizierern im Weka-toolkit (Hall & Witten, 2011) die besten Ergebnisse erzielt werden konnten. Da in den Trainingsdaten einige Topikdomänen nur sehr selten vorkommen und in diesen Fällen keine verlässlichen Generalisierungen des Klassifizierer zu erwarten sind, wurde auf zwei Versionen des Datensatzes trainiert: dem vollen Datensatz, und einem reduzierten Datensatz, bei dem schwach repräsentierte Kategorien ausgefiltert wurden. Die Abbildung 3 bis Abbildung 5 zeigen die Klassifikationsergebnisse (Genauigkeit, d.h. Anteil korrekt klassifizierter Dokumente) in Abhängigkeit von der Anzahl der induzierten LSI-Topiks, mit denen der Klassifizierer trainiert wurde. Dargestellt sind verschiedene Bedingungen: unterschiedliche Vorverarbeitungsvarianten, voller vs. reduzierter Datensatz, unterschiedliche Anzahl zusätzlicher Dokumente bei der Topikmodellierung (s.o.), Evaluation auf Basis von Trainingsdaten vs. 10-fache Kreuzvalidierung. Die jeweils besten Kombinationen dieser Parameter für die einzelnen Korpora (DECOW, DeReKo und DECOW ∪ DeReKo) zeigt Tabelle 4.



[Abbildung 3 hier (idsyb\_decow.pdf)]

Abbildung : Klassifikationsergebnisse (Genauigkeit) der Grundlage von unterschiedlichen Topikmodellen für DECOW-Dokumente



[Abbildung 4 hier (idsyb\_dereko.pdf)]

Abbildung : Klassifikationsergebnisse (Genauigkeit) auf der Grundlage von unterschiedlichen Topikmodellen für DeReKo-Dokumente



[Abbildung 5 hier (idsyb\_coreko.pdf)]

Abbildung : Klassifikationsergebnisse (Genauigkeit) auf der Grundlage von unterschiedlichen Topikmodellen für kombinierte DECOW- und DeReKo-Dokumente

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Korpus | zusätzl Dokumente | Attribute | Topiks | Genauigkeit | Prec. | Rec. | F1 |
| DECOW | 3200 | Token | 20 | 68,77% | 0,69 | 0,69 | 0,67 |
| DeReKo | 3600 | Lemma + POS | 40 | 73,00% | 0,73 | 0,73 | 0,70 |
| DECOW ∪ DeReKo | 0 | Lemma + POS | 30 | 51,87% | 0,43 | 0,52 | 0,42 |

Tabelle : Beste erreichbare Genauigkeit für die reduzierten Datensätze (d.h. ohne schwach repräsentierte Kategorien) bei 10-facher Kreuzvalidierung. Precision, Recall und F1-Score sind gewichtete Mittelwerte über alle Kategorien.

Aus den Ergebnissen ist ersichtlich, dass Topikmodelle, die aus den DECOW- und DeReKo-Daten jeweils für sich genommen erzeugt wurden, 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 vermuten, 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 ausgeprägte modale Kategorien, die den Klassifizierer dazu verleiten, einen Großteil der Dokumente diesen beiden Klassen zuzuordnen. Ein Blick auf die Konfusionsmatrizen[[4]](#footnote-4) (Tabellen 5-7) 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*. Man darf jedoch nicht aus dem Blick verlieren, dass die zum Training verwendeten Goldstandardkorpora 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.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **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** |

Tabelle : Konfusionsmatrix für die DECOW-Daten (bestes Klassifikationsergebnis)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **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** |

Tabelle : Konfusionsmatrix für die 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** |

Tabelle : Konfusionsmatrix für kombinierte DECOW- und DeReKo-Daten

# 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, und dies könnte ein Grund für das schwache Klassifikationsergebnis bei kombinierten Web- und Zeitungskorpora sein. Weitere Experimente werden zeigen, ob größere Trainingskorpora hier Abhilfe schaffen können.

Die relativ häufig auftretenden Fehlklassifikationen in den Kategorien *Life and Leisure* sowie *Politics and Society* legen nahe, dass diese als Topikdomänen 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.

(Noch einen Satz zum Schluss: encouraging, auf dem Weg zu praktikablem, empirisch fundiertem Annotationschema und dies ist eine der Voraussetzung 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. Jordan (2003): Latent dirichlet allocation. In: Journal of Ma- chine Learning Research 3, S. 993–1022.

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

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, PA, USA, S. 363—371.

Hall, Mark/Witten, Ian H. (2011): Data mining: practical machine learning tools and techniques. 3. Auflage. Burlington.

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. DigitalCom- mons@University of Nebraska - Lincoln.

Kilgarriff, Adam (2001): Comparing corpora. In: International Journal of Corpus Linguistics 6(1), S. 97–133.

Kupietz, Marc/Belica, Cyril/Keibel, Holger/Witt, Andreas (2010): The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In: Nicoletta Calzolari, et al. (Hg.): Proceedings of the Seventh 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 rep-resentation of knowledge. In: Psychological Review 104(2). S. 211–240.

Lee, David Y. W. (2001): Genres, registers, text types, domains and styles: Clarifying the concepts and nevigating a path through the BNC jungle. In: Language Learning and Technology 5(3), S. 37–72.

Mehler, Alexander/Sharoff, Serge/Santini, Marina (Hg.) (2010): Genres on the web: computational models and empirical studies. Text, speech and language technology, Bd. 42. New York.

Nelson, Robert K. (2016): Mining the Dispatch. <http://dsl.richmond.edu/dispatch>.

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. 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 effcient tool chain. In: Calzolari, Nicoletta, et al. (Hg.): Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12). Istanbul, S. 486–493.

Schäfer, Roland/Bildhauer, Felix (2013): Web Corpus Construction. Synthesis Lectures on Human Language Technologies. San Francisco.

Üstün, Bülent/Melssen, Willem J./ Buydens, Lutgarde M.C. (2006): Facilitating the application of Support Vector Regression by using a universal PearsonVII function based kernel. In: Chemometrics and Intelligent Laboratory Systems 81, S. 29–40.

1. Im Folgenden verwenden wir „Genre“ stellvertretend für Kategorien, die in der Literatur oft auch als „Register“ oder „Textsorte“ o.ä. behandelt werden. [↑](#footnote-ref-1)
2. Wir verwenden „Textthema“ und „Topik“ gleichbedeutend, d.h. „Topik“ hat hier nichts mit dem Begriff „(Satz-)topik“ zu hat, der in der Literatur zur Informationsstruktur eine Rolle spielt. Gleiches gilt für „Topikdomäne“. [↑](#footnote-ref-2)
3. REF [↑](#footnote-ref-3)
4. Die unterschiedliche Anzahl der Kategorien ergibt sich aus den jeweils ausgeschlossenen schwach repräsentierten Kategorien. [↑](#footnote-ref-4)