Induktive Topikmodellierung und extrinsische Topikdomänen

# 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, auch 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 soziodemographischen Angaben zum Verfasser oder zur Verfasserin in 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 als Kritikpunkt gesehen (z.B. Leech 2007). Dabei sollte man allerding nicht außer Acht lassen, dass die gewünschten Metadaten auch für traditionelle Korpora oft nicht vorliegen. Dies gilt insbesondere für so abstrakte Kategorien wie *Register*, *Genre*, *Textsorte* usw.[[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 werden Metadadaten-Kategorien wie *Kommunkationssituation* oder *Intention* gebraucht (s. Lee 2001 für einen Überblick). 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 et al. (ed.) 2007). Die vorgeschlagenen Kategorien sind meist prototypisch aufzufassen, 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 und Annotatorinnen oft unbefriedigend ist. Daher verwundert es nicht, dass auch die automatische Klassifikation von Genres —zumal für Webdaten— auch in rezenten Experimenten nur unbefriedigende Ergebnisse liefert. Egbert & Biber (2016) berichten, dass ihr automatischer Klassifikator eine Genauigkeit von 42,1% auf 32 Kategorien aufweist.

Darüberhinaus verwenden Methoden zur automatischen Klassifikation von Genres meist sprachliche (oft grammtische) 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 aus dem (gemeinsamen) Auftreten grammatischer Phänomene gewonnen 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 Genre G 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 auszugehen. enkönnen linguistischerrelevant seingeringer Die thematische Zusammensetzung eines Korpus ist zudem ein gutes Kriterium, anhand dessen Korpora miteinander verglichen werden können. Der Vergleich von Korpora ist ein vieldiskutiertes Thema in der Korpuslinguistik; siehe u.a. Kilgarriff 2001; Baroni & who????; Schäfer & Bildhauer 2013.

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 und 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 und eine Klassenbildung auf dessen Basis. Ein solches datengetriebenes Aufdecken von sogenannten *Topiks* (relativ speziellen Einzelthemen) ist objektiv, aber die resultierenden Kategorien sind i.d.R. schlecht interpretierbar (vgl. Abschnitt *Topikmodellierung*). 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.

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 Linguisten und Linguistinnen 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 von uns verwendeten Daten wurde ein ähnlicher Ansatz in Weiß (2005) bereits verfolgt. Unsere Methode unterscheidet sich von Weiß (2005) in mehreren Punkten. So verwenden wir ein Clusteringverfahren,, das dezidiert für die Verarbeitung von Sprache entwickelt wurde. Wir kombinieren darüberhinaus 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 die Konstruktion von Korpora für die linguistische Forschung nutzbar gemacht werden.

# Übersicht

Unser Experiment gliedert sich in drei Schritte. Zunächst wird eine Stichprobe von 1756 Dokumenten manuell nach Topikdomänen (thematischen Großbereiche) 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 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 zwei verschiedenen Korpora entnommen: 870 Dokumente stammen aus DECOW2014A, einem Korpus aus gecrawlten HTML-Dokumenten aus dem World Wide Web (ca. 17 Mio. Wörter; Schäfer und 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 in der Verteilung von Topiks zwischen den Korpora zu erwarten sind. 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 aus beiden Korpora wurden auf identische Weise aufbereitet (Tokenisierung, Wortarten-Tagging, Lemmatisierung, Eigennamenerkennung).

Die Dokumente wurden manuell nach dem COWCat-Klassifikationsschema[[2]](#footnote-4) annotiert, das wiederum auf Arbeiten von Sharoff (2006) aufbaut. Zielgröße beim Aufbau des Schemas war eine moderate Anzahl von ca. 10–20 Topikdomänen. Die Taxonmie wurde in mehreren Annotationsdurchläufen entwickelt und angepasst. Die von uns hier verwendete Version umfasst die 13 Kategorien in Tabelle 1:

|  |  |
| --- | --- |
| *∙ 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 Verteilung dieser Kategorien in den manuell annotierten Stichproben aus DECOW2014A und DeReKo-2014-II. Die Schriftgröße spiegelt den Anteil der jeweiligen Kategorie wider. In beiden Korpora dominieren Texte über Freizeitthemen (*Life and Leisure*). Bei DECOW2014A (links) sind darüber hinaus Texte über Kunst oder Wirtschaft auffällig häufig, während in DeReKo-Texten die Themen aus den Bereichen Politik, Gesellschaft und öffentliche Einrichtungen bzw. Infrastruktur vorherrschen.



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

Abbildung : Verteilung der COWCat-Topikdomänen in den Stichproben aus DECOW2014A (links) und DeReKo-2014-II

# 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 und deren Nutzen in zahlreichen Studien dokumentiert wurde (paar Beispiele). Die am meisten verbreiteten Verfahren dieser Art stützen sich ausschließlich auf die Vorkommenshäufigkeit von lexikalischen Wörtern in Texten und quantifizieren mit verschiedenen mathematischen Verfahren die thematische Ähnlichkeiten der Texte untereinander. Sehr bekannt ist *Latent Semantic Indexing* (LSI; Landauer und Dumais 1997). Ein anderes Verfahren ist *Latent Dirichlet Allocation* (LDA; Blei et al. 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 (s. Abschnitt *Motivation*). 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 dann im nächsten Schritt vorgebenen Topikdomänen zuzuordnen.

Wir verwenden die in der Software *gensim* (Řehůřek und Sojka 2010) implementierte Variante vom *Latent Semantic Indexing*. 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 (in der Form von Tupeln aus dem Lemma und der Wortart, z.B. *haus\_nn*, *franzkafka\_ne*). Weil die von uns verwendeten Stichproben für Topikmodellierung eher klein sind, 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). Solche und ähnliche Methoden werden in der Topikmodellierung öfter zur Stabilisierung des Verfahrens eingesetzt.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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

Das Ergebnis ist für jedes der 30 Topiks eine Liste von Wörtern, die für das jeweilige Topik charakteristisch sind. Tabelle 2 illustriert für einige Topiks die 14 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 den Punkt, 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 3 zeigt einen Ausschnitt 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. 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. 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 (logarithmierter) relativer Häufigkeiten 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 0 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 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 durch den 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. 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, Evaluation auf Basis von Trainingsdaten bzw. 10-facher 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, 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 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[[3]](#footnote-5) (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*. Es muss allerdings beachtet werden, 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, Mark/ Witten, Ian H. (2011): Data mining: practical machine learning tools and techniques. 3. Auflage. Burlington.

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. pages 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.

Radim Řehůřek and Petr Sojka. 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

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. REF Eigentlich Schäfer & Bildhauer (2012) und http://corporafromtheweb.org/cowcat/ [↑](#footnote-ref-4)
3. Die unterschiedliche Anzahl der Kategorien ergibt sich aus den jeweils ausgeschlossenen schwach repräsentierten Kategorien. [↑](#footnote-ref-5)