TOBIAS HODEL, DENNIS MÖBUS, INA SERIF

Von Inferenzen und Differenzen. Ein Vergleich von *Topic-Modeling-Engines* auf Grundlage historischer Korpora

1. Erkenntnisinteresse

Seit Jahren drängt der Begriff „Big Data" zunehmend in die öffentliche Diskussion. Die klassisch hermeneutisch arbeitenden Geschichtswissenschaften haben sich davon lange unbeeindruckt gezeigt, doch auch Historiker:innen kommen nicht mehr um die Auseinandersetzung mit neuen Heuristiken herum, um die stetig wachsende digitale Quellenlage zu erfassen, zu durchmessen und zu erschließen.[1] Eine weit verbreitete, auch für nicht-Data-Scientists verständliche und auf (historische) Texte anwendbare Methode ist das *Topic Modeling*. Dabei handelt es sich um ein Verfahren des Maschinellen Lernens (ML), das anhand von aus großen Textkorpora extrahierten Wortgruppen (*Topics*) eine inhaltliche Erschließung dieser Texte ermöglicht. Der Vorteil einer solchen Informationsextraktion ist, dass gewaltige Textsammlungen mit Millionen von Wörtern in kurzer Zeit kursorisch erschlossen werden können und dabei nicht nach isolierten Begriffen, sondern nach thematischen Zusammenhängen gesucht wird. Die oftmals zeitraubende, ermüdende und gezwungenermaßen lückenhafte manuelle Praxis des Querlesens in der Phase der Quellenrecherche kann somit umgangen oder zumindest ergänzt werden.

Dieser Beitrag plädiert für eine offene und transparente Kommunikation der eingesetzten Algorithmen im Sinne einer Methodenkritik, die den digitalen Werkzeugkasten ernst nimmt und als Moment des Beitrags zum Erkenntnisgewinn versteht. Im Fokus dieser Untersuchung wird der Vergleich verschiedener *Topic-Modeling-Engines* stehen, im Bewusstsein, dass eine Vielzahl von weiteren *Pre-* und *Postprocessing*-Arbeitsschritten das Resultat von *Topic Modeling* ebenfalls beeinflusst. Hier soll jedoch insbesondere die Theorie und Methode des *Topic Modeling* ausgelegt werden, um die Konsequenzen der zentralen algorithmischen

[1] Siehe zur Einführung: Shawn Graham/Ian Milligan/Scott B. Weingart: *Exploring Big Historical Data. The Historian's Macroscope*. London 2015.

Anwendungen abzuschätzen. Basis für die theoretische und methodische Reflexion ist eine empirische Untersuchung der Ergebnisse in einem Wechselspiel von *Distant* und *Close Readings*, also der quantitativ unterstützten Lektüre, die u.a. nach sich wiederholenden Strukturen in den extrahierten Ergebnissen sucht, und einer quellenkritischen Überprüfung dieser Ergebnisse durch Rückverfolgung in die zugrundeliegenden Texte.

Die Algorithmen, die dem Vergleich unterzogen werden, sind die *Topic-Modeling-Engines* Gensim (Python) und Mallet (Java), die zugrundeliegenden Daten sind drei sehr unterschiedliche historische Textkorpora: spätmittelalterliche Sammelhandschriften, Zürcher Ratsbeschlüsse aus dem 19. Jahrhundert und lebensgeschichtliche Interviews aus den 1980er-Jahren. Durch den Rückgriff auf die unterschiedlichen Korpora soll einerseits aufgezeigt werden, welche Entscheide im Anwendungsprozess Einfluss auf die letztendlich jeweils spezifisch generierten Themenfelder haben. Andererseits erlaubt der Ansatz eine kritisch-komparative Sicht auf die Anwendung des Verfahrens für ganz unterschiedliche geschichtswissenschaftliche Fragestellungen.[2]

Um die Prozesse nicht nur deskriptiv nachvollziehbar zu machen, wird ein kommentiertes Notebook (ein Skript zur Anwendung auf beliebige Korpora) parallel zum Kapitel publiziert und, wo rechtlich möglich, auch die Datengrundlage zugänglich gemacht. Dadurch können die Prozessschritte auch auf informatischer Ebene nachvollzogen werden.[3]

2.1 Von Quellen zu Korpora und Daten

In der Geschichtswissenschaft wird typischerweise mit Quellen, weniger mit Korpora und selten (die Wirtschaftsgeschichte ausgeklammert) mit Daten operiert. Bereits der Einsatz der Begrifflichkeiten zeigt, dass die Umwandlung von Quellen zu Daten im digitalen Raum mit einer

---

2 Eine Einführung, die stärker auf den praxisnahen Einsatz für die Geschichtswissenschaft abzielt, findet sich in: Philip Grant et al.: Topic modelling on archive documents from the 1970s. Global policies on refugees. In: *Digital Scholarship in the Humanities 36* 4/2021, S. 886-904 [https://doi.org/10.1093/llc/fqab018].

3 Online: https://github.com/moebusd/von_inferenzen_und_differenzen (25.3.2022).

epistemologischen Umdeutung einhergeht. Quellen sind nicht unbedingt Daten und Daten bilden nicht zwangsläufig alle Eigenheiten von Quellen ab. Eine Quellensammlung wird also zu einem mehr oder minder kohärenten Textkorpus, das jedoch zur Auswertung – in unserem Fall zur Themenextraktion – weiter aufbereitet werden muss. Im folgenden Teil beschäftigen wir uns mit der theoretischen Funktionsweise von *Topic-Modeling*-Ansätzen sowie mit der Aufbereitung der Korpora, die für unsere Versuche verwendet wurden.[4]

2.2 Theorie: Die Funktionsweise von *Topic-Modeling*-Algorithmen

Topic Modeling ist ein probabilistisches Verfahren des maschinellen Lernens, mit dem in großen Dokumentenkorpora Themenfelder bestimmt und extrahiert werden können. Durch Berechnung von Wahrscheinlichkeitsverteilungen werden automatisch Hypothesen aufgestellt und iterativ verfeinert sowie verifiziert. Präziser ausgedrückt, werden Vorannahmen über Zusammenhänge zwischen Wörtern (a priori-Wahrscheinlichkeit) durch beständiges Wiederholen der Berechnung dieser Zusammenhänge optimiert (a posteriori-Wahrscheinlichkeit). Das Verfahren zielt somit darauf ab, herauszuarbeiten, welche Wörter und Wortgruppen häufig im Zusammenhang mit anderen Wörtern und Wortgruppen auftauchen.[5] Zentral ist in diesem Zusammenhang die Vorstellung eines Sacks gefüllt mit (einzelnen) Wörtern, dem bag of words. Damit werden explizit sowohl die Reihung der Wörter als auch linguistische oder semantische Abhängigkeiten ignoriert. Als wichtig angesehen wird einzig, dass die Wörter im selben Dokument vorkommen. Der ausführende Algorithmus nutzt also kein semantisches Wissen, sondern errechnet, unabhängig von deren Bedeutung,

---

4 Fragen der Zeichenkodierung (etwa die Nutzung von Unicode/UTF-8) oder wie analoge bzw. mündliche Quellen im digitalen Raum abgebildet werden, beschäftigen uns auf diesen Seiten nicht. Selbstredend sind auch diese Umwandlungsvorgänge zu dokumentieren und gemäß standardisierten Vorgängen vorzunehmen.

5 Einführend zu Topic Modeling siehe Megan R. Brett: Topic Modeling. A Basic Introduction. In: *Journal of Digital Humanities* 2 1/2012, online: http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/ (27.10.2021). Für die mathematische Perspektive vgl.: David M. Blei: Introduction to Probabilistic Topic Models. In: *Communications of the ACM* 55 4/2012, S. 77-84 [https://doi.org/10.1145/2133806.2133826].

welche Wörter über ein Korpus hinweg in Beziehung zueinander stehen und bündelt diese zu Gruppen – den Topics.[6] Im Umkehrschluss bedeutet dies, dass in solchen Topics Wörter zusammen vorkommen können, die semantisch gesehen nichts gemeinsam haben. Ausgehend von den Topics ist es in einem nächsten Schritt möglich, zu analysieren, welche Teile des Korpus – welche Dokumente – einen Anteil an bestimmten Themenfeldern haben. Obwohl Topic Modeling bereits seit Längerem in der Literaturwissenschaft und auch in der Geschichtswissenschaft angewendet wird, erfolgt nur in Einzelfällen eine Reflexion darüber, welche Vorgehensweisen zu den extrahierten Themenfeldern führen.[7]

Auf diesen Seiten beschränken wir uns auf den quasi-Standard der *Latent Dirichlet Allocation* (LDA). Obwohl auch andere *Topic-Modeling*-Ansätze existieren, werden diese aufgrund der vielfältigen Forschungen und der bereits existierenden Literatur zu LDA seltener verwendet.[8] Aktuell werden in den Digital Humanities und der Informatik vorrangig zwei Programme für *Topic Modeling* genutzt: Einerseits Gensim, eine Pythonbibliothek, andererseits Mallet, eine Java-basierte Umsetzung. Beide *Engines* haben in den Geisteswissenschaften ein gutes Standing. Aufgrund der steilen Lernkurve der Programmiersprache Python bietet Gensim einen guten Einstieg in das *Topic Modeling*. Allerdings existieren für Mallet sowohl hervorragende Tutorien (u.a. programminghistorian.org)[9] als auch Umsetzungen mit *graphical user interfaces*

---

6 Für dieses Kapitel unterscheiden wir bewusst zwischen *Topic* und Themenfeld: Ersteres meint ein algorithmisch erstelltes Cluster von Wörtern, Zweiteres zielt auf semantisch kohärente Themenbündel ab.

7 Siehe für die Literaturwissenschaften etwa: Christof Schöch: Topic Modeling Genre. An Exploration of French Classical and Enlightenment Drama. In: *Digital Humanities Quarterly 11* 2/2017, online: http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html (27.10.2021).

8 Alternativen zu LDA sind bspw. LSA oder LSI, vgl. dazu: David M. Blei/XY/XY: Latent Dirichlet Allocation. In: *Journal of Machine Learning Research* 3/2003), S. 993-1022, hier: S. 993f.; online: https://dl.acm.org/doi/10.5555/944919.944937 (27.10.2021).

9 Shawn Graham/Scott Weingart/Ian Milligan: Getting Started with Topic Modeling and MALLET, online: https://programminghistorian.org/en/lessons/topic-modeling-and-mallet (27.10.2021).

(bspw. DARIAH Topics Explorer)[10]. Bei beiden Ansätzen ist der direkte Eingriff in den maschinellen Lernprozess möglich. Ein Vorteil für den Vergleich beider *Engines* ist, dass es mittlerweile einen *Mallet-Wrapper* für Gensim gibt, der Ausführung und *(Hyper-)Parametertuning* der Java-Engine innerhalb einer Python-Umgebung ermöglicht.[11]

2.3 Korpora

2.3.1 Erstes Korpus: Zürcher Regierungsratsbeschlüsse des 19. Jahrhunderts

Die Regierungsratsbeschlüsse des Kantons Zürich stellen eine umfangreiche Reihe dar, die sich im *Close reading* gar nicht analysieren lässt. Die Aufschlüsse zu den Vorgängen in der Exekutive einer Verwaltungseinheit im Wandel des 19. Jahrhunderts versprechen jedoch vielfältige Einblicke in administrative Abläufe, akute Probleme und Verhandlungsformen und machen das Korpus zu einem optimalen Untersuchungsgegenstand. Die Texte liegen als digitale Dokumente in maschinenlesbarer Form vor.[12]

Bei den aufbereiteten Quellen handelt es sich um eine Reihe, die nach dem Ende der französischen Besatzung und der Helvetischen Republik 1803 einsetzt und handschriftlich bis 1898 weitergeführt wurde. Darin enthalten sind die schriftlich festgehaltenen Beschlüsse des Zürcher Regierungsrats. Die Dokumente streifen inhaltlich ein weites Feld: von Infrastrukturprojekten über Entscheide zu Einbürgerungen und Ausweisungen bis zu Gratulationen zur Geburt von königlichem Nachwuchs in benachbarten Staaten. Insgesamt sehen wir uns mit mehr als 150.000 Beschlüssen konfrontiert. Die Dokumente sind alle in

---

10 Severin Simmler/Vitt Torsten/Steffen Pielström: Topic Modeling with Interactive Visualizations in a GUI Tool. In: *Proceedings of the Digital Humanities Conference*. Utrecht 2019, online: https://dev.clariah.nl/files/dh2019/boa/0637.html (27.10.2021).

11 Das für den Nachvollzug erarbeitete Notebook nutzt die Möglichkeit von Gensim, Korpora mit Mallet aufzubereiten (dafür wird ein *Wrapper* genutzt, der innerhalb der Codeblöcke einen externen Algorithmus ausführt).

12 Durch das Staatsarchiv Zürich sind die Beschlüsse online zugänglich: https://www.archives-quick-access.ch/search/stazh/rrb (27.10.2021).

Deutsch, wobei sich die Sprache über den Zeitraum wandelt und immer stärker einer Norm folgt.[13]

2.3.2 Zweites Korpus: Lebensgeschichtliche Interviews aus den 1980er-Jahren

„Lebensgeschichte und Sozialkultur im Ruhrgebiet" (LUSIR) war das erste große Oral-History-Projekt in der BRD. Zwischen 1980 und 1988 wurden etwa 300 lebensgeschichtliche Interviews mit Arbeiter:innen, Angestellten, Gewerkschaftsfunktionär:innen und Betriebsrät:innen aus den großen Industrieunternehmen des Ruhrgebiets durchgeführt. Das ursprüngliche Erkenntnisinteresse waren Faschismus-Erfahrungen und Sozialkultur in der Montanindustrie des Ruhrgebiets zwischen 1930 und 1980.[14] Die daraus entstandenen Quellen liegen heute als analoge Audiotapes und digitale Kopien im Archiv „Deutsches Gedächtnis" des Instituts für Geschichte und Biographie der FernUniversität in Hagen.[15] In vergangenen und laufenden Forschungsprojekten wurden bereits zahlreiche Interviews transkribiert, sodass für diesen Aufsatz 166 Volltexte herangezogen werden konnten.[16] Bei einer Laufzeit der Interviews von bis zu acht Stunden bedeutet das einerseits eine gewaltige Textmasse, andererseits bieten die Stegreiferzählungen der Interviewten eine Bühne für Ausschweifungen in verschiedenste Lebens- und damit Themenbereiche. Diese kursorisch zu überschauen oder einfach zu durchsuchen, ist aussichtslos, sodass *Topic Modeling* hier das Potential bietet, Inhalte freizulegen, die mit dem bloßen Auge – oder Ohr – kaum wahrnehmbar sind.

---

13 Im Vergleich zu modernem Deutsch lassen sich daher mehr Schreib- und Wortvarianten („hujus" → diesjährigen) identifizieren, und gewisse Schreibungen unterscheiden sich noch („bey" → bei, „Theil" → Teil).

14 Vgl. Lutz Niethammer (Hg.): „Die Jahre weiß man nicht, wo man die heute hinsetzen soll." Faschismuserfahrungen im Ruhrgebiet. Berlin/Bonn 1983.

15 Online: https://www.fernuni-hagen.de/geschichteundbiographie/deutschesgedaechtnis (27.10.2021)

16 Online: https://www.fernuni-hagen.de/geschichteundbiographie/forschung/projekte/KA3.shtml; https://www.oral-history.digital/ (27.10.2021).

2.3.3 Drittes Korpus: Spätmittelalterliche Chronikhandschriften

Die spätmittelalterliche Chronik des Straßburger Geistlichen Jakob Twinger von Königshofen ist in knapp 130 Handschriften überliefert, wobei rund 100 davon mehr bzw. weniger als den integralen Chroniktext überliefern und neben Twingers Werk noch zahlreiche andere Texte beinhalten. Über diese Mitüberlieferung können Manuskriptwanderungen sichtbar gemacht werden, wobei unvollständige Inhaltsbeschreibungen und nicht normierte Werktitel eine Analyse erschweren. Nach einer automatischen Texterkennung (HTR) können über die Anwendung von *Topic Modeling* Verbindungen zwischen einzelnen Handschriften auf sprachlicher Ebene deutlich gemacht werden, die über das Auszählen von Worthäufigkeiten allein oder die Übereinstimmung bestimmter Begriffe nicht sichtbar würden. Als Fallstudie wurden sieben Handschriften aus dem Korpus ausgewählt;[17] von diesen weisen vier neben der Chronik mehrere Texte gemeinsam auf bzw. behandeln hauptsächlich Ereignisse die Stadt Konstanz betreffend und sind in abhängigen Kopierprozessen entstanden.[18] Dieses Vorwissen hilft bei der Einordnung bzw. Interpretation der Ergebnisse und den Überlegungen zu einer grundsätzlichen Anwendung von *Topic Modeling* auf größere Korpora mit nicht-standardisiertem Textbestand.

---

17 Dresden, Landesbibliothek, Mscr. F 98 [Dre1]; Freiburg, Universitätsbibliothek, Hs. 471 [Fre2]; Heidelberg, Universitätsbibliothek, Cpg 116 [Hei2]; Heidelberg, Universitätsbibliothek, Cpg 475 [Hei4]; München, Staatsbibliothek, Cgm 568 [Mue5]; Stuttgart, Landesbibliothek, Cod. HB V 22 [Stu3]; Wolfenbüttel, Herzog August Bibliothek, Cod. 16.17 Aug. 4° [Wol2]. Beschreibungen zu den einzelnen Handschriften finden sich unter https://handschriftencensus.de/werke/1906, eine aktuell gehaltene Auflistung aller Textzeugen inkl. Siglen in: Ina Serif: Der zerstreute Chronist. Zur Überlieferung der deutschsprachigen Chronik Jakob Twingers von Königshofen. In: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte*. 5.12.2015, aktualisiert am 28.10.2021, online: https://mittelalter.hypotheses.org/7063 (27.10.2021).

18 In den drei Codices Fre2, Hei4 und Mue5 stehen neben der Twinger-Chronik die „Konstanzer Jahrgeschichten", eine Konstanzer Bischofsliste und ein Bericht über die Ermordung des Lausanner Bischofs Guillaume de Menthonay 1406. Stu3 ist in Konstanz entstanden und enthält u.a. die „Konstanzer Weltchronik".

Korpus	Regierungsrats-beschlüsse	Lebensgeschichte und Sozialkultur im Ruhrgebiet (LUSIR)	Chronikhandschriften
Entstehungszeit	1803–1887	1980-1988	15. Jahrhundert
Anzahl Dokumente/ Einheiten	166.658 Beschlüsse	166 lebensgeschichtliche Interviews	7 Handschriften
Anzahl Wörter (mit Stopwords)	1.158.029	3.761.023	835.295
Anzahl Wörter (ohne Stopwords)	713.393	729.780	405.916

Tabelle 1: Quantifizierender Vergleich der verwendeten Korpora

2.4 *Preprocessing*: Den Text aufbereiten

Für alle drei Korpora, unabhängig davon, welchen Zeitbereich diese abdecken oder wie umfangreich sie sind, ist eine Aufbereitung über die Digitalisierung der Quellen hinaus notwendig. Diese Aufbereitungsschritte werden gemeinhin als *Preprocessing* bezeichnet. Da diese Schritte eine Manipulation der historischen Quellen darstellen, wird dieses „data cleaning" in den Digital Humanities in letzter Zeit intensiv kritisch diskutiert.[19] Die hier durchlaufenen Schritte sollten entsprechend nicht als standardisiert eingesetzte Abfolge verstanden, sondern reflektiert angewendet werden. Üblicherweise gehört *lower casing* (Kleinschreibung aller Zeichen), Entfernung von Satzzeichen und *Stopwords* sowie unter Umständen die Lemmatisierung (Rückführung von Wortformen auf ihre Grundform) zu diesen Arbeitsschritten. Bei ausgesprochen langen Dokumenten, die sich nicht weiter segmentieren lassen, ist schließlich auch die Zerlegung in kleinere Teile sinnvoll, da nur auf diesem Weg eine feingranulare Themenfluktuation sichtbar wird. Der dazu notwendige Schritt wird als *chunking* bezeichnet.

Etwas ausführlicher lohnt es sich, auf sogenannte *Stopwords* einzugehen. *Topic-Modeling*-Algorithmen arbeiten im Kern mit Worthäufigkeiten. Um eine Ansammlung von *Topics* mit nichtssagenden Inhalten (etwa Artikel, Pronomen, Paratext oder korpusspezifische Wörter

[19] Siehe dazu Katie Rawson/Trevor Muñoz: Against Cleaning. In: Matthew K. Gold/Lauren F. Klein (Hg.): *Debates in the Digital Humanities*. Minneapolis 2019, online: https://dhdebates.gc.cuny.edu/read/untitled-f2acf72c-a469-49d8-be35-67f9ac1e3a60/section/07154de9-4903-428e-9c61-7a92a6f22e51 (27.10.2021).

ohne semantischen Gehalt) zu vermeiden, arbeitet man – wie bei diversen anderen Textanalyse-Verfahren auch – mit sogenannten *Stopword*-Listen, die Zeichenfolgen mit potentiell hohem Vorkommen und geringem oder gar keinem semantischen Gehalt entfernen. Aus epistemologisch-hermeneutischer Perspektive muss entsprechend die Frage gestellt werden, wo wir Sinn in textuellen Elementen identifizieren und welche Kriterien wir dazu anwenden.[20] Vor diesem Hintergrund ist auch das Erkenntnisinteresse von Bedeutung, da dieses die Einschätzung beeinflusst, ob eine Zeichenfolge a priori als *Stopword* definierbar ist oder ob anhand des gemeinsamen Vorkommens mit anderen Wörtern innerhalb eines *Topics* a posteriori eine Sinnaufladung möglich wird. Für die drei Korpora wurde daher jeweils auf unterschiedliche *Stopword*-Listen zurückgegriffen bzw. neue erstellt.

3. Qualitativer Vergleich der Ergebnisse

Die Leistungsfähigkeit und Grenzen von *Topic Modeling* sowie der jeweiligen Engine zeigen sich eindrücklich, sobald die Ausgabe von *Topic*-Listen durchgesehen und Vergleiche angestellt werden. Mit wenigen Kenntnissen über ein Korpus wird offensichtlich, welche *Topics* sich zu welchen Themenfeldern verdichten lassen. Gleichzeitig bieten die *Topic*-Listen eine Möglichkeit, die unterschiedlichen *Preprocessing*-Schritte visibel, nachvollzieh- und vor allem in ihren Auswirkungen vergleichbar zu machen.

3.1 Die Regierungsratsbeschlüsse

Die mehreren zehntausend Regierungsratsbeschlüsse sind aufgrund ihrer Digitalisierung zwar im Volltext durchsuchbar und für die Forschung und Öffentlichkeit nun weitestgehend verfügbar. Eine wichtige Frage, nämlich welche Themen überhaupt wann verhandelt wurden, lässt sich aufgrund der Größe des Korpus aber nicht so einfach beantworten und könnte höchstens umständlich über die Frequenzen von Schlagwörtern eruiert werden. Mittels *Topic Modeling* ändert sich die Ausgangslage, da nicht nur Themen, sondern auch Tendenzen und vor

[20] Vgl. Patrick J. Burns: Constructing Stoplists for Historical Languages. In: *Digital Classics Online 4* 2/2018, S. 4-20.

allem Frequenzen nachvollzogen werden können. Damit lässt sich bereits aufgrund dieser *Distant-Reading*-Methode aufzeigen, welche Themenbereiche zu welchen Zeiten auftauchen. Ein valider Startpunkt der Auseinandersetzung ist die Analyse der generierten *Topics*, um festzustellen, welche Themenbereiche identifiziert werden können.

Nr.	MALLET	GENSIM	Nr.
12	familie, kinder, gemeinde, frau, armenpflege, anna, barbara, kind, maria, ehe, mutter, vater, unterstützung, ehegericht, elisabetha	d, gemeinde, bezirksrathes, beschluß, u, recurs, m, innern, familie, überwiesen, armenpflege, v, bülach, j, regierungsrath	47
13	militärs, kriegsrathe, kriegsrathes, infanterie, kriegsrath, oberst, eidg, regierungsrath, zürich, mannschaft, ziegler, truppen, major, antrag, artillerie	kriegsrathe, regierungsrath, hauptmann, d, zürich, kriegsrathes, kriegsrath, eidg, infanterie, m, v, betreffenden, ernennung, heinrich, quartier	0
14	ad, hherren, acta, protokoll, legen, hherrn, conferenz, bericht, verdankt, dank, verlesen, escher, rathsherr, staatsrath, abgeordneten	d, regierung, l, acta, jacob, ad, verfügung, gelegt, obergericht, recepiße, barbara, heinrich, legen, m, bescheint	32
15	regierungsrath, schweiz, bundesrath, innern, direktion, dr, zürich, mittheilung, beschließt, eid, ii, zuschrift, kreisschreiben, einladung, prof	rordorf, u, mittheilung, bericht, kreisschreiben, regierung, prinzeßin, näf, prinzen, fcs, v, sr, italien, d, anzeige	20
16	direktion, arbeiten, öffentlichen, regierungsrath, beschließt, einsicht, mittheilung, antrages, ii, öff, ermächtigt, gefängnißwesens, berichtet, finanzdirektion, finanzen	direktion, regierungsrath, zürich, d, u, dr, mittheilung, j, i, einsicht, v, ii, finanzen, beschlossen, wahl	21
17	gesandtschaft, schweiz, französischen, paris, schweizerischen, zürich, mittheilung, heinrich, geschäftsträger, nachforschungen, gesandten, de, schweizerische, handen, bericht	k, gesandtschaft, d, m, u, geschäftsträger, l, winterthur, paris, v, oesterreichischen, h, jahr, j, zuschrift	27
18	rath, kleinen, commißion, seyn, kleine, ihren, gemeinden, betreffenden, dießfälligen, antrag, rathe, herren, solle, bereits, raths	l, regierung, standes, stand, rath, tagsatzung, st, stände, kleinen, hohen, gallen, canton, commißion, bern, mißiven	2
19	zürich, winterthur, stadtrath, stadt, stadtrathes, stadtrathe, rordorf, städtischen, weiningen, stadtgemeinde, generalobligationen, uebertragung, stadtraths, meier, bauordnung	regierungsrath, gemeinden, winterthur, rath, d, u, direktion, großen, bericht, zürich, gemeinde, unterhalt, einsicht, mittheilung, a	48
20	brandaßecuranzcommißion, gebäude, brandbeschädigten, schaden, jacob, scheune, ß, heinrich, haus, schatzung, gemeinde, gebrüder, haushaltungen, häuser, wiederkehr	u, gemeinde, gemeinden, regierungsrath, anstalt, b, direktion, gebäude, gefangenen, namentlich, aufsichtsbehörde, arbeiten, sträflinge, a, kosten	12
21	kanton, salz, zürich, zentner, rheinau, schweiz, rappen, abgabe, zölle, verkauf, maß, waaren, salzes, einführung, preise	d, regierungsrath, zürich, kanton, großen, rath, zentner, l, escher, beschloßen, salz, bericht, gemeinde, kleinen, m	40

Tabelle 2: Topic Model basierend auf den Zürcher Regierungsratsbeschlüssen. Vergleich zwischen Mallet und Gensim mit identischer Stopword-Liste. Die Gegenüberstellung der Themen ist mehr aus einem Eindruck erwachsen und bildet keine perfekten Themenbereichspaare ab.

Während die fünfzehn Wörter eines *Topics* in Mallet meist einen mehr oder minder engen Themenbereich nahelegen, so sind die mit Gensim generierten *Topic*-Listen schwieriger interpretierbar. Auch fällt auf, dass Mallet viel weniger auf einzelne Zeichen reagiert und ‚Wörter', die nur aus einem Zeichen bestehen, eliminiert. Somit fallen Zeichenketten, die in den Dokumenten vielfach zur Abkürzung gebraucht werden, nicht ins Gewicht.

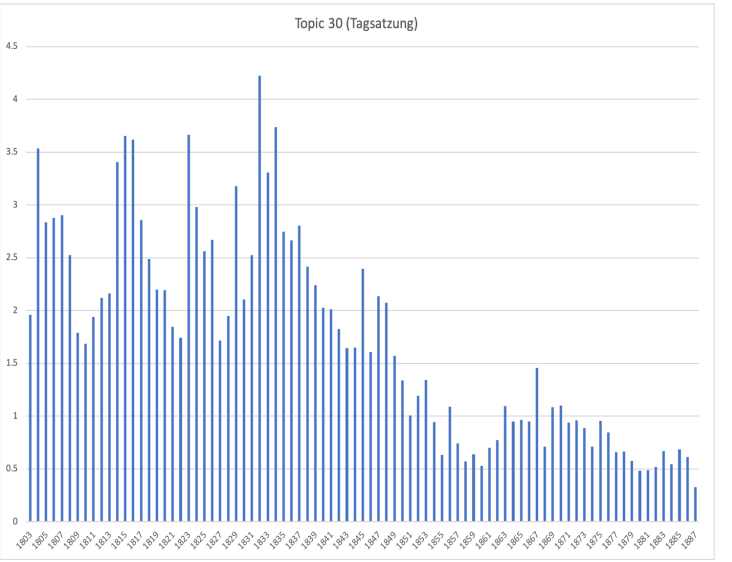

Abb. 1: Topic 30 (Mallet) zur Tagsatzung. Zusammengerechnetes Vorkommen pro Jahr.

Die bereits oben angesprochenen Tendenzen der Häufigkeiten von *Topics* über Jahre verteilt erlauben in den Regierungsratsprotokollen Rückschlüsse auf mehr oder minder diskutierte Themenfelder. Das kann beispielsweise anhand des *Topics* 30 demonstriert werden (mit Bezug zur Tagsatzung, der Vorgängerin der heutigen Bundesversammlung), das bis 1849 auftaucht und danach nur noch marginal (etwa aufgrund der Begrifflichkeit „eydgenössisch", die sich überschneidet) erscheint.

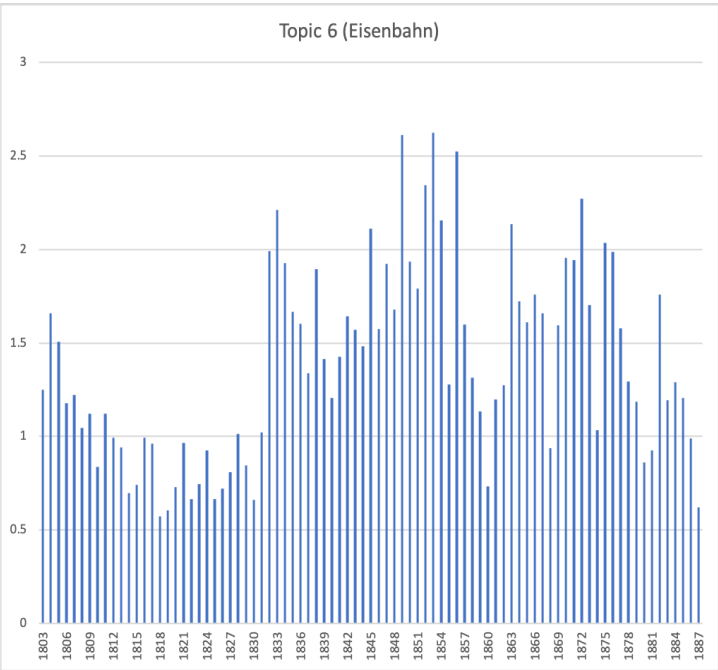

Abb. 2: Topic 6 (Mallet), mit Bezug zu Eisenbahn und Infrastruktur. Zusammengerechnetes Vorkommen pro Jahr (siehe dazu auch Fußnote 21).

Anhand des Balkendiagramms zu *Topic* 6 zum Eisenbahnbau und allgemeinen Infrastrukturprojekten lässt sich gut nachvollziehen, welche Zeiträume für diese Themen von besonderem Interesse sind bzw. in welchen Zeiten diese diskutiert wurden.
Der Vergleich von zwei *Topics* erlaubt über die Skala auch Aussagen zur Häufigkeit der Thematisierung in den Beschlüssen: *Topic* 6 erreicht knapp einen Wert von 2.5, während die häufiger diskutierte Tagsatzung über den Wert 4 reicht. Die Y-Achse wird bei dieser Visualisierung durch die Addition aller Vorkommen des Themas in einem Jahr errechnet (Kombination aller sog. *topic weights*).[21]

[21] Der Wert 1 bedeutet, dass ein *Topic* ein gesamtes Dokument ausmacht (was in der Praxis nie vorkommt). Typischerweise macht ein wichtiges *Topic* 10-25% (0.1-0.25) eines Dokuments aus.

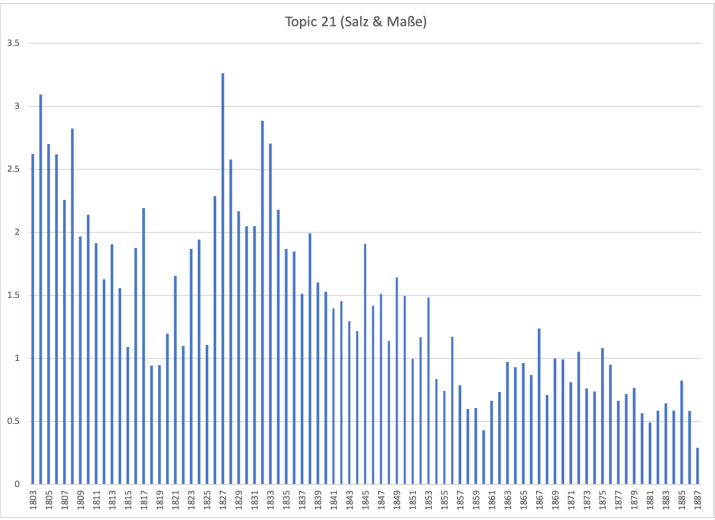

Abb. 3: Topic 21 (Mallet) mit Bezug zu Salz und Maßen.

Bereits auf einen Blick werden durch die Visualisierungen von *Topics* erste Aufschlüsse ermöglicht. So kann etwa nachvollzogen werden, dass mit der Gründung des Schweizer Bundesstaates (1848) die Hoheit über die Maßeinheiten und die wichtige Salzsteuer von den Kantonen an den Bund überging und diesbezüglich entsprechend weniger auf exekutiver Ebene des Kantons entschieden werden konnte.

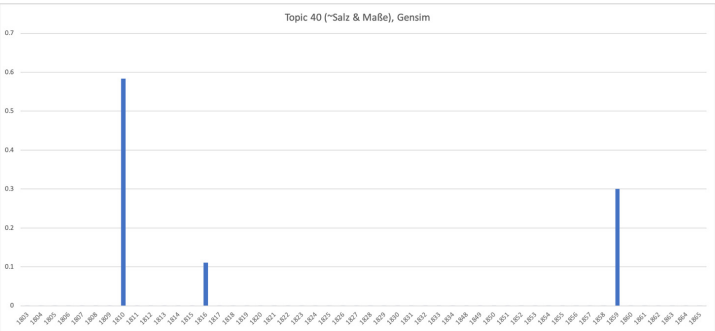

Abb. 4: Topic 40 mit Bezug zu Salz und Maßen, generiert aus Gensim-Daten. Das Topic kommt in jedem Jahr vor, häufig jedoch mit einem addierten Wert im Bereich von 0.00001 (Median) bis 0.02 (Mittelwert).

Im Vergleich dazu fehlt der aus Gensim gewonnenen Visualisierung die Aussagekraft. Das im Vergleich zu Mallet ähnlichste *Topic* kommt in den meisten Fällen weit weniger häufig vor, da Gensim von einigen wenigen, sehr generischen Themenfeldern dominiert wird. Die Ausschläge sind dann dafür umso stärker und stimmen nur partiell mit den Mallet-Graphen überein (ein besonders eklatantes Beispiel zeigt Abb. 4). Die Kombination der zeitlichen Ebene mit den aufsummierten Themenfeldern ergibt nochmals einen anderen Blick auf große Dokumentenmassen, der mit anderen Methoden nur schwierig zu erreichen ist. Aus der Vogelperspektive der Tendenzen drängt sich jedoch immer wieder der Blick in die Einzeldokumente auf, da sich keine Wertungen oder Haltungen aus den Balken herauslesen lassen. Dadurch lädt *Topic Modeling*, aufgeladen durch Visualisierungen, zum *Close Reading* und zur intensiven Beschäftigung mit ausgewählten Dokumenten ein.
Aufgrund der besseren Interpretierbarkeit und der zumindest oberflächlichen Geschlossenheit der Mallet-Themen wird die Weiterarbeit mit diesen Daten präferiert. Allein die Unterschiede der Resultate machen es gleichzeitig notwendig, weiter unten auf die Unterschiede der Algorithmen einzugehen, um erklären zu können, welche Differenzen bestehen.

3.2 Lebensgeschichtliche Interviews

Auch ein qualitativer Vergleich. der nun auf Grundlage der lebensgeschichtlichen Interviews exemplarisch demonstriert werden soll, kann mit einem *Distant Reading* beginnen, indem zunächst *Topic*-Listen miteinander verglichen werden. Es hat sich während der Evaluation von *Preprocessing* und *Parametertuning* herausgestellt, dass eine Aufteilung der umfangreichen Interviews (häufig mehrere Stunden lang) in kürzere Einheiten (Chunks) zu 10 (Gensim) und 25 Sätzen (Mallet) inhaltlich konsistente und aussagekräftige *Topics* nach sich zieht. Für die Bestimmung der optimalen Anzahl von *Topics* wurde eine einfache Metrik entwickelt, die in einem weiteren Aufsatz im Detail vorgestellt wird: Die

Keyword Diversity (KD) ist eine simplere und leicht nachvollziehbare Variante der verbreiteten *Exclusivity*-[22] oder *Relevance*-Metriken,[23] beide eng verwandt bzw. hergeleitet von der *Lift*-Methode.[24] Die KD berechnet prozentual, wie viele *Keywords* aus allen *Topics* einzigartig unter den ersten *n* Topwords der *Topics* vorkommen und sollte damit ein erster Indikator für die Trennschärfe der *Topics* sein. Es hat sich gezeigt, dass sich die *Keyword Diversity* bei etwa 50 bis 60 *Topics* auf einem Sockel einpendelt, daher werden hier die Modelle mit 50 *Topics* verglichen.
Wie verhalten sich nun die Ergebnisse von Gensim und Mallet zueinander? Bei der komparativen Durchsicht der Listen überzeugt Mallet. Ein vergleichender Blick allein auf das jeweils am höchsten gewichtete *Keyword* eines *Topics* zeigt, dass Gensim – das in der Evaluation stets eine wesentlich niedrigere KD aufwies – äußerst redundante Ergebnisse liefert: In fünf *Topics* steht das Wort „Krieg" an der Spitze, in acht weiteren an zweiter oder dritter Stelle. Insgesamt ist „Krieg" in 29 von 50 *Topics* unter den zehn gewichtigsten *Keywords* – bei Mallet hingegen lediglich in sechs, davon dreimal an erster Stelle. Ein weiterer Indikator für wenig Trennschärfe ist das doppelsinnige Wort „Essen", das einerseits die Ruhrmetropole, andererseits Nahrungsmittel bezeichnet. In den Gensim-Ergebnissen steht es in sechs *Topics* an erster, bei sieben an zweiter oder dritter Stelle der *Keywords*, insgesamt findet es sich in 22 *Topics* unter den ersten zehn *Keywords*. Bei Mallet findet man es nur einmal an der Spitze eines *Topics* und viermal an zweiter oder dritter Stelle, insgesamt ist es unter den zehn Top-*Keywords* der 50 *Topics* nur sechsmal vertreten. Der Vergleich bestätigt anhand dieser beiden Wörter die Evaluationsroutine der *Keyword Diversity*. Doch wie steht es um die Disambiguierung? „Essen" als Wohn- und Industriestandort ist in beiden Modellen weit verbreitet. Daneben steht es bei Gensim in vier

---

22 Jonathan M. Bischof/Edoardo M. Airoldi: Summarizing topical content with word frequency and exclusivity. In: *Proceedings of the 29th International Conference on Machine Learning*. Edinburgh 2012. Online: https://arxiv.org/abs/1206.4631v1 (27.10.2021).

23 Carson Sievert/Kenneth E. Shirley: LDAvis. A method for visualizing and interpreting topics. In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Baltimore 2014, S. 66, online: http://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf (27.10.2021).

24 Matthew A. Taddy: On Estimation and Selection for Topic Models. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Fort Lauderdale 2011, online: http://proceedings.mlr.press/v15/ (27.10.2021).

von 22 Bezügen in einem Nahrungsmittelkontext. Doch während diese *Topics* sowohl heterogen als auch redundant sind (vgl. Tabelle 3), zeichnen sich die drei Nahrungsmittel-*Topics* bei Mallet durch innerliche Geschlossenheit aus und grenzen sich durch verschiedene Kontexte voneinander ab: Nahrungsmittel, Essen im Rahmen bestimmter Anlässe wie „weihnachten", „geburtstag" oder „feiern" und Essen im Sinne des Akts einer Mahlzeit – „morgens", „mittags", „abends". Diese Ergebnisse liefern deutliche Hinweise auf algorithmische Unterschiede zwischen den beiden *Topic-Modeling-Engines*.

Wenn man ein Sample zur Arbeiterkultur im Ruhrgebiet untersucht, stellt sich zuerst die Frage nach der Abbildung von Montanindustrie und Bergbauerfahrung. Repräsentative *Topics* findet man bei beiden Modellen, dargestellt durch die Begriffe „zechen", „kohlen", „kumpel", aber auch durch Fachtermini wie „steiger", den aufsichthabenden Bergleuten unter Tage, und etwas abgeschlagen „gedinge", der leistungsbasierten Entlohnung im Bergbau. Während sich das einschlägigste *Topic* bei Mallet beinahe als systematisches Wortfeld darstellt, in dem sich unter den ersten 15 Wörtern auch „betriebsführer", „hauer" oder „schachtanlage" finden, tauchen im entsprechenden Gensim-*Topic* Begriffe wie „arzt", „kranken" und „morgens" auf.

Ein weiteres, für Zeitgeschichtsforschung und NS-Aufarbeitung bedeutsames Thema ist die Judenverfolgung und -vernichtung im Dritten Reich. Insbesondere das vielbeschworene „Schweigen" über das Dritte Reich in der frühen Nachkriegszeit kann durch die Darstellungen und Narrative der in den 1980er-Jahren mit eben dieser vom Krieg betroffenen Generation geführten Interviews *ex post* aufgearbeitet werden.[25] Doch bei näherer Durchsicht der Ergebnisse zeigt sich, dass nur Mallet ein konsistentes *Topic* zur Judenverfolgung zusammengestellt hat, während die thematischen Bezüge bei Gensim lediglich stellenweise durchschimmern. Dessen *Topic*, in dem auf dem Lemma „jude" basierende Wörter am höchsten gewichtet sind, lässt keine Engführung erkennen, die ersten *Keywords* sind „essen" und „haushalt", „juden" folgt an neunter Stelle und steht allenfalls mit dem fünfzehnten Begriff „gedenken" in Verbindung (vgl. Tab. 3). Demgegenüber finden sich

[25] Vgl. Gabriele Rosenthal: Kollektives Schweigen zu den Nazi-Verbrechen. Bedingungen der Institutionalisierung einer Abwehrhaltung. In: *psychosozial 51* 3/1992, S. 22-33.

bei Mallet weit über die abgebildeten 15 *Top-Keywords* hinaus eindeutige Bezüge zum Themenkomplex Antisemitismus, Judenverfolgung und Shoah – beispielsweise „geschäfte", „konzentrationslager", „synagoge", „reichskristallnacht", „kzs", „verschwinden" und „umbringen". Hier bietet sich ein guter Einstiegspunkt für ein abschließendes *Close Reading*, um die Ergebnisse direkt im Text zu überprüfen und in diesem besonderen Fall die Tragweite der Darstellungen als historisch signifikantes Exempel heranzuziehen.

Die Rückverfolgung in den Text liefert tatsächlich aussagekräftige Passagen, in denen die Interviewten von Antisemitismus und Judenverfolgung im Dritten Reich erzählen. Im Großen und Ganzen bestätigt sich das bekannte Bild mit Aussagen wie *„da war die Kristallnacht, aber wir hier haben nichts mitgekriegt"* oder *„von den Konzentrations…, dass sie die alle da versteckten haben wir nicht gewusst, nein."*[26] Daneben finden sich verschiedene Varianten von Relativierungen, etwa durch Vergleiche mit der DDR: *„Gehen sie mal in die DDR und sagen da was, kommen sie auch weg"*. Diese Stichproben bestätigen exemplarisch die inhaltliche Fokussierung der Mallet-Topics und untermauern das Potential von *Topic Modeling* zur explorativen Erschließung lebensgeschichtlicher Interviews.

[26] Vgl. hierzu: Peter Longerich: *„Davon haben wir nichts gewusst!" Die Deutschen und die Judenverfolgung 1933-1945*. München 2006 // Bernward Dörner: *Die Deutschen und der Holocaust. Was niemand wissen wollte, aber jeder wissen konnte*. Berlin 2007.

Nr.	MALLET Chunks à 25 Sätze, lemmatisiert, POS-Tags: Nomen, Eigennamen, Verben, Adjektive, Adverbien	GENSIM Chunks à 10 Sätze, lemmatisiert, POS-Tags: Nomen, Eigennamen, Verben, Adjektive, Adverbien	Nr.
	Krieg		
26	krieg, gedenken, angst, gott, passieren, reden, wussten, schwer, menschen, mensch, schlecht, lieb, arbeit, ehrlich, froh	krieg, essen, bunker, gott, helfen, passieren, versuchen, mensch, lieb, schlecht, leuten, kontakt, arbeit, deutsch, chef	2
28	krieg, einziehen, arbeitsdienst, soldat, bruder, krieges, weltkrieg, freiwillig, mitmachen, gefangenschaft, entlassen, zurückkommen, urlaub, russland, kriege	krieg, schule, erinnern, kind, deutsch, gebären, erleben, kinder, einziehen, amerikaner, mitmachen, schwester, lernen, besuchen, russland	6
39	krieg, erinnern, schlecht, zeiten, arbeitslos, 50er, erleben, menschen, ruhrgebiet, fünfziger, nachkriegszeit, arbeiter, verändern, normal, sachen	krieg, erinnern, kinder, moment, kontakt, bezahlen, wohnung, mädchen, gedenken, gott, arbeit, heiraten, eltern, verdienen, wohnen	39
	Essen: Disambiguierung		
8	essen, brot, kartoffeln, garen, bauern, pfund, butter, land, hamstern, backen, kinder, lebensmittel, milch, hunger, bauer	butter, essen, sachen, pfund, wohnen, fleisch, arbeit, schwer, mensch, meinung, abends, lernen, deutsch, leuten, fahrrad	8
	Bergbau		
7	zechen, bergbau, steiger, kohle, arbeit, verdienen, kumpel, bergmann, schicht, bergleute, betriebsführer, hauer, schacht, schachtanlage, kumpels	steiger, krieg, kohle, verdienen, kumpel, schicht, meter, bergbau, arbeit, erinnern, zechen, kohlen, arzt, kranken, morgens	49
	Judenverfolgung		
9	juden, berlin, jude, gewusst, jüdisch, jüdische, münchen, mitkriegen, deutsch, krieg, hitler, kristallnacht, amerika, wussten, politisch	essen, haushalt, mädchen, eltern, schwer, wohnen, wohnung, verdienen, juden, mensch, menschen, kind, zweit, schwiegermutter, gedenken	30

Tab. 3: Topic-Listen des LUSIR-Korpus.

3.3 Spätmittelalterliche Sammelhandschriften

Während die ersten beiden Korpora vollständig manuell transkribiert oder nach automatisierter Texterkennung nachkorrigiert wurden, fand beim Sample der texterkannten mittelalterlichen Handschriften keine Korrektur des Outputs statt. Für die Texterkennung wurde Transkribus mit generischen Modellen verwendet, d.h. für die jeweiligen Schreibhände ein möglichst passendes existierendes Modell gewählt, um ohne zusätzlichen Aufwand für ein allfälliges Training bzw. zur Verbesserung der Erkennung Volltexte für die sieben ausgewählten

Handschriften zu erhalten.[27] Bis auf das Auflösen diakritischer Zeichen wurden die Textdaten nicht bereinigt,[28] dies aus mehreren Gründen: Eine manuelle Korrektur wäre zu zeitaufwendig – die Codices haben einen Umfang von bis zu 580 Seiten –, ein automatisiertes *text cleaning*, bspw. über bestehende *Normalizer* für vormoderne Sprachen,[29] ist ohne ebenfalls sehr aufwendiges Training (noch) nicht verfügbar. Auch sollte das Sample als *Proof of Concept* dienen und zeigen, inwieweit eine solch imperfekte Datengrundlage dennoch bereits zu weiterführenden Ergebnissen führt. Vor allem aber ist das Erkenntnisinteresse bei der Anwendung von *Topic Modeling* bei diesem Korpus etwas anders gelagert als bei den anderen beiden Quellenbeständen: Während *Topic Modeling* auch hier – trotz fehlender Korrektur, Normalisierung oder Lemmatisierung – zu interpretierbaren bzw. sinnvoll erscheinenden *Topics* führt, liegt der Nutzen der Methode auch in einer potentiellen Sichtbarmachung von Abschreibevorgängen: Zwar werden in Handschriftenbeschreibungen oftmals Schreibdialekte bzw. Ursprungsorte/-regionen für Manuskripte vermerkt, allerdings nicht standardisiert und stark abhängig von der Expertise des/der jeweiligen Bearbeiter:in. Wenn also für vier der sieben Handschriften aus dem Sample die Schreibsprache als „schwäbisch" [Fre2], als „südliches Niederalemannisch" [Hei4], als „ostschwäbisch, beeinflusst von elsässischer und hochalemannischer Vorlage" [Mue5] und als „bodenseealemannisch-schwäbisch" [Stu3] bezeichnet wird, lassen sich diese Codices anders als durch einen menschlichen Blick auf die nicht-standardisierten Metadaten und durch Fachwissen um die Nähe der zugewiesenen Dialekte kaum in einen Zusammenhang bringen.

---

27 Siehe online: https://readcoop.eu/de/transkribus/oeffentliche-modelle/ (27.10.2021). Die Texterkennung wurde mit Credits aus dem Scholarship Programme von Transkribus durchgeführt, vgl. ebenfalls online: https://readcoop.eu/transkribus/scholarship.

28 Wörter mit Diakritika wurden bei der Texterkennung häufig zerteilt, d.h. aus „brůder" wurden die zwei Strings „brů" und „der". Nach Auflösung wurde daraus ein String, „bruoder". Ganz gelöscht wurden Informationen der digitalisierenden Institutionen, die z.T. im Digitalisat vorhanden sind und daher ebenfalls als Text erkannt wurden.

29 Der *Normalizer* „Norma" wird aktuell nicht weiterentwickelt, das letzte Release ist von 2017; vgl. online: https://www.linguistics.rub.de/comphist/resources/norma/index.html und https://github.com/comphist/norma (27.10.2021).

Über gemeinsame *Topics* lässt sich allerdings eine Verbindung erkennen, die sich bei den vier genannten Handschriften auch über geteilte Texte bzw. den inhaltlichen Fokus auf Konstanz erklären lässt; sie dürfte aber auch auf die Schreibsprache zurückzuführen sein. Da für diese gerade auch nicht-sinntragende Wörter sehr aufschlussreich sind, wurden für den Datensatz sowohl vor als auch nach dem Ausschluss von *Stopwords* Modelle trainiert.[30] Auch wurden Modelle für die integralen Dokumente ebenso wie für Chunks à 250, 500, 1000, 2500 und 5000 Wörter trainiert. Die Auswertung bzw. Ergebnissichtung erfolgte vor allem über *Keyword*-Listen und *Document-Topic-Heatmaps*, die das gewichtete Vorkommen eines *Topics* in einer Handschrift aufzeigen. Dabei ließ sich feststellen, dass Mallet durchgehend bessere Ergebnisse bei der Abgrenzung bzw. Zusammensetzung einzelner *Topics* lieferte; mit Gensim wurden einzelne *Topics* oftmals wiederholt. Komplette Nonsense-*Topics* wie Nr. 2 (Gensim) kamen bei Gensim häufiger vor (vgl. Tab. 4).

Nr.	MALLET	GENSIM	Nr.
0	swarber, sache, rasz, state, kolmer, gerittens, etwart, vr, ec, handen, gerant, uker, nuermeer, minat, künngen	stroszburg, guot, cc, stat, lant, stette, ccc, sant, vō, wenne, mā, wasz, heilge, zeichen, kint	7
1	xviiij, lya, erkante, cloester, begieng, scraffen, frùntschafft, gebun, frassent, rme, tytus, istahel, spysen, towe, romer	stat, künig, keyser, sant, mā, anno, statt, lant, bischoff, widder, stroszburg, geburt, keiser, dag, volck	8
4	anno, widder, weyssenburg, gem, statt, dag, stat, abt, ite, herren, landt, sagt, gefangen, herr, juncker	anno, weyssenburg, widder, stat, gem, statt, dag, sant, abt, ite, herren, landt, sagt, gefangen, juncker	10
5	küng, babst, kayser, volk, vō, ziten, it, herren, byschoff, richsete, machte, moyses, hiesz, lant, rom	küng, babst, kayser, land, vō, volk, rich, sant, stat, rom, starb, kung, it, gottes, ziten	0
7	mā, stat, keyser, künig, sant, lant, stroszburg, rome, keiser, bischoff, gottes, geburt, stette, volck, bobst	mā, stat, lant, keyser, künig, sant, bischoff, stroszburg, geburt, gottes, keiser, rome, volck, stette, rich	13
10	vō, babst, küng, kayser, volk, heren, recht, uolrich, hand, bischoff, rome, frawen, tod, reich, vater	küng, bapst, kayser, land, kung, statt, sant, rich, won, grosz, heren, bischoff, hundert, ziten, starb	1
16	bapst, won, wand, und, land, statt, lande, über, hand, kung, zechen, acht, zwen, ddas, imn	bapst, won, wand, und, get, land, hand, sant, zwen, statt, lande, über, kung, zechen, acht	14
17	mechtig, aund, sachent, wunder, meinan, kat, gebenen, handet, erss, dienen, klen, wi, goten, muest, guang	sant, stat, mā, keyser, künig, wand, bapst, won, lande, bischoff, lant, land, gottes, rome, statt]	15
19	mnt, wuchen, snen, seber, gi, warent, iser, enser, begent, bapse, diy, geboten, kanden, lte, eden	wvil, zehgne, zehft, zegteit, zegt, zeggost, zegen, zegein, zefsanen, zefet, zeen, zeei, zee, zediin, zedenstestzen	12

30 Wie bei den Regierungsratsprotokollen existiert keine direkt anwendbare *Stopword*-Liste für die mittelhochdeutschen Texte. Für das Korpus wurde eine bestehende Liste aus dem „Classical Language Toolkit", einer Python-Library für das *Natural Language Processing* vormoderner Sprachen, entsprechend erweitert; vgl. online: http://cltk.org/ bzw. https://github.com/cltk/cltk (27.10.2021).

Tab. 4: Topic Model basierend auf den sieben Handschriften. Vergleich zwischen Mallet und Gensim mit identischer Stopword-Liste auf den Volltext-Dokumenten, Ausgabe von zwanzig Topics.

Die Chronik Jakob Twingers, die in den meisten Handschriften den größten Umfang besitzt, gibt in den universalgeschichtlichen Kapiteln zwei und drei weltliche und geistliche Herrschergeschichte wieder, in den Kapiteln vier und fünf Straßburger Bischofs- und Stadtgeschichte. Ein Kreisen der *Topics* um Begriffe wie Kaiser, König, Papst, Stadt und Straßburg ist daher also wenig erstaunlich. Im Gegensatz zu Gensim werden mit Mallet aber auch kleinere *Topics* aufgedeckt, die aus der *Ulrichslegende* (in Handschrift Mue5) stammen (Mallet *Topic* Nr. 10) oder Ergänzungen zur Stadt Colmar (*kolmer*, in Handschrift Dre1) betreffen (Mallet *Topic* Nr. 0).
Die bereits bekannte inhaltliche Verbindung der drei bzw. vier Codices Fre2, Hei4, Mue5 und Stu3 wurde über die Heatmaps vor allem auf den gechunkten Dokumenten mit Gensim deutlich (in der Heatmap orange umrandet); das Aufzeigen der Nähe zweier weiterer Codices, Dre1 und Wol2, die nicht auf textlicher, sondern vielmehr dialektaler Ebene liegt, ließ sich hier besser erkennen (violett umrandet) als auf einer auf Mallet-*Topics* basierenden Heatmap:

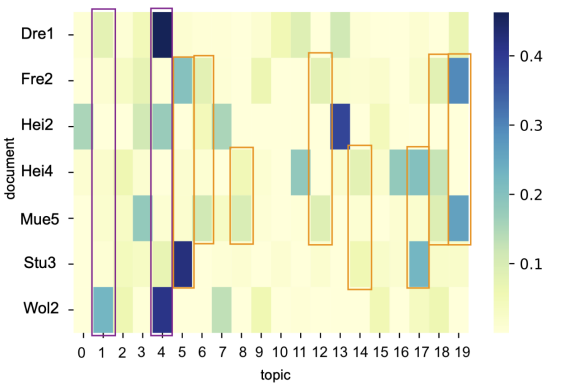

Abb. 5: Document-Topic-Matrix, Gensim, Chunks à 2500 Sätze, 20 Topics.

Die beiden Handschriften Dre1 und Wol2 haben, bis auf die Twinger-Chronik, keine gemeinsamen Texte; ihre Schreibsprachen wurden aber als „elsässisch" bzw. „nordelsässisch" klassifiziert – ein *Close Reading* der entsprechenden Auszüge aus der Chronik böte sich im Anschluss an die Ergebnisse des *Topic Modelings* an, um Rückschlüsse auf mögliche Abschreibeprozesse zu ziehen, die bisher übersehen wurden.
Die für diese spätmittelalterlichen Dokumente erzeugten *Topic*-Listen weisen, im Gegensatz zu den anderen beiden Korpora, einige Redundanzen auf – Formen wie „keyser", „kùnig" und „bapse" stehen neben „kayser", „kùng" und „babst". Eine Normalisierung, Lemmatisierung oder andere Art von *Preprocessing* würde diese dialektalen Eigenheiten in den einzelnen Dokumenten ausmerzen und das Ergebnis gänzlich verändern – mit Blick auf eine inhaltliche Erschließung der Sammelhandschriften würden sicher detailliertere und konzisere *Topics* ausgegeben. Dies wäre bei größeren Korpora bzw. Codices, deren Inhalte noch unzureichend erschlossen sind, äußerst aufschlussreich und könnte einen guten Ausgangspunkt für weiterführende Analysen darstellen. Der Verzicht auf eine Normalisierung macht hingegen Gemeinsamkeiten auf sprachlicher Ebene sichtbar und weist auf potentiell verwandte Handschriften, also Kopierprozesse hin, die in den meisten Fällen sonst nur durch mühsame Einzelvergleiche aufgedeckt werden können. *Topic Modeling* kann hier also vor allem auch als Hinweisgeber verstanden werden, gerade für noch wenig untersuchte Codices, wobei Mallet und Gensim jeweils unterschiedliche Vorteile besitzen: Während Mallet trennschärfere *Topics* erstellt und bei der inhaltlichen Erschließung hilfreich sein kann, werden über die mit Gensim erstellten *Document-Topic-Matrizen* sprachliche Verwandtschaften deutlicher. Ein Zusammenspiel beider Engines ist also je nach Untersuchungsgegenstand und -frage sinnvoll.

4. Verwendete Engines

In allen drei Korpora wurde offenbar, dass die Wahl der Engine unterschiedliche Resultate generiert. Die beiden Verfahren bauen zwar auf derselben Idee auf, dem *Clustering* von Wörtern, die in ähnlichen Kontexten vorkommen. Dieses *Clustering* erfolgt indes unterschiedlich und benötigt auch unterschiedlich viele Rechenressourcen. Ein Blick in die

Verfahren zeigt denn auch, dass bereits kleine Differenzen im Clusteringprozess zu stark divergierenden Resultaten führen.
Der Hauptunterschied zwischen den zwei Engines Mallet und Gensim ist die Approximation der *a posteriori*-Wahrscheinlichkeit in der Bayes'schen Wahrscheinlichkeitsrechnung, die von so-genannten Inferenz-Algorithmen errechnet wird. Nachdem der zufällige Ausgangspunkt als erster Prior festgelegt wird (typischerweise ein Wort und sein Verhältnis zu allen anderen Wörtern), wird in vielen kleinen Schritten versucht, daraus die *a posteriori*-Wahrscheinlichkeit des Auftretens mit anderen Wörtern zu errechnen. Zur Verbesserung des *Clustering* wird also ein anderes Wort zu Hilfe genommen und gemessen, wie wahrscheinlich es ist, dass diese zwei Wörter Teil eines Clusters sein könnten. Zur Eruierung dieser und aller weiteren Mini-Cluster gibt es unterschiedliche Inferenz-Methoden, im Kern die beiden Bereiche Sampling und Optimierungsfunktionen.[31]
Während Mallet *Gibbs Sampling* verwendet, stützt sich Gensim auf die ressourcenschonendere und schnellere, aber effektiv unpräzisere *VariationalBayes*-Methode.[32] Diese geht davon aus, dass alle *features* (im *Topic Modeling* „Wörter") unabhängig sind, und versucht in fortlaufenden Schritten die Optimierungsfunktion in Richtung der Minimierung der Kullback-Leibler-Divergenz zwischen *a priori*- und *a posteriori*-Verteilung aufzulösen.
Beim Gibbs Sampling hingegen, das in der Tradition der *Markov Chain Monte Carlo*-Verfahren (MCMC) steht, werden Stichproben gezogen. In einer Art ‚Spaziergang durch den Vektorraum' werden nach und nach zufällig gezogene Samples, also Textabschnitte, miteinander verglichen und so die *a posteriori*-Wahrscheinlichkeiten annäherungsweise berechnet.[33] Dabei wird – ähnlich wie bei *n*-Grammen – immer die letzte Berechnung zum Vergleich mit der nächsten herangezogen. Auf diese Weise wird eine sogenannte *Markov-Kette* generiert, die entsprechend rechenaufwendiger ist als das *Variational Bayes*-Verfahren.

---

31 Zur vergleichenden Einführung: Christopher M. Bishop: *Pattern Recognition and Machine Learning*. New York 2006, S. 461ff.; 523ff.

32 David M. Blei/Alp Kucukelbir/Jon D. McAuliffe: Variational Inference. A Review for Statisticians. In: *arXiv:1601.00670v9* S. 3; online: https://arxiv.org/abs/1601.00670 (27.10.2021).

33 Philip Resnik/Eric Hardisty: Gibbs Sampling for the Uninitiated, S. 7; online: https://drum.lib.umd.edu/handle/1903/10058 (27.10.2021).

Der große Wurf bei der *Latent Dirichlet Allocation* (LDA), die sich unter vielen *Topic-Modeling*-Algorithmen beinah als Standard durchgesetzt hat und das ‚Dach' dieser Verfahren bildet, ist, dass jedes Wort und jeder Textabschnitt nicht nur einem Cluster zugewiesen werden kann.[34] Die Offenlegung „latenter" Sinnstrukturen eines Textes, so Blei et al., sei nur dann zu haben, wenn jede Entität mehreren Clustern angehören kann. Das ermöglicht etwa eine bessere Disambiguierung mehrdeutiger Wörter oder die Abbildung kontextbedingten Bedeutungswandels von Wörtern. Die Qualität der semantischen Mehrschichtigkeit von mit LDA generierten *Topics* hängt letztlich von den Ergebnissen der Inferenz-Algorithmen ab. Die deutlichen Unterschiede zwischen Gensim und Mallet werden letztlich vor dieser Schichtung von Sinnebenen besonders sichtbar, die mangelhafte Konsistenz und Trennschärfe von *Topics* offenlegt. Es darf nicht unerwähnt bleiben, dass der Beweggrund für die Entwicklung von Gensim – Rechenaufwand zu reduzieren, um auch kleineren Forschungsprojekten *Topic Modeling* zu ermöglichen – für diese Studie keine Bewandtnis hatte. Die abgerufene Rechenleistung und Laufzeit beider Engines waren vergleichbar und äußerst gering. Bei dem Versuch, die Ergebnisse von Gensim denen von Mallet anzugleichen, ergab sich sogar eine gegenläufige Tendenz: Trotz bis zu zehnmal längerer Laufzeit durch extremes *Parametertuning* konnten die Gensim-*Topics* die Konsistenz der Mallet-*Topics* nicht erreichen.

5. Schluss

Obwohl dieser Beitrag nur ein Schlaglicht auf die Möglichkeiten von *Topic Modeling* wirft, zeigt sich anhand der drei vorgestellten Korpora, wie weitreichend bereits das unsupervisierte Auffinden von Clustern ist und wie auf diese Weise mit wenig Aufwand unstrukturierte Korpora exploriert und Themenfelder eruiert werden können. Je nach Korpus und Fragerichtung sind damit bereits erste Interpretationen, das Auffinden thematisch einschlägiger Passagen, aber auch der Nachvollzug von Abschreibeprozessen möglich. Unser Kapitel zeigt gleich-

[34] David M. Blei/Andrew Y. Ng/Michael I. Jordan: Latent Dirichlet Allocation, In: *Journal of Machine Learning Research* 3 (2003), 993-1022 . hier: S. 1000.

zeitig, wie abhängig die Ergebnisse des *Topic Modeling* von den genutzten Algorithmen/Engines sind. Bereits der Einsatz von unterschiedlichen Inferenz-Algorithmen beim Finden der Cluster führt zu gravierenden Unterschieden bei der Konsistenz und Trennschärfe der *Topics*. Schließlich sind *Topic Models* eng an die Korpora gebunden, die zur Berechnung genutzt wurden. Auch die Fragestellungen, die durch *Topic Modeling* beantwortet werden können oder zu (ersten) Interpretationen führen, sind kritisch abzuwägen und mit *Close-Reading*-Prozessen auf ihre Stringenz und Folgerichtigkeit zu überprüfen. Die hier aufgezeigten Ansätze kratzen erst an der Oberfläche, wenn die Methode breit als neue Form der Heuristik in den Geschichtswissenschaften eingesetzt werden soll. Das gesamte *Preprocessing* müsste einer kritischen Lektüre ausgesetzt und die dabei vorgenommenen Schritte gegeneinander abgewogen werden. Ebenso kritisch muss mit den in den Computer- und Informationswissenschaften entwickelten und verwendeten Messgrößen zur Evaluierung von *Topic Models* verfahren werden. Es ist jüngst angemerkt worden, wie oft die qualitative Auswertung der *Topics* auf der Strecke bleibt.[35] Die Geschichtswissenschaft sollte es sich in diesem Kontext zur Aufgabe machen, im Sinne einer Algorithmenkritik digitale Methoden kritisch zu analysieren und durch systematische Studien zum Verständnis und zur Transparenz der genutzten Verfahren beizutragen. Dabei darf nicht in ein starres *Distant* vs. *Close Reading* verfallen werden. Es gilt vielmehr, die Reziprozität beider Ansätze zu betonen und aufzuzeigen, welches Potential und welche Risiken die Symbiose von quantitativen und qualitativen Verfahren angesichts stetig wachsender Textkorpora bergen.

[35] James Dobson: Interpretable Outputs: Criteria for Machine Learning in the Humanities. In: *Digital Humanities Quarterly 15* 2/2021, online: http://www.digitalhumanities.org/dhq/vol/15/2/000555/000555.html (27.10.2021).