hosted by Hamburg University Press

apropos
[Perspektiven auf die Romania]

Sprache∫Literatur∫Kultur∫Geschichte∫Ideen∫Politik∫Gesellschaft

„Nuit, correspondance, sentiment"
Topic Modeling auf einem Korpus von französischen Romanen 1750-1800

Anne Klee & Julia Röttgermann

apropos [Perspektiven auf die Romania]
hosted by Hamburg University Press
2022, 9
pp. 57-86
ISSN: 2627-3446

Online
https://journals.sub.uni-hamburg.de/apropos/article/view/1888

Zitierweise

Klee, Anne & Julia Röttgermann. 2022. „`Nuit, correspondance, sentiment´. Topic Modeling auf einem Korpus von französischen Romanen 1750-1800." *apropos [Perspektiven auf die Romania]* 9/2022, 57-86.
doi: https://doi.org/10.15460/apropos.9.1888

2022, n°9
pp. 57-86
doi: https://doi.org/10.15460/apropos.9.1888

Anne Klee & Julia Röttgermann

„Nuit, correspondance, sentiment"

Topic Modeling auf einem Korpus von französischen Romanen 1750-1800

Anne Klee
ist wissenschaftliche Mitarbeiterin am *Trier Center for Digital Humanities* der Universität Trier.
klee@uni-trier.de

Julia Röttgermann
ist wissenschaftliche Mitarbeiterin am *Trier Center for Digital Humanities* der Universität Trier.
roettger@uni-trier.de

Keywords

18. Jahrhundert – Französische Literatur – Topic Modeling – Linked Open Data – Wikidata

1. Topic Modeling

Wie lassen sich große Korpora hinsichtlich ihrer literarischen Themen explorativ digital erforschen? Das Verfahren Topic Modeling, das in der hier vorgestellten Implementation auf der statistischen Methode *Latent Dirichlet Allocation* basiert (Blei, Ng, und Jordan 2003), kann zur Analyse verschiedener Arten von Daten eingesetzt werden: Neben beispielsweise Bildern, genetischen Daten oder Zeitungsarchiven (Blei 2011) lässt sich der Algorithmus mit entsprechenden Parametereinstellungen auch auf literarische Texte anwenden, wie zahlreiche Studien zeigen konnten (Underwood 2012; Rhody 2013; Jockers 2014). Die Methode, die von einer möglichst großen Datenmenge profitiert, berücksichtigt vor allem die Kookkurrenz von Wörtern und generiert auf dieser Grundlage Gruppen von semantisch verwandten Wörtern, die als Topics bezeichnet werden.

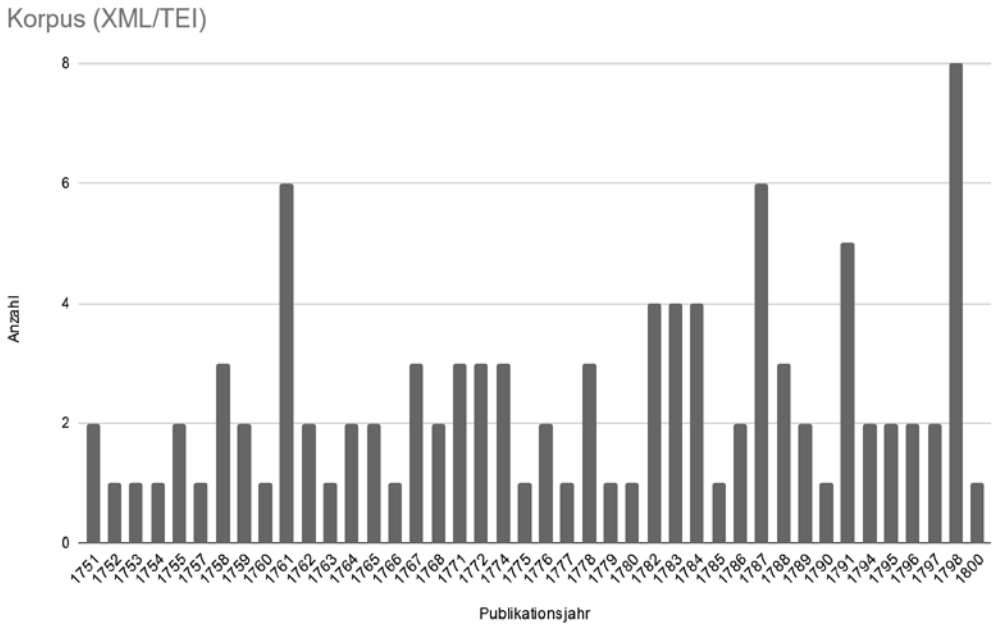
Beispiele für den erfolgreichen Einsatz der Methode Topic Modeling auf literarischen Texten sind Christof Schöchs Studie zu Topic Modeling auf französischen Dramen-Texten der Klassik und Aufklärung, die Topic Modeling Ergebnisse und Gattungen korreliert, oder auch Katherine Bodes Topic Modeling

Arbeiten auf fiktionalen Texten in australischen Zeitungen, in denen sie beispielsweise als zusätzlichen Parameter Gender analysiert (Bode 2018, 156-198; Schöch 2017).

Französische Texte aus dem 18. Jahrhundert haben Charakteristika, die in der Topic Modeling Pipeline durch entsprechende Preprocessing-Schritte berücksichtigt werden müssen. Forschungsprojekte wie ARTFL (American and French Research on the Treasury of the French Language) haben Topic Modeling bereits auf Texten des 18. Jahrhundert erprobt und die Artikel der *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers* mithilfe von Topic Modeling analysiert. Dabei unterstreichen sie vor allem die Möglichkeit, über Topic Modeling Diskurse abzubilden (Roe, Gladstone & Morrissey 2016). Elizabeth Andrews Bond und Robert M. Bond haben unlängst eine Studie zu Topic Modeling in vorrevolutionären französischen Presseartikeln vorgelegt (Bond & Bond 2020), wobei sie ihre Analyse mit dem stm-Package von R durchgeführt haben.

Im Kontext des Verbundprojekts „Mining and Modeling Text" (MiMoText), das vom *Trier Center for Digital Humanities* koordiniert wird, wurde Topic Modeling mit MALLET in Python auf das im Projekt entstehende Romankorpus „Collection de romans français du dix-huitième siècle (1750-1800)" in der Version 0.1.0 (kurz: roman18) angewendet. Die Topic Modeling-Pipeline basiert auf Schöch (Schöch 2020) und wurde an die Anforderungen des Projektes angepasst (Klee & Röttgermann 2020).

2. Datengrundlage: Romankorpus MiMoText



1 | MiMoText Romankorpus (XML/TEI): Werke pro Erstpublikationsdatum (Stand 13.10.2021)

Datengrundlage des Topic Modeling ist das roman18-Korpus an 80 französischen Romanen 1751-1800 (Röttgermann et al. 2020), das sich aus mehreren Quellen speist: eigene Volltextdigitalisierung per Double-Keying-Verfahren, eigene Volltext-digitalisierung mithilfe der Software OCR4all (Reul u. a. 2019), Wandlung von EPUB-Dateien aus den Quellen *Wikisource*[1], *Ebooks libres et gratuits*[2], *GoogleBooks*[3], *Frantext*[4] *und Rousseau Online*[5]. Alle Input-Dateien wurden in TEI-konformes XML nach den Richtlinien der *Text Encoding Initiative* (Burnard 2014) nach dem Schema der *European Literary Text Collection* (*ELTeC*) kodiert (Burnard & Odebrecht 2019). Mit Hilfe eines Python-Skripts[6], das historische Verbformen und Schaft-S in den Texten umwandelt, wurden die Texte teilmodernisiert, normalisiert und als Plaintext extrahiert.

Ein Vergleichshorizont der Daten ist zudem eine nahezu vollständige Dokumentation der literarischen Produktion französischer fiktionaler Werke 1751-1800 in Form einer Bibliographie (Martin, Mylne & Frautschi 1977), die uns dank wissenschaftlicher Vorarbeiten (Lüschow 2019) digitalisiert in Form eines RDF-Graphen vorliegt.

Aufgrund der statistischen Merkmale dieser bibliographischen Daten konnten wir uns hinsichtlich der Ausgewogenheit unseres Romankorpus den Proportionen in diesen Metadaten annähern und Merkmale wie Genderverteilung oder Erst-publikationsdaten approximativ im Romankorpus abbilden. Die Angabe zum Geschlecht der Autor:innen ist nicht explizit in den Metadaten enthalten, konnte jedoch über einen Abgleich mit VIAF ermittelt werden.

[1] *Wikisource* (ein Schwesterprojekt von Wikipedia) vereint Primärtexte in über 70 Sprachen. In einer auf Crowdsourcing basierenden Transkriptionsumgebung werden Faksimile und per Optical Character Recognition erkannter Text nebeneinander gestellt und von der Crowd korrigiert. Die Dateien durchlaufen verschiedene Qualitätsstufen. Wir haben uns dazu entschieden Texte aufzunehmen, die mindestens durch ein gelbes oder grünes Label ausgezeichnet sind, also vollständig transkribiert und von mindestens zwei verschiedenen Personen korrigiert wurden.

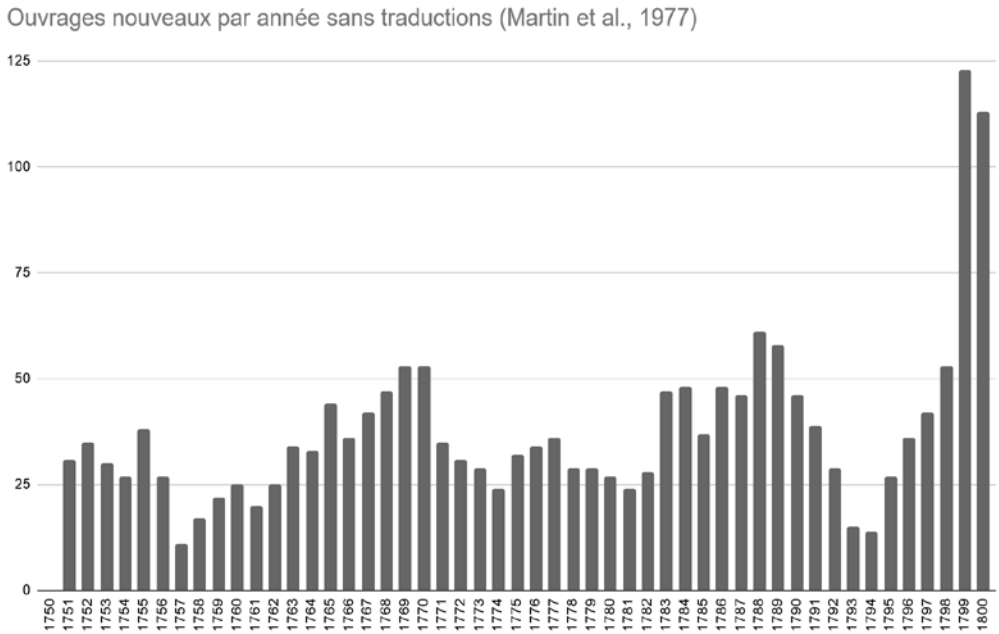
[2] Auf der Website *Ebooks libres et gratuits* konnten wir vor allem Werke kanonischer Autor:innen finden. Sie wurden als EPUB heruntergeladen und in TEI/XML konvertiert. Ein Nachteil der Plattform ist, dass sich die Provenienz der analogen Datenquelle (Print) leider nicht einsehen lässt.

[3] Die Datenqualität der französischen Romane, die wir im entsprechenden Publikationsdatum frei verfügbar als EPUB auf *GoogleBooks* finden konnten war sehr divers. Wir haben Werke nach der verfügbaren Qualität im Hinblick auf eine möglichst geringe OCR-Fehlerquote ausgewählt.

[4] *Frantext* enthält Primärtexte verschiedener Gattungen vom IX. bis XXI. Jahrhundert. Wir haben Texte verwendet, die den Kriterien Publikationsdatum 1750-1800, Gattung: "roman" und Lizenz "licence libre" entsprechen. Das auf Frantext vorliegende TEI/XML enthält eine Vielzahl an linguistischen Annotationen, die mit Hilfe eines Skripts entfernt wurden : <https://github.com/MiMoText/roman18/blob/master/Python-Scripts/transformation_frantext/Umwandlung_Frantext-Werke_in_TEI.py>, 28.01.2022.

[5] *rousseauonline.ch* bietet Zugang zu allen Werken von Jean-Jacques Rousseau (1712-1778) in ihrer ersten Referenzausgabe. Die Texte sind online zum Lesen zugänglich und stehen zum kostenlosen Download als PDF oder EPUB zur Verfügung. <https://www.rousseauonline.ch/>, 28.01.2022.

[6] <https://github.com/MiMoText/roman18/tree/master/Python-Scripts/modernization_and_transformation_to_plaintext>, 28.01.2022.



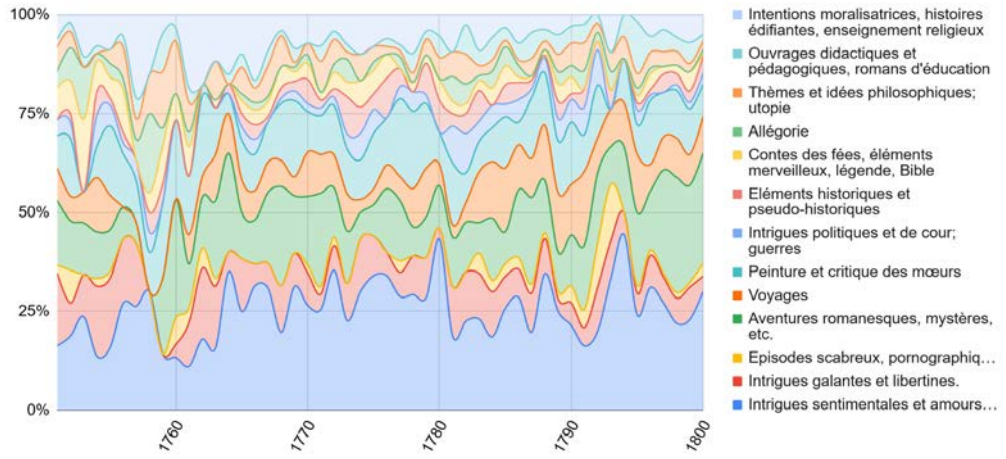
2 | Auswertung der Bibliographie du genre romanesque français 1751-1800 (Martin, Mylne & Frautschi 1977) hinsichtlich des Erstpublikationsdatums ohne Berücksichtigung der Übersetzungen.

Das Romankorpus wird fortlaufend um weitere erschlossene Quellen erweitert und wächst somit kontinuierlich. Der aktuelle Stand ist auf GitHub einsehbar. Die Grundlage des hier besprochenen Topic Modelings ist unter „Topic Model of roman18 corpus (November 2020)" auf Zenodo archiviert (Klee und Röttgermann 2020).

Die Besonderheit, dass die Grundgesamtheit[7] der literarischen Produktion der entsprechenden Dekaden innerhalb der Bibliographie (Martin, Mylne & Frautschi 1977) sorgfältig dokumentiert wurde und dass Schlagworte in Form thematischer Inhalte der Romane analysiert wurden, stellt für das Projekt einen reichen Datenschatz dar, der gerade im Vergleich mit den Ergebnissen des unsupervised Machine Learnings des Topic Modelings eine interessante Kontrastfolie bietet.

Das Korpus der bibliographischen Einträge 1751-1800 wurde von Seiten der Bibliograph:innen bezüglich der literarischen Themen und Handlung („thèmes et intrigues") in folgende Kategorien eingeteilt (s. Abb. 3).

[7] Zu möglicherweise fehlenden Werken cf. Dawson 1978, 497–508.



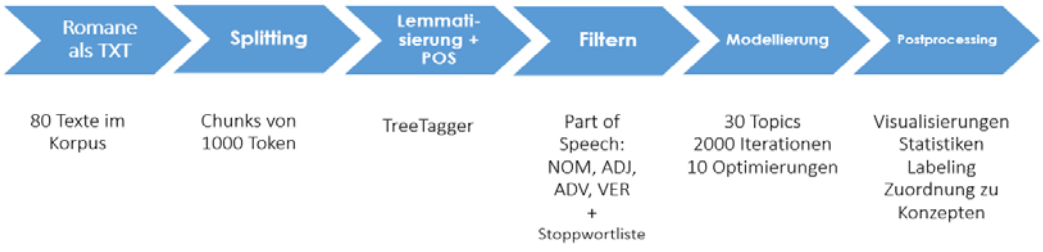
3 | Kategorien der Themen und Handlungselemente („thèmes et intrigues") des französischen Romans 1751-1800 im Zeitverlauf, Daten: Martin et al., 1977, S. xlviii.

Wir können in der Betrachtung der Themenkategorien im Zeitverlauf (Abb. 3) erkennen, dass religiöse und moralisierende Themen im französischen Roman der zweiten Hälfte des 18. Jahrhunderts eine abnehmende Tendenz aufweisen. Die Themen Erziehung („éducation") oder Reise („voyage") hingegen nehmen beispielsweise in den Dekaden ab 1780 prozentual zu.

Vor dem Hintergrund dieser 1977 publizierten Themenwerte für den französischen Roman der Zeit von 1751 bis 1800 in den bibliographischen Metadaten wollen wir nun betrachten, welche Topics im Gegenzug der Topic Modeling Algorithmus mit MALLET auf den Volltexten (Röttgermann et al. 2020) generiert.

3. Topic Modeling Workflow

Für die automatische Gewinnung von thematischen Aussagen aus dem MiMoText-Romankorpus wurde ein Topic Modeling-Workflow entwickelt, der mit Hilfe von Pythonskripten durchgeführt wird. Unsere Topic Modelling-Pipeline beinhaltet neben der Modellierung und verschiedenen Nachbearbeitungsschritten im Postprocessing zunächst einige Vorverarbeitungsschritte, die — unter der Berücksichtigung von Spezifika literarischer Prosatexte — dazu dienen, die Texte auf den Prozess des Topic Modelings vorzubereiten.



4 | Topic Modeling-Pipeline im Projekt „Mining and Modeling Text".

3.1 Input

Die Texte werden als Plaintext in die Pipeline eingespeist. Dabei sind diese bereits im Zuge der Korpuserstellung modernisiert und normiert worden. Dies beinhaltet die Anpassung historischer Wort- und Flexionsformen an die moderne Sprache sowie die Ersetzung historischer Schriftzeichen wie beispielsweise das Schaft-S.[8]

Da der Topic Modeling-Algorithmus Wörter in ihrem Kontext betrachtet, sollte dieser für die eingespeisten Texte nicht zu groß ausfallen und in einer einheitlichen Größe vorliegen. Die Romane stellen jedoch eine vergleichsweise umfangreiche Textsorte dar. Zusätzlich variieren die verschiedenen Texte im Korpus stark in ihrem Umfang. Aus diesem Grund werden die Inputdateien zu Beginn in Textstücke mit einer Wortlänge von 1000 Token gesplittet.

Den Textdateien beigefügt wird eine Metadatentabelle[9], welche später im Post-processingschritt bei der statistischen Auswertung und Visualisierung der Topic Modeling-Ergebnisse verwendet wird. Diese enthält Informationskategorien wie das Autorengeschlecht, das Publikationsjahr und die Erzählform.

3.2 Preprocessing: POS-Tagging und Filtern

Vor der eigentlichen Modellbildung werden die eingespeisten Textdateien mit üblichen Preprocessingschritten vorverarbeitet. In einem ersten Schritt werden die Texte lemmatisiert. Das bedeutet, die einzelnen Wortformen werden auf ihre Grundform zurückgeführt, sodass flektierte Formen eines Lemmas bei der Modellierung als gleiche Wörter behandelt werden. Für diese Wortformen wird zusätzlich ein *part of speech*-Tagging durchgeführt, bei dem sie mit ihrer Wortart annotiert werden. Dazu wird das Tool TreeTagger (Schmid 1994) verwendet, das Modelle für die Anwendung auf über 40 verschiedene Sprachen zur Verfügung stellt. Für die Wahl des geeigneten Modells wurden zwei Ausführungen erprobt: das von TreeTagger bereitgestellte Modell für die moderne, französische Sprache sowie das Presto-Modell (PRESTO 2014)[10], welches speziell für das Französisch des 16. und 17. Jahrhunderts auf der Grundlage historischer Texte trainiert wurde (Souvay & Pierrel 2009). Eine exemplarische Überprüfung der Tagging-Ergebnisse hat gezeigt, dass Presto zwar bessere Ergebnisse bei der Erkennung von historischen Wortformen liefert, sich jedoch im Vergleich zum TreeTagger-Modell durch eine deutlich unzuverlässigere POS-Klassifizierung auszeichnet (cf. Abb. 5). Zudem weist das Modell für die moderne Sprache eine bessere Named-Entity-Recognition auf, was beim Herausfiltern von Eigennamen eine wichtige Rolle spielt. Insgesamt überwiegen die Stärken des vom TreeTagger bereitgestellten Modells, weshalb dieses für die Vorverarbeitung in der Topic Modeling-Pipeline verwendet

[8] Dies erfolgt mit Hilfe eines Pythonskriptes und einer Modernisierungsliste (<https://github.com/MiMoText/roman18/tree/master/Python-Scripts/modernization_and _transformation_to_plaintext>). Für die Liste wurden mit einem Spellcheck und der Pythonbibliothek enchant (<http://pythonhosted.org/pyenchant/>) die im Korpus vorkommenden historischen Wortformen ermittelt.

[9] <https://github.com/MiMoText/roman18/blob/master/XML-TEI/xml-tei_metadata.tsv>, 28.01.2022.

[10] <http://presto.ens-lyon.fr/>.

wird. Seine Schwäche bei der Erkennung historischer Sprachformen wird durch den Modernisierungsschritt im Vorfeld aufgefangen.

	TreeTagger fr-Modell		Presto	
	POS	Lemma	POS	Lemma
étoit	Substantiv	étoit	Konjugiertes Verb (être & avoir)	être
errois	Adjektiv	errois	Konjugiertes Verb	errer
sçavois	Verb im Subjonctiv imperfect	sçavois	Konjugiertes Verb	savoir
connoissance	Substantiv	connoissance	Substantiv	connaissance
quinze	Numeral	quinze	Substantiv	quinze
vingt-quatre	Numeral	vingt-quatre	Adjektiv	vingt-quatre
dernier	Adjektiv	dernier	Substantiv	dernier

5 | TreeTagger und Presto in einem exemplarischen Vergleich. Die erste Spalte beinhaltet Wortformen, wie sie im Romantext auftreten, die Spalte „POS" beinhaltet die Wortklasse, die das Modell der jeweiligen Form zuordnet, und „Lemma" die Grundform, auf die die Wortform zurückgeführt wird.

Es konnte nachgewiesen werden, dass durch die Beschränkung auf ausgewählte Wortarten kohärentere Topics erzielt werden (cf. Uglanova & Gius 2020). Um semantisch aussagekräftige Topics zu generieren, werden die Texte deshalb im Vorfeld nach Wortarten gefiltert. In die Modellierung eingespeist werden nur Lemmata der Wortarten Substantiv, Adjektiv, Adverb und Verb.

Mit Hilfe zweier Stoppwortlisten werden zusätzlich sowohl Hilfsverben als auch die für die Textsorte Roman charakteristischen Figurennamen, die beim POS-Tagging nicht als Eigennamen erkannt wurden, herausgefiltert.[11] Beide Kategorien von Wörtern kommen in den Texten sehr häufig vor, haben aber wenig bis keinen semantischen Gehalt und tragen damit also nicht zur Gewinnung thematischer Muster bei.

3.3 Modellierung

Für die Durchführung des Modellierungsschrittes wurden zwei verschiedene Varianten getestet: die Pythonbibliothek Gensim (Rehurek & Sojka 2010)[12] und das Java-basierte Tool MALLET (McCallum 2002)[13].

[11] Die in den Texten vorkommenden Figurennamen wurden mit Hilfe von Named Entity Recognition mit SpaCy ermittelt.
[12] <https://pypi.org/project/gensim/>.
[13] <http://mallet.cs.umass.edu/topics.php>.

	Gensim	MALLET
0	cœur	cœur
1	voir	amour
2	point	point
3	aimer	aimer
4	amour	jamais
5	ami	voir
6	croire	rendre
7	tout	âme
8	jamais	sentiment
9	sentiment	bonheur
10	moins	ami

6 |Vergleich zwischen Gensim und MALLET bezogen auf die Verteilung der zehn wichtigsten Topicwörter des Topics amour_sentiment. Die Topics stammen aus vergleichbaren Modellen, die für 10 Topics mit 500 Iterationen vorgenommen wurden.

Da die Bibliothek von Gensim für die Anwendung auf sehr großen Textkorpora entwickelt wurde[14], war im Vorfeld zu erwarten, dass in unserem Fall eines vergleichsweise kleinen Korpus das Tool MALLET besser geeignet ist. Diese Annahme konnte im Abgleich der berechneten Modelle bestätigt werden. Unsere Untersuchungen haben gezeigt, dass die mit MALLET erstellten Modelle konsistentere — das heißt für die Ableitung von thematischen Aussagen besser interpretierbare — Topics[15] geliefert haben. Illustrieren lässt sich dies am Beispiel des Topics amour_sentiment. Wie in Abb. 6 zu sehen ist, finden sich in dem mit MALLET erstellten Topic unter den zehn wichtigsten Wörtern deutlich mehr solcher Wörter, die semantisch dem Themenfeld Liebe/Gefühl zuzuordnen sind.

Durch den Wrapper LdaMallet von Gensim[16] kann die Modellierung mit MALLET in die Python-Pipeline eingebunden werden. Nach Tests mit verschiedenen Modellgrößen wurde ein Topic Model mit 30 Topics trainiert. Diese Anzahl bietet bei der Korpusgröße von rund 80 Texten ein ausgewogenes Topic-Spektrum. Eine höhere Zahl Topics vergrößert zwar die Menge unterschiedlicher Topics, unter diesen finden sich jedoch in größerer Zahl generische Topics und es kommt vermehrt zu semantischen Überschneidungen. Des Weiteren handelt es sich zunehmend um sehr spezifische Topics, welche nur in einzelnen Werken vorkommen und welche wir ohnehin bei der Einspeisung in unser Wissensnetzwerk ignorieren. Eine kleinere Anzahl führt zu mehrdeutigen Topics und würde verhindern, dass manche Wortkookurrenzen überhaupt aufgedeckt werden.

[14] Cf. Rehurek & Sojka 2010, die Gensim auf ein Korpus von 61.293 Volltexten und insgesamt über 270 Mio. Token anwenden.

[15] Zu einem ähnlichen Ergebnis kommt auch Dipanjan Sarkar, der ein Korpus an wissenschaftlichen Artikeln mit Topic Modeling analysiert und die coherence scores von MALLET und Gensim miteinander vergleicht: "You can clearly see that the model from MALLET is much better (...) as compared to the default LDA model from Gensim." (Sarkar 2019, 402)

[16] <https://radimrehurek.com/gensim_3.8.3/models/wrappers/ldamallet.html>.

Die Anzahl der Iterationen und Optimierungen beim Machine Learning-Prozess sind zwei weitere festzulegende Parameter. Umso mehr Iterationen durchgeführt werden, desto stärker verlängert sich die Laufzeit, desto mehr steigert sich jedoch die Qualität des resultierenden Topic Models. Für das vorliegende Korpus wurden 2000 Iterationen und 10 Optimierungen für den Trainingsprozess als geeignete Größe ermittelt. Die Optimierungen führen zu einer differenzierteren Wahrscheinlichkeitsverteilung der Topics im berechneten Modell.[17]

3.4 Postprocessing

Das Postprocessing umfasst die Erzeugung verschiedener Statistiken und Visualisierungen der Topic Modeling-Ergebnisse, die der Analyse und Extraktion der thematischen Statements dienen, aber auch selbst den Nutzer:innen des Wissensnetzwerkes bereitgestellt werden.

Das berechnete Topic Model besteht aus einer zuvor definierten Anzahl von Topics, die aus einer Wahrscheinlichkeitsverteilung der eingespeisten Wörter bestehen, sowie einer Wahrscheinlichkeitsverteilung dieser Topics für jedes Textdokument des Korpus. Bestehend aus diesen Informationen werden verschiedene CSV-Dateien[18] erstellt. Auf der Basis der wahrscheinlichsten Wörter wird jedem Topic ein Label zugewiesen (vgl. 3.5). Zusammen mit dieser Information werden aus der Verteilung der Top-Topics pro eingespeistes Werk schließlich Themenaussagen abgeleitet. Dabei berücksichtigen wir die fünf wahrscheinlichsten Topics für jeden Roman, bei vorheriger Aussortierung aller Topics, die in weniger als 10 % und in mehr als 80% der Korpuswerke enthalten sind.[19] Es werden dadurch einerseits sehr seltene, zum Teil werkspezifische, und andererseits sehr häufige, in der Regel generische, Topics ausgeschlossen, da sie für einen werkübergreifenden Themenabgleich keinen Gewinn bringen. Es verbleiben damit 25 Topics, die bei der Generierung der Themenaussagen einbezogen werden.

Aufgrund der in der französischen Literatur des 18. Jahrhunderts vorherrschenden Themen ähneln sich viele Werke in Bezug auf die Top-Topics. Um dennoch Aussagen darüber treffen zu können, wie sich die einzelnen Werke thematisch voneinander unterscheiden, ist es hilfreich, die distinktiven Topics pro Werk zu ermitteln. Ein Topic ist distinktiv für ein Werk, wenn es in diesem überrepräsentiert

[17] Dazu passt der Algorithmus interne Hyperparameter fortlaufend so an, dass sowohl die Topics als auch die Wörter ihrer Zusammensetzung im resultierenden Modell stärker unterschiedlich gewichtet werden (cf. Steyvers & Griffiths 2007).

[18] *topicwords.csv* enthält die 50 Top-Wörter für jedes Topic, *wordprobs.csv* zeigt für jedes Lemma im Vokabular den Score für jedes Topic. *doc-topic-matrix.csv* enthält für jedes Textchunk eine Verteilung der Topics, welche in *chunkmatrix.csv* um relevante Metadaten pro Textstück ergänzt ist. Die Topicwahrscheinlichkeiten bezogen auf die Romane als Gesamttexte sind in der *mastermatrix.csv* zusammengefasst. *topicranking.csv* listet für jedes Werk die zehn wahrscheinlichsten Topics mit ihren Wahrscheinlichkeitswerten.

[19] Hier ist anzumerken, dass im Grunde jedes Topic in jedem Werk vorhanden ist. Es tritt jedoch erst ab einer bestimmten Wahrscheinlichkeit signifikant in Erscheinung, ab der wir vereinfacht davon sprechen, dass es in einem Werk vorkommt. Der Schwellwert ist von der Korpusgröße und Topicanzahl abhängig. Für das hier beschriebene Topic Model haben wir eine Wahrscheinlichkeit von 0.03 als Schwellwert angewandt. Mithilfe von diesem lässt sich berechnen, in wieviel Prozent der Texte jedes Topic vorkommt.

ist, also im Vergleich zum Gesamtkorpus eine überdurchschnittliche Wahrscheinlichkeit aufweist. Demnach definieren wir seltene Topics, die nur in wenigen Romanen des Korpus vorkommen und dadurch eine geringe Wahrscheinlichkeit bezogen auf das Gesamtkorpus aufweisen, für ein Einzelwerk distinktiv, sofern sie dort unter den Top-Topics auftreten.

MiMoText-ID	Top 5-Topics	Distinktive Topics = Seltene in Top 5
Senac_Emigre	Correspondance, Bonheur, Attraction_Personnalité, [temps paraître encore], Amour_Sentiment	
Maistre_ Voyage	Bonheur, Art, Nuit, Mort, Nature	Art
Sade_Aline	Philosophie, Interaction, Amour_Sentiment, Deuil, Migration_Voyage	
Sade_Justine	Terreur, Philospophie, Mort, Nuit, Interaction	
Bernadin_Paul	Nature, Bonheur, Art, Mort, Famille	Art
Laclos_Liaisons	Correspondance, Interaction, Amour_Sentiment, Bonheur, [temps paraître encore]	
Retif_Paysanne	Famille, Correspondance, Nuit, Interaction, Deuil	
Mercier_An	Philosophie, Monarchie, Art, Mort, Richesse	Monarchie, Art
Retif_AntiJustine	Nuit, Famille, [temps paraître encore], Relation amoureuse, Interaction	Relation amoureuse
Rousseau_Julie	[point voir moins], Amour_Sentiment, Éducation_Enfance, Bonheur, Deuil	
Voltaire_Candide	Alimentation_Sociabilité, Migration_Voyage, [point voir jour], Richesse, Deuil	Alimentation_ Sociabilité

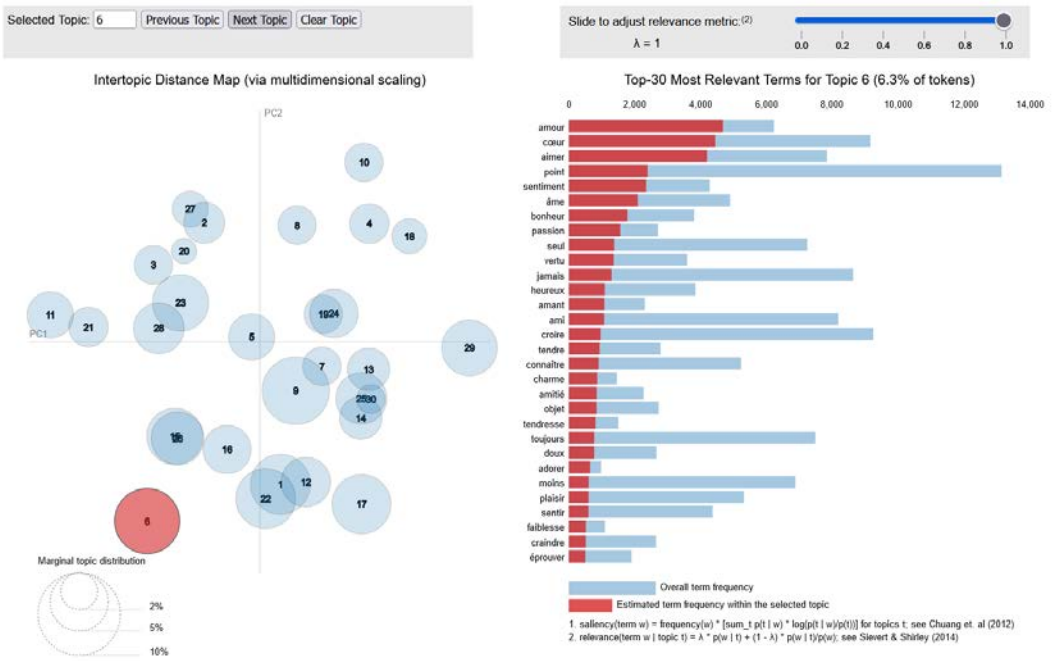
7 | Verteilung der distinktiven Topics bei den Pilotwerken (Ausschnitt aus dem gesamten Romankorpus). Topics, die keinem der Einträge des thematischen Vokabulars zugeordnet werden können, sind mit ihren drei wichtigsten Wörtern in eckigen Klammern gelabelt (vgl. Kapitel 3.5).

Mithilfe der Verteilung der Topics im Korpus können die seltenen Topics ermittelt werden. Als selten wurden hier Topics definiert, die nur in maximal 20% aller Romane im Korpus vorkommen. Die Schnittmenge mit den Top 5 Topics pro Werk zeigt schließlich, durch welche prävalenten Topics sich ein Werk von den übrigen abhebt.[20]

Auf der Grundlage der Statistiken werden zur Veranschaulichung der Topic Modeling-Ergebnisse verschiedene Visualisierungen erstellt.

[20] Zur Ermittlung von distinktiven Topics sind außerdem noch andere statistische Verfahren denkbar. Cf. dazu das Projekt "Zeta und Konsorten", welches verschiedene statistische Distinktivitätsmaße erforscht (cf. Schöch 2018, Du et al. 2021).

Mithilfe der Pythonbibliothek PyLDAVis[21] lässt sich eine interaktive HTML-Visualisierung[22] des Topic Models erstellen, mit welcher die Nutzer:innen die Verteilung und Zusammensetzung der Topics explorieren können.[23]



8 | Topic Modeling-Visualisierung in PyLDAVis. Auf der linken Seite sind die 30 Topics als Kreise zu sehen. Ihre Größe symbolisiert ihre Gewichtung im Korpus, ihre Lage zueinander bildet die Ähnlichkeit ihrer Wortverteilungen ab. Die Auswahl eines Topics links ermöglicht eine genauere Exploration seiner Wortverteilung auf der rechten Seite der interaktiven Visualisierung.

Einen Überblick über die häufigsten Wörter in jedem Topic bieten die mit der Pythonbibliothek wordcloud[24] erstellten Wortwolken. Die unterschiedlich großen Wahrscheinlichkeitswerte werden hier durch die Größe der Wörter abgebildet.

[21] <https://pyldavis.readthedocs.io/en/latest/#> [28.01.2022].

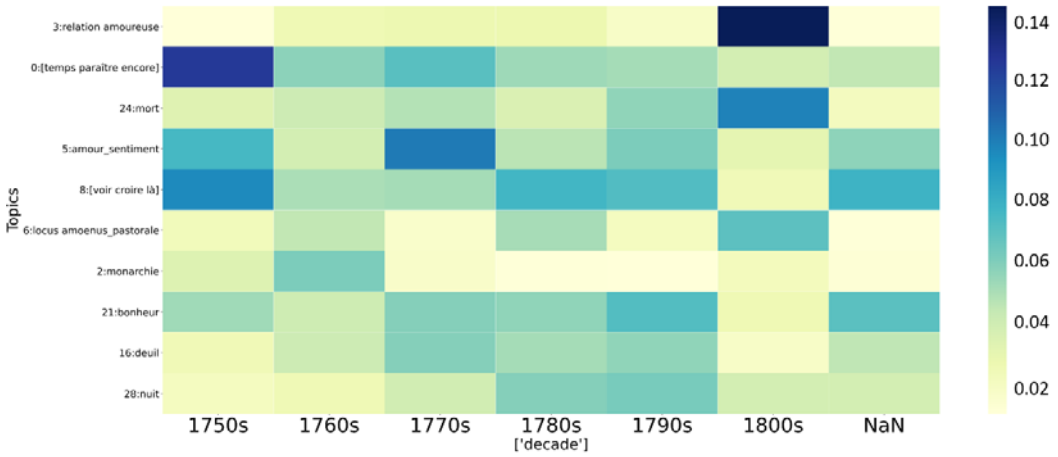
[22] Cf. die Visualisierung zum hier behandelten Topic Model: <https://github.com/MiMoText/mmt_2020-11-19_11-38/blob/main/results/mmt_2020-11-19_11-38/visualization.html> [21.10.21]. Um die Visualisierung wie in Abb. 8 anschauen zu können, ist es erforderlich, die verlinkte HTML-Datei zunächst lokal abzuspeichern und dann im Browser zu öffnen.

[23] Da PyLDAVis mit dem Output der Topic Modeling-Bibliothek von Gensim arbeitet, müssen die von MALLET erstellten Ergebnisse zunächst in das entsprechende Format transformiert werden: <https://github.com/MiMoText/topicmodeling/blob/master/scripts/make_overview.py> [28.01.2022].

[24] <https://amueller.github.io/word_cloud/> [28.01.2022].

9 | Wordcloud zum Topic „amour_sentiment".

Mithilfe der Bibliothek seaborn werden sogenannte Heatmaps zu einem bestimmten Metadaten-Parameter erstellt. Aktuell ist es möglich, Heatmaps zur Topicverteilung bezüglich der TextID, des Autor:innennamens, des Autor:innen-geschlechts, der Dekade oder der Erzählform zu generieren. Die Auswahl wird über die Parametereinstellungen im Script gesteuert.



10 | Heatmap mit der Verteilung von Topics im Romankorpus, dominante Topics pro Dekade.[25]

3.5 Labeling mit Hilfe eines thematischen Vokabulars

Um die aus dem Topic Modeling gewonnen Aussagen in das Wissensnetz einzuspeisen und mit den thematischen Informationen aus der Sekundärliteratur und der Bibliographie vergleichbar zu machen, ist es notwendig, die Topics zu labeln und zu normalisieren. Hierfür dient uns ein auf das Projekt zugeschnittenes kontrolliertes Themenvokabular. Bei der Erarbeitung waren bestimmte Eigenschaften von besonderer Wichtigkeit: Die Begriffe sollten die Themenkonzepte der französischen Aufklärung abdecken, ein gewisses Abstraktionslevel aufweisen, damit sie als kategorische Begriffe fungieren können, und ihre Zusammenstellung

[25] Die hier verwendete Anzahl an Romanen (80) der Version 0.1.0 des roman18-Korpus ist noch nicht groß genug, um robuste Aussagen hinsichtlich der Topics pro Dekade zu treffen. Wir sind dabei das Romankorpus im Laufe des Jahres 2022 auf bis zu 200 Volltexte zu erweitern. Die bibliographischen Metadaten enthalten derzeit ca. 2000 items, die in ca. 30.000 RDF-Tripeln resultieren und somit robustere Aussagen erlauben.

sollte transparent und nachvollziehbar sein. Eine erste Grundlage bildet das Themeninventar des *Dictionnaire européen des Lumières* (Delon 1997). Die Artikelstichwörter bieten eine gute Abdeckung an gesellschaftlich, politisch, ideengeschichtlich oder kulturell relevanten Themen der Epoche und stellen somit einen geeigneten Grundstock an möglichen Labeln für die in den Romanen vorkommenden Themen. Dennoch enthält die Ressource Begriffe, die entweder zu spezifisch (z.B. „pyrrhonisme") oder zu generisch (z.B. „fonction") sind, um durch sie literarische Themen zu beschreiben, weshalb diese für das Vokabular nicht berücksichtigt wurden. Ergänzt wurden die Begriffe um fehlende Konzepte zum Labeln der Topics, um thematische Schlagworte aus der Bibliographie (cf. Martin, Mylne & Frautschi 1977) sowie um Themenkonzepte, die in der Sekundärliteratur erwähnt werden, wenn diese anderweitig nicht repräsentiert waren. Das Vokabular ist nun konsolidiert, kann aber auch in Zukunft bei Bedarf erweitert werden.

Um die multilinguale Vergleichbarkeit zwischen französischsprachigen Primärtexten und deutschsprachiger Sekundärliteratur zu gewährleisten, und im Sinne der Anschlussfähigkeit an und Interoperabilität mit anderen Datenbeständen, werden die Themenkonzepte auf einen Normdatensatz (Wikidata) gemappt, wodurch das kontrollierte Vokabular konsolidiert und multilingual erfasst ist (siehe Abb. 11).[26]

théologie	theology	Theologie	Q34178	https://www.wikidata.org/wiki/Q34178	DEL
tolérance	toleration	Toleranz	Q183225	https://www.wikidata.org/wiki/Q183225	DEL
tradition	tradition	Tradition	Q82821	https://www.wikidata.org/wiki/Q82821	DEL
traduction	translation	Übersetzung	Q7553	https://www.wikidata.org/wiki/Q7553	DEL
tragédie	tragedy	Tragödie	Q80930	https://www.wikidata.org/wiki/Q80930	DEL
transport	transport	Transport	Q7590	https://www.wikidata.org/wiki/Q7590	DEL
travail	work	Arbeit	Q6958747	https://www.wikidata.org/wiki/Q6958747	DEL
troubadour	troubadour	Troubadour	Q186370	https://www.wikidata.org/wiki/Q186370	BGRF
tyrannie	tyranny	Tyrannei	Q22082330	https://www.wikidata.org/wiki/Q22082330	Seklit
universalisme	universalism	Universalismus	Q875797	https://www.wikidata.org/wiki/Q875797	Seklit
urbanisme	urbanism	Urbanistik	Q59950	https://www.wikidata.org/wiki/Q59950	DEL
utilitarisme	utilitarianism	Utilitarismus	Q160590	https://www.wikidata.org/wiki/Q160590	DEL
utopie	utopia	Utopie	Q131156	https://www.wikidata.org/wiki/Q131156	DEL
valeur	value	Wertvorstellung	Q194112	https://www.wikidata.org/wiki/Q194112	Seklit
vanité	vanity	Eitelkeit	Q1321250	https://www.wikidata.org/wiki/Q1321250	Seklit
vengeance	revenge	Rache	Q1712140	https://www.wikidata.org/wiki/Q1712140	Seklit
vérité	truth	Wahrheit	Q7949	https://www.wikidata.org/wiki/Q7949	Seklit
vertu	virtue	Tugend	Q157811	https://www.wikidata.org/wiki/Q157811	DEL
vie	life	Leben	Q3	https://www.wikidata.org/wiki/Q3	Seklit

11 |Ausschnitt aus dem kontrollierten Vokabular zur Extraktion thematischer Konzepte. „DEL" bezeichnet dabei Ressourcen aus dem *Dictionnaire européen des Lumières* (Delon 1997), „BGRF" Ressourcen aus den thematischen Schlagworten der *Bibliographie du genre romanesque français 1751-1800* (Martin, Mylne & Frautschi 1977) und „Seklit" Ressourcen aus der im Projekt ausgewerteten Sekundärliteratur.

Dieses aktuell 368 Einträge umfassende Vokabular der Themenbegriffe liefert die Konzept-Items für die Objektposition aller thematischen Statements. Zur Erstellung der Topic Modeling-Statements wird folglich jedem der Topics ein Label aus diesem Inventar zugewiesen. Dazu wurde zunächst ein automatischer Ansatz mit Einsatz von Word Embeddings (Mikolov u. a. 2013) und der Ermittlung von Topic-Zentroiden erprobt. Notwendig dafür ist ein Word Embedding Modell, welches auf

[26] Zur Dokumentation der Liste: <https://github.com/MiMoText/vocabularies>, letzter Zugriff 28.01.2022.

einer ausreichend großen Menge französischer Texte trainiert wurde. Hierfür finden sich bereits vortrainierte Modelle, die auf frei verfügbaren Texten im Web wie den Texten der französischen Wikipedia oder Nachrichtenkorpora basieren (cf. Fauconnier 2015) und für den Versuch des Labelings nachgenutzt werden können. Mithilfe des Modells wird für jedes Topic ein Vektor berechnet, der Topiczentroid, welcher sich aus dem Durchschnitt der Vektoren zu den wahrscheinlichsten Topicwörtern (z.B. die Top-20 Wörter) ergibt. Auch für jedes der Labelkandidaten, d.h. den Wörtern aus dem vorher definierten Labelvokabular, kann ein Vektor bestimmt werden. Über einen Distanzabgleich der Topicvektoren und Labelvektoren kann nun für jedes Topic ein zugehöriges Label ermittelt werden: Die Paarungen mit den geringsten Distanzen zeigen potentiell geeignete Label.

Ein weiterer Ansatz besteht in der Abbildung der Top-Topicwörter auf einen gemeinsamen generischen Begriff unter Einsatz eines lexikalisch-semantischen Netzes wie WordNet (<https://wordnet.princeton.edu/>). Mit dem automatischen Ansatz sind verschiedene Schwierigkeiten und Probleme verbunden: Klassische Word Embedding-Modelle enthalten nur Vektoren für Einzelwörter. Somit können mit dem dargelegten Verfahren keine Mehrwortbegriffe als Label zugeordnet werden. Die vortrainierten Modelle enthalten darüber hinaus in der Regel keine sehr speziellen Begriffe, die allerdings für die französische Literatur des 18. Jahrhunderts relevant sein könnten. Aus diesen Gründen wäre das Trainieren eines eigenen Modells anzudenken, welches einerseits Mehrwort-Lexeme beinhaltet und andererseits auch Texte zu Themen der Literatur der Aufklärung berücksichtigt. Allerdings stellte sich der Word Embedding-Ansatz noch aus einem anderen Grund als weniger gut geeignet heraus: Bei den Wörtern mit der geringsten Distanz zu den jeweiligen Topiczentroiden handelt es sich häufig um sehr spezifische Begriffe, die weniger gut geeignet sind, um ein Topic in seiner Ganzheit zu labeln. Beispielsweise würde „livre" als Label für ein Topic ausgewählt werden, das nicht rein literarische Aspekte enthält, sondern auch weitere Bereiche der Kunst wie die Musik und die Malerei abdeckt. Ähnlich gestaltet es sich beim WordNet-Ansatz schwierig, einen Oberbegriff für die Topicwörter mit einem geeigneten Abstraktionsniveau zu finden.

Aus diesen Gründen fiel die Entscheidung für eine manuelle Vergabe der Label. Die Zusammensetzung der Topicwörter einiger Topics lässt die Zuweisung eines eindeutigen Labels nicht zu, da hier verschiedene Konzepte repräsentiert sind. In diesen Fällen werden Doppellabel vergeben wie beim Topic „alimentation_sociabilité" in Abb. 12: Wörter wie „table", „manger", „boire" und „vin" deuten auf das Thema der Ernährung hin, daneben die vorkommenden Personen und Lebewesen „homme", „femme", „maître", „hôte" und „chien" auf eine gesellige Zusammenkunft. Das Topic wird daher durch die Verknüpfung zweier Begriffe des thematischen Vokabulars repräsentiert.

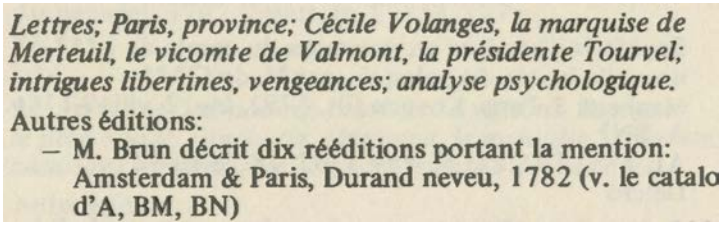
12 | Wordle des Topics „alimentation_sociabilité".

Da manche der Topics sehr generisch sind und nur eine geringe semantische Kohärenz aufweisen, ist es nicht immer möglich, ein passendes Label zu vergeben. In diesen Fällen werden die ersten drei Topicwörter als Label vergeben wie zum Beispiel „[point voir jour]".

4. Auswertung der Ergebnisse des Topic Modelings anhand von zwei Fallbeispielen

4.1 *Les Liaisons Dangereuses* von Choderlos de Laclos (1782)

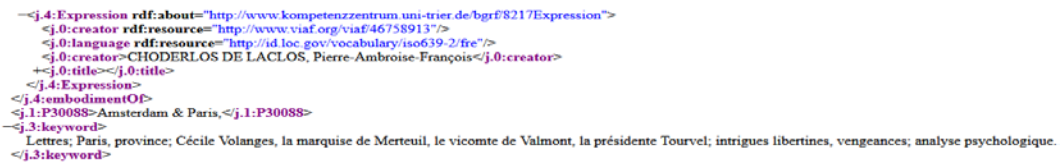
Anhand mehrerer Fallbeispiele seien hier die Ergebnisse des Topic Modelings analysiert und illustriert. Wir wählen als erstes Fallbeispiel ein Schlüsselwerk des 18. Jahrhunderts aus, um zu vergleichen, wie sich die Informationsextraktion aus der Bibliographie und aus dem Topic Modeling des Primärtextes hinsichtlich der extrahierten Themen verhalten: *Les Liaisons Dangereuses* von Choderlos de Laclos aus dem Jahr 1782. In der *Bibliographie du genre romanesque français 1751-1800* (Martin, Mylne & Frautschi 1977) werden von Seiten der Bibliograph:innen folgende Spezifikationen hinsichtlich der literarischen Themen des Werks vorgenommen:


Lettres; Paris, province; Cécile Volanges, la marquise de Merteuil, le vicomte de Valmont, la présidente Tourvel; intrigues libertines, vengeances; analyse psychologique.

Autres éditions:
– M. Brun décrit dix rééditions portant la mention: Amsterdam & Paris, Durand neveu, 1782 (v. le catalo d'A, BM, BN)

13 |Die Schlagworte zur Gattung, Handlungsort, Protagonisten und zu den Themen des Romans *Les Liaisons Dangereuses*, Screenshot aus der *Bibliographie du genre romanesque français 1751-1800* (Martin, Mylne & Frautschi 1977, 246).

Die Einträge der *Bibliographie du genre romanesque français 1751-1800* (Martin et al., 1977) wurden in einen RDF-Graphen überführt (Lüschow 2019), als Beispiel hier der Eintrag zu dem analysierten Werk:



14 | Der bibliographische Eintrag zu *Les Liaisons Dangereuses*, modelliert in RDF (Lüschow 2019)

Ausgehend von den Keywords innerhalb des RDF-Graphen wurden diese extrahiert und mit einem Python-Skript sortiert, mit dem Ziel die Einteilung der Bibliographen in die Kategorien Erzählform (hier „Lettres"), Handlungsort (hier: „Paris, province"), Protagonisten (hier „Cécile Volanges, la marquise de Merteuil, le vicomte de Valmont") Schlagworte der Handlung (hier: „intrigues libertines, vengeances") und Stil/Haltung/Tonalität (hier: „analyse psychologique") nachzuvollziehen.

ID	author	Label	narrative perspective_ string	Narrative location_string	characters_ string	plot_theme	style_attitude_ tonality
82.17	CHODERLOS DE LACLOS, Pierre-Ambroise-François	Les liaisons dangereuses ou lettres recueillies dans une société & publiées pour l'instruction de quelques autres, par M. C.....de L . . .	Lettres	Paris, province	Cécile Volanges, la marquise de Merteuil, le vicomte de Valmont, la présidente Tourvel	intrigues libertines, vengeances	analyse psychologique

Tab 1: Ergebnis der automatischen Keyword-Sortierung[27] für den Eintrag zu *Les Liaisons Dangereuses* (Laclos, 1782), die Metadaten stammen aus dem Eintrag 82.17 (Martin, Mylne & Frautschi 1977, 246).

Anhand dieser Kategorisierung innerhalb der XML-Dateien lassen sich die thematischen Schlagworte extrahieren. Relevant ist das Keyword-Feld „Themen und Handlung".

Nach einer Vereinheitlichung der Keywords mittels des kontrollierten Vokabulars[28] ergibt sich damit aus den bibliographischen Daten folgende Aussage zu literarischen Themen des Werks, hier formuliert in der Struktur eines RDF-Triples (Resource Description Framework Triples):

[27] Cf. <https://github.com/MiMoText/KeywordExtractor>.
[28] Cf. <https://github.com/MiMoText/vocabularies>.

Laclos_Liaisons	about	libertinage	http://zora.uni-trier.de:11000/wiki/Item:Q1
Laclos_Liaisons	about	vengeance	http://zora.uni-trier.de:11000/wiki/Item:Q1

Wie verhalten sich die extrahierten Aussagen aus dem Topic Modeling der Primärtexte dazu? Folgende Topics sind prävalent in *Les Liaisons Dangereuses*:

Laclos_Liaisons	about	correspondance	https://doi.org/10.5281/zenodo.4493224
Laclos_Liaisons	about	interaction	https://doi.org/10.5281/zenodo.4493224
Laclos_Liaisons	about	amour_sentiment	https://doi.org/10.5281/zenodo.4493224
Laclos_Liaisons	about	bonheur	https://doi.org/10.5281/zenodo.4493224
Laclos_Liaisons	about	[temps paraître encore]	https://doi.org/10.5281/zenodo.4493224

Das Topic „correspondance" umfasst Topicwörter wie Brief („lettre"), schreiben („écrire"), Antwort („réponse"), empfangen („recevoir") etc. und enthält einen Hinweis auf die Gattung von *Les Liaisons Dangereuses*: Es handelt sich um einen Briefroman. Die Bibliograph:innen der *Bibliographie du genre romanesque français 1751-1800* haben dies ebenfalls mit der Gattungszuschreibung „Lettres" analysiert (Martin, Mylne & Frautschi 1977, 246).

Topic „correspondance" und Topic „interaction"

15 |Wordles der Topics „correspondance" und „interaction", <https://doi.org/10.5281/zenodo.4493224>.

Das Topic, das sich im Ranking der Gewichtungen als nächstes für Choderlos Laclos' Werk als Ergebnis zeigt, ist das Topic „interaction". Es umfasst Topicwörter wie verlangen („demander"), verweigern („refuser"), geben („donner"), Mittel („moyen"), vorschlagen („proposer") etc. Das Label interaction fasst dieses semantische Cluster zusammen.

Es verweist auf Interaktionen und Prozesse des Aushandelns, die im Kontext des Romans auf psychologische Verhandlungen hindeuten. Die Bibliograph:innen

haben dies mit den Schlagworten Intrigen („intrigues") und Rache („vengeance") terminologisch stärker auf Plotmuster hin formuliert. Hier zeigt sich eine Eigenschaft des Topic Modeling Algorithmus: Er bewegt sich gewissermaßen an der Oberfläche der Wörter und bildet Verteilungen und Kookurrenzen ab, hat jedoch nicht die interpretatorische Kompetenz des menschlichen Lesers oder der Leserin, die beim Lesen der Briefe psychologische Zusammenhänge, die nicht explizit benannt werden, erschließen (das Konzept der Rache, der Intrige).

Topic „sentiment" und Topic „bonheur"

16 |Wordle der Topics „sentiment", „bonheur" (Klee & Röttgermann 2020).

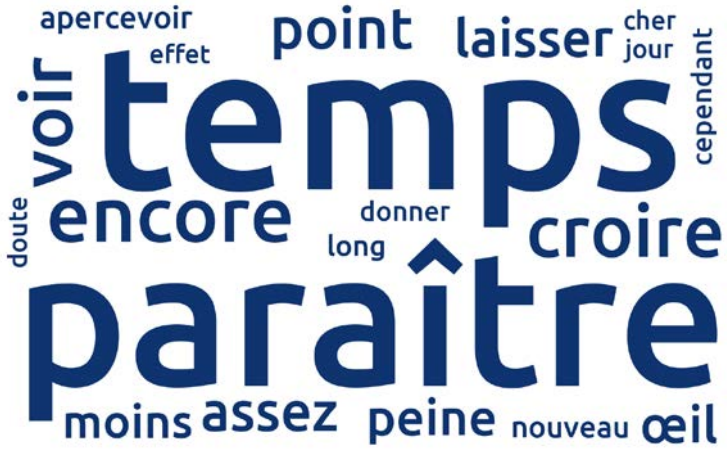
Das dritthäufigste Topic laut Topic Modeling Durchgang aus November 2020[29] für das Werk *Les Liaisons Dangereuses* ist das Topic „amour_sentiment", das Topicwörter wie Gefühl („sentiment"), Liebe („amour"), Herz („cœur"), Seele („âme") umfasst. In den statistischen Auswertungen der gesamten literarischen Produktion 1751-1800 von Seiten der Bibliograph:innen zeigt sich, dass für das Gesamtkorpus der bibliographischen Daten als vorherrschendes Thema mit 25,2 % das Thema der „intrigues sentimentales" den größten thematischen Raum im Korpus der Bibliographiedaten einnimmt (cf. dazu auch Abb. 3). Die „intrigues sentimentales" sind laut Daten aus der Bibliographie ein häufiger Topos des Romans des 18. Jahrhunderts.

[29] Topic Model des roman18 Korpus (Nov 2020), Release v0.1.0. Trier: Trier Center for Digital Humanities 2021. URL: <https://github.com/MiMoText/mmt_2020-11-19_11-38>. DOI: 10.5281/zenodo.4493224. Die hier dokumentierte Datengrundlage enthält 92 Dateien. Eingespeist in unser Wissensnetzwerk wurden jedoch nur Statements zu 79 Werken (Romanen). Die Differenz ergibt sich daraus, dass einige Files in Form von Einzelbänden vorlagen und in einem späteren Bearbeitungsschritt zu einer Datei fusioniert wurden.

Zu einem ähnlichen Schluss kommt man auch bei genauer Betrachtung der Auswertung der Topic Modeling Ergebnisse. Das Topic „amour_sentiment" hat sich bei der Ermittlung der distinktiven Topics[30] pro Werk als nicht distinktiv gezeigt. Das bedeutet, dass es ein Topic ist, welches in einer Vielzahl an Werken des Korpus prominent vorkommt. Bibliographische Daten und Topic Modeling kommen hier zu einem ähnlichen Ergebnis: Die „intrigues sentimentales" sind als literarisches Thema breit im Korpus der Bibliographie vertreten und das Topic „sentiment" ist im gesamten Romankorpus präsent.

Das Topic amour_sentiment beinhaltet Top-Topicwörter wie Liebe („amour"), Leidenschaft („passion") und Liebhaber („amant"), die semantische Überschneidungen mit dem Themenwert „libertinage", der von den Bibliograph:innen dem Werk zugeordnet wurde, aufweisen.

Topic [temps - paraître - encore]



17 | Wordcloud zum Topic [temps - paraître - encore], <https://doi.org/10.5281/zenodo.4493224>.

Die Problematik, dass einige Topics nicht ganz eindeutig zu labeln sind, zeigt sich am Beispiel des Topics [temps - paraître - encore] (Abb. 13) gut. Innerhalb der Liste der thematischen Labelkandidaten findet sich nicht nur keine Entsprechung, das gesamte Topic weist zudem eine geringe Kohärenz[31] auf. Einige Literaturwissenschaftler:innen wie zum Beispiel Ted Underwood sehen in schwer eindeutig zu labelnden, heterogenen Topics dennoch eine literaturwissenschaftlich interessante Ressource (Underwood 2012). Für das Wissensnetzwerk werden diese Topics nicht aussortiert, sondern mit einem Label, welches aus den Top-Topicwörtern besteht, eingespeist.

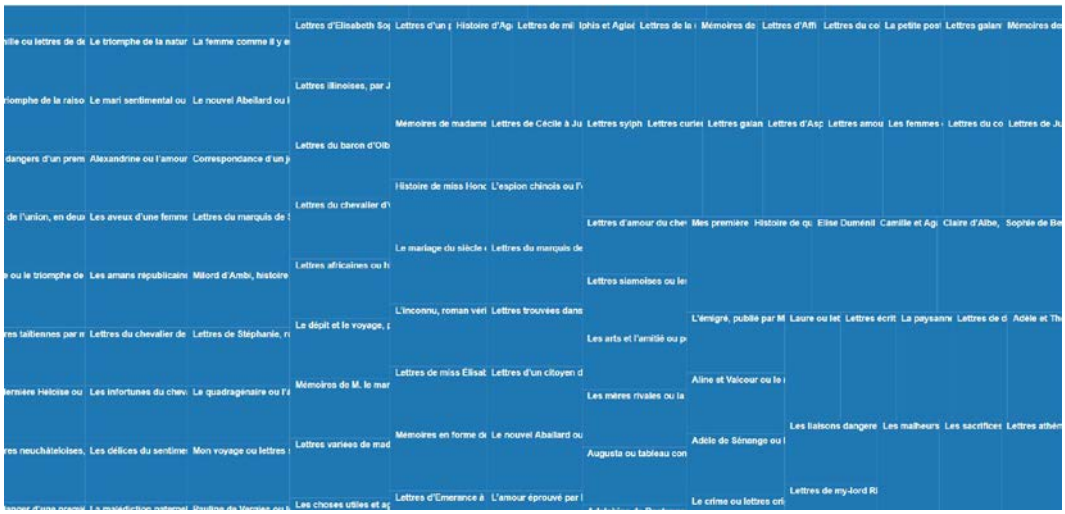
Im Abgleich der Daten zeigen sich für *Les Liaisons Dangereuses* teilweise semantische Überschneidungen (amour_sentiment/libertinage) zwischen den Ergebnissen des Topic Modelings und den bibliographischen Daten, teilweise zeigen sich aber auch komplementäre thematische Konzepte. Unterstreicht die

[30] Zur Ermittlung der distinktiven Topics vgl. Kapitel 3.4.3
[31] Zur Frage der menschlichen Bewertung der Kohärenz von Topics cf. (Chang et al. 2009).

Bibliographie die psychologischen Themen des Romans wie das Thema der Rache oder der Intrigen, extrahiert der Algorithmus des Topic Modelings in diesem Fall als weitere Topics „bonheur" und „interaction" und somit zwei Topics, die erkennen lassen, dass das doppelte Spiel der Protagonist:innen, Zynismus und Rache, nicht an der Oberfläche der Wörter erkennbar ist, sondern sich aus dem Handlungszusammenhang der Briefe und den Schlüssen des Lesenden ergeben.

Das Topic „correspondance" verweist hier auf die Gattung/Erzählform des Romans, eine Information, die auch über die bibliographischen Metadaten im MiMoText Wissensnetzwerk gespeichert ist. Interessiert sich ein Forschender für weitere Briefromane aus dem 18. Jahrhundert, lässt sich ein Überblick über die narrativen Formen im Korpus mit folgender SPARQL-Abfrage[32] im MiMoText-Wissensnetzwerk einsehen:

```
#defaultView:TreeMap
SELECT
   ?narrativeformLabel ?item ?itemLabel
WHERE
{?item wdt:P54 ?narrativeform.
 ?item wdt:P54 wd:Q3718. #?item has narrative perspective "epistolary novel"
  SERVICE wikibase:label { bd:serviceParam wikibase:language "fr". }
  }
```



18 | SPARQL-Abfrage zu narrativen Formen mit defaultView „Treemap". In der treemap werden (anklickbar: <https://tinyurl.com/2oxxeyt2>) weitere Briefromane ausgegeben. Diese Informationen basieren auf Daten aus der *Bibliographie du genre romanesque français, 1751-1800* (Martin, Mylne & Frautschi 1977).

Die Analyse des Einzelromans *Les Liaisons Dangereuses* hinsichtlich der Topics zeigt insgesamt, dass sich sinnstiftende und plausible thematische Konzepte maschinell extrahieren lassen, die jedoch in diesem Beispieltext teilweise abweichend von den von Menschenhand extrahierten thematischen Konzepten des Romans sind.

[32] Link zum projektinternen SPARQL-Endpoint: <https://query.mimotext.uni-trier.de/>Weitere Informationen zu SPARQL und ein Tutorial findet man hier : <https://mimotext.github.io/MiMoTextBase_Tutorial/>.

4.2 *Voyage autour de ma chambre* von Xavier de Maistre (1794)

> Charmant pays de l'imagination, toi que l'Être
> bienfaisant par excellence a livré aux hommes
> pour les consoler de la réalité
> (De Maistre 1794, 104)

Voyage autour de ma chambre (1794) von Xavier de Maistre ist ein Werk, das im Zuge der weltweiten Coronapandemie, die für viele Menschen im Lockdown oder „confinement" damit einherging, sich vermehrt auf die eigenen Privaträume zurückzuziehen, als Roman des 18. Jahrhunderts eine Brücke zu den Lesenden des 21. Jahrhundert schlagen kann (Laurentin 2020; Villa Ramirez & Gartner Restrepo 2020). Das zentrale Thema des Romans ist ein 42-tägiger Hausarrest, den das erzählende Ich allein in seinem Zimmer verbringt.

Schon Blaise Pascal hatte ein Jahrhundert zuvor in seinen *Pensées* bemerkt, dass alles Unglück des Menschen davon herrühre, dass er nicht ruhig in einem Zimmer zu bleiben vermöge: „Tout le malheur des hommes vient de ne savoir pas demeurer en repos, dans une chambre." (Pascal 1670, 200) Jedwede Ablenkung, jedwedes „divertissement" wie Jagd oder Tänze diene dazu, den Menschen von seiner eigenen Sterblichkeit abzulenken. Daher, so Blaise Pascal, sei auch das Gefängnis eine Bestrafung, da es den Menschen dazu verdamme, alleine ohne Ablenkung in seinem Zimmer zu bleiben (Pascal 1670, 200).

In *Voyage autour de ma chambre (1794)* verlässt der Protagonist sein Zimmer nicht, lässt jedoch in Parodie auf das im 18. Jahrhundert beliebte Genre des Reiseromans seine Phantasie schweifen und reist so imaginär aus seinem Zimmer heraus in ferne Länder. Die Struktur des Romans in 42 Kapiteln spiegelt die Dauer der Quarantäne[33] (42 Tage) wider. Auch De Maistre selbst saß, als er den Roman schrieb, aufgrund eines Duells in einem 42-tägigen Hausarrest in Turin fest.

Die Bibliograph:innen haben den Plot des Werks folgendermaßen resümiert: „Récit fantaisiste: les objets que l'auteur voit dans sa chambre évoquent des souvenirs, inspirent des réflexions." (Martin, Mylne & Frautschi 1977, 376) Zu welchem Schluss kommt der Topic Modeling Algorithmus? Die dominanten Topics werden in folgenden Statements deutlich:

[33] Im 18. Jahrhundert bezeichnet die Quarantäne noch gemäß der etymologischen Herkunft eine Dauer von 40 Tagen, analog zur religiösen 40-tägigen Einteilung der Fastenzeit. Der Eintrag QUARANTAINE der *Enyclopédie* nennt einerseits die 40-tägige Quarantäne von Schiffen zur Verhinderungen der Übertragung von Seuchen, andererseits die Quarantäne in der Rechtsprechung (D'Alembert und Diderot 1751, 658), <https://fr.wikisource.org/wiki/L%E2%80%99Encyclop%C3%A9die/1re_%C3%A9dition/QUARANTAINE>.

Maistre_Voyage	about	bonheur	https://doi.org/10.5281/zenodo.4493224
Maistre_Voyage	about	art	https://doi.org/10.5281/zenodo.4493224
Maistre_Voyage	about	nuit	https://doi.org/10.5281/zenodo.4493224
Maistre_Voyage	about	mort	https://doi.org/10.5281/zenodo.4493224
Maistre_Voyage	about	nature	https://doi.org/10.5281/zenodo.4493224

Tab.2 |Struktur der thematischen Statements zu Werken im MiMoText Knowledgegraph; zur Erläuterung der Modellierung vgl. „5. Modellierung der Themenaussagen in RDF".

Betrachten wir das Topic „art" genauer am Beispiel der Visualisierung als Wordcloud. Die Größe der Schriftart spiegelt die Gewichtung der Topicwörter im Topic wider.

19 | Wordle-Visualisierung der Topics „art", <https://doi.org/10.5281/zenodo.4493224>.

Das Topic „art" setzt sich aus Top-Topicwörtern wie Werk („ouvrage"), Buch („livre"), Gemälde („tableau"), Musik („musique"), Theater („théâtre"), Geschmack („goût") etc. zusammen. Wertet man das Werk im Close Reading aus, so bewahrheitet sich, dass es in den philosophischen Betrachtungen und Reflektionen um Glück und Kunst geht, auch um das Verhältnis der verschiedenen Künste zueinander. So berichtet das erzählende Ich beispielsweise in Kapitel XXV, dass Mme de Hautcastel Überlegungen anstellt, dass sie die Musik von Cherubini (s. im Anhang Abb. 15) und Cimarosa im Vergleich zur „alten Musik" tief berühre und dass die bildende Kunst nur von einer sehr kleinen Klasse goutiert werde, während die Musik jedwedes Lebewesen verzaubere:

> Mais que m'importe à moi, me dit un jour Mme de Hautcastel, que la musique de Cherubini ou de Cimarosa diffère de celle de leurs prédécesseurs? Que m'importe que l'ancienne musique me fasse rire, pourvu que la nouvelle m'attendrisse délicieusement? Est-il donc nécessaire à mon bonheur que mes plaisirs ressemblent à ceux de ma trisaïeule? Que me parlez-vous de peinture, d'un art qui n'est goûté que par une classe très-peu nombreuse de personnes, tandis que la musique enchante tout ce qui respire? (De Maistre 1794, 71)

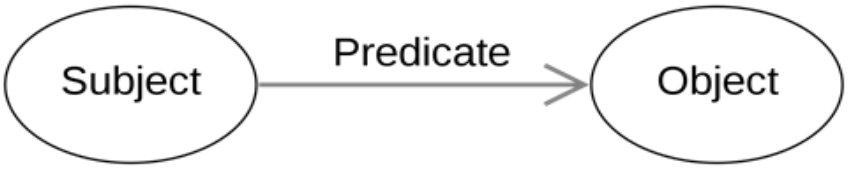
Das Close Reading bestätigt eine Diskussion des Kunstbegriffs, die sich im Topic „art" andeutete. Weit davon entfernt, nur Topics an der Oberfläche des Textes zu erkennen, konnte im diskutierten Werk mit Topic Modeling auch die psychologische und philosophische Dimension des Textes in Topics („mort", „nature", „art", „bonheur") abgebildet werden.

Am Ende von *Voyage autour de ma chambre* verlässt das erzählende Ich den geschützten Raum („la chambre") seiner Phantasiereisen und begibt sich in die Außenwelt. Denn – und hier ähnelt sich der Befund von Blaise Pascal und Xavier de Maistre – „la solitude ressemble à la mort." (De Maistre 1794, 105)

5. Modellierung der Themenaussagen in RDF

Wie in Kapitel 3 und 4 beschrieben haben wir ein Korpus an 80 Romanen der zweiten Hälfte des 18. Jahrhunderts mithilfe von Topic Modeling mit MALLET analysiert. Im Ergebnis erhalten wir 30 Topics, die mithilfe eines kontrollierten Vokabulars[34] gelabelt wurden und somit vergleichbar mit einer weiteren Informationsquelle, den bibliographischen Metadaten, sind. Die Topics wurden unter Berücksichtigung der distinktiven Topics in Relation zu den Werken zunächst in einer Liste erfasst, die alle Zuordnungen aus Top Topics und Werken enthält. Als Cutoff der berücksichtigten Topics – denn theoretisch ist jedes Topic in einer bestimmten Wahrscheinlichkeit in jedem Werk enthalten – haben wir die ersten fünf Top Topics gewählt. Die Entscheidung beruht darauf, auch hier eine Vergleichbarkeit zu den weiteren Quellen des Textminings zu erreichen (die Metadaten nennen auch bis zu fünf Themen pro Werk).

Die Ergebnisse modellieren wir sodann in Form von RDF (Resource Description Framework)[35]-Tripeln, deren Struktur sich aus drei Elementen zusammensetzt: Subjekt, Prädikat und Objekt.



20 | Grundlegendes RDF-Diagram (Wikicommons, Basic RDF Graph-en.svg, User: cmplstofB, Lizenz: WTFPL).

Übertragen auf die Modellierung unserer Themenwerte für Romane hier ein Beispiel zur Veranschaulichung:

Subjekt = Voyage autour de ma chambre

Prädikat = About

Objekt = art

[34] <https://github.com/MiMoText/vocabularies>, 24.9.2021.
[35] <https://www.w3.org/RDF/>, 24.9.2021.

In folgender Beispieltabelle (ergänzt um weitere Aussagen) bildet jede Zeile ein Tripel:

Subjekt	Prädikat	Objekt
Maistre_Voyage	about	bonheur
Maistre_Voyage	about	art

Tab. 3 | Jede Zeile bildet ein RDF-Tripel (hier in menschenlesbarer Form angedeutet).

Bezüglich des Prädikats (hier: „About") nutzen wir nach Möglichkeit bereits vorhandene Ontologien nach. „About" wurde als Prädikat aus dem Vokabular schema.org nachgenutzt.[36]

Das Projekt „Mining and Modeling Text" folgt dem Prinzip von Linked Open Data (Berners-Lee u. a. 2006), das davon ausgeht analog zu Hyperlinks in Dokumenten nun Datenobjekte bzw. Entitäten miteinander zu vernetzen. Dazu müssen Daten so ausgezeichnet werden, dass sie anschlussfähig an die Linked Open Data Cloud sind. Daher wurden im Projekt alle Themenwerte mit Identifiern aus dem Linked Open Data Hub Wikidata verknüpft. Dies ermöglicht eine Disambiguierung und bietet die Möglichkeit bei Bedarf über diese einheitliche ID automatisiert weitere Informationen aus der Linked Open Data Cloud abzurufen, beispielsweise das Label des Themenkonzepts in einer anderen Sprache (hier: Englisch).

Subjekt	Prädikat	Objekt	Wikidata ID	Label: engl
Maistre_Voyage	about	bonheur	https://www.wikidata.org/wiki/Q8	-> happiness
Maistre_Voyage	about	art	https://www.wikidata.org/wiki/Q735	-> art

Tab.4 |Über Wikidata-Identifier erreichen wir eine Disambiguierung und können weitere Informationen zu den Themenwerten abrufen, beispielsweise Label in anderen Sprachen.

Die RDF-Tripel, die auf den Ergebnissen des Topic Modelings basieren, werden aggregiert und in unsere lokale Instanz der offenen und freien Software Wikibase importiert. Dort liegen Sie nach dem Import in einer wikifizierten Form vor und können über den projekteigenen SPARQL-Endpoint abgefragt werden.

[36] <https://schema.org/about>, 28.01.2022.

21 | Projektinterne Wikibase-Instanz <http://data.mimotext.uni-trier.de>, hier der Eintrag zu *Voyage autour de ma chambre* (1794).

Durch die Aggregation tausender RDF-Tripel[37] zu den französischen Romanen 1751-1800 bildet sich ein Graph, der sich via SPARQL abfragen lässt. Es ließe sich zum Beispiel fragen, welche Romane der Zeit das Thema „famille" enthalten oder ob sich das Thema „éducation" im Zeitverlauf 1751 bis 1800 verändert, beispielsweise vor und nach 1789. Auch Kombinationen an Abfragen sind möglich: Enthalten Romane der Kategorie „Brief" vorrangig ein bestimmtes Thema? Zeige mir Romane, die von Frauen geschrieben wurden und das Topic „nature" enthalten etc. Das übergeordnete Ziel der hier vorgestellten Extraktion von Topics aus französischen Volltexten ist es, diese im Zusammenspiel mit aus weiteren Quellen (insbesondere Sekundärliteratur und bibliographische Metadaten) extrahierten Aussagen im Sinne einer "data-rich literary history" (Bode 2018, 37–57) als Wissensgraphen zu modellieren.

Die Annäherung an den Wissensgraphen über die thematischen RDF-Tripel ermöglichte es uns, unseren technischen Workflow zu etablieren und einen Grundstock an relevanten Aussagen zu den Primärtexten in unser Netzwerk einzuspeisen. Insbesondere der Vergleich der Ergebnisse aus unüberwachtem maschinellem Lernen (Topic Modeling) mit den Ergebnissen aus der Analyse der bibliographischen Metadaten ist aufschlussreich, da sie einen Mensch-Maschine-Vergleich ermöglicht.

Nachdem die erste Projektphase von MiMoText vom Korpusaufbau und dem Einspeisen von Tripeln zu Erzählformen, Textlänge, Themen, Publikationsdatum und Autor:innen getragen war, planen wir als nächste Schritte weitere RDF-Tripel

[37] Aktuell sind ca. 30.000 Tripel zu Autor:innen und Werken der französischen Prosa 1751-1800 eingespeist (Stand 28.01.2022).

zu Figuren, Handlungsorten und zur Tonalität der Romane zu erheben und auf diese Weise das Wissensnetzwerk kontinuierlich weiter anzureichern.

Hinweise

Die verwendeten Daten, Zwischenergebnisse und Visualisierungen sind auf GitHub verfügbar.

ROMANKORPUS
<https://github.com/MiMoText/roman18/>
DOI: 10.5281/ZENODO.4061904.
TOPIC MODELING WORKFLOW
<https://github.com/MiMoText/topicmodeling>
doi:10.5281/zenodo.4493223.

Das Projekt „Mining and Modeling Text" wird an der Universität Trier durchgeführt und durch die Forschungsinitiative des Landes Rheinland-Pfalz 2019-2023 gefördert.

Bibliographie

Software

KELLY, Ryan. 2004-2011. "PyEnchant".
<https://github.com/pyenchant/pyenchant>, 24.8.2021.

MCCALLUM, Andrew Kachites. 2002. "MALLET: A Machine Learning for Language Toolkit."
<http://mallet.cs.umass.edu/topics.php>, 3.12.2020.

PRESTO TEAM. 2014. "PRESTO. Projet ANR/DFG: L'évolution du système prépositionnel du français."
<http://presto.ens-lyon.fr/?page_id=197>, 14.1.2020.

ŘEHŮŘEK, Radim, und Petr Sojka. 2010. "Gensim: topic modelling for humans."
<https://pypi.org/project/gensim/>, 3.12.2020.

ŘEHŮŘEK, Radim, und Petr Sojka. 2010. "Gensim KeyedVectors."
<https://radimrehurek.com/gensim/models/keyedvectors.html>, 17.12.2020.

ŘEHŮŘEK, Radim, und Petr Sojka. 2010. "Gensim mallet wrapper."
<https://radimrehurek.com/gensim/models/wrappers/ldamallet.html >, 3.12.2020.

REUL, Christian et al. 2019. "OCR4all — An open-source tool providing a (semi-) automatic OCR workflow for historical printings." *Applied Sciences* 9 (22).
<https://github.com/OCR4all/getting_started>, 4.12.2020.

SCHMID, Helmut. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees." *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK: Association for Computational Linguistics.
<https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>, 3.12.2020.

SCHÖCH, Christof. 2020. *Simple Topic Modeling pipeline using TextBlob and gensim.*
<https://github.com/dh-trier/topicmodeling/>, 21.01.2022

Datensätze

FAUCONNIER, Jean-Philippe. 2015. *French Word Embedding Models*.
<https://fauconnier.github.io/>.

KLEE, Anne. & Röttgermann, Julia. 2020. *Doing topic modeling on French 18th century novels in the context of MiMoText project [Data set]*.

<https://github.com/MiMoText/topicmodeling>.
<https://doi.org/10.5281/ZENODO.4493223>.

LÜSCHOW, Andreas. 2019. *Bibliographie du genre romanesque français 1751-1800: RDF model [Data set]* [French]. Trier University.
<http://doi.org/10.5281/zenodo.3401428>.

RÖTTGERMANN, Julia (ed.), contributors: Dudar, J., Klee, A., Konstanciak, J., A., Ondraszek S., Probst, A., Schöch, C. 2020. *Collection de romans français du dix-huitième siècle (1750-1800) / Eighteenth-Century French Novels (1750-1800) [Data set]*, Release v0.1.0. Trier: TCDH, 2020.
URL: <https://github.com/mimotext/roman18>.
DOI: <https://doi.org/10.5281/zenodo.4061903>.

Referenzen

BERNERS-LEE, Tim et al. 2006. „A Framework for Web Science". *Foundations and Trends in Web Science* 1, Nr. 1, 1–130.
<https://doi.org/10.1561/1800000001>.

BLEI, D. M. 2011. "Introduction to Probabilistic Topic Models". *Communications of the ACM*, 1–16.

BLEI, D. M. et al. 2003. "Latent dirichlet allocation." *The Journal of Machine Learning Research*, 3, 993–1022.

BODE, Katherine. 2018. „`Man people woman life´/`Creek sheep cattle horses´: Influence, Distinction, and Literary Traditions". In *A World of Fiction: Digital Collections and the Future of Literary History*, 157–98. University of Michigan Press.
<https://www.jstor.org/stable/j.ctvdtpj1d.10>.

BODE, Katherine. 2018. *A World of Fiction: Digital Collections and the Future of Literary History*. University of Michigan Press.

BOND, Elizabeth Andrews & Robert M. Bond. 2020. „Topic Modelling the French Pre-Revolutionary Press". In *Digitizing Enlightenment : Digital Humanities and the Transformation of Eighteenth-Century Studies*, ed. Simon Burrows und Glenn Roe, 247–76. Oxford: Liverpool Univ. Press.

BURNARD, Lou. 2014. "What Is the Text Encoding Initiative? : How to Add Intelligent Markup to Digital Resources." *Encyclopédie Numérique*. Marseille: OpenEdition Press.
<http://books.openedition.org/oep/426>.

BURNARD, Lou & Carolin Odebrecht. 2019. „COST-ELTeC/Schemas: level0 and level1 release." *Zenodo*.
<https://doi.org/10.5281/zenodo.3490758>.

CHANG, Jonathan et al. 2009. „Reading Tea Leaves: How Humans Interpret Topic Models." In *NIPS*.
<http://books.nips.cc/papers/files/nips22/NIPS2009_0125.pdf>.

D'ALEMBERT, Jean Le Rond & Denis Diderot. 1751. *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers*. Paris: Briasson.
<https://fr.wikisource.org/wiki/Encyclop%C3%A9die,_ou_Dictionnaire_raisonn%C3%A9_des_sciences,_des_arts_et_des_m%C3%A9tiers>.

DAWSON, Robert L. 1978. „Review :The Martin, Mylne, Frautschi Bibliographie du genre romanesque français." Eighteenth-Century Studies 11, Nr. 4, 497–508.
<https://doi.org/10.2307/2737969>.

DELON, Michel. 1997. *Dictionnaire européen des Lumières*. Paris: PUF.

DE MAISTRE, Xavier. 1794. *Voyage autour de ma chambre*. Paris: Firmin-Didot et Cie.
<https://fr.wikisource.org/wiki/Voyage_autour_de_ma_chambre>.

DU, Keli et al. 2021. „Zeta & Eta: An Exploration and Evaluation of two Dispersion-based Measures of Distinctiveness." In *Proceedings of the*

Conference on Computational Humanities Research 2021. Amsterdam. <http://ceur-ws.org/Vol-2989/short_paper11.pdf>.

JOCKERS, Matthew L. 2014. „Topic Modeling." In *Text Analysis with R for Students of Literature*, ed. Matthew L. Jockers, 135–59. Quantitative Methods in the Humanities and Social Sciences. Cham: Springer International Publishing. <https://doi.org/10.1007/978-3-319-03164-4_13>.

LAURENTIN, Emmanuel. 2020. „Lire ‚Voyage autour de ma chambre', un texte ô combien d'actualité !" *France Culture,* 22.03. <https://www.franceculture.fr/litterature/lire-voyage-autour-de-ma-chambre-un-texte-o-combien-dactualite>.

MARTIN, A., V. Mylne & R. L. Frautschi. 1977. *Bibliographie du genre romanesque français, 1751-1800*. London: Mansell.

MIKOLOV, T. et al. 2013. *Efficient Estimation of Word Representations in Vector Space*. arXiv:1301.3781 [cs], 6.09. <http://arxiv.org/abs/1301.3781>.

PASCAL, Blaise. 1670. *Pensées de M. Pascal sur la religion et sur quelques autres sujets, qui ont esté trouvées après sa mort parmy ses papiers*. Paris: Guillaume Desprez. <https://fr.wikisource.org/wiki/Livre:Pascal_-_Pens%C3%A9es,_%C3%A9dition_de_Port-Royal,_1670.djvu>.

ŘEHŮŘEK, Radim & Petr Sojka. 2010. "Software framework for topic modelling with large corpora.", In *Proceedings of the LREC 2010 Workshop on new Challenges for NLP Frameworks*, 45-50. <https://is.muni.cz/publication/884893/lrec2010-rehurek-sojka.pdf>.

REUL, Christian et al. 2019. OCR4all -- An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. arXiv:1909.04032 [cs], September. <http://arxiv.org/abs/1909.04032>.

RHODY, Lisa M. 2013. "Topic Modeling and Figurative Language." *Journal of Digital Humanities*. 7 April. <http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody>.

ROE, Glenn, Clovis Gladstone & Robert Morrissey. 2016. „Discourses and Disciplines in the Enlightenment: Topic Modeling the French Encyclopédie." *Frontiers in Digital Humanities 2*, Lausanne: Frontiers Media SA. <https://doi.org/10.3389/fdigh.2015.00008>.

SARKAR, Dipanjan. 2019. *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing*. Second edition. New York: Apress.

SCHÖCH, Christof. 2017. „Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama." *Digital Humanities Quarterly* 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html >.

SOUVAY, G. & J.-M Pierre. 2009. "LGeRM: Lemmatisation de mots en moyen français." *Traitement Automatique des Langues*, 50 (2). <https://halshs.archives-ouvertes.fr/halshs-00396452/document>.

STEYVERS, M. & T. Griffiths. 2007. „Probabilistic topic models." In *Handbook of latent semantic analysis*, ed. Landauer, T. K. et al., 427–448, Mahwah: Lawrence Erlbaum Associates Publishers.

UNDERWOOD, Ted. 2012. „What kinds of "topics" does topic modeling actually produce?" *The Stone and the Shell* (blog). <http://tedunderwood.com/2012/04/01/what-kinds-of-topics-does-topic-modeling-actually-produce/>.

UGLANOVA, Inna & Evelyn Gius. 2020. "The Order of Things. A Study on Topic Modelling of Literary Texts." *CHR 2020: Workshop on Computational Humanities Research*, November 18–20, 2020, Amsterdam, The

Netherlands.
<http://ceur-ws.org/Vol-2723/long7.pdf>.

VILLA RAMIREZ, Oscar Jhony & Carolina Fernanda Gartner Restrepo. 2020. „Voyage autour de ma chambre dans le temps du covid-19: confrontant la réalité." *Miguilim - Revista Eletrônica do Netlli* 9, Nr. 3, 331–41.
<https://doi.org/10.47295/mgren.v9i3.2582>, 7.12.2020.

Zusammenfassung

Wie lassen sich romanistische Korpora hinsichtlich ihrer literarischen Themen mit digitalen Methoden explorativ erforschen? Im Kontext des Verbundprojekts „Mining and Modeling Text" wurde Topic Modeling mit MALLET (McCallum 2002) auf ein Korpus von 80 französischen Romanen aus der Zeit von 1750 bis 1800 (Röttgermann et al. 2020) angewandt. Ziel des Topic Modeling-Ansatzes ist es dabei, Aussagen über die Themen von Werken zu treffen, die in Form von RDF-Tripeln in ein auf Wikibase basierendes Wissensnetzwerk einfließen. Die übergeordnete, interdisziplinäre und neuartige Idee ist es dabei, datenbasierte Literaturgeschichtsschreibung zu betreiben. Neben der Informationsextraktion aus Primärtexten speist sich das Wissensnetzwerk auch aus bereits digitalisierten bibliographischen Daten (Martin, Mylne & Frautschi 1977; Lüschow 2019). Im Zusammenspiel dieser beiden Informationsflüsse lässt sich über ein gemeinsames kontrolliertes Vokabular ein aufschlussreicher Datenabgleich vollziehen: Welche Themen der Werke wurden durch die Bibliograph:innen identifiziert und welche Topics treten durch den Topic Modeling-Algorithmus zutage? Zwei Fallstudien zu Choderlos de Laclos und Xavier de Maistre exemplifizieren die Vorgehensweise und das Potential dieses Ansatzes.

Abstract

How can Romance corpora be digitally researched exploratively with regard to their literary topics? In the context of the project "Mining and Modeling Text", topic modeling with MALLET (McCallum 2002) was applied to a corpus of 80 French novels 1750-1800 (Röttgermann et al. 2020). The aim of the topic modeling approach is to generate statements about the topics of the novels, which then are imported into a knowledge graph based on Wikibase. The overriding, interdisciplinary and novel idea is to practice data-based literary historiography. In addition to information extraction on primary texts, the knowledge network is also fed from digitized bibliographic data (Martin, Mylne & Frautschi 1977; Lüschow 2019). In the interplay of these two data types, a comparison can be carried out: Which "topics" of the novels were identified by the bibliographers and which topics are revealed by the algorithm? Two case studies on Choderlos de Laclos and Xavier de Maistre exemplify this.