Comedia - Comédie: Topic Modeling als Perspektive auf das spanische und französische Theater des 17. Jahrhunderts

Abstract für die zweite Jahrestagung des Verbandes der *Digital Humanities im deutsprachigen Raum* zum Thema “Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation”, 23.-27. Februar 2015 an der Universität Graz

Christof Schöch
Universität Würzburg

Nanette Rißler-Pipka
Universität Siegen

1. Hintergrund

Im Europa des 17. Jahrhunderts entwickelten sich zeitgleich verschiedene Formen des Theaters. Trotz unterschiedlicher sozialer und poetologischer Kontexte weisen das spanische und französische Theater viele (stoffliche / stilistische) Verbindungen auf, die auf eine gemeinsame europäische Theatergeschichte hindeuten (Couderc 2013). Die Frage, ob man dazu nicht Methoden der Digital Humanities nutzen sollte, stellte sich bisher in der Romanistik nicht (anders als in weiteren Philologien, vgl. Rybicki 2012, Jockers 2013, Eder 2014). Allgemein gilt, dass sprachübergreifende, quantitative Textanalysen eine Herausforderung bleiben (vgl. Steinberger 2009, Eder/Rybicki 2011).

In romanistischer Tradition über Sprachgrenzen hinweg quantitative Verfahren anzuwenden, scheint mit Topic Modelling möglich zu sein: Die Topics mehrerer einheitlich strukturierter Textsammlungen können zunächst unabhängig voneinander modelliert werden, um dann auf der Grundlage von Topic-Labels und strukturellen Merkmalen Ähnlichkeiten und Unterschiede zu ermitteln.

2. Fragestellungen

Anstelle von Einzelhypothesen und konkreter Passagenvergleiche sind hier zwei spanische und französische Textsammlungen durch Topic Modeling verglichen worden. Welche Arten von Topics liegen vor, und wie verhalten sie sich zueinander? Welche Relation besteht zwischen den Topics und Kategorien wie Untergattungen (Komödie / Tragödie)? Wie gestaltet sich dies im Vergleich des spanischen und französischen Theaters?

Über diese Fragen hinaus soll die Eignung der Methode für Theaterstücke geprüft werden. Wie verhalten sich die "Topics" zu theaterwissenschaftlich relevanten "Themen"? Welche Perspektivenverschiebung ergibt sich durch ein quantitatives Verfahren wie Topic Modeling?

3. Textsammlungen

Die spanische Textsammlung enthält 145 Theaterstücke von sechs Autoren. Die Stücke sind zwischen 1585 und 1688 erschienen. Die Untergattungen sind "drama", "comedia" und "auto sacramental". Die Texte stammen von www.comedias.org, Wikisource und der Biblioteca Cervantes.

Die französische Textsammlung enthält 143 Theaterstücke von neun Autoren. Die Stücke sind zwischen 1630 und 1708 erschienen, und stammen von www.theatre-classique.fr. Die Untergattungen sind "comédie", "tragédie" "tragi-comédie" und "pastorale".

4. Methode: Topic Modeling

Topic Modeling ist ein quantitativer Ansatz, um in größeren Textsammlungen thematische Muster zu entdecken (Blei 2003; Anwendungen in den DH: Blevins 2010, Rhody 2012, Jockers 2013). Mathematisch gesehen sind „Topics“ Verteilungen von Auftretenswahrscheinlichkeiten von Wörtern. Die Wörter eines Topic mit der höchsten Auftretenswahrscheinlichkeit sind sich semantisch (oder anderweitig) ähnlich (vgl. Blei 2011 und Steyvers & Griffiths 2007). Durch Verknüpfung mit Metadaten können thematische Trends über einen Zeitverlauf oder thematische Differenzen zwischen Textgattungen entdeckt werden.

Wichtige Parameter sind das Präprozessieren der Texte (bspw. Lemmatisierung), die Auswahl der zu berücksichtigenden Wörter (nach Wortarten, Wortfrequenzen oder Stoplist), die Textsegmentierung sowie die Anzahl der Topics, die gefunden werden sollen. Für diese Studie wurden die Texte mit TreeTagger (Schmidt 1994) lemmatisiert und nach Wortarten annotiert. Es wurden verschiedene Textfassungen generiert, die bspw. nur Substantive und Verben enthalten, die Texte in Segmente von 40 Lemmata zerlegt und Topic Modeling mit MALLET (McCallum 2009) durchgeführt. Die Anzahl der Topics wurde auf 50 bzw. 200 festgelegt.

5. Ergebnisse und Diskussion

5.1 Die ermittelten Topics

Die ermittelten 50 Topics lassen sich meist mit einem Begriff zusammenfassen, der die inhaltliche Gemeinsamkeit der wichtigsten Worte im Topic fasst. Es gibt allgemeinere und spezifische Topics mit unterschiedlichem Gewicht in der Textsammlung, was hier am Beispiel der französischen Topics gezeigt wird (Abb. 1).

Topic „Label“	Topic-Score	Topic-Worte mit höchstem Score
Topic 14: „Liebe“	0.154	aimer amour cœur amant haïr œil âme flamme feu
Topic 34: „Komödie“	0.014	comédie pièce jouer monsieur trouver monde auteur rôle comédien
Topic 33: „Suchen-Finden“	0.180	trouver attendre sortir lieu chercher heure temps quitter ami
Topic 1: „Vergnügen“	0.100	homme esprit trouver gens femme monde rire plaire discours

Abb. 1: Auswahl von Topics aus der französischen Textsammlung.

Einige Topics betreffen allgemein gefasste, erwartbare Themen, wie bspw. Liebe / Intrigen (5 der 6 wichtigsten Topics gehören in diesen Themenbereich). Das Liebestopic enthält oft ein Element des Schmerzes und Hasses, das auf die Tragödie hindeutet (Topic 14). Dagegen lässt sich ein inhaltlich typisches Komödientopic nur durch selbstreferentiellen Begriffe erkennen (Topic 34). Faktisch am distinktivsten für die Komödie sind dagegen ein relativ unbestimmtes Topic (33, "Suchen-Finden") sowie Topic 01 ("Vergnügen"; vgl. 5.3).

Andere Topics sind spezifischer (bspw. Topic 11, "Gefahr" oder Topic 24 "Geheimnis") und könnten vermuten lassen, dass sie mit bestimmten Untergattungen des Theaters verknüpft sind (bspw. Topics 30 und 38, "Verbrechen"). Allerdings zeigt ein Vergleich von Topics und Textklassen (vgl. 5.3.), dass Topic 38 zwar der Tragödie zugeordnet werden kann, es in Topic 30 aber offenbar um ein “Verbrechen” geht, das sich in Komödien abspielt.

5.2 Die Topics im Vergleich

Vergleicht man die Topics der französischen und der spanischen Textsammlung miteinander, stellt man einige Übereinstimmungen und Unterschiede fest (Abb. 2).

Französische Sammlung		Spanische Sammlung	
Topic 41: „Liebe-Hoffnung“	cœur **amour aimer** oser **espoir** gloire **âme** vœu souffrir	**amor** celo **amar alma** amantar olvidar ver **esperanza** favor	Topic 35 „Liebe-Hoffnung“
Topic 43: „Krieg“	**guerre soldat** armée bataille chef ennemi camp champ muraille	**soldado guerra** arma valor gente tocar caja vencer armar	Topic 13 „Krieg“
Topic 7 „Arzt“	**médecin** mal **remède guérir** monsieur **maladie** fille demander homme	fingir pinzón **médico** doctor **curar** salud **enfermedad cura remedio**	Topic 17 „Arzt“
(keine Entsprechung)	-/-	justicia rey juez castigo delito mandar muerte sentencia prisión	Topic 16 „Gericht-König“
Topic 38 „Verbrechen“	crime venger mort punir sang haine vengeance perdre fureur	-/-	(keine Entsprechung)

Abb. 2: Topics im Sprachvergleich

In beiden Textsammlungen präsent sind allgemeine Topics, wie diejenigen um das Thema “Liebe” (bspw. Topic 41fr vs. Topic 35sp). Zwar kann man, abgesehen von einer leichten Tendenz in Richtung Lust (“celo, ver”) im spanischen und Leid (“souffrir”) im französischen Topic, kaum von einer semantischen Differenz sprechen. Dennoch kann Topic 35sp in der Topicverteilung nach Gattungen (vgl. 5.3) der (spanischen) "Comedia" zugeordnet werden, während Topic 41fr der (französischen) Tragödie zugeordnet wird. Die ähnlich große Wichtigkeit beider Topics in den jeweiligen Korpora belegt die stoffgeschichtliche Verwandtschaft des Theaters beider Länder.

Auch bei noch spezifischeren Topics gibt es zahlreiche Übereinstimmung, bspw. Topic 4fr und Topic 2sp, die beide mit dem Titel “Gnade-Gottes” versehen werden könnten, oder sehr konkrete Topics wie "Krieg" (Topic 46fr und 13sp) oder "Arzt" (Topic 7fr und 17sp), die mit fast identischen Wörtern vorkommen.

Topics in der spanischen Sammlung ohne Übereinstimmung in der französischen Sammlung sind bspw. Topic 45 (“Schuld-Unschuld”) oder 16 (“Gericht-König”). Umgekehrt sind Topics in der französischen Sammlung ohne Übereinstimmung in der spanischen Sammlung bspw. Topic 18 "Gehorsam" oder 38 "Verbrechen"). Diese Ergebnisse bieten Ausgangspunkte für einen Abgleich mit Erkenntnissen der Literaturgeschichte.

5.3 Topics und Textklassen

Mit unterschiedlicher Ausprägung zeigt sich in beiden Textsammlungen, dass die Untergattungen jeweils mindestens einen charakteristischen Topics besitzen (Abb. 3 und 4).

topics	AutoS	Comedia	Drama	(sd)
tp45	0.083	0.002	0.003	0.047
tp35	0.025	0.095	0.064	0.035
tp09	0.083	0.023	0.034	0.032
tp11	0.065	0.007	0.023	0.030
tp21	0.062	0.005	0.021	0.030
tp44	0.064	0.016	0.018	0.027
tp00	0.032	0.060	0.061	0.017
tp12	0.068	0.098	0.084	0.015
tp46	0.007	0.020	0.035	0.014
tp36	0.032	0.008	0.011	0.013
tp38	0.010	0.032	0.017	0.011
tp48	0.017	0.039	0.030	0.011
tp49	0.007	0.027	0.009	0.011
tp37	0.019	0.023	0.038	0.010
tp41	0.040	0.045	0.057	0.008
tp07	0.039	0.042	0.054	0.008
tp34	0.010	0.021	0.009	0.007
tp03	0.016	0.028	0.017	0.007
tp27	0.029	0.017	0.029	0.007
tp19	0.017	0.029	0.022	0.006

Abb. 3: Heatmap für Topic-Scores in Genres (Spanisch)
(20 Topics mit größter Varianz, gemessen als Standardabweichung)

Der wesentliche Kontrast bei den spanischen Stücken (Abb. 3) liegt zwischen "Auto sacramentales" einerseits, “Comedias” und “Dramas”, andererseits. Letztere haben schwächer kontrastive Topics, bspw. Topic 35 ("Liebe-Hoffnung") oder, auf niedrigerem Niveau, Topic 49 (unklar).

Bei den französischen Stücken hat jede Untergattung zumindest einen charakteristischen Topic (Abb. 4): Topic 33 ("Suchen-Finden") für die Komödie, Topic 41 ("Liebe-Hoffnung") für die Tragödie, Topic 08 ("Liebe-Schönheit") für die Pastorale. Ausnahme ist die Tragikomödie.

topic	Comedie	Tragedie	Tracom.	Pastorale	(sd)
tp08	0.052	0.029	0.087	0.129	0.044
tp41	0.059	0.120	0.061	0.050	0.032
tp19	0.027	0.076	0.057	0.015	0.028
tp33	0.088	0.029	0.051	0.051	0.024
tp38	0.026	0.078	0.043	0.029	0.024
tp01	0.061	0.012	0.016	0.019	0.023
tp03	0.002	0.002	0.004	0.044	0.021
tp28	0.041	0.004	0.006	0.007	0.018
tp26	0.028	0.044	0.069	0.049	0.017
tp14	0.042	0.051	0.049	0.078	0.016
tp11	0.026	0.058	0.033	0.029	0.015
tp20	0.045	0.068	0.044	0.036	0.014
tp00	0.040	0.011	0.015	0.015	0.013
tp48	0.013	0.031	0.033	0.011	0.012
tp49	0.025	0.002	0.003	0.003	0.011
tp02	0.038	0.055	0.063	0.059	0.011
tp22	0.008	0.010	0.010	0.031	0.011
tp09	0.010	0.028	0.016	0.006	0.010
tp05	0.030	0.012	0.015	0.028	0.009
tp24	0.038	0.047	0.029	0.030	0.008

Abb. 4: Heatmap für Topic-Scores in Genres (Französisch)
(20 Topics mit größter Varianz, gemessen als Standardabweichung)

Insgesamt scheint die gattungsbezogene Trennschärfe in den französischen Texten deutlicher als in den spanischen Texten. Dieser Befund entspricht den unterschiedlichen französischen und spanischen Poetiken der Zeit.

5.4 Gruppierung auf Grundlage von "topic scores"

Es ist nicht auszuschließen, dass die verwendeten Gattungsbezeichnungen tatsächlich vorhandene Differenzierungen verdecken. Ohne vorgängige Kategorien, nur auf Grundlage der Ähnlichkeit von Stücken nach der Verteilungen von 200 Topics sollten daher mit Principal Component Analysis Strukturen in den Textsammlungen gefunden werden.

Die räumliche Verteilung der Stücke zeigt für die spanischen Texte kaum Struktur und bildet eine recht einheitliche Wolke (Abb. 6). Die französischen Texten (Abb. 5) zeigen mehr Struktur: ein kompakterer, leicht separierter Bereich rechts oben sowie ein weiterer, besonders dichter Bereich links oben. Die in den ersten beiden Komponenten enthaltene Varianz der Daten ist mit zusammen 17,3% (französisch) und 9,2% (spanisch) verhältnismäßig gering.

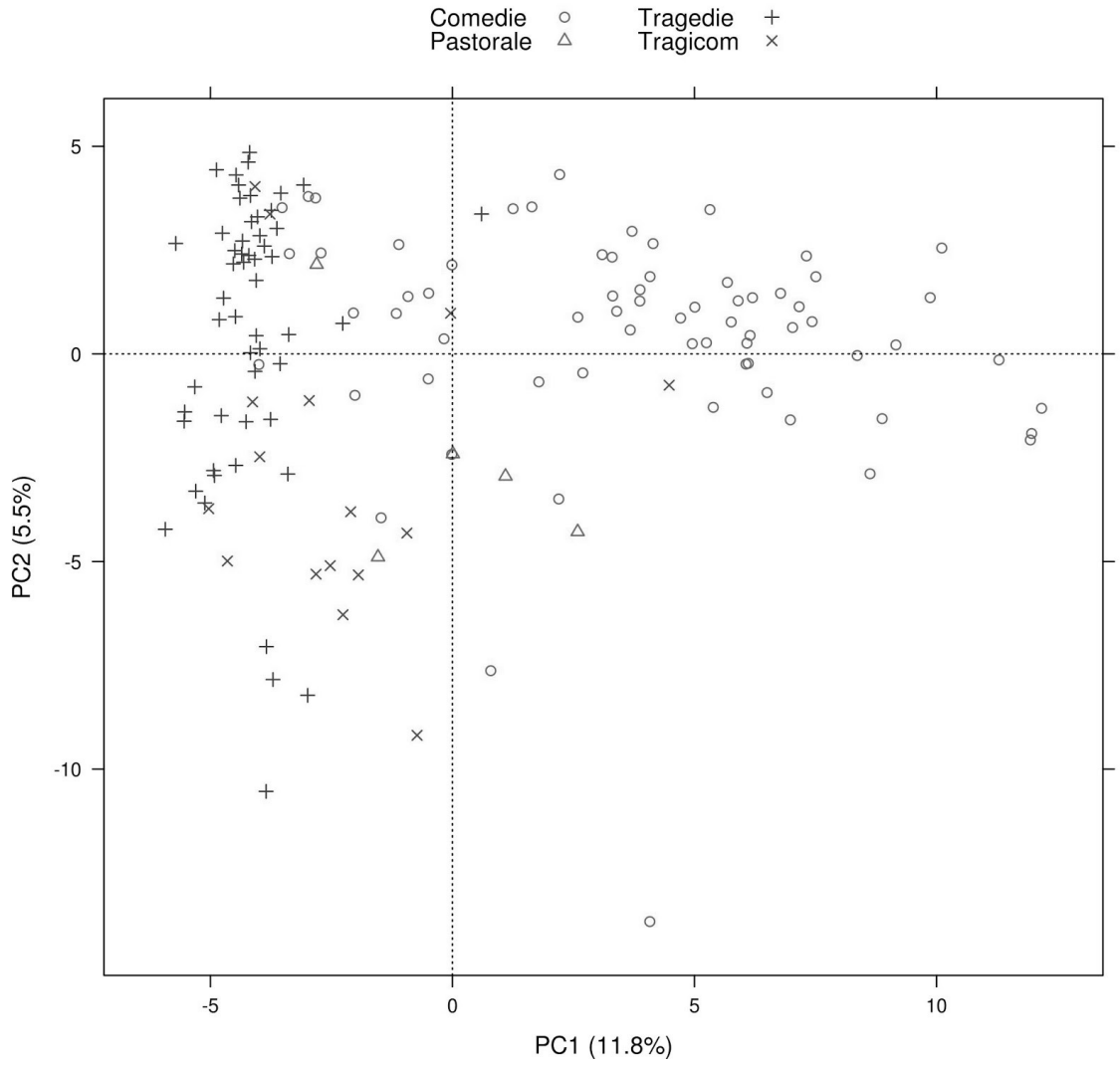

Abb. 5: PCA-Plot auf Grundlage von 200 topic scores (französische Sammlung, Genre-Labels)

Die Verteilung der Gattungssymbole zeigt, dass die französischen Texte nach Gattungen gruppiert sind: rechts oben die Komödien, links oben die Tragödien; die stärker verteilten Tragikomödien überlappen vor allem mit den Tragödien.

Bei den spanischen Texten gibt es ebenfalls Gruppen: die “Auto Sacramentales” im linken unteren Quadranten, die “Comedias” eher in der rechten Hälfte, die Dramen breit gestreut in der Mitte.

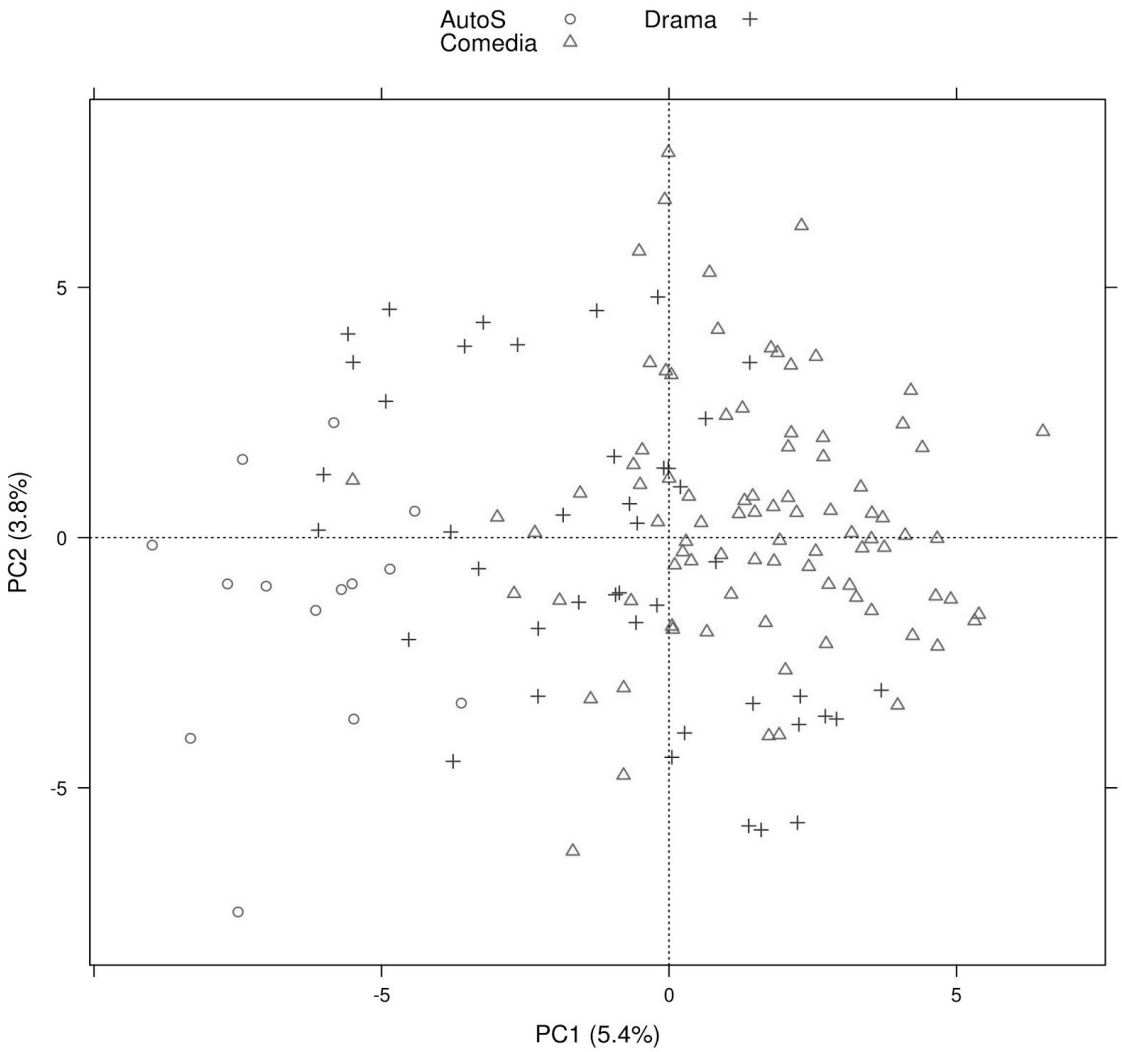

Abb. 6: PCA-Plot auf Grundlage von 200 topic scores (spanische Sammlung, Genre-Labels)

Einfache Korrelationstests bestätigen den Gesamteindruck (Abb. 7). In den französischen Stücken korreliert Genre sehr deutlich nur mit PC1, Autorschaft dagegen vor allem mit PC2. Bei den spanischen Stücken ist nur die Korrelation zwischen Autorschaft und PC1 stark.

	Französische Stücke			**Spanische Stücke**		
	PC1	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
Korrelation mit Autorschaft	0.33 ***	-0.56 ***	-0.05 ns	0.72 ***	-0.28 ***	-0.21 *
Korrelation mit Gattung	0.72 ***	0.18 *	-0.18 *	0.14 *	0.22 ns	0.04 ns
Varianz (sd)	4.86	3.31	2.89	3.28	2.75	2.61

Abb. 7: Korrelationtests zwischen Principal Components und Autorschaft bzw. Gattungszugehörigkeit

Die thematische Differenzierung der Stücke ist also in der französischen Textsammlung stärker ausgeprägt und korreliert auch stärker mit den vorhandenen Gattungs-Kategorien als in der spanischen Textsammlung.

Bilanz und nächste Schritte

Zahlreiche Einzelergebnissen zum Verhältnis der inhaltlichen Bestimmung einzelner Topics und ihrer eventuellen Zuordnung zu Untergattungen des Theaters zeigen, dass sich spanisches und französisches Theater auf Grundlage der Topic-Verteilungen auf eine Weise unterscheiden, die gattungspoetischen Positionen der Zeit entspricht und an vorhandene literaturwissenschaftliche Erkenntnisse anschlussfähig ist.

Außerdem zeigen die Ergebnisse den Unterschied zwischen “Topics” und “Themen” im literaturwissenschaftlichen Sinn. Der semantische Gehalt des Topics, der in einem Begriff wie “Liebe-Leidenschaft” (Topic 14) gebündelt werden kann, beschreibt *nicht* unbedingt das zentrale Thema der sich dahinter verbergenden Theaterstücke (vgl. die Diskussion der Tragödien-, Komödien und Pastoralentopics). Diese vermeintliche Kluft zwischen Topics und literaturwissenschaftlichen Themen ist aber eher eine Chance als ein Dilemma: so lassen sich vorschnelle Interpretationsansätze überprüfen und neue Erkenntnisse gewinnen.

Methodisch wird deutlich, dass Topic Modeling selbst nur ein Schritt in der Analyse- und Interpretationskette sein kann, der durch linguistische Annotation und Metadaten vorbereitet werden muss, und dessen Ergebnisse durch weitere Verarbeitung und Kontextualisierung erst bedeutungsvoll werden.

Als nächste Schritte könnte die Textsammlung erweitert und um weitere Metadaten ergänzt werden, um die Vergleichbarkeit der Textsammlungen zu erhöhen. Es könnte mit "Multilingual Topic Modeling" (Boyd-Graber & Blei 2009) operiert werden, das unmittelbar thematische Bezüge zwischen Dokumenten in unterschiedlichen Sprachen ermittelt. Alternativ wäre ein algorithmisches Verfahren zur Ähnlichkeitsbestimmung verschiedensprachiger Topics zu entwickeln (vgl. Pouliquen 2006).

Bibliographie

Blei, David M. 2011. “Introduction to Probabilistic Topic Models.” *Communication of the ACM*.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3,March: 993–1022.

Blevins, Cameron. 2010. “Topic Modeling Martha Ballard’s Diary.” *Historying*. http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/.

Boyd-Graber, Jordan, and David M. Blei. 2009. “Multilingual Topic Models for Unaligned Text.” In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 75–82. UAI ’09. Arlington, Virginia, United States: AUAI Press. http://dl.acm.org/citation.cfm?id=1795114.1795124.

Couderc, Christophe 2013: *La tragédie Espagnole et son contexte Européen : XVIe - XVIIe siècles*, Paris : Presses Sorbonne Nouvelle.

Eder, M. 2014. Stylometry, network analysis and Latin literature. In: *Digital Humanities 2014: Book of Abstracts*, EPFL-UNIL, Lausanne, pp. 457-58. http://dharchive.org/paper/DH2014/Poster-324.xml

Jockers, Matthew L. 2013. *Macroanalysis: Digital Methods and Literary History*. Topics in the Digital Humanities. University of Illinois Press.

McCallum, Andrew K. 2002. *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu.

Pouliquen, Bruno, Ralf Steinberger, and Camelia Ignat. 2006. “Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus.” *arXiv:cs/0609059*, September. http://arxiv.org/abs/cs/0609059.

Rhody, Lisa M. 2012. “Topic Modeling and Figurative Language.” *Journal of Digital Humanities* 2,1. http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/.

Rybicki, Jan. 2012. The great mystery of the (almost) invisible translator: stylometry in translation. In M. Oakley and M. Ji (eds.), *Quantitative Methods in Corpus-Based Translation Studies*. Amsterdam: John Benjamins, pp. 231-248.

Rybicki, Jan, and Maciej Eder. 2011. Deeper Delta across genres and languages: do we really need the most frequent words?Literary and Linguistic Computing 26(3), 315-21.

Schmid, Helmut. 1994. “Probabilistic Part-of-Speech Tagging Using Decision Trees.” In *Proceedings of International Conference on New Methods in Language Processing*. Manchester.

Steyvers, Mark, and Tom Griffiths. 2006. “Probabilistic Topic Models.” In *Latent Semantic Analysis: A Road to Meaning*, edited by T. Landauer, D. McNamara, S. Dennis, and W. Kintsch. Laurence Erlbaum.