Topic, Genre, Text

,
christof.schoech@uni-wuerzburg.de
Universität Würzburg, Deutschland

,
ulrike.henny@uni-wuerzburg.de
Universität Würzburg, Deutschland

,
jose.calvo@uni-wuerzburg.de
Universität Würzburg, Deutschland

,
daniel.schloer@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

,
stefanie.popp@uni-wuerzburg.de
Universität Würzburg, Deutschland

Einleitung

Der Beitrag möchte zeigen, wie die Berücksichtigung detaillierter, gattungsbezogener Metadaten auf produktive Weise mit dem Verfahren des Topic Modeling verbunden werden kann, um bisher nicht bekannte thematische Strukturen im Textverlauf in einer Sammlung spanischer und hispanoamerikanischer Romane zu entdecken. Ausgangshypothese ist, dass die Wichtigkeit bestimmter Topics nicht nur im Textverlauf variiert, sondern dies auch in verschiedenen Untergattungen auf unterschiedliche Weise tut. Eine Pilotstudie wurde im März 2015 beim Workshop zu Computational Narratology bei der DHd-Tagung in Graz vorgestellt. Im Rahmen der interdisziplinären Würzburger eHumanities-Nachwuchsgruppe "Computergestützte literarische Gattungsstilistik (CLiGS) wurde dieser Fragestellung nun mit weiter entwickelten Methodik sowie einer neu erstellten Sammlung spanischsprachiger Romane aus Spanien und Hispanoamerika nachgegangen.

Stand der Forschung und Fragestellung

Die Frage nach dem Text- oder Handlungsverlauf in narrativen literarischen Texten hat jüngst zunehmende Aufmerksamkeit in der digitalen Literaturwissenschaft erhalten. Matthew Jockers kam durch Sentiment Analysis im Verlauf zahlreicher Romane zu dem (kontrovers diskutierten) Ergebnis, es gäbe sechs oder sieben grundlegende Plotstrukturen (Jockers 2015). Ben Schmidt hat unter anderem den Verlauf von Topic-Wahrscheinlichkeiten in der "screen time" amerikanischer Fernsehserien verfolgt (Schmidt 2014). Der vorliegende Beitrag verbindet die Frage nach dem Textverlauf mit der nach den Untergattungen, seine zentrale Fragestellung lautet: Können wir nach Untergattung unterschiedliche Verlaufsmuster für bestimmte Topics über den Textverlauf hinweg feststellen?

Daten

Die Textsammlung enthält 150 spanische und hispanoamerikanische Romantexte aus der Zeit von 1880 bis 1930 (für den spanischen Roman: Altisent 2008; de Nora 1963, für den hispanoamerikanischen Roman: Gallo 1981; Williams 2009). Die Texte sind in TEI aufbereitet und mit detaillierten Metadaten versehen worden. Es wurden vier weit gefasste Untergattungen gewählt, um die Romane miteinander vergleichen zu können: *novela sentimental*, *novela histórica*, *novela político-social* und *novela de tendencia subjetiva*. Die Auswahl der Texte ist auch von der Verfügbarkeit als digitaler Volltext beeinflusst und daher nicht unbedingt repräsentativ. Abbildung 1 zeigt die Verteilung der Romane nach ausgewählten Metadaten.

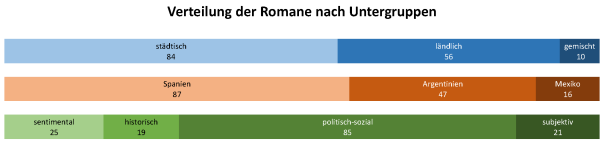

***Abb. 1**: Verteilung der Romane nach Metadaten*

Methode

Topic Modeling ist eine unüberwachte, nicht-deterministische Methode aus dem Bereich des *Natural Language Processing*, die auf Annahmen aus der distributionellen Semantik basiert und verborgene semantische Strukturen in großen Textsammlungen aufdeckt (einführend Blei 2011, grundlegend Blei 2003). Gruppen semantisch verwandter Wörter werden insbesondere aufgrund ihres häufigen gemeinsamen Auftretens in den untersuchten Dokumenten entdeckt. Ein Topic ist eine Wahrscheinlichkeitsverteilung von Wörtern; ein Dokument wird als Wahrscheinlichkeitsverteilung von Topics beschrieben. Topic Modeling ist eine in den DH äußerst beliebte Methode (Anwendungsbeispiele: Blevins 2010; Rhody 2012; Jockers 2013; Schöch 2015).

Hier wurde Topic Modeling als Teil eines umfassenden, weitgehend automatischen Arbeitsablaufes als Serie von Python-Skripten implementiert: Präprozessieren der

Texte (Segmentierung, Binning, Lemmatisierung, POS-Tagging), das eigentliche Topic Modeling (mit Mallet, siehe McCallum 2002), Aufbereitung des Mallet-Outputs, zahlreiche Visualisierungen als Perspektiven auf die Ergebnisse. Die wichtigsten Parameter: Berücksichtigung ausschließlich der Substantive, Weglassung der 70 häufigsten Substantive, Romansegmente von ca. 600 Wörtern (unter Berücksichtigung von Absatzgrenzen), Anzahl von 70 Topics. Die Python-Skripte sind frei verfügbar und ausführlich dokumentiert, Begleitmaterialien (Skripte, Parameterdatei, Metadaten, Abbildungen) sind unter https://github.com/cligs/projects/tree/master/2016/dhd einsehbar.

Ergebnisse und Diskussion

Es werden zunächst die Topics selbst dargestellt, dann Unterschiede in den Topic-Verteilungen nach Untergattungen, über den Textverlauf hinweg und schließlich über den Textverlauf in Abhängigkeit der Untergattung.

Topics

Die Mehrheit der erhobenen Topics beinhaltet konkrete typische Themen und Motive des spanischsprachigen Romans der Epoche. Man erkennt eine klare semantische Beziehung der Wörter: ein konkreter Bereich menschlicher Tätigkeiten, wie in Topic 19 (maestro-colegi o-escuela, dt. "Lehrer-Schule-Schule") oder Topic 23 (sangre-golpe-arma, dt. "Blut-Schlag-Waffe"); oder abstrakte Begriffe und Gefühle, wie bei Topic 69 (conciencia-honor-crimen, dt. "Gewissen-Ehre-Verbrechen"). Weniger kohärent ist Topic 45 (marido-rato-chico, dt. "Ehegatte-Weile-Junge"). Die folgenden Wordclouds (Abbildung 2) veranschaulichen die erwähnten Topics.

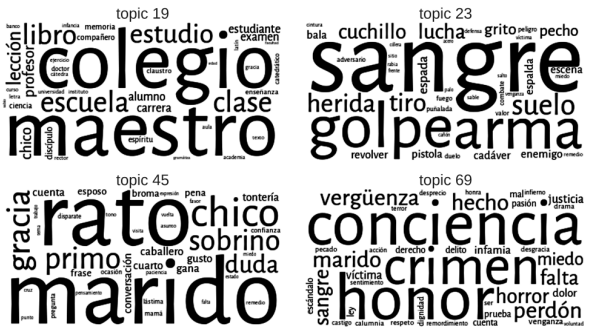

***Abb. 2**: Wordclouds für ausgewählte Topics.*

Untergattungen und Topics

Die folgende Heatmap (Abbildung 3) zeigt die Verteilung der durchschnittlichen Topic-Wahrscheinlichkeiten in den vier Untergattungen für diejenigen 20 Topics, deren Werte zwischen den Untergattungen besonders stark schwanken (nach Standardabweichung). Besonders distinktive Topics existieren für die *novela de tendencia subjetiva* (Topic 11: mirada-huerto-silencio, dt. "Blick-Garten-Stille") und die *novela sentimental* (Topic 45). Wenig überraschend auch, dass die *novela histórica* als distinktiven Topic unter anderem Topic 57 hat (rey-caballero-príncipe, dt. "König-Ritter-Prinz"). Für die *novela histórico-social*, für die aufgrund der großen Zahl von Beispielen eine größere Bandbreite an Topic-Verteilungen zu erwarten ist, gibt es keinen vergleichbar stark distinktiven Topic. Dennoch sind die Untergattungen ein wichtiger Faktor für die Verteilung der Topics in der Sammlung und die thematische Komponente spielt für die Definition der Untergattungen tatsächlich eine wesentliche Rolle.

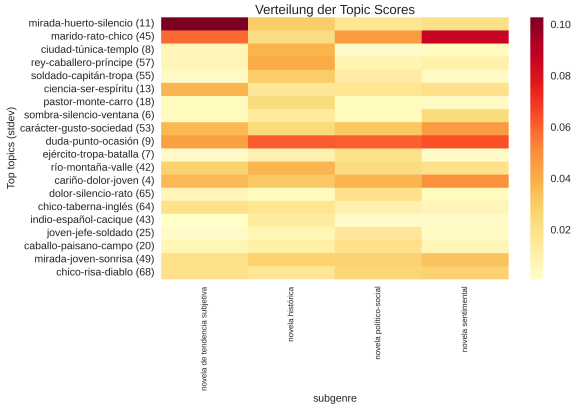

***Abb. 3**: Verteilung von Topic-Scores nach Untergattungen.*

Topics im Textverlauf

Die Ausprägung der Topics variiert nicht nur hinsichtlich der Untergattungen, sondern auch über den Textverlauf hinweg. So gibt es einige Topics, deren Vorkommen am Anfang der Romane besonders wahrscheinlich ist (Abbildung 4a). Dazu zählen Topic 10 (vino-plato-pan, dt. "Wein-Teller-Brot"), Topic 17 (sombrero-ropa-bota, dt. "Hut-Kleidung-Stiefel") und Topic 19, welche auf die Beschreibung von Ambiente, Situation und Personen hindeuten. Gegen Ende der Romane sind andere Topics wahrscheinlicher (Abbildung 4b), z. B. Topic 2 (pecado-caridad-conciencia, dt. "Sünde-Wohltätigkeit-Gewissen"), Topic 23 und Topic 69, also abstraktere Themen oder solche, die sich auf Wertvorstellungen beziehen. Dies deutet darauf hin, dass in den Romanen am Ende Bilanz gezogen wird, die Handlung einen drastischen Ausgang nimmt oder das im Textverlauf Behandelte in gesellschaftliche oder religiöse Diskurse eingebunden wird.

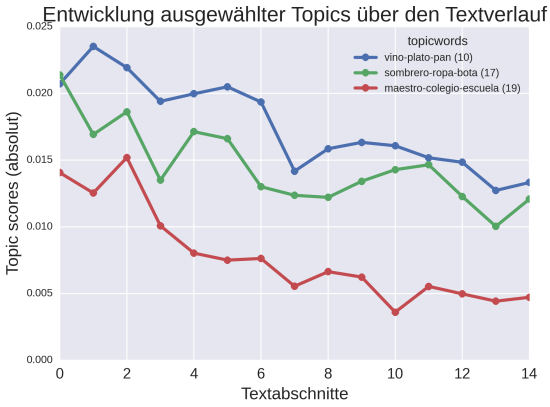

***Abb. 4a**: Verteilung von Topics im Textverlauf (fallend).*

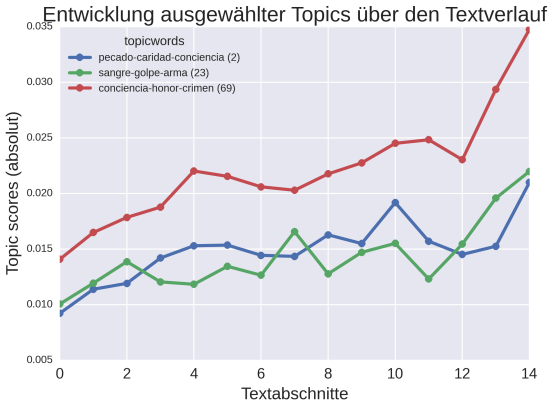

***Abb. 4b**: Verteilung von Topics im Textverlauf (steigend).*

Textverlauf abhängig von den Untergattungen

Für einige der genannten Topics, die in bestimmten Bereichen des Textverlaufs wahrscheinlicher sind, kann die Tendenz über alle Untergattungen hinweg bestätigt werden (bspw. bei Topic 10 und 17, siehe oben). Es gibt aber auch Themen, bei denen sich durch die Betrachtung des Verlaufs in den einzelnen Untergattungen ein differenzierteres Bild ergibt. Die Wahrscheinlichkeit von Topic 23 beispielsweise nimmt nur für die *novela político-social* zum Ende hin zu (Abbidung 5a):

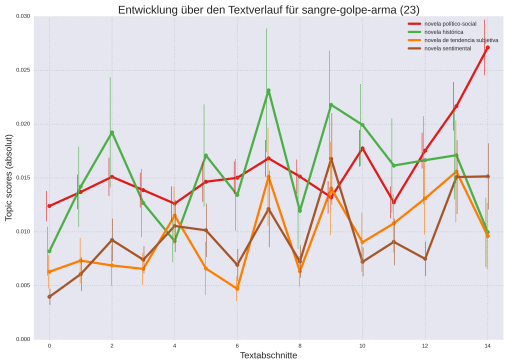

***Abb. 5a**: Topic 23 nach Textverlauf und Untergattung.*

Das kann so interpretiert werden, dass die *novela político-social* im Gegensatz zu den anderen Untergattungen dazu tendiert, am Ende des Textes mit einer gewalttätigen Szene und einem Umbruch zu schließen. Topic 19 ist nicht in allen Untergattungen zu Beginn des Textverlaufes stark ausgeprägt, sondern nur bei der *novela de tendencia subjetiva*. Dies erklärt sich, weil bei diesen Romanen das Schulthema als Teil einer autofiktionalen Erzählung zu Beginn erscheint (Abbildung 5b):

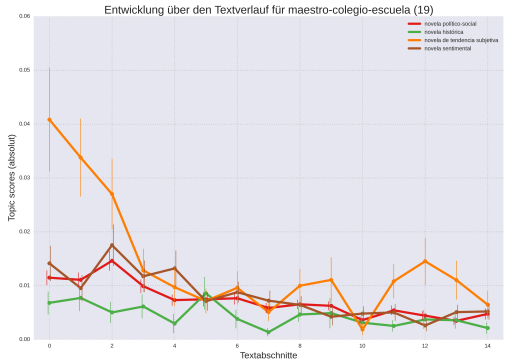

***Abb. 5b**: Topic 19 nach Textverlauf und Untergattung*

Allgemein gilt, dass die Untergattungen sich in ihrer Topicverteilung im Textverlauf auch dann deutlich unterscheiden können, wenn dies für alle Untergattungen zusammengenommen nicht der Fall ist und so leicht übersehen werden könnte.

Für die Berechnung wurden die Romansegmente von 600 Wörtern bezüglich des Textverlaufs auf 15 Romanabschnitte (Bins) verteilt, um die unterschiedliche Romanlänge zu berücksichtigen. Diese Bins wurden hinsichtlich der Untergattung gruppiert und jeweils das arithmetische Mittel bestimmt. Die in den Plots eingezeichneten Kurven entsprechen der linearen Interpolation dieser gemittelten Werte. Zusätzlich wurde der Standardfehler vertikal um den jeweiligen Kurvenpunkt eingezeichnet, der deutlich macht, wie sehr die jeweiligen

dem Mittelwert zugrunde liegenden Werte streuen, also wie gut der Mittelwert die Gesamtheit der Segmentwerte repräsentiert.

Die Ergebnisse im literaturgeschichtlichen Kontext

Insgesamt zeigen sich verschiedene Zusammenhänge: Zwischen bestimmten Topics und einzelnen Roman-Untergattungen, zwischen Topics und dem Textverlauf, und dies zum Teil dann auch wieder in Abhängigkeit von den Untergattungen. Aus literaturgeschichtlicher Perspektive betrachtet erweisen sich die in die Untersuchung einbezogenen Metadaten für eine Einordnung der Topic-Resultate als nützlich. Topics sind für die Romangattungen im vorliegenden Korpus ein wichtiger Faktor, ähnlich wie dies für Gattungen wie die klassische Komödie und Tragödie bereits gezeigt werden konnte (Schöch 2015).

Ein detaillierterer Blick zeigt beispielsweise Folgendes: Topic 11, welches typisch für die *novela de tendencia subjetiva* ist, ist vor allem in den 1910er- und 1920er-Jahren wichtig sowie für bestimmte Autoren. Interessanterweise ist dieses bei spanischen und hispanoamerikanischen Modernisten vorkommende Thema auch bei der früher wirkenden Schriftstellerin Juana Manuela Gorriti schon wichtig, die offenbar thematische Präferenzen späterer Autoren vorweggenommen hat. Außerdem kommt Topic 11 bei Larreta in einem (modernistischen) historischen Roman vor, obwohl es ansonsten vor allem für die Romane subjektiver Tendenz typisch ist. Es ist anzunehmen, dass für dieses spezielle Thema eher die literarische Strömung bestimmend ist als die Untergattung. Der Topic enthält einige für die modernistische Strömung typische Wörter, etwa zu Sinneseindrücken (azul, dt. "blau", olor, dt. "Geruch") und Zurückgezogenheit (huerto, silencio, campo, soledad, dt. "Garten, Ruhe, Land, Einsamkeit").

Fazit und Ausblick

Die Nutzung von Topic Modeling als Methode kann für die digitale Literaturwissenschaft verbessert werden, wenn spezifisch literaturwissenschaftliche Metadaten in die Betrachtungen einbezogen werden und die Textstruktur - hier als Sequenz von Textverlaufseinheiten - berücksichtigt wird. Verschiedene Visualisierungsstrategien erweisen sich als entscheidende "Interfaces" zu den Daten (im Sinne von Doueihi 2012), die Muster sichtbar machen und den Blick lenken. Die Ergebnisse des Topic Modelings können differenzierter und aus verschiedenen Perspektiven betrachtet und mit literaturhistorischem Wissen in Verbindung gebracht werden. Die Ergebnisse ergänzen und erweitern etablierte hermeneutische Lektürestrategien, insofern sie einen synthetisierenden Blick auf sehr umfangreiche Textsammlungen erlauben.

Nächste Schritte betreffen insbesondere die weitere Auseinandersetzung mit der Signifikanz von Unterschieden in den Topic-Wahrscheinlichkeiten im Textverlauf, deren Berechnung u. a. durch die mangelnde Normalverteilung der Werte nicht trivial ist. Zusätzlich zu den Untergattungen sollen auch Kategorien wie das Setting modelliert werden. Zudem sollen die Textverlaufs-Daten für die automatische Klassifikation von Romanen nach Untergattungen genutzt werden. Schließlich wird bereits an der Erweiterung der Textsammlung gearbeitet, insbesondere mit Blick auf den Umfang und ein ausgeglicheneres Verhältnis der Untergattungen.

Bibliography

Altisent, Marta E. (2008): *A Companion to the Twentieth-Century Spanish Novel*. Woodbridge: Tamesis.

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

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

Blevins, Cameron (2010): “Topic Modeling Martha Ballard’s Diary,” in: *Historying* http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/ [letzter Zugriff 16. Februar 2016].

Doueihi, Milad (2012): *Pour un humanisme numérique (2011)*. Paris: Seuil.

Gallo, Marta (1981): *La Novela Hispanoamericana En El Siglo XIX*. Madrid: La Muralla.

García de Nora, Eugenio (1963): *La Novela Española Contemporánea*. Madrid: Gredos.

Jockers, Matthew L. (2013): *Macroanalysis - Digital Methods and Literary History*. Champaign, IL: University of Illinois Press.

Jockers, Matthew L. (2015): “Revealing Sentiment and Plot Arcs with the Syuzhet Package” in: *Matthew. L. Jockers* http://www.matthewjockers.net/2015/02/02/syuzhet/ [letzter Zugriff 09. Februar 2016].

McCallum, Andrew K. (2002): *MALLET: A Machine Learning for Language Toolkit* http://mallet.cs.umass.edu [letzter Zugriff 09. Februar 2016].

Nachwuchsgruppe CLiGS (o.J.): *Computergestützte literarische Gattungsstilistik* http://cligs.hypotheses.org/ [letzter Zugriff 16. Februar 2016].

Rhody, Lisa M. (2012): “Topic Modeling and Figurative Language,” in: *Journal of Digital Humanities* 2 http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody [letzter Zugriff 09. Februar 2016].

Schmidt, Benjamin M. (2014): “Typical TV Episodes: Visualizing Topics in Screen Time,” in: *Sapping Attention* http://sappingattention.blogspot.de/2014/12/typical-tv-episodes-visualizing-topics.html [letzter Zugriff 09. Februar 2016].

Schöch, Christof (2015): "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama [submitted]", in: *Digital Humanities Quarterly*.

Williams, Raymond L. (2009): *The Twentieth-Century Spanish American Novel*. Austin, Texas: University of Texas Press.