COMEDIA - COMÉDIE

Topic Modeling als Perspektive auf das spanische und französische Theater des 17. Jahrhunderts

Christof Schöch (Univ. Würzburg)
Nanette Rißler-Pipka (Univ. Siegen)

DHd-Tagung, 23.-27.2.2015, Univ. Graz

1 Vergleich – 2 Perspektiven

Traditionelle Literaturwissenschaft

- 2 kontroverse poetologische Konzepte an ausgewählten Beispielen vergleichen
- Stoffgeschichtliches Vorgehen: z.B. „Don Juan" oder Bibliographieren einzelner Verwandtschaftsbeziehungen zwischen frz. und sp. Stücken (Losada Goya 1999)

Quantitativ mit Topic Modelling

- Überblick über die Masse der Theaterstücke, die im 17. Jhd. vor allem in Spanien, aber auch in Frankreich entstehen
- Sprachübergreifender Vergleich mit **Topic Modelling**: 2 Textsammlungen einzeln ausgewertet und Vergleich der Topic-Labels, Genreverteilung, Autorenverteilung

Die Fragestellung

- Wie unterscheiden sich französische und spanische Theaterstücke in der jeweiligen Blütezeit (Siglo de Oro / siècle classique)?
- Gibt es Genrebesonderheiten, stoffliche Vorlieben, Alleinstellungsmerkmale bestimmter Autoren?
- Können durch Topic Modelling dazu wissenschaftlich fundierte Antworten gefunden werden?
- Wie verläuft das semantische Spannungsfeld zwischen algorithmisch ermittelten Topics und literaturwissenschaftlichen Themen?

Die Daten

2 Textsammlungen

145 sp. Stücke: 6 Autoren; 3 Untergattungen (Comedia, Drama, Auto sacramental); Zeitrahmen: ca. 1585-1688
[Quelle: comedias.org; es.wikisource.org; cervantesvirtual.com]

143 frz. Stücke: 9 Autoren; 4 Untergattungen (Comédie, Tragédie, Tragicomédie, Pastorale); Zeitrahmen: ca. 1630-1703
[Quelle: theatre-classique.fr]

Die Methode

- Topic Modelling mit MALLET
- Vorbereitung und Auswertung mit Python

- Vorbereitung

1. Lemmatisierung und Annotation nach Wortarten
2. Selektion der Verben + Substantive (+ Stoplist)
3. Segmentierung der Texte (40 Lemmata)

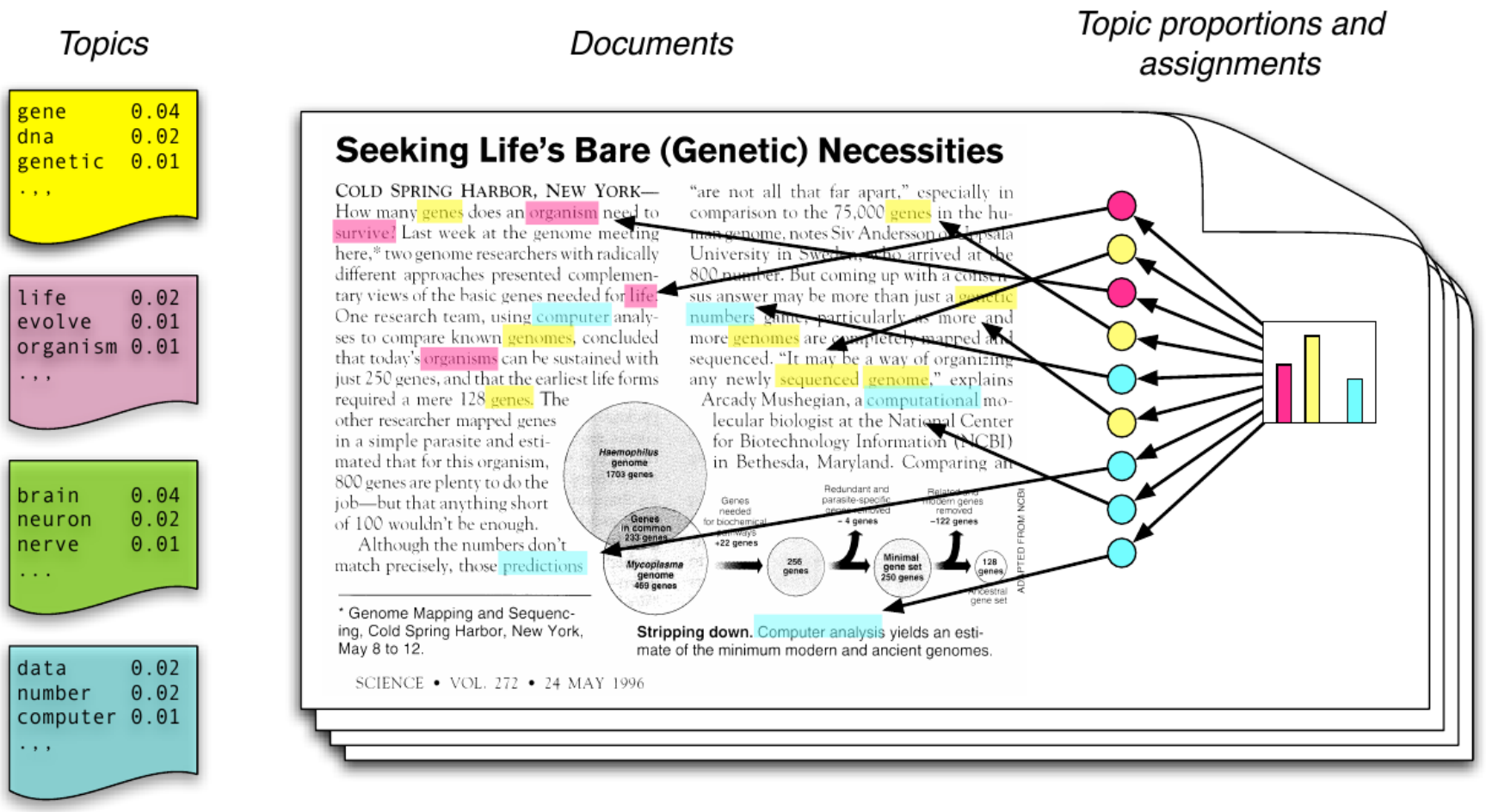
- Parameter

1. Anzahl der Topics: 50 sowie 200

- Auswertung

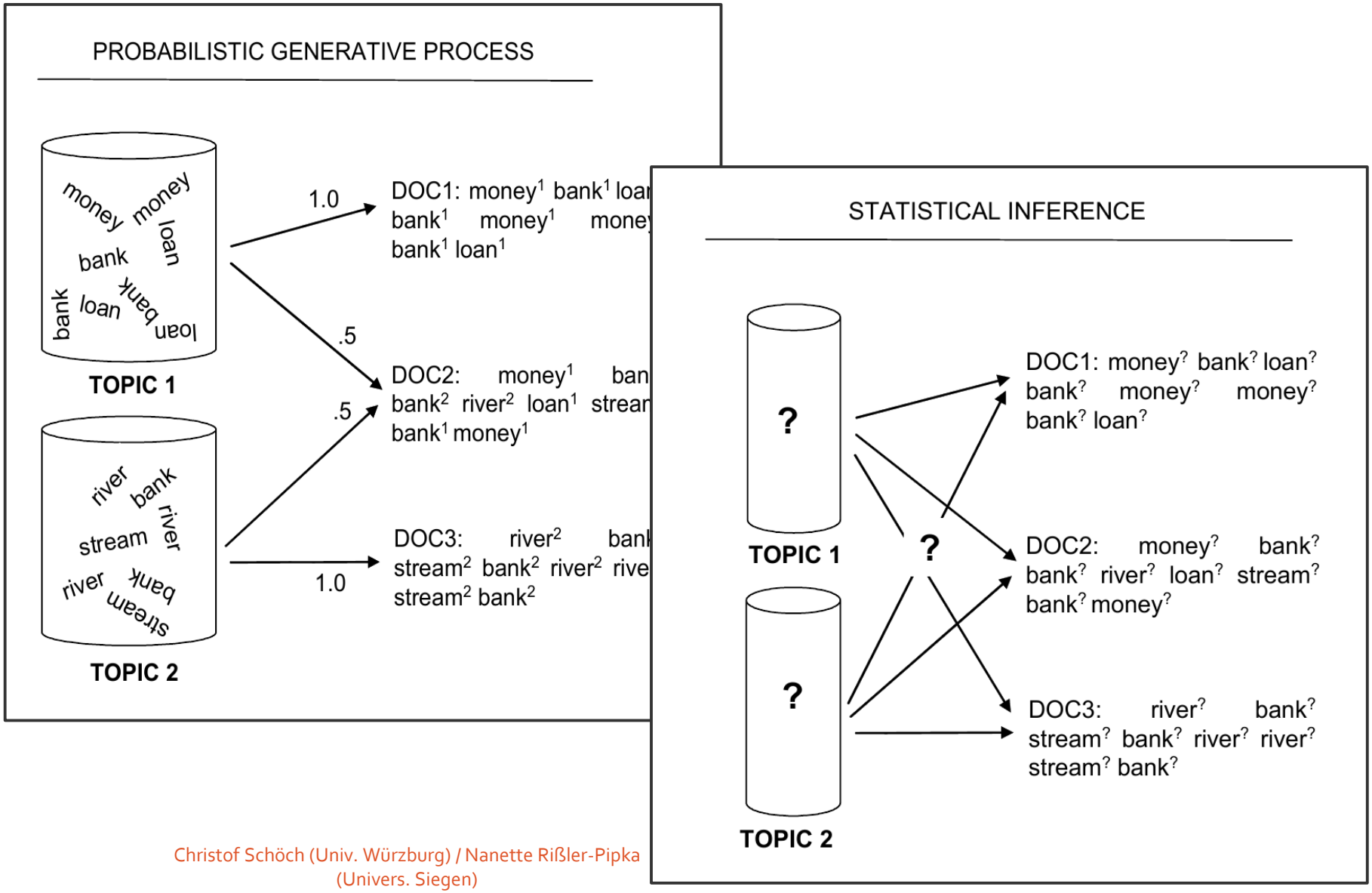
1. Aggregation nach Metadaten

Die Methode: Topic Modeling



Christof Schöch (Univ. Würzburg) / Nanette Rißler-Pipka (Univers. Siegen)

Die Methode: Topic Modeling



Christof Schöch (Univ. Würzburg) / Nanette Rißler-Pipka (Univers. Siegen)

Das Problem

- Mit „Comedia" können im Spanischen fast alle Theaterstücke des 17. Jhds. bezeichnet werden
- Im klassischen Theater Frankreichs gibt es eine strikte Trennung zwischen Comédie und Tragédie (+ das Untergenre der Tagicomédie)
- Vergleichen wir also Äpfel mit Birnen?
- Nein – der Reiz des Vergleichs liegt gerade darin, zu überprüfen, ob sich auch in den Ergebnissen des Topic Modelling der Gegensatz manifestiert:
  Spanien: Genremischung – Frankreich: Genretrennung

Der Kompromiss

- Für die spanische Textsammlung nehmen wir pro forma eine Einteilung nach Tragödie vs. Komödie vor (analog zu spanischen Bezeichnungen wie „comedia/drama de honor" (Tragödie) und „comedia de capa y espada (Komödie))
- So erhalten wir zwei vergleichbare Textsammlungen, die nicht nur formal gleich modelliert wurden (durch Lemmatisierung, etc.), sondern auch in ihren Gattungsbezeichnungen vergleichbar werden – letzteres betrachten wir als vorläufigen Kompromiss

Die Ergebnisse

- Pro Textsammlung 50 Topics (deren erste 9 Wörter betrachtet werden)
- **3-fache Analyse**

1. Topics und ihr semantischer Gehalt: Vergabe der Label

2. Zusammenhang zwischen Topics und Kategorien von „Autorschaft" / „Untergattung"

3. Clustering auf Grundlage der Topic Scores (mit 200 Topics)

Die Topics

- Beispiel frz. Sammlung: allgemeine und spezifische Topics
- Höchster Topic-Score für Liebes-Topics: bestätigt die Thesen

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

- Frage: Tragödie oder Komödie?
- Semantischer Trugschluss: Topic 14 = Tragödie vs. Topic 34 = Komödie
- Distinktiv für Tragödie: Topic 41 „Liebe-Hoffnung“: „cœur amour aimer oser espoir gloire âme vœu souffrir“
- Distinktiv für Komödie: Topic 33 und 1

Die Topics

- **Fazit: Unterschied zwischen "Topics" und "Themen"**

1. Topic-Label oder -Wörter ≠ Themen, der darunter fallenden Stücke

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

2. Genaue Abgrenzung zwischen einzelnen Topics nicht möglich (z.B. zwischen Liebe-Hass vs. Liebe-Hoffnung)
3. Dennoch erkennbarer Unterschied zwischen Tragödien-Topic (Nr. 41) und Komödien-Topic (Nr. 33)

Die Topics – im Sprachenvergleich

- Fast wörtliche Übereinstimmgen

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)

- Topics ohne Entsprechung: könnten als Besonderheiten des jeweiligen Theaters interpretiert werden: aber Relevanz (Topic Score) ist zu beachten

Die Topics – im Sprachenvergleich

- **Fazit: belegbare Gemeinsamkeiten –Stoffvergleich möglich**

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)

- Interessante Brüche: Topic 41 (frz.) für Tragödie distinktiv – fast gleiches Topic 35 (spa.) für Komödie distinktiv

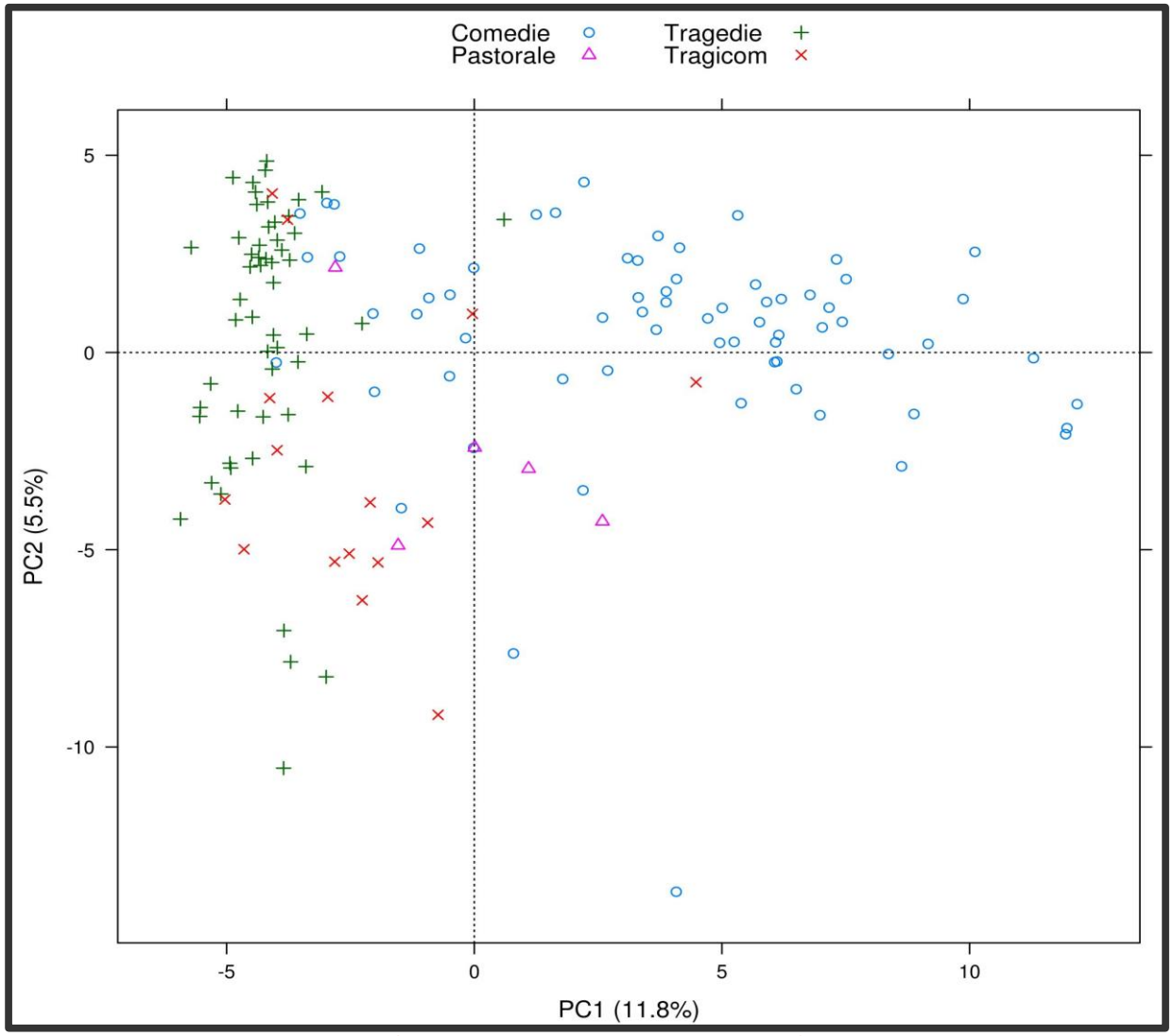
Topics & Genres: französisch

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

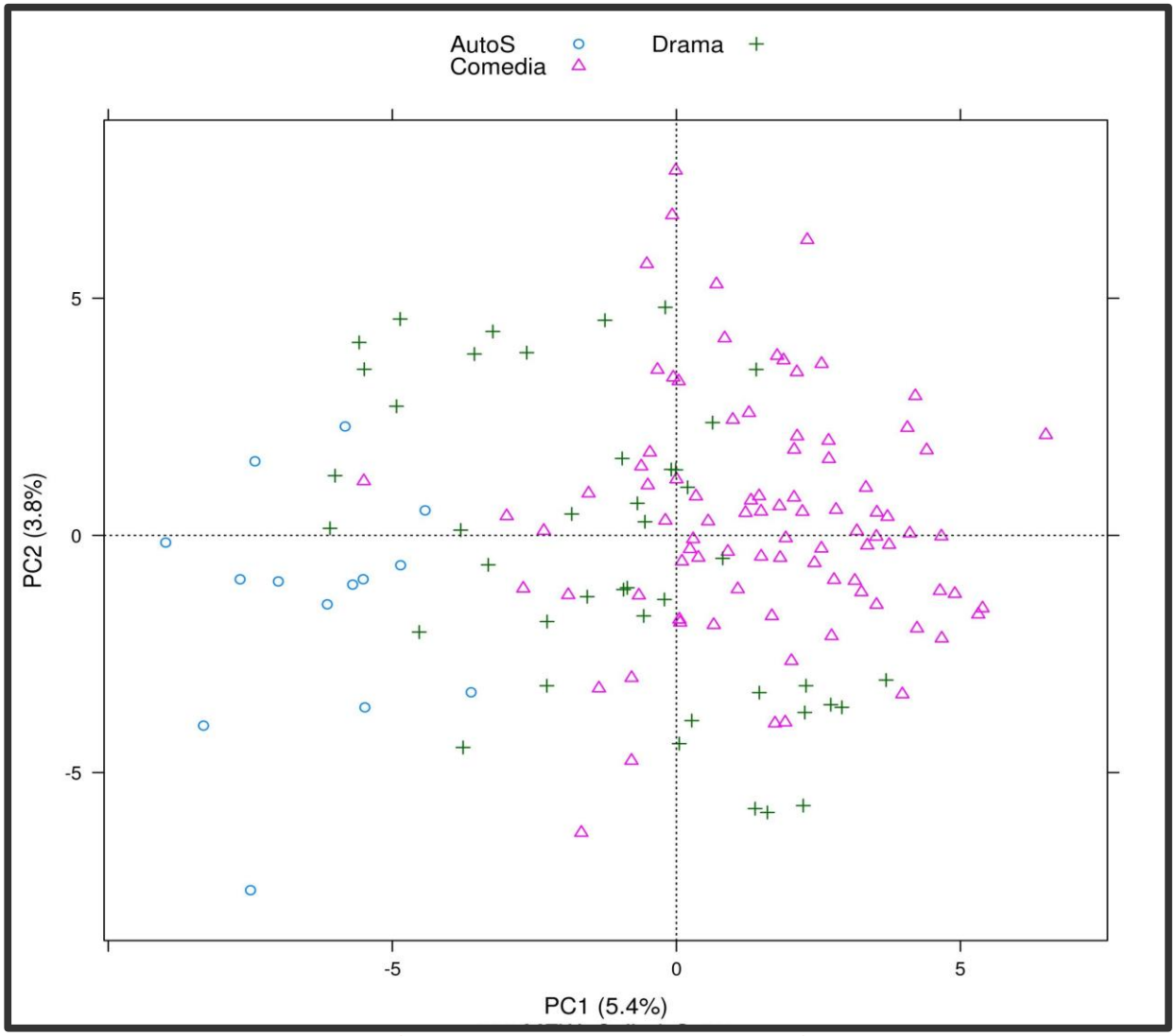
Topics & Genres: spanisch

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

PCA mit topic scores: französ.



PCA mit topic scores: spanisch



Conclusion

- **Bilanz**

1. Interessante, belegte Ähnlichkeiten und Unterschiede zwischen spanischer und französischer Dramen-Tradition
2. Topics sind nicht gleich Themen

- **Nächste Schritte**

1. Erweiterung der Textsammlung
2. Detailliertere Metadaten (Vergleichbarkeit)
3. Ähnlichkeitsbestimmung verschiedensprachiger Topics
4. Multilingual Topic Modeling

Bibliographie

- Blei, David M. 2011. "Introduction to Probabilistic Topic Models." Communication of the ACM.
- Couderc, Christophe. 2013: La tragédie Espagnole et son contexte Européen : XVIe – XVIIe siècles, Paris : Presses Sorbonne Nouvelle.
- McCallum, Andrew K. 2002. MALLET: A Machine Learning for Language Toolkit.
- http://mallet.cs.umass.edu.
- 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.
- 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.
- Pouliquen, Bruno, Ralf Steinberger, and Camelia Ignat. 2006. "Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus." arXiv:cs, September. http://arxiv.org/abs/cs/0609059.