

# Que faire avec 740 romans ?

**Exploration einer Sammlung französischer Romane mit Topic  
Modeling**

---

Christof Schöch

Nachwuchsgruppe

"Computergestützte literarische Gattungsstilistik" (CLiGS)

Lehrstuhl für Computerphilologie, Universität Würzburg

***Digital Humanities Kolloquium***

**Trier, 28. Oktober 2015**

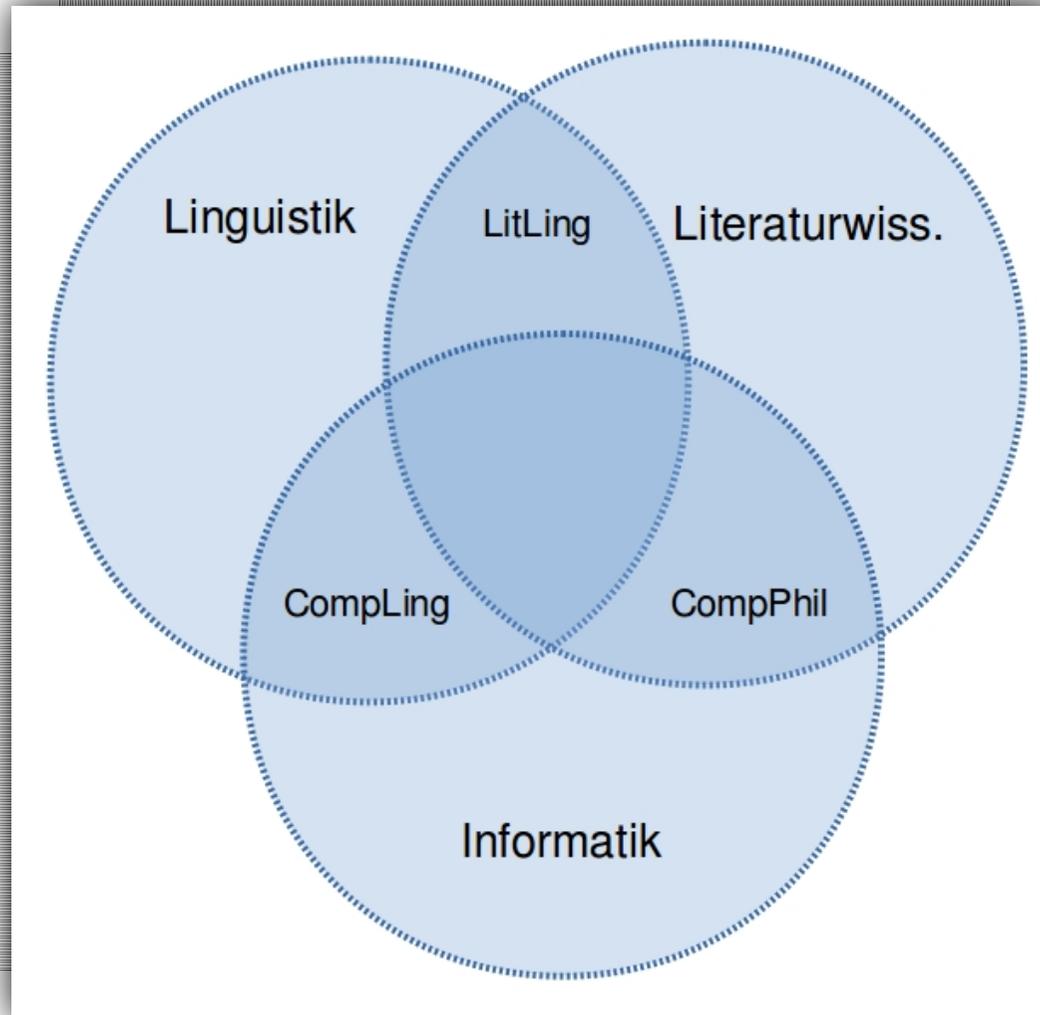
Gefördert vom BMBF - FKZ 01UG1408/1508

# 1. Einleitung

# Überblick

- Einleitung: Kontext und Fragestellungen
- Textsammlung: 740 französische Romane
- Methode: Topic Modeling Workflow
- Ergebnisse: Perspektiven auf die Daten
- Fazit

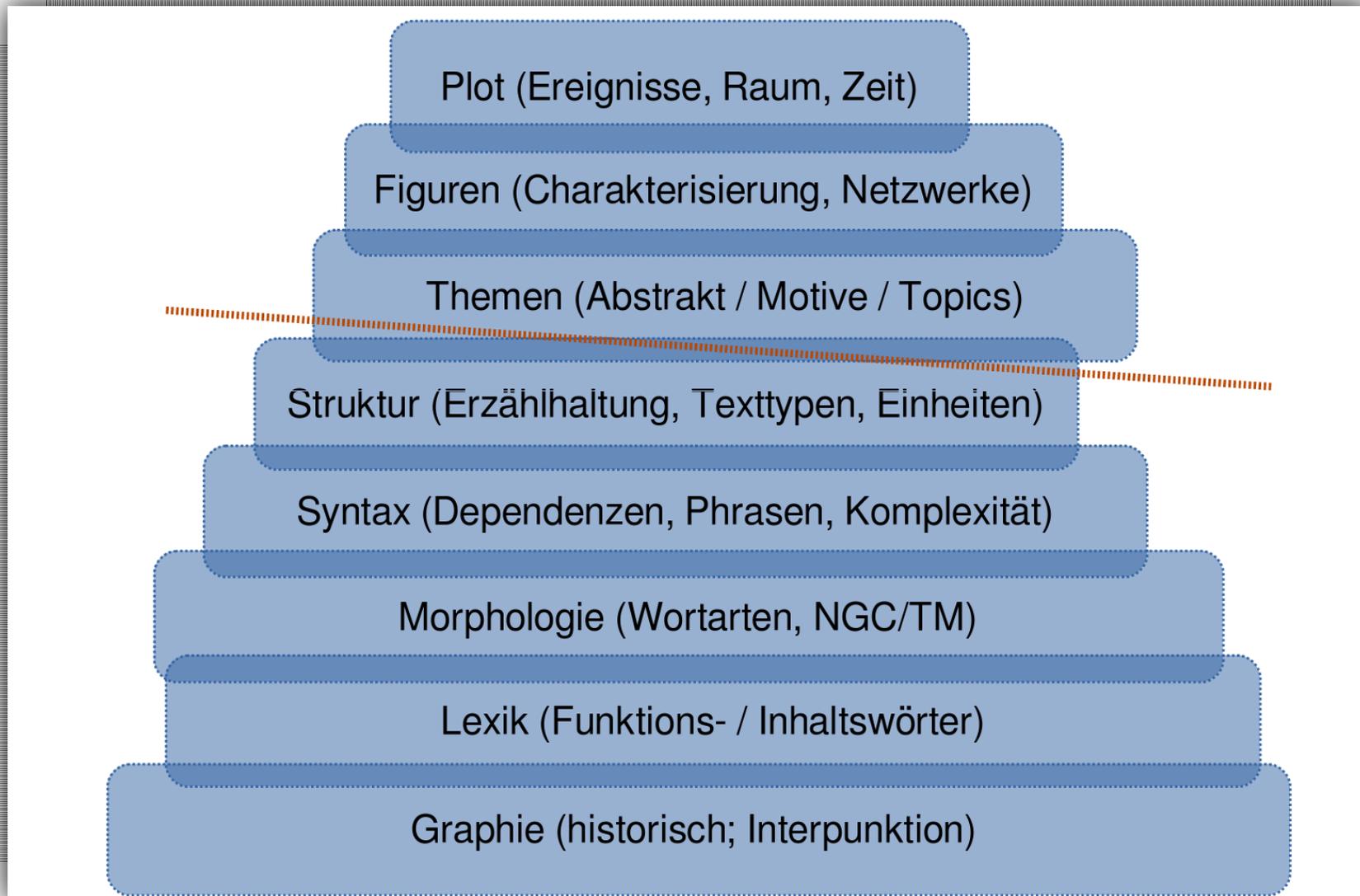
# Kontext Digitale Geisteswissenschaften



# Kontext "CLiGS"

- CLiGS = Computergestützte literarische Gattungsstilistik, [cligs.hypotheses.org](http://cligs.hypotheses.org)
- BMBF-geförderte Nachwuchsgruppe (Romanistik und Informatik), 2014-2019
- Übergeordnete Ziele:
  - Beschreibung von Untergattungen des (französischen und spanischen) Romans und des Theaters auf Grundlage großer Textsammlungen
  - Adaptation von Methoden aus der Informatik für die Literaturwissenschaften
  - Reflexion über das Konzept der Gattungen

# Beschreibungsebenen von Gattung



# Fragestellungen

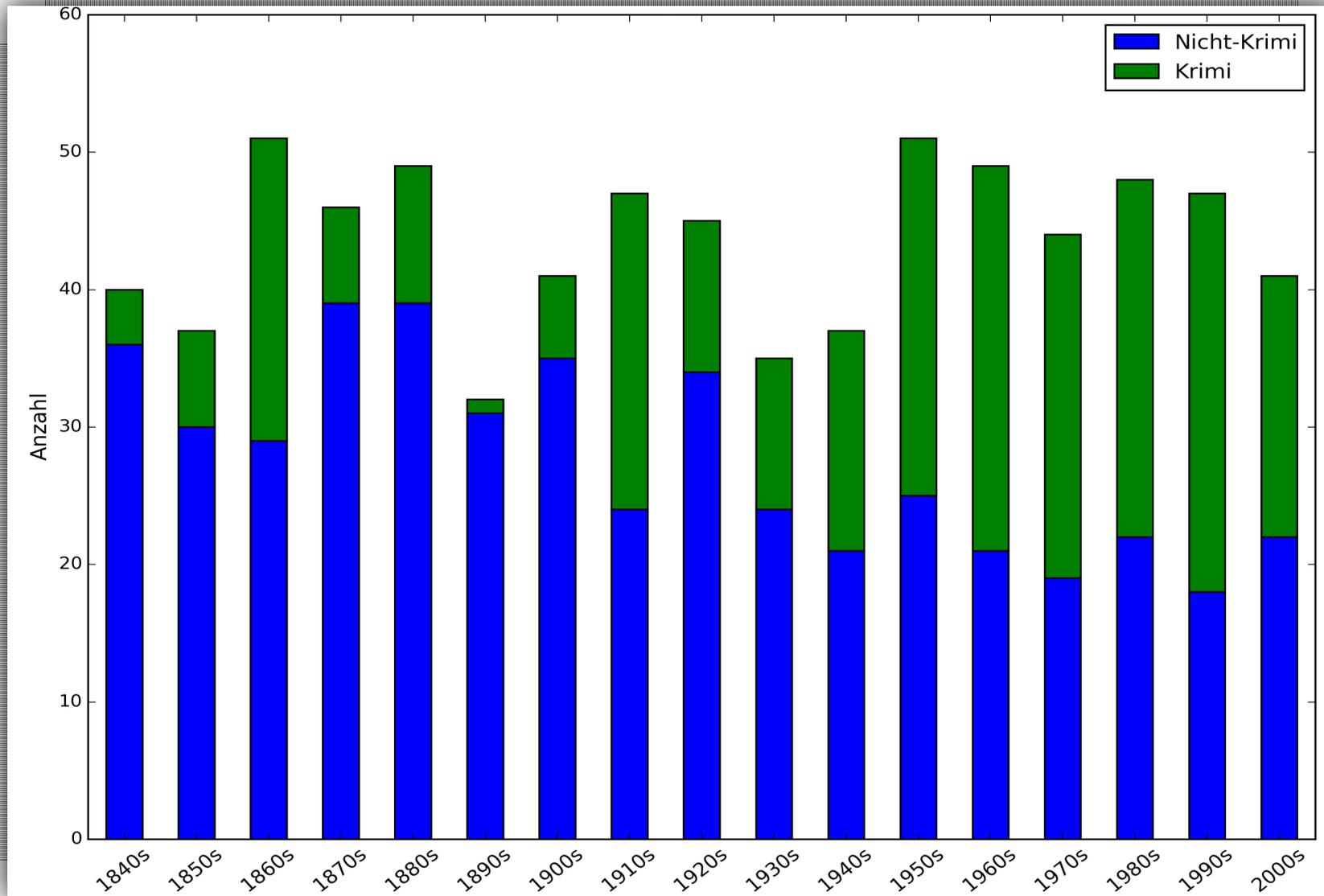
- Worum geht es im französischen Roman des 19. und 20. Jahrhunderts?
- Welche Typen von Topics gibt es überhaupt? (Was sind 'Topics'?)
- Welche Topics sind typisch für den Kriminalroman, welche für Nicht-Kriminalromane?
- Fokus auf den Kriminalroman: Wie gestaltet sich die Binnendifferenzierung der Gattung?
- Erlaubt Topic Modeling einen Zugang zu gattungsrelevanten Informationen?
- Welche Rolle spielen Visualisierungen der Daten für ihre Interpretation?

# 2. Die Textsammlung

# Die Textsammlung

- 740 Romane, davon 271 Kriminalromane und 469 Nicht-Krimis;  
Textmenge: ca. 50 Mio. Tokens / 314 MB Text ; mehr in Vorbereitung
- Publiziert zwischen 1840 und 2009;  
durchschnittlich 41 Romane pro Jahrzehnt
- Liegen mit Metadaten und im TEI-Format vor (Text Encoding Initiative)

# Textsammlung im Überblick



# Metadaten

idno	author-name	gender	title	year	subgenre	subsubgenre	narration	protagonist	setting
rf0021	Daudet	male	PortTarascon	1890	blanche	adventure	heterodiegetic	n.av.	n.av.
rf0623	FevalPP	male	ReineEpees	1852	blanche	adventure	heterodiegetic	n.av.	n.av.
rf0622	FevalPP	male	Quittance2Galerie	1846	blanche	adventure	heterodiegetic	n.av.	n.av.
rf0621	FevalPP	male	Quittance1Heritiere	1846	blanche	adventure	n.av.	n.av.	n.av.
rf0636	FevalPP	male	Louve1	1855	blanche	adventure	heterodiegetic	n.av.	n.av.
rf0615	FevalPP	male	LoupBlanc	1843	blanche	adventure	n.av.	n.av.	rural
rf0610	FevalPP	male	FilsDiable1	1846	blanche	adventure	heterodiegetic	n.av.	n.av.
rf0633	FevalPP	male	Bossu	1857	blanche	adventure	n.av.	n.av.	n.av.
rf0487	Giono	male	HussardToit	1951	blanche	adventure	heterodiegetic	criminal	rural
rf0176	Verne	male	TestamentExcentrique	1899	blanche	adventure	heterodiegetic	n.av.	mixed
rf0731	Barbara	male	AssassinatPontRouge	1855	policier	archaique	n.av.	n.av.	n.av.
rf1263	BoilNarc	male	SermentLupin	1987	policier	archaique	heterodiegetic	criminal	n.av.
rf1262	BoilNarc	male	SecretEunerville	1986	policier	archaique	heterodiegetic	mixed	rural
rf1261	BoilNarc	male	SecondVisageLupin	1986	policier	archaique	heterodiegetic	criminal	n.av.
rf1260	BoilNarc	male	Poudriere	1987	policier	archaique	heterodiegetic	criminal	n.av.
rf1184	Boisgobey	male	RubisOngle	1886	policier	archaique	heterodiegetic	victime	urban
rf1003	Boisgobey	male	DoubleBlancB	1889	policier	archaique	heterodiegetic	criminal	urban
rf0628	FevalPP	male	MysteresLondres3	1843	policier	archaique	heterodiegetic	criminal	urban
rf0627	FevalPP	male	MysteresLondres2	1843	policier	archaique	heterodiegetic	criminal	n.av.
rf0614	FevalPP	male	HommeSansBras	1881	policier	archaique	heterodiegetic	victim	n.av.
rf0607	FevalPP	male	ErrantsNuit	1857	policier	archaique	heterodiegetic	victim	rural
rf1159	Gaboriau	male	MonsieurLecoq2	1869	policier	archaique	heterodiegetic	detective	urban
rf1158	Gaboriau	male	MonsieurLecoq1	1869	policier	archaique	heterodiegetic	detective	urban
rf1155	Gaboriau	male	CrimeOrcival	1866	policier	archaique	heterodiegetic	detective	rural
rf1152	Gaboriau	male	CordeCou2	1873	policier	archaique	heterodiegetic	suspect	rural
rf1151	Gaboriau	male	CordeCou1	1873	policier	archaique	heterodiegetic	suspect	rural
rf1150	Gaboriau	male	AmoursEmpoisonneuse	1881	policier	archaique	heterodiegetic	criminal	n.av.
rf0361	Galopin	male	MemoiresCambrioleur	1922	policier	archaique	homodiegetic	criminal	urban

# **3. Methode: Topic Modeling**

# Topic Modeling

## Topics

gene 0.04  
dna 0.02  
genetic 0.01  
...

life 0.02  
evolve 0.01  
organism 0.01  
...

brain 0.04  
neuron 0.02  
nerve 0.01  
...

data 0.02  
number 0.02  
computer 0.01  
...

## Documents

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those predictions

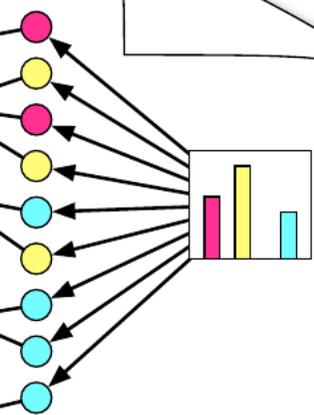
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

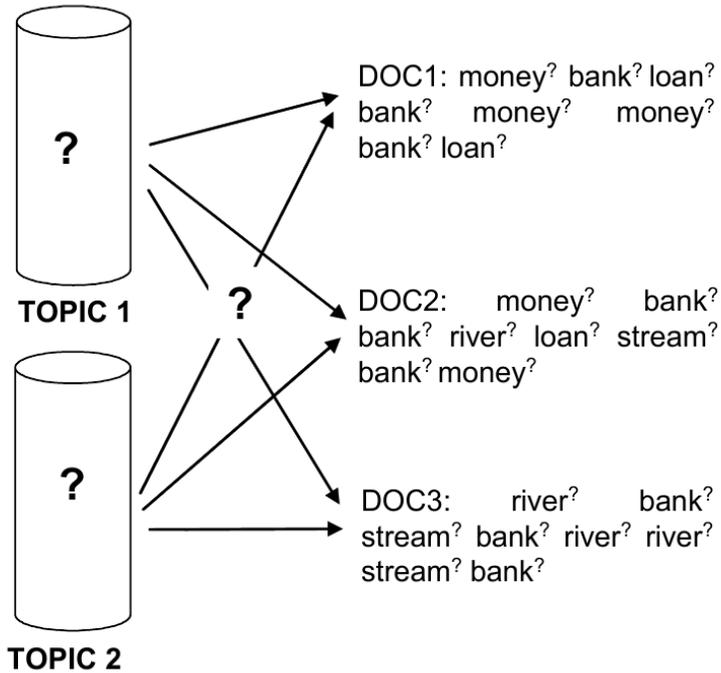
## Topic proportions and assignments



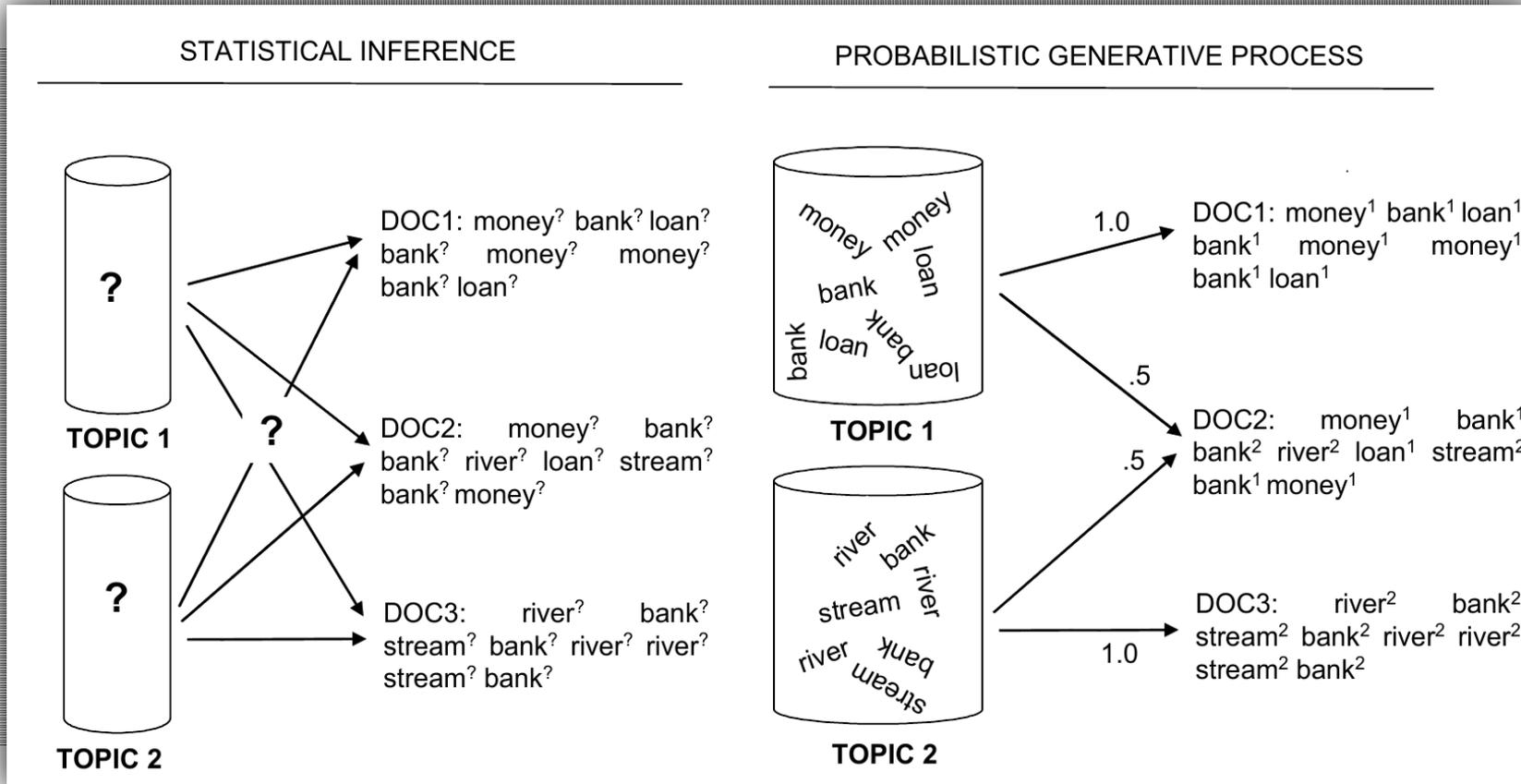
# Topic Modeling

STATISTICAL INFERENCE

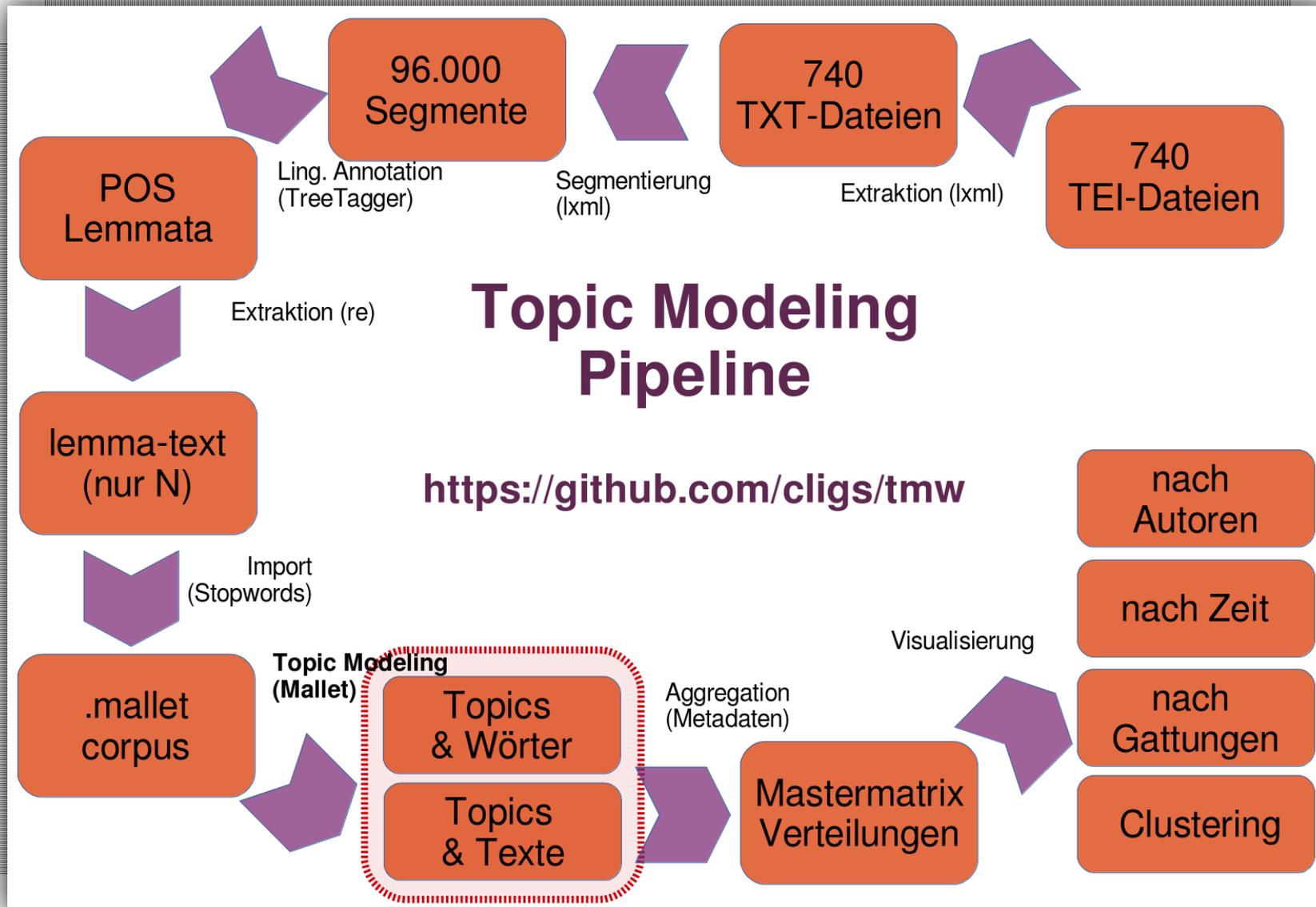
---



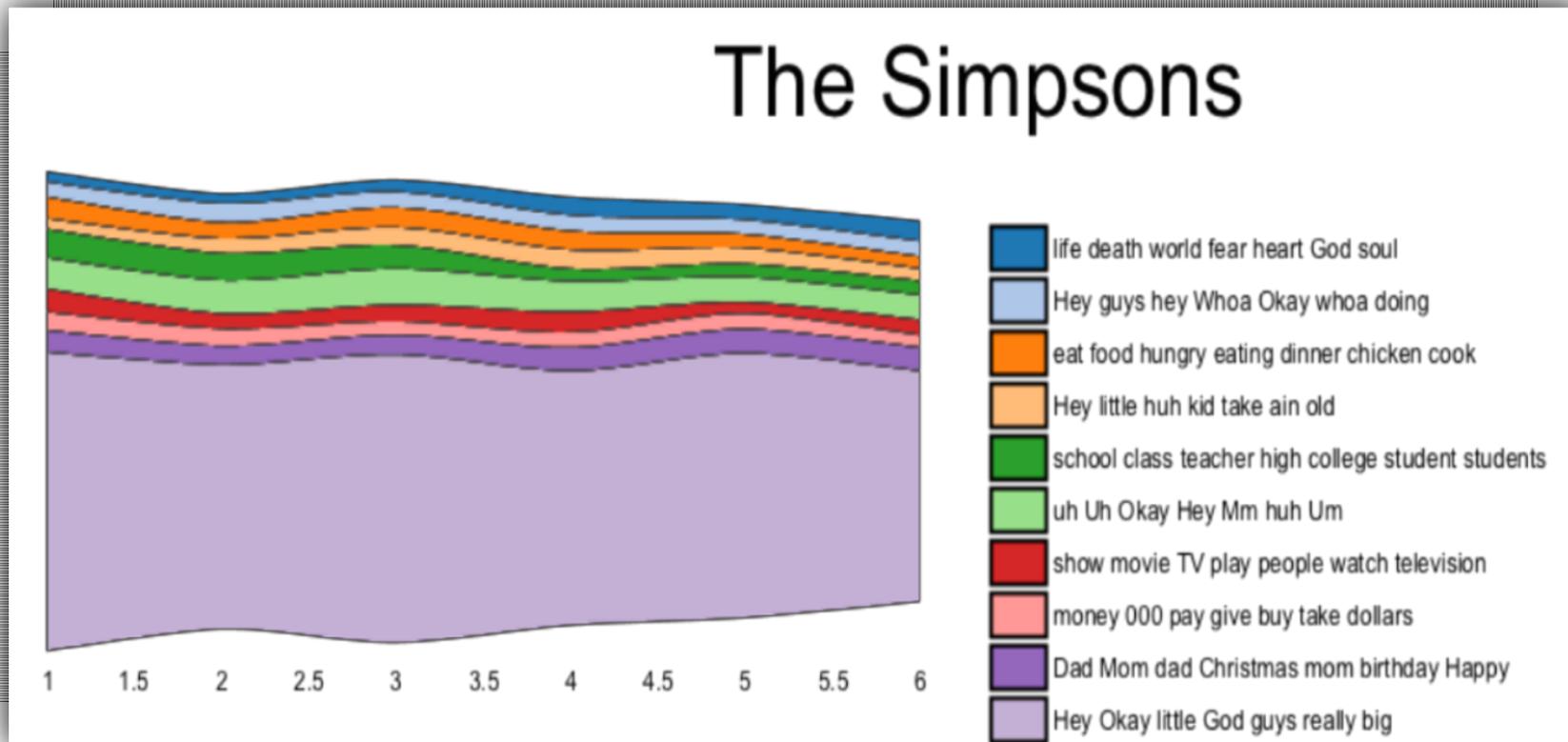
# Topic Modeling



# tmw (Python-Skripte)



# Topic Modeling: Anwendungen



<http://sappingattention.blogspot.com/> (Ben Schmidt)

# Topic Modeling: Anwendungen

The screenshot shows the 'Signs @ 40' website's topic modeling interface. The main header features the 'Signs @ 40 1975 2014' logo and navigation links: Home, Topic Model, Curated Contents, Cotation Network, Cover Art, and Editorial Comments. A left sidebar contains navigation options: Overviews, Topic (selected), Article, Word, Bibliography, Word index, Interpreting the model, Settings, and Getting started.

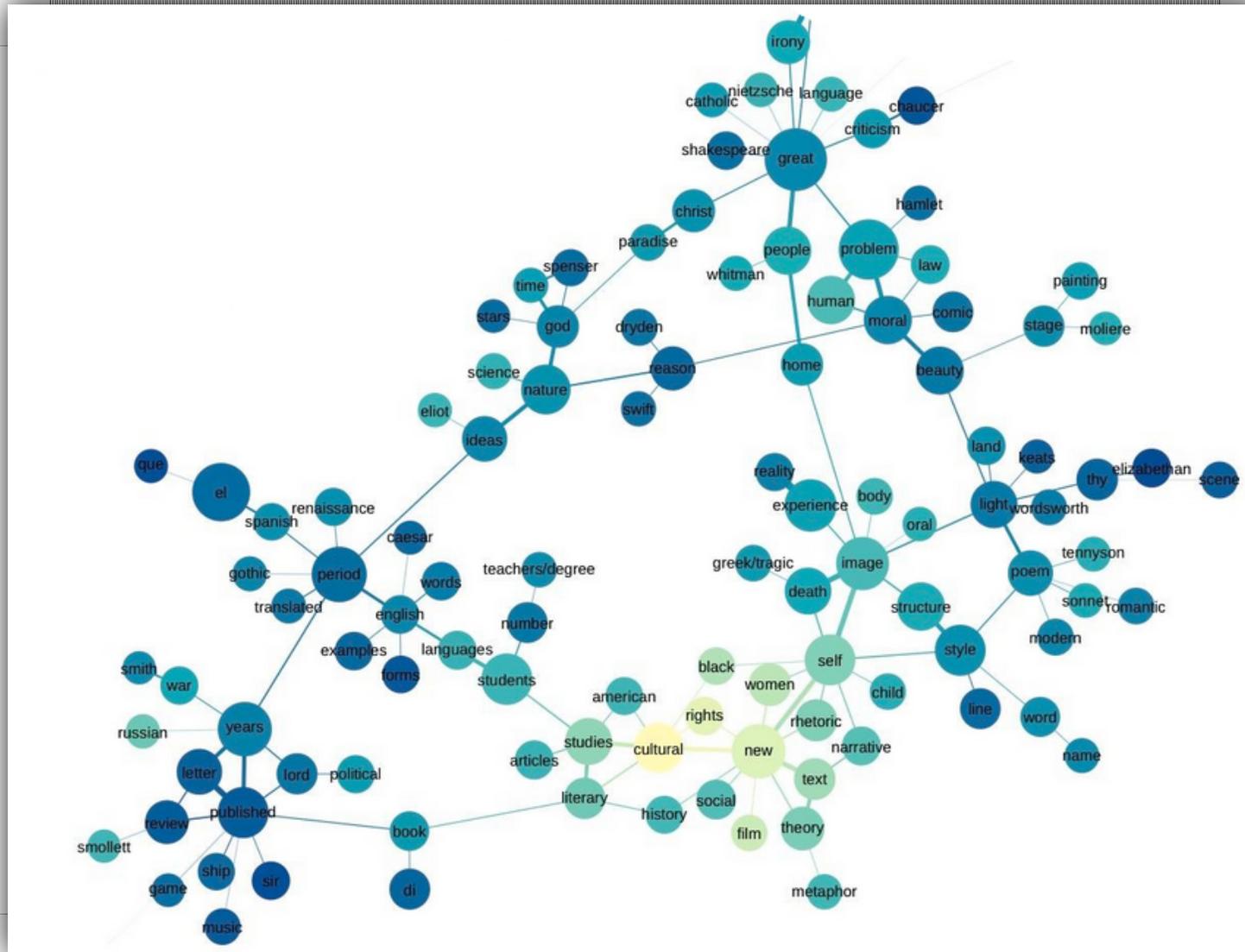
The main content area is titled 'Bodies' and includes a note: 'Joanna Kempner discusses this topic in her comment.' Below this, there are three main sections:

- Top words:** A horizontal bar chart showing the weight of various words. The most prominent words are 'body' (approx. 5.0%), 'subject' (approx. 4.5%), 'bodies' (approx. 4.0%), 'space' (approx. 3.5%), 'time' (approx. 3.0%), 'history' (approx. 2.5%), 'object' (approx. 2.0%), 'relation' (approx. 1.5%), 'place' (approx. 1.0%), 'desire' (approx. 0.5%), 'butler' (approx. 0.5%), 'logic' (approx. 0.5%), 'future' (approx. 0.5%), 'symbolic' (approx. 0.5%), 'bodily' (approx. 0.5%), 'objects' (approx. 0.5%), 'representation' (approx. 0.5%), 'site' (approx. 0.5%), 'moment' (approx. 0.5%), 'past' (approx. 0.5%), 'matter' (approx. 0.5%), 'act' (approx. 0.5%), 'material' (approx. 0.5%), 'imaginary' (approx. 0.5%), 'mode' (approx. 0.5%), and 'embodied' (approx. 0.5%).
- Yearly proportion of words in topic:** A bar chart showing the percentage of the corpus in total for each year from 1975 to 2014. The chart shows a general upward trend, with a significant peak around 2003 (approx. 4.5%) and another peak around 2014 (approx. 3.5%).
- Top articles:** A table listing the top articles for this topic, including the article title, journal issue, and the percentage of tokens.

Article	% Tokens
Bray, Abigail, and Claire Colebrook. "The Haunted Flesh: Corporeal Feminism and the Politics of (Dis)Embodiment." <i>Signs</i> 24, no. 1 (Autumn 1998): 35–67.	41.0%
Barad, Karen. "Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter." <i>Signs</i> 28, no. 3 (Spring 2003): 801–831.	37.1%
Söderbäck, Fanny. "Revolutionary Time: Revolt as Temporal Return." <i>Signs</i> 37, no. 2 (Winter 2012): 301–324.	36.3%
Grosz, Elizabeth. "Histories of a Feminist Future." <i>Signs</i> 25, no. 4 (Summer 2000): 1017–1021.	36.0%
Meijer, Irene Costera, and Baukje Prins. "How Bodies Come to Matter: An Interview with Judith Butler." <i>Signs</i> 23, no. 2 (Winter 1998): 275–286.	28.1%
Eng, David L. "Melancholia in the Late Twentieth Century." <i>Signs</i> 25, no. 4 (Summer 2000): 1275–1281.	27.7%
Jones, Amelia. "The 'Eternal Return': Self-Portrait Photography as a Technology of Embodiment." <i>Signs</i> 27, no. 4 (Summer 2002): 947–978.	26.2%

<http://signsat40.signsjournal.org/topic-model/> (Andrew Goldstone)

# Topic Modeling: Anwendungen



<http://pmla.site44.com/> (Ted Underwood)

# **4. Perspektiven auf die Gesamtsammlung**

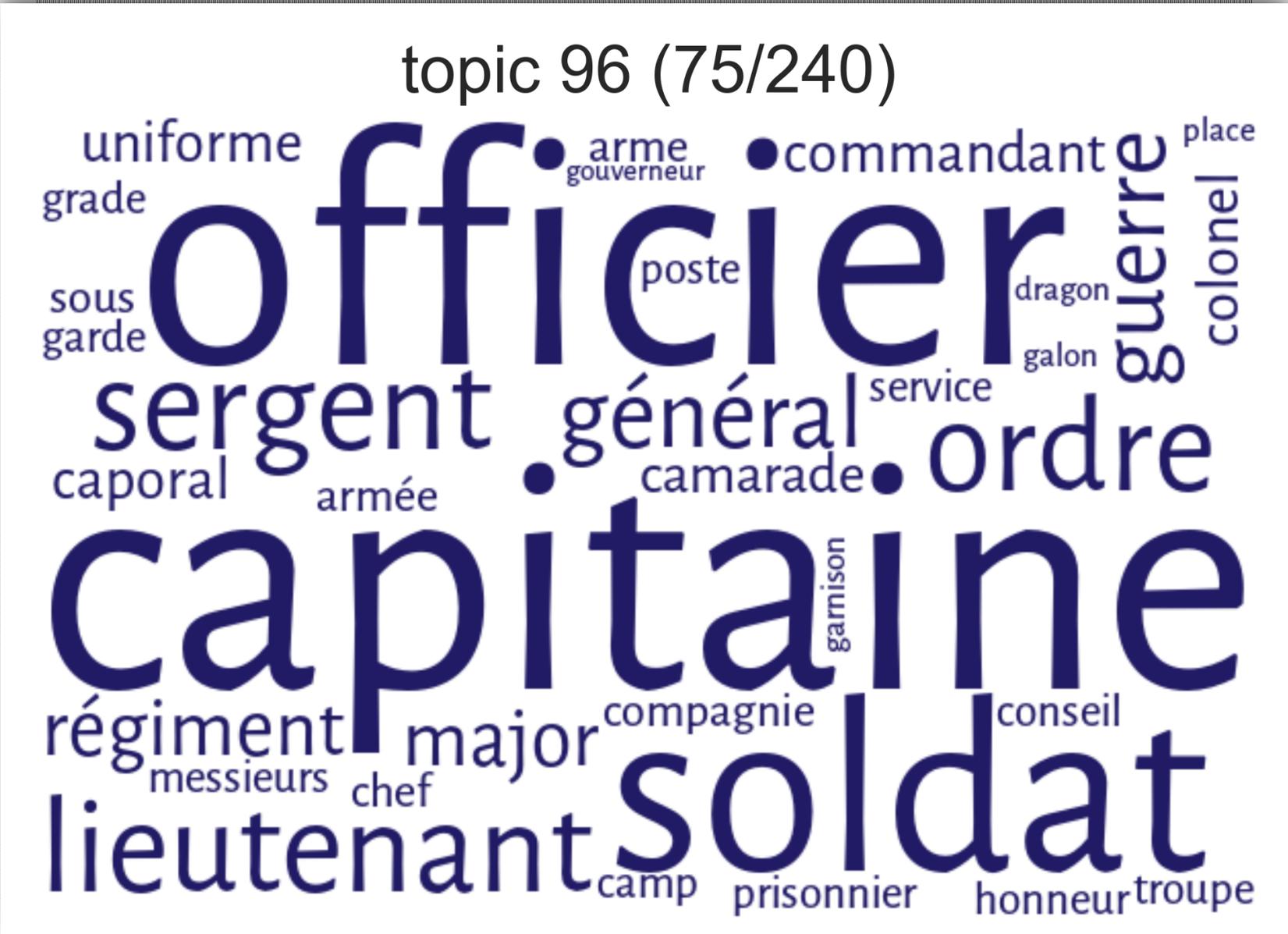
# Überblick

- Typen von Topics
  - nach Semantik
  - nach Struktur
  - nach Verteilung
- Ähnlichkeit von Topics
- Topics über die Zeit
- Romane nach Topics

# Typen von Topics

# Personal

topic 96 (75/240)



Offizier, Hauptmann, Soldat

# Personal

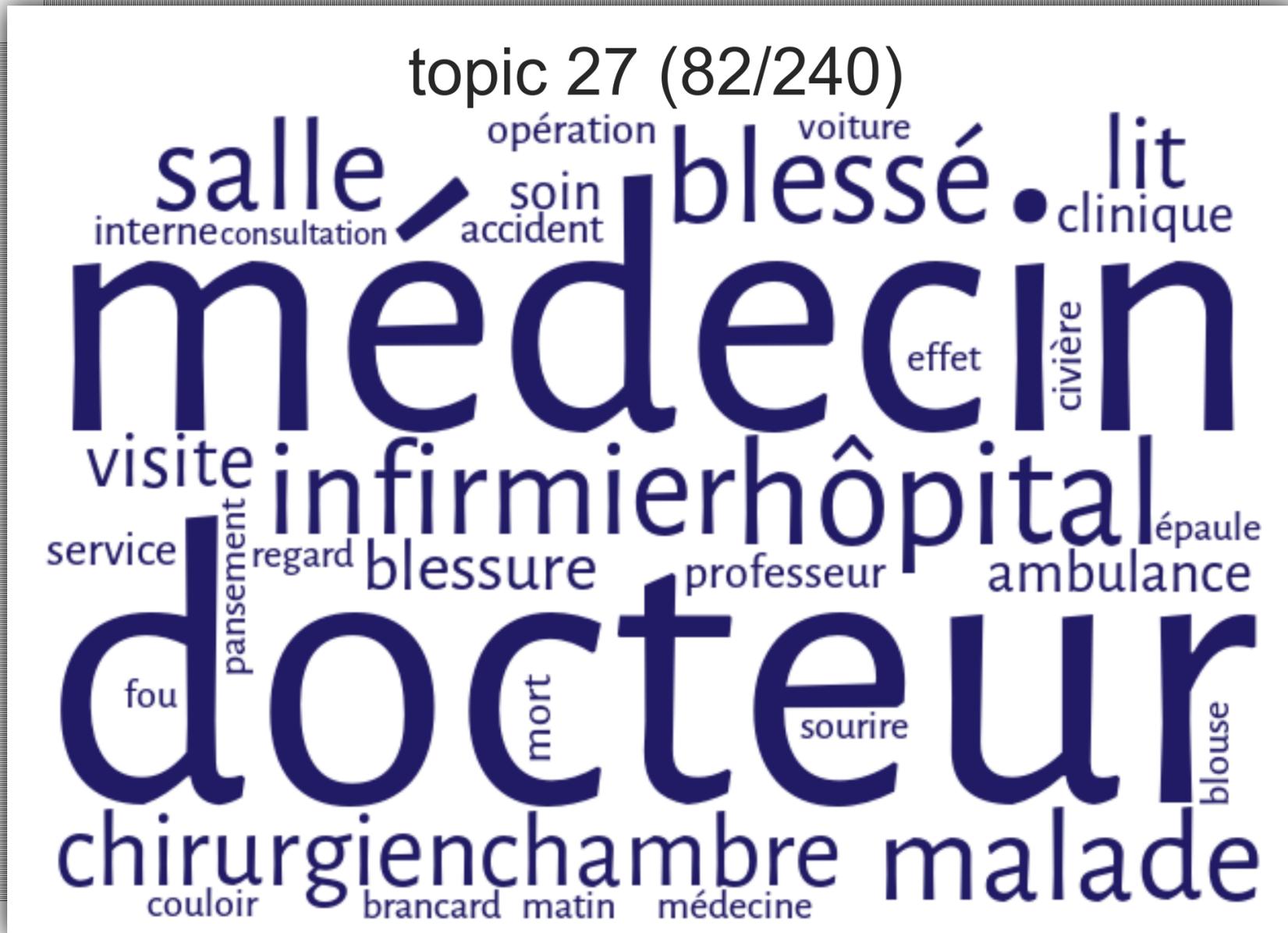
topic 107 (56/240)







# Handlungsmotiv







# Künste

topic 125 (104/240)



Dichter, Werk, Autor, Schriftsteller

# Meta-Topics

topic 144 (221/240)



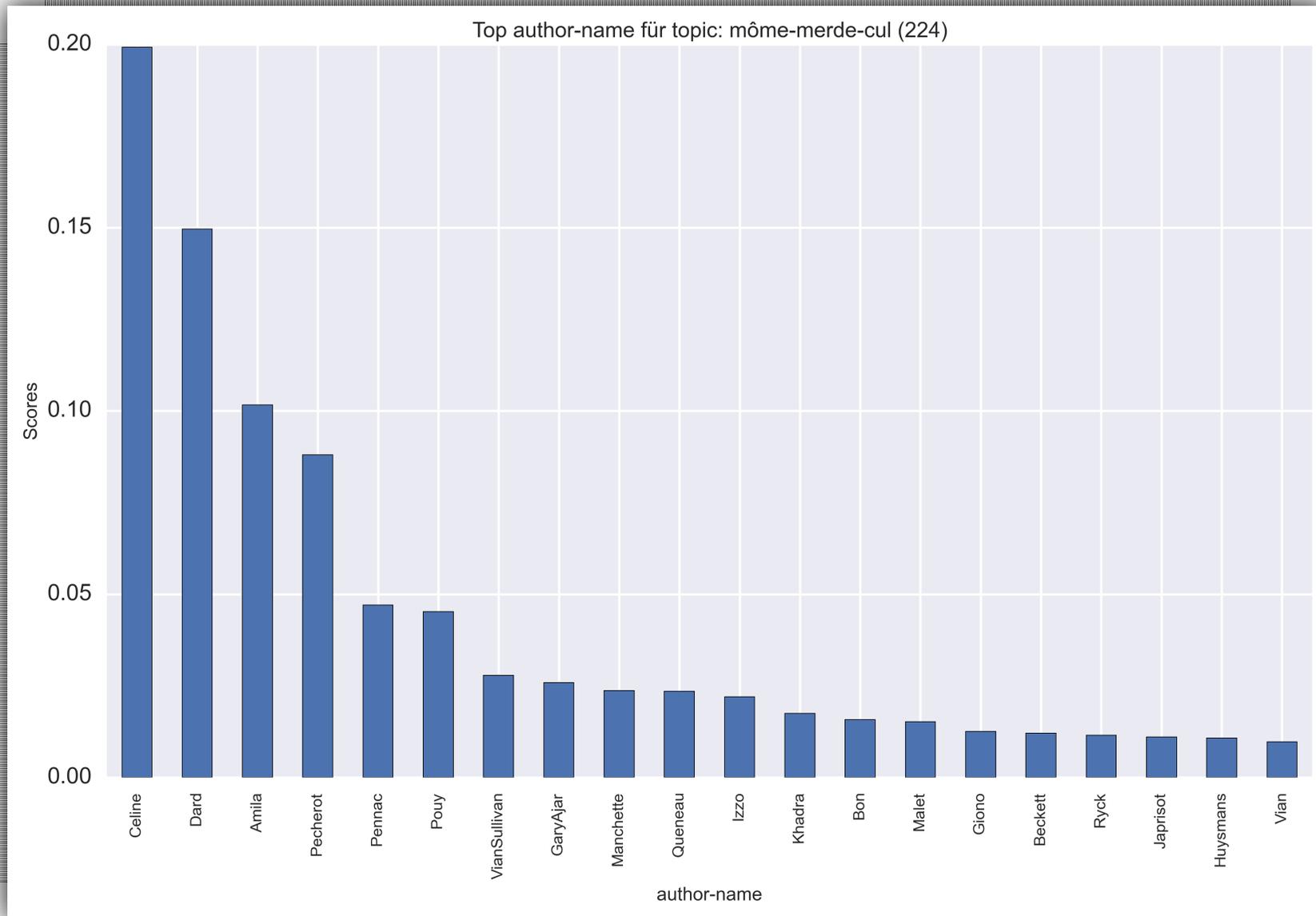
# Meta-Topics

topic 219 (32/240)



Geschichte, Erzählung, Detail

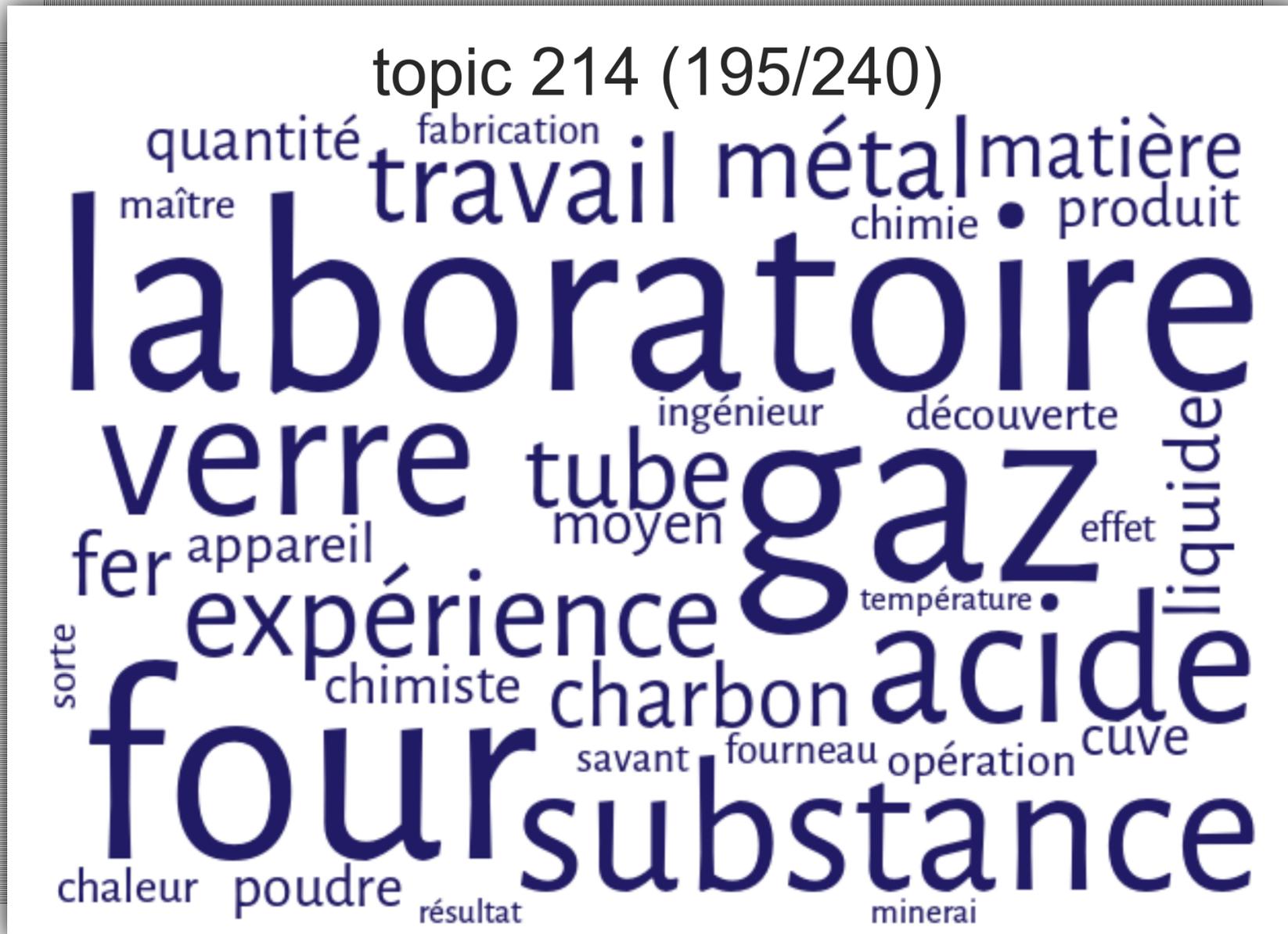
# Register (umgangssprachlich)



Kleine/r, Hintern, Sch\*\*ße



# Themen



Ofen, Gaz, Labor, Substanz

# Semantik: Polysemie

15 **billet** franc argent poche banque portefeuille pièce papier liasse monnaie  
chèque dollar caisse numéro besoin paquet compte table valeur louis carnet  
carte geste caissier vol million garçon minute voleur attention ...

59 train gare wagon quai compartiment voyageur voie station fer voyage valise  
rail employé minute chemin voiture couloir ligne vitre place tunnel départ  
locomotive convoi chef portière banquette **billet** machine classe ...

108 lettre enveloppe papier écriture adresse poste ligne matin courrier **billet**  
doute réponse cachet poche soir correspondance bureau lecture signature  
timbre mois table facteur pli cœur message rendez-vous paquet ...

(frz. billet = Geldschein, Fahrschein, kurzer Brief)

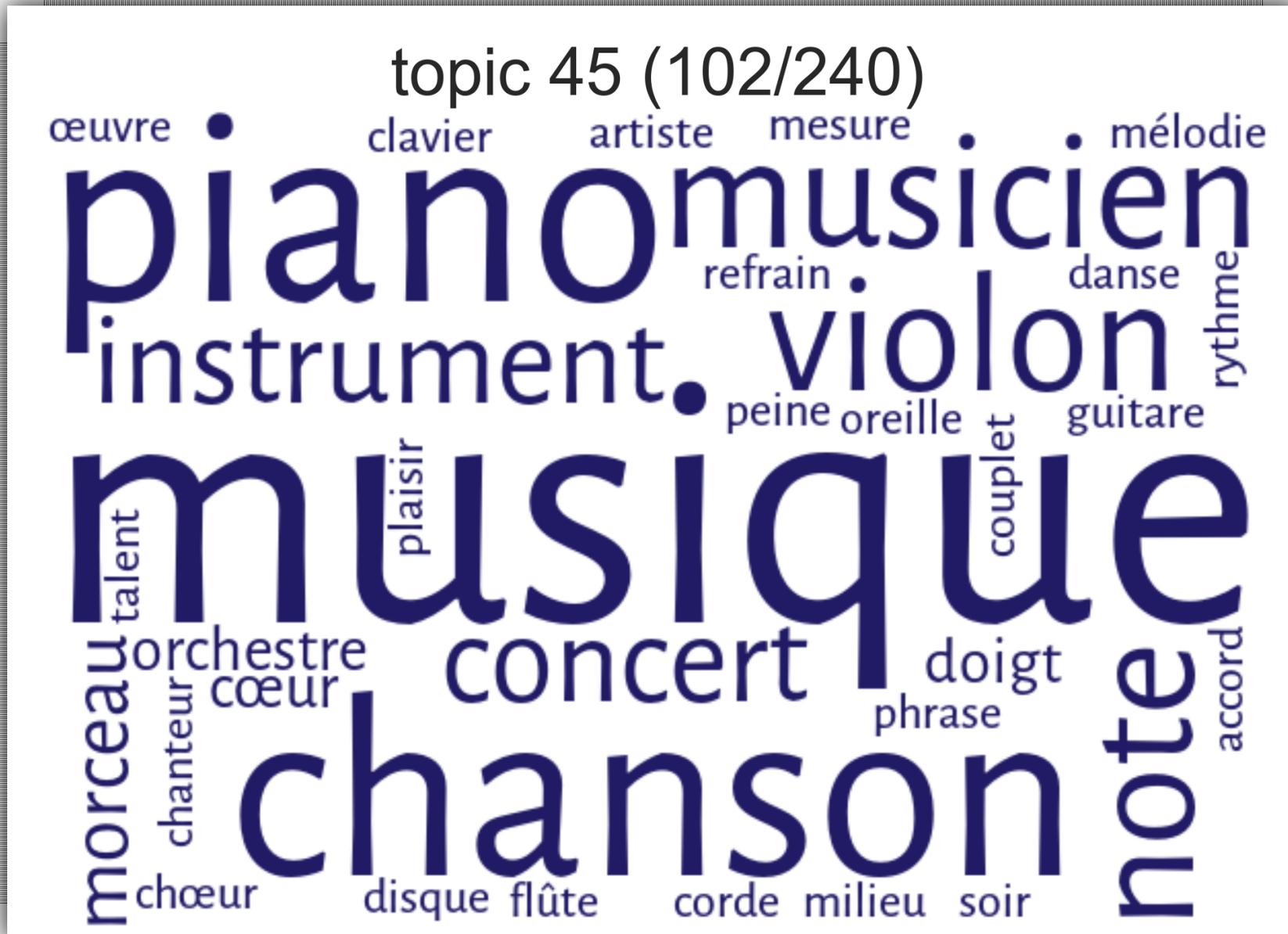


# Struktur: steil abfallend



Schlüssel, Tresor, Schloss, Schublade

# Struktur: flach abfallend



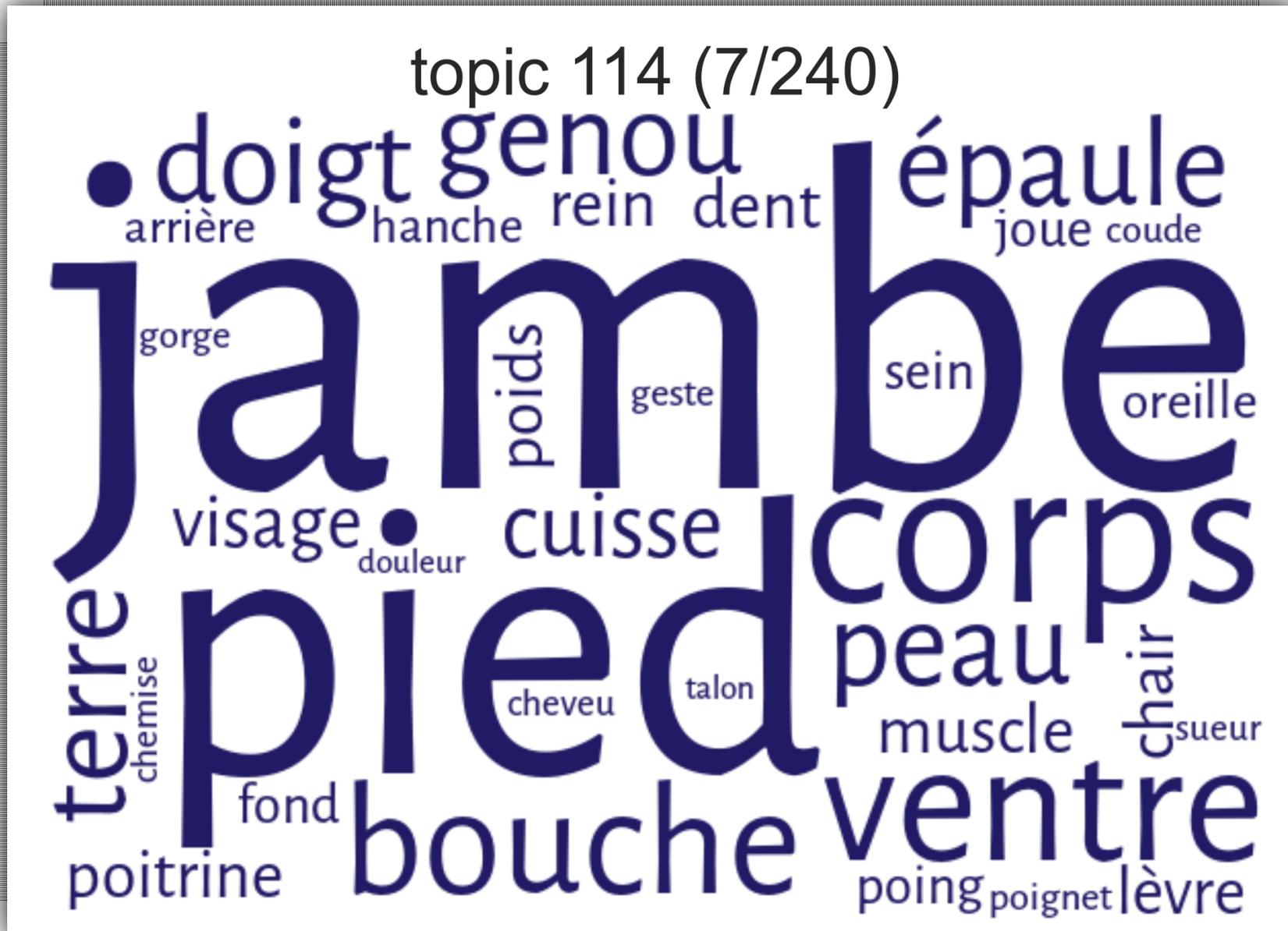
# Struktur: flach abfallend



# Verteilung: breit

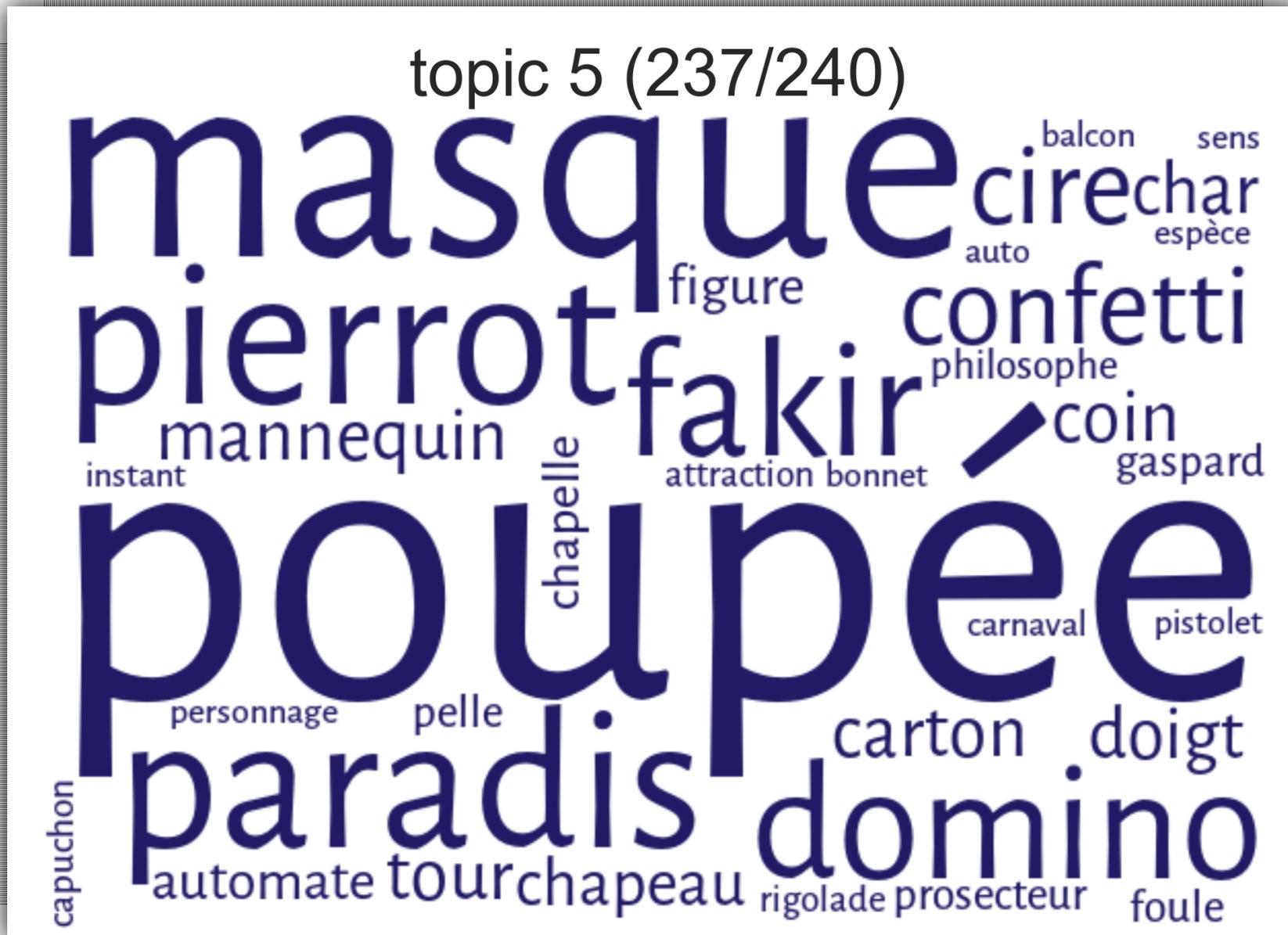


# Verteilung: breit



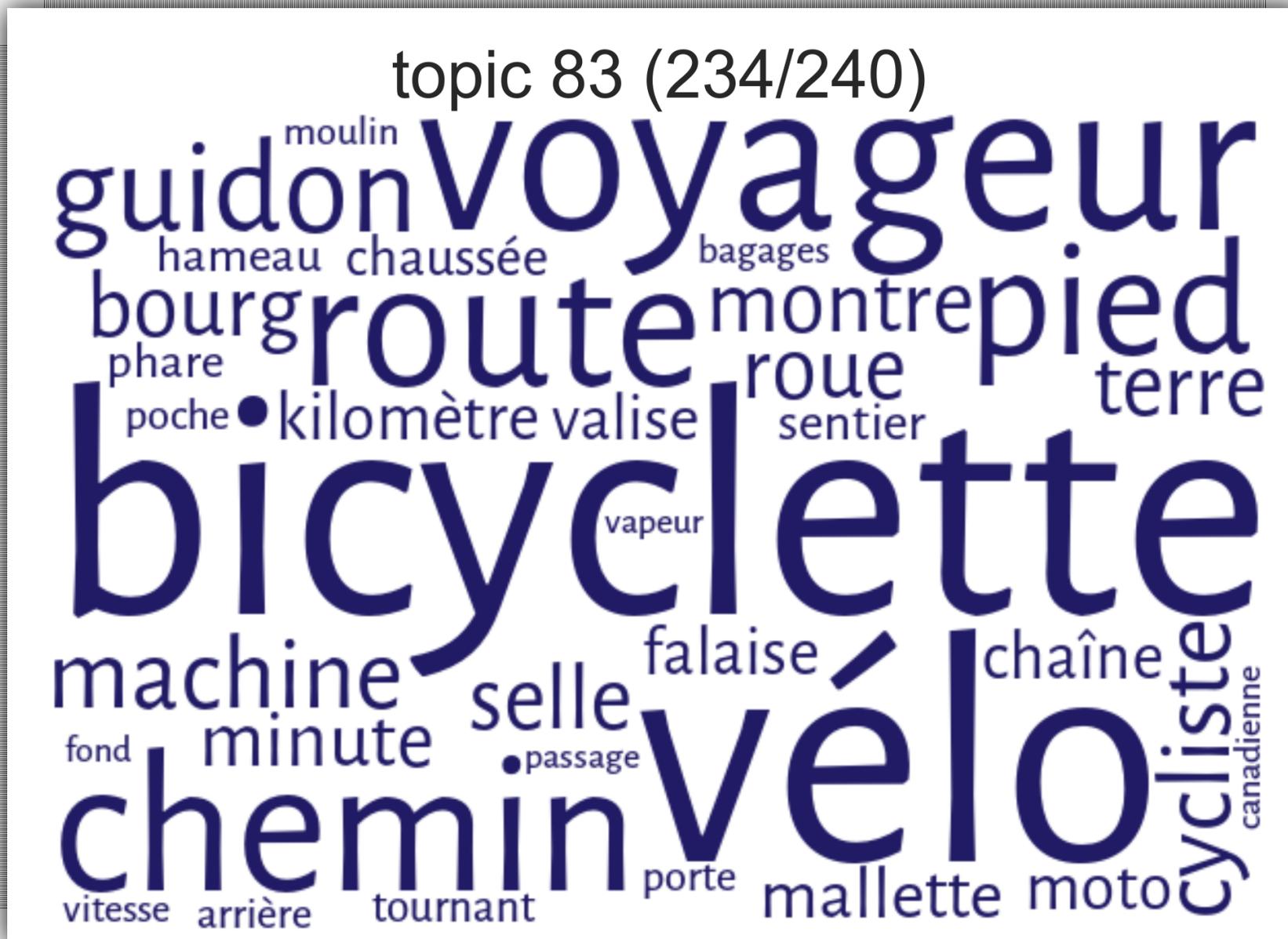
Bein, Fuß, Körper

# Verteilung: spezifisch

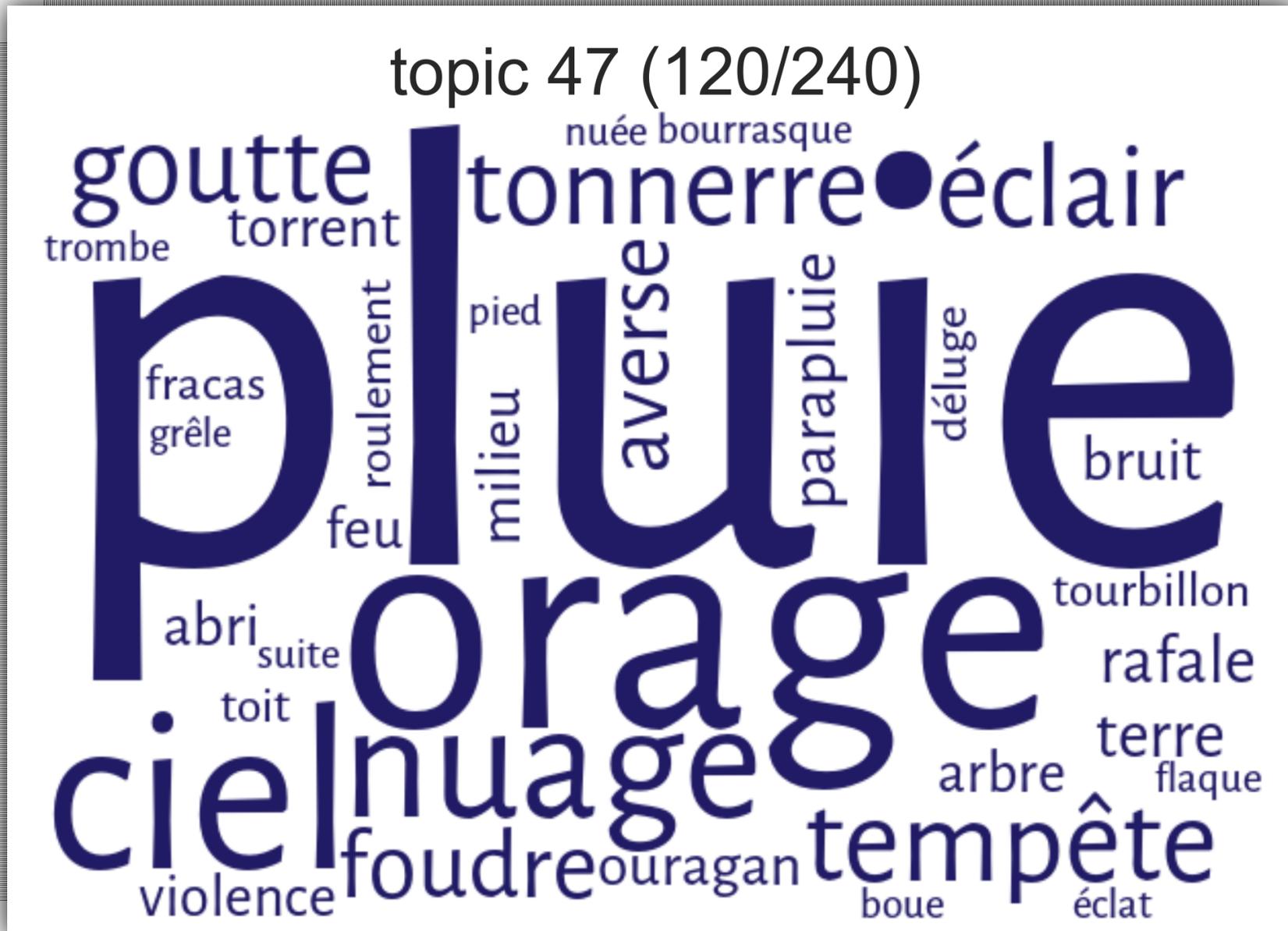


Puppe, Maske, Pierrot (Clown)

# Verteilung: spezifisch

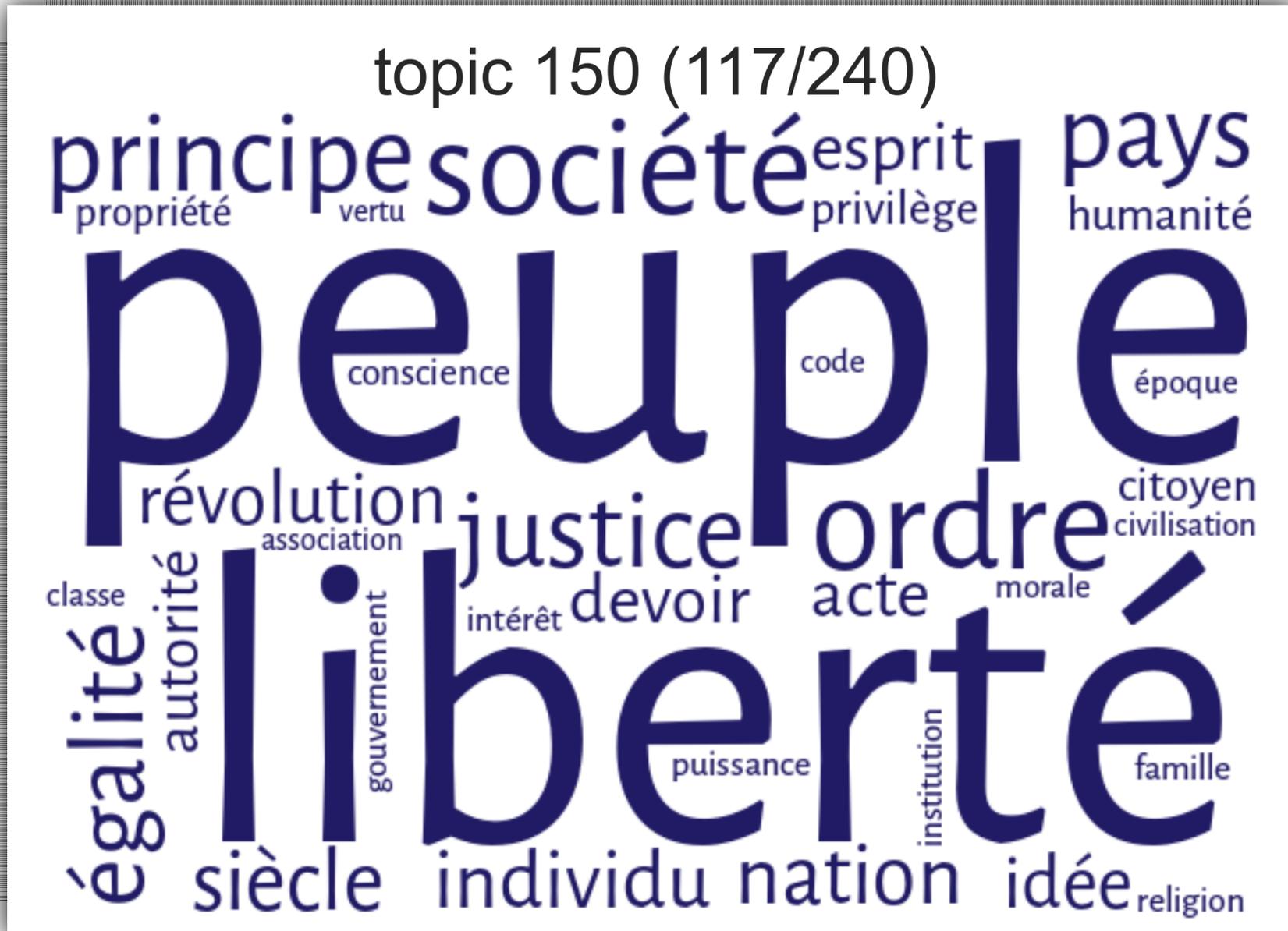


# Verteilung: mittel



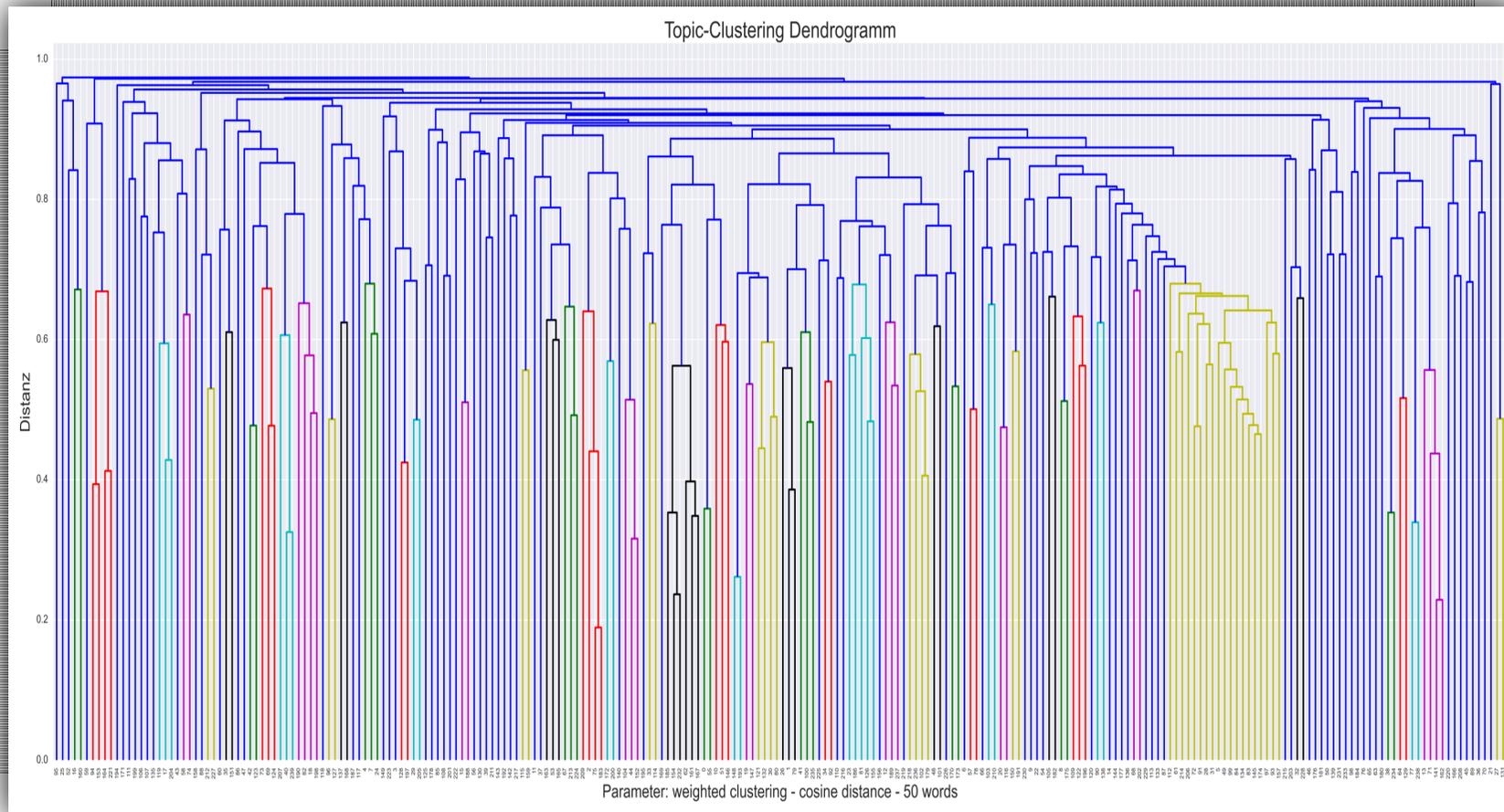
Regen, Gewitter, Himmel

# Verteilung: mittel



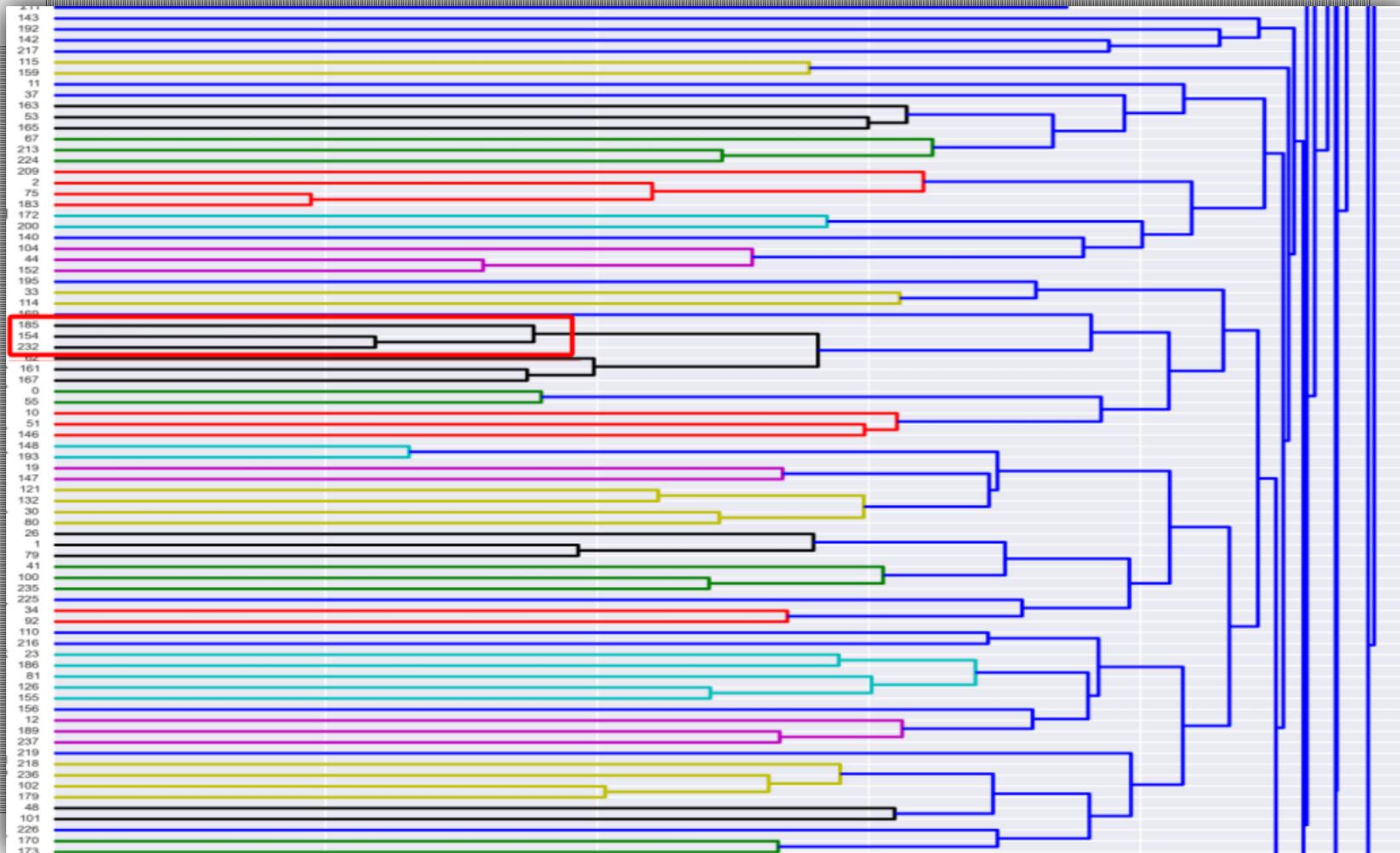
# Ähnlichkeit von Topics

# Ähnlichkeit: Clustering



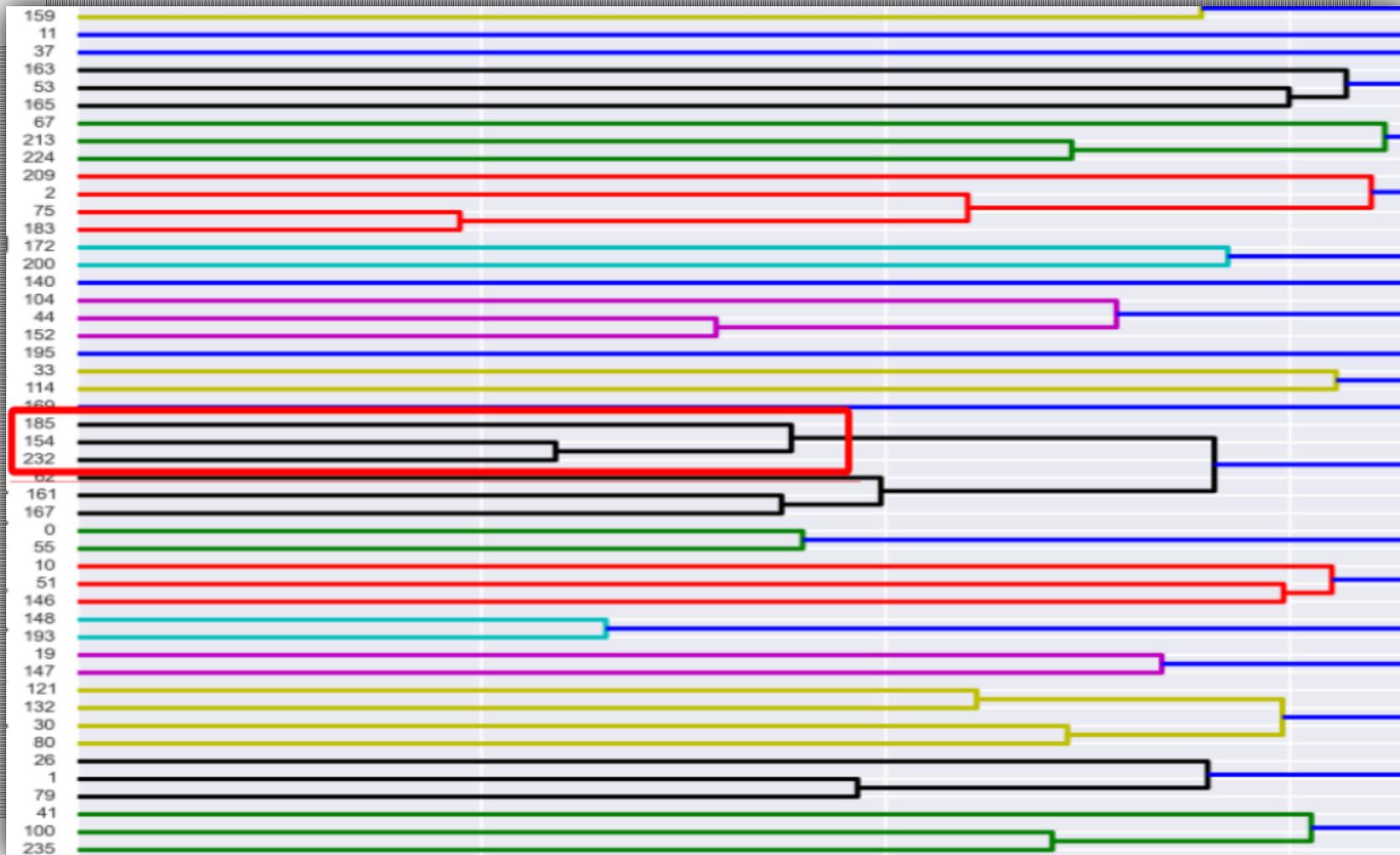
(50 Wörter, Cosine-Distanz, weighted-Clustering)

# Clustering (Detail)



(50 Wörter, Cosine-Distanz, weighted-Clustering)

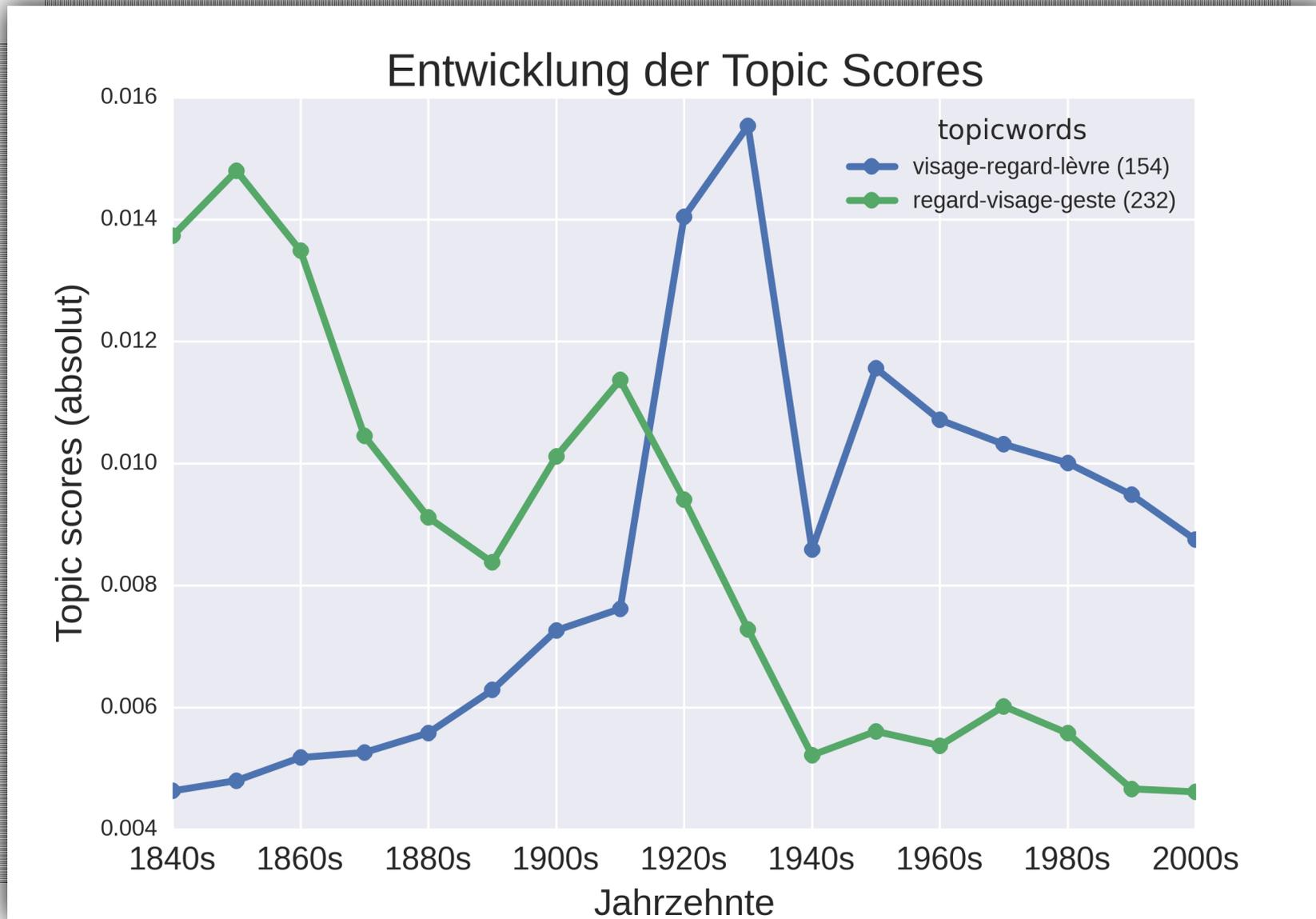
# Clustering (Detail)



(50 Wörter, Cosine-Distanz, weighted-Clustering)

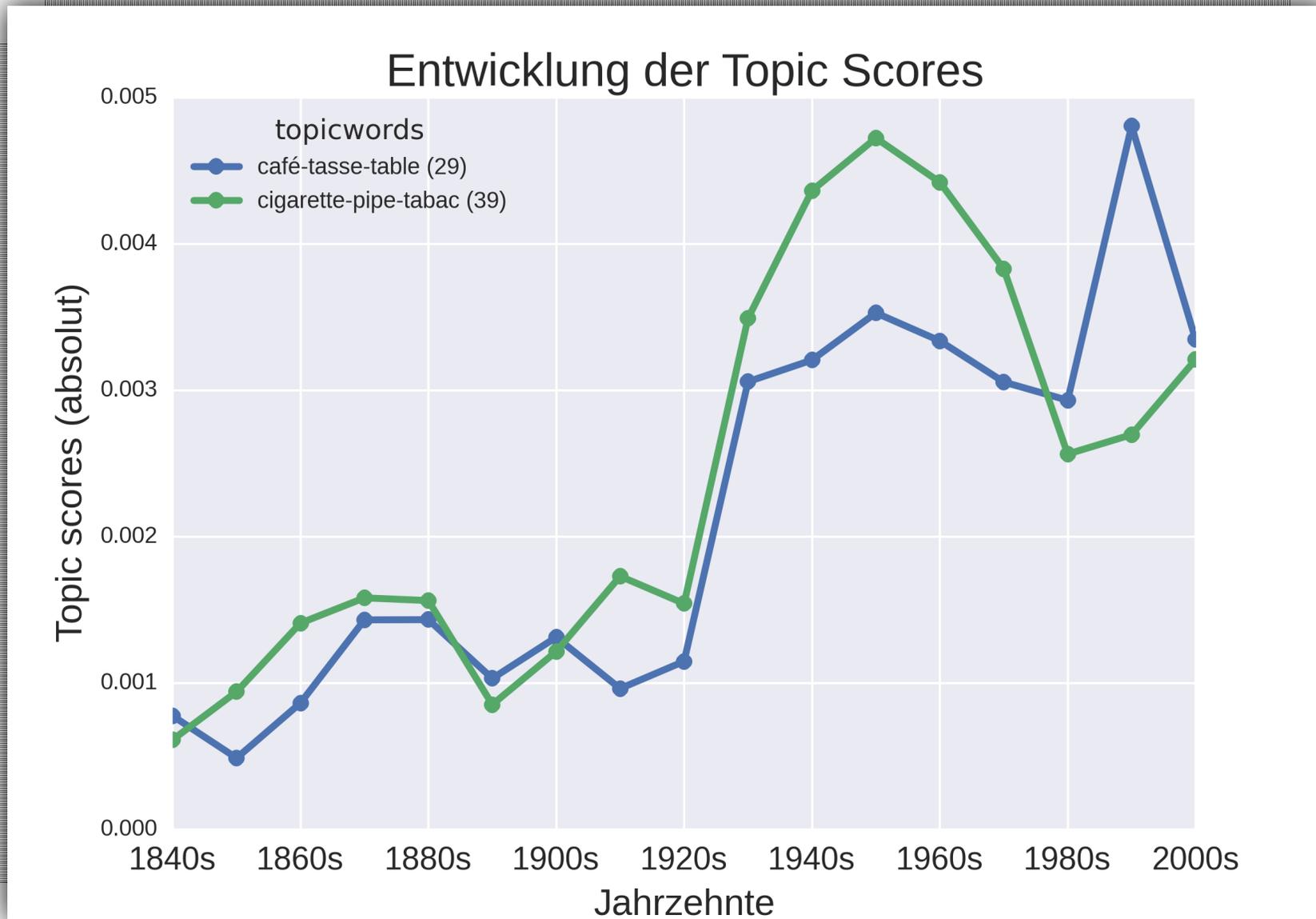
# Topics über die Zeit

# Sich ablösende Topics



Gesicht, Blick, Lippe / Blick, Gesicht, Geste

# Analoge Topics: Korrelation



Kaffee, Tasse, Tisch / Zigarette, Pfeife, Tabak

# **5. Perspektiven auf die Kriminalromane**

# Überblick

- Krimi-Topics
  - inhaltlich
  - statistisch
- Topics nach Untergattungen
  - Überblick
  - Gruppierung nach Topic-Ähnlichkeit

# **Krimi-Topics: inhaltlich und statistisch**

# Inhaltlich typisch



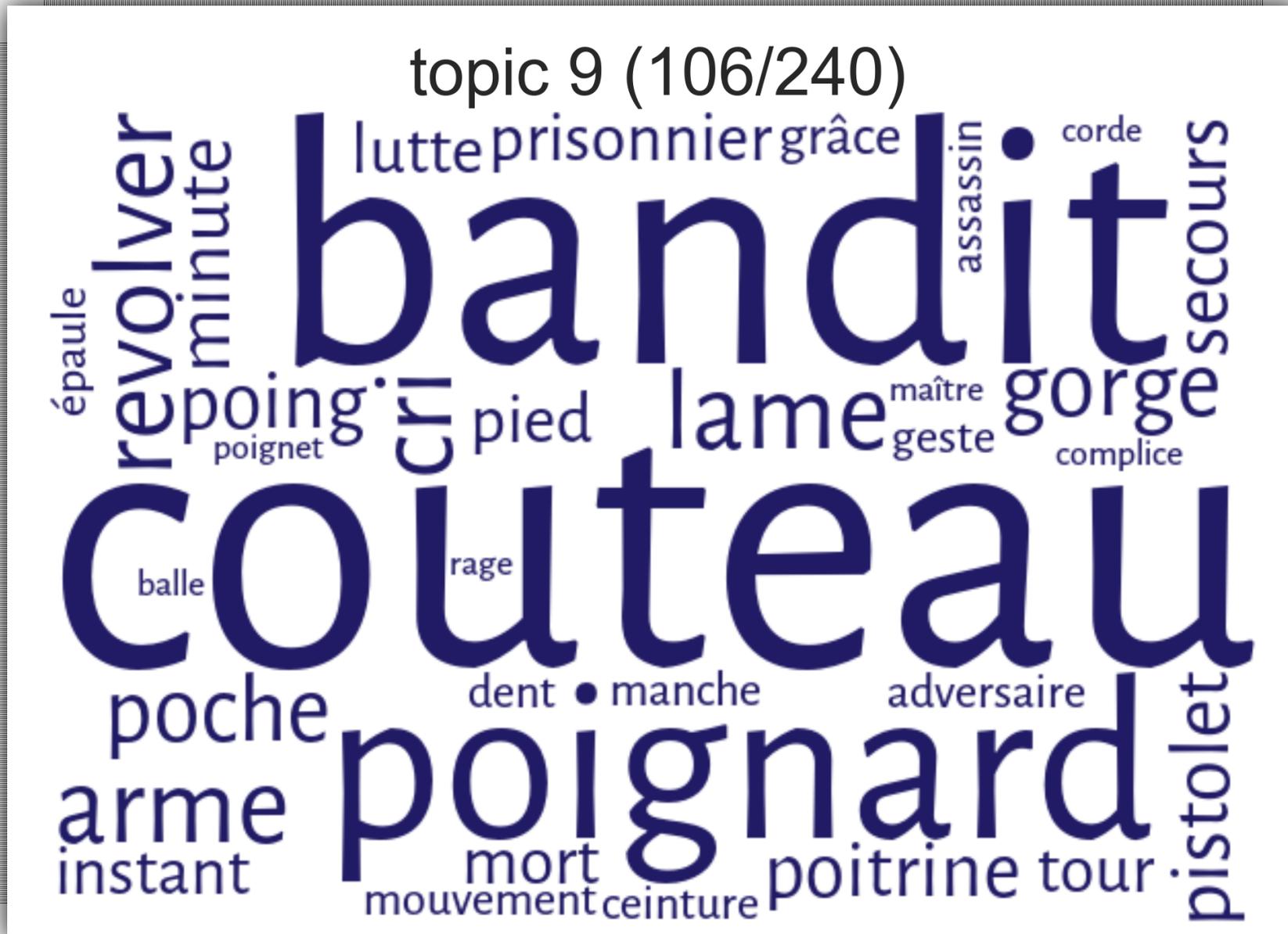
Polizei, Kommissar, Agent / Beamter

# Inhaltlich typisch



Verbrechen, Justiz/Gerechtigkeit, Tod

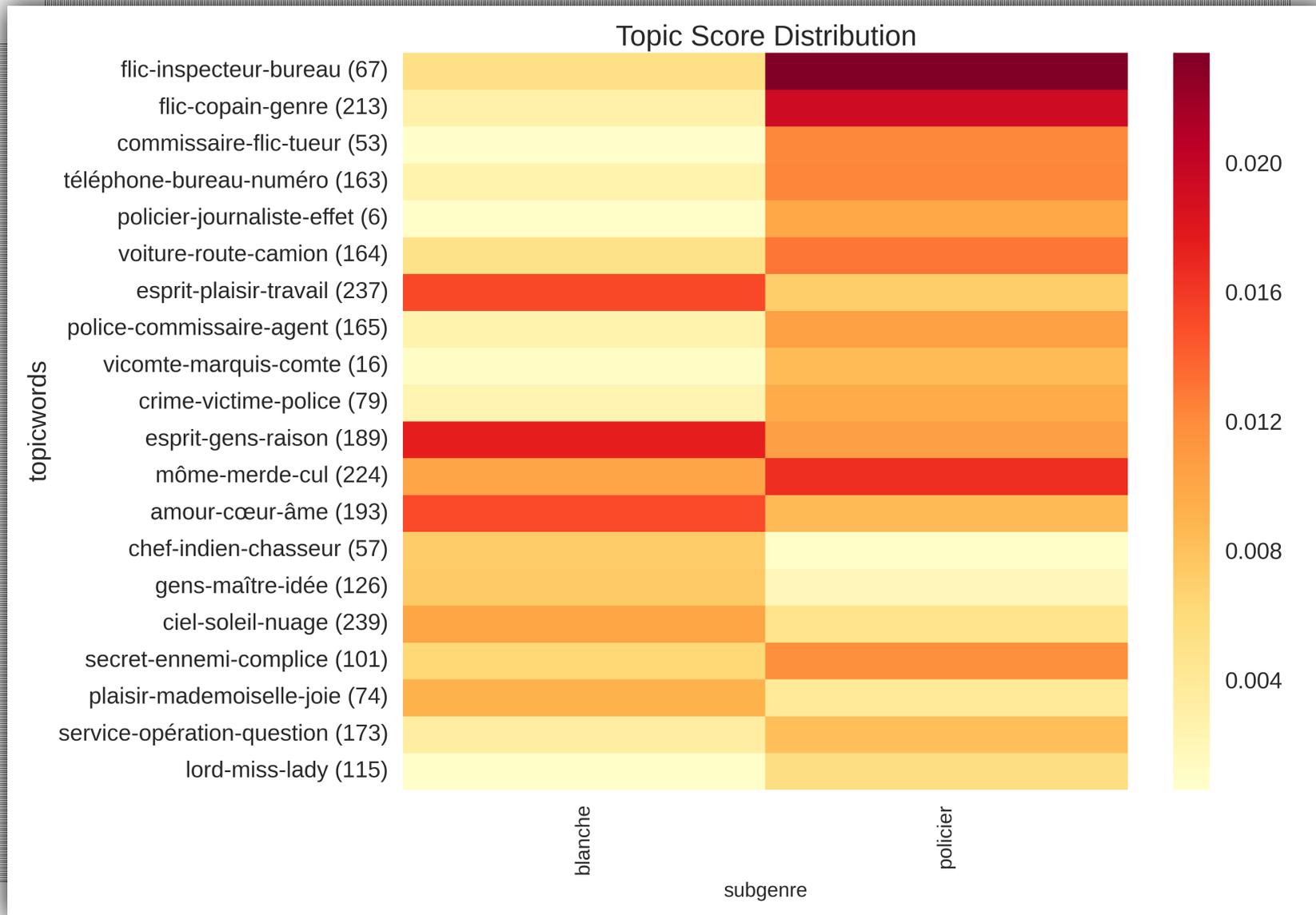
# Inhaltlich typisch?



Räuber, Messer, Dolch



# Statistisch: Überblick



(z-score-Normalisierung - Sortierung nach Standardabweichung)



# Statistisch: Nicht-Krimis

topic 193 (6/240)



Liebe, Herz, Seele

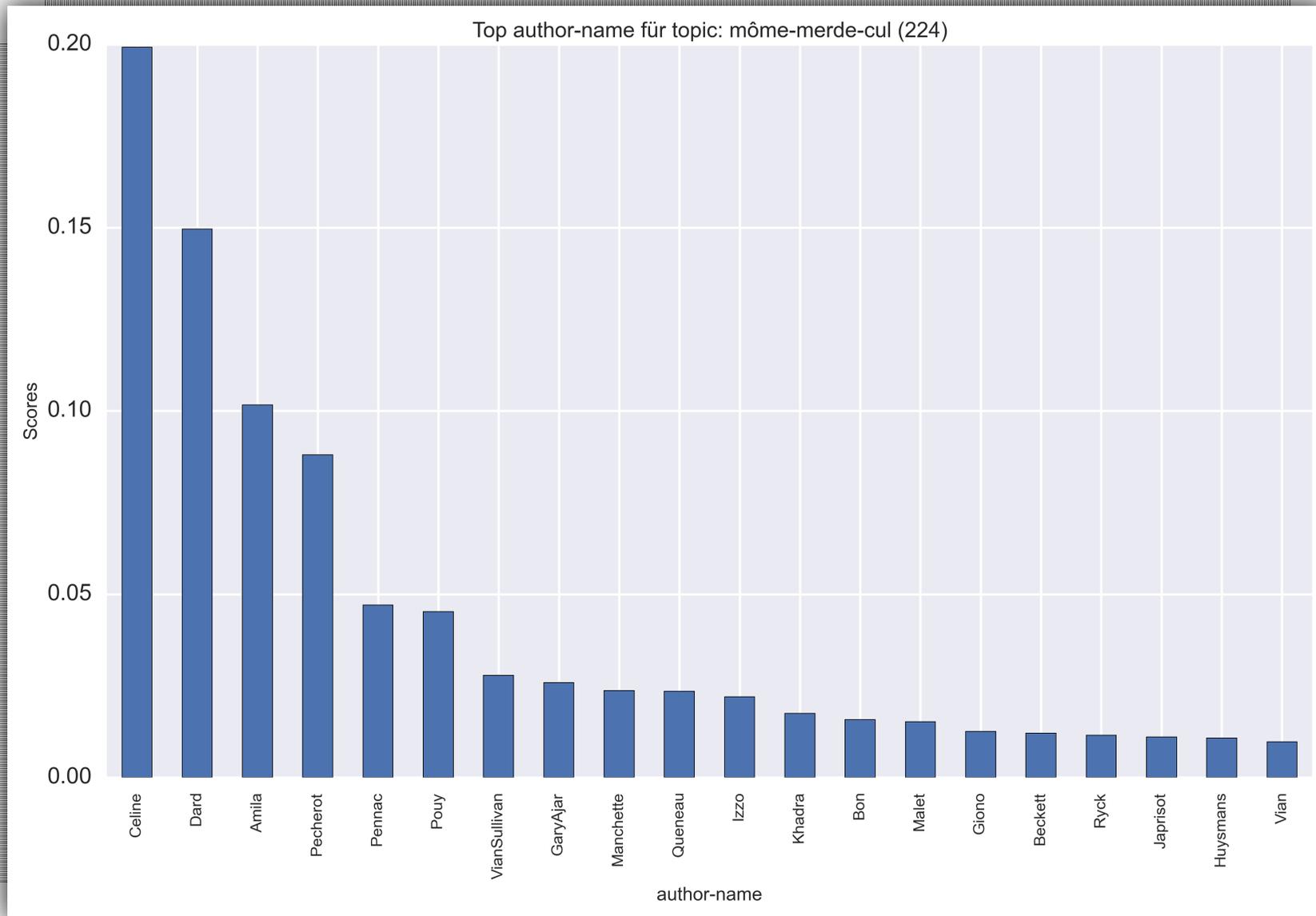
# Statistisch: Krimis



Polizist, Inspektor, Büro



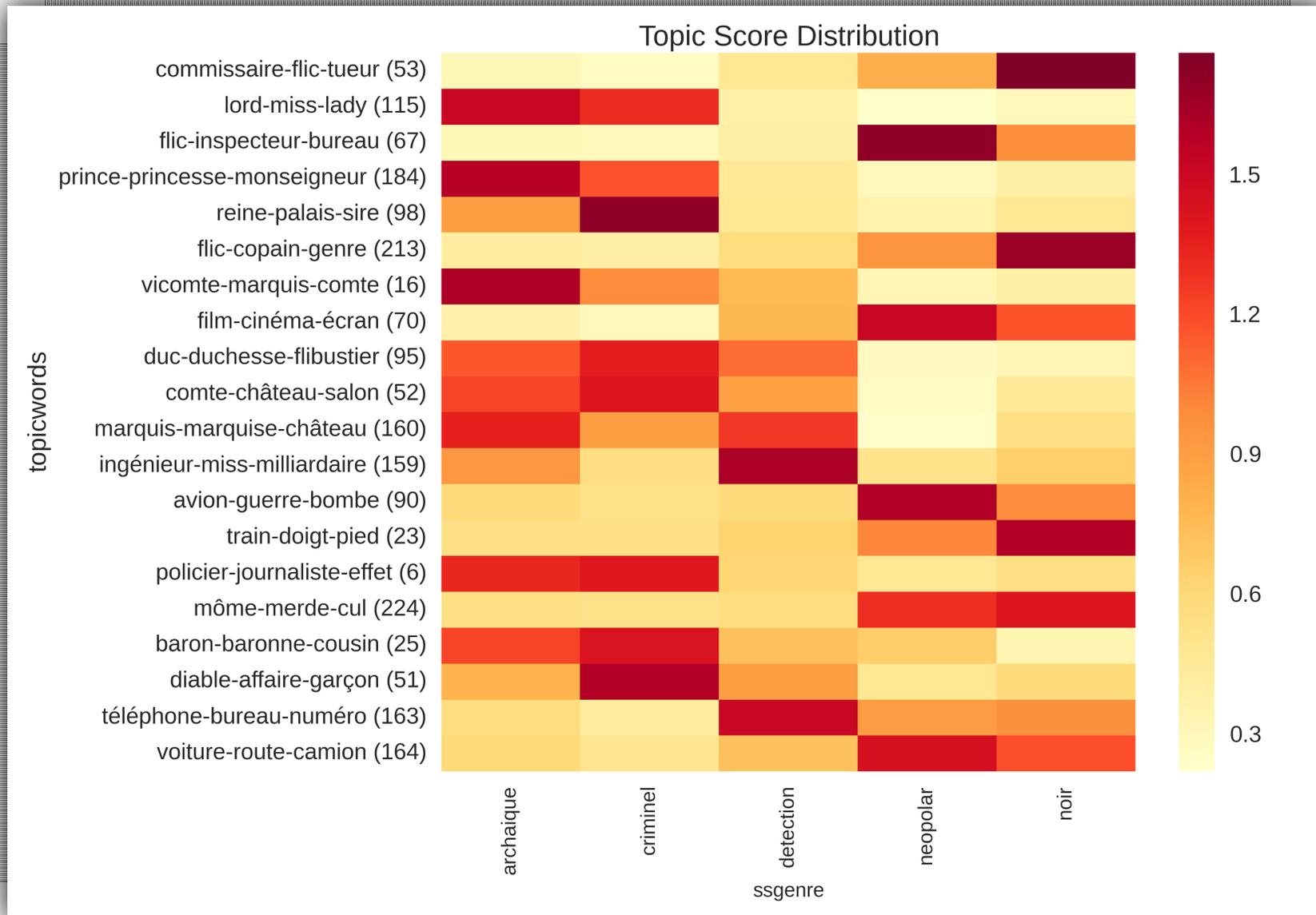
# Topic 224: Autoren?



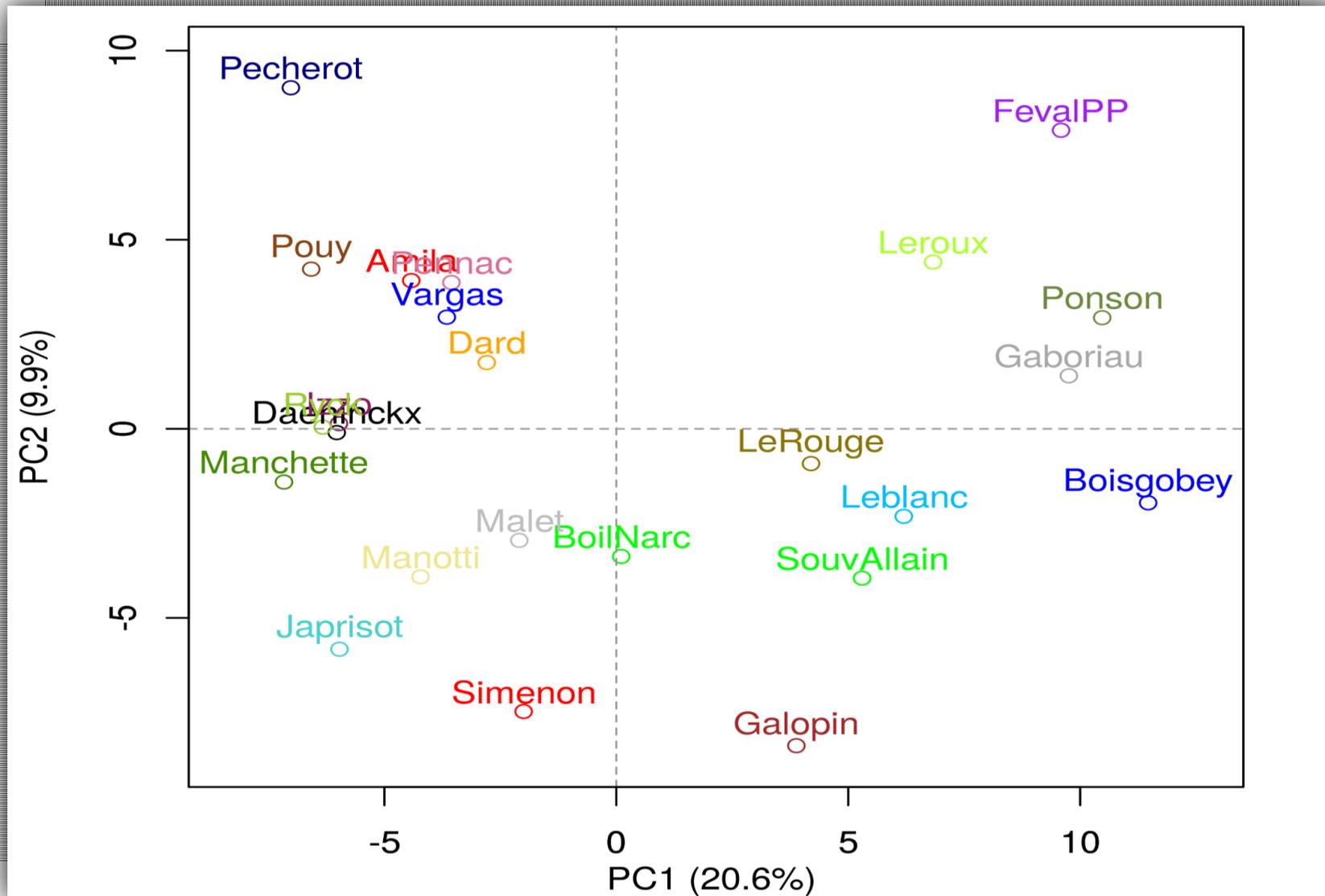
Kleine/r, Hintern, Sch\*\*ße

# Distinktivität vs. Clustering

# Distinktive Topics (nur Krimis)

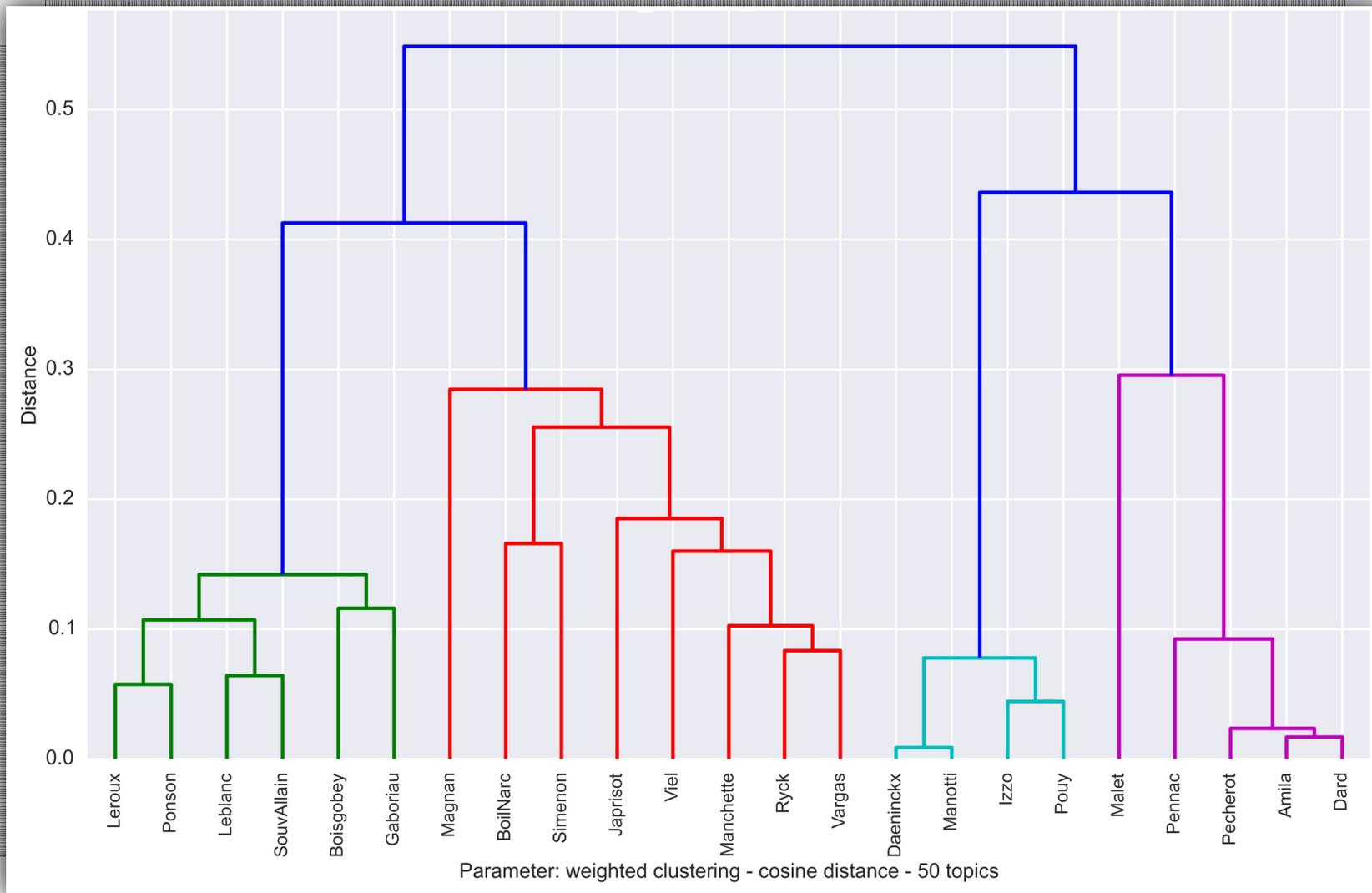


# PCA der (Krimi-)Autoren



(Grundlage: Topics 20-220 aus 240)

# Distanz-Clustering



# Fazit

# Wesentliche Aspekte

- Topics  $\neq$  Themen, sondern auch: Personal, Setting, Motive, Metatopics
- Inhaltlich typisch  $\neq$  statistisch distinktiv
- Gesamtsammlung: max. 10% "Krimi-Topics"; große Topic-Gemeinsamkeiten
- Clustering nach Topic-Werten: Zeit stärkster Faktor!
- Visualisierung als Instrument der "Makroanalyse" (Jockers)
- Dynamische, interaktive Verknüpfung der Visualisierungen
- Viel über wenige Texte sagen vs. wenig über viele Texte sagen?

**Vielen Dank**

<http://www.christof-schoech.de>

<https://twitter.com/christof77>

## **Lektürehinweise**

- Blei, David M. « Probabilistic Topic Models ». *Communications of the ACM* 55, n° 4 (2012): 77-84.
- Dubois, Jacques. *Le roman policier, ou la modernité*. Paris: Colin, 2005.
- Jockers, Matthew. *Macroanalysis - Digital Methods and Literary History*. Champaign, IL: University of Illinois Press, 2013.
- Lits, Marc. *Le roman policier. Introduction à la théorie et l'histoire d'un genre littéraire*. Liège: Éd. du CÉFAL, 1993.
- Schöch, Christof. « Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama ». *Digital Humanities Quarterly*, 2015 (angenommen).
- Steyvers, Mark, et Tom Griffiths. « Probabilistic Topic Models ». In *Latent Semantic Analysis: A Road to Meaning*, ed. T. Landauer et al. Laurence Erlbaum, 2006.
- Todorov, Tzvetan. « Typologie du roman policier » (1966). *Poétique de la prose*, 55–65. Paris: Seuil, 1971.