Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

CLiGS

Topic Modeling Genre
An Exploration of French Classical and Enlightenment Drama

Dr. Christof Schöch
JRG "Computational Literary Genre Stylistics" (CLiGS)
Department for Literary Computing
University of Würzburg, Germany

#gddh15 **#dayofdh2015**

Göttingen Dialog in Digital Humanities
Göttingen, May 19, 2015

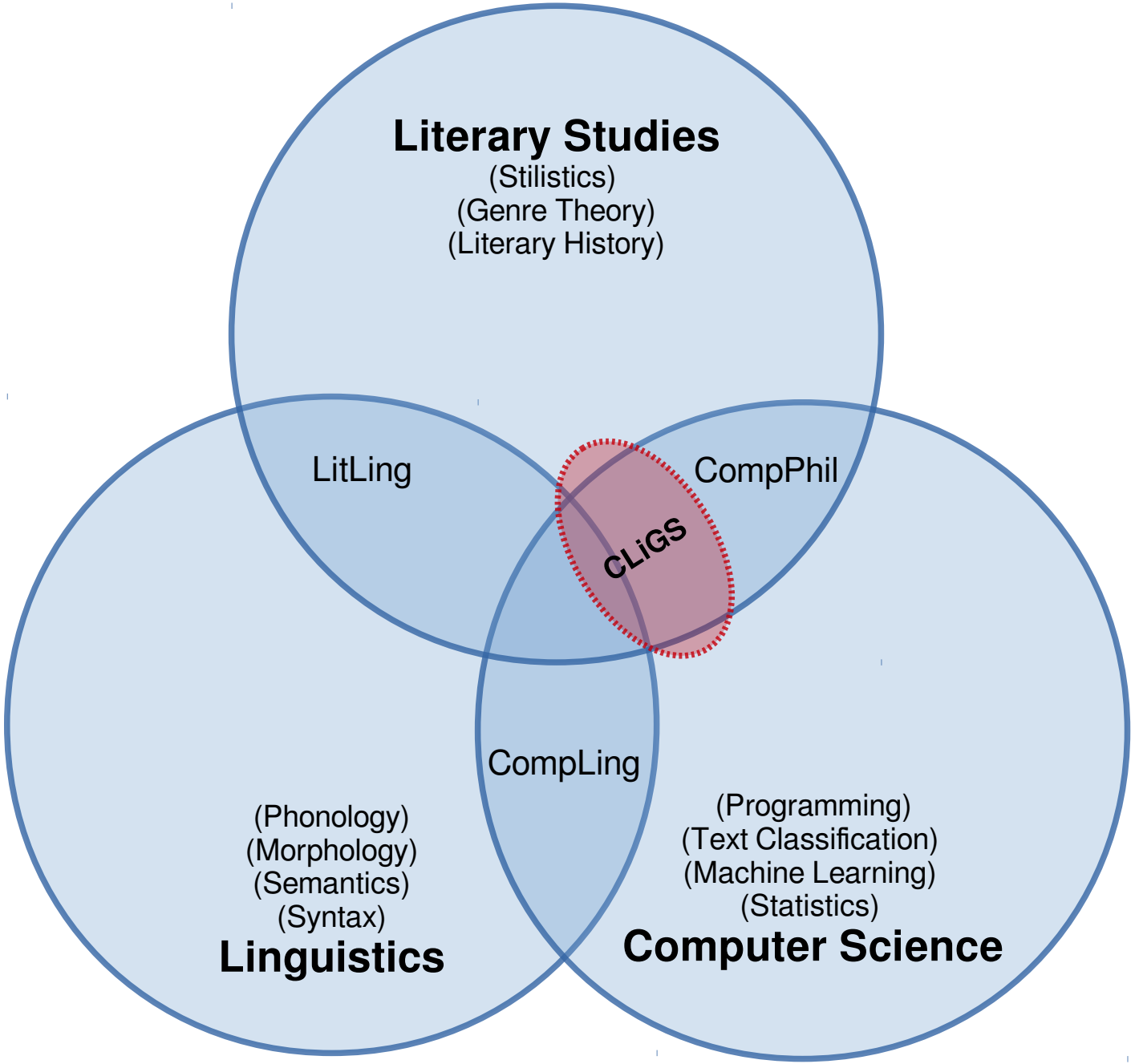
GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Overview

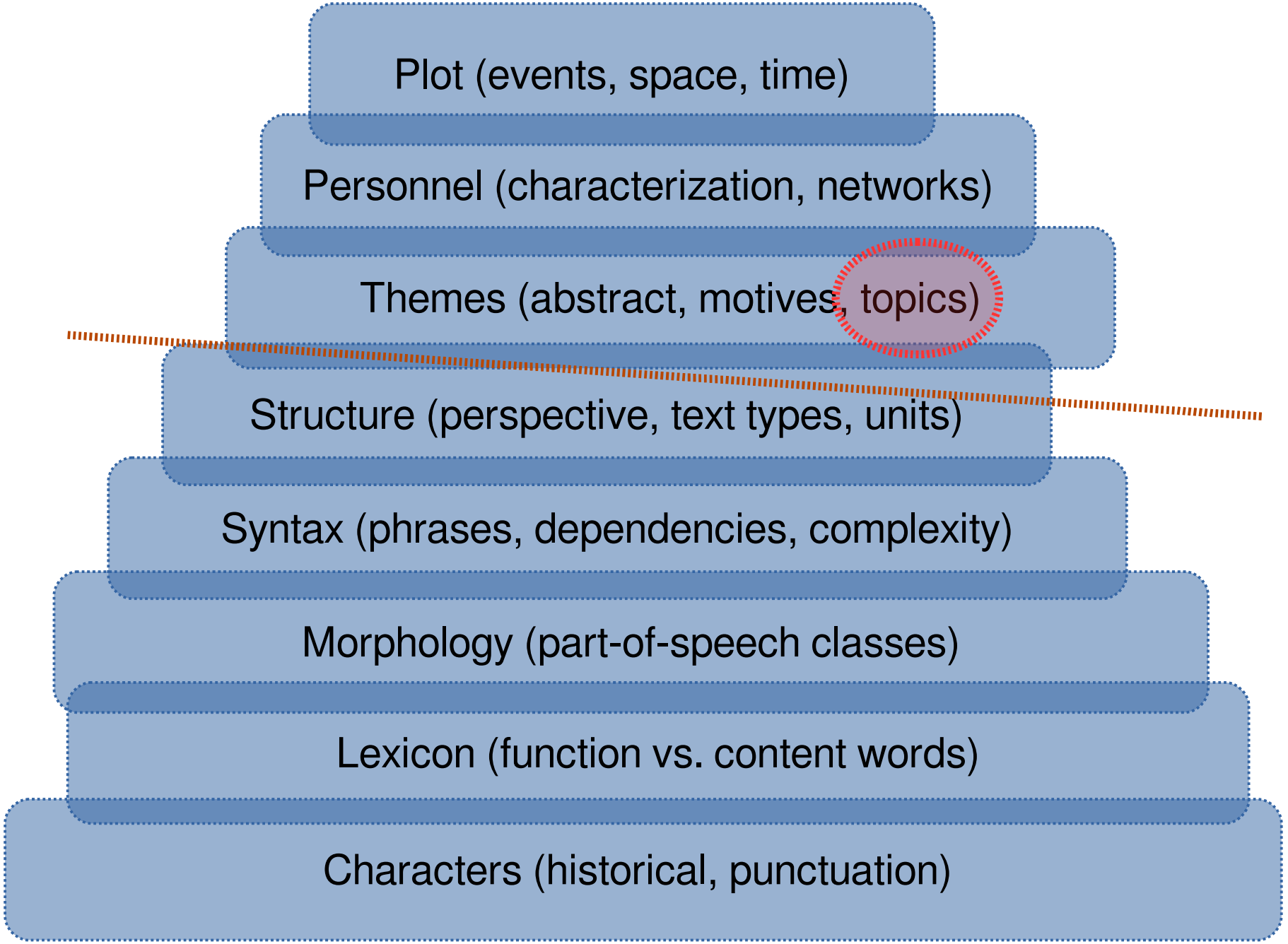
- **Introduction**
- **Data**
- **Hypotheses**
- **Method**
- **Results and Discussion**
  - Topic words, topic structure
  - Association with existing labels
  - Topic-based clustering
- **Conclusions**

Introduction

Computational Literary Text Analysis



Levels of description of genre

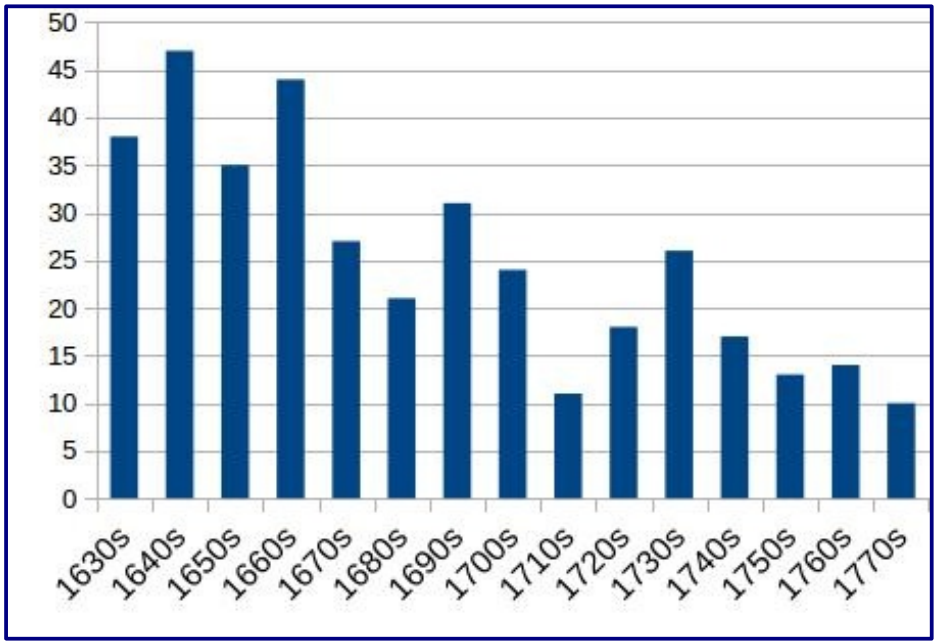


Data

Data: théâtre-classique.fr

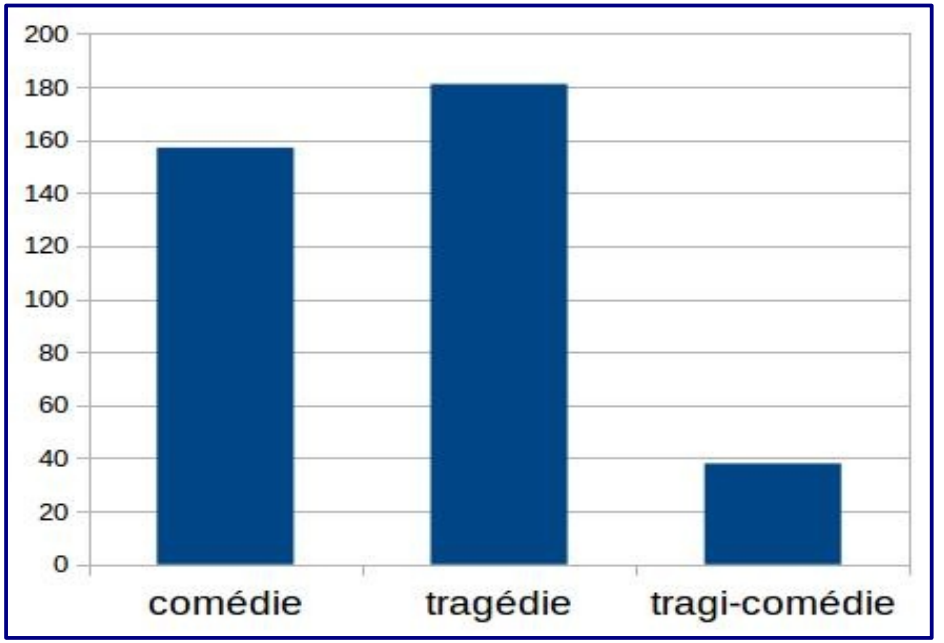
Théâtre classique

- ed. Paul Fièvre (Paris-IV)
- 1610 to 1810
- 740 plays
- no critical texts
- (quite) reliable text
- modernised spelling
- structural markup (TEI P4)
- rich metadata



Today's subcorpus

- 1630-1779
- three genres
- plays with 3/5 acts
- 375 plays
- speaker text only
- 5.3 mio. tokens / 30 MB
- metadata



Hypotheses

Hypotheses / Questions

Topics and genre

- Dramatic genres being (in part) defined on the basis of their themes, topic modeling should bring out genre-related patterns in the data

Genres' distinctive topics

- Which will be the topics most distinctive of comedies and tragedies? Will they be clearly thematic? Will they be expected?

Topics and plot

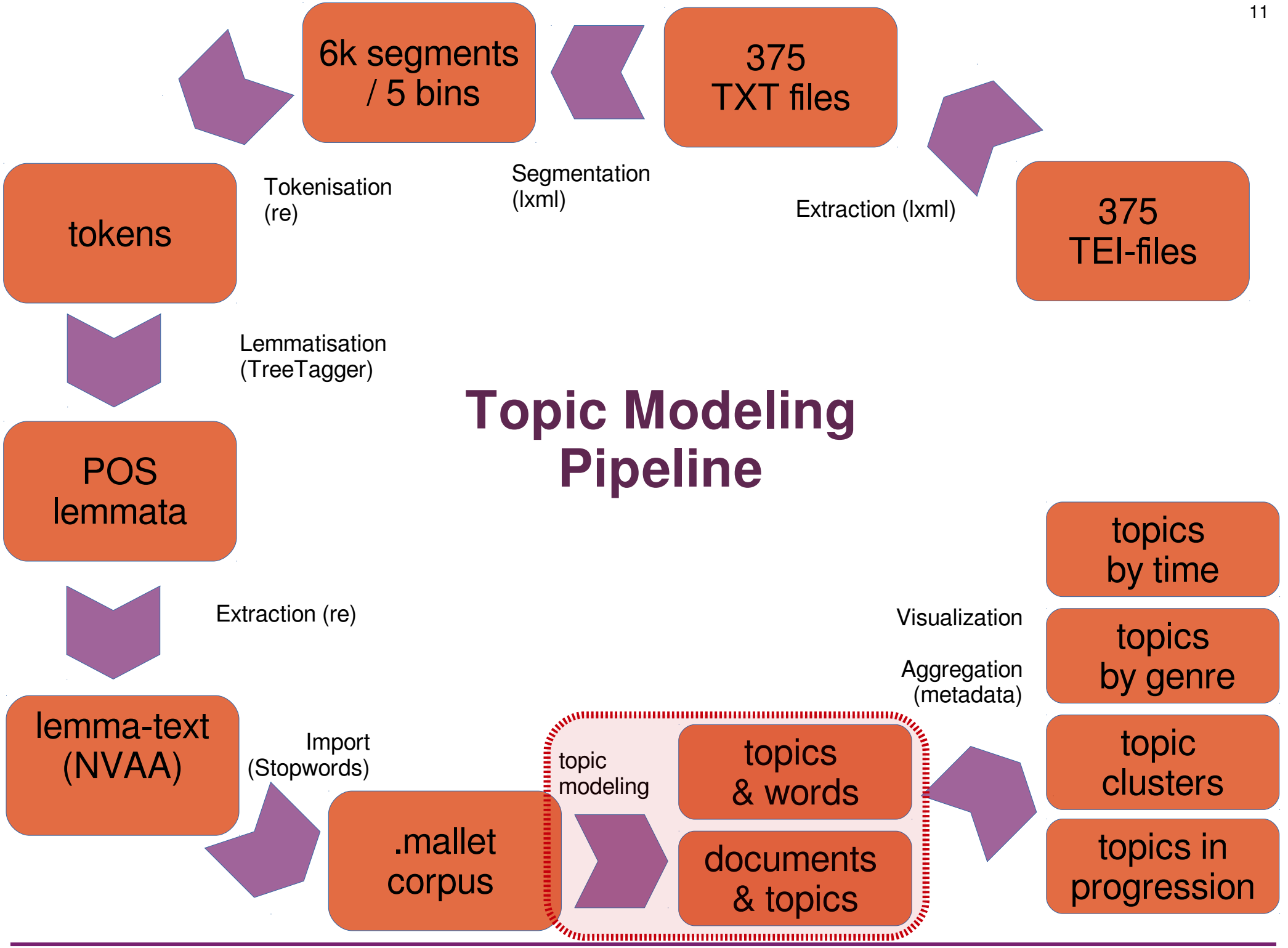
- Some topics should show genre-related plot patterns (i.e., show trends over textual progression)

Topics vs. MFW

- Topic-based clustering should be more genre-related than MFW-based clustering

Method

Topic Modeling Pipeline



Topic Modeling (1)

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

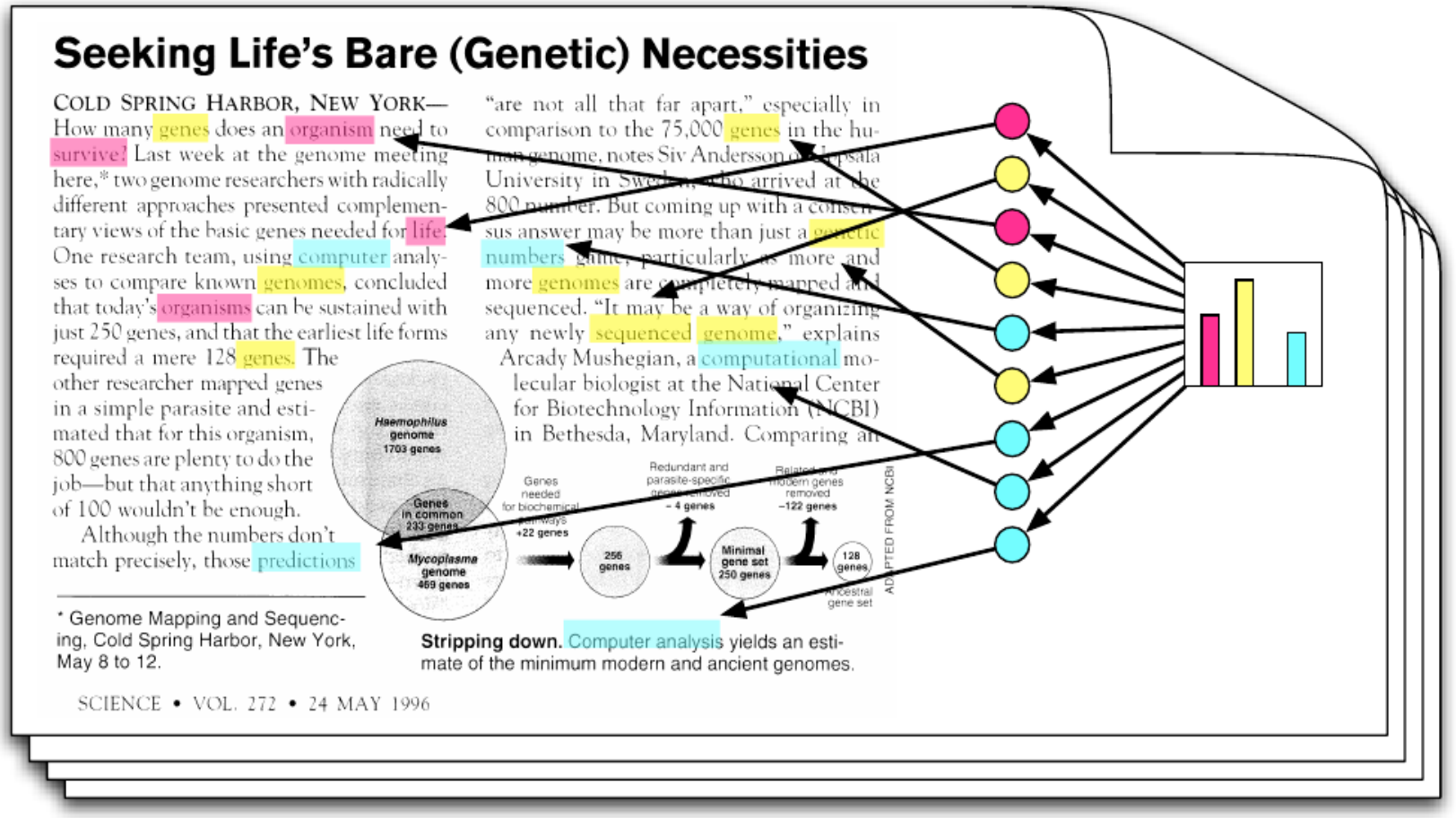
life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments



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

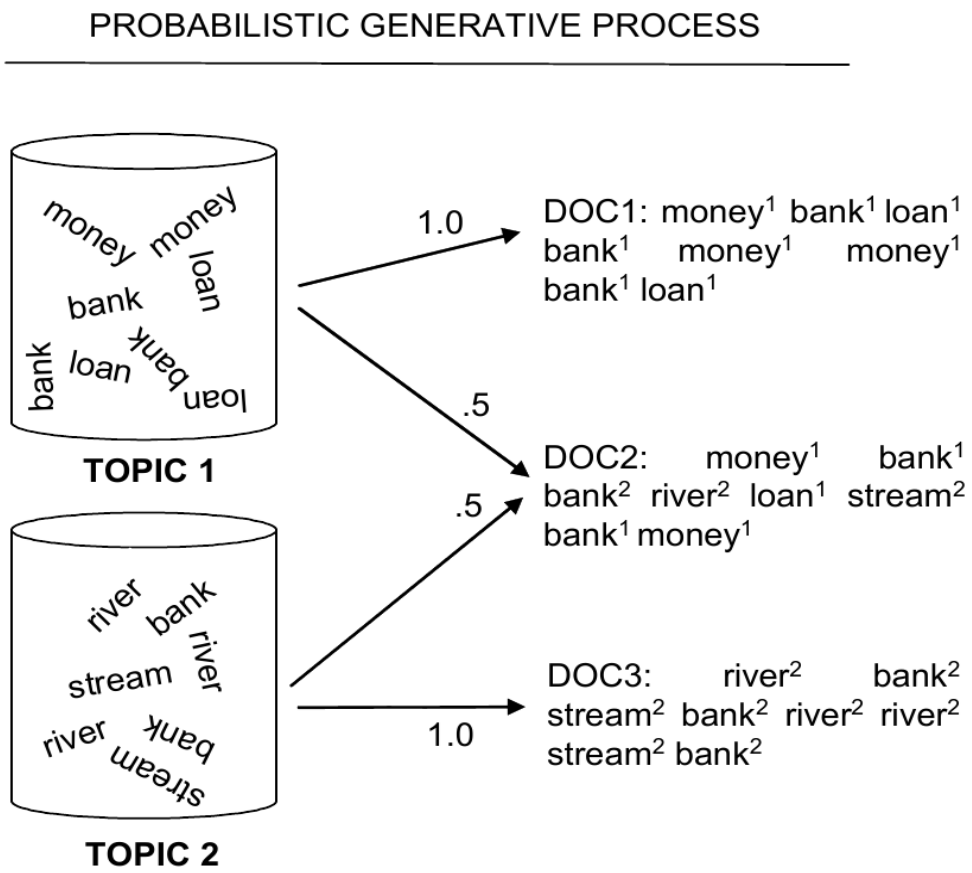
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

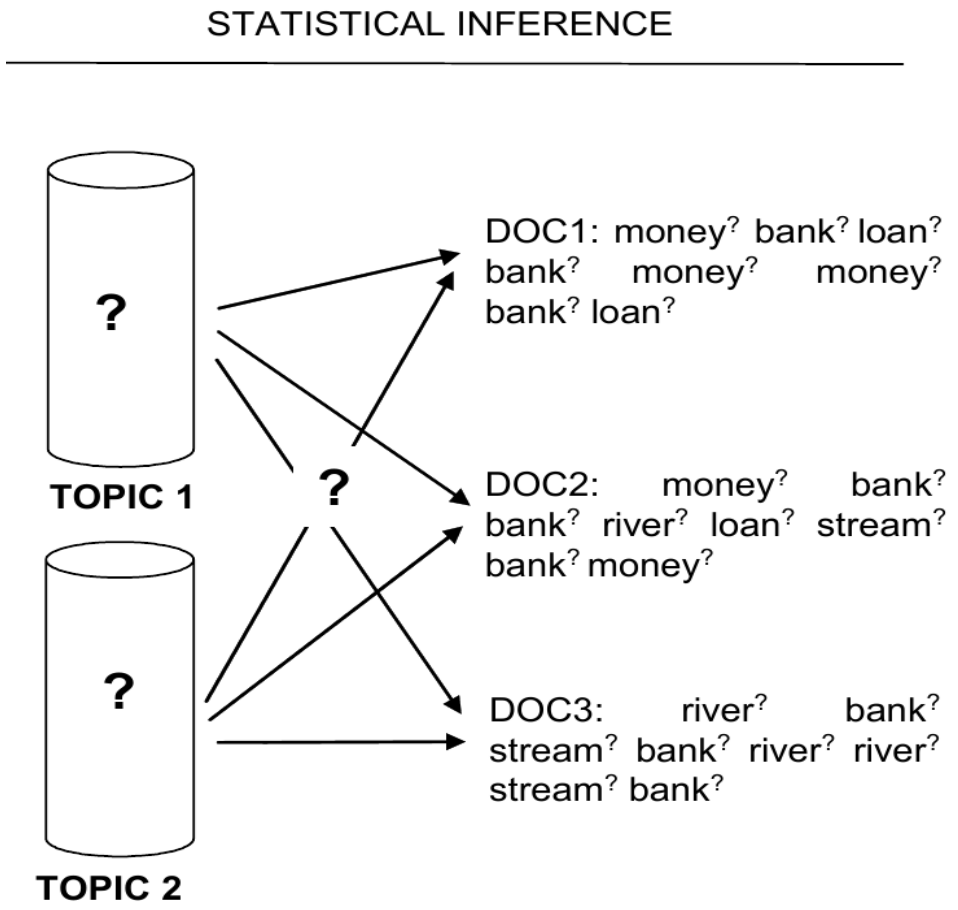
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Quelle :
Blei 2011

Topic Modeling (2)





Source :
Steyvers & Griffith 2006

Results

1. Topics
(topic words and structure)

2. Class-driven
(distinctive topics by genre / plot)

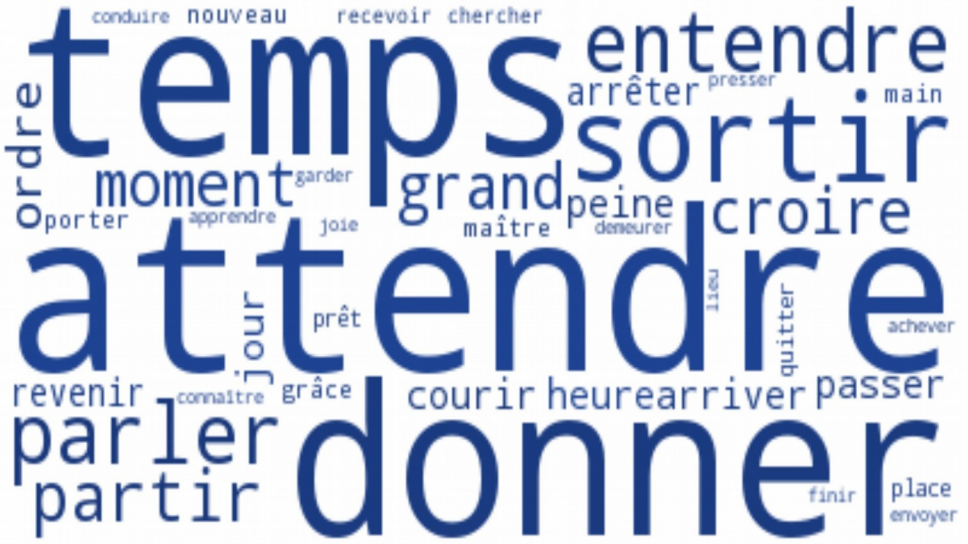
3. Data-driven
(topic-based clustering)

Topics: high and low topic probability

topic 55

topic 54

topic 32



topic 37

Topics: internal structure

topic 13

topic 35

topic 42

topic 51

Topics: expected and surprising

topic 21

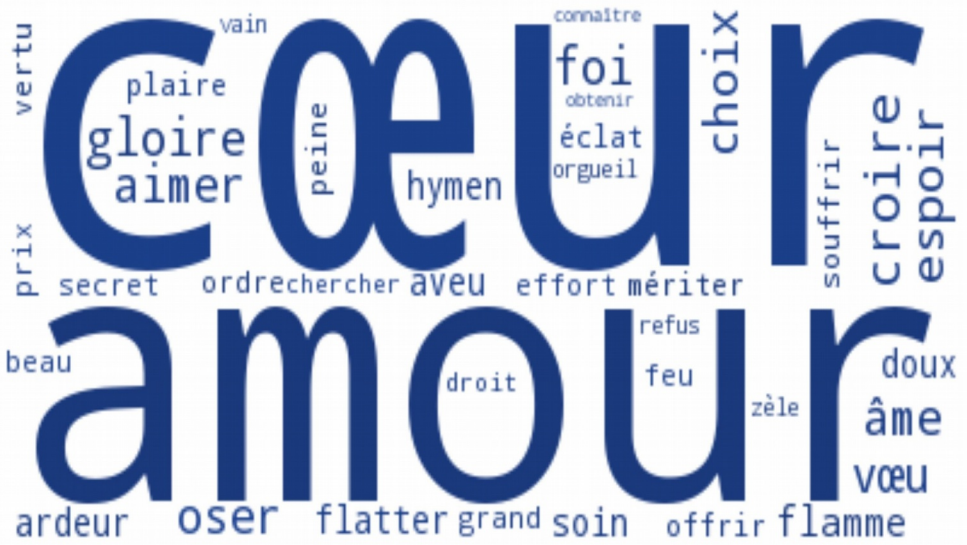
topic 75

topic 78

topic 45

Topics: love, love, love, love?

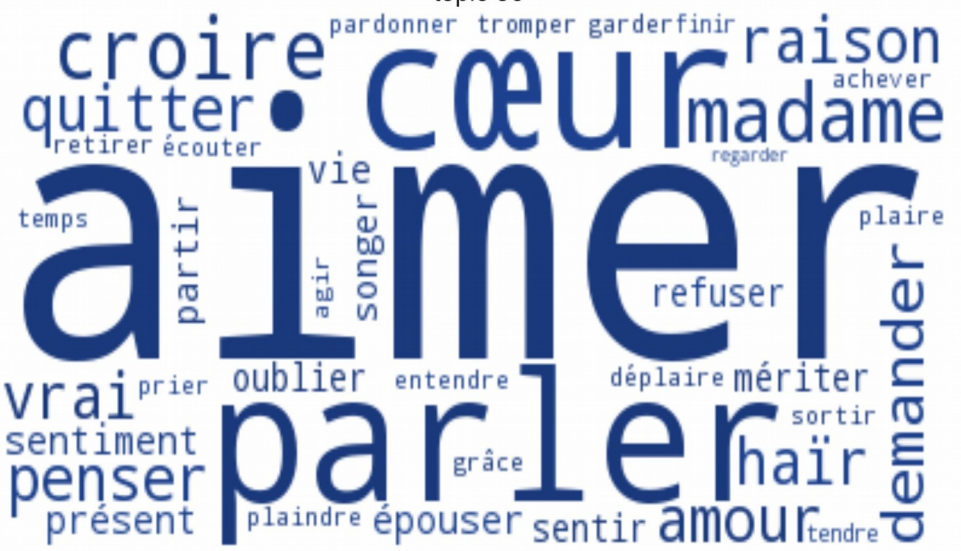
topic 15



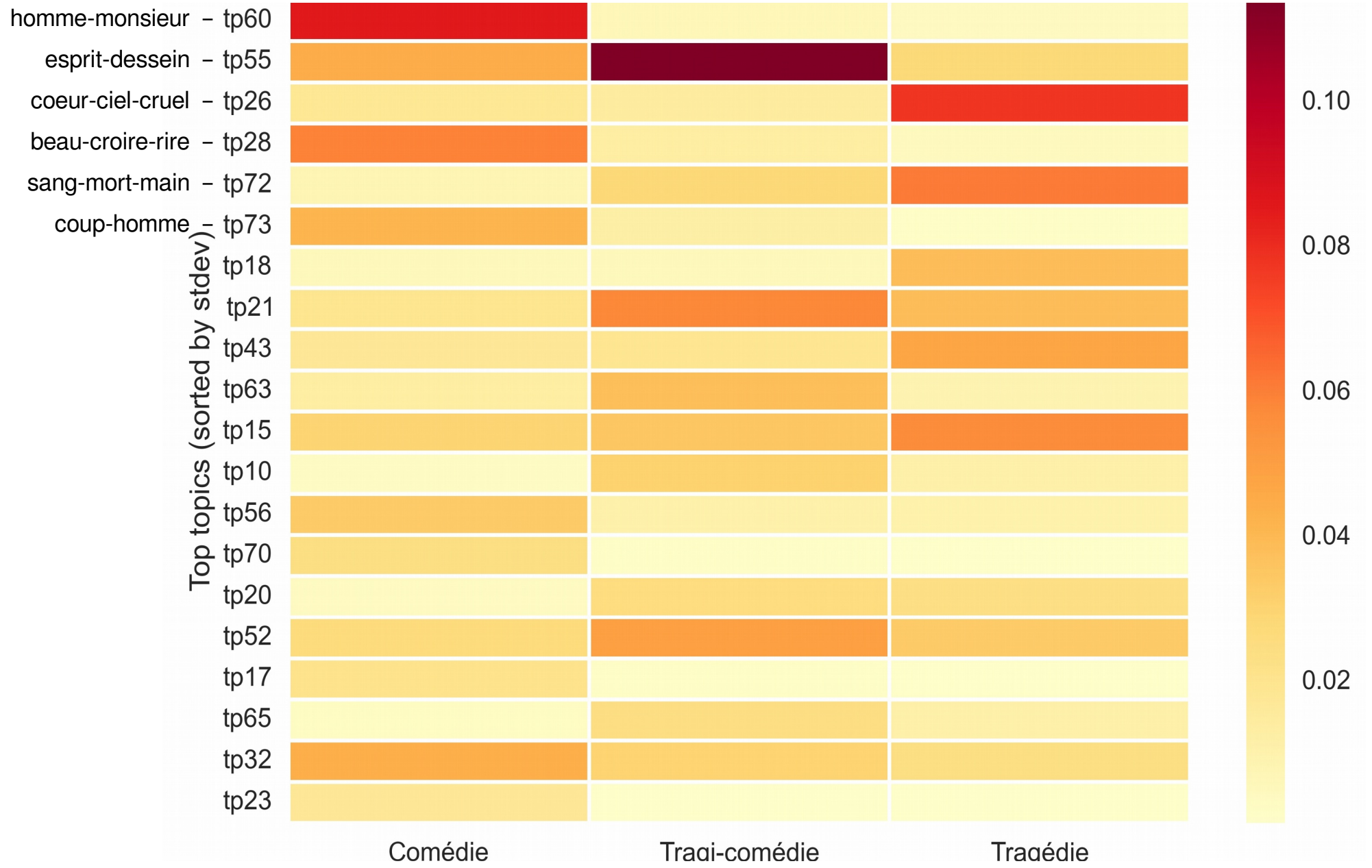
topic 52

topic 26

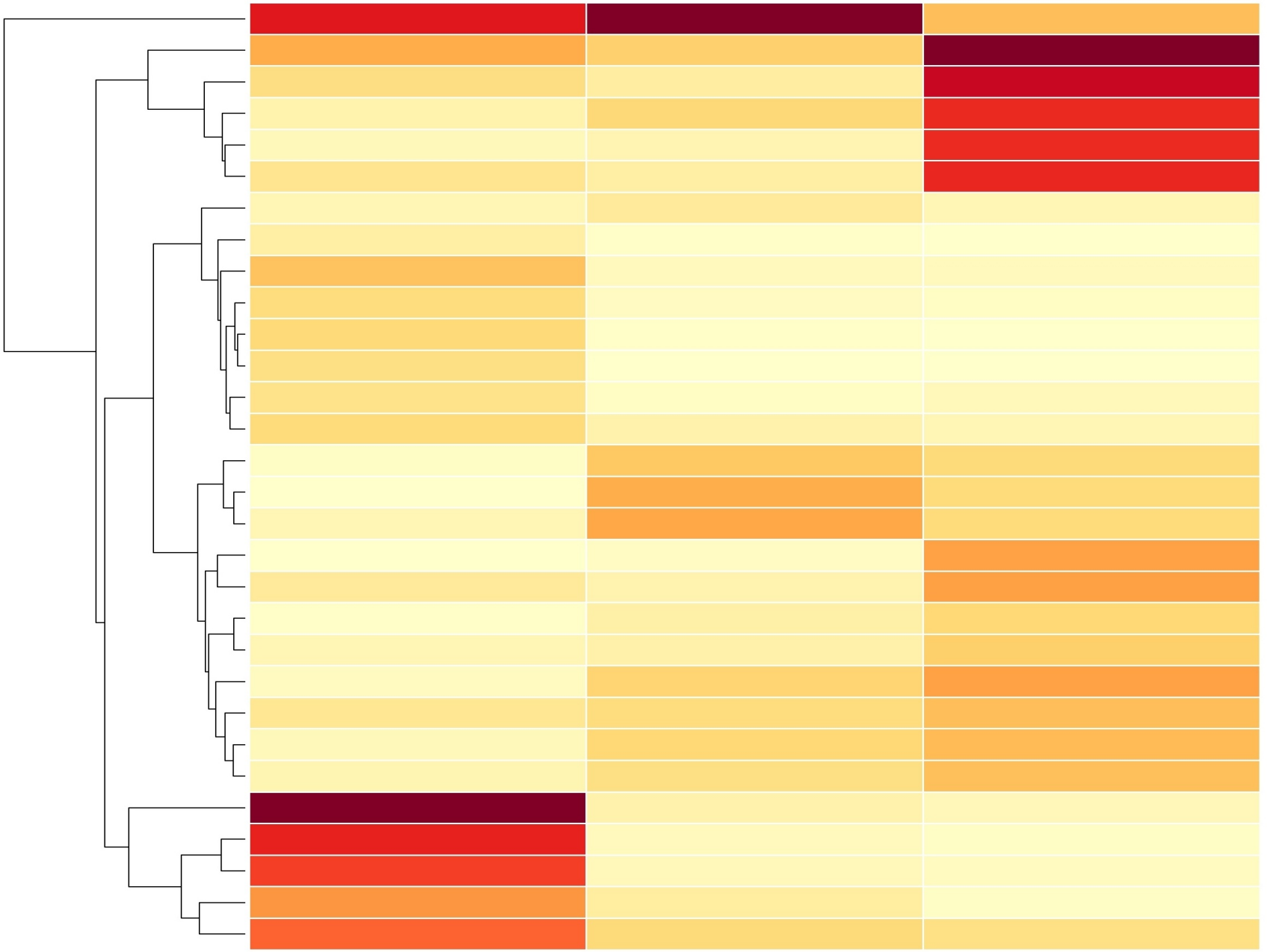
topic 56



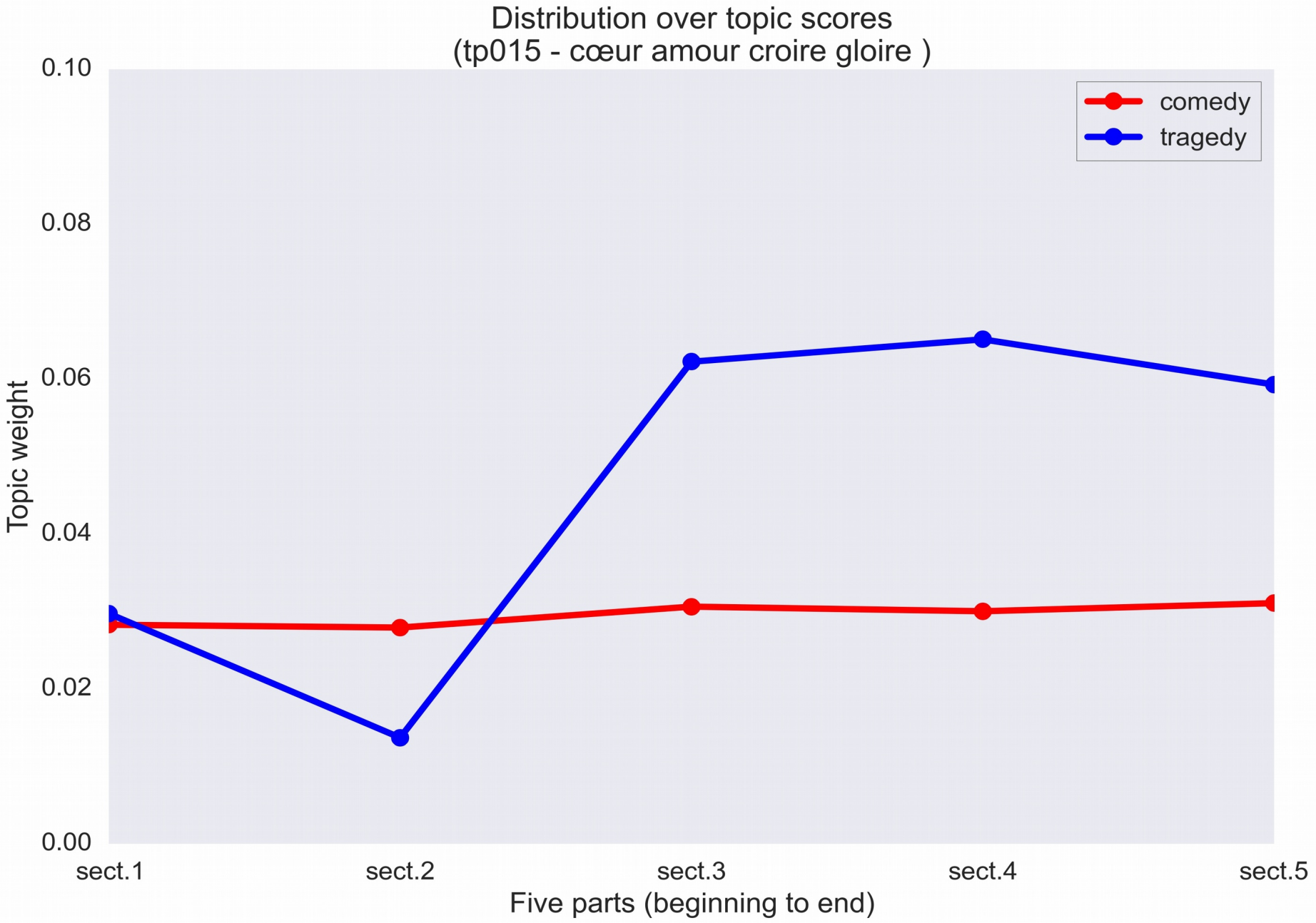
Distinctive Genre Topics (stdev)



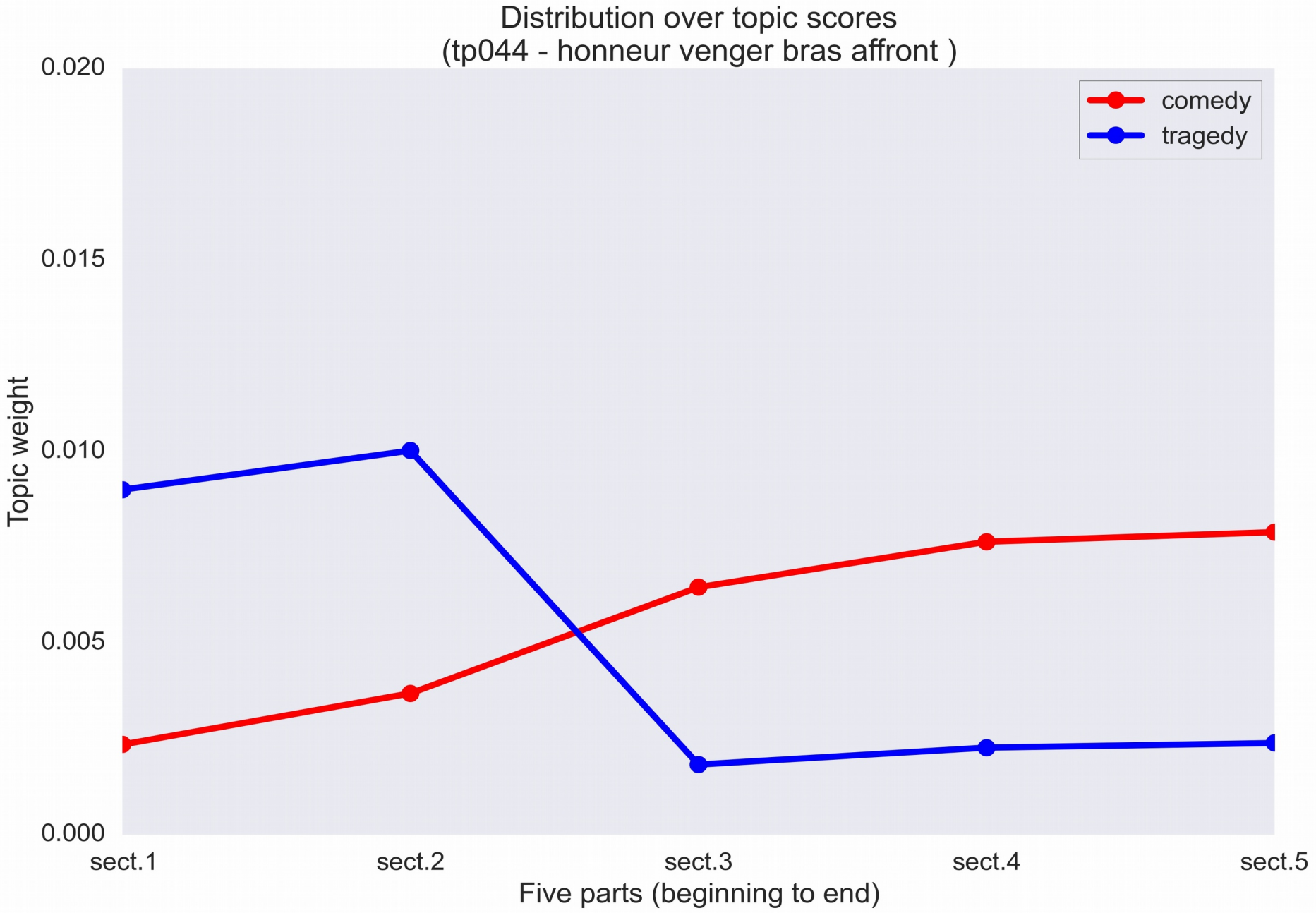
Topic Clustering (with genre)



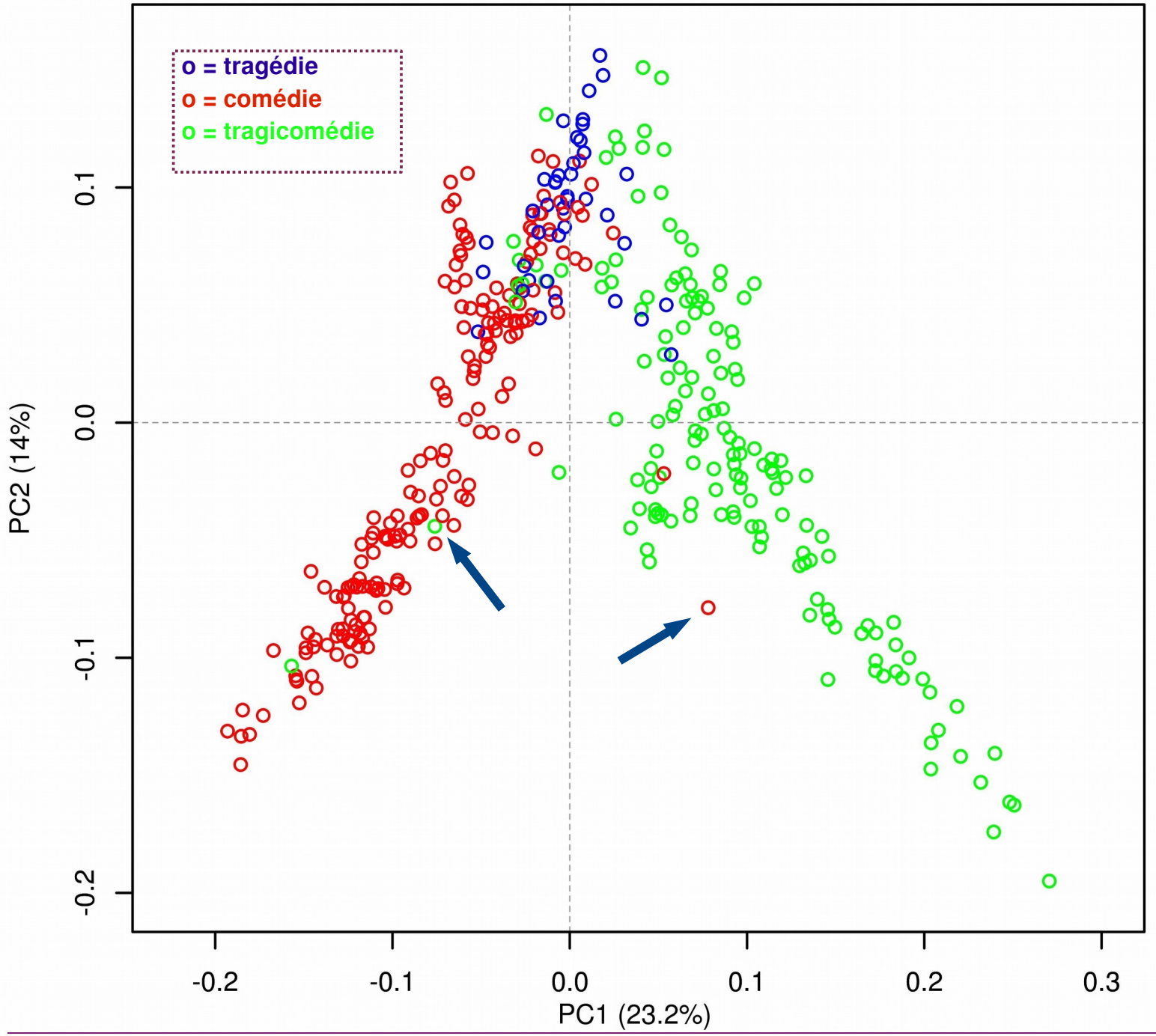
Topics in Textual Progression, by Genre (1)



Topics in Textual Progression, by Genre (2)



Topic-based clustering (plays, by genre)



Stray plays

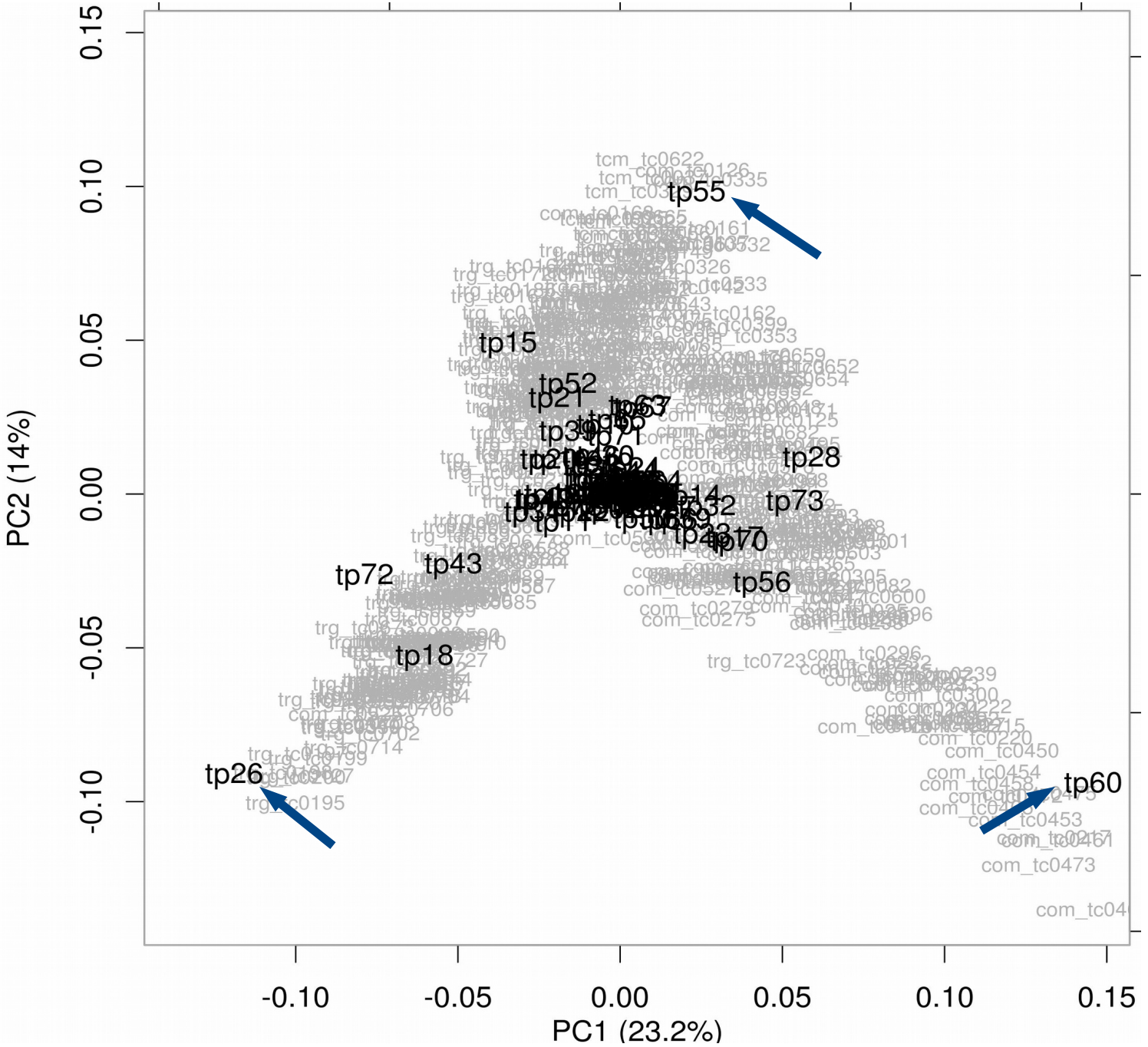
in comédie:

- for example: Voltaire, *Socrate* (tc0723): tragedy in prose

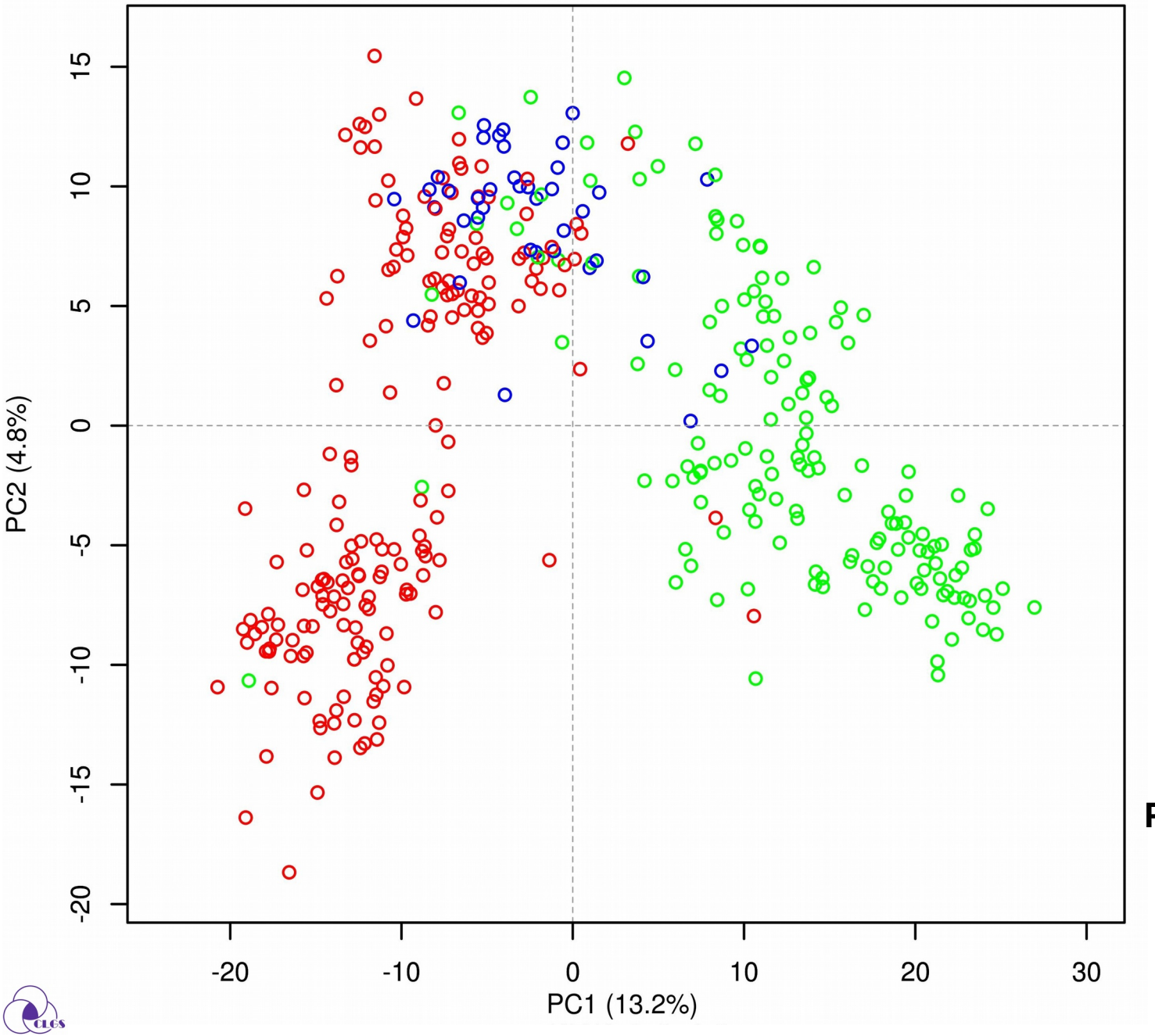
in tragédie:

- for example: Boissy, La vie est un songe (tc0055): comédie héroique

Topic-based clustering (loadings, by genre)



MFW-based clustering (by genre)



PCA
- 1200 MFW
- correlation matrix

Conclusion

Findings and challenges

The topics

- Most of the topics are quite coherent (subjectively)
- Some topics are abstract themes, others are motives / setting-related

Topics and genre

- Strong genre signal connected to dramatic sub-genres (as expected)
- Results suggest there may be two subtypes of tragedy

Topics and plot

- Topic scores across text progression suggest link between (several) topics and genre/plot

Topics vs. MFW

- PCA based on topics and on MFW yields very similar results (suprisingly) – genre signal strong in corpus (too many authors/decades)

Findings and challenges

Some challenges / future work

- Lack of sufficient numbers of texts (375 for 150 years!)
- Some results confirm existing knowledge, but there are a lot more trends/patterns
- Lack of experience (or knowledge) concerning topic-based clustering (influence of top words on topic score?)
- How to determine whether a change in topic weight over time, across genres, in textual progression, is really significant?
- With regard to textual progression: smarter way of splitting plays into smaller segments (trade-off: scene boundaries / similar length)
- Move from metadata-based averages to logistical regression or even to supervised / labeled LDA / sequential LDA

References

- Blei, David M. 2011. “Introduction to Probabilistic Topic Models.” *Communication of the ACM*.
- Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013.
- Larthomas, Pierre. “La notion de genre littéraire en stylistique.” *Le Français moderne* 32 (1964).
- McCallum, Andrew. 2002 "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. University of Massachusetts Amherst.
- Scherer, Jacques. *La dramaturgie classique en France*. Paris: Nizet, 2001 (1950).
- Schöch, Christof. “Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik.” In *Literaturwissenschaft im digitalen Medienwandel*, ed. Christof Schöch & Lars Schneider, 130–57, Beiheft 7 von *Philologie im Netz,* 2014. http://web.fu-berlin.de/phin/beiheft7/b7t08.pdf.
- Steyvers, Mark, and Tom Griffiths. “Probabilistic Topic Models.” *Latent Semantic Analysis: A Road to Meaning*, ed. T. Landauer et al. Laurence Erlbaum, 2006.

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

CLiGS

Thank you!

Dr. Christof Schöch
JRG “Computational Literary Genre Stylistics” (CLiGS)
Department for Literary Computing
University of Würzburg, Germany

#gddh15 **#dayofdh2015**

Göttingen Dialog in Digital Humanities
Göttingen, May 19, 2015

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung