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
Overview

- Introduction
- Data
- Hypotheses
- Method

**Results and Discussion**
- Topic words, topic structure
- Association with existing labels
- Topic-based clustering

- Conclusions
Introduction
Levels of description of genre

Plot (events, space, time)
Personnel (characterization, networks)
Themes (abstract, motives, topics)
Structure (perspective, text types, units)
Syntax (phrases, dependencies, complexity)
Morphology (part-of-speech classes)
Lexicon (function vs. content words)
Characters (historical, punctuation)
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

- 375 TXT files
- 375 TEI-files
- 6k segments / 5 bins

**Tokens**
- Tokenisation (lxml)
- Segmentation (lxml)
- Extraction (lxml)

**POS lemmata**
- Lemmatisation (TreeTagger)
- Extraction (re)

**lemma-text (NVAA)**
- Import (Stopwords)

**.mallet corpus**
- Extraction (re)

**topic modeling**
- topics & words
- documents & topics

**Visualization**
- topics by time
- topics by genre

**Aggregation (metadata)**
- topic clusters
- topics in progression
Topic Modeling (1)

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 70,000 genes in the human genome, notes Siv Anderson, an entomologist at Harvard University in Cambridge. But coming up with a consensus answer may be more than just a matter of numbers. Other, particularly more and more genomes are being sequenced and compared. “It may be a way of organizing any newly sequenced genome,” explains Arvind Mudhakar, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing and

Quelle: Blei 2011
Topic Modeling (2)

PROBABILISTIC GENERATIVE PROCESS

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
Topics: internal structure
Topics: expected and surprising
Topics: love, love, love, love?
Distinctive Genre Topics (stdev)

homme-monsieur - tp60
esprit-dessein - tp55
coeur-ciel-cruel - tp26
beau-croire-rire - tp28
sang-mort-main - tp72
coup-homme - tp73

Top topics (sorted by stdev)

tp18
tp21
tp43
tp63
tp15
tp10
tp56
tp70
tp20
tp52
tp17
tp65
tp32
tp23

Comédie | Traï-comédie | Tracédie

0.10
0.08
0.06
0.04
0.02
Topic Clustering (with genre)
Topics in Textual Progression, by Genre (1)

Distribution over topic scores
(tp015 - cœur amour croire gloire)

- Comedy
- Tragedy

Five parts (beginning to end)
Distribution over topic scores
(tp044 - honneur venger bras affront)

- Comedy
- Tragedy

Five parts (beginning to end)
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 héroïque
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

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