**SCHÖCH — Topic**

<7 figures>

**Topic Modeling French Crime Fiction**

Schöch, C.

This study applies topic modeling to a collection of French crime fiction novels in order to discover topic-related patterns. The results show both expected and unexpected patterns related to authors, subgenres, and time period. Topic modeling proves highly useful for investigating the history of French crime fiction.

# French Crime Fiction

Crime fiction is a type of narrative prose fiction involving the elucidation of a (usually) violent crime through a (more or less) rational investigation (often) taking place in an urban setting and (typically) involving (one or several) investigators, victims, witnesses, suspects, and criminals. French crime fictionʼs rich history goes back to the 1860s and has many highly prolific proponents. Prototypical detective fiction is easily recognized, but the boundaries of the genre and its internal division into subgenres remain controversial (see Todorov, 1971; Lits, 1993; Colin, 1999; Lavergne, 2009). The abundant material is a challenge to any readerʼs memory but an opportunity for quantitative methods.

# Research Questions

This study addresses the following questions: How prevalent are expected, genre-related topics such as crime and investigation, and which other topics are important? What relations exist between topics and categories like authorship, subgenre, or time period? What kind of groupings of novels does one obtain based on topic similarity? What new insights into the history of crime fiction does topic modeling allow?

# Data

For this study, a collection of 270 French novels published between 1858 and 2012 was created. The vast majority are crime fiction novels, but some non–crime fiction novels have also been included. The collection includes novels pertaining to seven subgenres and written by 14 different authors. It has around 16 million word tokens. Texts in the public domain have been obtained online (from ebooksgratuits.com), while additional texts have been obtained by full-text digitization.

# Method

Topic modeling is an unsupervised method of discovering latent semantic structure in large collections of texts. Technically, a topic is a probability distribution over word frequencies, and each text is characterized by a distribution over topics (see Steyvers and Griffiths, 2006). In practice, the words with the highest scores in a given topic are mostly semantically related words; the topics with the highest scores in a text represent the textʼs major themes or motives.

The most widely known algorithm is Latent Dirichlet Allocation (LDA; see Blei et al., 2003), but several precursors and alternatives exist—for instance, Non-Negative Matrix Factorization (Lee and Seung, 1999). Several tools are available, like MALLET (McCallum, 2002) or gensim (Rehurek and Sojka, 2010), as well as tutorials (e.g., Graham et al., 2012, Riddell, 2014). Topic modeling has proven immensely popular in digital humanities (e.g., Blevins, 2010; Rhody, 2012; Jockers, 2013).

For the results reported here, the following parameters have been used: Lemmatization has been applied to the texts, because French is a highly inflected language. After POS-tagging with TreeTagger (Schmid, 1994), nouns have exclusively been selected for analysis. Each novel has then been split into segments of approximately 150 nouns each. MALLET has been run with 60 target topics and 10,000 iterations.

# Results and Discussion

## Topics Obtained

The topics obtained can manually be labeled by their dominant semantic trait (see Figure 1).

### 

Figure 1. Selection from the 60 topics obtained (with topic ID and 20 top-ranked words).

The subjective topic coherence is very high: few top-ranked words do not share semantic traits, and few topics are hard to interpret (but see Chang et al., 2009; Schmidt, 2012). Many topics could appear in any type of novels, such as topics #28, #38, and #01 (labeled ‘family’, ‘money’, ‘train’). Only nine out of 60 topics are related to crime fiction, such as #15, #33, and #53 (labeled ‘investigators’, ‘fire arms’, ‘jewelry’). Judged by topic composition alone, crime fiction appears to be a less distinctive novelistic sub-genre than expected. Note that topics are based on various types of similarity: topic #44 (‘interiors’) is related to a recurrent setting, topic #22 (“informal1”) to a specific register.

## Authorship, Subgenre, and Time

Topic scores per text segment can be aggregated and averages obtained, for instance, at the document, author, genre, or time period levels.

On the author level (see Figure 2), it appears that several (but not all) authors have a distinctive ‘signature topic’: a topic with a particularly high score in comparison both to other topics for the same author and to the same topic for other authors (i.e., across rows and columns).

### 

### Figure 2. Distribution of topic scores at the author level (15 topics with the largest variation across authors, measured in standard deviations).

Gaboriauʼs signature topic is very general (#11, ‘bourgeoisie’) while Simenonʼs is more genre-specific (#29, ‘office’) and Maletʼs is not thematic (#22, ‘informal1’). For some authors (Leroux, Manchette, Ponson), no clear signature topic emerges.

Compared to authors, most subgenres have less marked characteristic topics (Figure 3).

### 

### Figure 3. Topic scores aggregated to the subgenre level (20 topics with the largest variation, measured in standard deviations, across genres).

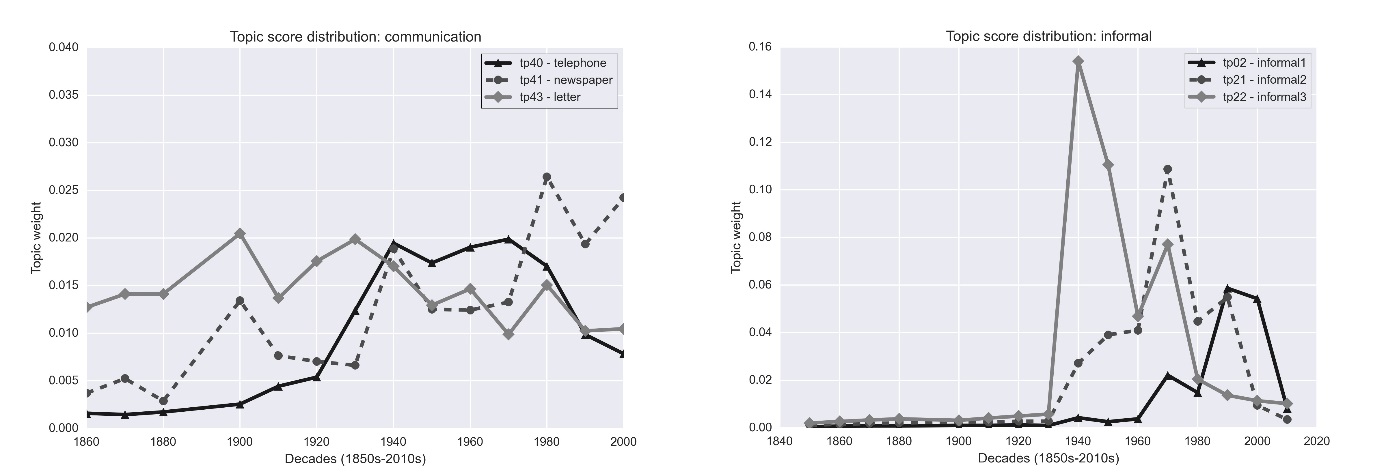
For example, the classical detective novelʼs most characteristic topic is #29 (‘office’). However, the traditional genre labels used here are problematic, because they tend to refer to periods rather than structural types and correlate strongly with authorship.

When average topic scores are obtained across all novels for each successive tenths of novels, and topic score progressions are compared across subgenres, genre-specific patterns appear (see Figure 4).

### 

Figure 4. Topic score progression for topic #26 (average topic scores per text segment and subgenre).

For instance, topic #26 (‘twilight’) decreases over the course of the average detective fiction novel, but remains stable at a higher level over the course of ‘roman noir’. In the former, darkness and uncertainty are being dispelled and order restored, while in the latter, they are not.

The distribution of topics at the level of time period (Figure 5) shows that, as expected, some topics gain while others lose in importance over time. Some very similar topics ‘take turns’, as it were, with successive peaks (‘informal’, right). Others reflect larger societal changes (different means of communication, left).

### Figure 5. Topic scores per decade for two topic groups.

For the ‘informal’ topics, each peak is associated with one author and their period of activity: Malet, Dard, and Vargas (see Figure 2). The topicʼs rapid rise in the 1940s underlines how bold Maletʼs use of informal language was, but also shows the (problematic) impact of individual authors on these temporal patterns.

Beyond individual topic development over time, the cumulated rate of change in topic scores from one decade to the next gives an insight into the thematic innovation cycles of crime fiction (Figure 6).

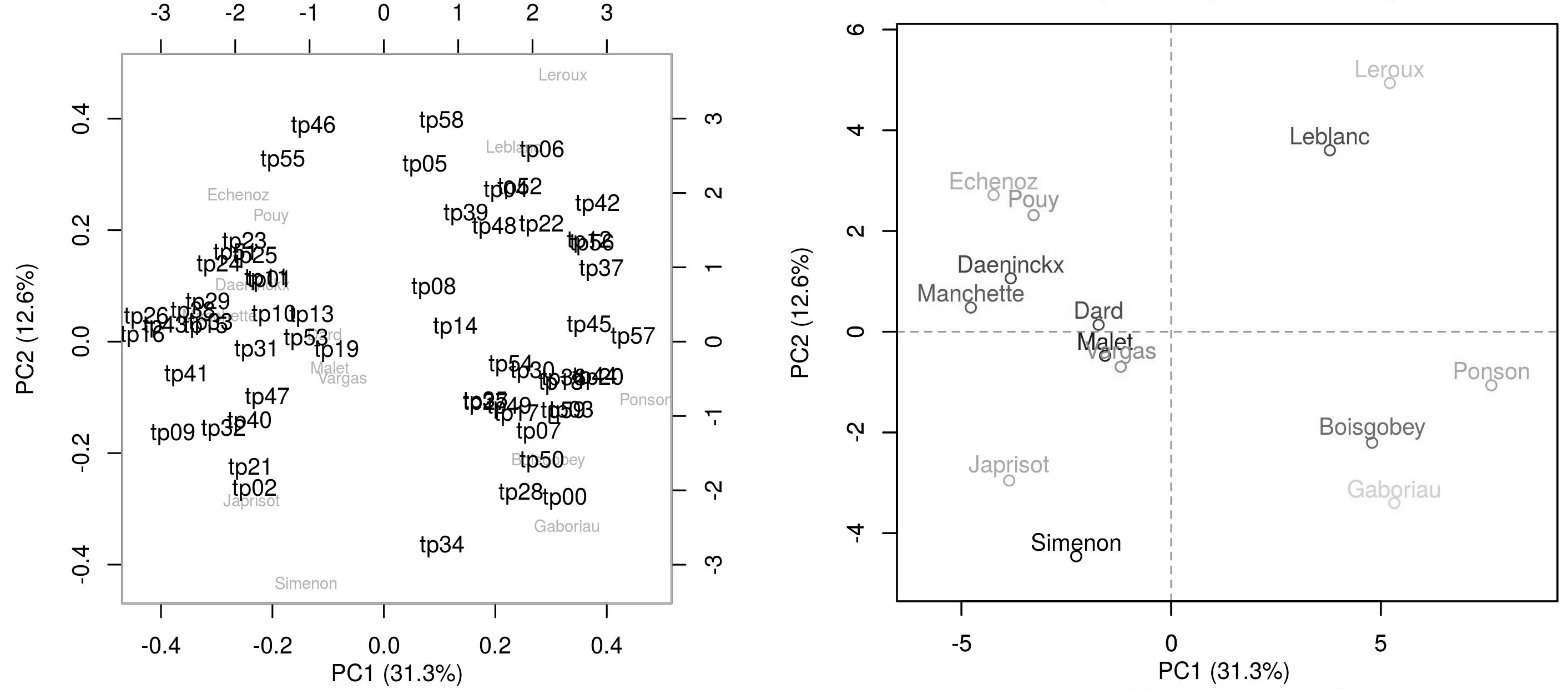
### 

Figure 6. Cumulated absolute differences between topic scores from one decade to the next.

Values above trendline indicate periods of intense topic-related innovation and, possibly, generational shifts (notably, 1880s–1900s and 1930s–1940s). Values below indicate periods of relative continuity (notably, 1910s–1920s and 1940s–1960s). Such results provide a fresh perspective on periodization in literary history.

## Author Similarity

Based on the topic scores per novel, author, or genre, Principal Component Analysis (see Joliffe, 2002) yields groupings of topically similar items, independently of preexisting classifications. The PCA plots in Figure 7 are based on topic scores aggregated to the author

level and have been obtained using the stylo package for R (Eder et al., 2013). The first two components retain large parts of the variation in the data (31.3% and 12.6%, respectively).

### Figure 7. PCA plot of aggregated topic scores per author: authors (right) and loadings (left).

### Not surprisingly for a collection of texts spanning 150 years, the first component is correlated

with time period: authors active before 1930 are located to the east of the plot, those active from 1930 to 1960 close to the middle, and those active after 1960 to the west. A notable exception is Fred Vargas: although writing in the late twentieth century, she appears close to Malet and Dard: the topic-based grouping reveals a link between the now classic ‘roman noir’ and its later reinterpretation by Vargas. Note that although the three authors each have an ‘informal’ topic as their ‘signature topic’ (see above), their proximity remains unchanged even when these topics are removed. Japrisotʼs unique work rightly stands out, but his relative proximity to Simenon remains to be explained.

# Conclusions and Future Work

The results obtained here shed a new light on the history of French crime fiction. Authors, subgenres, and time periods each have distinctive topic characteristics. Some well-understood facts about the genre can be confirmed (e.g., author groupings), but new insights into the genreʼs thematic history also become possible (e.g., thematic innovation cycles). Topic modeling proves to be a valuable tool, providing a fresh perspective on literary history and prompting new interrogations about periodization and the contours of subgenres.

Future work will involve two areas: The text collection will be expanded to reduce correlation between authors, subgenres, and time period. Also, the precise relation between topics and certain categories (e.g., subgenres) will be further investigated using supervised/labeled topic modeling (see McAuliffe and Blei, 2008; Ramage et al., 2009).

# References

**Blei, D. M., Griffiths, T., Jordan, M. I. and Tenenbaum, J. B.** (2004). Hierarchical Topic Models and the Nested Chinese Restaurant Process. In Thrun, S., Saul, L. K. and Schölkopf, B. (eds), *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA.

**Blei, D. M., Ng, A. Y. and Jordan, M. I.** (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**: 993–1022.

**Blevins, C.** (2010). Topic Modeling Martha Ballard’s Diary. http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/.

**Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C. and Blei, D. M.** (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems 22*, pp. 288–96.

**Colin, J.-P.** (1999). *La belle époque du roman policier français. Aux origines d’un genre romanesque*. Delachaux et Niestlé, Lausanne.

**Eder, M., Kestemont, M. and Rybicki, J.** (2013). Stylometry with R: A Suite of Tools. In *Digital Humanities 2013: Conference Abstracts*. Lincoln: University of Nebraska, pp. 487–89.

**Graham, S., Weingart, S. and Milligan, I.** (2012). Getting Started with Topic Modeling and MALLET. The Programming Historian, http://programminghistorian.org/lessons/topic-modeling-and-mallet.

**Jockers, M. L.** (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

**Joliffe, I. T.** (2002). *Principal Component Analysis*. 2nd ed. Springer.

**Lavergne, E.** (2009). *La naissance du roman policier français*. Garnier, Paris.

**Lee, D.D. and Seung, H. S.** (1999). Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature,* **401**(6755): 788–91.

**Lits, M.** (1993). *Le roman policier. Introduction à la théorie et l’histoire d’un genre littéraire*. CÉFAL, Liège.

**Mcauliffe, J. D. and Blei, D. M.** (2008). Supervised Topic Models. In Platt, J. C., Koller, D. Singer, Y. and Roweis, S. T. (eds), *Advances in Neural Information Processing Systems 20*, http://papers.nips.cc/paper/3328-supervised-topic-models.pdf, pp. 121–28.

**McCallum, A. K.** (2002). *MALLET: A Machine Learning for Language Toolkit.* http://mallet.cs.umass.edu.

**Mesplède, C.** (ed.). (2007). *Dictionnaire des littératures policières*. Joseph K., Paris.

**Ramage, D., Hall, D., Nallapati, R. and Manning, C. D.** (2009). Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, pp. 248–56.

**Rehurek, R., Sojka, P.** (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta: ELRA, pp. 45–50.

**Rhody, L. M.** (2012). Topic Modeling and Figurative Language. *Journal of Digital Humanities,* **2**(1), http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/.

**Riddell, A. B.** (2014). Text Analysis with Topic Models for the Humanities and Social Sciences (TAToM). DARIAH-DE, https://de.dariah.eu/tatom/.

**Rubin, T. N., Chambers, A., Smyth, P. and Steyvers, M.** (2012). Statistical Topic Models for Multi-Label Document Classification. *Machine Learning,* **88**(1–2): 157–208, doi:10.1007/s10994-011-5272-5.

**Schmid, H.** (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

**Schmidt, B. M.** (2012). Words Alone: Dismantling Topic Models in the Humanities. *Journal of Digital Humanities,* **2**(1), http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/.

**Steyvers, M. and Griffiths, T.** (2006). Probabilistic Topic Models. In Landauer, T., McNamara, D., Dennis, S. and Kintsch, W. (eds), *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.

**Todorov, T.** (1971). Typologie du roman policier. In *Poétique de la prose*. Paris: Seuil, pp. 55–65.