Topic Patterns in an Academic Literary Journal: The Case Of *Teksty Drugie*

Maciej Maryl
maciej.maryl@ibl.waw.pl
Institute of Literary Research
Polish Academy of Sciences, Poland

Maciej Eder
maciejeder@gmail.com
Institute of Polish Language
Polish Academy of Sciences, Poland

Modelling Literary Scholarship

The availability of digitised full-text resources, as well as bibliographical data in standard database format, has recently opened a new chapter in the sociology of literature by revaluating empirical approaches and data-driven scholarship. The road to this "empirical turn" in literary scholarship has been paved by such scholars as Franco Moretti (2005, 2013) and Matthew Jockers (2013), who showed how empirical data like bibliographical records, annotations, title words, genre categorization, etc., may help in generating new knowledge about literary periods. This approach gathered its momentum as other works exploring the possibility of using such data to answer particular research questions emerged. Due to the shortage of space we will name just a few that have the most influence on this paper, dividing them into three research strands. Firstly, the use of bibliographical data for statistical inferences on literary processes, e.g Bode's (2012) rereading of Australian literary history through the data from AustLit (Australian Literary Bibliography). Secondly, the study of author co-occurrences and mutual references, e.g. visualising literary circles on the basis of such data by Long and So (2013a, 2013b). Thirdly, the application of topic modelling to uncover pertinent issues in literary scholarship, e.g. Goldstone and Underwood's analyses of the evolution of American literary scholarship on the example of PMLA (2012) and seven major literary journals (2014). In combining those approaches into a macroanalytical study of *Teksty Drugie*, we also adopted the rationale introduced by the 40th anniversary internet edition of *Signs*, a literary journal dedicated to feminist criticism.

Aim

The aim of this study is to apply macroanalytical methods to trace the chronology of transformations of Polish literary studies using the example of *Teksty Drugie.* We hypothesise that the collection of papers published in a leading academic journal on literary scholarship can serve as a reliable approximation to chronological changes and/or breaks in Polish literary theory at the turn of the 20th century. We will first trace the topics present in the journal and then analyse them in diachronic perspective. We will focus on the influence of extra-textual events and phenomena on literary scholarship.

We believe that 25 years is a sufficient timespan to observe linguistic differences which are not caused by regular language change. Other projects conducted by the authors of this paper show that a language change (in Polish) typically spans many decades rather than a mere 25 years (e.g. Eder & Górski, 2016). Furthermore, we deal with conventionalised language of scholarship, so the use of certain terms often relates to a given paradigm rather than to a language in general. Nevertheless, we are aware of possible changes of meaning of keywords while interpreting topic models.

Material

Teksty Drugie is a Polish literary journal dedicated to literary scholarship. It has been published since 1990 by the Institute of Literary Research of the Polish Academy of Sciences. It focuses on literary theory, criticism and cultural studies, while also publishing articles by authors from neighbouring disciplines (philosophy, sociology, anthropology). The journal publishes monographic issues dedicated to particular topics or approaches within literary and cultural studies. All those features make it a good example for exploring the vicissitudes of Polish literary scholarship.

The corpus consists of the entire collection of papers published in *Teksty Drugie* (excluding letters, surveys, notes, etc.) in the years 1990–2014 (2,553 texts, 11,310,638 words). The material covering the years 1990–1998 was digitised, OCR-ed, and then manually edited, in order to exclude running heads, editorial comments, and so forth. Obviously, some textual noise – e.g. a certain number of misspelled characters – could not be neutralised. The material from 1999 onwards was digitally-born, but even though a small number of textual issues might have occurred. We believe, however, that distant reading techniques are resistant to small amounts of systematic noise (Eder, 2013).

Given the nature of Polish, which is highly inflected, lemmatization was necessary for a reliable processing of texts. The corpus has been lemmatised with LEM 1.0. (Literary Exploration Machine) developed by CLARIN-PL (see: Piasecki, Walkowiak, Maryl 2017).

Method

To scrutinise the formulated hypothesis, we applied one of the methods of information retrieval that recently attracts a good share of attention in Digital Humanities circles, namely topic modelling in its classical variant known as Latent Dirichlet Allocation (LDA). The method, introduced by Blei (2012), allows for finding co-occurring cohorts of words that presumably reveal (latent) semantic relations.

The experiments were performed using a tailored script in the R programming language, supplemented by the package 'stylo' (Eder et al., 2016) for text pre-processing, and the package 'mallet' (McCallum, 2002) for the actual LDA analysis. A bimodal network of the relations between topics were produced using the software Gephi (Bastian et al., 2009).

Topic modelling relies on the assumption that particular topics are defined by words co-appearing in a given context. Hence, the definition of "context" is crucial to allow for any reliable observations. A few different solutions have been suggested (e.g. Blei, 2012; Jockers, 2013). In our approach, we did not split input texts into smaller samples, which was motivated by the fact that the vast majority of the studies published in Teksty Drugie are rather short.

Other parameters used in the study included: a stop word list containing 327 words (mostly function words, numerals, and very common adverbs), 100 topics extracted in 1,000 iterations, with the obvious caveat that this choice was arbitrary.

Results

A general overview of the obtained results shows a few interesting patterns. Firstly, we analysed and categorised the topics on the basis of their predominant words. The categories are as follows: literary theory (e.g. literature, fiction, text), poetics (e.g. verse, novel, short story, rhetoric) and methodological approaches (e.g. deconstruction, comparative literature, postcolonial studies, psychoanalysis); history of literature (e.g. romanticism, contemporary poets) and cross-cutting research themes (e.g. death, politics, literacy).

A thorough exploration of such models requires a topographical visualisation capable of showing the connections between various topics, which often share a key word (cf. Goldstone and Underwood, 2012). The network (Fig. 1) is too large to be adequately rendered in this paper (a higher resolution image of Figure 1 is available online), yet even without the knowledge about concrete topics presented, we may see (partly thanks to ForceAtlas2 layout, which highlighted this feature) that groups of topics in our corpus are concentrically distributed. This onion-like distribution allows us to distinguish between the central topics (i.e. those who appear in many different papers) and those who appear less often or sporadically and hence are not particularly well-connected with other topics. For instance, in the geometrical centre of the network we may find topics and words pertinent to literary scholarship: literature, literary, comparative literature, national literatures, Jewish studies, fiction, together with some names of contemporary authors. Outliers are also interesting, and could be assigned to 3 groups: (1) expressions in foreign languages, (2) particular research topics or discourses which introduce quite a hermetic language, not shared in other topics, (3) noise (e.g. word bits generated through some errors in OCR).

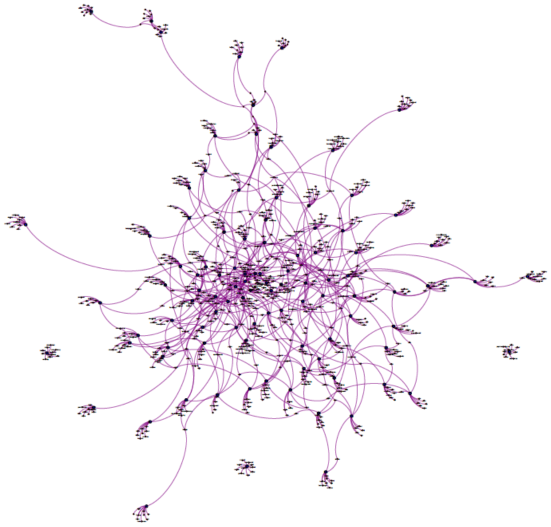

Figure 1. Relationships between topics in Teksty Drugie.

Yet it has to be noted that even the most accurate rendering of the topical distribution is still only a static snapshot insensitive to changes. In order to see the evolution of topics, we need to visualise them on a temporal axis. Due to a shortage of space we present here only a few examples, to show the application of our method. All dot plots are presented below with a trend line based on two period moving average.

Fig. 2 represents the gradual shift of interest from more literature-oriented approaches, to the cultural ones. Both red (topic 19: literature, literary, writer, work) and green (topic 5: literature, research, theory) seem to be dominating until approx. 2007, when the blue line (topic 49: culture, cultural, social) overtakes the green line for the first time. Three years later it becomes the dominant approach, marking the shift in the overall content of Teksty Drugie.

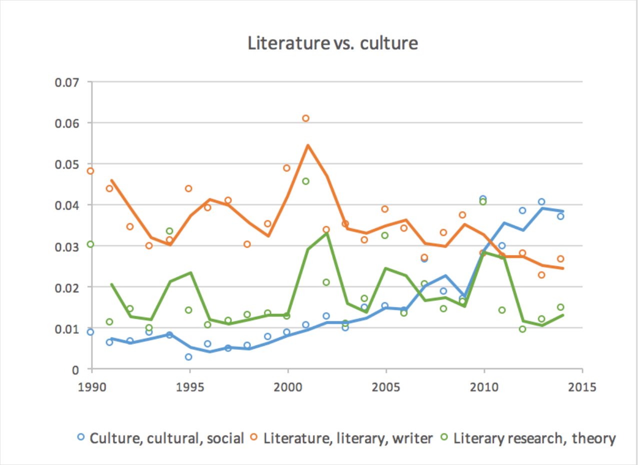

Fig. 2. A temporal distribution of three topics related to literature and culture.

Topic analysis allows us to not only trace the evolution of the journal itself but also to see how the real-world events shape the topics undertaken by literary scholars. Fig. 3 shows the influence of the political transformation in Poland on the content of Teksty Drugie. We see a similar pattern in trends of all topics presented: grey (topic 60: power, society, state, fight, war, law), red (topic 36: political, communism, Polish People's Republic), blue (topic 7: Polish, Pole, national), yellow (topic 94: censorship, exile, novel, positivism, country, London, political). All of them are quite important in the early 1990s and the interest gradually fades until the end of this decade. The spikes around 2001/2002 are caused by the publication of monographic issues which make certain topic more dominant. E.g. Issue No.1-2/2000 was dedicated to socialist realism hence the spike of "communism-related" issue in that year.

This trend shows how political events (namely the transformation and forming of the new democracy) are dominating even the literary scholarship. It could be also the case that more politically charged issues (e.g. history of censorship in Poland) could have been published only after the fall of the communism, hence so many articles in that period.

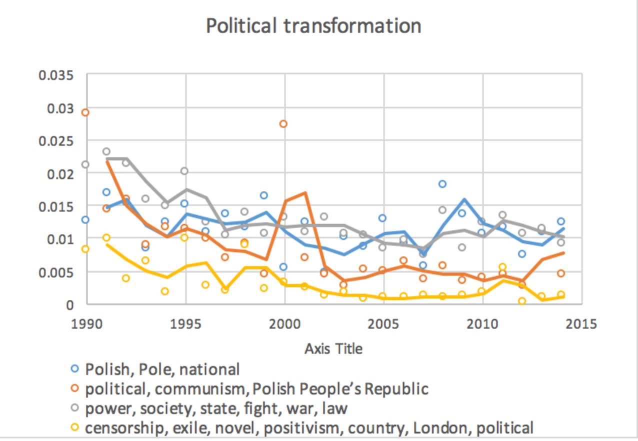

Fig. 3. Temporal shift of topics related to politics.

The last trend we would like to discuss is the emergence of the Holocaust studies in Teksty Drugie. As we can see in the Fig. 4, the red trend line (topic 59: Jew, Jewish, antisemitic) is visible on the fairly same level all through the 25 years, whereas the blue one (topic 18: testimony, Holocaust) is virtually non-existent until 2001.

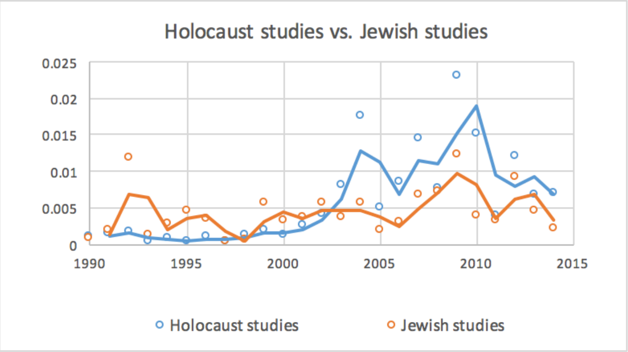

Fig. 4. Temporal distribution of topics related to Jewish studies and the Holocaust.

This sudden boom can be linked to the publishing of the Polish edition of Neighbors by Jan Gross (2000) and the investigation into the role of Polish civilians in the genocide perpetrated in the city of Jedwabne during the World War II. This case opened a long process of re-investigating the troubled Polish-Jewish past, which could be traced also in the issues of *Teksty Drugie*.

Conclusions

In this study we tried to show how extra-textual events influence the content of literary scholarship on the example of Holocaust studies and political transformation, which entailed the prevalence of topics related to politics, power, society, state, and communism in the early 1990s. In the subsequent studies we plan to compare the results of topic modelling with bibliographical data in order to check whether the dominance of a certain topic stems from the large number of scholars who pursue it, or if it instead depends on the fact that a small group of authors published more often than others.

Acknowledgement

The authors wish to thank Dr Tomasz Walkowiak for his extensive help with cleaning and lemmatizing the corpus.

Bibliography

Bastian, M., Heymann, S. and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proceedings of the Third International ICWSM Conference*. San Jose, pp. 361–62.

Bode, K. (2012). *Reading by Numbers: Recalibrating the Literary Field*. London & New York: Anthem Press.

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, 55(4): 77–84.

Eder, M. (2013). Mind your corpus: systematic errors in authorship attribution. *Literary and Linguistic Computing,* 28(4): 603–14.

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal,* 8(1): 107–21.

Eder, M., Górski, R. (2016). Historical Linguistics' New Toys, or Stylometry Applied to the Study of Language Change. In *Digital Humanities 2016: Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków, pp. 182-184.

Goldstone, A. and Underwood, T. (2012). What can topic models of PMLA teach us about the history of literary scholarship?. *Journal of Digital Humanities*, 2(1).

Goldstone, A. and Underwood, T. (2014). The quiet transformations of literary studies: What thirteen thousand scholars could tell us. *New Literary History*, 45(3): 359–84.

Gross, J. T. (2000). *Sąsiedzi: Historia zagłady żydowskiego miasteczka.* Sejny: Fundacja Pogranicze.

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

Long, H. and So, R. (2013a). Network science and literary history. *Leonardo*, 46(3): 274–274.

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu/.

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. New York: Verso.

Moretti, F. (2013). Distant Reading. New York: Verso Books.

Piasecki, M., Walkowiak, T., Maryl, M. (2017). Literary Exploration Machine (LEM 1.0) - New Tool for Distant Readers of Polish Literature Collections. Paper accepted for presentation at ADHO Digital Humanities conference at McGill Universiy, Montreal.

So, R. and Long, H. (2013b). Network analysis and the sociology of Modernism. *Boundary* 2, 40(2): 147–82.