Exploring Textual and Social Hierarchies in Czech Sociological Articles

Radim Hladík[*, **]
[*] Institute of Philosophy of the Czech Academy of Sciences
Prague, Czech Republic
[**] National Institute of Informatics
Tokyo, Japan
radim.hladik@fulbrightmail.org

Abstract

This paper empirically explores the entanglement between rhetoric of scientific texts and intradisciplinary hierarchies in the domain of sociology that is said to embody the dilemma of writing in either more literary or more scientific style. For this purpose, a corpus of literary short fiction is injected into the dataset of research papers published in the journal *Czech Sociological Review*. The sociological articles are subsequently classified into two groups based on the confusion matrix obtained by hierarchical clustering over the bag-of-verbs vector space model of the combined corpus. The sociological papers that are clustered with the works of fiction tend to have higher proportion of female first authors and lower average citation counts. These results are consistent with expectations about the duality in sociological rhetoric and the concomitant social attributes of sociological journal articles.

Keywords: sociology, literature, scientific writing, classification

1. Introduction and related work

Using texts from a Czech sociological journal, this paper sets out to empirically assess the claims about the interactions between rhetoric of scholarly writing and academic power in sociology. Sociology, as a social science, is situated in between two major textual formations: literature and (natural) sciences (Lepenies 1988). According to Bourdieu (1988), the duality of sociological writing – a trade-off between literary elegance and scientific rhetoric of numbers, graphs, and impersonal style – stems from its "double subordination" in the system of scholarly disciplines. Wolfe (1990) suggested that the "two faces" of sociology have institutionalized two publication cultures with different prestige assigned to books and articles. The duality of scientific writing has also been raised in terms of gendered organization of public textuality (Smith 1990). Some sociologists thus explicitly call for the recognition of narrative and personally tinted style as more truthful to the idiosyncratic social experiences that sociology often captures (Adler and Adler, 2008; Richardson, 2002).

Sociological writing has been scarcely studied in quantitative ways by sociologists themselves. Abbott and Barman (1997) made a pioneering study of the rhetorical structure of sociological papers on a small sample of less than 100 papers representing 70 years of a prominent journal's history. Traditional content analysis of sociological journals is a more typical approach (Abend et. al. 2013). The uptake of methods from computational linguistics by sociologists is relatively lagging behind (DiMaggio et. al., 2013). In computational humanities, the vector space model (VSM) of texts explored by the means of principal component analysis (PCA) and hierarchical clustering (HC) are becoming standard techniques to inquire about genres (Schöch, 2017; Vierthaler, 2016). Riguet and Mpouli (2017) explored scientific inspiration of literary criticism. In computational linguistics, scientific writing has been used to test predictions about gender of authors and types of paper (Bergsma et al., 2012) or detecting speculative phrases (Özgür and Radev, 2009). Bag-of-verbs VSM was used for argument mining of legal texts (Falakmasir and Ashley 2017). Mengle et al., (2008) promoted the discovery of related categories through misclassification approach.

2. Data

Central to this paper is the corpus consisting of sociological articles published from 1993 to 2016 in *Český sociologický časopis/Czech Sociological Review* (CSR), a "core journal" of Czech sociology. The texts and metadata were scraped from the website of the journal. Its editorial office provided missing full-texts. Original research papers had to be distinguished from other journal matter (e.g. book reviews or translations) using metadata and manual inspection. Aspects such as an existing abstract, a list of references, or length were taken into account. Ultimately, 523 articles were included in the corpus. A variable capturing the sex of the first authors was added using the gendered suffix of Czech last names and verified manually. Citation data were retrieved from the *Web of Science* database, where 500 of the 523 articles successfully matched their corresponding records against the last names of first authors, year of publication, and the beginning page. The sociological texts were lemmatized and morphologically tagged using the MorphoDiTa software (Straková et al., 2014).

The short fiction texts used in this study to gauge the rhetoric features of sociological writing were extracted from a ready-made corpus of written Czech (Křen et al., 2016). All texts marked as non-translated in the category of short fiction were chosen.

The final corpus was assembled from the combination of the two corpora (Table 1). For each document, word tokens (numerals, proper names, and punctuation marks were removed), their lemma and part-of-speech tags are known.

Corpus	Docs	Tokens	Lemma	Years
Short fiction	153	7977791	79401	1991-2014
CSR articles	523	2472222	46401	1993-2016

Table 1: Short Fiction & CSR Articles: Combined Corpus

3. Analysis and Results

The initial steps in the analysis involved finding a representation of the unified corpus that would sufficiently capture differences in rhetorical style but would not dichotomize its two main constituents so as to effectively preempt any misclassification. The suitability of models was assessed through examination of results

derived from principal component analysis (PCA). A simple bag-of-verbs representation with terms 80% or more sparse removed satisfied the research design (Figure 1) and resulted in mere 10 verb features. The reduction of features through the removal of sparse terms emphasizes communalities among texts, while verbs, as content words, maintain considerable discriminatory power. Still, the non-sparse verbs – with notable presence of modals – arguably serve to organize textual narrative and arguments rather than bearing directly on topics.

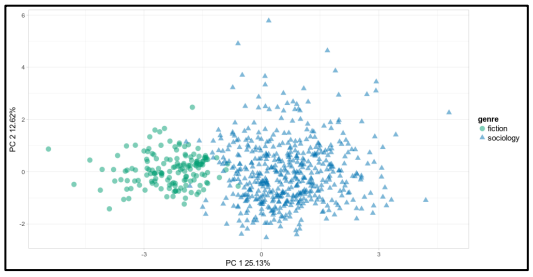

Figure 1: PCA of the Short Fiction and CSR Articles Combined Corpus (Tf-Idf VSM, 10 Verbs)

For exploratory purposes, hierarchical clustering (HC) is complementary to PCA. In this particular implementation, HC based on Euclidean distances and Ward's algorithm separates the sociological corpus into two groups (Table 2). The cluster merging rule for class membership is simple: an article belongs to *literary sociology* (n=124) if it is found in a cluster where works of fiction also appear. Conversely, an article falls in the group of *exclusively sociology* (n=399) if it sits in a cluster containing only sociological articles. For the herein reported results, the hyperparameter *k* was set to 10, i.e. the combined corpus was divided into 10 clusters. Experiments with Manhattan distance measure and different levels of *k* for HC did not substantially impact the trends described in this paper.

Cluster	1	2	3	4	5	6	7	8	9	10
Fiction	52	62	23	16	0	0	0	0	0	0
Sociology	1	25	1	97	58	92	79	63	92	15

Table 2: Confusion Matrix of Two Genres after HC (Merged Clusters of CSR Articles Are Color Coded)

In the remainder of the paper, the focus shifts to the differences between the two groups of sociological articles established above. The space constraints limit reporting to two factors: sex of the first author and citation counts. Both of the attributes indicate intradisciplinary hierarchies and it is expected that the less distinctively sociological rhetoric would be marginalized.

1st Author Group	Male	Female	Total	% Female
Exclusively sociology	283	116	399	29.07
Literary sociology	75	49	124	39.52
Total	358	165	523	31.55

Table 3: The Distribution of First Authors by Gender (Fisher's Exact Test for the Shaded 2x2 Area: $p < 0.05$)

Before evaluating the distribution of the sex attribute, a baseline must be established. Table 3 shows that men dominate as first authors over the entire sociological corpus, with less than a third of first authors being women, indicating that men constitute a priviledged category of authors. Although the monitored groups of articles were separated solely on the basis of their particular usage of 10 common verbs, Table 3 reveals that the percentage of female authors is higher than the baseline in the group designated as *literary sociology*.

Another perspective on the same issue of intradisciplinary hierarchy can be taken by considering the distribution of specifically scientific prestige between the two groups. We resort to citation counts as a standardly used proxy for the recognition that publications receive from the scientific community. The data had to be reduced to 500 articles for which the citation records were obtained.

In Figure 2, citation counts for both groups of CSR articles observe similar power law pattern, with most articles receiving zero or only a few citations. However, the patterns are not entirely identical. *Exclusively sociological* (n=380) articles display a heavier and longer tail in the distribution. It has the upper hand in the ranks of higher citation counts and includes nearly all highly cited papers. The *literary* group (n=120) prevails in the lower ranks of citation counts; articles with 0 or 1 citation account for more than 58% of the *literary* styled papers. Citation counts thus reinforce our previous notion of the hierarchy between the two groups. The average citation rates per article are 1.84 and 3.48 for the *literary* and *exclusively sociological* groups respectively. Mann-Whitney U test confirms the two groups as significantly distinct at the level of p-value < 0.005.[1]

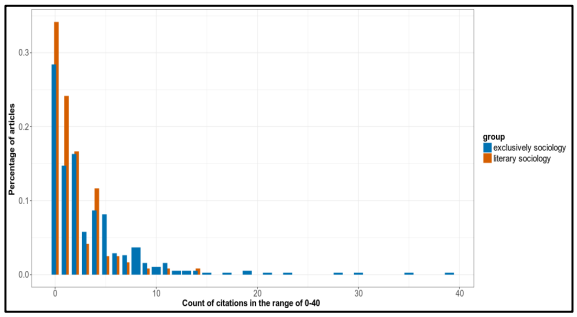

Figure 2: Proportions of Articles by Citation Counts in the Two Groups of CSR Articles

4. Conclusion

The analysis demonstrated that even in the absence of labels for rhetorical categories of interest in sociological writing, these can be detected with the help of a confounding factor introduced into the bag-of-verbs vector space model in the form of another corpus made up of literary works. The results advance our understanding of sociological corpora as potentially containing specific subgenres of disciplinary research papers. Further research could lead to building predictive models to distinguish the heuristically identified two types of sociological writing. The obtained knowledge is possibly extensible to other academic disciplines as the paper reminds us about the socially constructed character of categories and boundaries that are sometimes taken for granted in machine learning literature and other applications of classificatory approaches.

[1] Although the study deals with what are essentially population data and the reported values are, in that sense, exact, significance tests were included upon a reviewer's request. The author considers it admissible under a superpopulation assumption.

5. Acknowledgements

The author conducted the work as the overseas researcher under Postdoctoral Fellowship of Japan Society for the Promotion of Science and with the support by Grant-in-Aid for JSPS Fellows no. 17F17769.

6. Bibliographical References

Abbott, A., & Barman, E. (1997). Sequence Comparison Via Alignment and Gibbs Sampling: A Formal Analysis of the Emergence of the Modern Sociological Article. *Sociological Methodology*, *27*(1): 47–87.

Abend, G., Petre, C., & Sauder, M. (2013). Styles of Causal Thought: An Empirical Investigation. *American Journal of Sociology*, *119*(3): 602–654.

Adler, P. A., & Adler, P. (2008). Of Rhetoric and Representation: The Four Faces of Ethnography. *Sociological Quarterly*, *49*(1): 1–30.

Bourdieu, P. (1988). *Homo Academicus*. Cambridge, UK: Polity Press.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding. *Poetics*, *41*(6): 570–606.

Falakmasir, M. H., & Ashley, K. D. (2017). Utilizing Vector Space Models for Identifying Legal Factors from Text. *Frontiers in Artificial Intelligence and Applications*, 183–192.

Lepenies, W. (1988). *Between Literature and Science: The Rise of Sociology*. Cambridge, UK: Cambridge University Press.

Mengle, S. S. R., Goharian, N., & Platt, A. (2008). Discovering Relationships Among Categories Using Misclassification Information. In *Proceedings of the 2008 ACM Symposium on Applied Computing* (pp. 932–937). New York, NY, USA: ACM.

Richardson, L. (2002). Writing Sociology. *Cultural Studies ↔ Critical Methodologies*, *2*(3): 414–422.

Riguet, M., & Mpouli, S. (2017). At the Crossroads Between the Scientific and the Literary Discourse: Comparison as a Figure of Dialogism. *Digital Scholarship in the Humanities*, *32*(suppl_2): ii60-ii77.

Schöch, C. (2017). Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. *Digital Humanities Quar*terly, *11*(2).

Smith, D. E. (1990). *The Conceptual Practices of Power: A Feminist Sociology of Knowledge*. Boston: Northeastern University Press.

Vierthaler, P. (2016). Fiction and History: Polarity and Stylistic Gradience in Late Imperial Chinese Literature. *Journal of Cultural Analytics*.C

Wolfe, A. (1990). Books vs. Articles: Two Ways of Publishing Sociology. *Sociological Forum*, *5*(3): 477–489.

7. Language Resource References

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., … Zasina, A. (2016). *SYN v4: Large Corpus of Written Czech*. LINDAT/CLARIN Digital Library at the Institute of Formal and Applied Linguistics, Charles University.

Straková, J.J., Straka, M., & Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 13–18). Baltimore, Maryland: ACL.