Where Close and Distant Readings Meet: Text Clustering Methods in Literary Analysis of Weblog Genres

Maciej Maryl, Maciej Piasecki, Ksenia Młynarczyk

Where Close and Distant Readings Meet: Text Clustering Methods in Literary Analysis of Weblog Genres

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

Maciej Piasecki
maciej.piasecki@pwr.edu.pl
Wrocław University of Technology

Ksenia Młynarczyk
ksenia.mlynarczyk@gmail.com
Wrocław University of Technology

Problem: towards a non-topical classification of weblog genres

The existing typologies of weblog genres - both popular and academic - are based on the blog topic, e.g. cooking blogs, travels, business (cf. Morrison 2008) or its medium, e.g. vlogs, picture logs (cf. Herring et al. 2005). In order to go beyond topical distinctions, Maryl, Niewiadomski and Kidawa (2016) conducted an interpretive study on the sample of 322 popular Polish blogs. They adopted a new-rhetorical approach, basing on Carolyn Miller's (1994) concept of genre as a social action, concentrating mostly on the blog's communicative purpose and functions. Following the principles of the grounded theory (cf. Lonkila 1999) the team interpreted those blogs and created an empirical-conceptual typology which entailed following genres: diaries (subjective, self-referential discourse), reflection (subjective discourse on universal matters), criticism (subjective and expert discourse on general issues), information (objective facts), filter (gateway to the existing web content), advice (subjective and expert instructions on particular issues), modelling (serving as a role model for readers) and fictionality (description of fictional events). Weblogs in the sample were coded by three separate coders with 69% average pairwise percent agreement and Cohen's kappa of .622[1]. Such a moderate agreement could be attributed to the fact that the resulting genres are ideal types, and most of the actual blogs share features of more than one genre.

This subsequent study aims at supplementing this close-reading typology with a distant-reading perspective (Moretti 2013), based on selected tools for language processing and text clustering. We explore the style of those genres, adapting the definition proposed by Herrmann et al.: "Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively" (2015:41). We chose this approach due to its stress on mixed methods, as we are combining linguistic and literary criteria of selecting style markers to discriminate between blog genres (Leech & Short 2007,57-58). Current research in the field of computational literary genre stylistics focuses on Most Frequent Words (e.g. Schöch and Pielström 2014; Jannidis & Lauer 2014) or functional linguistic categories (or both) (e.g. Allison et. al 2011). Yet, this study applies similar methods to emergent and uncategorised forms of writing. The quantitative methods are incorporated into the qualitative research workflow in order to create a productive feedback loop.

Corpus

The corpus of blogs was collected with the use of *BlogReader* - an extension of a corpus gathering system developed in CLARIN-PL (Oleksy et al. 2014) on a basis of open components: *jusText* and*Onion* (Pomikálek, 2011). From the initial set analysed by Maryl et al., 250 blogs were selected for processing as being long enough and included clean text (comments were omitted). We intentionally left out blogs with exceptionally large or small amount of text in order to balance the sample. The selected subcorpus includes: Diaries (44 blogs), Reflection (12), Criticism (73), Information (10), Filter (11), Advice (59), Modelling (24), Fictionality (5), and 10 'Unblogs', i.e. websites or portals using the label of blogs. Posts from the one blog were merged together into a single text document per a blog that was saved in the CCL corpus format (Broda et al., 2012).

Processing

We followed the blueprint of stylometry to find groups of blogs, e.g. (Burrows 2002), (Stamatatos, 2009) or (Eder, 2011). Blogs were described by feature vectors whose initial values were frequencies of the selected elements. They were next filtered or transformed. The transformed vectors were clustered into a number of groups that could be presented as automatically identified blog types or compared with the original types.

According to the criteria considered for the typology of blogs, we assumed that the interesting distinctions are not of semantic character. Thus we tried to define descriptive features that are not sensitive to the semantics of the blog contents. As a consequence, we have analysed features based on frequencies of lemmas, grammatical classes and sequences of grammatical classes. The brief description below will be elaborated in the presentation:

1. We have selected the 500 most frequent lemmas from the *Polish National Corpus* (Przepiórkowski et al., 2012) and in the series of experiments on the corpus of novels we reduced it to 212 lemmas that did not trigger semantic grouping (e.g. filtering out most of nouns and verbs).

2. Grammatical classes (as defined in the *Polish National Corpus* tagset) were recognised by *WCRFT* morpho-syntactic tagger (Radziszewski, 2013).

3. Features were defined and extracted with the help of the *Fextor* system (Broda et al., 2013).

4. Raw feature values were transformed by measures returning positive results for those features which contribute the significant amount of information to the document description. *SuperMatrix* system (Broda & Piasecki, 2013) for Distributional Semantics was applied during the transformation.

5. Similarity of the transformed vectors were computed by the cosine and ratio measures. The first is not sensitive to the differences in the document lengths that was the case of the analysed collection. The ratio as a heuristic measure that is aimed at comparing how much information is shared by the two vectors:

$$\mathrm{ratio}(V,U) = 2{*}\mathrm{sum}((Vi + Ui)/\max(Vi, Ui) - 1) / (\mathrm{length}(V) + \mathrm{length}(U))$$

6. Clustering was performed by the *Cluto* package for text data clustering (Zhao & Karypis 2005). In addition, *Stylo* package (Eder et al., 2013) for stylometry was used in experiments with visualisation of the possible blog clusters.

In order to understand the clusters better, most significant features for each cluster were identified and ranked. From several tests the Mann-Whitney U nonparametric test was chosen. For each feature its values in the documents of the given cluster were compared with its values in documents from the rest of the collection.

Experiments

We have performed several experiments that can be divided into three main groups:

1. *lexical level analysis*, based solely on the selected most frequent lemmas and punctuation marks and aimed at testing whether those properties can serve as a basis for automated identification of the blog types;

2. *lexico-syntactic level analysis* featured grammatical classes in combination with the lexical features of the lexical analysis in order to assess whether blog styles result in syntactic properties. On both levels we set the expected number of clusters to 20, in order to give algorithm more 'freedom';

3. extraction of significant features for the blog types with the help of the Mann-Whitney U nonparametric test.

Discussion

The generated groups represented relatively high average of clusters purity: 54%-60,4%, i.e. more than 50% blogs in a cluster are of same type. Entropy was higher than expected: 0.438-0.481, i.e. besides dominating types in clusters blogs of other types were scattered (especially smaller types). However, the obtained clusters did not match very well the qualitatively defined types. Lexical analysis combined with the ratio measure produced results that were closest to the qualitative types: entropy of 0.467 and 58% purity, see Figure 1. Yet, lexico-syntactic analysis (lexical features together with grammatical classes and bigrams) yielded better results: 0.438 of entropy and 60.4% of purity, see Figure 2. A slightly worse result: 0.481 of entropy and 54% of purity, was obtained with trigrams instead of bigrams - groups became too small and too specific.

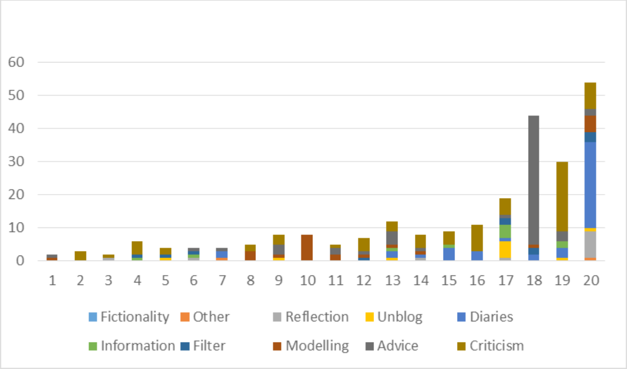

Figure 1. Results of the lexical analysis (features: 212 selected frequent lemmas, punctuation marks), PMI weighting, the ration similarity and, graph clustering algorithm from Cluto

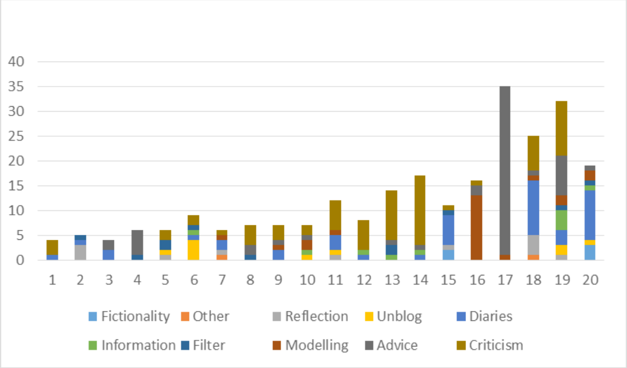

Figure 2. Results of the lexico-syntactic analysis (lexical features plus grammatical classes and bigrams), PMI weighting, ration similarity, graph clustering algorithm

Such genres as advice, criticism and, to certain extent, diaries and modelling were clustered together with others present in multiple clusters. It was caused by distinctive language features of those genres, especially of the advice, which employs instructional vocabulary, or criticism, due to its essayistic style with compound sentences and conjunctions reflecting logical reasoning. Diaries tend to use narrative language, whereas modelling blogs are clearly concentrated on expressing the author's self.

Those differences were further explored through the extraction of blog types' significant features with the use of Mann-Whitney U statistic. The results were in line with the definitions of classes, but provided more detailed information about the linguistic cues in those genres, some of which are presented in Table 1.

Conclusions

This study showed how close readings (literary interpretative practices) and distant readings (computational approaches to genre analysis) could be integrated in a non-topical analysis of the emerging genres. The novelty of the presented approach lies in the fact that we do not aim at assessing existing genres but rather at developing tools and procedures for the analysis and classification of new genres. The automated methods are used not only to verify the qualitative findings, but rather to enhance them by pointing towards the attributes which might have been overlooked by human coders who were able to read only a sample of each of 332 blogs. The aim is not to cluster texts automatically but rather to support human interpretation in an integrated research design.

Recurring problems with clustering genres other than advice could be attributed to the fact that individual blogs within one class may consists of posts which follow different genre conventions. Hence, further studies should explore the genre problem by comparing individual posts (rather than entire blogs) by different authors in order to find stylistic similarities.

Genre	Linguistic features
Advice	infinitive, passive adjectival participle, numerals, measurements (“about”, “large”, “small”)
Criticism	subjective vocabulary: „I”, „mine”; conjunctions pointing to logical reasoning, e.g. “if”, “that”, “given”, “hence”, “but”
Diaries	1st & 2nd person; vocabulary: “self”, “to be”; specific words and verb forms pointing out to a narrative: “certain”, “there”
Fictionality	past tense, 3rd person
Filter	punctuation, substantives
Information	impersonal verb forms, 3rd person
Modelling	interjections (e.g. “eh”), exclamation marks, 1st & 2nd person, vocabulary: “mine”, “thing”, “new”, “why”, “because”
Reflection	1st & 2nd person, vocabulary: “self”, “always”, “everything”

Table 1. Selected linguistic features of weblog genres (Mann-Whitney U)

Bibliography

Allison, S., Heuser, R., Jockers, M. L., Moretti, F. and Witmore, M. (2011). *Quantitative Formalism: An Experiment*, Pamphlet 1. Stanford Literary Lab.

Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski, A. and Wardyński, A. (2012). *KPWr: Towards a Free Corpus of Polish*, *Proceedings of LREC'12*. Istanbul, Turkey.

Broda, B., Kędzia, P., Marcińczuk, M., Radziszewski, A., Ramocki, R. and Wardyński, A. (2013). Fextor: A Feature Extraction Framework for Natural Language Processing: A Case Study in Word Sense Disambiguation, Relation Recognition and Anaphora Resolution. In Przepiórkowski, A., Piasecki, M., Jassem, K. and Fuglewicz, P. (eds), *Computational Linguistics. Applications*, volume 458 of Studies in Computational Intelligence, Berlin: Springer Verlag, pp. 41–62.

Broda, B. and Piasecki, M. (2013). Parallel, Massive Processing in SuperMatrix – a General Tool for Distributional Semantic Analysis of Corpora. *International Journal of Data Mining, Modelling and Management*, **5**(1): 1–19.

Burrows, J. F. (2002). ‘Delta’: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Eder, M. (2011). Style-markers in authorship attribution: a cross-language study of authorial fingerprint. *Studies in Polish Linguistics*, **6**: 99-114.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. *Digital Humanities 2013: Conference Abstracts*. University of Nebraska-Lincoln, NE, pp. 487-89.

Freelon, D. (2010). ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science*, **5**(1): 20-33.

Herring, S., Shedit, L. A., Writh, E. and Bonus, S. (2005). Weblogs as a bridging genre, *Information, Technology & People*, **18**(2): 142-71.

Herrmann, J. B., van Dalen-Oskam, K. and Schöch, Ch. (2015). Revisiting Style, a Key Concept in Literary Studies, *Journal of Literary Theory*, **1**(9): 25-52.

Jannidis, F. and Lauer, G. (2014). Burrows's Delta and Its Use in German Literary History. In Erlin,M. and Tatlock, L. (eds), *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*, New York: Camden House, pp. 29-54.

Leech, G. N. and Short, M. (2007). *Style in fiction: A linguistic introduction to English fictional prose*. Harlow: Pearson Longman.

Lonkila, M. (1999). Grounded theory as an emerging paradigm for computer-assisted qualitative data analysis. In Kelle, U. (ed), *Computer-Aided Qualitative Data Analysis: Theory, Methods and Practice*. London: Sage, pp. 41-51.

Maryl, M., Niewiadomski, K. and Kidawa, M. (2016 - forthcoming). Empirically Generated Typology of Weblog Genres. *CLCWeb: Comparative Literature and Culture*, **18**(2).

Miller, C. R. (1994). Genre as Social Action. In Freedman, A. and Medway, P. (eds), *Genre and the New Rhetoric*. London: Taylor & Francis, pp. 57-66.

Moretti, F. (2013). *Distant reading*. London: Verso.

Morrison, A. (2008). Blogs and Blogging: Text and Practice. In Siemens, R. and Schreibman, S. (eds), *A Companion to Digital Literary Studies*. Oxford: Blackwell, pp. 369-87.

Oleksy, M., Kocoń, J., Maryl, M. and Piasecki, M. (2014). Linguistic analysis of weblog genres, *Practical Applications of Linguistic Corpora Conference*, PALC'14, Łódź.

Pomikálek, J. (2011). *Removing Boilerplate and Duplicate Content from Web Corpora*. PhD Thesis, Faculty of Informatics, Masaryk University, Brno. http://is.muni.cz/th/45523/fi_d/phdthesis.pdf (accessed 29 February 2016).

Przepiórkowski, A., Bańko, M., Górski, R. L. and Lewandowska-Tomaszczyk, B. (eds). (2012). *Narodowy Korpus Języka Polskiego*. Warszawa: PWN.