Quantitative Analysis of Character Networks in Polish XIX and XX Century Novels

Marek Kubis
mkubis@amu.edu.pl
Adam Mickiewicz University, Poland

final version available at:
https://dev.clariah.nl/files/dh2019/boa/0843.html

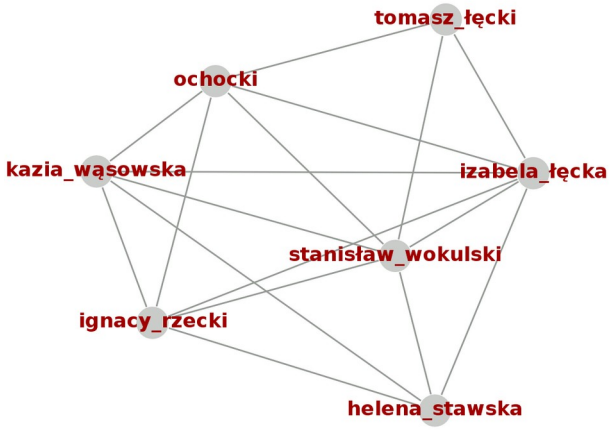

Figure 1. Main characters of *Lalka* by B. Prus

Introduction

From qualitative work of Moretti (2013) on Shakespeare's plays and Chinese novels, through quantitative works on the 19th century English literary fiction by Elson et al. (2010) and Jayannavar et al. (2015) to the investigation of dynamic plots of German plays by Fischer et al. (2017), the analysis of social networks induced from literary works became a valuable tool in digital humanities research. This paper presents a study on induction and quantitative analysis of character networks inferred from Polish novels. The corpus gathered for this study is an order of magnitude larger than the collection of novels used by Elson et al. (2010) and Jayannavar et al (2015). It contains primarily novels from the second half of the 19th century and the first half of the 20th century. The main goal of this paper is to present novel results on systematic differences between the 19th century and 20th century Polish prose with respect to the collected corpus. The two by-products of this research are:

1. The development of fully automatized, quantitative pipeline that leads from raw Polish text to the set of testable hypotheses.
2. The reproduction of the observations of Elson et al. (2010) and Jayannavar et al. (2015) on a larger, more demanding corpus of a different language origin that contains both the 19th century and 20th century works.

Data collection

In order to build the corpus I utilized two digital libraries that offer texts in Polish. The Polona digital library (Polona, 2018) which is maintained by the National Library of Poland offers digitized copies of printed books. I have managed to fetch from Polona around 3000 volumes that are in public domain and are available online in a form of OCR-ed text. The multi-volume editions of novels fetched from Polona were merged resulting in ca. 2300 complete pieces of literary fiction. The second source of texts for the corpus is the Wolne Lektury library (Wolne Lektury, 2018) that focuses on school readings and offers carefully revised electronic editions of books that are in public domain. Around 230 novels were available for download from Wolne Lektury at the time of writing.

In order to make the corpus representative I decided to select exactly one (the most recent) edition of every novel that has authorship attributed, resulting in 1555 unique pieces of work. Due to the sparsity of available data I restricted my attention to novels created between 1800 and 1945, obtaining 1443 volumes in result. Since the focus of this research is on Polish novels only I have selected from the corpus the books that have Polish origin according to the catalog of the National Library of Poland (Biblioteka Narodowa, 2018). Thus, the corpus used in this paper consists of 930 novels (392 from the 19th century and 538 from the first half of the 20th century).

Before feeding the texts from the corpus into the network induction procedure described

in the following section some preliminary processing is required. In case of Polona texts word segmentation errors introduced by OCR have been fixed by a custom normalization script. Furthermore, since the part-of-speech (POS) and named entity recognition (NER) taggers used for the network induction are trained on a corpus of modern Polish language (Przepiórkowski et al., 2012), I have applied a diachronic normalizer (Graliński, 2018; Jassem et al., 2017) in order to contemporize the Polona texts for the purpose of improving the part-of-speech and named entity recognition accuracy. Finally, the novels are split into paragraphs by another script. The texts from Wolne Lektury are checked for errors and contemporized by library editors before publication, hence beside splitting them into paragraphs in accordance to the XML schema no further processing is required.

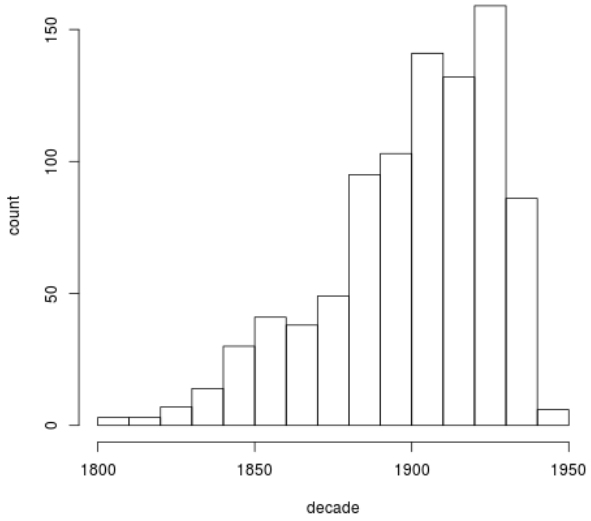

Figure 2. Novels in corpus by decade

Network induction

Before the conversation network can be inferred from a novel, the text has to be passed to the annotation pipeline that appends additional data necessary for the network induction procedure. The annotation pipeline splits the paragraphs of text into sentences and tokens. Then, the text is lemmatized with help of the Polimorf dictionary (Woliński et al., 2012). Afterwards, the text is annotated by POS and NER modules that were trained using the manually annotated 1-million word subcorpus of the NKJP corpus (Przepiórkowski et al., 2012). The final step of the annotation pipeline is the detection of dialog boundaries. A script with hand-crafted rules that take into consideration possible shapes of beginnings, endings and internal paragraphs of dialogs is used for this purpose. The script failed to extract dialogs from 23 books, thus the networks have been inferred for 383 19th century novels and 524 20th century novels, respectively.

The network induction procedure iterates over dialogs identified in the novel. The dialog turns are surface parsed in order to detect speakers and distinguish them from other named entities that are referenced in the dialog, but do not talk. Since characters can be referred in text in different ways[1], the detected speaker mentions are passed to the entity resolution module which is responsible for assigning a common identifier to all mentions of the same character on the basis of the dialog history and the plot. Finally, a (conversational) link is created in the network for every pair of identified speakers that participate in the dialog.

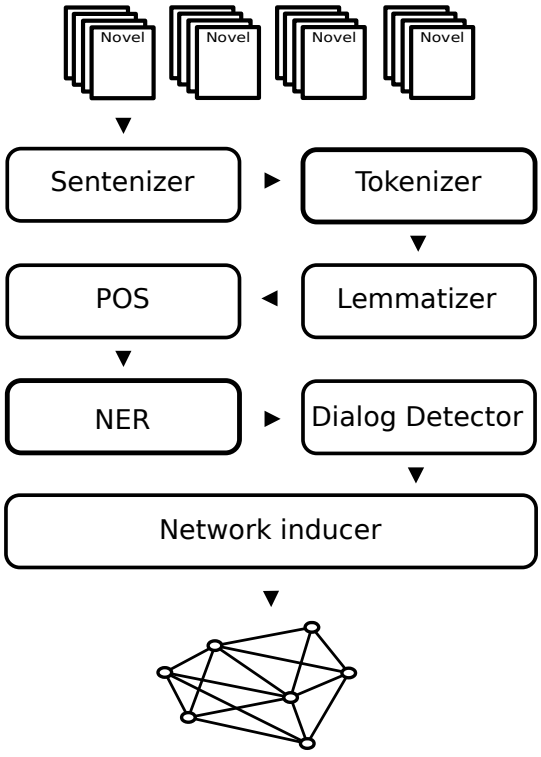

Figure 3. Annotation pipeline

Results

Elson et al. (2010) made a distinction between characters and speakers (characters that took

1 E.g. by their names, surnames, first names, diminutives, honorifics.

part in at least one conversation). I decided to focus the study on speakers only since the NER module identified many entities that are not relevant to the conversations such as historical figures.[2]

The conversation networks induced from the entire corpus have 34 (σ=25.42) characters on average. The mean number of dialogs is 191 and the average number of conversation links between characters is 80. The average number of conversations that a character is involved into[3] is 2.28**.** I have run the Walktrap community detection algorithm (Pons and Latapy, 2005) and found the mean number of communities to be 4.68 and the average size of a community to be 5.68[4].

Network metric	19th century	20th century	All
character count	36.87 ±27.48	31.48 ±23.56	33.75 ±25.42
dialog count	182.60 ±168.07	196.27 ±158.08	190.50 ±162.42
link count	89.56 ± 109.23	73.16 ±94.08	80.09 ±101.03
average degree	2.35 ±1.81	2.23 ±1.52	2.28 ±1.65
community count	5.13 ±3.49	4.35 ±2.83	4.68 ±3.14
community size	5.83 ±2.94	5.56 ±2.55	5.68 ±2.73

Table 1. Network properties

(Elson et al., 2010) and (Jayannavar et al., 2015) reported that the number of characters in the novel is correlated with the properties of the inferred network. As can be expected the same observation holds for the conversation networks induced from Polish novels. The number of characters is strongly correlated to the number of dialogs (r=0.73) and the number of conversational links (r=0.86). Furthermore, The number of communities is strongly correlated to the number of characters (r=0.83). In contrast to (Jayannavar et al., 2015) I found the number of characters to be positively[5] correlated to the average number of interlocutors that a character has (r=0.44) and the average community size (r=0.43). This result may be due to the definitional difference between conversational links used in this paper and interaction links used by Jayannavar et al.

Network metric	19th century	20th century	All
dialog count	0.74	0.73	0.73
link count	0.84	0.88	0.86
average degree	0.35	0.54	0.44
community count	0.83	0.83	0.83
community size	0.40	0.46	0.43

Table 2. Network metrics correlated with character number

Having the corpus that contains comparable number of the 19th and 20th century novels, I decided to check if properties of networks change systematically between the centuries. Since network metrics are not normally distributed[6], I have used the Mann-Whitney test to verify the hypothesis that it is equally likely that a randomly selected novel from the 19th century subcorpus has a lower or higher value of the network metric being tested than a randomly selected novel from the 20th century subcorpus. This hypothesis has been rejected in case of the character number, dialog count, link count and community count metrics and maintained in case of the average degree and average community size (cf. Table 3). These results suggest that (at least with regard to the collected corpus) the prose of the first half of the 20th century became richer in dialogue, but at the same time focused on smaller sets of characters.

Hypo-thesis	19th century median	20th century median	p-value	0.95 conf. interval
character count	30.0	26.0	**0.00098**	[2.00, 6.00]

2 Plus some false positives that definitely are not named entities.

3 Average node degree according to graph terminology.

4 Excluding communities of size 1 that are found by the walktrap algorithm.

5 Instead of negatively.

6 Normality of all network metrics discussed in the paper is rejected acording to Shapiro-Wilk test.

Hypo-thesis	19th century median	20th century median	p-value	0.95 conf. interval
dialog count	132.0	161.5	**0.00671**	[-34.0, -6.0]
link count	54.0	43.0	**0.00370**	[3.00, 15.00]
average degree	1.9	1.9	0.68892	[-0.14, 0.21]
comm. count	4.0	4.0	**0.00020**	[0.000068, 1.00]
comm. size	5.2	5.0	0.26845	[-0.117, 0.49]

Table 3. Network metric change between centuries

Final Remarks

The systematic differences between the 19th and 20th century Polish novels presented in this paper are interesting on their own, but they can also initiate further computational investigations of the character networks. One aspect of the problem that is not covered by this study and should be examined in the future is the impact of indirect speech on the structure of character networks. Another issue that should be taken into consideration is the verification if the proposed hypotheses still hold in presence of the constantly growing corpus of digitized literary works of Polish origin[7].

References

Biblioteka Narodowa (2018). Katalogi Biblioteki Narodowej. Available at: http://katalogi.bn.org.pl/ (Accessed: 27 November 2018).

Elson, D. K., Dames, N. and McKeown, K. R. (2010). Extracting social networks from literary fiction. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 138–47.

Fischer, F., Göbel, M., Kampkaspar, D., Kittel, C. and Trilcke, P. (2017). Network Dynamics, Plot Analysis: Approaching the Progressive Structuration of Literary Texts. *Digital Humanities 2017*. Montréal: McGill University, pp. 437-40.

Graliński, F. (2018). lucene-transducers. Available at: https://gonito.net/gitlist/lucene-transducers.git/ (Accessed: 27 November 2018).

Jassem, K., Graliński, F., Obrębski, T. (2017). Pros and Cons of Normalizing Text with Thrax. *Proceedings of the 8th Language and Technology Conference*, Poznan, Poland, Fundacja Uniwersytetu im. Adama Mickiewicza, pp. 230-235.

Jayannavar, P., Agarwal, A., Ju, M. and Rambow, O. (2015). Validating Literary Theories Using Automatic Social Network Extraction. *Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature*, pp. 32–41

Moretti, F. (2013). Network Theory, Plot Analysis. *Distant Reading.* London: Verso.

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

Polona (2018). About Polona website. Available at: https://polona.pl/page/about-polona/ (Accessed: 27 November 2018).

Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks, Available at: http://arxiv.org/abs/physics/0512106 (Accessed: 27 November 2018).

Woliński, M., Miłkowski, M., Ogrodniczuk, M., Przepiórkowski, A. and Szałkiewicz, Ł. (2012). PoliMorf: a (not so) new open morphological dictionary for Polish. *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul: ELRA, pp. 860–64

Wolne Lektury (2018). About the project. Available at: https://wolnelektury.pl/info/o-projekcie/ (Accessed: 27 November 2018).

7 Polona claims to add up to 2000 digital objects such as books, photographs and postcards on the daily basis (Polona, 2018).