Towards a contextualised spatial-diachronic history of literature: mapping emotional representations of the city and the country in Polish fiction from 1864 to 1939

Agnieszka Karlińska
Institute of Literary Research of the Polish Academy of Sciences

Cezary Rosiński
Institute of Literary Research of the Polish Academy of Sciences

Jan Wieczorek
Wroclaw University of Science and Technology

Patryk Hubar
Institute of Literary Research of the Polish Academy of Sciences

Jan Kocoń
Wroclaw University of Science and Technology

Marek Kubis
Adam Mickiewicz University in Poznan

Stanisław Woźniak
Wroclaw University of Science and Technology

Arkadiusz Margraf
Institute of Bioorganic Chemistry of the Polish Academy of Sciences

Wiktor Walentynowicz
Wroclaw University of Science and Technology

Abstract

In this article, we discuss the conditions surrounding the building of historical and literary corpora. We describe the assumptions and method of making the original corpus of the Polish novel (1864-1939). Then we present the research procedure aimed at demonstrating the variability of the emotional value of the concept of 'the city' and 'the country' in the texts included in our corpus. The proposed method considers the complex socio-political nature of Central and Eastern Europe, especially the fact that there was no unified Polish state during this period. The method can be easily replicated in the studies of the literature of countries with similar specificities.

1 Introduction

The main objective of our paper is to introduce a comprehensive workflow employing NLP methods and Linguistic Linked Open Data (LLOD) that allows for conducting literary research in temporal and spatial dimensions while taking into account local specificities arising from historical, socio-cultural, and infrastructural factors. The proposal addresses, on the one hand, the current criticism towards Digital Humanities (DH), in particular distant reading, and on the other hand, the call for a new way of describing the complex relationship between fiction and imaginary geography put forward in non-digital literary studies.

Criticism of the distant reading approach revolves around the gap between the development of methods and tools and the use of their potential to address new research questions or discover new phenomena. It has been argued that applications of NLP in literary research, while spectacular, are often limited to confirming already known insights (Brennan, 2017), and that much of the research is aimed solely at the tools' validation (Hammond, 2017). Computational analysis of literary texts has also been criticised for reductionism, observation triviality, ahistoricism and the disregard of the socio-cultural determinants of the patterns detected (Bode, 2017).

Addressing the call arising from the above criticism, to embed computational analyses within broader disciplinary contexts and knowledge (Underwood, 2019), the workflow we developed was tailored to current debates and trends in humanities research. We draw on studies within the horizons of geography of literature and literary affect studies, trends highly influential in contemporary humanities that emerged from the topographical and affective turns (Peraldo, 2016; Rybicka, 2014; Ahern, 2019; Nycz et al., 2015). Our goal is to take into account local circumstances and to trace the impact of historical and spatial factors on the dynamics of literary processes. At the current stage, we focus on Central and Eastern Europe (CEEC). However, the workflow was designed for reusabil-

Proceedings of the 6th LaTeCH-CLfL Workshop, pages 115–125

ity and should be adaptable to other local contexts, including non-European.

Studies referring to CEEC as a region undergoing extensive geopolitical changes, highly diverse in terms of nationality, language and culture, build more and more on linguistic resources and metadata for embedding literary texts in socio-cultural realities. Although some CEEC languages are on track to achieve a well-resourced status, there are still many gaps to bridge (Vetulani and Vetulani, 2020; Goldhahn et al., 2016), especially in terms of historical resources. This applies not only to language technology (LT), but also to DH in general. There are very few solutions suited to the processing and analysis of literary texts, e.g. Named Entity Recognition or stylometric analysis. Moreover, the metadata produced by CEEC institutions is incomplete and there is no single relevant and reliable metadata retrieving source. It leads to the necessity of combining various resources, mappings, and harmonisation of disparate data types (Király, 2019).

In the research presented in this paper, the workflow was applied to reconstruct the urban-rural dichotomy, i.e., the attribution of distinguishing values and emotions to urban and rural geo-entities, as a prominent example of reflection from the field of literary geography. This topic's long-standing presence in non-digital literary studies (e.g. (Rybicka, 2003; Williams, 1975)) resulted in the burden of cliché interpretations. Not only do we intendWe intend to verify these interpretations, but also to broaden the scope of investigation by taking into account both historical and geographical dimensions, crucial for CEEC-oriented research and neglected in literary studies. This will allow for a thorough evaluation of the workflow regarding its strengths and shortcomings.

We surveyed Polish novels from 1864 to 1939, representing three consecutive literary periods, Positivism, Young Poland, and the Interwar Period. At the end of the 18th century, Poland ceased to exist as a sovereign state. The Polish territories were divided into three partitions and remained under the control of the Habsburg Monarchy, the Kingdom of Prussia and the Russian Empire until 1918. The partitions differed substantially in terms of the pace of socio-economic development, the extent of urbanisation, and civil liberties (Kaczynska, 1970). These contrasts led to differences in the imaginary, including the dominant discourses of urbanity and rurality, which persisted even after Poland regained independence (Chwalba, 2009). To reconstruct the evolution of the Polish variant of the urban-rural dichotomy, we posed three research questions: (i) Did the level of discrepancy between urban and rural depictions change over time?; (ii) How have the emotional representations of the city and the country changed over time; and (iii) Did the form of the urban-rural dichotomy and the valuation of geo-entities vary according to the partition in which the unit was located?

Related works

The establishment of Spatial Humanities and the rapid growth of Digital Literary Cartography (Cooper et al., 2016; Gregory et al., 2015) on the one hand, and new developments in sentiment analysis for computational literary studies (Jacobs, 2019; Kim and Klinger, 2018), on the other, have not yet translated into systematic research on emotion in relation to fictional representations of space and place. While the recognition, disambiguation, and mapping of toponyms in literary works have been relatively well explored, methodological support for the literary geography of emotions is still lacking and no comprehensive framework has been developed to serve as a reference point (Morariu, 2020). Attempts in this direction have been made in two projects: 'The Emotions of London' (Heuser et al., 2016) and 'High Mountains Low Arousal? Distant Reading Topographies of Sentiment in German-Swiss Novels in the early 20th Century' (Herrmann et al., 2022). The former was a crowdsourcing experiment combining quantitative and qualitative methods of literary geography and focused specifically on the fictional representation of London. The latter (still ongoing) has a much broader scope. It aims to explore whether representations of the landscape can be regarded as a part of the construction of different national identities. The project is based on a comparative analysis of German-language novels from the early 20th century, employing methods well established in computational literary studies, i.e. sentiment analysis and Named-Entity Recognition (NER). The authors do not focus on creating new solutions and workflows; instead they mainly use and validate existing tools and resources. Although their approach may work well for regions that have not experienced significant geopolitical transitions or for well-resourced languages, it is overly simplistic to apply to CEEC.

In the absence of methodological support for the literary geography of emotions, it is necessary to develop our own solutions, combining several components, ranging from the creation of historical corpora through NER, Named-Entity Disambiguation (NED), and Named-Entity Linking (NEL) to sentiment analysis. To compile an optimal procedure, we reviewed the solutions and identified potential challenges.

The corpus linguistics literature points out that the fulfillment of the balanced source selection postulate (Biber, 1993) encounters many more obstacles in the case of historical corpora than in the case of contemporary text corpora (Gruszczyński et al., 2020). A fundamental and unfathomable problem is the limited knowledge of the writing of a particular era, which makes the decisions during corpora compilation arbitrary and often based on speculation with purely theoretical assumptions. The balance of the corpus is complicated by the disproportion between the number of extant texts from earlier eras and the number of texts from later ones (Górski, 2018). Projects aiming to build literary corpora that are as representative and balanced as possible (e.g., KOLIMO(Herrmann and Lauer, 2020) or dProse 1870-1920, (Gius et al., 2021)) have reduced speculativity by using data from bibliographic records on the production and reception of texts.

One of the most recent and most ambitious projects of this type is the European Literary Text Collection (ELTeC) composed of 11 national subcorpora, each containing 100 novels published between 1840 and 1920. It was assumed that each sub-corpus should fulfill the same compositional criteria with some level of flexibility (Schöch et al., 2021). However, the construction of the corpus raises some concerns. Foremost, the arbitrary categorisation of the texts into four twenty-year time periods, which do not correspond to the caesuras marked by significant socio-political events, both on a pan-European and regional level, is questionable. In the case of the Polish-language sub-corpus (ELTeC, 2021), the incorporation of texts published after 1920 and the lack of a description of the procedure for obtaining metadata, crucial for assessing the quality of the corpus, can also be considered problematic.

In the case of Polish, we have several NER solutions. Liner2 (Marcinczuk et al., 2017) is a tool based on the Conditional Random Fields (CRF) method, which uses morphological information and sets of manually prepared features to identify named entities. The latest method dedicated to Polish is PolDeepNer2 (Marcinczuk and Radom, 2021). It is a neuronal model based on a RoBERTa-type language model. Another tool for the NER task with support for Polish is spaCy (Honnibal and Montani, 2017). While very fast and adapted to industrial language work, it is marginally less efficient than PolDeepNer2.

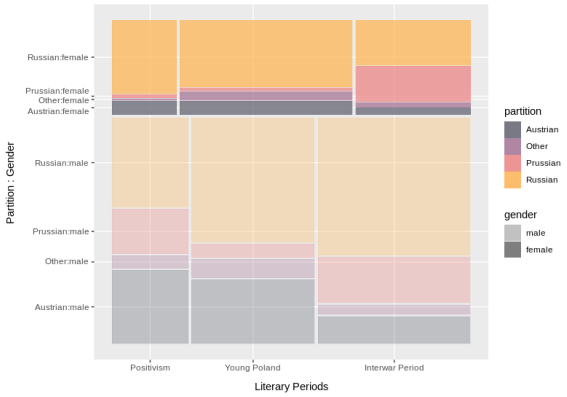

Figure 1: Mosaic diagram of the relationships between the characteristics of the novels in 19/20MetaPNC: the gender of the author, the partition in which the novel was published, and the literary period.

There are several sentiment analysis systems for Polish described in (Wawer, 2019), but none of them is available as an open service. One open system that can be used for sentiment analysis is the MultiEmo service[1] (Kocoń et al., 2021), available through the CLARIN-PL project. It is a new solution based on transformer-type models, available for more than 100 languages, thanks to the LaBSE (Language Agnostic BERT Sentence Embeddings) model (Feng et al., 2022).

Data

Currently, there is no representative and balanced historical corpus of novels in Polish that could function as a referential corpus for various research purposes. For Polish, there are literary corpora that do not meet the standard criteria for composition and have not been robustly described with metadata. An exception is the Polish-language ELTeC subcorpus, which can function as a benchmark for the development of historical literary corpora, despite the aforementioned drawbacks (see Related works). Sampled literary texts are also a

[1] https://ws.clarin-pl.eu/multiemo

component of the general corpora of Polish, such as NKJP (Przepiórkowski et al., 2012) and KPWr (Marcińczuk et al., 2016).

For the requirements of spatial-diachronic literary research, it was necessary to design a new corpus, reusable, historically and geographically balanced, precisely described with the possibly complete metadata, which would enable the selection of predefined subcorpora for comparative purposes. We named the designed collection "Metadata-enriched Polish Novel Corpus from the 19th and 20th centuries" (19/20MetaPNC). The specific nature of the geopolitical and socio-cultural context of the Polish territories in the second half of the 19th and first half of the 20th century determined the metadata structure and content as well as the criteria for balancing the corpus. Due to the impossibility of precisely defining the population of texts and the lack of data on literary production and reception of the period covered by the study, our efforts were focused on the aspect of proper balancing and precise description with regard to the metadata of the available textual resources. The basic criteria for the selection of texts, in addition to the time horizon adopted in the project, were the genre and language of the text — we decided to include in the corpus novels originally written in Polish and first published as books between 1864 and 1939. An additional criterion was the time of the plot, which could not be earlier than 1815. This was the year of the Congress of Vienna, which defined the national borders that remained in force with minor modifications for more than 100 years. We assumed that the corpus should be balanced with regard to the individual novel's belonging to one of the three literary eras distinguished in Polish literary studies — Positivism (1864-1890), Young Poland (1890-1918) and the Interwar Period (1918-1939) — determined by the date of first publication, the Partition in which the novel was first published, the gender of the author and the level of reception. Following the same approach as in ELTeC (Schöch et al., 2021), we decided that the corpus should include both: novels that can be considered part of the contemporary canon and works that have been mostly forgotten. As a measure of the level of reception, we took the number of reissues of a given publication.

The process of text selection for the corpus was conducted in three stages. In the first stage, we identified and sourced potential candidate texts. The collected texts come from several distinct sources. We started with 100 novels gathered in the ELTeC that are encoded in TEI format. Next, we included 193 texts from the Wolne Lektury library (Modern Poland Foundation, 2022), an online repository that is primarily focused on school readings and offers contemporised editions of novels that have fallen into the public domain. The data in Wolne Lektury is available in a custom XML format that preserves information about paragraph boundaries. Afterwards, the 225 novels from the Polish edition of the Wikisource project (Wikimedia Foundation, 2022) were added. These texts are transcriptions of printed books whose copyright has expired. They are encoded in the MediaWiki format and proofread by Wikisource editors, but contrary to the Wolne Lektury volumes, the original spelling is preserved, hence orthographic forms that do not appear in modern Polish can be observed. The last and most demanding source of texts for our corpus is the Polona digital library maintained by the National Library of Poland (2022). Polona offers scans of printed books along with the OCR-derived textual layer. The raw texts from Polona are neither proof-read nor contemporised, but the volume of available data is an order of magnitude greater than in the other resources. We downloaded approx. 6,000 digitised volumes from Polona. After merging multi-volume editions of novels, we obtained 4,808 complete Polona texts. From the 5,326 pieces of literary fiction that formed our initial dataset we selected exactly one edition of every novel. For the purpose of further processing the texts from Wolne Lektury, Wikisource and Polona were converted into a uniform, tab-separated format inspired by CONLL-U Plus representation[2].

In the second phase of corpus construction, we focused on completing the metadata of the collected texts. The work was carried out in an automated and manual procedure. We linked metadata of digital copies of texts to metadata from library catalogues, using the services of the National Library of Poland, and then enriched the entities with permanent identifiers (PIDs) of widely used databases: VIAF, Wikidata, and Geonames. From the data available in the authority databases, we extracted information required for corpus balancing and relevant from the perspective of spatial-diachronic literary research. Simultaneously, the

[2] https://universaldependencies.org/ext-format.html

collection of texts was manually annotated, which covered the time of the novel's action (before or after 1815) and also verified for original language and genre (not always correctly described in the National Library). On this basis, we have again made a selection of texts, rejecting texts that are not novels, written before or after the period covered by the evaluation, and those set before 1815. We also identified and removed duplicates, thus obtaining a database of 1,707 unique novels. The texts were described with the following metadata: author, author's gender, author's Wikidata ID (if available), author's place of birth with coordinates (this information was available for 1,198 items), title, year of first publication, place of first publication, first publication place coordinates and geopolitical territories (Russian, Austrian, and Prussian Partition or foreign countries), number of reprints, number of tokens.

In the third stage, we balanced the corpus. Because it was impossible to keep equal proportions between the classes, we determined the minimum and maximum share of a particular text class in the corpus. We gave priority to balancing by date and place of publication. We assumed that each of the three partitions should be represented by at least 15% of the texts, while each of the three literary eras should be represented by at least 20% of the texts. Following the approach of the ELTeC authors, we determined that at least 10% and a maximum of 50% of the titles should have a female author, at least 30% of the titles should have a low (no more than 2 reprints) and at least 30% a high (2 reprints and more) reception. The proportion of titles for each balance criterion is presented in Fig. 1. 19/20MetaPNC will be published by the end of 2022 in the CLARIN-PL Repository.

Methods

Taking into consideration that the novels gathered in our corpus come from sources that vary in quality, the preliminary steps undertaken to process the collected texts depend on their origin. The proposed workflow for contextualised spatial-diachronic literary research is presented in Fig.2. In the case of ELTeC and Wolne Lektury data we simply split texts into paragraphs and sentences and perform tokenization. OCR-derived texts from Polona are additionally pre-processed by a normalisation script that determines proper word segmentation by looking up correct word forms in the

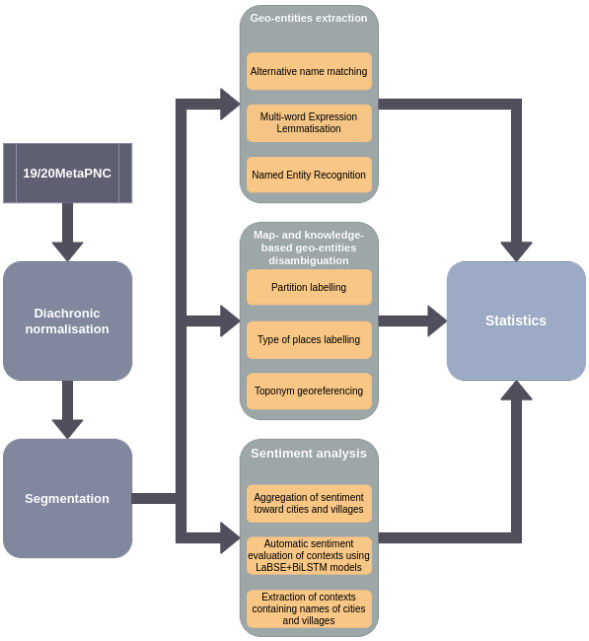

Figure 2: The workflow for retrieving statistics for literary research in temporal and spatial dimensions.

PoliMorf dictionary (Woliński et al., 2012) following the algorithm for OCR gap elimination outlined in (Kubis, 2021). The texts from Wikisource and Polona are contemporised with the use of a diachronic normalizer (Jassem et al., 2017) in order to improve the performance of NLP tools designed for modern Polish that are used in the following steps.

In order to retrieve named entities in the text, we used the PolDeepNer2 system[3] and its pretrained model, learned from the KPWr corpus (Marcinczuk, 2020). We decided to implement this model because it is the best available model for PolDeepNer2 in terms of general texts. After the process of recognising named entities, a pre-selection of the entities of interest was made based on their class. All classes related to the administrative names of locations and verb entities from the names of locations were included. The data extracted in this way was lemmatised with the Polem (Marcinczuk, 2017) tool[4]. The authors of the publication report the effectiveness of the PolDeepNer2 systems as 0.899 F1-Score measure on the PolEval 2018 set, and the Polem as 0.979 F1-Score measure (with a refinement of 0.846 F1-Score measure for NER) on KPWr corpus. Since the names of cities and villages in CEEC changed

[3] https://gitlab.clarin-pl.eu/information-extraction/poldeepner2

[4] https://github.com/CLARIN-PL/Polem

with geopolitical transformations, the final step in preparing the data for the standardisation stage was to include alternative names and varieties for the locations. We used geographic-historical registers, directories and dictionaries selected on the basis of the completeness of the sources and the territorial coverage of the Three Partitions of Poland as data sources. Data extraction involved an OCR process. To identify place names in the documents' page area, we developed an original solution using hierarchical agglomerative clustering (HCA) algorithm to identify the structure of historical documents (tables, indices, and appendices) regardless of their digital copy quality. We specified that the algorithm would analyse clusters in one dimension and applied the Euclidean distance measure. This approach allowed for visualising the performance of the algorithm and facilitated controlling the selection of other parameters.

To identify and standardise geographic entities (geo-entities), we applied an experimental three-stage toponym disambiguation workflow, based on leading approaches in Geographical Information Retrieval (GIR) (Buscaldi, 2011; Derungs and Purves, 2014). We used mainly the Geonames database supplemented with additional data sources (i.e., Wikidata and Wikipedia). The goal of the first stage was to unambiguously assign records from the Geonames database for the geo-entities identified in the text. We used a list of historical name variants prepared in the first stage to query the Geonames database and pre-filter the search results. We included only those geo-entities for which we found the corresponding names in the 'name' or 'alternateName' fields of a Geonames record. If we received only one record after initial filtering of the results, we retrieved its complete information and assigned it to geo-entity. If we obtained several Geonames records, we selected the geo-entity whose coordinates were closest to the area mapped using the coordinates of other locations identified in a given text.

In the second stage, we determined whether the name refers to a city or a village. For this purpose, we used the records from stage one, since they contained the dates of granting municipal rights, as well as contextual information extracted automatically from the text (the terms 'city' and 'village' and their synonyms occurring in the immediate vicinity of the named entity).

The third stage of the disambiguation process was to determine the partition in which a village or a city was located, using a map-based approach. For this purpose, we used historical maps of Polish territories under the post-1815 partitions. Given the raster form of the maps, we used the open source software QGIS that allows georeferencing of the maps in a form that allows automatic processing and the OpenStreetMap resources. Then, we plotted three polygons on the map corresponding to each partition, which allowed us to determine the precise coordinates of their borders. Once the coordinates of the partitions and geoentities were determined, it was possible to assign the geoentity's affiliation to a particular partition.

Next, we automatically determined the sentiment of contexts containing proper names representing cities and villages. For this purpose, we used the MultiEmo tool (Kocoń et al., 2021) trained on sentences annotated with sentiment within the PolEmo 2.0 corpus (Kocoń et al., 2019). This corpus contains more than 8k consumer reviews and more than 50k sentences. Both texts and sentences were annotated using four sentiment labels: positive, negative, neutral and ambivalent. The model achieves very good quality, i.e. an F1-score of about 85%. Next, we evaluated all sentences representing proper names pertaining to cities and the countries. We decided to use a sentence-level sentiment model because the model is more domain-independent and there were more training examples than for a model dedicated to analysing entire reviews.

Results

We identified 130,635 mentions of places in the corpus, for 86,568 (66.3%) mentions we found matches in the Geonames service, which provided information about the partition in which the place was located, and we included these mentions in further analysis. 51.2% of them referred to cities, 48.8% to villages.

This proportion varied over the years and between partitions. The share of mentions reffering to city (Fig. 3) was the lowest and most stable over the years in the Russian partition. In the case of the Austrian and Prussian partitions, there are substantial fluctuations over time, although it is difficult to determine trends. In the former case, a large increase in the share of city mentions was observed at the beginnig of the 20th century and immediately after World War I, while in the latter case, a large

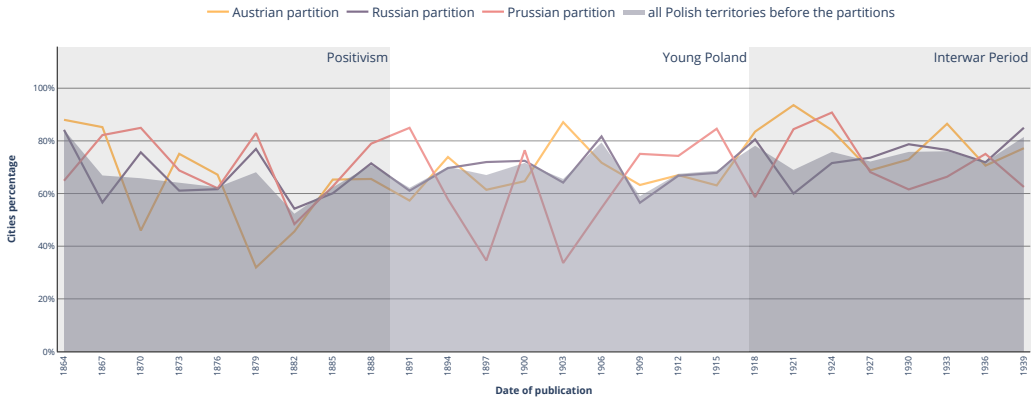

Figure 3: Percentage of city mentions in successive years.

	Estimate	Std. Error	z value	Pr(>|z|)	0.95 Conf. Interval	
					Lower	Upper
(Intercept)	11.4263	0.6777	16.86	0.0000	10.10	12.75
status: village	0.0593	0.0180	3.29	0.0010	0.02	0.09
partition: Austrian	-0.0403	0.0277	-1.46	0.1455	-0.09	0.01
partition: Prussian	0.0241	0.0338	0.71	0.4749	-0.04	0.09
partition: Russian	0.1024	0.0206	4.96	0.0000	0.06	0.14
year	-0.0065	0.0004	-18.19	0.0000	-0.01	-0.01
gender: male	0.0153	0.0174	0.88	0.3774	-0.02	0.05

Table 1: Regression coefficients of the negative sentiment prediction model.

increase was observed in the early 1880s, during World War I and in the mid-1920s. Overall, city mentions dominated over village mentions.

Most mentions (68%) had neutral sentiment, 31% were negative, 1.2% positive and 0.01% ambivalent. The results of the sentiment analysis confirm the urban-rural dichotomy. The difference in the proportion for each sentiment dimension (positive and negative) between cities and villages, decreases with time (Fig. 4). In the case of positive sentiment, the decrease is visible from the last decade of the 19th century (the beginning of Young Poland), in the case of negative sentiment the decrease is less evident, but it appears that from the beginning of the 20th century the dichotomy between cities and villages begins to decrease, although the disproportion is still greater than for positive sentiment.

Mentions with positive sentiment refer to cities and villages equally, with the majority of positive mentions beginning to refer to cities during the Interwar Period. In the case of negative sentiment, the dominance of the villages is clear and this does not change over the literary periods.

While in the case of negative mentions in which a given place was located, the urban-rural dichotomy decreased in subsequent years (with some fluctuations) in all partitions, in the case of positive mentions the dichotomy clearly decreased in the Russian partition, however in the Austrian and Prussian partitions a decrease occurred in Young Poland, and increased in the Interwar Period.

Among the negative mentions referring to places located in the Russian partition, mentions of villages definitely dominate. Among the positive ones, on the other hand, this pattern persists during Positivism but reverses in subsequent literary periods. The mentions relating to places in the Prussian and Austrian partitions demonstrate similar shifts: the positive mentions are dominated by cities, but there are years when the dominance of villages is very apparent. In the case of negative mentions relating to the Prussian partition, in Positivism mentions of villages prevail, and from the end of Positivism mentions of cities are more prevalent (not without exceptions). In the case of negative mentions relating to the Austrian partition mentions of villages tend to dominate until the end of Young Poland, while in the Interwar Period mentions of cities are predominant.

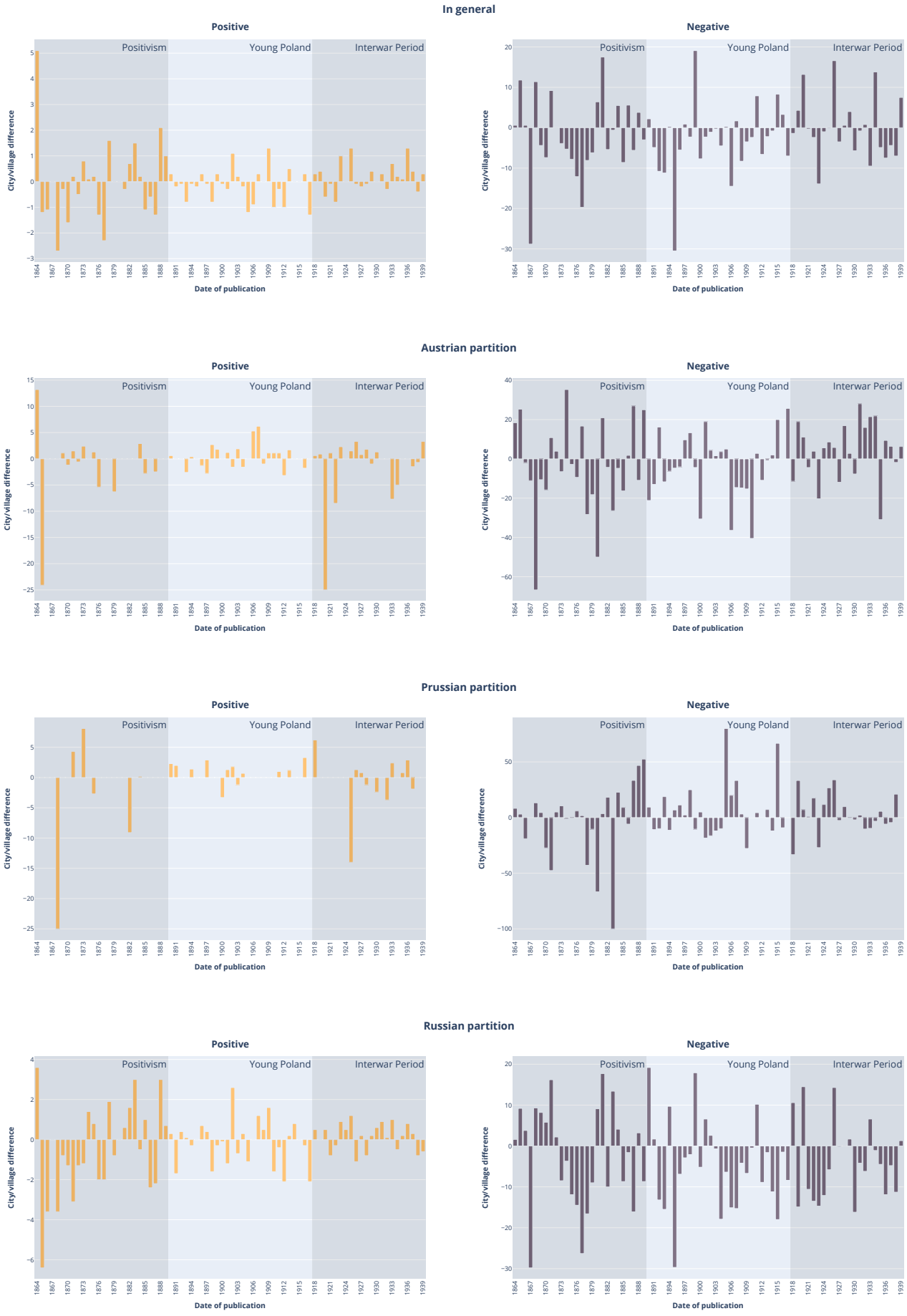

Figure 4: The difference in the proportion for each sentiment dimension (positive and negative) between cities and villages, both in total and by partition, in successive years. A score above zero means that a dimension dominates for the cities, and below zero for the villages.

Taking into account disproportion in the number of sentences belonging to particular sentiment categories we decided to further investigate the impact of location on the sentiment by constructing a model that discriminates between the negative sentiment and all the other categories considered jointly. For this purpose we fitted a binomial logistic regression model with the degree of negativity being the independent variable and location status (village or city), partition (Austrian, Prussian, Russian or abroad), publication year and author's gender being dependent variables. Table 1 presents the regression coefficients of the fitted model. The status of a village and belonging to the Russian partition contributes to the negative sentiment towards the location significantly. Furthermore, the sentiment tends to be less negative for more recent publications.

Conclusions and future work

The results of our study confirmed the urban/rural dichotomy manifested in the different valorisation of urban and rural geo-entities in literary texts. We also confirmed Rybicka's (Rybicka, 2003, 2014), statement that this dichotomy decreased over time, although the changes we observed occurred slightly later than the author assumed. We found that the negative sentiment of place mentions also decreased over time. However, we did not confirm the thesis of the dominance of the anti-urban myth (Rybicka, 2003) in depictions of the city and the country. On the contrary, we showed that villages were more negatively portrayed than cities. Nevertheless, this conclusion needs further research and stronger confirmation. What can be recognised as an important achievement is the group of conclusions concerning the differentiation between the partitions, which clearly indicates the need to include the spatial dimension apart from the temporal dimension.

The workflow component currently raising the most doubts is the sentiment analysis. An analysis with a tool trained on literary data, optimally historical data, would provide more precise results. It would be worthwhile to study the contexts of the occurrence of proper names representing cities and villages using neuro-symbolic models for sentiment (Kocoń et al., 2022), which may yield better results for texts from domains other than those used to train the sentiment model. However, as this postulate requires extensive efforts, in the nearer term it is more effective to refine other aspects of the proposed scheme. Firstly, following (Herrmann et al., 2022) it is worth preparing a dictionary of the names of the objects related to city and village (e.g. 'sawmill', 'manor', 'tenement', 'town hall') and use it to specify the geo-entity status. Secondly, due to the focus on determining the coordinates of recognised geo-entities, we excluded imaginary places from the analysis. They will be covered in subsequent project stages. Thirdly, in the paper we did not consider the phenomena of anaphora and coreference. Efforts are underway to adapt the tools for anaphora and coreference resolution to the specificities of literary texts. Fourthly, in order to address the bias associated with the potential overrepresentation of particularly long novels, we will balance the corpus by length of texts. Fifthly, we will take into account authors' biobibliographical data (e.g. place of birth, work and education) and address the question of mobility in the analysis.

Acknowledgements

The work was co-financed by the European Regional Development Fund as a part of the 2014–2020 Smart Growth Operational Programme, Digital research infrastructure for the humanities and arts sciences DARIAH-PL, agreement no. POIR.04.02.00-D0006/18-00 dated 28/12/2020.

The work was co-financed by the European Regional Development Fund as a part of the 2014–2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19.

The work was co-financed as part of the investment: "CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure (period: 2022-2023) funded by the Polish Ministry of Education and Science (Programme: "Support for the participation of Polish scientific teams in international research infrastructure projects"), Agreement number 2022/WK/09.

References

Stephen Ahern. 2019. *Affect theory and literary critical practice: a feel for the text*. Palgrave Macmillan, Cham.

Douglas Biber. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4):243–257.

Katherine Bode. 2017. The equivalence of "close" and "distant" reading; or, toward a new object for data-rich literary history. *Modern Language Quarterly*, 78:77–106.

Timothy Brennan. 2017. The Digital-Humanities Bust. *The Chronicle of Higher Education*, 64(8).

Davide Buscaldi. 2011. Approaches to disambiguating toponyms. *SIGSPATIAL Special*, 3(2):16–19.

Andrzej Chwalba. 2009. Dziedzictwo zaborów. In *Polski wiek XX*, volume 1, pages 7–24. Bellona i Muzeum Historii Polski, Warszawa.

David Cooper, Christopher Donaldson, and Patricia Murrieta-Flores. 2016. *Literary Mapping in the Digital Age*. Routledge, Warszawa.

Curdin Derungs and Ross S. Purves. 2014. From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus. *International Journal of Geographical Information Science*, 28(6):1272–1293.

ELTeC. 2021. Polish novel collection (ELTeC-pol). Ed. by Joanna Byszuk. COST Action Distant Reading for European Literary History.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Evelyn Gius, Svenja Guhr, and Inna Uglanova. 2021. "d-Prose 1870–1920" a Collection of German Prose Texts from 1870 to 1920. *Journal of Open Humanities Data*, 7(0):11.

Dirk Goldhahn, Maciej Janicki, and Uwe Quasthoff. 2016. Corpus collection for under-resourced languages with more than one million speakers. Workshop on Collaboration and Computing for Under-Resourced Languages (CCURL), LREC.

Ian N. Gregory, Christopher Donaldson, Patricia Murrieta-Flores, and Paul Rayson. 2015. Geoparsing, GIS, and Textual Analysis: Current developments in spatial humanities research. *Int. J. Humanit. Arts Comput.*, 9:1–14.

Włodzimierz Gruszczyński, Dorota Adamiec, Renata Bronikowska, and Aleksandra Wieczorek. 2020. Elektroniczny korpus tekstów polskich z XVII i XVIII w.– problemy teoretyczne i warsztatowe. *Poradnik Językowy*, (0):32–51.

Rafał L. Górski. 2018. Metody korpusowe i kwantytatywne w językoznawstwie historycznym. In *Metodologie językoznawstwa. Od diachronii do panchronii*, pages 65–81. Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Adam Hammond. 2017. The double bind of validation: distant reading and the digital humanities' "trough of disillusionment". *Literature Compass*, 14(8):e12402.

Berenike Herrmann, Giulia Grisot, and Simone Rebora. 2022. High mountains low arousal? Distant reading topographies of sentiment in German-Swiss novels in the early 20th century. https://mountain-sentiment.github.io/. Accessed: 2022-07-10.

Berenike Herrmann and Gerhard Lauer. 2020. Kolimo. a corpus of literary modernism for comparative analysis. https://kolimo.uni-goettingen.de/about. Accessed: 2022-07-10.

Ryan Heuser, Mark Andrew Algee-Hewitt, and Anna Lockhart. 2016. Mapping the Emotions of London in Fiction, 1700–1900: A Crowdsourcing Experiment. In *Literary Mapping in the Digital Age*, pages 43–64. Routledge.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Arthur M. Jacobs. 2019. Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics. *Frontiers in Robotics and AI*, 6.

Krzysztof Jassem, Filip Graliński, and Tomasz Obrębski. 2017. Pros and Cons of Normalizing Text with Thrax. In *Proceedings of 8th Language & Technology Conference*, pages 230–235.

Elżbieta Kaczynska. 1970. *Dzieje robotników przemysłowych w Polsce pod zaborami*. Warszawa.

Evgeny Kim and Roman Klinger. 2018. A survey on sentiment and emotion analysis for computational literary studies. *ArXiv*, abs/1808.03137.

Péter Király. 2019. *Measuring Metadata Quality*. [Doctoral's thesis, Faculty of Humanities of the Georg-August-Universität Göttingen].

Jan Kocoń, Joanna Baran, Marcin Gruza, Arkadiusz Janz, Michał Kajstura, Przemysław Kazienko, Wojciech Korczyński, Piotr Miłkowski, Maciej Piasecki, and Joanna Szołomicka. 2022. Neuro-symbolic models for sentiment analysis. In *International Conference on Computational Science*, pages 667–681. Springer.

Jan Kocoń, Piotr Miłkowski, and Kamil Kanclerz. 2021. Multiemo: Multilingual, multilevel, multidomain sentiment analysis corpus of consumer reviews. In *International Conference on Computational Science*, pages 297–312. Springer.

Jan Kocoń, Piotr Miłkowski, and Monika Zaśko-Zielińska. 2019. Multi-level sentiment analysis of

Polemo 2.0: Extended corpus of multi-domain consumer reviews. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 980–991.

Marek Kubis. 2021. Quantitative analysis of character networks in Polish 19th- and 20th-century novels. *Digital Scholarship in the Humanities*, 36(Supplement_2):ii175–ii181.

Michal Marcinczuk. 2017. Lemmatization of multiword common noun phrases and named entities in polish. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 483–491. INCOMA Ltd.

Michal Marcinczuk. 2020. KPWr n82 NER model (on polish RoBERTa base).

Michal Marcinczuk, Jan Kocoń, and Marcin Oleksy. 2017. Liner2 — a Generic Framework for Named Entity Recognition. In *BSNLP@EACL*.

Michał Marcińczuk, Marcin Oleksy, Marek Maziarz, Jan Wieczorek, Dominika Fikus, Agnieszka Turek, Michał Wolski, Tomasz Bernaś, Jan Kocoń, and Paweł Kędzia. 2016. Polish Corpus of Wrocław University of Technology 1.2. CLARIN-PL digital repository.

Michal Marcinczuk and Jarema Radom. 2021. A single-run Recognition of Nested Named Entities with Transformers. In *KES*.

Modern Poland Foundation. 2022. About the Project. https://wolnelektury.pl/info/o-projekcie/. Accessed: 2022-07-10.

David Morariu. 2020. The affective geography of paris in the 19th century romanian novel: Between admiration and aversion. *Metacritic Journal for Comparative Studies and Theory*, 6:129–147.

National Library of Poland. 2022. About Polona Website. https://polona.pl/page/about-polona/. Accessed: 2022-07-10.

Ryszard Nycz, Anna Łebkowska, and Agnieszka Dauksza, editors. 2015. *Kultura afektu - afekty w kulturze : humanistyka po zwrocie afektywnym*. Nowa Humanistyka, t. 19. Wydawnictwo Instytutu Badań Literackich PAN, Warszawa.

Emmanuelle Peraldo. 2016. *Literature and geography: the writing of space throughout history*. Cambridge Scholars Publishing, Newcastle upon Tyne.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy korpus języka polskiego: praca zbiorowa*. Wydawnictwo Naukowe PWN, Warszawa.

Elżbieta Rybicka. 2003. *Modernizowanie miasta: zarys problematyki urbanistycznej w nowoczesnej literaturze polskiej*. Universitas, Kraków.

Elżbieta Rybicka. 2014. *Geopoetyka: przestrzeń i miejsce we współczesnych teoriach i praktykach literackich*. Towarzystwo Autorów i Wydawców Prac Naukowych "Universitas", Kraków.

Christof Schöch, Roxana Patras, Tomaž Erjavec, and Diana Santos. 2021. Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*, (1):25. Number: 1 Publisher: Liverpool University Press.

Ted Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, Chicago, IL.

Zygmunt Vetulani and Grazyna Vetulani. 2020. The case of Polish on its Way to Become a Well-Resourced-Language. In *Proceedings of LT4All*, Paris. European Language Resources Association.

Aleksander Wawer. 2019. Sentiment analysis for polish. *Poznan Studies in Contemporary Linguistics*, 55(2):445–468.

Wikimedia Foundation. 2022. About Wikisource. https://wikisource.org/wiki/Wikisource:About_Wikisource. Accessed: 2022-07-10.

Raymond Williams. 1975. *The country and the city*. Oxford University Press, New York.

Marcin Woliński, Marcin Miłkowski, Maciej Ogrodniczuk, and Adam Przepiórkowski. 2012. PoliMorf: a (not so) new open morphological dictionary for Polish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 860–864, Istanbul, Turkey. European Language Resources Association (ELRA).