Analysis of Toponyms from the Polish National Bibliography

Adam Pawłowski[1], Tomasz Walkowiak[2]

[1]Institute of Information and Library Science, University of Wrocław, pl. Uniwersytecki 1, 50-137 Wrocław, Poland
[2]Faculty of Information and Communication Technology, Wrocław University of Science and Technology, 27 Wybrzeże Wyspiańskiego St., 50-370 Wrocław, Poland

Abstract

The paper describes the process of extraction and analysis of toponyms included in the records of Polish National Library catalogs from 1801 to 2019. Using NLP and data science techniques, toponyms in bibliographic records were automatically identified and disambiguated. Web applications were prepared to visualize the projection of toponyms from databases in MARC format onto maps of Europe and the world. In the course of the research, the main geographical and cultural areas present in Polish publishing between 1801 and 2019 were identified and quantitatively analyzed. According to the culturomics paradigm and the *longue durée* theory, the hypothesis that large bibliographies are a faithful mirror of the historical development of the culture, economy, and political situation of a country in the long term was validated.

Keywords

bibliography, text mining, geolocalization, name entity recognition, toponym

1. Introduction

In most countries national libraries create and maintain large-scale bibliographies that are meant to represent the entire body of writing in a given language, and additionally works written by authors from this country in other languages over a long period. Technically, national bibliographies are "just" databases that contain text (e.g., titles), geolinguistics data (e.g., toponyms in titles), anthroponyms (author's name, and/or gender), and dates (e.g., year of publication). For decades, their function was reduced to practical tasks such as recording and/or retrieving publications. But from a cognitive perspective, they are coherent and extremely rich resources of knowledge, reflecting a nation's culture, the interests of its citizens, intellectual fashions, long-standing cultural trends, etc. – in other words, all that falls within the field of culturomics. This makes large national bibliographies a unique subject for transdisciplinary research in history, language sciences, cultural anthropology, and data science.

The paper is structured as follows: in the second section we present the research hypotheses, in the third one we discuss the data (formats, volume), in the fourth one the

HistoInformatics 2021 – 6th International Workshop on Computational History, September 30, 2021, online

adam.pawlowski@uwr.edu.pl (A. Pawłowski); tomasz.walkowiak@pwr.edu.pl (T. Walkowiak)
0000-0002-0804-5698 (A. Pawłowski); 0000-0002-7749-4251 (T. Walkowiak)

CEUR Workshop Proceedings (CEUR-WS.org)

methods used, and in the fifth one we present the results. Issues of the earlier research in biblio- and data science are addressed in section two.

2. Goals, Hypotheses, Previous Research

According to the customary practice, research in culturomics and data science is primarily exploratory and descriptive in nature. The researcher, armed with the tools of big data processing, seeks patterns and schemes that are indirectly expected to explain processes occurring in reality. Moreover, it is possible to advance and verify hypotheses that are based on premises external to the proposed model and data set.

The aim of our research was to present the possibility of modeling the spatial structure of toponyms contained in large-scale databases, using NLP, geolinguistic tools, and numerical methods. As a result, we have developed dynamic models in the form of online applications, displaying the effects of mapping automatically recognized toponyms. We have also verified the hypothesis that the toponyms in the titles of publications express long-standing, historical trends in culture and geopolitics, provided a large volume of data is available. In the case of Poland, this is manifested in the shifting of locations included in the titles and descriptions of publications between eastern, western, and partly southern areas – generally in conjunction with changes in the country's national borders and political alignment in the 20th century. This approach is transdisciplinary in that it combines NLP, data science, cultural anthropology, and quantitative history (in particular the theory of long durée).

Much and more research is now concerned with exploratory analysis of library catalogs around the world, for example [1, 2]. The analysis of data from the Polish National Library is rarer and was undertaken in [3, 4]. The problem of automatic recognition of named entities (including toponyms) is solved by Name Entity Recognizers (NER). In the case of Polish, the state-of-the-art solution for the last ten years was LiNER2 [5], a tool based on conditional random fields. However, in recent years we observe the rise of deep learning applications in the NLP field. And in the case of NERs for Polish such solutions are PolDeepNer [6] based on bidirectional LSTM [7] and PolDeepNer2 based on XMLRoberta architecture [8]. On the one hand, the problem of toponyms' geo-localization, due to the existence of such databases as GeoNames[1], could be considered scientifically solved. However, due to a lack of historical data and the ubiquity of toponyms, it is still a virtually unexplored area of scientific research – especially in less spoken languages [9].

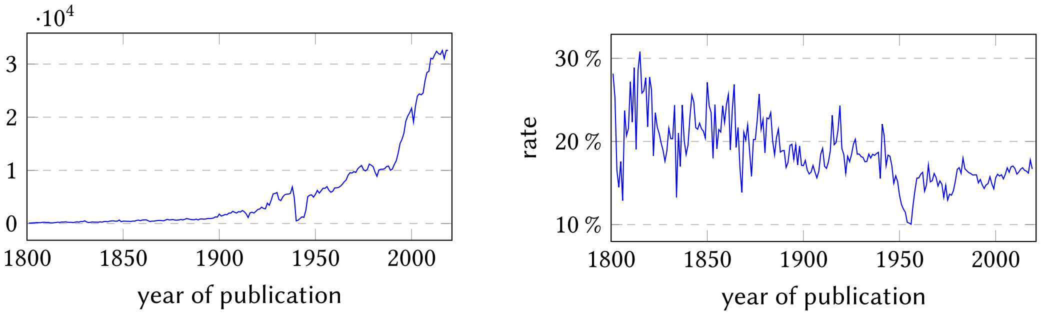
3. Data

3.1. Bibliographic Records

The research was conducted on records extracted from the catalogs of the Polish National Library[2], stored in the MARC format [10]. These contain bibliographic descriptions of

[1] https://www.geonames.org

[2] http://data.bn.org.pl/db/bibs-ksiazka.marc



(a) Number of records

(b) Rate of records with toponyms in book titles

Figure 1: Number of records and rate of records with toponyms meeting the analysis criteria (non-empty fields of title, place of release, and publication date between 1801 and 2019, Polish language).

books published in the 19th, 20th, and 21st centuries (some of them may have been written earlier). Only those records that met the following conditions were used for analysis: they referred to works in Polish (field 008, positions 35–37 of the MARC format), they contained a non-empty title and place of publication fields (subfields $a and $b of field 245, 260$a), the date ranging from 1801 to 2020 (field 008 positions 7–11). Conference materials (field 655) and other publications that are not books were discarded. The total number of records retrieved was 1.9 million. The number of records meeting the above conditions was 1.26 million.

In the relevant records, toponyms included in titles and annotated as publication places (field 260) were automatically recognized and disambiguated. This process required considerable effort, as it was necessary to identify those character strings that were toponyms (e.g., 'York' may be a person or a geographical name) and to standardize the transcriptions of thousands of toponyms appearing over a period of 200 years. The extraction of toponyms was carried out using Name Entity Recognizer as described in section 4.1. The contents of fields 203 and 651, which indicate the region of origin of the publication, were omitted from the analysis because the way it was filled in changed over the years and the data were not fully reliable.

Figure 1a shows the number of publications meeting the criteria of the analysis by years. The result obtained is a very good validation of the hypothesis formulated earlier, stating that book publications are a reliable measure of the dynamics of the development of a nation and its culture. The line of the graph shows relatively stable periods (stagnant development during Russian-German-Austrian colonization of Poland from 1801 to the outbreak of World War I), the years of the sovereign state (1918-1939) with a "swing" of the curve around 1928, when the National Library was reactivated (this change intensified the creation of records), the catastrophe of World War II, slow increase during the years of communism until 1980, the period of martial law in Poland (stagnation), and finally the explosion of the editorial movement after the fall of communism and the lifting of censorship in 1989, coinciding with the period of rapid advances in information technology. These historical periods also become apparent when other parameters are analyzed.

4. Methods

4.1. Toponym Extraction from Titles

Automatic extraction of toponyms from titles requires a named entity recognition tool for Polish. We have used PolDeepNer2[3]. This is a NER based on XMLRoberta [8] architecture. We used as a starting model the HerBERT [11] pre-trained transformer-based language model for Polish. The XMLRobetra was trained to detect NE boundary detection and fine-grained categorization (82 types) [5]. The detected categories include more than 20 types connected with localization (i.e., country, city, lake, mountain), and in our experiments, they were assumed as candidates for toponyms. Since Polish is highly inflectional and the same toponym extracted by PolDeepNer2 could appear in various (inflected) word forms, it was necessary to lemmatize detected name entities. We have used the Polem lemmatizer [12]. Since Polem requires morphological information to lemmatize NE we used KRNNT [13], a morphological tagger for Polish based on recurrent neural networks. Finally, around 205,000 titles include toponyms, with a total number of ca. 260,000 toponyms and ca. 34,000 unique toponyms. Figure 1b shows the rate of books having toponyms in titles to the total number of books as presented in Figure 1a.

4.2. Geo-localization of Toponyms and Grouping

Visualization of toponyms on the map requires the mapping between a toponym and its geographical location (longitude and latitude). In the case of publication places, i.e., cities, geo-localization was performed using GeoNames city database. In the case of cities not found in GeoNames and having a frequency of appearance higher than 50, we made a mapping using Google Maps. In the case of the second set of toponyms analyzed in this paper, i.e., toponyms extracted from titles, GeoNames was also used but with *AllCountries* database. However, it required manual errata, especially for lost locations (such as the USSR or the GDR) and small towns.

One of the proposed methods of toponyms' analysis is to group them into regions. We have distinguished ten geographical regions: seven continents, and three regions specific for Polish history: Poland itself, the USSR, and Near East. Automatic assignment of toponyms to the regions was carried out in three steps. Firstly, geo-localization of toponyms was performed. Secondly, toponyms were linked with specific countries using the countries' polygons[4] and an algorithm that allows checking, if a point is inside it. In step three, mapping countries into regions was performed. Moreover, ca. 300 toponyms that were not localized in step one and had a frequency of appearance above 50, were mapped onto regions manually by the authors.

4.3. Toponyms in Directions

Finally, we prepared software to visualise geographical directions where the identified toponyms have been located from 1801 to 2019, assuming Warsaw as the reference

[3] https://github.com/CLARIN-PL/PolDeepNer2

[4] https://datahub.io/core/geo-countries

point. It was carried out by calculating the number and the average distance from Warsaw of toponymes in consecutive years. To perform such an analysis, it was first necessary to determine and ignore toponyms located on the Polish territory. In the case of the remaining toponyms, their azimuth was calculated, taking Warsaw as a reference point. This was performed by inverse geodetic computation (implemented in pyproj[5] library) that allows determining the forward and back azimuths, as well as distances, given the latitudes and longitudes of two points using the world geodetic system WGS84. Since the locations of toponyms are discrete and sparse, we needed to smooth the directions. Therefore, we have defined the direction as a beam of +- 15 degrees (in 1-degree steps) covering +- 3 years. The shape of the beam was smoothed by Hamming window (commonly used in signal processing for signal smoothing). The weighted sum of toponyms in each beam was multiplied by an arbitrary constant and using forward geodetic computation mapped on longitude and latitude, representing the importance of each direction in time. Similarly, by calculating the weighted average distance, we could calculate and visualise the average distance of toponyms in each direction.

5. Results

The results of the research include, in the first place, the development of algorithms for processing large bibliographies in MARC format and verification of specific hypotheses, concerning some aspects of the recent history of Poland and Central Europe (section 5.1 – distribution of publication locations over time and sections 5.3, 5.3 – distribution of toponyms recognized in titles). Web applications[6] that show the dynamics of these phenomena and that allow users independent explorations should be considered an integral part of this project.

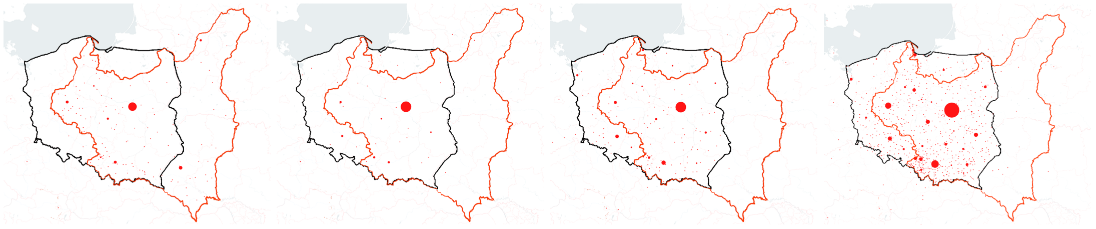
5.1. Distribution of Publication Places Over Time

The main result of the project is the application visualizing the projection of large bibliographic data in MARC format onto a map. A scalable map[7] displays in the way easy to read points which are locations of book editions split by years, moreover the cursor shows in a given location its name and the number of books published. The application has been designed in such a way that it could serve as a tool supporting any user's research. In this paper we interpret as examples screenshots of four selected time points (Figure 2). The chosen years indicate significant or breakthrough moments in Polish history. 1938 is the last year of independence of the state reborn in 1918 (in 1939 the Germans and Soviets invade and occupy Poland). 1956 is the moment of political breakthrough after the Stalinist era, called the "thaw". The year 1989 is formally the end of communism in Poland and the return to democracy. The year 2010 is representative of the modern time, when Poland is a member of NATO and the EU. This choice is

[5] https://github.com/pyproj4/pyproj

[6] https://ws.clarin-pl.eu/public/geo2

[7] https://ws.clarin-pl.eu/public/geo2/time_PUB.html



(a) 1938 (b) 1956 (c) 1989 (d) 2010

Figure 2: Publication places of books in Polish in 1938, 1956, 1989, and 2010 (red circles). The size of a circle is proportional to the number of books published at a given location. Results for any date range can be generated at: https://ws.clarin-pl.eu/public/geo2/time_PUB.html.

intended to illustrate and verify the hypotheses and goals outlined in section 2. The contours of Poland's borders from 1918-1939 and after 1945 were plotted on the map afterwards.

The distribution of points in 1938 lies to a large extent within the borders of the then Republic of Poland. The eastern provinces, now within the borders of Lithuania, Ukraine, and Belarus (after 1945 – USSR), are visible. The points indicating the greatest publication activity are, of course, Warsaw (the capital), but also Lviv, Vilnius, and Cracow (regional capitals). The 1956 map reveals two significant features. The first one is the shift of the country's borders: the places of Polish cultural publishing activity in the east practically disappear, while such places appear in the western part (e.g., Wrocaw, Szczecin). The reason for this change is that in 1945 Poland lost to the USSR about one third of its territory in the east in exchange for much smaller areas in the west (mainly Silesia and Pomerania). The second feature of the whole system of Polish culture during the period of Stalinist totalitarianism of the 1950s was sweeping centralization. The analysis of the data on the map shows that during this period almost only the biggest centers were active, while the regions were to a large extent excluded from editorial activity.

The map from 1989 shows a much denser publishing network fitting well within the Polish state border. Smaller cities are also editorial centers, but the four largest dominate, i.e., Warsaw, Cracow, Pozna and Wrocaw, which proves the persistence of the centralistic tendency. Finally, the map from 2010 shows a completely new cultural landscape. The boundaries set by the places of publication have not changed, but the technological revolution (e.g., desktop publishing) and the lifting of censorship after the fall of communism have unleashed the creative forces of many previously invisible groups, triggering an avalanche of publications in regional centers, where previously such activity could not flourish. The editorial image of Poland became "dense" and the contours of the country's borders even more distinct. This result thus confirms the hypothesis, which in the spirit of culturomics states that "toponyms in the titles of publications and their descriptions express long-standing, historical trends in culture and geopolitics".

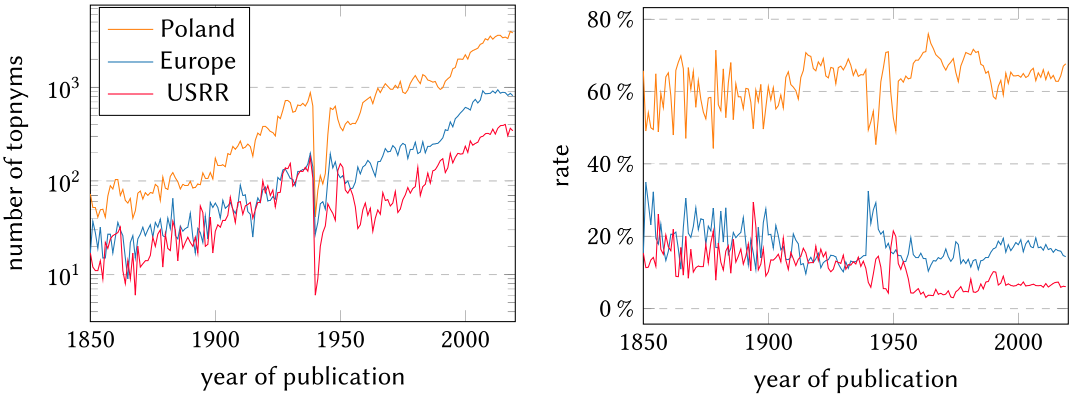


Figure 3: Increase in the number of toponyms from selected geographical areas recognized automatically in the titles of Polish books (nominal and relative value). Poland's and the USSR's territory in the 1945-1989 borders, Europe without Poland and the USSR.

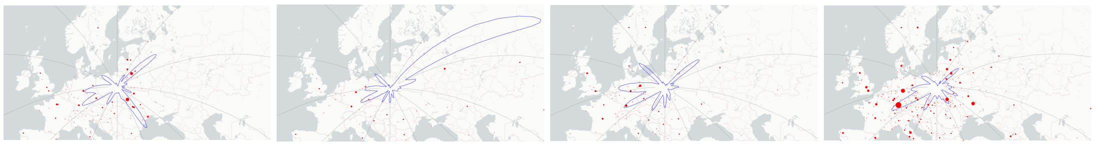
5.2. Areal Distribution of Toponyms in Book Titles

Figure 3 shows histograms of the growth in the number of toponyms located in the territory of Poland, Europe, and the USSR, included in book titles between 1850 and 2019. The attribution of toponyms was carried out automatically (as described in section 4) based on available sources and this was the only rational solution when analyzing large-scale data. However, as authors, we are aware that the attribution of certain place names to large geographical units over a long period of time is conventional and subject to the risk of error.

The graph of absolute values is closely correlated with the overall increase in the number of publications during the period under study (cf. Figure 1), and shows fluctuations at critical moments (e.g., 1939-1945). Between 1945 and 1990, a slight decrease in the share of "external" toponyms in titles can be observed. This was most likely due to Poland's isolation from foreign countries during the communist period and its quasi-colonial dependence on the USSR. The graphs of the mutual proportions of toponyms are rather surprising. The basic cultural areas from the perspective of Warsaw, that is, Poland, Europe, and the great eastern territories (the USSR), have a similar percentage share in the entire set of toponyms since 1850. These proportions seem to be a very stable characteristic of the Polish worldview, independent of historical conditions. Significant disruptions are apparent only during World War II (a drastic decrease in references to toponyms from the Polish territory during the Nazi occupation), and then after its end – during the installment of communism.

5.3. Mapping Toponyms in Book Titles and Visualizing Directions

The analysis of the distribution of toponyms in the titles raised many problems due to the existence of homographs denoting both common and proper names (here referred



(a) 1938 (b) 1951 (c) 1970 (d) 2015

Figure 4: Historical directions of interest of Polish culture based on the occurrences of toponyms in book titles. Results for any date range can be generated at: https://ws.clarin-pl.eu/public/geo2/distance_NER.html.

to as named entities). The developed web application[8] allows to dynamically project toponyms on a map and can be used independently. To analyze the interest in geographic and cultural directions that prevailed in different periods, an algorithm has been prepared that ignores internal references, affiliated as Poland within its contemporary borders. The external directions are displayed with a line moving over time beyond Poland's borders, Warsaw being the center and point of reference.

The years selected for interpretation illustrate changes in Poland's cultural orientation and are consistent with historical knowledge on the subject (see Figure 4). In 1938 (the last year of independence before the outbreak of WW2), two large "tongues" are visible, oriented to the east and southeast. This is due to the existence of strong cultural centers in regions that lay within Polish territory until 1939. The year 1951 (the culmination of stalin's repression in Poland) shows a very high level of sovietization of Polish culture, imposed during the Stalinist period (a very large "tongue" directed to the east). The next map (1970 - one of the political breakthroughs in communist Poland) contains two interesting "tongues" pointing west. The upper one is a trace of toponyms located in the GDR (toponyms from West Germany were much rarer), while the lower one points to Europe. It can be argued that while the earlier years were dominated by an eastern orientation in Polish culture, in 1971 the balance between east and west is already marked (a distinct eastern "tongue" is still visible). Finally, the map from 2015 (modern years) shows a strong reorientation of Polish culture's interests towards the west and the dominance of Europe in relation to Germany[9] [10].

The small share of northern and southern references is also characteristic. This is a result, and at the same time a confirmation, of the decades-long mutual ignoring of the cultures of Central Europe (primarily Poland) and the countries of the Scandinavian Peninsula, separated by the Baltic Sea. As for the south (southwest), this picture indicates a slowly emerging interest in the forming identity of Central Europe, which, however, did not result yet in an increased number of publications until 2015.

[8] https://ws.clarin-pl.eu/public/geo2/distance_NER.html

[9] https://ipn.gov.pl/en/brief-history-of-poland#1945

[10] https://guides.loc.gov/poland-manuscripts/timeline

6. Conclusions

The goal of this paper was a comprehensive analysis of the spatial distribution of toponyms included in bibliographic records of Polish National Library catalogs (1801–2019). The research was also a demonstration and test of an online tool that allows analysis of large-scale bibliographic databases. The authors focused on place names contained in the "place of publication" field and toponyms automatically recognized in the titles. The research allowed us to achieve the objectives set and to validate the advanced hypotheses. It has been shown that large-scale bibliographic databases are repositories of knowledge and mirrors of culture, showing over long periods of time civilization trends, but also political and economic changes. The maps created by projecting toponyms generated from the records, as well as histograms (time series) of quantitative parameters of publications spectacularly reveal the processes taking place in the social and political reality of Poland and Europe. The transdisciplinary quantitative approach, integrating methods of NLP, data science, linguistics and cultural anthropology, constituting a new paradigm of culturomics, thus gives very good cognitive results also from the point of view of social history. Moreover, the web applications produced during the work on the project allow for independent analyses and interpretations of large bibliographic databases of the Polish National Library. Last but not least – we intend to continue the research presented here on other, much larger library catalogues, e.g., British Library, Library of Congress, Deutsche Nationalbibliothek etc.

Acknowledgments

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.

References

[1] L. Lahti, J. Marjanen, H. Roivainen, M. Tolonen, Bibliographic data science and the history of the book (c. 15001800), Cataloging & Classification Quarterly 57 (2019) 5–23. doi:10.1080/01639374.2018.1543747.

[2] M. Tolonen, L. Lahti, H. Roivainen, J. Marjanen, A quantitative approach to book-printing in sweden and finland, 16401828, Historical Methods: A Journal of Quantitative and Interdisciplinary History 52 (2019) 57–78. doi:10.1080/01615440.2018.1526657.

[3] A. Pawłowski, T. Walkowiak, Automatic recognition of gender and genre in a corpus of microtexts, in: Theory and Applications of Dependable Computer Systems, Springer International Publishing, Cham, 2020, pp. 472–481. doi:10.1007/978-3-030-48256-5_46.

[4] W. Wysota, K. Trzaska, Correlation of bibliographic records for omnis project, in: Theory and Engineering of Dependable Computer Systems and Networks, Springer

International Publishing, Cham, 2021, pp. 487–495. doi:10.1007/978-3-030-76773-0_47.

[5] M. Marcińczuk, J. Kocoń, M. Oleksy, Liner2 — a generic framework for named entity recognition, in: Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 86–91. URL: https://aclanthology.org/W17-1413. doi:10.18653/v1/W17-1413.

[6] M. Marciczuk, J. Koco, M. Gawor, Recognition of named entities for polish-comparison of deep learning and conditional random fields approaches, in: Proceedings of the PolEval 2018 Workshop, Polish Academy of Science, 2018, pp. 77–92.

[7] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing 45 (1997) 2673–2681.

[8] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[9] M. Gritta, M. T. Pilehvar, N. Collier, Which Melbourne? augmenting geocoding with maps, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 1285–1296. URL: https://aclanthology.org/P18-1119. doi:10.18653/v1/P18-1119.

[10] J. Thomale, Interpreting marc: Wheres the bibliographic data?, Code4Lib Journal 11 (2010). URL: https://journal.code4lib.org/articles/3832.

[11] R. Mroczkowski, P. Rybak, A. Wróblewska, I. Gawlik, HerBERT: Efficiently pre-trained transformer-based language model for Polish, in: Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing, Association for Computational Linguistics, Kiyv, Ukraine, 2021, pp. 1–10. URL: https://aclanthology.org/2021.bsnlp-1.1.

[12] M. Marcińczuk, Lemmatization of multi-word common noun phrases and named entities in Polish, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, INCOMA Ltd., Varna, Bulgaria, 2017, pp. 483–491. doi:10.26615/978-954-452-049-6_064.

[13] K. Wróbel, Krnnt: Polish recurrent neural network tagger, in: Z. Vetulani, P. Paroubek (Eds.), Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu, 2017, pp. 386–391.