Journal of Open Humanities Data

Opening a Free Path to Analyze the Discourse Shift in the Soviet Belarusian Newspaper *Zviazda* after the Molotov-Ribbentrop Pact

RESEARCH PAPER

LOÏC BOIZOU

ubiquity press

ABSTRACT

This paper attempts to develop a pipeline designed to convert graphical PDF files of the newspaper *Zviazda* into usable text data in the Belarusian language with search and visualization options. Apart from punctual conversion scripts to allow navigating between formats, the pipeline relies on freely available resources in order to process this relatively under-resourced language (at least for freely available resources). This pipeline was designed to include a graph database and to be compatible with data visualization tools. The ultimate goal is to develop a resource to analyze the political discourse in the Soviet Belarusian press during the Second World War. With a view to validating the pipeline, a pilot study was carried out: it aims to visualize some simple manifestations of the Soviet rhetorical shift about Nazi Germany after the signing of the Molotov-Ribbentrop Pact in order to prove that some useful phenomenon can be revealed even with quite noisy data.

CORRESPONDING AUTHOR:
Loïc Boizou
Research Institute of Natural and Technological Sciences, Vytautas Magnus University, Kaunas, Lithuania
lboizou@gmail.com

KEYWORDS:
NLP; Belarusian language; Graph databases; Discourse; Soviet press

TO CITE THIS ARTICLE:
Boizou, L. (2023). Opening a Free Path to Analyze the Discourse Shift in the Soviet Belarusian Newspaper *Zviazda* after the Molotov-Ribbentrop Pact. *Journal of Open Humanities Data*,9: 23, pp. 1–13. DOI: https://doi.org/10.5334/johd.133

1 INTRODUCTION

In the background of the current Russian-Ukrainian war, the Second World War is pervasive in the political discourse. This study is a first step towards the future goal of comparing the political discourse in wartime in a publication that can be considered the voice of the Belarusian authorities, with a resource allowing to articulate both close and distant reading.

The main aim of the present work is to evaluate a set of resources in order to construct a full pipeline for Belarusian texts from the digitized images of newspaper pages to a database with extended query and visualization capabilities. In order to facilitate working with scarce resources, the whole pipeline is to be based on freely available tools compatible with Linux. This exploratory study also allows identification of some missing or weak parts that will need additional work in the designed pipeline. As the task is defined, it covers both language non-specific components and Belarusian language components, either tools specifically designed for Belarusian or Belarusian models for non-specific tools.

In order to evaluate the adequacy of the pipeline, there is a second practical goal. This article aims to initiate a limited pilot study to reveal the drastic discourse shift in a totalitarian system. Indeed, one can expect that a press organ in a tightly controlled system can undergo a radical discourse reorientation, that would be far more progressive in a more free and polyphonic media space. The current analysis will focus on the association with the adjective 'German' in the Belarusian Soviet press before and after the Treaty of Non-Aggression of August 1939 between Germany and the Union of Soviet Socialist Republics. We would like to note that we do not consider the information treatment of this event in a Belarusian publication to be significantly different from the common Soviet discourse. Given the extreme authoritarianism of the Stalinist organization, a wider multilingual analysis would probably show identical results in all the languages of the Soviet press.

This article will successively present similar or related projects, the selected data, the structure of the pipeline from the rough PDF files to the exploitable corpus, and the results of an attempt to use the database in a minimal pilot study about the Soviet discourse shift after August 1939.

2 SIMILAR STUDIES

The topic of building and using pipelines from images to structured text data is standard in Digital Humanities, hence only several references among the many ones available were selected for this contribution. For example, Nundloll, Smail, Stevens, and Blair (2022) present an ambitious pipeline from rough data to usable text data and propose some ways to improve the quality of the OCR (optical character recognition) output. In addition, this article lists other related text-driven projects in several fields. The present article is very similar to Schätzle, Hund, Dennig, Butt, and Keim (2017) with a tool presentation and a brief pilot study, but with a focus on a concrete and specifically designed tool and not a combination of available tools.

Given the relatively easy access to data, there are many newspaper corpora (as it can be seen in the Clarin main repository).[1] Among other relevant examples related to the totalitarian communist period, Chronopress for Polish (Pawłowski, 2016), DDR-Zeitungsportal[2] for German or the digitized version of *Izvestiia*[3] for Russian deserve mention. A wider perspective is provided in the introduction of Ehrmann, Romanello, Clematide, Ströbel, and Barman (2020).

The issue of OCR and its correction is discussed in Alex and Burns (2014) and Nguyen, Jatowt, Coustaty, Nguyen, and Doucet (2019). The question of the correction of OCR errors and the use of ML (machine learning) techniques for that goal is not systematically addressed in the current pipeline. It is probably of crucial importance for future versions of the dataset.

The question of the potential of tools and models based on graphs, such as graph databases, is also increasingly discussed since it is directly in line with the development of linked data, for example in Perak (2020) (these proceedings are specifically about this topic) or McGillivray, Cassotti, Basile, Di Pierro, and Ferilli (2023). This type of database offers a reliable option for

1 https://www.clarin.eu/resource-families/newspaper-corpora (last accessed: 17 October 2023).

2 https://zefys.staatsbibliothek-berlin.de/ddr-presse/ (last accessed: 17 October 2023).

3 https://www.library.ucsb.edu/node/5839 (last accessed: 17 October 2023).

representing corpora (Efer, 2015) or structuring digital editions (Spadini, Tomasi, & Vogeler, 2021). Their main advantage is to facilitate the development of relatively independent and overlapping information layers (Pęzik, 2013, p. 1; Sippl, Burghardt, & Wolff, 2021, p. 181). The graph schema used for the present database is relatively similar to Efer (2015). Although the most frequently mentioned graph database appear to be Neo4j[4] (when no tool is specifically designed), the present work relies on the more recent Memgraph.[5]

In addition to a graph database, the work presented by Perak (2020) used UDPipe for morphological and syntactic annotation, like the present pipeline. For this pipeline, the priority was given to morphological and syntactic annotation over identification of named entities, while the usual priority with newspaper corpora tends to be the opposite (Ehrmann et al., 2020). Identification of named entities was not considered for this initial Belarusian pipeline.

Although the visualization task remains limited in the present work, the planning of future development is nonetheless relevant. As a way to summarize complex data, visualization is an important topic in Digital Humanities, for example in Lamirel, Dugué, and Cuxac (2016), Allen (2017) or Beck and Butt (2020). In this last reference, the authors also briefly discuss the question of data (mis)consistency, which goes far beyond the issue of visualization. Applications in the field of literary studies as in Scrivner and Davis (2017) could give useful insights for rhetorical and discourse analysis.

3 DATASET DESCRIPTION

The CoNLL-U files[6] that were generated and used for the pilot study can be find under the following reference:

Object name Annotated files of the Soviet Belarusian newspaper *Zviazda* (years 1938, 1939, 1940) (0.1.1)

Direct link https://zenodo.org/record/8424771

Format names and versions CoNNL-U

Creation dates 2023-06-17

Dataset creators Loïc Boizou

Language Belarusian

License CC-BY

Repository name Zenodo

Publication date 2023-10-09

4 THE DATA

The corpus consists of Issues of *Zviazda* (in Belarusian Звязда).[7] This newspaper, which still exists as a pro-government publication in Belarus, was the official publication of the Central Committee and the Minsk regional branch of the Communist Party (Bolsheviks) of Belarus. As such, it is the direct voice of the authorities. Nevertheless, it must be noted that in a situation where the whole press was tightly controlled by the state, all kinds of publications had to speak in unison to a very large extent. "Provincial Russian papers took their cues about wording and descriptions of events from these major publications. Following a pattern developed in 1917, what appeared in *Pravda* one day was likely to appear in *Izvestiia* on the same day or on the next day, and a day or so later in the regional or specialized newspapers and magazines" (Thompson, 1991, p. 388). But the same trend affected *Zviazda* and Soviet newspapers in all other languages as well.

4 https://neo4j.com/ (last accessed: 17 October 2023).

5 https://memgraph.com/ (last accessed: 17 October 2023).

6 The CoNLL-U files are not the final format of the data, but the format used to store the data in the database is less useful for most usage cases.

7 Звязда means 'star', but it is generally considered a Russianism for the standard Belarusian word зорка (zorka).

As a rule, the newspaper consisted of four pages, but some special Issues were longer (e.g., 8 pages for Issue 281 of 1939). It was issued about five times a week, with no publication on fluctuating days. *Zviazda* was first published in Russian, and then, due to the Soviet *korenizatsiia* ('indigenization') policy (in Belarusian Карэнізацыя) of the Twenties, it was published in both Russian and Belarusian from 1925, before finally adopting Belarusian as its only language in the summer of 1927 (entry Звязда in Энцыклапедыя Гісторыі Беларусі (Encyclopedia of Belarusian History), vol. 3, 1996).

The selected Issues cover the years 1938, 1939 and 1940. Apart from the year of the signing of the Treaty of Non-Aggression between Germany and the Union of Soviet Socialist Republics, it was decided to lengthen the time period to include the previous and following years. The year 1938 was significant for its multiple prewar crises, with a breakthrough year for the Spanish civil war that paved the way for the Republican defeat in 1939, the Anschluss and the Munich Conference. Throughout this year, the Soviet Union and Germany were obviously antagonistic powers in the diplomatic arena. The year 1940 was the main year of peaceful coexistence between the two former ideological enemies. This time span allows the Soviet discourse before and after the Treaty to be fully captured.

All these Issues were downloaded from the website of the Presidential Library of Belarus.[8] They constitute a subset of a wider amount of Issues from 1918 to 1945.[9] The Issues are digitized as PDF documents (resolution of 300 ppi) and cannot be used in this form to extract and process textual information. For the years 1938 and 1940 each page was scanned as one PDF page, but for the year 1939 pages were scanned in two parts (the upper and the lower part) with the middle part of the text being repeated in both scans (see Figure 1), except Issue 59/1939.

A certain number of Issues are missing among the documents provided by the Presidential Library (see Table 1). Some original documents are also physically damaged to a certain extent (e.g., Issue 152 of 1940). The number of Issues and pages is provided in Table 2. Thirteen Issues do not have their usual four pages. Some special Issues are longer (281/1940 – 8 p., 1/1939, 58/1939 and 62/1939 – 6 p.). Issues 270/1938 and 38/1939 have 5 pages because one page was scanned twice. In some cases, Issues are not complete (134/1939 – 3.5 p., 54/1939, 187/1939 and 280/1939 – 3 p., 26/1939 and 251/1939 – 2 p.). For one Issue, a single article was scanned (11/1939).

The processing of these documents is explained in the following section about the pipeline.

5 THE PIPELINE

The main task of the present study was to develop a pipeline from the PDF files to the searchable database. As a rule, all components had to be free and to carry out the bulk of the work, with minimal *ad hoc* developed functions (mainly for format conversion). The pipeline was developed on Fedora 38.[10]

5.1 TEXT EXTRACTION AND PREPROCESSING

The conversion from image-based PDF to image files (TIFF or JPG) relies on Poppler[11] while the OCR relies on Tesseract 5.3.0.[12] Given the uneven quality of the PDF images of old newspapers, the OCR is relatively difficult, especially without any attempt to improve the recognition process (either by improving the quality of pictures, by providing a lexicon or by using additional tools like ocrd_tesserocr). We made several attempts to perform OCR with both TIFF and JPG as input and with two different page segmentation modes (with the default mode –pms -3 and with –pms -6). As it is now, the pipeline goes through JPG images with the default page segmentation mode because it recognizes the text flow through different columns. Nonetheless, it is unable

8 https://opac.preslib.org.by:8080/abis/ (last accessed: 17 October 2023).

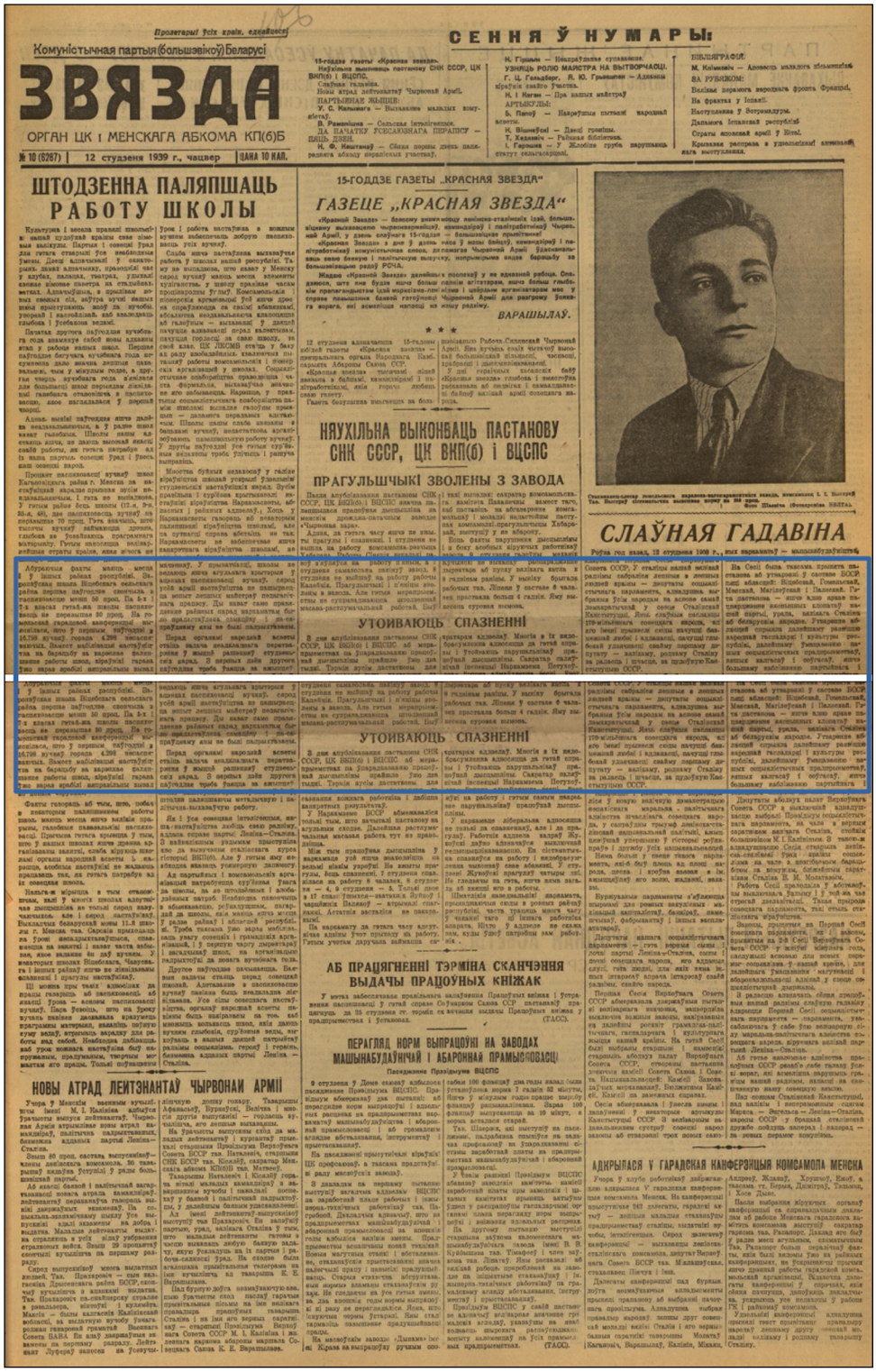
9 *Zviazda* was almost totally discontinued due to the German occupation of Soviet Belarussia from the middle of August 1941 to late January 1943.

10 https://fedoraproject.org/ (last accessed: 17 October 2023).

11 https://poppler.freedesktop.org/ (last accessed: 17 October 2023).

12 https://tesseract-ocr.github.io/ (last accessed: 17 October 2023).

DOI: 10.5334/johd.133



Прозетарыі ўсіх краін, еднайцеся!

Камуністычная партыя (большэвікоў) Беларусі

ЗВЯЗДА

ОРГАН ЦК і МЕНСКАГА АБКОМА КП(б)Б

№ 10 (6287) | 12 студзеня 1939 г., чацвер | ЦАНА 10 КАП.

СЕННЯ Ў НУМАРЫ:

ШТОДЗЕННА ПАЛЯПШАЦЬ РАБОТУ ШКОЛЫ

НОВЫ АТРАД ЛЕЙТЭНАНТАЎ ЧЫРВОНАЙ АРМІІ

15-ГОДДЗЕ ГАЗЕТЫ „КРАСНАЯ ЗВЕЗДА"

ГАЗЕЦЕ „КРАСНАЯ ЗВЕЗДА"

НЯЎХІЛЬНА ВЫКОНВАЦЬ ПАСТАНОВУ СНК СССР, ЦК ВКП(б) і ВЦСПС

ПРАГУЛЬШЧЫКІ ЗВОЛЕНЫ З ЗАВОДА

УТОІВАЮЦЬ СПАЗНЕННІ

АБ ПРАЦЯГНЕННІ ТЭРМІНА СКАНЧЭННЯ ВЫДАЧЫ ПРАЦОЎНЫХ КНІЖАК

ПЕРАГЛЯД НОРМ ВЫРАБОТКІ НА ЗАВОДАХ МАШЫНАБУДАЎНІЧАЙ І АБАРОННАЙ ПРАМЫСЛОВАСЦІ

Пасяджэнне Прэзідыума ВЦСПС

СЛАЎНАЯ ГАДАВІНА

АДКРЫЛАСЯ У ГАРАДСКАЯ КАНФЕРЭНЦЫЯ КОМСАМОЛА МЕНСКА

Figure 1 Example of overlapping PDF pages.

1938	151, 190, 221, 228, 255
1939	37, 50, 66, 114, 118, 119, 127, 129, 206, 226, 241, 271, 284, 296
1940	59, 65, 88, 111, 204, 282, 294

Table 1 Missing Issues by year.

YEAR	PAGES (NUMBER)	ISSUES (NUMBER)
1938	1185	296
1939	1148	288
	(2288 PDF half pages + 4 full pages)	
1940	1188	296
total	3521	880

Table 2 Basic statistics related to Issues.

to properly recognize the article structure on a given page, therefore some unrelated parts of the page can appear as following one another. Early attempts were made to link the flow text with the spatial structure of the page through –psm -6 but this approach needs to be tested further. An important missing step in the pipeline is the merging of both half pages and the deletion of the duplicated middle part. As a consequence, the text flow is always interrupted in the middle of the page. A similar issue arises when an article appears across several pages.

A limited attempt to give an approximate measure of the OCR quality was performed on the files corresponding to the first page of the Issue 200/1938 and to the first half page of the Issue 100/1939. Table 3 shows the word error rate for these two extracted text files.[13]

Table 3 OCR word errors.

	NUMBER OF TOKENS (ORIGINAL TEXT)	NUMBER OF TOKENS (OCRED FILE)	COMMON TOKENS (NUMBER)	MISSPELLED OR MISSING TOKENS
200/1938 (first page)	4223	4217	3795	428 (10%)
100/1939 (first half page)	1864	1602	1352	512 (27%)

The WERs (word error rates) are respectively 10% and 27%. A significant part of the misspelled words shows a single character mistake that might be recoverable with some specific techniques of OCR correction. The most problematic issue manifests itself in the second file (100/1939), where the number of words is notably smaller in the OCRed file than in the original text: all the lowest lines of the PDF page that appeared in a darker area in the PDF page (over 250 words out of 1864 taking into account the number of tokens) were totally missed by the OCR tool, hence the lower WER for the second file (given the token numbers, the WER for the OCRed parts of the page might be around 13%, very close to the WER for the first file). It means that some parts of the content are missing in the extracted data and this serious shortcoming needs further investigations. The WER evaluated for the current data are extremely similar to the rates mentioned in Nguyen et al. (2019) (in Section 2), from 9% to 27%, for a dataset that is mostly related to older newspapers, but consists of texts in English. Moreover, the OCR task was performed at least partly with commercial software.

Two limited cleaning operations were performed on the OCR results. First, a rudimentary script was written to dehyphenate the text and thus to improve the text flow. The result is segmented into sentences and words and converted into a CoNLL-U file by UDPipe 1[14] with the HSE Belarusian model[15] (Shishkina & Lyashevskaya, 2021). Second, the Hunspell Python library[16] was used to correct some country names and nationality nouns or adjectives, like Germany, German, Belarus, Belarusian and Soviet in the CONLLU files. It is very problematic to use a fully automatic spelling correction since all unknown words would be replaced by known ones. Our approach to minimize the risks (except for case endings) was to accept the correction only if one of our keywords (either a country name or a nationality noun/adjective) is among the proposed Hunspell suggestions. For future steps, a different decision possibly involving a partly manual correction of words unrecognized by Hunspell will be necessary. Since the OCR by Tesseract with different inputs and parameters results in different mistakes, it may be possible to combine the different outputs of the same text in order to reduce the number of mistakes, but such an approach has not been tested. However, the main factor for improvement is probably an increase in the quality of the OCR process.

5.2 MORPHOLOGICAL AND SYNTACTIC ANALYSIS

The morphological and syntactic analysis was also performed by UDPipe with the already mentioned HSE Belarusian model. We were not able to find freely available analyzers or Belarusian language models (although Zubov (2019) mentions that such tasks were performed

13 At this point, the number of tokens was calculated without dehyphenation: it means that a word split across two lines is treated as two tokens. As a consequence, the corresponding number of tokens in the morphologically annotated files is lower.

14 https://ufal.mff.cuni.cz/udpipe (last accessed: 17 October 2023).

15 https://github.com/UniversalDependencies/UD_Belarusian-HSE (last accessed: 17 October 2023).

16 https://pypi.org/project/hunspell/ (last accessed: 17 October 2023).

in the Belarusian Academy of Sciences). The HSE Belarusian model is based on a relatively small treebank of about 30,000 tokens and only half of the data were used to train the model (with the second part used for testing). As a result, the quality of the linguistic analysis is insufficient.

Hunspell was considered as another option for tagging the data but the morphological information is totally absent in the standard Hunspell Belarusian dictionary,[17] so it cannot be used as a basic morphological analyzer for Belarusian. Furthermore, even lemmatization is practically impossible since the Belarusian Hunspell dictionary relies on a very fragmented description of the lexicon. Due to the concatenative nature of Hunspell, it does not allow the possible vowel alterations that are frequent in Belarusian stems (in relation with the movement of the lexical stress) to be easily dealt with in the traditional way, that is with a strong approach with word (lemma) and paradigm.[18] It led the dictionary authors to divide single lexical units into several base forms when there are alterations in the stem.[19] Separate singular and plural base forms for nouns are quite frequent (e.g., год 'year', гады 'years'), but some very variable nouns are split into even more base forms (e.g., дзень/дня/дні/дзён 'day(s)'). Other parts of speech can be split as well, e.g., маю and мець for the verb 'to have'. Such an approach proved to be suitable for spell-checking, which is the core purpose of Hunspell, but it does not allow using this tool directly for other goals in Belarusian.

A preliminary evaluation of the results was performed on a very small sample of the first 100 sentences of the Issue 200/1938. It consists of 1093 tokens in total, but the evaluation was limited to the 864 alphabetic tokens since punctuation signs and numeric expressions are almost always correctly lemmatized and tagged for parts of speech.

The results for lemmatization are given in Table 4. Regarding the correct word forms, lemmatization was successful for about 79% of the alphabetic tokens and even close to 85% if the cases in which lemmatization was correct except for capitalization (either incorrectly assigned to the lemma or missing when needed) are included. When the word forms are incorrect (because of mistakes in OCR or dehyphenation), lemmatization was partly correct[20] in 1/4 of the cases.

Table 4 Preliminary results (lemmatization).

LEMMATIZATION OUTCOME	ON CORRECT WORD FORM		ON INCORRECT WORD FORM	
Correct	604	78.9%	24	24.5%
Correct except for capitalization	43	5.6%		
Incorrect	119	15.5%	74	75.5%
Total	766	100%	98	100%

The POS (part-of-speech) tagging was evaluated on the same list of tokens. The results (see Table 5) are relatively similar to the lemmatization results: correct POS assignment[21] amounts to 84% of the alphabetic tokens with correctly provided word forms, when the proportion of correct POS tagging for incorrect word forms is slightly higher and reaches 35%. This is the direct consequence of a greater role of the context in determining the POS, while lemmatization is more strongly lexicon-based.

Table 5 Preliminary results (POS tagging).

POS TAGGING RESULT	ON CORRECT WORD FORM		ON INCORRECT WORD FORM	
Correct	644	84.1%	35	35.7%
Incorrect	122	15.9%	63	64.3%
Total	766	100%	98	100%

17 https://github.com/375gnu/hunspell-be (last accessed: 17 October 2023).

18 Furthermore, the mentioned Belarusian Hunspell dictionary does not use the morphological description fields (such as *stem*, *is*, *po* etc.) that would allow modeling morphology explicitly.

19 In addition, the choice of the base form can vary, e.g., the first person singular (шукаю 'search (first person present)') or the infinitive (рабіць 'to do') for verbs.

20 The term 'partly correct' is used for lemmas that are correct except for a letter wrongly identified during the OCR process.

21 For both lemmatization and POS tagging, the evaluation was considered as correct if the result was acceptable, without consideration for the validity of the decisions concerning the lemmas or the POS tags in borderline examples (e.g., participles, deadjectival adverbs, determiners, substantivized adjectives).

As for the parsing, a quick overview of the 100 mentioned sentences seems to show that about half of the sentences are not properly segmented. The mistakes are especially obvious in the part of the page where headers announce the content of the Issue 200/1938. In more massive paragraphs the segmentation tends to perform better, provided that the sequence of the text lines is correct. When the sentence segmentation is correct or almost correct, the syntactic analysis appears good or relatively good in about half of the sentences. The quality of the parsing does not seem to be related to the sentence length only: nominal sentences were often incorrectly parsed while several longer sentences were surprisingly well analyzed.

Given the small size of the sample, all these preliminary results have to be considered with caution. Nonetheless, it proves that at the current stage, the data are excessively noisy. Besides the mistakes that come from the OCR step, the quality of the syntactic and morphological analysis is too low and this issue must be addressed in the future. Some steps are underway to develop a morphological analyzer on the basis of the Hunspell Belarusian lexicon. An option for improving parsing could be to attempt using a larger Russian dependency model given the high syntactic similarity between Russian and Belarusian.

5.3 CORPUS STORAGE IN A GRAPH DATABASE

The data were finally stored in a database. With a view to develop linked information layers, the decision was made to store the data in a graph database. After trying several options, Memgraph[22] was selected for its reliability and usability with relatively big databases, its ability to import CSV files at high speed, its graph visualization function allowing quick exploration of the database and the option to query the database through Python programs. The CoNLL-U files were converted to suitable CSV files by an *ad hoc* script.

In the current database schema (see Figure 2), vertices (or nodes) have the following properties:

- :Token: position in the document (Int), id field, form, lemma, part of speech, grammatical features, glue (all are strings)
- :Sentence: id field (String)
- :Document: id field (String), year (Int), issue number (Int)

The edges (or relationships) express the relation between each token and the document it belongs to (IS_IN_DOC), the relation between each token and the sentence it belongs to (IS_IN_SENT), the sequential relation between tokens in the text flow (IS_NEXT), the dependency relation between words (IS_DEP) and the relation between the root and its sentence (IS_ROOT). Only the dependency relation has a property, the syntactic type of the dependency relation. The relation IS_ROOT between the root and the sentence is largely redundant with the relation IS_IN_SENT and should probably be replaced by the boolean property is_root for each token.

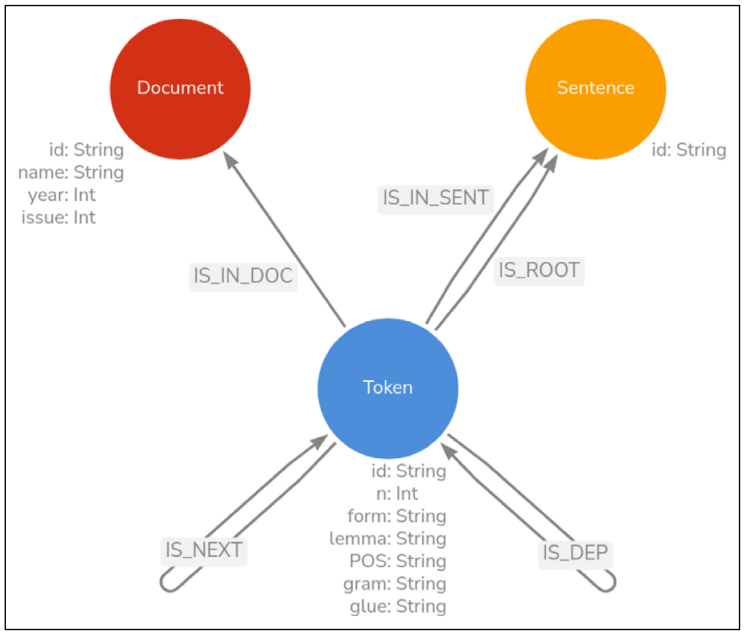


Figure 2 The database schema (visualization generated by Arrows, https://arrows.app).

22 https://memgraph.com/ (last accessed: 17 October 2023).

At its present stage, article, page and paragraph structures are not expressed in the database (although information about newlines could be retrieved through the field *glue*). The quantitative information about the present database is summarized in Table 6.

Table 6 Database summary.

VERTICES (NODES)		EDGES (RELATIONSHIPS)	
tokens	14,297,480	IS_IN_DOC	14,297,480
sentences	1,404,085	IS_IN_SENT	14,297,480
documents	880	IS_NEXT	14,296,600
		IS_DEP	12,893,395
		IS_ROOT	1,404,085
total	15,702,445	total	57,189,040

5.4 VISUALIZATION

While Memgraph Lab allows a quick view of the data, it is also very limited in the way it can visualize these data, only as relation graphs. Given its wide range of possible visualizations, the Python library Plotly[23] was selected to be included in this pipeline. For the current study, it was restricted to simple time diagrams and to static use of the data, but Plotly will make it possible to develop dynamic visualizations in connection with the Memgraph database through Python scripts.

The next part presents how these data, despite the above-mentioned shortcomings, can be used for a concrete small-scale study related to the radical Soviet discourse shift of 1939.

6 A MINIMAL PILOT STUDY

As a minimal exploratory study, the time distribution of some pejorative collocates of the adjective германскі (hermanski) 'German'[24] was extracted from the database. Given the low quality of the parsing, this short study relies on a simple collocation approach with a span of five tokens before the adjective and five tokens after it. This means that it could have been realized through a traditional text database such as BlackLab[25] or NoSketch[26] and that the potential of the graph database is underutilized at this point.

After a quick manual check of the 32,230 collocates of the word германскі to identify a set of potentially negative words with a minimum frequency (at least four occurrences), the following list was created: фашыст 'fascist' (and related words like фашызм 'fascism'), агрэсія 'aggression' (and related words like агрэсар 'aggressor'), правакацыя 'provocation' (and related words), тэрор 'terror' (and related words). The latter three words have a direct negative connotation, while the first word names a major antagonistic ideological system (at some point *the* main one). Two more words, which are related to supposed or actual subversive activities and used to reinforce the feeling of an omnipresent threat, were added to the list: шпіён 'spy' (and related words) and агент 'agent'[27]. The extraction of the selected collocates was performed by regular expressions (e.g., *фашы* '*fasci*') in order to partly recover some incorrect tokens due to OCR errors, hyphenation issues or incorrect lemmatization.

These selected collocates were visualized diachronically for the years 1938, 1939 and 1940. Beside a usual diagram with monthly frequency for each selected collocate (Figure 3), an unusual alternative is provided as Figure 4.

23 https://plotly.com/ (last accessed: 17 October 2023).

24 There is another Belarusian synonym нямецкі, but it is far less frequent in the corpus.

25 https://inl.github.io/BlackLab/ (last accessed: 17 October 2023).

26 https://nlp.fi.muni.cz/trac/noske (last accessed: 17 October 2023).

27 The infamous connotation of this word in the Soviet or post-Soviet context is now widely known, thanks to the widely discussed Russian foreign agent law.

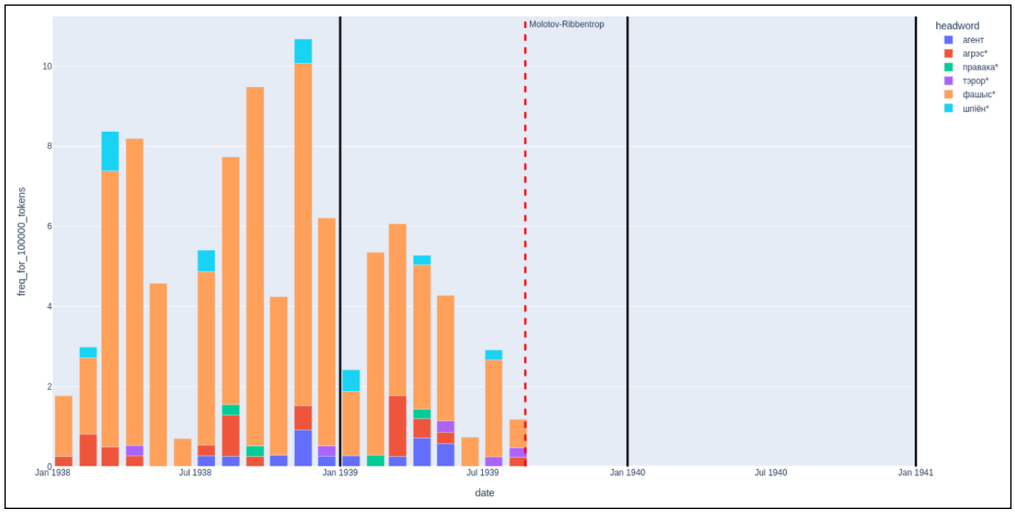


Figure 3 The monthly frequency of the selected colocates of германскі (visualized with Plotly).

This specific co-occurrence schema attempts to represent the appearance of collocates by Issue, in order to obtain a very fine sequential alignment (by date or Issue number), even for collocates that are meaningful not by their higher frequency, but by their only presence or absence. Since the frequency of the selected collocates in each Issue is mostly 0 or 1, rarely more, the (raw or normalized) frequency on the y-axis would have been hardly readable when two or more selected collocates appear once in an Issue. In general, the line would have been almost flat between 0 and 1 with a few peaks to 2. Instead, the ordinate was used to add the total frequency in the given period: each occurrence by Issue[28] as abscissa is disposed on the line representing the total frequency in the whole selected period. Since this number is constant, the dots appear on a straight line, but such a decision allows one to read all the values (except in an improbable case in which the total frequency is exactly the same for two collocates) and to have a measure of the relative weight of each collocated in addition to their temporal distribution. In order to not lose information about eventual multiple occurrences of a collocate in an Issue, the size of the dots varies accordingly.

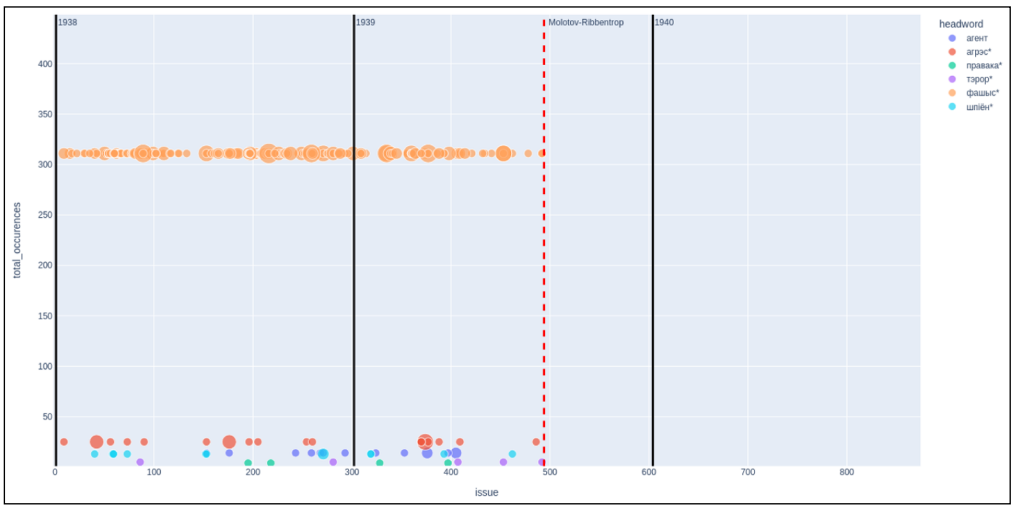


Figure 4 The co-occurrence schema of германскі (visualized with Plotly).

The data in Figures 3 and 4 show that the negatively connotated words stopped being used abruptly immediately after the Soviet-German pact and that this situation continued until the end of the described period (December 1940). It must be emphasized that the later German invasions of Poland, Denmark, Norway, the Netherlands, Belgium, Luxembourg and France did not provoke any return to the previous inflammatory rhetoric against Germany. In the Issues of *Zviazda*, the events related to the tipping point appear in the following way:

- Issue 191 (492 in Figure 4) of August 21: last pejorative mentions of Germany (фашыст 'fascist', фашызм 'fascism', тэрар 'terror').

28 For the sake of simplicity, the Issue numbers in Figure 4 appear as a continuous sequence in the selected period, thus the first Issue of 1939 is numbered as 302 (following the 301 Issues of 1938) and the first Issue of 1940 as 604.

- Issue 192 (493 in Figure 4) of August 22: nothing remarkable.
- Issue 193 (494 in Figure 4) of August 23: the arrival of Ribbentrop in Moscow is announced.
- Issue 194 (495 in Figure 4) of August 24: the signing of the Soviet-German Treaty is announced.

The data confirm at least the first two of the following statements by Ewa M. Thompson: "The signing of the Molotov-Ribbentrop Pact reversed the tone spectacularly. The word *fascist* was eliminated and virtually overnight the press adopted a pro-Nazi point of view regarding Europe" (Thompson 1991, p. 389). The suddenness of such a radical shift illustrates how a totalitarian system can easily switch on or off a certain tone.[29] While the negatively connotated words were not used any more, it does not mean that German topics disappeared from *Zviazda*, as shown in Figure 5.

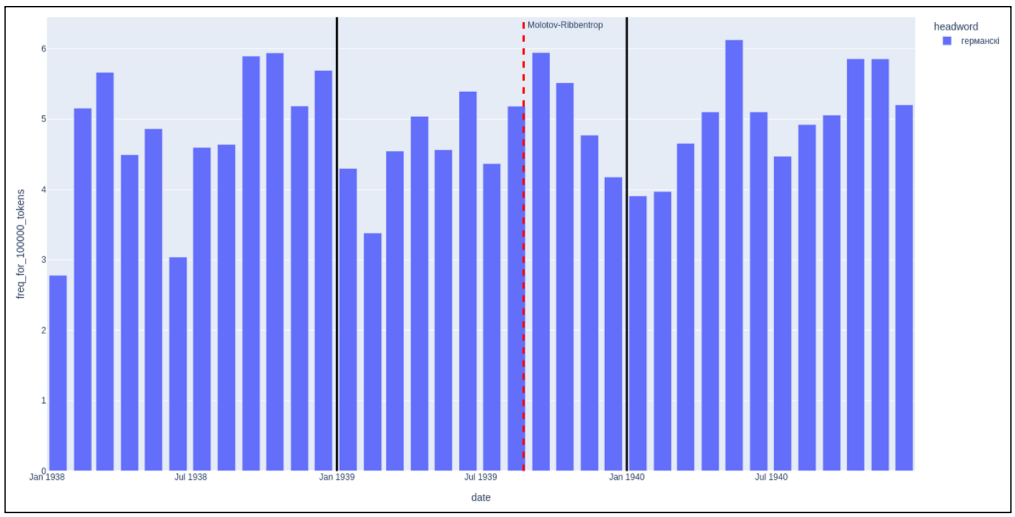


Figure 5 The monthly frequency of германскі (visualized with Plotly).

If we observe the period before the Soviet-German Pact, there is no obvious clue that a political shift was coming in relation to the events that are considered by historians as potential turning points, namely Stalin's speech of March 10, 1939 or the replacement of Litvinov by Molotov on May 3, 1939 as the People's Commissar for Foreign Affairs of the Soviet Union (Weinberg, 1989, Leem, 1998, Haslam, 1997). Nonetheless, the frequency of rhetorical attacks seems to decrease about three months before the Pact, from the middle of May 1939, that is, not long after Molotov's nomination as the Commissar for Foreign Affairs.

7 FINAL REMARKS

This paper shows that it is possible to build a full pipeline from the PDF copies of a Soviet Belarusian newspaper from the Second World War to an annotated corpus with the prospect of developing future richer layers of information and visualizations. Despite the obvious weakness of the data, the minimal exploratory analysis clearly illustrates the radical discourse shift in the Soviet press when the critical coverage of Germany was abandoned "overnight" after the Molotov-Ribbentrop Pact. From this point of view, the exploratory study can be deemed successful.

The approach and the data open the way for a larger outlook on this rhetorical shift. For example, do collocates with a positive connotation appear after the signing of the Molotov-Ribbentrop Pact or is the tone about Germany strictly neutral? Can we observe an opposite trend for the countries that were considered as prospective circumstantial allies before this Treaty, France in particular? In addition, the data allow us to discuss and explore more broadly the topic of public discourse in the Stalinist context and to compare it with other spaces and periods. Some ideas and hypotheses from de Leeuwe, Azrout, Rekker, and Van Spanje (2020) could perhaps be used to analyze the influence of the Soviet legacy on the present Belarusian government press, which is still under strict control.

29 It reminds us of George Orwell's *Nineteen Eighty-Four* where a former existential foe is suddenly presented as a faithful ally (with the crucial difference that Soviet newspapers were not rewritten afterward), what could be explained by the importance of Stalinist totalitarianism in the genesis of this novel.

Nevertheless, the resulting data are extremely noisy, to the extent that some information layers are not usable yet. The Belarusian linguistic component, as far as freely available resources are considered, is still underdeveloped. It is necessary to significantly enhance these resources and to make them easily available. Improving the lemmatization and the syntactic analysis would increase the usability of the graph database, which remains underutilized. As it stands, the same analysis could have been realized with a simpler tool such as AntConc concordancer,[30] but one of the main goals was to explore a sustainable pipeline.

In general, more layers, for example with the rhetorical structure, the annotation of named entities or ideologically marked terms, would also provide this graph database with more options for distant reading, but the priority is clearly to improve the OCR and the morphological and syntactic analysis before enriching the information contained in the *Zviazda* corpus.

ACKNOWLEDGEMENTS

The author is grateful to the Editorial Board, the Reviewers and the Copyeditor for their contributions.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR CONTRIBUTIONS

L. Boizou: Conceptualization, Data curation, Formal Analysis, Visualization, Writing – original draft.

AUTHOR AFFILIATIONS


Loïc Boizou orcid.org/0000-0001-7729-7533
Research Institute of Natural and Technological Sciences, Vytautas Magnus University, Kaunas, Lithuania

REFERENCES

Alex, B., & **Burns, J.** (2014, 05). Estimating and rating the quality of optically character recognised text. In *Datech '14: Proceedings of the first international conference on digital access to textual cultural heritage* (pp. 97–102). DOI: https://doi.org/10.1145/2595188.2595214

Allen, W. (2017). Making corpus data visible: visualising text with research intermediaries. *Corpora, 12*(3), 459–482. DOI: https://doi.org/10.3366/cor.2017.0128

Beck, C., & **Butt, M.** (2020). Visual analytics for historical linguistics: opportunities and challenges. *Journal of Data Mining and Digital Humanities* (pp. 1–23). DOI: https://doi.org/10.46298/jdmdh.6707

de Leeuwe, S. E., Azrout, R., Rekker, R. S. B., & **Van Spanje, J. H. P.** (2020). After all this time? the impact of media and authoritarian history on political news coverage in twelve Western countries. *Journal of Communication, 70*(5), 744–767. DOI: https://doi.org/10.1093/joc/jqaa029

Efer, T. (2015). Text mining with graph databases: Traversal of persisted token-level representations for flexible on-demand processing. In *Autonomous systems – proceedings of the 8th GI conference, VDI Verlag.* Retrieved from http://asv.informatik.uni-leipzig.de/publication/file/332/autsys2015-efer.pdf (last accessed: 17 October 2023).

Ehrmann, M., Romanello, M., Clematide, S., Ströbel, B. P., & **Barman, R.** (2020, 05). Language resources for historical newspapers: the Impresso collection. In *Proceedings of the 12th conference on language resources and evaluation (LREC 2020)* (pp. 958–968). Retrieved from https://aclanthology.org/2020.lrec-1.121 (last accessed: 17 October 2023).

Haslam, J. (1997). Review: Soviet-German relations and the origins of the Second World War: The jury is still out. *The Journal of Modern History, 69*(4), 785–797. DOI: https://doi.org/10.1086/245594

Lamirel, J.-C., Dugué, N., & **Cuxac, P.** (2016). Performing and visualizing temporal analysis of large text data issued for open sources: Past and future methods. In *12th IEEE International conference: Beyond databases, architectures and structures (BDAS'2016)* (pp. 56–76). DOI: https://doi.org/10.1007/978-3-319-34099-9_4

30 https://www.laurenceanthony.net/software/antconc/ (last accessed: 17 October 2023).

Leem, K. H. (1998). The origins of the Nazi-Soviet Non-aggression Pact of 1939. *러시아연구* (*Russian Studies), 8*(1/2), 204–233. Retrieved from https://s-space.snu.ac.kr/bitstream/10371/88002/1/8.%201939%EB%85%84%20%EB%8F%85-%EC%86%8C%20%EB%B6%88%EA%B0%80%EC%B9%A8%20%EC%A1%B0%EC%95%BD%EC%9D%98%20%EA%B8%B0%EC%9B%90.pdf (last accessed: 17 October 2023).

McGillivray, B., Cassotti, P., Basile, P., Di Pierro, D., & **Ferilli, S.** (2023). Using graph databases for historical language data: Challenges and opportunities. In *IRCDL 2023 – information and research science connecting to digital and library science 2023* (pp. 88–96). Retrieved from https://ceur-ws.org/Vol-3365/short7.pdf (last accessed: 17 October 2023).

Nguyen, T.-T.-H., Jatowt, A., Coustaty, M., Nguyen, N.-V., & **Doucet, A.** (2019, 06). Deep statistical analysis of OCR errors for effective post-OCR processing. In *JCDL '19: Proceedings of the 18th joint conference on digital libraries* (pp. 29–38). DOI: https://doi.org/10.1109/JCDL.2019.00015

Nundloll, V., Smail, R., Stevens, C., & **Blair, G.** (2022). Automating the extraction of information from a historical text and building a linked data model for the domain of ecology and conservation science. *Heliyon, 8*(10), e10710. DOI: https://doi.org/10.1016/j.heliyon.2022.e10710

Pawłowski, A. (2016). Chronological corpora: Challenges and opportunities of sequential analysis. the example of ChronoPress corpus of Polish. In *In digital humanities 2016: Conference abstracts.* (pp. 311–313). Retrieved from https://www.researchgate.net/publication/311536886_Chronological_corpora_Challenges_and_opportunities_of_sequential_analysis_The_example_of_ChronoPress_corpus_of_Polish (last accessed: 17 October 2023).

Perak, B. (2020). Modeling semantic relations from a dependency-based graph: A corpus-based network analysis of Croatian parliamentary debates. In *Graph technologies in the humanities – proceedings 2020* (pp. 172–192). Retrieved from https://ceur-ws.org/Vol-3110/paper9.pdf (last accessed: 17 October 2023).

Pęzik, P. (2013, 10). Indexed graph databases for querying rich TEI annotation. In *Perspectives on querying TEI-annotated data*. Retrieved from https://master.dl.sourceforge.net/project/lingsig/Documents/queryTEI-2013/abstracts/Pezik_2013_QueryTEI-abstract.pdf?viasf=1 (last accessed: 17 October 2023).

Schätzle, C., Hund, M., Dennig, F. L., Butt, M., & **Keim, D. A.** (2017). HistoBankVis: Detecting language change via data visualization. In *Proceedings of the NoDaLiDa 2017 workshop on processing historical language* (pp. 32–39). Retrieved from https://aclanthology.org/W17-0507.pdf (last accessed: 17 October 2023).

Scrivner, O., & **Davis, J.** (2017). Interactive text mining suite: Data visualization for literary studies. In *Proceedings of the workshop on corpora in the digital humanities (CDH 2017)* (pp. 29–38). Retrieved from https://ceur-ws.org/Vol-1786/scrivner.pdf (last accessed: 17 October 2023).

Shishkina, Y., & **Lyashevskaya, O.** (2021). Sculpting enhanced dependencies for Belarusian. In *Analysis of images, social networks and texts: 10th international conference* (pp. 137–147). DOI: https://doi.org/10.1007/978-3-031-16500-9_12

Sippl, C., Burghardt, M., & **Wolff, C.** (2021). Modelling cross-document interdependencies in medieval charters of the St. Katharinenspital in Regensburg. In *Graph data-models and semantic web technologies in scholarly digital editing* (pp. 181–203). Retrieved from https://kups.ub.uni-koeln.de/55234/1/SipplEtAl.pdf (last accessed: 17 October 2023).

Spadini, E., Tomasi, F., & **Vogeler, G.** (2021). *Graph data-models and semantic web technologies in scholarly digital editing.* Retrieved from https://www.i-d-e.de/publikationen/schriften/bd-15-graph-data-models/ (last accessed: 17 October 2023).

Thompson, E. M. (1991). Nationalist propaganda in the Soviet Russian press, 1939–1941. *Slavic Review, 50*(2), 385–399. DOI: https://doi.org/10.2307/2500213

Weinberg, G. L. (1989). The Nazi-Soviet pacts: A half-century later. *Foreign Affairs, 68*(4), 175–189. DOI: https://doi.org/10.2307/20044116

Zubov, A. V. (2019). The creation of the large corpus of Belarussian language and the use of it for the invesigation the Belarussian language and its connection with the different languages of european (in Russian). In *Proceedings of the international conference «Corpus linguistics–2019»* (pp. 23–29). Retrieved from https://events.spbu.ru/eventsContent/events/2019/corpora/corp_sborn.pdf (last accessed: 17 October 2023).

TO CITE THIS ARTICLE:
Boizou, L. (2023). Opening a Free Path to Analyze the Discourse Shift in the Soviet Belarusian Newspaper *Zviazda* after the Molotov-Ribbentrop Pact. *Journal of Open Humanities Data*, 9: 23, pp. 1–13. DOI: https://doi.org/10.5334/johd.133

Submitted: 18 August 2023
Accepted: 12 October 2023
Published: 10 November 2023

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.