TNA ACTIVITY REPORT

Exploring classical Belarusian fiction with the methods of corpus and computational linguistics

Author: Alyaxey Yaskevich

Current position: PhD Student

Affiliation: University of Warsaw

Host institution: Charles University, Prague

Mentor(s): Michal Křen

Period of stay: 3 November 2024 until 25 January 2025

Introduction

Motivation

The idea of the project is to build a corpus of the texts of Belarusian classical fiction of the first half of 20th century.

As for 2024, still the corpus projects on Belarusian can be counted on one hand. Belarusian literary tradition is not so much studied, as the written heritage of other Slavic nations that have similar history. One can say that Belarusian literature is practically absent on the map of the projects implemented in the framework of the digital humanities. The literary studies on Belarusian fiction lack the good cases of applying up-to-date approaches to this material.

Project idea & research focus

The period from the beginning of the 20th century till the end of the World War II could be considered as classical age of Belarusian literature. The writers of that epoch mostly were original talents. They were not studied Belarusian at the school, since there were no Belarusian schools at that time. The language and style they used were not influenced by politics of Russianization and Sovietization implemented by the institute of literary editors and self-censorship, as it take place after the war (however, the process was started in 30th).

The idea of the project involved gathering the texts of that epoch, building a corpus

CLS INFRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

and providing public access to it. Since after this work is completed I get access to newly built corpus, I planned to do a general data-driven research on the corpus. It can be formulated as “What would we see if we take a look at the literary works of the whole age through the lens of computational methods?”

Project implementation steps

Collection of the corpus & data processing

I made a list of 17 authors who were active mostly in that time. Among them the only female writer is Tsiotka.
Here is the list. The number means the quantity of the texts of that author in the corpus.

Yakub Kolas	193
Zmitrok Biadula	80
Uladzislau Galubok	68
Mikhas’ Lyn’kou	47
Mikhas’ Zaretski	45
Maksim Garetski	42

Kuz’ma Chorny	40
Yadvigin Sha	40
Yanka Maur	39
Tsishka Gartny	23
Lukash Kaliuga	14
Andrej Mryi	14

Tsiotka	10
Mikhas’ Charot	10
Yanka Niomanski	9
Yazep Liosik	7
Maksim Bagdanovich	3

I used those online libraries as the text sources:
1) Беларуская Палічка (Little Belarusian Bookshelf, knihi.com)
2) Родныя вобразы (Native Scenery, rv-blr.com)
3) Вікікрыніцы/Wikisources (be.wikisource.org)

The original formats of the files available on mentioned platforms were Epub, Fiction Book 2 and HTML. So, I wrote converters pipeline to preprocess the texts, to extract available metadata and to store them as plain text, as it was required for next steps of making the corpus.

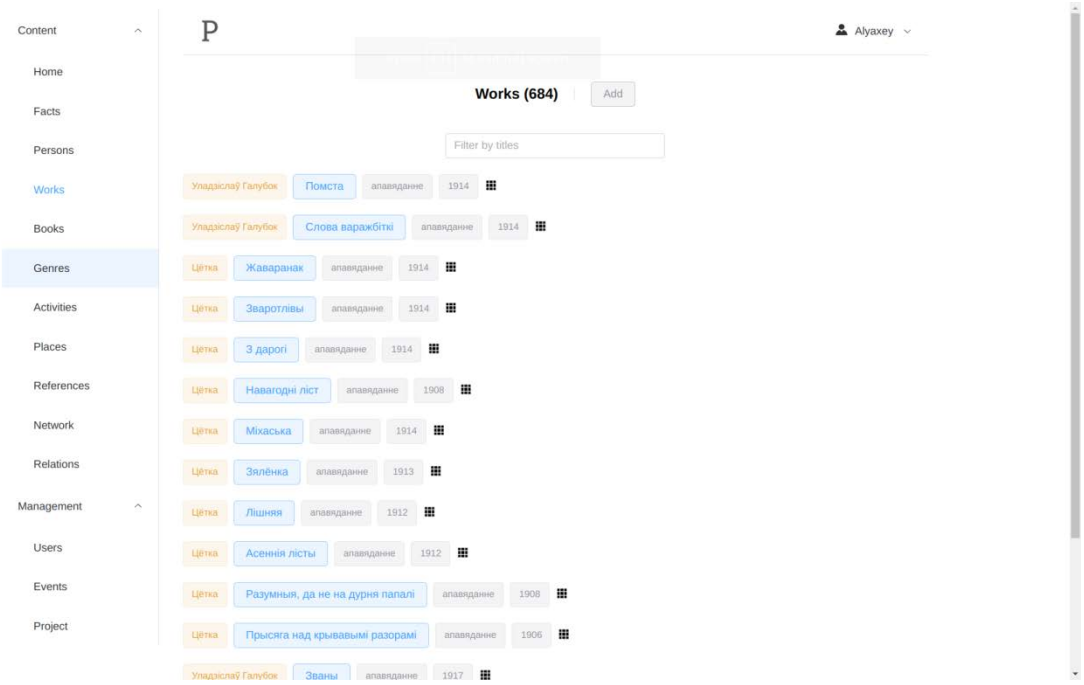
Unfortunately, the quality of the data from all mentioned sources was far from excellent. Some texts contained artifacts from the non-Unicode age, OCR-caused typos. Despite the online platforms were positioned as libraries, the attached text attributes were inconsistent (e.g. genre) and often were missed (year).
The case of dating is important for me, since I set strict time-frame for the texts. However, from this list of Belarusian classical writers before the end of the World War II:

- 6 were executed by Soviet NKVD as Belarusian nationalists
- 3 were arrested and died in prison
- 5 died from illness.

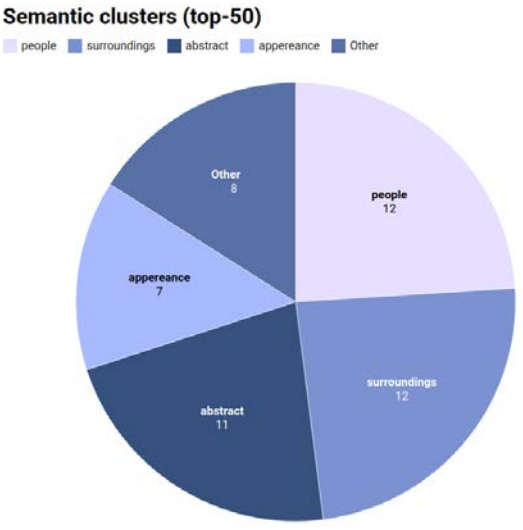
Only 3 persons outlived the Stalin’s repressions and the war, and their works are well documented (regarding the dates).

Data management & corpus publishing

The feature of this project is that I stored all the data in the web platform *Persona* I develop. Originally, it was designed for organizing metainformation only – as a prosopographical database (that contains biographical data of creative persons and information about their works). But I added the functions to manage texts (see the screenshot of the platform below).



Every title was annotated with basic metadata (author, genre, year of completion or first publishing). After that, the texts were tagged with UDPipe (*belarusian-hse-ud-2.15-241121* model) and resulted CONLL-U files were converted into XML-based format for Kontext (the corpus management platform of the Czech National Corpus). The final annotation of every token in corpus contains lemma, part-of-speech and syntactic information. The full annotated corpus has about 4 mil tokens from 664 texts. The real number of the collected texts is higher (684), but they were filtered by date (till 1950) and genre (only fiction) for corpus.
Unfortunately, the quality of part-of-speech annotation and produced lemmas for the tokens from the texts of the corpus is not really good enough. The reason is that UDPipe was trained on a small corpus containing some issues in its annotation. Except of that, the texts in my corpus are written not in standard Belarusian of 21st century, they have many forms that now would be classified as archaic or dialectal. Obviously, this kind of data was not presented in the Belarusian train set for UDPipe.



Data exploration

Nouns frequencies

The simplest way to see what is a text about is to look at the word frequencies. The semantic of clusters top-50 of the corpus are: 1. people (gender-based nominations and types of family relations, 12), 2. surroundings of the peasant's house (12), abstract phenomena (11), body parts and physical features (7), nature (4) and time units (4). Even such a distant reading of the whole corpus says that the texts are

mostly devoted to everyday life of Belarusian peasants, their relationships in a family, interaction with a surroundings and reflections about all that. The texts of the first half of the century are still more into the life before the Russian revolution since the word *pan* (*пан, 'mister/master')* on the 18th place, when *tavarysh* (*таварыш, 'comrade')* is on 34th. They both are used for addressing an adult male person: first on is a marker of "old" prerevolutionary life, when second one is a strong Soviet marker. Also, the gender disproportion is evident: the texts tell us rather about men, than about women.

чалавек	man	people
час	time (age)	surroundings
рука	hand	body parts
вока	eye	body parts
дзень	day	nature
хата	(peasant's) house	surroundings
раз	time (round)	time
слова	word	abstract
бацька	father	people
галава	head	body parts
год	year	time
зямля	soil	surroundings
справа	business/case	abstract
жыццё	life	abstract
бок	side	abstract
дарога	way	surroundings
месца	place	surroundings
пан	mister/master	people
лес	forest	nature
твар	face	body parts
голас	voice	body parts
думка	thought	abstract
дзед	grandfather	people
нага	leg	body parts
свет	world	surroundings

вёска	village	surroundings
хлопец	boy	people
сэрца	heart	body parts
праўда	truth	abstract
дзверы	doors	surroundings
жонка	wife	people
вуліца	street	surroundings
работа	work (labor)	abstract
таварыш	comrade	people
брат	brother	people
маці	mother	people
двор	homestead	surroundings
ноч	night	time
вада	water	nature
сонца	sun	nature
дзіця	baby	people
сын	son	people
праца	work (job)	abstract
бог	God	abstract
сіла	energy/power	abstract
хвіліна	minute/moment	time
дзяўчына	girl	people
стол	table	surroundings
дом	house	surroundings
душа	soul	abstract

Place names frequencies

If common nouns tell us about the things and phenomena an author focuses, geographical proper nouns show how one reflects the space of real world in writing.

Here is the list of countries and regions (first number is frequency of a name, second one is number of texts it was mentioned).

Belarus	294	81
Germany	221	33
Russia	170	45
Poland	166	45
America	114	32

Palessie	95	12
Europe	53	18
Siberia	46	29
Ukraine	32	20

France	25	16
England	24	15
USSR	24	15

Such high frequency of *Germany* is related to the fact that Belarus significantly suffered from combat operations of the World War I. It is interesting that *Russia*

competes with *Poland*. Unfortunately, I don't have comparable corpus for modern fiction, but I wouldn't expect the similar proportion now. In 21st century - in information space - *Russia* became "closer" to *Belarus* than it was even during first decades of the Soviet rule. Moreover, before 1939, Western Belarus was part of Poland and it obviously attracted attention of the people of Soviet Belarus. *America* was more popular than *Europe*, because many Belarusians immigrated to the United States in the beginning of 20th century, while Europe was not so attractive destination being destroyed by the war. During 19th century Siberia was a region where political dissents were sent into exile from western part of Russian Empire. In modern texts I would say it would be mentioned not really often, now it's a just distant region of Russia. As to *Ukraine*, that took 9th place, even lower than *Palessie* (specific ethnic region of Belarus), it is really significant difference with what we have nowadays. I would say that now one may expect *Russia* and *Ukraine* to hold first places. Then, however, Ukraine was just a newly-formed neighboring state that was not really influencing Belarusian life -- neither politically, nor culturally.

The top list of the cities mentioned in Belarusian classical fiction is even more impressive than one of the countries. It captured completely different political situation and the system of cultural interaction that we have today. *Vilnius* (in Belarusian – *Вільня*, *Vilnia*) was an important cultural center of Belarusian culture (and Polish as well, by the way). *Moscow* got high numbers, because it received capital city status in Soviet Russia (*Saint Petersburg* was a capital of Russian Empire). 4th place of *Warsaw* it is a mark that then the Polish capital was still close and important for Belarusians, even after Polish-Bolshevik war. What happened in Warsaw was important for Belarusians after the Polish-Soviet War, because of the abovementioned case of Western Belarus and the attitude of Polish government to Belarusians. *Smolensk* (in Belarusian – *Smalensk*) came to Belaruisn novels because it was treated as Belarusian town before the Russian revolution, as well as it was a staging area during the war. Soviet Belarus was proclaimed in December 1918 in Smalensk. *Homiel* was and is a border city close both to Russia and Ukraine, it is a symbol of Southern border of Belarus.

Minsk	427	71
Vilnius	239	25
Moscow	171	51
Warsaw	72	27
Saint Petersburg	40	16
Smalensk	31	13
Homiel	24	12

What I would expect in today's books? Probably, Vilnius and Warsaw would be mentioned, but Belarusian cities would be more frequent, so Lithuanian and Polish capitals would not make it to the top. *Smolensk* disappears from the top list, and *Moscow* extends it distance from *Saint Petersburg*.

Initially, I planned to look at the "map" of the cities – to visualize their distances in vector space, but the number of occurrences was not high enough. Anyway, I experimented with making embedding on the basis of this corpus that looked too small for that task: *word2vec* algorithm via Python package *gensim* (5-token window, vector dimensionality 100).

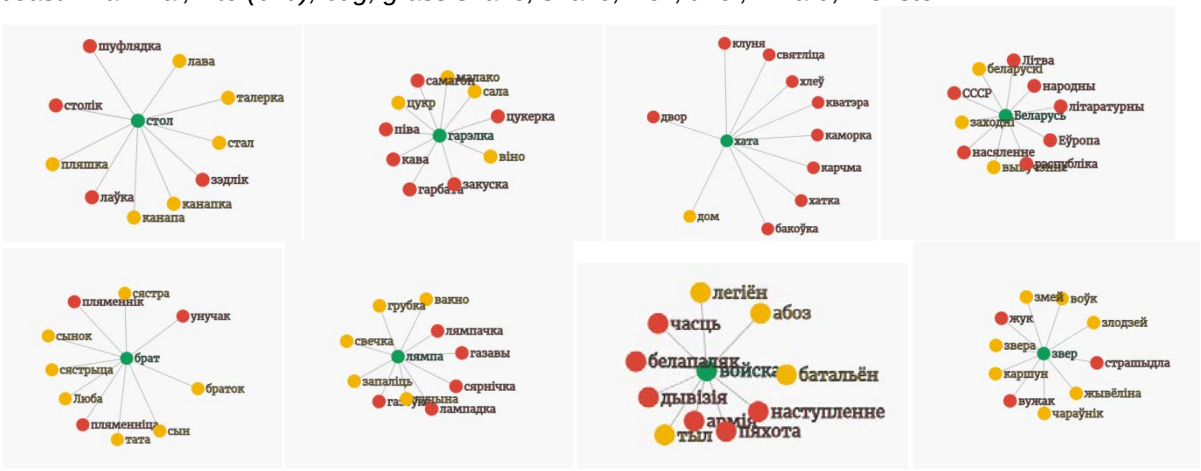
Semantics in vector space

I explored the model with visualization tool I developed. Surprisingly, some topical clusters look meaningful. It means that this resource and the tool could be used for research as well in this manner.

From left to right (center → neighbors):

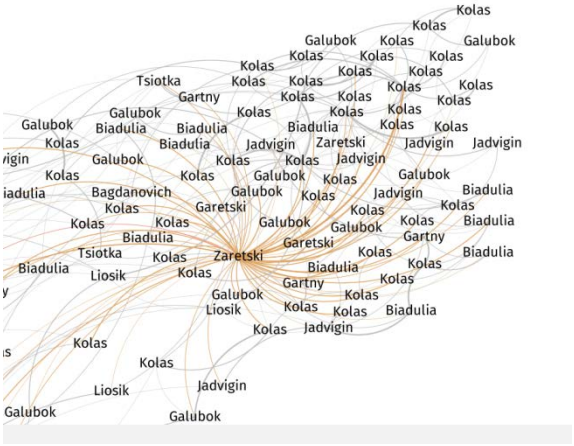
table → plate, bank, small table, bottle, drawer, couch, seat
vodka → wine, sweets, sugar, beer, tea, milk, appetizer, coffee, bacon
(peasants's) house → enclosure, house, side room, barn, apartment, tavern

Belarus → people, literature, Europe, republic, population, Western, USSR, Belarusian, Lithuania, study
brother → sister, grandson, son, daddy, nephew (she), nephew (he)
lamp → gas, bulb*, lantern, to light, stove, window*
army → offensive, unit, battalion, division, home front, , infantry
beast → animal, kite (bird), bug, grass snake, snake, wolf, thief, wizard, monster



Stylometry

It was not possible to omit such a wide-spread approach to deal with literary corpora as stylometry. With *stylo* tool one can take a look into similarity of author style. I prepared a corpus in a format required by *stylo*, and run it with its default settings (Classical Delta). I put here the picture of the clustering results produced with *stylo* and processed with *Gephi*.



Zoom on nortwestern cluster

There are quite independent clusters as Chorny, Lyn'kou, Haretski, Biadula, Zaretski. However, all the writers have some texts that are closer to other writers than to their core oeuvre, especially remarkable in this regard Yakub Kolas (the most well-known Belarusian writer). He has works almost in every cluster of other authors. Probably, it could be explained either that he wrote quite different texts during his semicentenary career, or that he influenced other authors.

The visualization shows that Lukash Kaliuga should be discussed separately. His cluster is significantly distant from other ones. I got to know the name of Kaliuga only when I started to work on this project. He was a first person from this list who was repressed - first arrest was in February 1933. It means that his books were removed from the libraries, his name was erased from the encyclopedias and textbooks, and this process started earlier than it affected others. So, as the tool suggests, he is a very genuine writer. But his name and works are almost forgotten.

Although the stylometric experiment produced interesting results, it was just a first touch; my explanation is not final. This material needs further investigation and deeper look both into 2D/3D visualization and the text content. It could be crucial to see other dimensions, such as dates. Maybe, texts written at the same time could share something in common, that's why some works are closer to the cores of the authors. Maybe, one has to check other clustering methods and see what they will give.

Outcomes

The new Belarusian 4-million corpus is available at the web platform of Czech National Corpus (https://www.korpus.cz/kontext/).

During the work on my project I used the tools:

- **Persona** (developed by me, customized its functions to the project, https://github.com/yaskevich/persona)
- **UDPipe** (developed by Charles University, https://github.com/ufal/udpipe, available as a web-service at https://lindat.mff.cuni.cz/services/udpipe/)
- **Kontext** (developed by Charles University, https://github.com/czcorpus/kontext, available as web-platform at https://www.korpus.cz/kontext/)
- **Stylo** (https://github.com/computationalstylistics/stylo)
- **Gephi** (https://gephi.org/)
- **Gensim** (https://radimrehurek.com/gensim/)
- **Semantic Vector Visualization Tool** (developed by me, https://github.com/yaskevich/vectoresjs)

Both *UDPipe* and *Kontext* are available at the infrastructure of the Charles University.

Kontext is the most complicated piece of software, being a set of several applications. Without help from the host institution it would be very hard to manage this system. I appreciate supportiveness and hospitality of my colleagues from the Department of Linguistics (former Institute of Czech National Corpus) of the Charles University in Prague, Czech Republic.

During my stay in Prague I attended seminars at the Institute of Formal and Applied Linguistics and at the Institute of Czech National Corpus of the Charles University. Also, I presented my project at the seminar of the Institute of Czech National Corpus.

Future work

The project became a step into underexplored lands. It showed that one can get completely new look into literature by means of the lens of computational methods. In terms of this project I used approaches from digital humanities and corpus/computational linguistics. Practically, here we have a system of **prosopographic database for metadata related to the texts and the authors** connected with **linguistic corpus (morphology + syntax) providing comprehensive toolset for text search**.

Unfortunately, to extend this project from this proof-of-concept level to its true potential one would need a research team and proper funding.

- Still, much work should be done on the versions: I have the texts in my database, but it is not clear which editions do they represent. In online libraries the texts are attributed with the year of writing/first publishing, despite the same text with the same dating but from other online source may differ, because they actually represent separate editions (early/late). After that work is done, the attitude behind the editions could be uncovered: whether another version was just improved or it was ideologically corrected/censored? Are newer versions are actually “better” than older?
- Another minor issue is dealing with different scripts (Latin/Cyrillic) and orthographies of the publication. They should be stored in corpus in some unified form.
- The quality of the language model for automatic text annotation/lemmatization is not good enough. There is a need in gold standard for Belarusian. And separate task is customizing it (or producing another annotated train set) especially for the texts of first quarter of 20th century.
- The tools I developed are rather MVP and some efforts should be invested into them to make production-ready. What is good is that they are not designed for specific datasets like I dealt with, they are universal. Thus, that entire pipeline could be reused for texts of other language and culture.
- The database of authors and texts needs completion. Biography of the authors, their bibliography, bibliography on them, and the events of their lives – it’s a big work to do. However, if it’s done, one would be able to query the database and make subcorpora based on the queries (e.g. authors who born in the same decade or who belong to the same political party or association). The literary historians will get direct access to annotated texts, derived metrics and visualizations.
- The corpus can be extended with poetry and drama; the upper bound of the timeframe could be raised.
- The comparable balanced modern corpus should be compiled to allow make conclusions about word frequency differences (and other metrics) supported with data.
- More sophisticated methods to analyze the data could be applied (extraction of network of characters, detecting collocations, keywords and specific words for texts/authors, sentiment analysis, topic modeling etc.).