UDC 81–131

D.B. Slavenskoj

Comenius University, Bratislava, Slovakia
(E-mail: danslav.slavenskoj@uniba.sk)

Corpus Methods of Analysis of Slavic Language Literature: A Small Corpus of Svetlana Alekseyevich's Works

Examining the state of corpus-based approaches to the analysis of Slavic language literature, the rarity of such research is identified. The advantages of such an approach are noted, and previous research by citing Koteyko is identified. Given that few studies are currently done in Slavic languages or using Slavic language texts using corpus based approaches, the author suggests further research is required. A small corpus of the literary works of Svetlana Alekseyevich was created. The corpus was then analyzed and revealed several tendencies and the focus of the works of Svetlana Alekseyevich. This shows how such an analysis of even a small corpus can be a useful tool in the research of literature.

Key words: slavic language literature**,** corpus methods of analysis, Svetlana Alekseyevich, autor srtyle.

Corpus-based approaches to East Slavic literature of any kind are not at present, commonly undertaken, even though, according to scholars of literature of other languages, such research methods yields superior results. According to Biber [1], «a corpus provides the best way to represent a textual domain, and corpus analysis is the most powerful empirical approach for analyzing the patterns of language use in that domain (emphasis added)» Koteyko.

How has the language of East Slavic literatures in the decades following 1991 changed? Koteyko, while focusing on post-Soviet Russia, identifies this time period as one of «rapid changes in political and social life» that «were accompanied by dramatic shifts on the socio-linguistic landscape» [2; 32]. It is the exploration of this shift in political language and its reflection in literature, in Ukraine and Belarus, that concerns the proposed research. According to Ryazanova-Clarke [3; 117] in the period following 1991, with «the disappearance of the despotic Soviet state and, together with it, the ritual Communist Party rhetoric and the ideological prevalence over public speech could not but incur deep shifts in the Russian language», a change which affected «lexis, word formation and morphology». A corpus based study of contemporary East Slavic literature may reveal the ways in which language itself is changing: certain word forms, first appearing in the political or legal realm, make their way into the discourse of non-fiction.

According to scholars of English literature, Almela and Keshabyan [4; i], corpus-based, computational investigations of literature, using computational tools, yield superior results to traditional methods of research, because corpus-based «analyses of literary texts can provide a notable degree of confidence, since this kind of investigation is not related to the mere interpretation of the meaning through readers' perception». Cantos-Gómez and Sánchez [5] note that «since samples of language in digital format are nowadays easily accessible and computers allow for a quick processing of huge amounts of data, the change of the research paradigm is shifting from theoretically based constructs to data based ones.»

However, scholars of literature have been «traditionally reluctant to objective and formal analysis,» even though previous «theoretical constructs have proved to be rather precarious» [4; vii]. Applying such a methodology to examine modern East Slavic literature, has not been widely adopted. However, in keeping with these developments in the study of literature, it is important to note the availability of digital sources, which could for the basis for further corpus based research.

Biber notes three corpus approaches [1; 15]: «'keyword' analysis, identifying typical extended lexical phrases, and collocational analysis». Of these, according to Biber, 'keyword' analysis remains the most common, and the most straight forward approach. A variety of computational tools exist for performing this analysis.

Literary texts themselves are likewise often available in a digital format, and where they are not readily available, they may be digitized for the purposes of this research project. If a digitized works are not readily available, a significant selection of literary works from each year may be taken into consideration based on criteria to be identified. Koteyko, in her corpus-assisted research into post-Soviet Russian political discourse, showed that it is possible for a researcher to create a corpus for targeted study [2; 55].

There are two other important elements limiting the scope of the proposed work: language, and time frame. Posokhin notes [6], among «characteristic traits of present-day literary Belarus», «the dominance of Russian language on the book market.» Likewise, Russian, according to Krouglov [7] «remains one of the major languages in Ukraine by the number of speakers and its use in the media». However, while the Russian language remains significant throughout Ukraine and Belarus, Ukrainian and Belarusian have strong literary traditions of their own, and such a study would have the advantage of providing a more complete grasp of the problem if it considered all East Slavic languages. The proposed work will therefore be limited in scope to East Slavic language works.

The time period of the post 1991 decade provides us ample data which may have already been digitized, as well as a dynamic period of political change on the path of social transformation after independence following the collapse of the Soviet Union in 1991. According to Koteyko [2; 3], «the link between language change and politics is particularly acute during the time of social upheaval, the post-Soviet period represents a great opportunity to explore the process of discursive change and stability.»

Since at present «the vast majority of corpus-based analyses tend to rely on texts written in English or other languages of the European Union» [2; 5], corpus-based methods in the research of East Slavic texts deserves more scholarly attention.

As an example of such a corpus-based approach, we analyzed the works of Svetlana Alekseyevich, a Belarusian author, and winner of the Noble Prize in Literature in 2015, and was the first author from Belarus to receive this award. Although she is from Belarus, reflective of the linguistic situation in the country as noted by Posokhin [6], her works are in Russian.

A small corpus of her published novels, namely: Зачарованные смертью (Zacharovannye Smertyu, Enchanted with Death) (Belarusian: 1993, Russian: 1994); Последние свидетели: сто недетских колыбельных (Poslednie svideteli: sto nedetskikh kolybelnykh, The Last Witnesses: A Hundred of Unchildlike Lullabys), Moscow, Palmira, 2004, ISBN 5–94957–040–5 (first edition: Moscow: Molodaya Gvardiya, 1985); У войны не женское лицо (U voyny ne zhenskoe litso, War Does Not Have a Woman's Face), Minsk: Mastatskaya litaratura, 1985; Цинковые мальчики (Tsinkovye malchiki, Boys in Zinc), Moscow: Molodaya Gvardiya, 1991; and Чернобыльская молитва (Chernobylskaya molitva, Chernobyl Prayer), Moscow: Ostozhye, 1997. ISBN 5–86095–088–8, were used to create the corpus.

These works were selected because they a readily available in a digitized full-text format and provide us with a snapshot of the work of a noted author in the time period 1991–2000, which was identified above as a time of significant social change in the region.

Using the concordance program CasualConc, the corpus was compiled and analyzed. CasualConc <https://sites.google.com/site/casualconc/> can be used to generate concordance lines, analyze word clusters, perform collocation analysis, and do simple word counts. For the purposes of this study, keywords relating to Russian verbs of motion were used as input. There were no issues handling the Russian language text, however, when searching, word morphology had to be taken into account, as there is no built in function that takes care of this, as the program was initially designed for the analysis of Japanese and English language texts.

Although pure keyword searches, as well as all the other methods of analysis using a corpus based approach were possible, with such a small corpus, the most interesting results were from to be found from collocation analysis.

The significant of any research is to provides insight into both the way language is used by an author, however, it may also be used to further reveal the degree of external influences that has been exerted over authors in the region at a time of great social change. Importantly, unlike more ephemeral forms of expression, literature has the potential to leave a lasting impression and survive as an active medium of culture for future generations. Therefore, the effects of any influence on literature, may be felt for years after the work has been written and may continue to affect a society in the future.

Collocation analysis is a corpus-linguistics statistical tool that allows us to identify words that occur together in the corpus. Using this tool, we can identify the frequency of such collocations, and these collocations can reveal tendencies in the author's use of the language.

Focusing on the keyword of the verb быть 'to be' and the verb мочь 'to be able to', and человек 'human', we were able to identify the tendency of Svetlana Alekseyevich towards a pessimistic outlook towards what a human can be or is able to achieve, even without a close reading of the texts themselves.

Table 1

Cluster List Output быть ‘to be’
2-word cluster: 384 / 672 found in 5 files

Cluster		Frequency	Translation
1	не будет	63	will not be
2	все будет	14	everything will be
2	что будет	14	what will be
4	это будет	13	this will be
5	он будет	10	he will be
6	будет жить	8	will live
6	будет не	8	will not

Table 2

Cluster List Output мочь ‘to be able’
2-word cluster: 968 / 1980 found in 5 files

Cluster		Frequency	Translation
1	не могу	189	I can’t
2	не могла	142	I couldn’t
3	не мог	83	couldn’t
4	не могли	61	they couldn’t
5	я могу	44	I can
6	я могла	25	I could [f]
7	я мог	20	I could [m]
8	мог бы	16	could have
9	он мог	15	he could

Table 3

Cluster List Output человек ‘human’
2-word cluster: 1,318 / 2150 found in 5 files

Cluster		Frequency	Translation
1	человек не	21	person not [verb]
2	человек и	18	person and
2	человека в	18	person
4	одного человека	16	of one person
5	несколько человек	14	several people
5	один человек	14	one person
5	человек с	14	person with

As can be seen from the above cluster frequencies of the verb to be, to be able to, and the word person reveal that the negative aspect of these words prevails statistically. This established, we can explore what it is that Svetlana Alekseyevich’s works are telling us people can’t be, or aren’t able to do? A search of the phrase ‘человек не может’ (a person can’t) shows us the following:

Table 4

Collocation Result Output: 2016–04–06 10:20:19
2,742 with 2,471 items in 5 files

	Context Word	Translation
1	с ружьем	with a rifle, firearm
2	ни одного	not one
3	на войне	at war
3	не может	can’t
5	что-то	something

5	я не	I not
7	а я	and I
8	может быть	maybe
8	никогда не	never not
8	о настоящем	about the present
8	у которого	who has
8	у нас 5	our
13	а не	and not
13	воевал в	fought with
13	еще несколько	several more
13	который воевал	who fought
13	над другим	above others
18	был не	was not
18	в одном	in one
18	в себе	in oneself

Thus, using such a corpus based approach, we have identified that Svetlana Alekseyevich's works have a tendency towards use of negative language that shows us not what a person can, but what he or she can't and perhaps, shouldn't, primarily in relationship with firearms. While this may seem trivial, it demonstrates the power of a corpus based statistical approach to quickly reveal and conveniently present us with tendencies spread through a large sample of works.

References

1 *Biber D.* Corpus Linguistics and the Study of Literature: Back to the Future? // Scientific Study of Literature. — 2011. — Vol. 1. — No. 1. — P. 15–23.

2 *Koteyko N.* Language and Politics in Post-Soviet Russia: A Corpus-Assisted Approach. — Publ. 1. — Hondsmills, Palgrave Macmillan. — 2014. — 191 p.

3 *Ryazanova-Clarke L.* The Russian Language Outside the Nation: Speakers and Identities. Macmillan. — 2014. — 224 p.

4 *Almela A., Keshabyan I.* Introduction // International Journal of English Studies. — 2012. — Vol. 12. — No. 2.

5 *Cantos-Gómez P., Sánchez A.* Recent and Applied Corpus-Based Studies // International Journal of English Studies. — 2009. — Vol. 9. — No. 3.

6 *Posokhin I.* The Literary Situation in Contemporary Belarus // Eastern Partnership Literary Review. — 2014. — Vol. 1. — No. 1. — P. 13–17.

7 *Krouglov A.* War and Peace: Ukrainian and Russian in Ukraine // Journal of Language and Politics. — 2002. — Vol. 1. — No. 2. — P. 221–239.

Д.Б. Славенской

Славян тіліндегі мәтіндерді корпустық талдаудың әдістемесі: Светлана Алексеевич шығармаларының шағын корпусы

Мақалада корпустық лингвистика әдісі аясында Светлана Алексеевичтің әдеби шығармалары талданған. Автор әдеби мәтіндерді талдауда бұл зерттеу әдісі басымдыққа ие және сирек кездеседі. Осы әдісті негізге ала отырып, автор жоғарыда аталған жазушының шығармаларын жан-жақты сараптаған. Зерттеу әдісінің тиімділігі мен болашағы бар екендігі дәлелденген.

Д.Б. Славенской

Методика корпусного анализа славяноязычных текстов: малый корпус произведений Светланы Алексеевич

В статье предпринят анализ литературных произведений Светланы Алексеевич в аспекте метода корпусной лингвистики. Автором отмечены редкость и в то же время преимущества данного исследовательского метода по отношению к литературным текстам. На основании данного метода автором был создан и проанализирован с точки зрения языковых особенностей малый корпус литературных произведений Светланы Алексеевич. В статье доказаны эффективность и перспективность подобного метода исследования.