н.п.	DNG	R	L	Ba	Вже давно думала про дорогу. Ввижалася їй нескінченним білим змієм, бо простягалась у визорі звідсіля, від їхнього передмістя, туди, в далекий безмежний світ, думати про який було страшно й лячно.

Таким чином, спостерігаються кореляції F із L і, відповідно, A / P із M / B.

При цьому S = Ba, S = Ab, Ba = Ab = S.

Відповідно до зазначеного, корпусне анотування концептів має такий вигляд: ***Наша присутність — це і є та дорога, яку я так часто останнім часом бачу {DNG/R/F/A}*** («Дім на горі», концепт «Шлях», переносне значення, активне переносне значення).

Таким чином, концептні смисли реалізують значення S (Ba, Ab) та FA, тобто символічні значення і активні переносні значення. Можливо, символічні значення, визначені як поєднання активних переносних і основних прямих значень, є прикметою ідеостилю Валерія Шевчука, який свідомо накладає абстрактні глибинні смисли на конкретні образи у своїх творах. Тому ефективність пропонованої схеми аналізу на матеріалі творів інших авторів потребує перевірки.

Література

1. Автоматизация анализа научного текста / В. А. Вербицкий, Т. А. Грязнухина, Н. П. Дарчук и др. – К.: Наукова думка, 1984. – 256 с.
2. Демська-Кульчицька О. Основи національного корпусу української мови. – К., 2005. – 219 с.
3. Іващенко В. Концепт-символ «кобзар» у життєтворчості Т. Г. Шевченка // Материалы IV Международного семинара «Шевченковский Петербург». – СПб., 2005. – 186 с.
4. Никитин М. Курс лингвистической семантики: Учебное пособие для студентов, аспирантов и преподавателей лингвистических дисциплин в школах, лицеях, колледжах и вузах. – СПб., 1996. – 760 с.

Ł. Grabowski*

Institute of East-Slavonic Studies University of Opole (Opole, Poland)

УДК 81’322

COMPARATIVE CHARACTERISTICS OF POLISH AND RUSSIAN LANGUAGE CORPORA

The aim of this paper is to present and compare the most representative language corpora of Polish and Russian according to selected criteria determining both their potential as a source of linguistic data for various types of linguistic analyses and their availability for researchers. Moreover, the present paper indicates areas for improvement as far as the possibilities offered by the corpora and access to them are concerned.

1. Introduction

Corpus linguistics, which is both a branch of linguistics and a methodology of research, concentrates on the study of texts with the use of dedicated computer software [4:143]. Thus, any collection of texts stored in electronic form is called a corpus. The linguistic data stored in the corpus is the actual collection of electronic words, which are classified in terms of types and tokens [2:16]. This typology is important in that a token is every running word-form (segment) the corpus is composed of, whereas a type is a group of the same tokens (eg. the sentence ‘*Моя мама живёт в городе, а моя тётя живёт в деревне’* consists of 8 types and 11 tokens).

Since language corpora, unlike single and coherent texts, are objects of complex and multidimensional character, the answer to the question how similar or different they are will indubitably be complex and multidimensional itself. Thus, the study aimed at comparing corpora that represent two different languages will be even more challenging and complex. Corpora as objects of comparison will always be similar in some aspects and different in others. For the comparison to render objective observations, one has to adopt a framework that would serve as criteria for such a study. Although there have been multiple quantitative methods based on word frequencies and ngram frequencies [1], purely statistical approach to a comparison of different multilingual corpora falls beyond the scope of the current paper. Thus, its aim is to present possibilities offered by the most representative commercial corpora representing Polish and

Russian languages as far as the following criteria are concerned: a) the overall number of tokens collected in corpora; b) availability of meta-linguistic annotation and lemmatisation of a corpus data; c) types of queries offered by corpora; d) availability of corpora and access to them. The Polish corpora are represented by Korpus Języka Polskiego Wydawnictwa Naukowego PWN (henceforth PWN Corpus), Korpus Instytutu Podstaw Informatyki Polskiej Akademii Nauk (henceforth IPI PAN Corpus) and the PELCRA Corpus. The Russian ones are the following: the Russian National Corpus (henceforth the RNC), Comparable Corpus of English and Russian News Texts (henceforth CCERNT Corpus) and Computational Corpus of Russian Newspaper Texts at the End of the Twentieth Century (henceforth CCRNTE20 Corpus). Such a selection is not accidental since the aforementioned linguistic resources are the most representative corpora compiled for Polish and Russian, respectively.

2. Characteristics of Selected Polish Corpora

PWN Corpus is a synchronic and balanced corpus compiled by the commercial institution Wydawnictwo Naukowe PWN S. A., one of the largest publishing houses operating on the Polish market. The corpus in question is available on the CD-ROM, on the Internet [7] and at the headquarters of the compilers. The access differentiation has a bearing on the corpus sample which is available to researchers. The overall corpus available at the PWN S. A. headquarters consists of 100 million tokens, out of which 70 million tokens account for the balanced corpus, which includes modern Polish literature (41 %), press articles (45.5 %), dialogues, leaflets and manuals, Internet websites (13.5 %); the remaining part includes Polish literature and press archives. The opportunities provided to the user of the Internet are limited, however. When accessing the corpus on-line, one can choose either a web sample of the corpus (*wersja sieciowa*) or the demonstrative version (*wersja demonstracyjna*). The difference between the two is crucial since the web sample provides access to the corpus of 40 million tokens (out of which 22 million constitutes the balanced corpus), whereas the demonstrative version sports a corpus sample of 7.5 million whereby only 3.5 million of tokens is a balanced collection. Although both samples are available through a concordancer placed on the above website, the difference between the two is that the very access to the web sample is chargeable, whereas the demonstrative version is free-available. Moreover, the queries which may be executed in the demonstrative version disable to regulate the width of the left and right-hand context of the key-words subject to search. As for the annotation of the corpus data, it contains only meta-situational and meta-textual information in a format that conforms to the Text Encoding Initiative (TEI) guidelines. Although such annotation renders part-of-speech queries unavailable, the undeniable advantage of the concordancer is its ability to display all inflected forms of the key-word subject to search. Finally, the corpus sample available on the CD-ROM, which was distributed to all Institutes of Polish Studies at Polish universities free of charge is the same corpus sample as a demonstrative one available on the Internet.

The next corpus subject to this study is IPI PAN Corpus, which was developed at the Institute of Computer Science of the Polish Academy of Sciences. The full version of the corpus comprises almost 300 million segments available through the Poliqarp concordancer, which serves as an integral tool for browsing the corpus on-line or it can be downloaded as a compressed tar file. As for the subcorpora available free of charge from the http://korpus.pl website, three of them require further presentation. The source version (*próbka źródłowa*) of the IPI PAN Corpus comprises 100 million tokens which renders 286, 000 types. The preliminary sample (*próbka wstępna*) comprises 70 million segments which renders 364, 000 types. The sample of the corpus available on-line is composed of 15 million tokens which renders 217, 000 types and which further accounts for the opportunistic corpus whose 90 % fall into the category of modern Polish literature and socio-political journalism. All the above samples are downloadable from the corpus website in a form of tar archive files, which can be decompressed with the use of 7Zip freeware application. As far as the annotation is concerned, the corpus data are enriched with morphosyntactic tags whose common format is the slightly modified XML version of Corpus Encoding Initiative. As for the meta-textual and meta-situational information, it is still incomplete and subject to major improvements. The Poliqarp search engine and concordancer (available in three versions: on-line, graphical and GNU/Linux, where the latter one is the most advanced one) is equipped with a user-friendly interface, which enables researchers to request multifarious types of queries, which may contain standard regular expressions: the base-form query, grammatical class query (based on grammatical tags specifying the values of the part of speech), grammatical category query, query with

constraining matches to sentences or paragraphs as well as query with constraining matches to meta-situational or meta-textual information [5; 6].

The last Polish corpus subject to this description is the PELCRA Corpus, which is developed at the Institute of English at the University of Lodz in co-operation with the University of Lancaster. The Reference Corpus of Polish, as the major subcorpus within the framework of PELCRA project, comprises 93, 129, 588 tokens. Methodology used in the compilation of this corpus was very similar to the one adopted for the British National Corpus. It is a synchronic corpus of modern written and spoken Polish whereby the latter one comprises only 600, 000 tokens, which is the only major difference from the BNC. The target figure for spoken subcorpus, however, is 1 million tokens. As for the annotation of the corpus in question, its source form is the binary XML-annotated corpus text and its target form is a collection of data stored in the relational database MySQL. As a result, the meta-textual, meta-situational and meta-linguistic annotation that the corpus is equipped with is of hierarchical character and the corpus data are stored in the tables and columns of the MySQL database [3: 107]. Such a solution renders possible the extensive use of SQL (Simple Query Language) whose benefit is the opportunity to carry out multiple queries using the search tool available on the corpus website. The most important types of queries, which may again contain standard regular expressions, are the following: the base-form query, inflection query and phrase query, collocation query and the MI3 collocation query. The queries may be constrained by the meta-situational and meta-textual information, which enable researchers to specify the features of corpus samples subject to search. Moreover, the search tool offers a wide variety of statistical analyses concerning, in particular, the frequencies of words and collocations. However, the availability of the corpus for rank-and-file user is limited to executing a small number of queries and full access to the corpus data is chargeable. It has to be emphasized, however, that since the Reference Corpus of Polish described here is available on the World Wide Web, its interface is user-friendly and easy to manipulate.

3. Characteristics of Selected Russian Corpora

The description of commercially available corpora of Russian starts with the Russian National Corpus (*Национальный корпус русского языка*) which has been compiled at the V. Vinogradov Institute of Russian of Russian Academy of Sciences in Moscow. This corpus, which is available through the website http://ruscorpora.ru, includes 120 million tokens (recorded as of February 7, 2006) which make up the meta-situationally, meta-textually and meta-linguistically annotated representative collection of Russian texts in the electronic form. A particular emphasis shall be put on the meta-linguistic annotation, which is very detailed and comprehensive in that it comprises grammatical and semantic tagging. The former covers information concerning part of speech of the tokens, their case, gender, degrees of comparison, number, tense, aspect etc., whereas the semantic annotation is even more extended and covers semantic categories, taxonomy and axiology. As a result, the concordance search tool available on the website allows one to execute two types of queries: the 'exact word-forms query' (*Поиск точных форм*) and the 'lexico-grammatical query'(*Лексико-грамматический поиск*). The latter one enables researchers to constrain the queries to specified grammatical properties (which facilitates detailed searches by determining specific morphological properties of key-words subject to search) and semantic properties (by determining semantic properties of key-words) which may contain standard regular expressions. Moreover, the search tool offers an additional option which enables one to constrain the search to a specific genre or linguistic environments a given text functions in. This corpus has been equipped with the option of 'reduced homonymy' (*снятие омонимии*) or 'non-reduced homonymy' which allows one to display either the lemmatised or non-lemmatised results of the query. The former one reduces tokens (lemmatas) to specific types (lemmas) representing the same part of speech which enables to make a distinction between inflectional forms of the same lemma. The latter one allows inaccuracy in the case of multifunctional words, such as a Russian lemma *печь*, which may function either as a verb or a noun. Moreover, the sample of the corpus with reduced homonymy allows for a display of the concordances with marked accents, which is essential for many language teachers browsing the corpus for linguistic material to be used during their classes. Apart from that, one has access to the subcorpus of spoken Russian and to the parallel Russian-English and English-Russian subcorpora, invaluable collections of texts for translators and lexicographers comprising aligned text-units (sentences and paragraphs). These separate corpora with non-reduced homonymy are equipped with concordance search tool, which allows researchers to execute lexico-grammatical search therein.

The Comparable Corpus of English and Russian News Texts was compiled at the University of Leeds under the supervision of Sergei Sharoff. It consists of multiple subcorpora which are available on the website. Russian part consists of a morphologically-tagged (the grammatical properties include: part-of-speech, case, gender, animation, number, person, degree of comparison, aspect and tense) and lemmatised collection of articles from Izvestia daily newspaper (issued between 2000-2001), which add up to 14 million tokens. The search tools allow one to constrain the search to grammatical properties of the word-forms and to display rich statistical data on frequency and distribution of the key-words subject to search. The interactivity of the interface enables corpus-oriented researchers to execute searches in the Russian National Corpus (in addition to the Izvestia Corpus) and in the small corpus of modern Russian fiction (500, 000 tokens), of which the latter one is also compiled in Leeds. The search results are displayed as concordances (whose width can be extended) enriched with meta-textual and meta-situational information after double-clicking on the selected concordance. For the users who log in, having received the access code to the corpus, it is available free of charge.

The last Russian corpus subject to this description is Computational Corpus of Russian Newspaper Texts at the End of the Twentieth Century (*Компьютерный корпус текстов русских газет конца XX-ого века*). This synchronic and opportunistic corpus, which can be accessed through the http://www.philol.msu.ru/~lex/corpus website, is available either on-line (205, 000 tokens in 446 press articles) or it can be researched at the Institute of General and Computational Lexicography and Lexicology at Moscow State University, where it was compiled. The full version of the corpus comprises 11, 401, 479 tokens in 23, 110 press articles. Although the size of the corpus is limited, it contains meta-linguistic (morphosyntactic), meta-situational and meta-textual annotation. This collection of texts is also lemmatized and comprises samples of full issues of thirteen Russian newspapers dating from 1994 -- 1997. The search tool offers users two types of queries which may be constrained to the selected properties of annotation: the 'exact word-form query' (*Буквальное совпадение*) and the 'indirect query' (*Подстрока*), the latter allowing researchers to search for a sequence of characters which may either account for the full key-word or may be its integral part. The results of the queries are displayed as either concordances or lists of word-forms.

4. Conclusions

Having been familiarised with general characteristics of the Polish and Russian corpora and having laid down four criteria for above description, it is possible to arrive at the following observations.

As far as the overall number of tokens is concerned, three Polish corpora oscillate around the figure of 100 million, which may be the result of treating the British National Corpus as a reference point (the BNC has 100, 106, 008 tokens). Although IPI PAN Corpus is outstanding in that it includes 300 million tokens, its actual availability starts with a source sample which includes 100 million tokens. Moreover, the sheer number is not conclusive in that all Polish corpora have not been fully completed so far. What is crucial, however, are the issues concerning access to them for researchers working outside Poland and the question: which corpus to use if one wants to avoid charges and to access it on-line? In this respect IPI PAN Corpus comes to the fore as its demonstrative version available on-line includes 15 million tokens. Although PELCRA Corpus offers over 93 million tokens, the search is in practice limited to one query at a time. Another advantage of IPI PAN Corpus is that one can download all three samples (cf. the source, preliminary and on-line versions) together with Poliqarp search tool designed specifically to access the corpus. As a result, after decompressing the samples stored in tar files, one may execute searches directly on the hard drive. As for the Russian corpora, the biggest one is obviously the RNC, which comprises 120 million tokens available on-line and free of charge, which makes it a promising corpus for the size of data and access. If one conducts quantitative research, the Comparable Corpus of English and Russian News Texts compiled at Leeds makes up a good reference point as it provides extended statistical information about the key-words subject to search. The last Russian corpus, the Computational Corpus of Russian Newspaper Texts at the End of the Twentieth Century, is advantageous in that it is a specialist collection of texts but its overall size, which is relatively small (over 11 million tokens and only 200, 500 tokens available on-line), and opportunistic nature make it custom-designed for corpus-oriented researchers studying the language of modern Russian press.

As for the meta-linguistic annotation, the IPI PAN Corpus and PELCRA Corpus are the best annotated Polish corpora because they are equipped with all three types of annotation (meta-

situational, meta-textual and meta-linguistic) which extends the scope of executable queries. It is perfectly visible on the example of PELCRA Corpus, whose user-friendly interface in Polish and English allows one to execute multiple types of queries (as referred to above), which proves that designers and compilers of this corpus aptly took advantage of its rich annotation. Among Russian corpora, the RNC is the most promising one as its detailed meta-linguistic annotation includes morphological, semantic and axiological features, which widen the scope of executable queries. As a result, one may constrain any lexico-grammatical query to all grammatical and semantic properties allowing narrow and specified search to be executed. The two remaining corpora of Russian also use three major types of annotation, but the meta-linguistic one is limited to morphological properties.

Finally, access to both Polish and Russian corpora renders an interesting observation in that nearly all of them rely on the on-line access which is equivalent to browsing the corpus by means of a search engine, the fashion similar to looking for information on the Internet using such search engines as Google, Altavista, etc. In other words, the idea of developing a unique dedicated corpus client as a separate software application designed to be installed on any computer has not found much appeal for the analysed corpora. The exception to this rule is IPI PAN Corpus of Polish, which allows one to download Poliqarp, a dedicated search engine and concordancer to access and browse the corpus. This solution reflects the one adopted by the compilers of the British National Corpus, which can be accessed either on-line or from the hard drive by means of SARA client, a search engine and concordancer offering, among others, multiple options and queries to be executed (even building up a single query from multiple types of queries), printing the results straight away, manipulating the data and their format, etc.

The only corpus available on the CD-ROM is PWN Corpus of Polish. This is another interesting observation since the process of distributing corpora on CD-ROMs is both profitable for their compilers and convenient for armchair researchers, because it enables them to enjoy fast and efficient access to electronic collections of linguistic data, even if they do not have access to the Internet.

Summing up, the present comparison of selected Polish and Russian corpora shows that corpora vary and even using specific criteria does not guarantee that one arrives at objective and comprehensive results. Nevertheless, having been familiarised with above collections of texts, researchers interested in extensive linguistic analyses of Polish and Russian using corpus methodology are indubitably in a better position to choose the corpus which would meet their requirements and would correspond with goals of their research.

References

1. Kilgarriff, A. (2001). “Comparing Corpora”. Retrieved on 10 Jan. 2006 from: http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/publications.html
2. Mason, O. (2000). Programming for Corpus Linguistics. Edinburgh: University Press.
3. Pęzik, P., Uzar, R., Levin, E. (2005). Zastosowania baz danych w językoznawstwie. In: B. Lewandowska-Tomaszczyk, Podstawy językoznawstwa korpusowego (p. 95-115). Łódź: Wydawnictwo Uniwersytetu Łódzkiego.
4. Piotrowski, T. (2003). Językoznawstwo korpusowe: wprowadzenie do problematyki. In S. Gajda (Ed.), Językoznawstwo w Polsce. Stan i perspektywy (p.143-154). Opole: Wydawnictwo Uniwersytetu Opolskiego.
5. Przepiórkowski, A. (2004). “The IPI PAN Corpus in Numbers”. Retrieved on 11 Feb. 2006 from: http://nlp.ipipan.waw.pl/~adamp/Papers/2005-ltc-numbers/
6. Przepiórkowski, A. (2005). “The Potential of IPI PAN Corpus”. Retrieved on 11 Feb. 2006 from: http://nlp.ipipan.waw.pl/~adamp/Papers/2005-psicl-numbers/
7. Korpus Języka Polskiego Wydawnictwa Naukowego PWN. Retrieved on 12 Feb. 2006 from: http://www.korpus.pwn.pl
8. Korpus Języka Polskiego IPI PAN. Retrieved on 12 Feb. 2006 from: http://korpus.pl
9. The PELCRA Reference Corpus of Polish. Retrieved on 12 Feb. 2006 from: http://korpus.ia.uni.lodz.pl
10. Национальный корпус русского языка. Retrieved on Feb. 14 2006 from: http://ruscorpora.ru
11. Компьютерный корпус текстов русских газет конца XX-ого века. Retrieved on 14 Feb. 2006 from: http://www.philol.msu.ru/~lex/corpus
12. The Comparable Corpus of English and Russian News Texts. Retrieved on Feb. 14 2006 from: http://corpus.leeds.ac.uk