Anna Caldrová, Brno
413405@mail.muni.cz

The Use of the Czech National Corpus in the Analysis of Russicisms in Czech

Das Tschechische Nationalcorpus zeigt die unterschiedlichen Formen der tschechischen Sprache und bietet eine Reihe nützlicher Tools um linguistische Phänomene, Textarten oder Wortbeziehungen zu suchen, zu vergleichen und zu analysieren. Im vorliegenden Beitrag wird das Korpus anhand einer Frequenzanalyse von Russizismen im Tschechischen im Zeitraum von 1998 bis 2021 vorgestellt.

The Czech National Corpus maps the various forms of the Czech language and offers useful tools for searching, comparing and analysing linguistic phenomena, text types or relationships between words. In this paper, we will introduce this unique tool and demonstrate the possibilities of its use through a frequency analysis of the occurrence of russicisms in Czech in the period from 1998 to 2021.

Electronic language corpora have become an indispensable tool in linguistic research. These large collections of authentic written and spoken texts make it possible to study linguistic phenomena in ways that would have been unthinkable in the past. Without corpora, we could hardly work effectively with the frequency of millions of words and phrases in use and determine whether they are at the centre or the periphery of the vocabulary. Finding and analysing the natural context of a word, its variations over time and across genres, would be limited by the lengthy and necessarily subjective collection of material without corpus data.

Linguistic corpora are therefore an invaluable tool, especially when studying very specific linguistic phenomena, such as russicisms, i.e., words that have been borrowed from Russian into Czech. And it is these that we will focus on in this paper.

Before presenting the results of the frequency analysis of russicisms, we would like to devote a few lines to the tool without which we would not have been able to carry out this research – the Czech National Corpus.

The Czech National Corpus (CNC) was founded in 1994 at the Faculty of Arts of Charles University in cooperation with other institutions and universities. It is continuously built and developed by the Institute of the Czech National Corpus in cooperation with the departments of Charles University, Masaryk University, the Academy of Sciences of the Czech Republic and others. The CNC project (korpus.cz) is freely available for non-commercial purposes and is therefore used not only by linguists, dictionary and textbook authors, but also by sociologists, translators, teachers, students and the general public. It contains a number of corpora that map and monitor various forms of the Czech language and offers many useful tools for searching, comparing and analysing various phenomena, types of texts or relationships between words. The CNC has been compiled with a view to research and is a non-selective data source that presents phenomena in their natural context. It provides us with information about common and preferred usage and allows us to discover new behaviours and meanings of lexemes that are not recorded in dictionaries.

Corpora are usually defined as structured, large, technically processed sets of linguistic data in electronic form. In addition, CNC corpora are annotated with bibliographical and sociolinguistic information, lemmatised and morphologically tagged. As a result, they can be searched for word forms and specific parameters such as year of publication, country of origin, text type, genre, or gender of author.

The CNC consists of several components covering different forms of language of different nature and scope for all kinds of (not only) linguistic research. They represent both written Czech (from a synchronic or diachronic perspective) and spoken Czech. The SYN (synchronic) series contains corpora of written Czech, produced regularly every five years and covering the period since 1989. These corpora mainly cover the language of the last five years before their publication and each of them contains 100 million text words. The corpora are mutually disjunctive, genre-balanced and consist of printed and publicly published texts in the following genres:

- fiction (including belles-lettres in the broadest sense: prose, poetry and drama);
- specialised literature (including scientific, popular, educational, professional and administrative texts);
- journalism (including traditional and leisure journalism, i.e., the national and regional daily press and thematic periodicals, as well as non-periodic journalistic texts).

They are complemented by the extensive SYN PUB synchronous corpora, which consist exclusively of written journalism. The individual versions of the SYN corpus are not representative, as their composition is dominated by easily accessible journalism. All corpora in the SYN series can be searched simultaneously, giving access to more than 5 billion words (cf. Cvrček / Richterová 2022d).

The ORAL (spoken) series consists of transcripts of authentic spoken Czech, including spontaneous dialogues between friends, in the family, etc., recorded over ten years (2002–2011) throughout the Czech Republic. In total, it contains about 5.3 million words, which corresponds to 582 hours of recordings (cf. Cvrček / Richterová 2022c). This series is followed by the ORTOFON corpus (103 hours of recordings from 2012–2017) and the DIALEKT corpus, which maps territorial dialects on the territory of the Czech Republic (data were collected in two layers, from the 1950s to the 1980s and from the 1990s to the present). The corpora are lemmatised and morphologically tagged, as are the written corpora (cf. Cvrček / Richterová 2022a).

The InterCorp parallel multilingual corpus contains texts in Czech and 41 foreign languages (in the InterCorp 15 version of 2022), aligned sentence by sentence. The corpus contains fictional texts, film subtitles, European Union legal texts and European Parliament minutes, as well as journalistic articles and Bible translations. It serves as a source of data for theoretical and lexicographical studies, as well as a tool for the study and teaching of foreign languages or translation (cf. Cvrček / Richterová 2022b).

Also worth mentioning are the NET and ONLINE web corpora, which contain texts available on the Internet and are suitable for studying discourse or the specifics of online communication. There are also a number of specialised corpora – these are devoted, for example, to Czech as a second language (CzeSL), specific texts and language areas (e.g., LINK, Totalita) or specific authors and works (e.g., ORWELL, CAPEK; cf. Cvrček / Richterová 2021a).

The SYN version 11 corpus is the most suitable tool for studying changes in frequency and shifts in meaning of russicisms, as it serves as a very clear and up-to-date source of texts and information. The source texts within the created sub-corpus were limited by the date of the first edition (1998–2021). The main reason for this limitation is that the texts from the given period are well represented in the CNC, thus reducing the risk of bias in the frequency analysis. To compare the frequency of russicisms in different periods, we then used the so-called relative frequency expressed in instances per million (i.p.m.), which expresses the average number of occurrences of a unit or word in a hypothetical text of 1 million words (cf. Cvrček / Richterová 2021b).

In Czech explanatory dictionaries and numerous academic papers describing various periods of our history, russicisms are well documented – we can find over 1000 russicisms,[1] i.e., lexemes marked as words of Russian origin or used in the Russian environment or adopted through Russian.

Among them are words that are well known to us today and that we would hardly recognise as words of foreign origin if we did not know the historical development of the Czech language. These words entered the vocabulary of the Czech language mainly at the beginning of the 19th century, during the period of national revival. A considerable number of poetic expressions, as well as contemporary specialist terminology – botanical, zoological and military – date from this period.

A separate category is formed by russicisms that appeared in Czech during the period of the previous regime after 1945, the so-called sovietisms. These include words from the industrial, economic, administrative and socio-political spheres, where we can often observe shifts in meaning to the satirical or even pejorative level. We should also mention Russian realia in the field of art and culture, way of life or social organisation. Today we find them not only in dictionaries, but also in the works of contemporary authors and in journalism.

With the help of the corpus, we are able to identify the most frequent russicisms in contemporary Czech. As the corpus data shows, the most frequently used russicisms are those that have become a common part of our language and are no longer perceived as words of foreign origin. However, these lexemes are not numerous and make up only a few percent of the total number of russicisms. We will mention at least some of them. Below we list the entry, its translation, relative frequency (i.p.m.),[2] and a source identifying the word as a russicism.[3]

---

1 Data on the total number of russicisms in Czech vary considerably. For example, Lilich (1982: 74) lists 1000 russicisms only in the works and dictionaries of Jungmann, while Bláha (2016: 89) includes in his glossary of russicisms a total of 917 lexemes from dozens of sources, including the works of Lilich. In the course of our research, we have so far identified about 1300 lexical russicisms, which are analysed in more detail in our other studies.

2 It is important to note that some lexemes already existing in Czech have expanded their meaning under the influence of Russian. An example is the lexeme *hovořit*, which in its main definition means *to talk to sb about sth*. However, under the influence of Russian it has expanded its meaning to *report sth* (according to *говорит Москва*, i.e., *Moscow (radio) reports*. However, for lexemes with a relatively high frequency, such as the lexeme *hovořit* it is not possible to filter the individual results and determine the frequency for only one of the given meanings. It is therefore necessary to take this frequency with a certain amount of caution. We therefore mark such cases with an asterisk after the frequency data.

3 Some russicisms are found in more than one source. We list mainly large dictionaries for clarity.

›řešit‹, *verb*: fulfil *sth*; 19,97* i.p.m.; Mejstřík 1965: 116
›postupně‹, *adverb*: gradually, progressively; 18,84 i.p.m.; Lilič 1982: 44
›vnímat‹, *verb*: sense, feel *sth as sth*; 16,44 i.p.m.; Trávníček 1952: 1660
›hovořit‹, *verb*: report *sth* (*on the radio*); 16,07* i.p.m.; Havránek 1971: 639
›příroda‹, -y, f.: nature; 16,06 i.p.m.; Trávníček 1952: 1284
›vzduch‹, -u, m.: air; 15,15 i.p.m.; Machek 1997: 706
›průmysl‹, -u, m.: industry; 12,59 i.p.m.; Machek 1997: 488
›soulad‹, -u, m.: harmony; 10,31 i.p.m.; Šmilauer 1941: 66
›čaj‹, -e, m.: tea; 8,81 i.p.m.; Havránek 1971: 235
›lyže‹, -e, f.: ski(s); 7,84 i.p.m.; Machek 1997: 345
›soustava‹, -y, f.: (e.g., decimal, solar, central nervous) system; 7,65; Šmilauer 1941: 66
›duševní‹, *adj.*: mental, psychological; 7,41 i.p.m.; Lilič 1982: 46
›paluba‹, -y, f.: (*boat*) deck; 6,85 i.p.m.; Havránek 1971: 493
›vkus‹, -u, m.: taste *in sth*; 6,13 i.p.m.; Machek 1997: 693
›nudný‹, *adj.*: boring; 5,17 i.p.m.; Havránek 1936: 91

However, more than half of the russicisms are not used at all or very rarely in contemporary Czech. These are mostly russicisms that are recorded in dictionaries but are no longer relevant to speakers, have been replaced by new terms, or refer to phenomena that are not very widespread. A large number of russicisms borrowed at the time of the national revival or by the Czechoslovak legions during the First World War are now classified as archaisms. And the aforementioned sovietisms from the (not so distant) period of the previous regime were banished to the fringes of the Czech vocabulary immediately after the November 1989 revolution. Let's take a few examples of russicisms in the field of realia, poeticisms, terminology and slang that are not very well known among speakers, hardly used anymore or even completely forgotten.

›moloděc‹, -dce, m.: brave young man (*colloq. expr.*); 0,05 i.p.m.; Havránek 1971: 1270
›bárišňa‹, -ni, f.: (in the Russian pre-revolutionary environment) young lady (*colloq. expr.*); 0,01 i.p.m.; Havránek 1971: 83
›soljanka‹, -y, f.: thick sour fish or meat soup; 0,02 i.p.m.; Havránek 1971: 437
›gorodky‹, -ů, m.: a game of throwing sticks to knock over small logs; 0,02 i.p.m.; Havránek 1971: 547
›chandra‹, -y, f.: wistfulness, ennui (*archaic*); 0,17 i.p.m.; Havránek 1971: 682
›jeseň‹, -ně, f.: autumn (*archaic*); 0,1 i.p.m. Lilič 1982: 46
›dušesloví‹, -í, n.: psychology (*archaic*); 0 i.p.m.; Havránek 1936: 91

›sovchoz‹, -u, m.: abbreviation of sovetskoye chozyaystvo, i.e., Soviet (agricultural) economy; 0,03 i.p.m. Havránek 1971: 455

›komandýrovka‹, -y, f.: travel order (*military slang*); 0 i.p.m. Mejstřík 1963: 265

›broněvik‹, -u, m.: name for an armoured car or train (*military slang*); 0 i.p.m.; Havránek 1971: 170

In addition to relative frequency, the CNK tools also allow us to follow the evolution of the frequency of individual words over time. While the established and well-known russicisms do not show any particular fluctuations in the period from 1998 to 2021, we can observe a downward tendency for some russicisms (especially from the Sovietism category). In quite exceptional cases, we find an upward tendency. We will illustrate these two cases with the lexemes ›bolševismus‹ (-u, m.: radical communist faction within the Russian Social Democratic Labour Party; 0,34 i.p.m.; Havránek 1971: 149) and ›boršč‹ (-e, m.: traditional soup made with beetroot and meat; 0,32 i.p.m.; Havránek 1971: 152). (Fig. 1 and Fig. 2)

The Czech National Corpus offers many useful tools for (not only) linguistic research of various phenomena in the Czech language – from comparing word variants and word phrases to searching for translation equivalents and word-forming relationships. Thanks to its ease of use, it can be used by linguists, teachers, students, translators, copywriters, or anyone interested in a deeper understanding of the language.

Relative frequency analysis, as we have presented in this paper, is then useful, for example, when translating Russian literature into Czech, or when creating popular educational or leisure content about the Russian environment. Thanks to the corpus data, we can easily find out which of the Russian lexical loanwords and Russian realia are likely to be familiar to the contemporary Czech reader and which, on the contrary, will have to be supplemented by descriptions, additions or even omitted from the text altogether.

It should be borne in mind that the experience of Russian and the Russian environment varies considerably between different generations of Czech readers and writers. In addition to journalists who grew up and lived in close contact with Russia during the socialist period, there is now a younger generation that can no longer do without explanations of the meaning of previously familiar Russian words. In view of the current situation and the considerable distancing from Russia and its culture, we can assume that the level of knowledge of the Russian environment and thus the frequency of russicisms will continue to decline and that the russicisms that are rarely used today will disappear from the active vocabulary for good in a few decades.

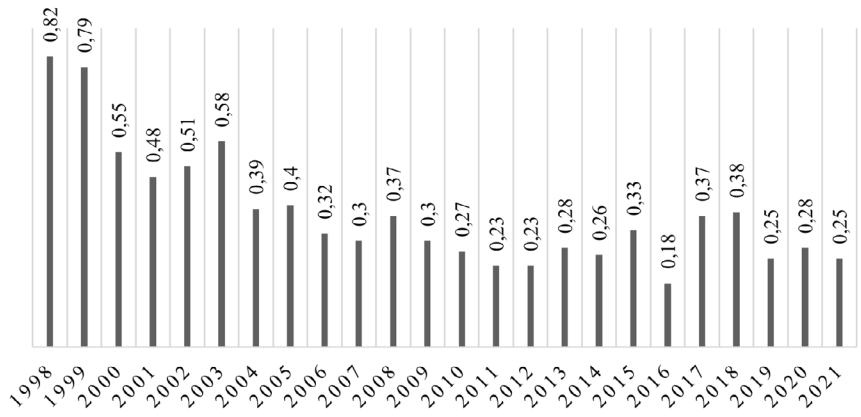

Fig. 1: ›Bolševismus‹ (downward tendency).

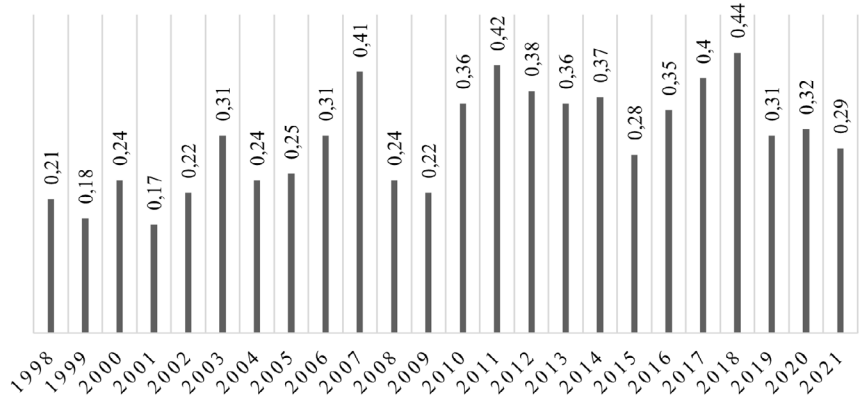

Fig. 2: ›Borščˇ (upward tendency).

References

Bláha, O. / Dudek, P. / Heklová, K. / Horáčková, K. (2016): *Lexikální rusismy v současné češtině*. Olomouc.

Cvrček, V. / Richterová, O. (2021a): »cnk:struktura«. In: *Příručka ČNK*, http://wiki.korpus.cz/doku.php?id=cnk:struktura&rev=1613589024 (accessed: 21 May 2023).

Cvrček, V. / Richterová, O. (2021b): »pojmy:ipm«. In: *Příručka ČNK*, http://wiki.korpus.cz/doku.php?id=pojmy:ipm&rev=1614020110 (accessed: 21 May 2023).

Cvrček, V. / Richterová, O. (2022a): »cnk:dialekt«. In: *Příručka ČNK*, http://wiki.korpus.cz/doku.php?id=cnk:dialekt&rev=1661786718 (accessed: 21 May 2023).

Cvrček, V. / Richterová, O. (2022b): »cnk:intercorp:verze15«. In: *Příručka ČNK*, http://wiki.korpus.cz/doku.php?id=cnk:intercorp:verze15&rev=1669153562 (accessed: 21 May 2023).

Cvrček, V. / Richterová, O. (2022c): »cnk:oral«. In: *Příručka ČNK*, http://wiki.korpus.cz/doku.php?id=cnk:oral&rev=1661786671 (accessed: 21 May 2023).

Cvrček, V. / Richterová, O. (2022d): »cnk:syn«. In: *Příručka ČNK,* http://wiki.korpus.cz/doku.php?id=cnk:syn&rev=1671624499 (accessed: 21 May 2023).

Havránek, B. (1936): »Vývoj spisovného jazyka českého«. In: *Československá vlastivěda* (2), pp. 1–144.

Havránek, B. (1971): *Slovník spisovného jazyka českého*. Praha.

Lilich, G. A. (1982): *Rol' russkogo yazyka v razvitii slovarnogo sostava cheshskogo literaturnogo yazyka*. Leningrad.

Machek, V. (1971): *Etymologický slovník jazyka českého*. Praha.

Mejstřík, V. (1963): »Z knih, časopisů a novin«. In: *Naše řeč* (5), pp. 263–265, http://nase-rec.ujc.cas.cz/archiv.php?art=4984 (accessed: 21 May 2023).

Mejstřík, V. (1965): »Z 26. sešitu ›Slovníku spisovného jazyka českého‹«. In: *Naše řeč* (2), pp. 114–116, http://nase-rec.ujc.cas.cz/archiv.php?art=5114 (accessed: 21 May 2023).

Šmilauer, V. (1941): »Ruské vlivy na češtinu«. In: *Naše řeč* (3), pp. 65–69, http://nase-rec.ujc.cas.cz/archiv.php?art=3571 (accessed: 21 May 2023).

Trávníček, F. (1920): »Příspěvek k mluvě naší sibiřské armády«. In: *Naše řeč* (6–7), pp. 203–209, http://nase-rec.ujc.cas.cz/archiv.php?art=662 (accessed: 21 May 2023).

Trávníček, F. (1952): *Slovník jazyka českého*. Praha.

Open Access