[**OPERA** SLAVICA **XXX / 2020 / 4** (35—47)]
https://doi.org/10.5817/OS2020-4-3

A Comparative Frequency Analysis of Russicisms

Anna Caldrová

(Brno, Czech Republic)

Abstract:

Every language is a living organism, it is constantly evolving, changing and most importantly adapting to the needs of its speakers. Old words are falling into disuse, new words are entering lexicon. There is a specific category of words which are adopted from other languages. This paper focuses on so-called russicisms and their usage in the Czech language during the last thirty years. I use a classification of lexemes excerpted from Václav Machek's *Etymological Dictionary of the Czech Language* to compare the frequency of each category's occurrence between 1989 and 2018, using the Czech National Corpus as source for texts and frequency data. The article reveals what types of russicisms we may encounter in Czech and what tendencies they show in analyzed period, which will get us a better idea how influence of Russian to Czech developed.

Key words:

russicisms; corpus; etymological dictionary; semantic statistics; frequency analysis; lemma

Introduction

Although Russia and the Czech Republic are geographically distant countries, Russian--Czech cultural and linguistic contacts cannot be considered negligible. Of course, in different periods of our history, these contacts have changed depending on the current needs of both nations. The Russian-Czech contacts gradually shifted from

noncontinuous meetings caused by necessary business and diplomatic contacts to intensive and entirely conscious contacts that were aimed to bring the Czech and Russian languages closer. Linguistic elements from Russian to Czech were adopted both consciously—due to enrichment of the language, especially in the field of technical literature, and randomly—by naturalization of russicisms that were used for example in translations or novels. The development of Russian-Czech linguistics contacts and influence of the Russian on the Czech is discussed and described in works of various authors, for example B. Havránek, G. A. Lilič, J. Vlček, V. Šmilauer, M. Giger or J. Filipec[1]. Obviously, the adopted linguistics elements, like any other, are changing according to the needs of the speakers, depending on economic, social and political changes. A large number of russicisms are not used in contemporary Czech, their meaning has shifted, or they are used only in a certain area of language (e. g. technical terms), while others have entered the generally-used language and we no longer consider them as foreign words at all.

Method

The aim of the following study is to compare the frequency of occurrence of russicisms from the Etymological dictionary of the Czech language in different time periods and create its semantic classification. In this case, the Czech National Corpus (Český národní korpus)[2] was used as a source of texts and frequency data, and the *Etymological dictionary of the Czech language* by Václav Machek (1997) as the source of analyzed lexemes. The Etymological dictionary is interesting mainly because it contains not only standard words but also vernacular words. The analysis was carried out using the 3rd edition of the dictionary from 1971[3] (photocopy reprint), which offered an

1 HAVRÁNEK, B.: *Vývoj spisovného jazyka českého.* Praha: Československá vlastivěda, řada II, 1936; LILIČ, G. A.: *Rol' russkogo jazyka v razvitii slovarnogo sostava češskogo literaturnogo jazyka (konec XVIII – načalo XIX veka).* 1982; VLČEK, J.: *Úskalí ruské slovní zásoby: slovník rusko-české homonymie a paronymie.* Praha: Svět sovětů, 1966; VLČEK, J.: *Porovnání slovní zásoby ruského jazyka se slovní zásobou českého jazyka.* Praha: Univerzita Karlova, 1986; ŠMILAUER, V.: *Ruské vlivy na češtinu.* Naše řeč, roč. 25, 1941; ŠMILAUER, V.: *Obohacování slovní zásoby: kurs pořádaný kruhem přátel českého jazyka v Praze.* Praha: Kruh přátel českého jazyka v Praze, 1953; GIGER, M., SUTTER-VOUTOVA, K.: *Transparency of morphological structures as a feature of language contact among closely related languages: Examples from Bulgarian and Czech contact with Russian.* Boston, De Gruyte: Besters-Dilger J., Pfänder S., Rabus A., Dermarkar C.: Congruence in contact-induced language change, 2014; FILIPEC, J., ČERMÁK, F.: Česká lexikologie. Praha: Academia, 1985.

2 The Czech National Corpus is an academic project founded at the Charles University's Faculty of Arts in cooperation with other institutions and universities in 1994. CNC corpora are used not only by linguists and experts from other fields but also by students and the general public.

3 First published in 1957 under the title Etymological Dictionary of Czech and Slovak language (Etymologický slovník jazyka českého a slovenského), articles about Slovak words were omitted in later editions.

interesting insight into the evolution of the use of the lexemes over time. The analysis of the data in the Czech National Corpus was limited to sources that belong to the non-translated Czech literature and were first published between 1989 and 2018. As a source of texts Corpus SYN version 8 which contains all the synchronic written corpora of the SYN series[4] was used for analysis. It is important to mention that the SYN corpus is not representative—it contains mainly journalistic texts due to their easy accessibility. However, as social and political changes are most clearly reflected in the language of journalism, this corpus seems to be a suitable source. Since SYN series corpora are using lemmatization, lemma, as a representative form of word, was used during analysis. To find relative frequency (in ipm—instances per million words) based on corpus size the built-in function was used. The "First hits in documents" filter was also used to obtain more accurate results. For the semantic classification purposes were analyzed lexemes divided into three main groups: realia[5], general language lexemes, archaisms and vernacular words, and afterwards to the following semantic categories according to their use in specific fields of human activity[6]:

1. ethnography (objects typical for everyday life, culture and work),
2. politics and society (political and social life, organs and functions)
3. natural science (geographic and geologic objects, names of plants and animals, body parts),
4. unclassified vocabulary.

Results

A total of 128 russicisms were extracted from the Etymological dictionary of the Czech language (the dictionary contains over 8000 etymological entries). What stands out in following tables is the ratio between the different groups of lexemes:

Table 1: Number of lexemes in the main groups

	General language	Realia	Archaisms and vernacular words
Total N	103	9	16
Total %	80,47 %	7,03 %	12,5 %

4 All synchronic written corpora of the SYN series are disjunctive, therefore corpus SYN version 8 contains 4.5 billion words in total.

5 Realia are language-specific lexemes without equivalents which reflect culture-specific facts in a certain culture.

6 Described classification is based on the Vlakhov's and Florin's classification of cultural realia.

Table 2: Number of lexemes in the semantic categories

	Ethnography	Politics and society	Natural science	Unclassified vocabulary
Total N	24	6	65	33
Total %	18,75 %	4,69 %	50,78 %	25,78 %

As can be seen, the classification results are unambiguous. Most lexemes from Etymological dictionary belong to the general language—this category includes lexemes such as *průmysl (industry)*, *maják (lighthouse)*, *vějíř (fan)*, *lyže (ski)*, *sopka (volcano)* etc. Only a small number of lexemes was categorized as realia, e. g. *azbuka (Cyrillic alphabet)*, *bohatýr (Russian heroic warrior)*, *láptě (bast shoes)*, or archaism/vernacular word, e. g. *nekošník (evil spirit)*, *čuma (plague)*, *hulati (make merry)*. Among semantic categories, lexemes that denote subjects from the natural science field, predominate. The names of plants and animals are most frequently represented—*baklažán (aubergine)*, *kambala (flounder)*, *klikva (cranberry)*, *lumík (lemming)*, *saranče (locust)* etc. The rest of analyzed lexemes belong to the category of unclassified vocabulary containing lexemes that could not be assigned to a group due to excessive diversity, e. g. *strohý (curt)*, *tlupa (troop)*, *nářečí (dialect)*, *sloh (style)*, *vesna (spring, literary)*, *žesť (metal sheet)*, and to the category of lexemes referring to ethnography such as *presto (sacrificial altar)*, *žertva (religious sacrifice)*, *žrec (priest in antient Slavic religion)*, *knuta (scourge)* etc. Only a few lexemes were categorized as related to politics and society: *bojar (Russian aristocrat)*, *bolševik (bolshevik)*, *car (tsar)*, *kulak (peasant in Russian empire)*.

The lexemes were also divided into groups according to their relative frequency:

Table 3: Number of lexemes in the semantic categories

Frequency (ipm)	≤ 0,09	0,1–0,99	1–4,99	5 ≤
Total N	67	35	20	6
Total %	52,34	27,34	15,63	4,69

As shown, analyzed russicisms are not high-frequency words. The category of low-frequency lexemes includes, inter alia, all archaisms that occurred in the corpus with zero (or very low) relative frequency. On the contrary subsequent table shows 15 most frequent lexemes, which—with one exception—belongs to general language:

The results of observing tendency in frequency were quite varied. Out of a total of 128 lexemes, a total of 22 lexemes, e. g. *drožka (hackney)*, *sumka (bullet case)*, *chrabrý*

Table 4: 15 most frequent russicisms

	Lexeme	Translation	Frequency (ipm)	Semantic group
1.	vzduch	air	14,66	Natural science
2.	průmysl	industry	12,62	Unclassified vocabulary
3.	lyže	ski	7,95	Ethnography
4.	paluba	deck	6,91	Ethnography
5.	vkus	taste *in sth*	6,17	Unclassified vocabulary
6.	nudit	bore *sb*	6,16	Unclassified vocabulary
7.	smršť	whirlwind	3,51	Natural science
8.	spět	be approaching *literary*	3,02	Unclassified vocabulary
9.	strohý	curt	2,79	Unclassified vocabulary
10.	maják	lighthouse	2,45	Ethnography
11.	sopka	volcano	2,32	Natural science
12.	kormidlo	helm	2,09	Ethnography
13.	sloh	style *(architectual (style)*	1,98	Unclassified vocabulary
14.	útes	cliff	1,87	Natural science
15.	bolševik	bolshevik	1,82	Politics and society—realia

(valiant), have shown a downward frequency trend and only 6 lexemes—*kustovnice (Lyceum Chinese)*, *ladný (graceful)*, *maják (lighthouse)*, *orobinec (Typha)*, *pyl (pollen)*, *rakytník (sea-buckthron)*—have shown an upward frequency trend. Other lexemes showed irregular changes or relative stability and it was not possible to determine the trend tendency in their usage. A total of 36 lexemes showed a zero frequency and therefore were not included in the analysis.

Table 5: Frequency trend tendency analysis

	Downward trend	Upward trend	Zero trend tendency	Zero frequency
Total N	22	6	64	36
Total %	17,19	4,69	50	28,13

Table 6: Downward trend lexeme examples (1st table)

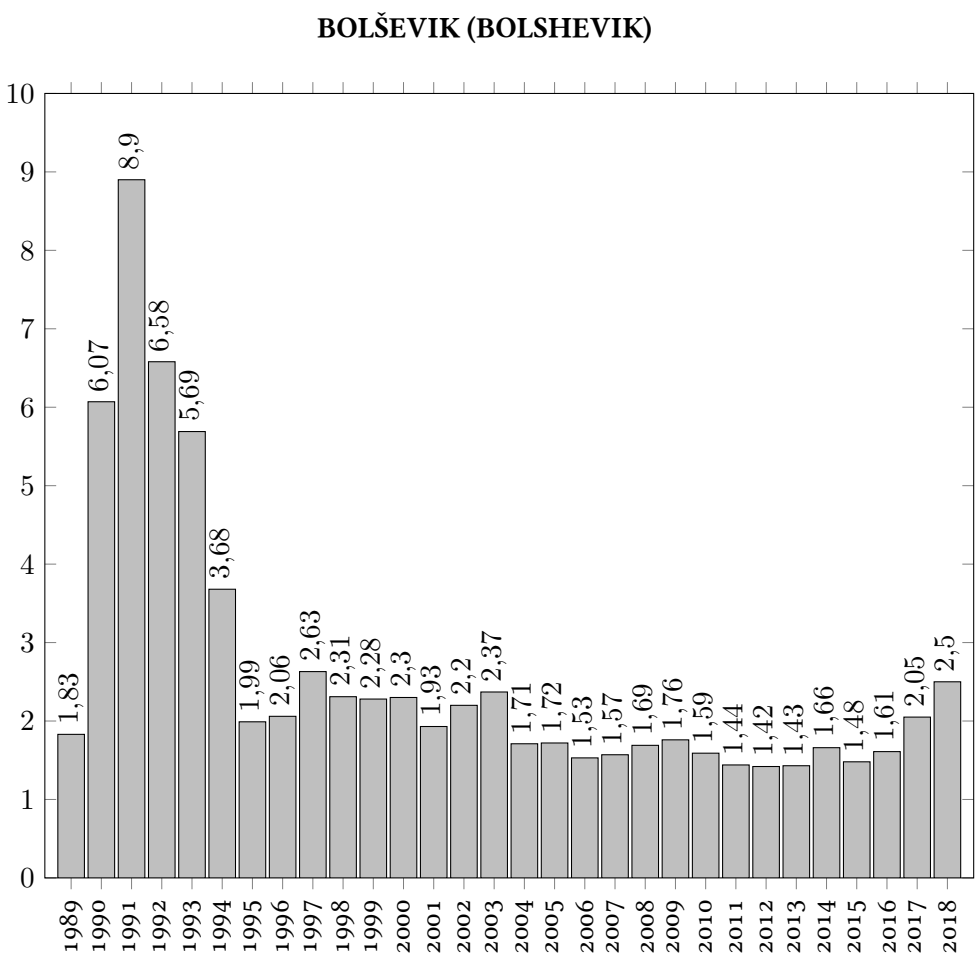


Table 6: Downward trend lexeme examples (2nd table)

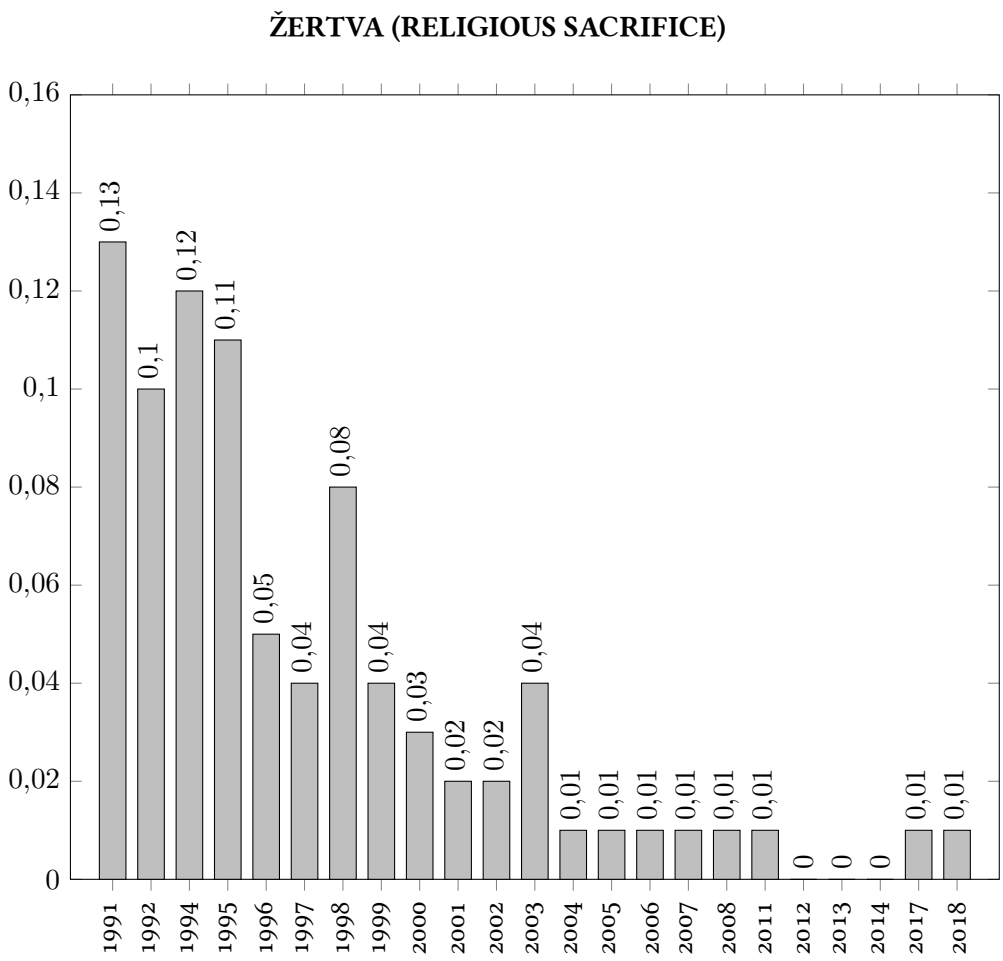


Table 7: Upward trend lexeme examples (1st table)

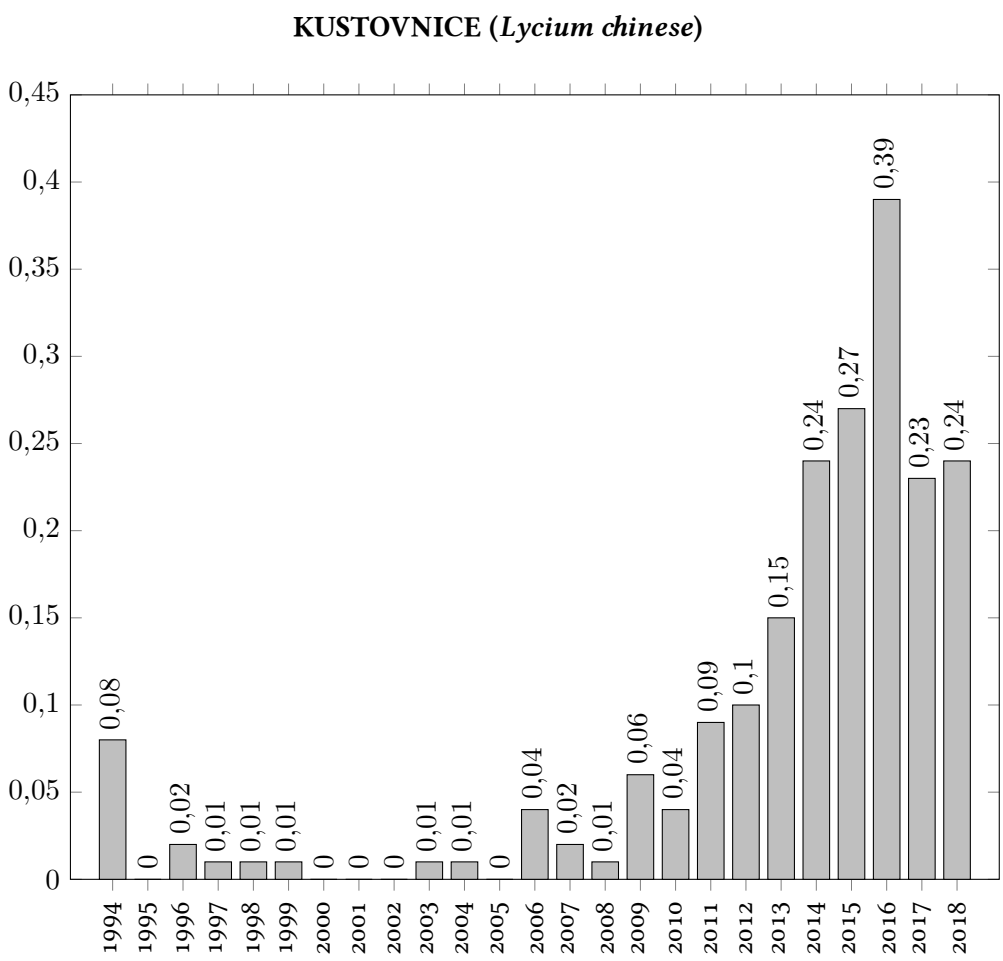


Table 7: Upward trend lexeme examples (2nd table)

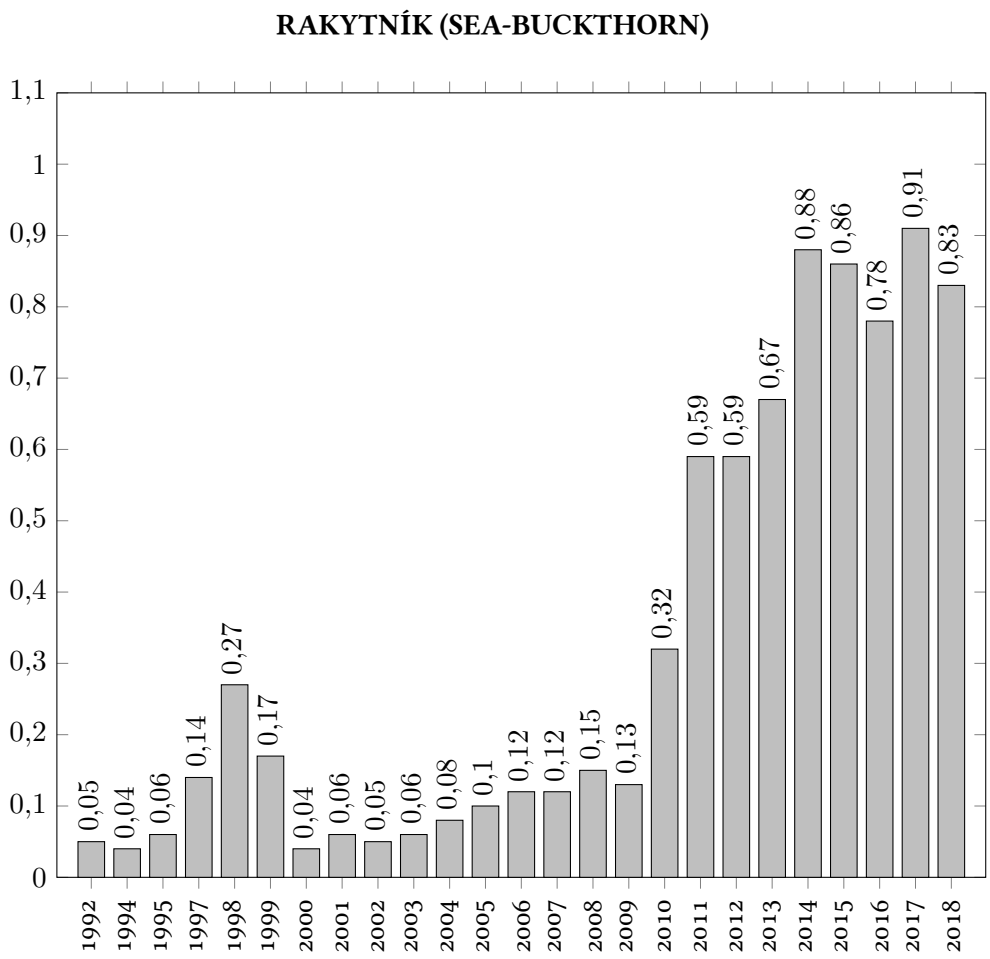


Table 8: Zero trend frequency lexemes example (1st table)

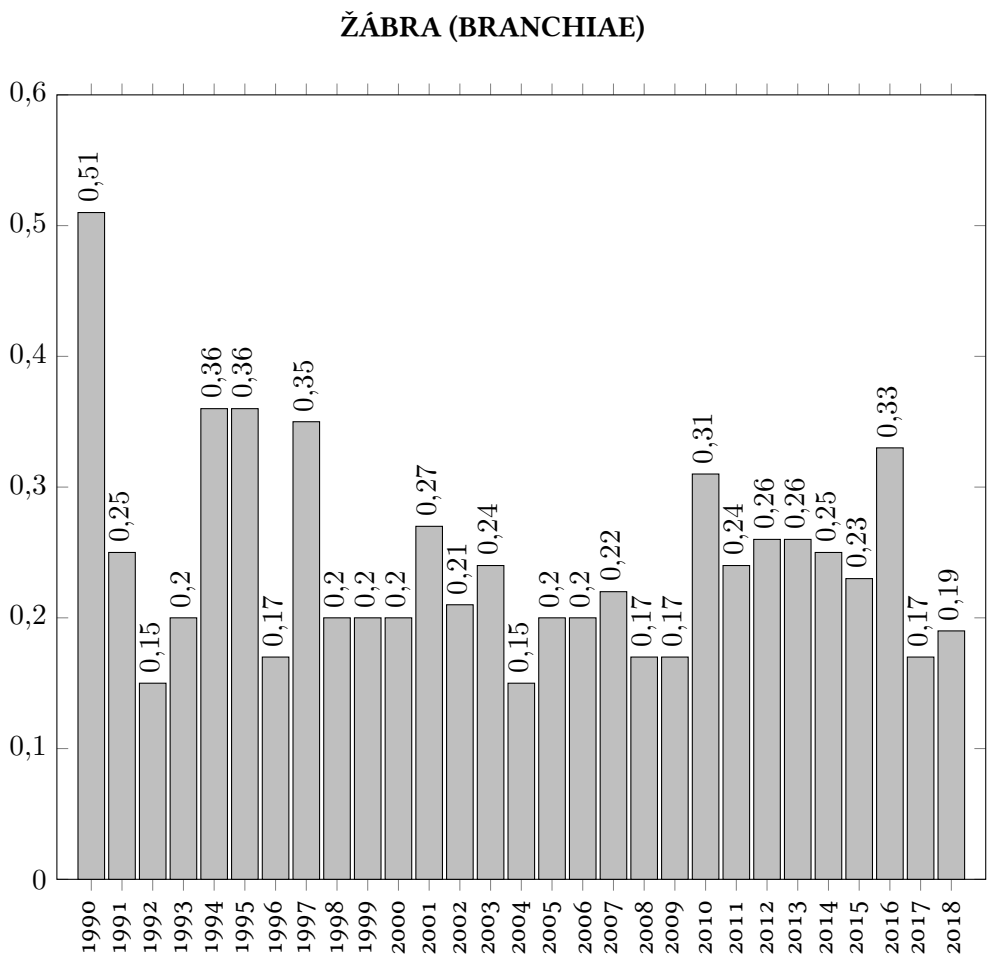
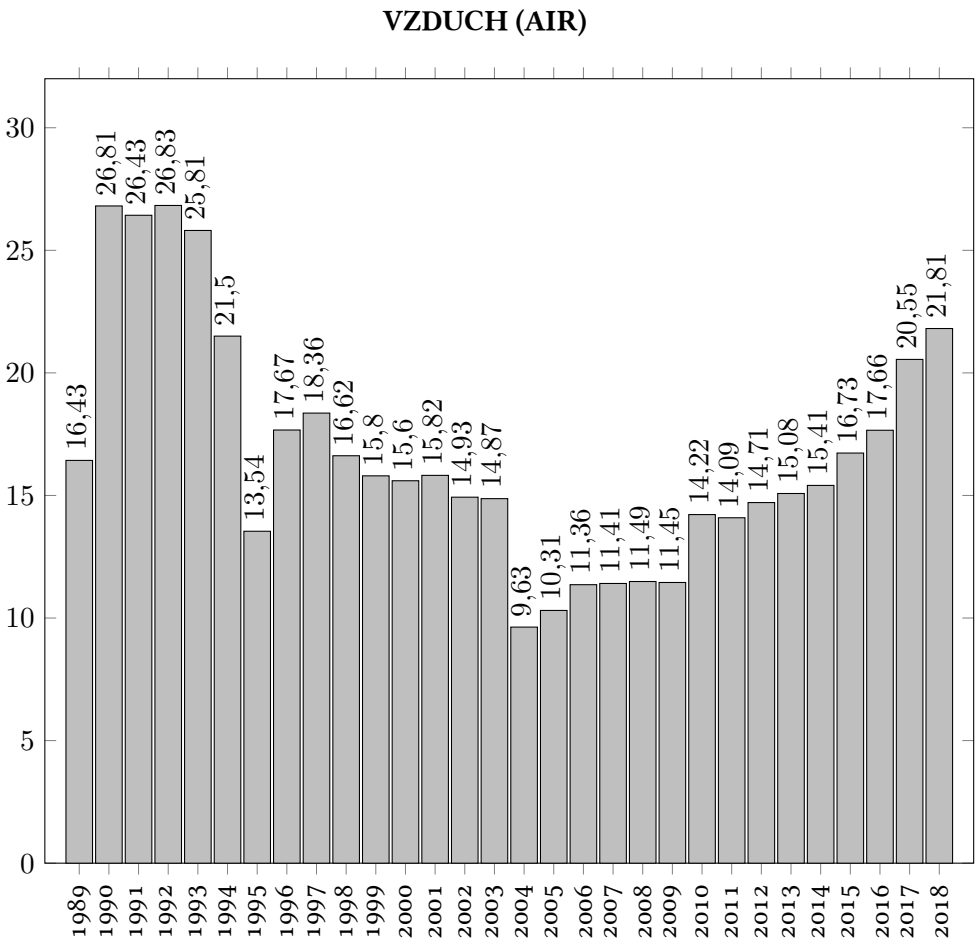


Table 8: Zero trend frequency lexemes example (2nd table)



Conclusion

The Czech language has borrowed many words from other languages and cultures. Russicisms (and the other loanwords) are so common, that most Czech speakers will not realize that they were borrowed from another language—especially if they were added to Czech a long time ago. The present study aimed at analyzing lexemes from Václav Machek's *Etymological Dictionary of the Czech Language* and comparing its frequency of occurrence during last thirty years. I studied the composition of words in the dictionary and created semantic classification of excerpted lexemes. Regarding the date of publication and composition of the dictionary, I assumed that a significant part of the lexemes would fall into category of archaisms, which are not high-frequency words. I also assumed that it would not contain many well-known russicisms, such as sovietisms. On the other hand, I also expected increased occurrence of plant and animals' names[7]. Semantic analysis confirmed these assumptions and showed us that excerpted lexemes are relatively diverse and therefore difficult to classify. Subsequent frequency analysis was based on excerpting frequency data from CNC which showed us the relative frequency of analyzed lexemes during analyzed period. The results of frequency analysis confirmed that above mentioned archaisms and vernacular words and some of lesser-known plant and animals' names are only minimally represented in corpora or not at all. Most remaining lexemes the showed zero trend tendency and only few lexemes showed a certain downward or upward trend. From a methodological point of view, I can state that it turned out to be appropriate to apply "First hits in documents" because it helped me to obtain more accurate results with regard to the composition of the corpus. In conclusion the results obtained demonstrate what impact Russian has had on Czech in the last thirty years and they also provide us with the typology of russicisms. What is more the results can be used to show how to use the CNC to perform a frequency analysis of lexemes in works of similar focus.

References

en:cnk:syn:verze8. (2019, Dec 18). In: Příručka ČNK. <http://wiki.korpus.cz/doku.php?id=en:cnk:syn:verze8&rev=1576679698>. [online]. [cit. 14:52, 11. 4. 2020].

FILIPEC, J., ČERMÁK, F.: *Česká lexikologie.* Praha: Academia, 1985.

GIGER, M., SUTTER-VOUTOVA, K.: *Transparency of morphological structures as a feature of language contact among closely related languages: Examples from*

7 Václav Machek dedicated part of his scientific work to zoological and botanical nomenclature, e. g. *Czech and Slovak names of plants* (*Česká a slovenská jména rostlin*, 1954).

Bulgarian and Czech contact with Russian. Boston, De Gruyte: Besters-Dilger, J., Pfänder, S., Rabus, A., Dermarkar, C.: Congruence in contact-induced language change, 2014.

HAVRÁNEK, B.: *Vývoj spisovného jazyka českého.* Praha: Československá vlastivěda, řada II, 1936.

LILIČ, G. A.: *Rol' russkogo jazyka v razvitii slovarnogo sostava češskogo literaturnogo jazyka (konec XVIII – načalo XIX veka).* 1982.

MACHEK, V.: *Etymologický slovník jazyka českého.* Praha: Nakladatelství Lidové noviny, 1997.

ŠMILAUER, V.: *Obohacování slovní zásoby: kurs pořádaný kruhem přátel českého jazyka v Praze.* Praha: Kruh přátel českého jazyka v Praze, 1953.

ŠMILAUER, V.: *Ruské vlivy na češtinu.* Naše řeč, roč. 25, 1941.

VLACHOV, S., FLORIN, S.: *Neperevodimoje v perevode.* Moskva: Meždunarodnyje otnošenija, 1980.

VLČEK, J.: *Porovnání slovní zásoby ruského jazyka se slovní zásobou českého jazyka.* Praha: Univerzita Karlova, 1986.

VLČEK J.: *Úskalí ruské slovní zásoby: slovník rusko-české homonymie a paronymie.* Praha: Svět sovětů, 1966.

About the author

Anna Caldrová
Masaryk University, Faculty of Arts, Department of Slavonic Studies, Brno, Czech Republic
413405@mail.muni.cz