A Quantitative Study of Russian Colour Terms *Buryj* and *Koričnevyj* in the Google Books Ngram Corpus

Vladimir V. Bochkarev[a], Anna V. Shevlyakova[a], Galina V. Parameі[b] and Ekaterina V. Rakhilina[c]

[a] *Kazan Federal University, Kremlyovskaya str. 18, Kazan, 420008, Russia*
[b] *Liverpool Hope University, Hope Park, Liverpool, L16 9JD, United Kingdom*
[c] *Higher School of Economics, Staraya Basmannaya str, 21/4, Moscow, 105066, Russia*

Abstract

We report a microdiachronic investigation of distributional semantics of two competing Russian colour terms (CTs) for 'brown', *buryj* (12th cent.) and *koričnevyj* (17th cent.), while using Russian subcorpus of Google Books Ngram. By conducting time-series analysis (1800–2009) of bigrams containing either of these terms, we estimated frequency of occurrences of the two "Russian browns" and explored changes in the extent of the terms' combinability with nouns signifying objects (N=259). Results provide evidence that in total frequency of use, *koričnevyj* overtook *buryj* at the beginning of 1920s and unequivocally prevails from the beginning of 1960s. Furthermore, the perplexity index indicates significant increase in the scope of objects whose denotations collocate with *koričnevyj*. This is complemented by the observed increase of the Jensen-Shannon divergence between frequency distributions of *buryj* and *koričnevyj*, with both phenomena being particularly manifested from 1960s. The obtained estimates of distributional semantics corroborate the status *koričnevyj* as the basic CT for 'brown' in modern Russian. The present diachronic corpus analysis provides novel insights into linguistic evolution of an emergent basic CT – by revealing the process of it gradually supplanting an old term with a similar colour meaning, the timescale of the new term's increase in usage, and significant expansion in its distributional semantics.

Keywords

Computational Linguistics, Google Books Ngram, linguistic evolution of colour terms, Russian terms for 'brown', *buryj* and *koričnevyj*, combinability, collocations, diachronic distributional analysis, frequency distribution, Jensen-Shannon divergence

1. Introduction

In the present study we explore linguistic evolution of the two competing Russian terms for 'brown' *бурый* / *buryj* and *коричневый* / *koričnevyj* by methods of diachronic computational analysis. The two terms differ in the time of their emergence and lexical origin, and in colour space together fill the slot termed *brown* in English or its counterparts in other European languages [1].

In modern Russian, *koričnevyj* is considered basic colour term (BCT) for 'brown', according to the criteria provided in the seminal work of Berlin and Kay [2]. As such, *koričnevyj* is attested in numerous linguistic and psycholinguistic studies [e.g. 1, 3–6]. The term emerged in the 17th century as a derivative of Russian word *korica* 'cinnamon', which in turn was derived from *kora* 'bark' [7–9].

Along with it, a significant scope of objects is still named by Russians by the old term *buryj* 'dust/greyish brown, brownish black' [10]. Studies in historical linguistics attest emergence of *buryj* in Old Russian in the 12th century [3, 7–9]. According to Herne [7], it is cognate of Mongolian *bürüj* 'dark-coloured' and is related to Persian **bōr* 'red, colour of pistachio' and Turkish **bur* 'fox-red'.

Proceedings of the Linguistic Forum 2020: Language and Artificial Intelligence, November 12-14, 2020, Moscow, Russia
EMAIL: vbochkarev@mail.ru (VVB); anna_ling@mail.ru (AVS); parameg@hope.ac.uk (GVP); rakhilina@gmail.com (EVR)
ORCID: 0000-0001-8792-1491 (VVB); 0000-0002-2659-1887 (AVS); 0000-0003-2611-253X (GVP); 0000-0002-7126-0905 (EVR)

The earliest dictionary examples of expressions containing *buryj* usually refer to (i) horse coat, defined as being between russet and (dark) brown, and (ii) silver (lit. black-*buryj*) fox. In modern Russian, the term exclusively collocates with 'bear', 'coal', 'ore', 'wheat' etc.

Here we investigated contextualised linguistic behaviour of each of the two "Russian brown" terms, specifically, their combinatorial lexical typology. Unlike broadly used (decontextualized) psycholinguistic analysis of denotative meanings of CTs, linguistic analysis focusses on collocational possibilities, structural semantics of the nouns denoting objects, and the contexts CTs are used in [11]. In diachronic studies, the linguistic approach provides insights into innate mechanisms and driving forces underlying the ways of re-structuring of colour categorisation system in order to capture optimisation of evolutionary dynamics of (basic) colour categories. Statistical changes in the co-occurrences of colour concepts over time in a large corpus reflect changes in distributional semantics, namely, the drift of colour concepts over time stipulated by sociocultural processes [12, 13]. Among the latter are local practices, technology and aesthetics [11], as well as sociocultural incentives that transpire in colour terminology: pragmatic and semantic distinctions [14]; cultural symbolism and values accorded to particular colour due to the workmanship that coloured objects received, and the distances that materials travelled [15], thus, bestowing the colour prestige and making its name a marker of social identity [16].

In the present study we undertook a diachronic computational analysis of Russian subcorpus of Google Books Ngram [17] that contains Russian books spanning more than four centuries, to explore frequency of occurrences of the two "Russian browns" and combinability of *buryj* and *koričnevyj* with nouns signifying certain objects. In pursuing this, we tested Rakhilina's [18, 19] hypothesis that an incipient colour term (here: *koričnevyj*) gradually expands in the realm of nouns signifying objects, increasingly supplanting the old term (here: *buryj*) in collocations, to finally becoming entrenched as a BCT. In our analysis we leaned upon indicative results of our previous study that ascertained frequencies of collocations of the two "Russian browns" with nouns denoting objects, where we employed the National Corpus of Russian Language (18^{th}–21^{st} centuries) [20].

2. Method

2.1 Dataset source and data cleansing

Russian subcorpus of Google Books Ngram (GBN) was employed [17], which contains data on frequencies of individual words, as well as *n*-grams, contiguous sequences of *n* words, with $n = 2, 3, 4,$ or 5. In the present study, the second version of the GBN corpus [21] was used, which includes texts of 591,310 books published in Russian between 1607–2009, with the total number of words amounting to more than 67 billion words. The GBN corpus was criticized by some as being unbalanced [22, 23]. In spite of this, the exceptionally large size of the corpus makes it a valuable tool for studies of language evolution, addressed, for example, in [24, 25]. Notably, the majority of the books contained in the Russian subcorpus of GBN were published after the beginning of the 19^{th} century. Here the analysed period in effect comprised about 200 years, since the data for distributional analysis of *buryj* and *koričnevyj* becomes sufficient and representative starting from 1830.

For the analysis, we extracted frequencies of all 2-grams (bigrams) corresponding to attributive constructions with *buryj* and *koričnevyj* (including their inflectional forms). Bigrams of the *Noun+colour* and *colour+Noun* types were selected automatically. Noteworthy, GBN is a part-of-speech (POS) tagged corpus, however, it contains numerous POS-tagging errors. To rectify inaccuracies, to lemmatize the nouns that collocate with the terms *buryj* and/or *koričnevyj*, POS-tagged data were verified using the OpenCorpora morphological dictionary (OC) [26, 27]. The OpenCorpora is one of the largest electronic dictionaries of the Russian language, which currently contains 391,800 lemmas that include 5,140,000-word forms. Finally, to ensure that only the target bigrams were selected, in some cases a manual check was performed in addition. In total, 2,621 bigrams were selected, including words related to 796 different lemmas. Selection of the bigram lists and their lemmatization, extraction of bigram frequencies from the GBN subcorpus, and statistical analysis were performed using scripts written in the Matlab environment.

2.2 Data analysis

Analysis of changes in distribution of the terms *buryj* and *koričnevyj* was performed (along with other methods) as vector representation of the word meaning, the method broadly applied for ascertaining distributive semantics [28–31]. Recently this approach was also used to estimate diachronic changes in word semantics and reveal new word meaning(s) [32–35].

For semantic computation, all referred to works utilized frequencies of the word in question in various contexts. However, different methods of computing word-representing vectors were employed, e.g. Pointwise Mutual Information [36] or Lexicographer's Mutual Information [37]. In addition, for estimating semantic similarity (distance) between words different metrics are applied.

In the present study we applied the explicit word vectors by using relative frequencies of *buryj*- and *koričnevyj*-bigrams that occur in different contexts, i.e. collocate with various nouns. Presaging the results reported below, 259 nouns were found to collocate with both *buryj* and *koričnevyj*, hence, the dimensionality of vector representation was *d*=259. Further, for each year and each of the two 'brown' terms, frequency vectors were computed and normalized to 1. Finally, differences in distributional semantics of *buryj* and *koričnevyj* were estimated by the Jensen-Shannon divergence (JSD) [38]:

$$JSD(p||q) = \frac{1}{2}\sum_i p_i \log_2 \frac{p_i}{(p_i+q_i)/2} + \frac{1}{2}\sum_i q_i \log_2 \frac{q_i}{(p_i+q_i)/2}, \quad (1)$$

where p_i and q_i are components of the two compared vectors for the *i*-th context.

A simple technique was proposed in [39] that allows one to estimate the contribution of each context to the obtained distance estimation. Note that each term in formula (1) reflects the contribution of only the *i*-th component of the compared distributions *p* and *q*. The values for each of the components, separately, are calculated as follows:

$$\frac{1}{2}\sum_i p_i \log_2 \frac{p_i}{(p_i+q_i)/2} + \frac{1}{2}\sum_i q_i \log_2 \frac{q_i}{(p_i+q_i)/2} \quad (2)$$

Further, one can sort the contexts in descending order of this value to determine by this means which contexts contributed most to the JSD value. A similar approach was used in the present analysis to identify specific collocations whose frequency change most strongly affects the change in the JSD values over time. To do this, we calculated increments of the values defined by formula (2) and sorted the nouns (bigram constituents) in the descending order of the values of these increments.

3. Results

3.1. Dynamics of frequency distribution of *buryj* and *koričnevyj*

As indicated in 2.2. above, 796 nouns were found in the corpus to collocate with the terms *buryj* and *koričnevyj*, whereby 133 co-occur only with *buryj*, 404 only with *koričnevyj*, and 259 nouns appear in combinations with either. Thus, over the entire period between 1607–2009 the term *koričnevyj* generally collocates with more nouns designating various objects than *buryj*.

Figure 1 shows the change of frequency of the terms *buryj* and *koričnevyj* (in all inflectional forms) over time. It is apparent that at the beginning of the 19th century frequency of the term *buryj* is significantly higher than that of *koričnevyj*. However, after a long period of competition, from the beginning of the 1980s, in the 20th century frequency of *koričnevyj* started to prevail. It is worth bearing in mind that in GBN some objects are mentioned quite often in combination with "Russian brown" terms, whereas others are mentioned only several times per century. Hence, the observed dependencies can be due to co-occurrences of each of the two terms with a relatively small number of frequently used nouns, and the obtained dependencies might disguise the ongoing competition of the "Russian brown" terms in typical cases.

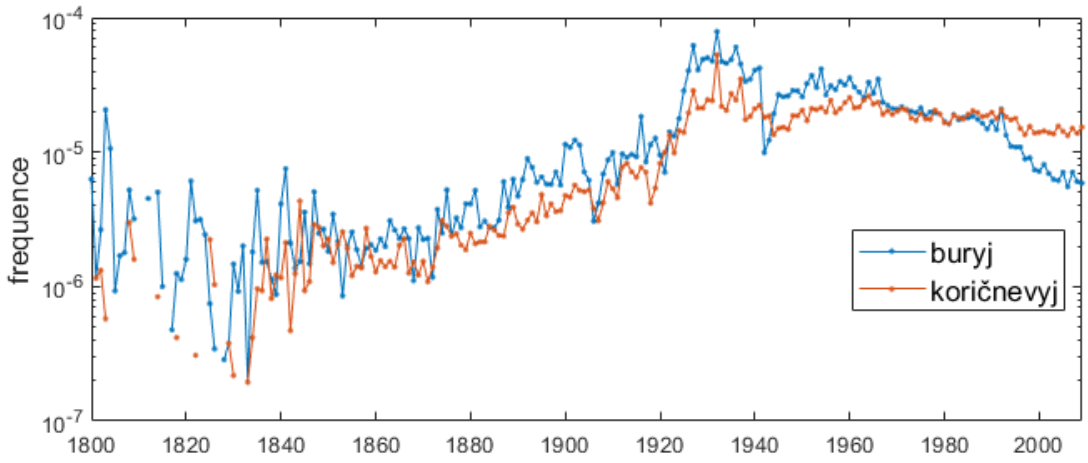

Figure 1: Total frequency of use of colour terms *buryj* and *koričnevyj* in GBN between 1800–2009

We scrutinised diachronic dynamics of the diversity of the "Russian brown" terms' usage by computing information entropy (h) of each term's frequency distribution in various contexts, i.e. in bigrams with different nouns. Expressed in bits, entropy is not particularly telling. Therefore, in Figure 2 we present its more instructive derivative – perplexity of frequency distribution [40], equal to 2^h, that reflects the number of frequently used alternatives, i.e. nouns collocating with either *buryj* or *koričnevyj*. Note that before 1840 the GBN corpus has data insufficient for a reliable analysis of the distribution of the terms *buryj* and *koričnevyj*.

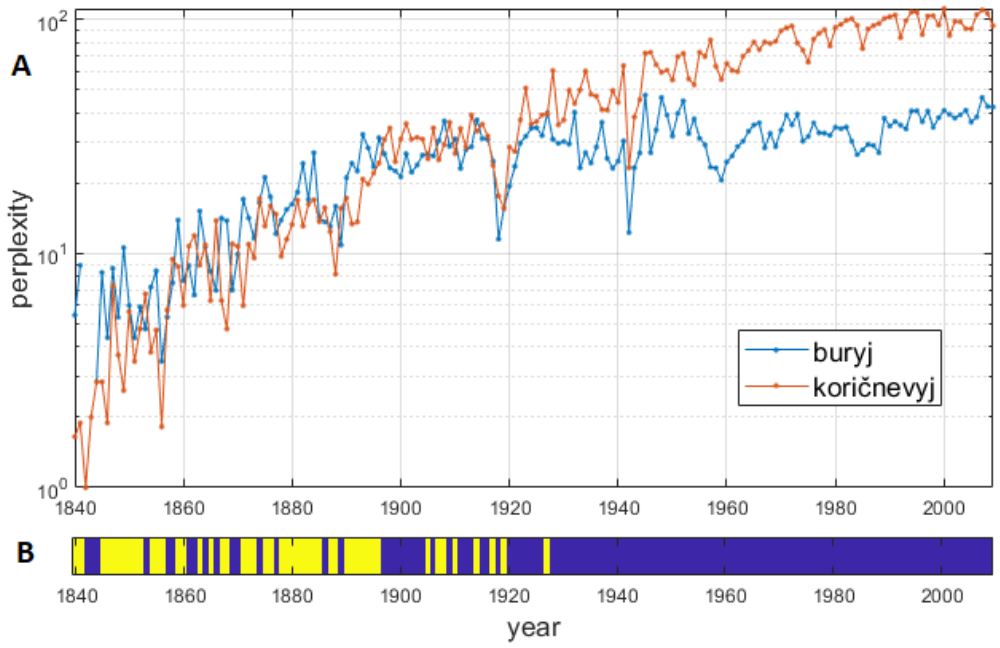

Figure 2: A: Perplexity of frequency distribution of bigrams containing a noun and either *buryj* or *koričnevyj*, computed year-by-year. B: The prevalent term perplexity is colour-coded by yellow for *buryj* and by blue for *koričnevyj*

As illustrated by Figure 2A, frequencies of both *buryj* and *koričnevyj* increase after 1840 – primarily due to the growth of the corpus size (as a manifestation of the Heaps' law [41, 42]). It is also apparent that initially, more objects collocate with *buryj* than with *koričnevyj*. However, after a long-period competition, from 1920s combinations of denoted objects with *koričnevyj* start to prevail. Furthermore, in the post-WWII period, the diversity of objects combined with *koričnevyj* become even greater compared to those with *buryj*. Dynamics of the prevalence of perplexity of the two "Russian browns",

on year-by-year basis, is presented in an alternative form in Figure 2B: it reveals an initial greater collocational diversity of *buryj* (until ca. 1900), the ensuing process of competition of the two terms (around 1900–1920), followed by overtaking the combinability diversity by *koričnevyj* in 1920s and its further incremental raise from mid-1940s.

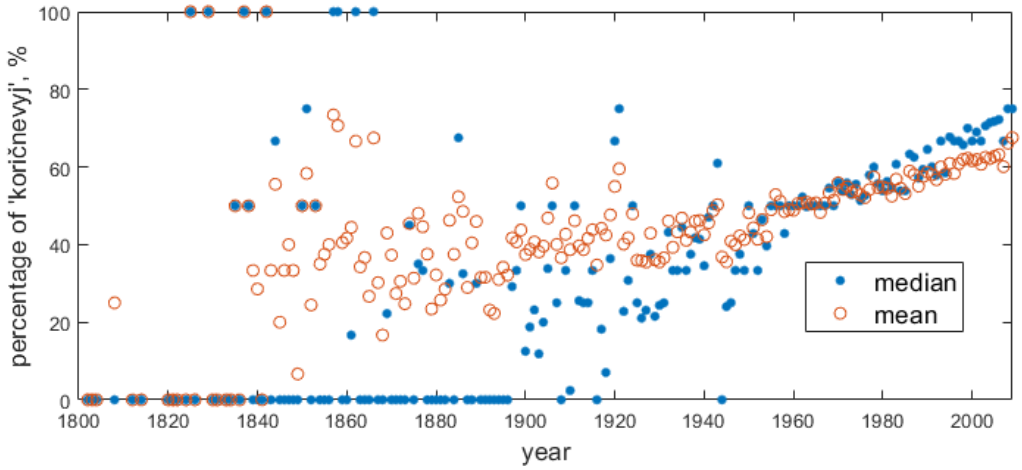

Figure 3: Mean and median percentage of collocations with *koričnevyj* from the total number of combinations with nouns (N=259) that in GBN co-occur with either *buryj* or *koričnevyj*

We undertook scrutiny of the competition process between *buryj* and *koričnevyj* by ascertaining the term prevalence for the subsample of object-denoting nouns (N=259) that collocate with either of the terms. Specifically, for each year and each of the colour terms we estimated the frequency of collocation with individual nouns and, from the total number of combinations with *buryj* and *koričnevyj*, calculated the proportion of objects collocating with *koričnevyj*. This proportion was estimated by two measures – as a mean over 259 nouns for each year and median. The latter measure is probably be more indicative since a small number of spurts in corpus individual nouns might significantly bias resulting mean values. Figure 3 prompts the tendency of an increase in the proportion of *koričnevyj* – from ca. 30% at

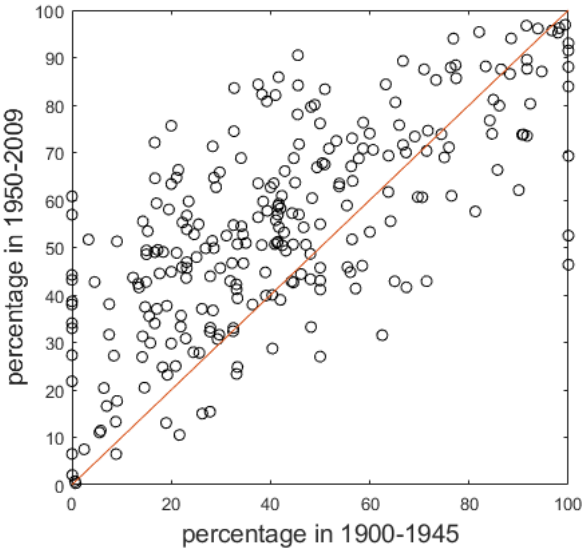

Figure 4: Average proportion of combinations with *koričnevyj* from the total number of nouns (N=259) in GBN collocating with either *buryj* or *koričnevyj* in the first and second half of the 20th century

the end of the 19th century to about 70% by the beginning of the 21st century. Furthermore, post-WWII the (ongoing) process of *buryj* being supplanted by *koričnevyj* is unidirectional.

Although, in tendency, the proportion of *koričnevyj*-combinations increases, this does not imply that it uniformly increases for all considered 259 nouns of objects. Figure 4 shows a scatterplot contrasting

values of *koričnevyj*-proportion for the (approximately) first vs. second halves of the 20th century. Apparent is the general tendency of the increase in the share of *koričnevyj* between 1950–2009, which amounts to 74.1% of the cases, although for some objects the proportion of *koričnevyj* slightly decreases (manifested by points below the diagonal in Figure 4).

3.2. Differences in distribution of *buryj* and *koričnevyj*

We further explored dynamics of distributional semantics of the two "Russian browns" by estimating the JSD between the frequency distributions of the terms *buryj* and *koričnevyj* for the 259 nouns combining with either of them. The outcome is presented in Figure 5. Large values reaching 1.0 (maximum possible JSD value) for the early analysis period (ca. 1850–1870) might result from depleted amount of corpus data for those years. For the following century, approximately from 1870s till 1960s, the JSD values decrease, to then revealing a slow ongoing increase. It is conceivable that the incessant JSD decrease before 1960s reflects the expansion of *koričnevyj* use in combination with nouns for the objects that hitherto had combined solely or predominantly with *buryj*. Conversely, subsequent JSD increase, from 1960s till the beginning of the 21st century, may manifest contraction of the scope of objects named *buryj*.

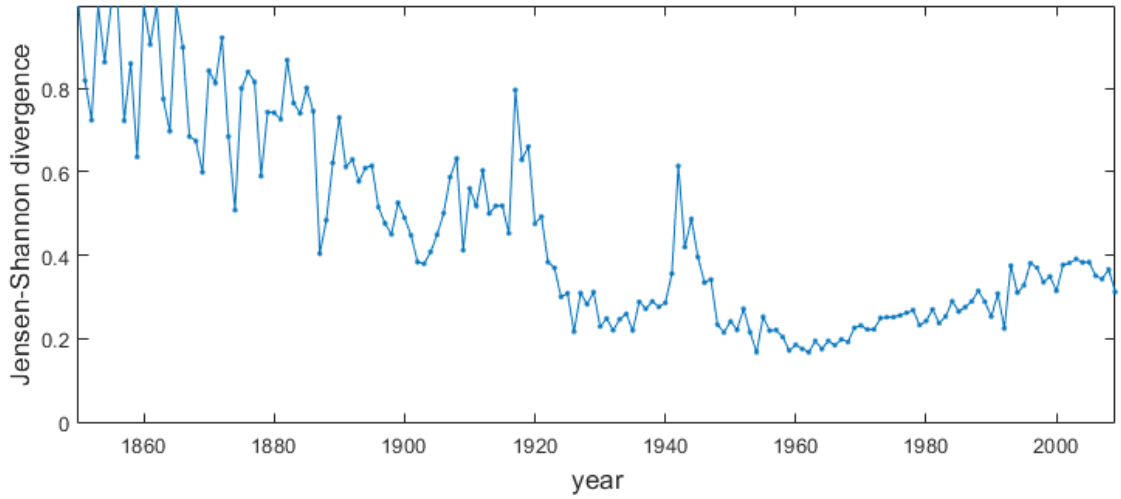

Figure 5: The Jensen-Shannon divergence between frequency distributions of *buryj* and *koričnevyj* with nouns (N=259) collocating with either term; GBN, 1840–2009

3.3. Exploring the impact of the corpus size on the values of distributional divergence of *buryj* and *koričnevyj*

We are cognisant though that, while reflecting a genuine change in the two terms' distributional semantics, i.e. the diachronic linguistic phenomenon, the JSD decrease might, in addition, result from a confounding factor – growth of the yearly corpus size, which is likely to bias JSD estimates [43]. For the Russian GBN subcorpus, in particular, it is known that until 1960 its size was rapidly increasing. To ensure that statistical significance of the observed changes in distributional semantics of *buryj* and *koričnevyj* are veridical, we examined whether the JSD estimates depended on the yearly amount of corpus data by performing a statistical modelling using the bootstrap-like procedure developed in [44].

The algorithm included the following steps [44]:

- Choosing the timespan, within which the frequency distribution of the target-word combinations is unlikely to change.
- Computing relative frequencies of the word combinations, i.e. the frequencies independent of the corpus size. For this, empirical frequencies of the word combinations in the considered year were normalized by the total corpus size in that year.

- Selecting one of the year values of the relative frequency from the chosen timespan for each component (independent of other components) of the frequency vector (frequency of the word combination).
- Computing JSD between the vectors generated at the previous steps.

The implementation of this algorithm allowed to simulate an empirical distribution of the JSD.

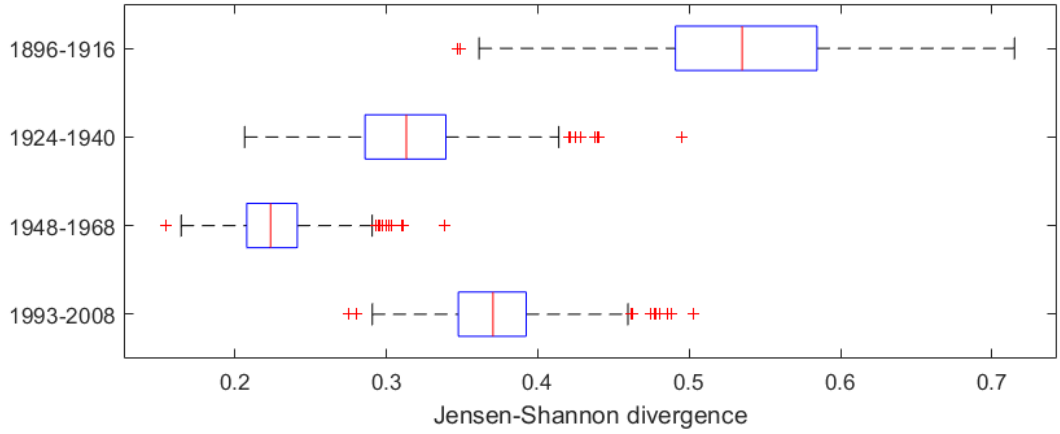

Figure 6: Boxplots of simulated values of the Jensen-Shannon divergence between frequency distributions of *buryj* and *koričnevyj* with nouns (N=259) collocating with either term, for four selected timespans, of comparable length, between 1896–2008

Leaning upon the results presented in Figure 5, we selected four timespans of comparable length, wherein the JSD values vary within relatively small limits and do not show signs of a clear trend: 1896–1916, 1924–1940, 1948–1968, and 1993–2008. Empirical distribution of the JSD was simulated for each of these timespans using bootstrapping (samples of 1,000 values generated). The simulation results are shown in Figure 6. As one can see, for each pair of the subsequent timespans the ranges of the obtained JSD values hardly overlap, implying that the obtained JSD differences definitely are statistically significant.

We were still aware of a possible impact on the JSD values of the variation in the corpus size, as well as changes in frequency of the compared terms (see Table 1). A calculation scheme that allows one to consider this effect is described in [44]. Following the proposed algorithm [44], we generated a vector representation of each "Russian brown" term constructed from relative frequencies of the combinations that include the given term. The so obtained vector was then multiplied by the mean value of the term absolute frequency within the timespan. The expected values of the frequencies of the term combinations were rounded down, and the obtained frequency vectors were normalized to 1.0.

Table 1

Average yearly number of occurrences of *buryj* and *koričnevyj* in combination with nouns (N=259) collocating with either term, in each of the four selected timespans between 1896–2008

Timespan	*buryj*	*koričnevyj*
1896–1916	224.3	132.4
1924–1940	2,249	1,252
1948–1968	6,007	5,417
1993–2008	2,413	4,014

For the timespan 1896–1916, the average yearly number of occurrences of the terms *buryj* and *koričnevyj* is 224.3 and 132.4 respectively (see Table 1); in comparison, for the timespan 1924–1940, it equals 2,249 and 1,252. Let us assume, as the null hypothesis, that the distribution of both colour terms actually is the same in both timespans, and the observed differences are associated solely with the change in the corpus size and the absolute frequency of the compared terms. Let us further assume that the relative frequencies of the noun combinations with *buryj* and *koričnevyj* are equal to their empirical values for the later of the two timespans (1924–1940). The calculation shows that, under the

assumption that the average yearly absolute frequencies of *buryj* and *koričnevyj* are 2,249 and 1,252 respectively, the JSD median value is 0.3200. However, if the average yearly absolute frequencies are 224.3 and 132.4 respectively, as in the earlier of the two timespans (1896–1916), the JSD median value increases to 0.3607, with the JSD standard deviation 0.0525. This comparison indicates that the change in the corpus size can indeed cause an estimate bias, however, the small discrepancy is not sufficient to explain the observed large differences between the two corresponding JSD values for the two compared timespans.

The calculations for the two other timespans, 1924–1940 and 1948–1968, were carried out in a similar way. The average yearly absolute frequencies of both colour terms for the later timespan (1948–1968) are 6,007 and 5,417 respectively. The modelled estimation of the JSD median value for this case is 0.2296. Provided the average yearly absolute frequencies of 2,249 and 1,252 (as in 1924–1940) are set in the model, under the same frequency distributions, the JSD median value increases to 0.2474, while the standard deviation of the JSD is 0.0278. Thus, as in the first instance, the bias is observed in this case, too, but, again, it is not significant and cannot explain the observed differences.

4. Conclusions

The present diachronic computational analysis of the two competing "Russian browns", the old term *buryj* (12th cent.) and the historically newer term *koričnevyj* (17th cent.), explored dynamics of the terms' linguistic behaviour using Russian subcorpus of the second version of Google Books Ngram. By conducting time-series analysis spanning 1800–2009, we ascertained and compared combinability of the two terms, in bigrams, with nouns signifying various objects. We were particularly interested in frequencies of occurrences *buryj* and *koričnevyj* in bigrams, which revealed the noun (N=259) collocational possibility with either of the 'brown' terms. The results provide evidence that in total frequency of use, *koričnevyj* overtook *buryj* at the beginning of 1920s, to progressively prevail from the beginning of 1960s. Furthermore, the perplexity index indicates significant increase in the scope of objects, whose denotations collocate with *koričnevyj*, in initially at the beginning of 1920s, with another upsurge mid-1940s, i.e. time windows following the two sociocultural upheavals, the WWI/October revolution and civil war in Russia, and WWII respectively. The findings on the expansion of *koričnevyj* collocational potential is complemented by the gradual increase of the Jensen-Shannon divergence between frequency distributions of *buryj* and *koričnevyj* observed from 1960s. The obtained estimates of distributional semantics corroborate the status *koričnevyj* as the basic CT for 'brown' in modern Russian. Together the findings provide convincing evidence supporting Rakhilina's [18, 19] hypothesis that an incipient colour term, *koričnevyj*, entrenches as a basic gradually, by expanding to the realm of nouns signifying objects with a colour previously named by the old term, *buryj*. Beyond this, the reported diachronic corpus analysis offers novel insights into linguistic evolution of an emergent basic CT – by revealing the process and timescale of the new term's increase in usage, significant expansion in its distributional semantics, and increasingly supplanting an old term in collocations, where the two terms compete.

5. Acknowledgements

The project was supported by the Russian Science Foundation (Grant No. 20-18-00206 to VVB and AVS). The authors are grateful to V.D. Solovyev for insightful comments on data analysis and outcome interpretation.

6. References

[1] G. G. Corbett, G. Morgan, Colour terms in Russian: Reflections of typological constraints in a single language, Journal of Linguistics 24 (1988) 31–64. https://www.jstor.org/stable/4175920

[2] B. Berlin, P. Kay, Basic Color Terms: Their Universality and Evolution, University of California Press, Berkeley, CA, 1969/1991.

[3] M. Vasmer, Russisches etymologisches Wörterbuch I–III, Carl Winter Universitätsverlag, Heidelberg, 1953.
[4] R. M. Frumkina, Cvet, smysl, sxodstvo. Aspekty psixolingvističeskogo analiza [Colour, Meaning, and Similarity: Aspects of a Psycholinguistic Analysis], Nauka, Moscow, 1984 (in Russian).
[5] I. R. L. Davies, G. G. Corbett, The basic color terms of Russian, Linguistics 32 (1994) 65–89.
[6] G. V. Paramei, Y. A. Griber, D. Mylonas, An online color naming experiment in Russian using Munsell color samples, Color Research and Application 43 (2018) 358–374. doi:10.1002/col.22190
[7] G. Herne, Die Slavischen Farbenbenennungen. Eine semasiologisch-etymologische Untersuchung, Almqvist & Wiksells Boktryckeri AB, Uppsala, 1954.
[8] N. B. Bakhilina, Istorija cvetooboznačenij v russkom jazyke [History of Colour Terms in Russian], Nauka, Moscow, 1975 (in Russian).
[9] Slovari russkogo jazyka 11-17 vekov: Ètimologija i istorija slov russkogo jazyka [Dictionaries of the Russian Language of the 11th-17th Centuries: Etymology and History of Russian Words], Nauka, Moscow, 1975–, pp. 314, 358 (in Russian).
[10] P. S. Falla, The Oxford English–Russian Dictionary, Clarendon Press, Oxford, 1984.
[11] S. C. Levinson, Yélî Dnye and the theory of basic color terms, Journal of Linguistic Anthropology 10 (2000) 3–55. doi:10.1525/jlin.2000.10.1.3
[12] C. P. Biggam, The Semantics of Colour: A Historical Approach, Cambridge University Press Cambridge, UK, 2012.
[13] L. Decock, Conceptual change and conceptual engineering: The case of colour concepts. Inquiry (2020). doi:10.1080/0020174X.2020.1784783
[14] L. Steels, Modeling the cultural evolution of language, Physics of Life Reviews 8 (2011) 339–356. doi:10.1016/j.plrev.2011.10.014
[15] J. Gage, What meaning had colour in early societies? Cambridge Archaeological Journal 9 (1999) 109–126. doi:10.1017/S0959774300015237
[16] R. E. MacLaury, Social and cognitive motivations of change: Measuring variability in color semantics, Language 67 (1991) 34–62. https://www.jstor.org/stable/415538
[17] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, Quantitative analysis of culture using millions of digitized books, Science 331 (2011) 176–182. doi:10.1126/science.1199644
[18] E. V. Rakhilina, Kognitivnyj analiz predmetnyx imen: semantika i sočetaemost' [Cognitive Analysis of Object Names: Semantics and Combinability], Russkie slovari, Moscow, 2000/2008 (in Russian).
[19] E. V. Rakhilina, Linguistic construal of colors: The case of Russian, in: R. E. MacLaury, G. V. Paramei, D. Dedrick (Eds.), Anthropology of Color: Interdisciplinary Multilevel Modeling, John Benjamins, Amsterdam/Philadelphia, 2007, pp. 363–377. doi:10.1075/z.137.24rak
[20] E. V. Rakhilina, G. V. Paramei, Colour terms: Evolution via expansion of taxonomic constraints, in: C. P. Biggam, C. A. Hough, C. J. Kay, D. R. Simmons (Eds.), New Directions in Colour Studies, John Benjamins, Amsterdam/Philadelphia, 2011, pp. 121–131. doi:10.1075/z.167.15rak
[21] Y. Lin, J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, S. Petrov, Syntactic annotations for the Google Books Ngram corpus, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Vol. 2, 2012, pp. 238–242.
[22] E. A. Pechenick, C. M. Danforth, P. S. Dodds, Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution, PLoS ONE 10 (2015) e0137041. doi:10.1371/journal.pone.0137041
[23] A. Koplenig, The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets — Reconstructing the composition of the German corpus in times of WWII, Digital Scholarship in the Humanities 32 (2017) 169–188. doi:10.1093/llc/fqv037
[24] V. D. Solovyev, V. V. Bochkarev, S. S. Akhtyamova, Google Books Ngram: Problems of representativeness and data reliability, in: A. Elizarov, B. Novikov, S. Stupnikov (Eds.), Data Analytics and Management in Data Intensive Domains. Communications in Computer and Information Science, volume 1223, Springer, Cham, 2019, pp. 147–162. doi:10.1007/978-3-030-51913-1_10
[25] S. Richey, J. Taylor, Google Books Ngrams and political science: Two validity tests for a novel data source, PS: Political Science & Politics 53 (2020) 72-77. doi:10.1017/S1049096519001318

[26] OpenCorpora, n.d. http://opencorpora.org/dict.php
[27] V. V. Bocharov, S. V. Alexeeva, D. V. Granovsky, E. V. Protopopova, M. E. Stepanova, A. V. Surikov, Crowdsourcing morphological annotation, Computational Linguistics and Intellectual Technologies 13(1) (2013) 109–114. http://www.dialog-21.ru/media/1308/dialog_2013_vol1web.pdf
[28] J. Weeds, D. Weir, D. McCarthy, Characterising measures of lexical distributional similarity, in: Proceedings of the 20th International Conference on Computational Linguistics, 2004, pp. 1015–1021.
[29] P. Pantel, Inducing ontological co-occurrence vectors, in: Proceedings of the 43rd Conference of the Association for Computational Linguistics, 2005, pp. 125–132.
[30] J. A. Bullinaria, J. P. Levy, Extracting semantic representations from word co-occurrence statistics: A computational study, Behavior Research Methods 39 (2007) 510–526. doi:10.3758/BF03193020
[31] M. Sahlgren, The distributional hypothesis, Italian Journal of Disability Studies 20 (2008), 33–53.
[32] K. Gulordava, M. Baroni, A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus, in: Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP, 2011, pp. 67–71.
[33] V. Kulkarni, R. Al-Rfou, B., Perozzi, S. Skiena, Statistically significant detection of linguistic change, in: WWW '15: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 625–635. doi:10.1145/2736277.2741627
[34] X. Tang, W. Qu, X. Chen, Semantic change computation: A successive approach, World Wide Web 19 (2016) 375–415. doi:10.1007/s11280-014-0316-y
[35] X. Tang, A state-of-the-art of semantic change computation, arXiv:1801.09872 [cs.CL] (2018). doi:10.1017/S1351324918000220
[36] R. M. Fano, Transmission of Information: A Statistical Theory of Communication, MIT Press, Cambridge, MA, 1961.
[37] S. Mitra, R. Mitra, S. K. Maity, M. Riedl, C. Biemann, P. Goyal, A. Mukherjee, An automatic approach to identify word sense changes in text media across timescales, Natural Language Engineering 21 (2015) 773–798. doi:10.1017/S135132491500011X
[38] D. M. Endres, J. E. Schindelin, A new metric for probability distributions, IEEE Transactions on Information Theory 49 (2003) 1858–1860. doi:10.1109/TIT.2003.813506
[39] V. Bochkarev, A. Shevlyakova, V. Solovyev, A method of semantic change detection using diachronic corpora data, in: W. M. P. van der Aalst et al. (Eds.), Analysis of Images, Social Networks and Texts. AIST 2019. Communications in Computer and Information Science, volume 108, Springer, Cham, 2020, pp. 94–106. doi:10.1007/978-3-030-39575-9_10
[40] P. F. Brown, V. J. Della Pietra, R. L. Mercer, S. A Della Pietra, J. C. Lai, An estimate of an upper bound for the entropy of English, Computational Linguistics 18 (1992) 31–40.
[41] H. S. Heaps, Information Retrieval: Computational and Theoretical Aspects, Academic Press, New York, 1978, pp. 206–208.
[42] D. C. van Leijenhorst, Th. P. van der Weide, A formal derivation of Heaps' Law, Information Sciences 170 (2005) 263–272. doi:10.1016/j.ins.2004.03.006
[43] C. Wartena, Distributional similarity of words with different frequencies, in: Proceedings of the Dutch-Belgian Information Retrieval Workshop, 2013, pp. 8–11. https://serwiss.bib.hs-hannover.de/frontdoor/index/index/docId/335
[44] V. V. Bochkarev, A. V. Shevlyakova, Calculation of a confidence interval of semantic distance estimates obtained using a large diachronic corpus, Journal of Physics: Conference Series 1730 (2021).