How Orthography affects Preposition Use in Russian Social Media Names

Evan Chapple

12/15/21

Аннотация

An analysis of bigrams made up of a preposition and a social media name between orthographies was conducted in a subcorpus of Russian National Media from the new Russian National Corpus. It was found that certain social media names (SMNs) had a higher level of variance of usage between orthographies of preposition use than others. It was further shown that this difference in variance was not directly correlated to the penetration rate of the social media names inside of Russia.

1 Introduction

The Cyrillic writing system was specially designed for Slavic languages, including Russian, and has been around for over a thousand years [Iliev(2013)]. However, in the modern day, Russian has absorbed words from languages written in the Latin alphabet. Despite Russian being one of the Slavic languages that still uses the Cyrillic alphabet, through the effects of globalization and English, a language written in the Latin alphabet, being the undisputed Lingua Franca of the modern world, Americanisms/Englishisms have now penetrated Russia [Keller(1988)].

Due to glocalization, there has been an increase in the use of Latin writing in Cyrillic documents [Ustinova and Bhatia(2005)]. One specific area where this has spawned is when references are made to foreign entities. A specific subset of companies that are commonly referenced would be social media companies due to there large-scale growth in recent years. When referring to foreign based companies whose names are not originally in Russian/Cyrillic, there are a number of steps that need to be taken before such words enter the lexicon.

When words are imported into a language, there are a few decisions that have to be made so as to integrate the word well into the language's orthography, morphology, and syntax [Holthenrichs(2019)]. In regards to the first of these three, one must decide how to write the word. There are two options; either maintain the original orthography, or transliterate the word. In regards to SMNs in Russian, it was found that both of these processes are common. In regards to morphology, it is important to decide on declesion, whether a new word should be declined (кошмар -> кошмара, этаж -> этаже, etc.), or if it should remain in the dictionary form throughout (e.g., меню, кафе, etc.) [Brown(1996)]. The latter is not very common in regard to social media sites in Russian, as most appear masculine and then are declined as such. The final important area for importing a word into a language is fitting it into the syntax. In Russian, there are two prepositions, "в" and "на", which occupy similar semantic spaces as "in" and "on" respectively in English. There, however, exists a specific pair when there is no additional information beyond the locative being given by the preposition between all words in Russian and either "в" or "на" [Holthenrichs(2019)]. The distinction between these two is somewhat confusing for native English speakers and perhaps Russian speakers as well. A somewhat shallow definition of these two could be that "в" is used when the boundaries of something are well defined, such as a park [Brown(1996)]. However, "на" is used when the object does not have such well defined boundaries, such as a street or an island. This idea of boundaries becomes very uncertain when the object is something as intangible as a site on the web, thus the aforementioned rules do not prescribe a specific preposition. It is then, by extension, hypothesized that when certain social media sites are used more, the more concrete the preposition pair becomes and the less disruption a change in script would cause to the usage of the "correct" pair.

A previous study [Holthenrichs(2019)], looked at what was the "correct" preposition pair was for each SMN across orthographies. The results can be found in Figure 1. It was found that the "correct" preposition for most SMNs in Russian in "в" with YouTube being a notable exception.
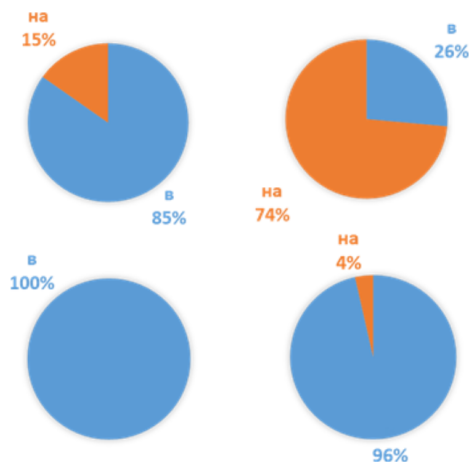
Figure 1: Pie graphs showing the average proportion of preposition use with SMNs. The SMNs from left to right, top to bottom are Facebook, Twitter, Youtube, and Instagram. Image from [Holthenrichs(2019)]

2 Corpus

The corpus used in this study was the new Russian National Corpus (nRNC) [Савчук(2011)]. The nRNC was developed by varying Russian groups including the Institute of Russian Language. It was found that the usage of social media terms were too low in the old Russian National Corpus while the nRNC returned many more hits. The nRNC is made up of many individual subcorpora including but not limited to media/newspaper, accentological corpus, and an educational corpus. This project was limited to the media/newspaper subcorpus for two main reasons. The first being that as newspapers are commonly considered a very revised and prescriptive text-type, one would expect overwhelmingly grammatical constructions within this text-type. The second being that within this subcorpora, a high number of SMNs were found.

3 Method

In order to test the variance of the different usages of prepositions before a social media name (SMN), the nRNC was searched for the bigram with the Latin or Russian spellings of a SMN, preceded by one of the two prepositions. It was assumed that the Latinized spellings would not decline; thus, "Facebook" was used and not "Facebook(j)e". The reverse was also not included; SMNs that were written in Cyrillic but were not declined were not used in the statistical analyses as they were five to ten times less common than the

declined forms. There were a few instances with the SMN being used in compound words, e.g., "YouTube-канале"; These were ignored as they did not make up a significant proportion of the results, nor were these results indicative of the SMNs' preposition pair. Instances of "во" were collapsed into the results for "в" as they have the same underlying form. The Latin and Cyrillic transcriptions of the SMNs are found in Table 1 and examples from the corpus with translations are found in Table 2.

Latin	Cyrillic
Twitter	Твиттер
Facebook	Фейсбук
YouTube	Ютуб
Instagram	Инстаграм

Table 1: List of SMNs of interest in dictionary form

Latin Examples
"в своем посте в Twitter"
-In her post on Twitter-
"написал он в Facebook"
-He wrote on Facebook-
"авторский канал на YouTube"
-Author channel on YouTube-
"ее пост в Instagram"
-Her post on Instagram-
Cyrillic Examples
"на своей странице в Твиттере"
-On her page on Twitter-
"что прочла рекламу в Фейсбуке"
-That she read the ad on Facebook-
"Миллион просмотров на ютубе"
-A million views on YouTube-
"на фото в Инстаграме"
-In the photo on Instagram-

Table 2: Examples from the nRNC showing the SMNs in context with translations

Less strict searches were also used, where the corpus was searched for the co-occurrence of the SMN and one of the prepositions. This approach led to a number of results where the head of the NP was not the SMN but another word, usually a form of the word "site" (e.g., "на сайте «YouTube»"). This approach lead to a stark increase in results, but each one would need to be manually vetted to check the head of the NP. This approach was not employed due to the time-constraint and thus not included in the statistical analyses.

4   Results

A two proportion z-test was then used to assess if the number of instances of the incorrect preposition being used for a specific social media site differed significantly when the social media name was in Latin vs Cyrillic form. The "correct" preposition is judged to be the preposition that is used the most often with the form (Latin or Cyrillic) with the highest n. These results also concur with the majority found in the paper by Holthenrichs (2019). For each z-score calculated, we are able to assign a certain confidence level to reject the null hypothesis, the hypothesis that the proportions of "correct" preposition to "incorrect" preposition between orthographies are drawn from the same population. This was performed for each of the SMNs under study, specifically Twitter, Facebook, Instagram, and YouTube.

For the SMN Facebook, it was decided that "в" was the main preposition (MP) so any instance of "в" being used was considered "correct" and any of "на" was considered "incorrect" by the test. When the SMN was in the Latin alphabet the proportion of incorrect usage was 0.0555 and when the SMN was in the Cyrillic alphabet the proportion of incorrect usage was .0508. This produced a z-score of 0.807 and with p=.21. For Facebook, the alphabet of the SMN showed no significant difference in incorrect preposition use at p=.05.

For the SMN Twitter, it was decided that "в" was the MP so any instance of "в" being used was considered "correct" and any of "на" was considered "incorrect" by the test. When the SMN was in the Latin alphabet the proportion of incorrect usage was 0.00378 and when the SMN was in the Cyrillic alphabet the proportion of incorrect usage was .00642. This produced a z-score of 1.740 and with p=.0409. For Twitter, the alphabet of the SMN showed a significant difference in incorrect preposition use at p=.05.

For the SMN YouTube, it was decided that "на" was the MP so any instance of "в" being used was considered "correct" and any of "в" was considered "incorrect" by the test. When the SMN was in the Latin alphabet the proportion

of incorrect usage was 0.159 and when the SMN was in the Cyrillic alphabet the proportion of incorrect usage was .500. This produced a z-score of 3.481 and with p=.0003. For Twitter, the alphabet of the SMN showed a significant difference in incorrect preposition use at p=.05.

For the SMN Instagram, it was decided that "в" was the MP so any instance of "в" being used was considered "correct" and any of "на" was considered "incorrect" by the test. When the SMN was in the Latin alphabet the proportion of incorrect usage was 0.0159 and when the SMN was in the Cyrillic alphabet the proportion of incorrect usage was .198. This produced a z-score of 12.786 and with p<.00001. For Twitter, the alphabet of the SMN showed a significant difference in incorrect preposition use at p=.05.

The results of the above tests are listed below in Table 3.

SMN	MP	Z-score	Confidence
Facebook	в	0.807	Too Low
Twitter	в	1.740	>90%
YouTube	на	3.481	>99%
Instagram	в	12.786	>99%

Table 3: SMN with their main preposition, z-score on the two proportion z-test and confidence interval

A further analysis was conducted to see if there were some semantic differences in the situations where one preposition was used over the other. This study did not produce any results as both prepositions were used in overlapping semantic space.

5   Discussion

The purpose of this study was to investigate the relationship between script and preposition use in Russian SMNs. It was found that a three out of the four SMNs under study had significant differences in the proportions of preposition use between scripts. This could suggest a couple of things. The first is that these SMNs do not have a 1:1 preposition to nominal pairing as other nouns do. This seems unlikely as this is an innate quality of Russian. The second is that SMNs with less penetration and usage in Russia, the less set-in-stone the pairings are.

In Figure 2, one can see the leading social media sites in Russia by penetration rate. This graph does not show what was previously hypothesized

as the highest social media site by penetration was YouTube, which was found to have the most disparate proportions of preposition use between orthographies.
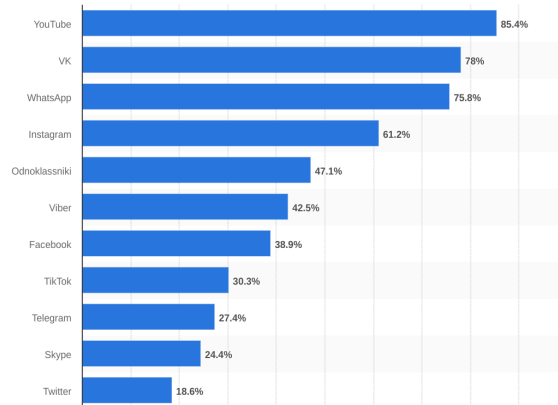

Figure 2: Graph of the usage of social media sites in Russia by penetration rates. Image from [Melkadze(2021)].

It is interesting to note that for the SMNs, YouTube and Instagram, when written in Cyrillic there is a higher proportion of "incorrect" preposition use than when written in the Latin alphabet. This could perhaps be due to the fact that these SMNs are most often written in the Latin alphabet, or atleast more so than in the Cyrillic alphabet.

Limitations with this study mostly had to do with the corpus used and the abilities to work within it. If one were to use an annotated corpus, it would be trivial to add in results in which the SMN was the head of the NP inside of a PP. Future works could add in this functionality if there is an annotated corpus with enough SMNs within.

Another direction for future work could be a cross-linguistic study to compare this difference between other Slavic languages as other Slavic languages may or may not have this variance. Ukrainian, for example, has a trichotomy of such prepositions, "на", "у", and "в", that could be used with SMNs [Holthenrichs(2019)]. On the other hand, Polish almost always uses the preposition "na" for SMNs, however Polish is written in the Latin alphabet, so the methods found within this study would not necessarily be applicable. All together, this study has led to an interesting discovery in regards to how a prescribed text-type in Russian still retains certain ambiguities as to how imported words are treated within the syntax and has opened doors to how these kind of words are incorporated into other languages and how effectively this happens.

Список литературы

[Brown(1996)] N. J. Brown. 1996. Russian Course; a complete course for Beginners. Penguin.

[Holthenrichs(2019)] Van Holthenrichs. 2019. How slavic languages reference social media. Unpublished paper.

[Iliev(2013)] Ivan G. Iliev. 2013. Short History of the Cyrillic Alphabet. International Journal of Russian Studies.

[Keller(1988)] Bill Keller. 1988. https://www.nytimes.com/1988/08/28/world/for-russians-the-americanisms-are-coming.html For russians, the americanisms are coming. The New York Times.

[Melkadze(2021)] A. Melkadze. 2021. https://www.statista.com/statistics/867549/top-active-social-media-platforms-in-russia/ Leading social media platforms in russia as of 3rd quarter of 2020*, by penetration rate.

[Ustinova and Bhatia(2005)] Irina P. Ustinova and Tej K. Bhatia. 2005. https://doi.org/https://doi.org/10.1111/j.0883-2919.2005.00433.x Convergence of english in russian tv commercials. World Englishes, 24:495–508.

[Савчук(2011)] С.О. Савчук. 2011. Корпус современной русской прессы: из опыта создания и использования.