Poljarnyj Vestnik: Norwegian Journal of Slavic Studies, vol. 24, 2021, pp. 1–12
https://doi.org/10.7557/6.5931

Subject-verb Agreement in Constructions with Quantifiers in Russian

Anna Aksenova

1. Introduction

The agreement of subject and predicate in Russian is less trivial than it might seem at first glance. One of the most controversial and complicated cases involves quantified subjects. For numerical quantifiers, which denote the exact number of objects, like тройка (group of three), четверка (group of four) etc. combined with a noun both singular and plural form markers can be triggered on the predicate (1-2). I will not discuss independent uses of numerals like (3), but limit myself to examples where the quantifier is followed by a noun phrase.

(1) Во всем была виновата эта *тройка лгунов*.
'This threesome of liars be-Pst.Sg to blame for everything. '
(2) *Тройка ребят* контужены, но легко.
'A threesome of guys be-Pst.Pl shell-shocked but not hard.'
(3) К машине подошли *трое*.
'Three (people) approached the car.'

The grammars of Russian usually treat this type of agreement as optional and describe them as the patterns, rather than categorical rules. However, even the patterns are not always defined clearly. Rozental' et al. (2005) state that for words два, три, четыре and двое, трое, четверо the predicate usually has plural markers, however for stative verbs the number can be varied. For words like тройка, четверка, сотня the agreement should be singular.

In this paper, I will analyze the factors that might influence the choice of the verb form in sentences with quantified subjects. I discuss 9 quantifiers, namely двое (collective numeral for group of 2), трое (collective numeral for group of 3), пара (group of 2), тройка (group of 3), десяток (group of 10), сотня (group of 100), тысяча (1000), миллион (1000000), миллиард (1000000000). In section 2 I explore previous research that has been conducted in this area. Section 3 is devoted to the data and variables that I used as a basis for quantitative analysis. In section 4, I discuss the results of my statistical investigation based on a number of models. Section 5 presents a summary of the study and introduces perspectives for further investigation.

2. Literature Overview

Relations between the subject and the predicate in Russian are quite intriguing. In general, Russian has syntactic agreement, whereby the predicate agrees with the subject in number and person (non-past tense) or gender (past tense). However, a complicating factor is so-called semantic agreement, i.e., the use of the agreed word in a form that is either not predetermined by the grammatical and morphological features of the noun or is only partially predetermined by them. Moreover, while maintaining the general nature of mutual influence, subject-verb relationships can vary depending on the formal characteristics of the subject and predicate (Valgina 1991).

In my research, I focus on the behaviour of quantifier phrases in the subject position. I define a quantifier phrase as a construction (Endresen, Janda 2020) with the following structure: X Y-GEN. The X-slot is filled with a quantifier, which can either exhibit purely numerical properties or behave like a noun. The Y-slot may be filled with a noun or pronoun phrase, but in the present study I focus only on noun phrases.

There are a number of studies analyzing adjective agreement in these constructions, and all of them have showed that both grammatical and semantic properties of the noun in the Y position have a significant impact on the agreement patterns. Nesset (2020) analyzed adjective agreement in constructions with paucal numbers (two, three, four). As one of the independent variables that has proven to influence the choice of the ending, he suggests the position of the verb (before the target phrase or after, Nesset 2020:8). It is proposed that the number and position of the verb might influence the choice of agreement marker, but statistical analysis shows that it is not the most influential factor. However, I believe that for my study the position of the verb can be informative since the word order influences the interpretation of the sentences and syntactic relationship between subject and predicate (Dyakonova 2009).

Another factor that has been shown to be relevant is the animacy of the noun. Baranova (2016) introduces an animacy hierarchy that can account for the behavior of nominal quantifiers in Russian. It is mentioned that singular verb agreement makes the subject less animate (Baranova 2016:93), however, this idea is just mentioned and not discussed.

Moreover, the crucial role of the numerical component of the construction cannot be denied. The type of quantifier can influence the verb agreement since different quantifiers exhibit different morphological patterns. G. Corbett showed that simple numbers in Russian are located on a scale from the closest to the adjectives to the most similar to nouns. Moreover, the smaller the quantitative value of a numeral, the more adjectival properties it shows (Corbett 1978). This scale can be summarized as follows:

один (one) > оба (both) > два (two) > три (three) & четыре (four) > пять (five) (six, etc.) > сто (one hundred) > тысяча (one thousand).

Depending on the point of view, the extreme members of this scale may or may not be included in the grammatical class of numerals. Mikaèljan (2012) suggests that the split of the scale takes place between four and five and there is an opposition between small numbers (1-4) on the one hand and big numbers (5 and bigger) on the other. For the small numbers, collective variants (e.g., двое instead of два) are used more often and an animate type of agreement is preferred. She also points out that the bigger the number, the more

variation in adjective agreement it shows (Mikaèljan 2012:432-433). Taking this research into account, I decided to investigate quantifiers from both parts of the scale.

As I have already mentioned, there is variation in the use of collective and non-collective quantifiers. Dobrushina, Panteleeva (2008) studied the factors that can influence the choice of the numeral. According to these authors, collective numerals are usually used with animate masculine nouns. Moreover, collective numerals are attracted to the nouns that denote groups and organizations. As one can see, there are quite a few properties of the arguments of the numerical constructions that interact and determine the choice of the adjective agreement. However, no proposals have been made about subject-verb agreement. Thus, I believe that to achieve a better understanding of the quantifier constructions' nature, one has to consider the predicate agreement.

3. Data and Methodology

For this study, I collected the data from the Russian National Corpus (RNC, ruscorpora.ru). It was decided that the analysis will only be run over the most modern examples, that is why I downloaded a subsample of the sentences from the XXI century subcorpus. There were not enough examples in the RNC, so I added some examples from RuTenTen, a Russian internet corpus (https://www.sketchengine.eu). All sentences contain a quantifier construction in the subject position and a predicate. I downloaded samples for words двое, трое, пара, тройка, десяток, сотня, тысяча, миллион, миллиард.

I manually deleted all examples that are irrelevant for the purposes of my study (e.g., the word пара can denote a couple of living creatures and function as an independent noun) and annotated the sample with seven levels that are discussed below.

3.1 Keyword

This factor refers to the quantifier in the subject position. The hypothesis is that the type of the quantifier has a considerable influence on the number agreement of the predicate. The distribution of the sentences for each keyword in my sample is presented in Table 1.

Quantifiers	Number of Examples
десяток	81
двое	145
миллиард	22
миллион	65
пара	164
сотня	43
трое	151
тройка	99
тысяча	88

Table 1. Keyword distribution in the dataset.

I should emphasize that the data includes two collective numerals (двое, трое) and seven non-collective quantifiers (пара, тройка, десяток, сотня, тысяча, миллион, миллиард). I

decided to use the word пара (couple) instead of двойка (group of 2) as a non-collective equivalent of двое, since двойка is used as a quantifier in the only context (4-6).

(4) двойка лошадей
a pair of horses
(5) *двойка парней
a pair of guys
(6) пара парней
a couple of guys

Based on previous research, I expect there to be two types of opposition: collective vs. non-collective quantifiers and small vs. big numbers.

3.2 Source

The variable Source shows from which source the sentence was taken. As can be seen from the Table 2, I annotated four types of the source: books, Internet, magazine, newspaper. I consider this feature to be relevant since it might reflect the style of texts in which plural or singular agreement markers are preferable. I expect that for the examples from the Internet more marginal forms will appear since 'online' language is close to oral speech where novel forms or constructions are born.

Source	Number of Examples
Book	199
Internet	187
Magazine	364
Newspaper	108

Table 2. Source distribution in the dataset

3.3 Genre

One more factor that might influence the use of certain types of agreement is the genre of the text. I divided all the examples into two groups, namely fiction and non-fiction. According to Table 3, there are twice as many non-fictional examples as fictional ones. The difference can be explained by the fact that there are quite many examples with the label 'Internet' and all of them were considered non-fictional. I must note that Genre and Source are correlated. I decided to annotate for both because in some cases the more fine-grained classification is helpful, but in other cases, it might create extra noise.

Genre	Number of examples
Fiction	275
Non-fiction	583

Table 3. Genre distribution in the dataset

3.4 Animacy

The feature animacy shows whether the modified noun in the subject position denotes an animate or inanimate participant. I expect this feature to be quite important for the agreement, since previously it has been demonstrated that the animacy of dependent nouns affects the behaviour of the head quantifier.

Animacy	Number of Examples
Animate	544
Inanimate	314

Table 4. Animacy distribution in the dataset

3.5 QP Position

I assume that the postverbal Quantifier Phrases (QP) can show more marginal agreement examples because the VS-order is not the global default for Russian (Bivon 1971). Admittedly, VS-order is not infrequent in Russian, e.g., for unaccusative verbs, but in such cases VS-order is motivated by the information structure (Dyakonova 2009).Despite the fact that usually the verb follows the subject, there are quite many cases of inversion in our data (see Table 5).

QP Position	Number of Examples
Postverbal	355
Preverbal	503

Table 5. Quantifier phrase position distribution in the dataset

3.6 Verb Number

Russian verbs usually agree in grammatical number (singular or plural) with the subject of the sentence. However, as mentioned, there are cases of semantic agreement when the verb takes the plural form because semantically the subject denotes a group. Moreover, there are also examples of agreement failure when the verb should be in the plural form but takes a singular marker instead. The verb number is the dependent variable of my analysis.

Verb Number	Number of Examples
Plural	361
Singular	497

Table 6. Verb number distribution in the data

4. Analysis

Looking at the plot of Verb Number distribution in our data (Figure 1), one can notice that collective numerals behave differently. They tend to trigger more plural agreement while for the non-collective ones, singular agreement is more widespread. Moreover, there is a noticeable variability within the non-collective quantifiers.

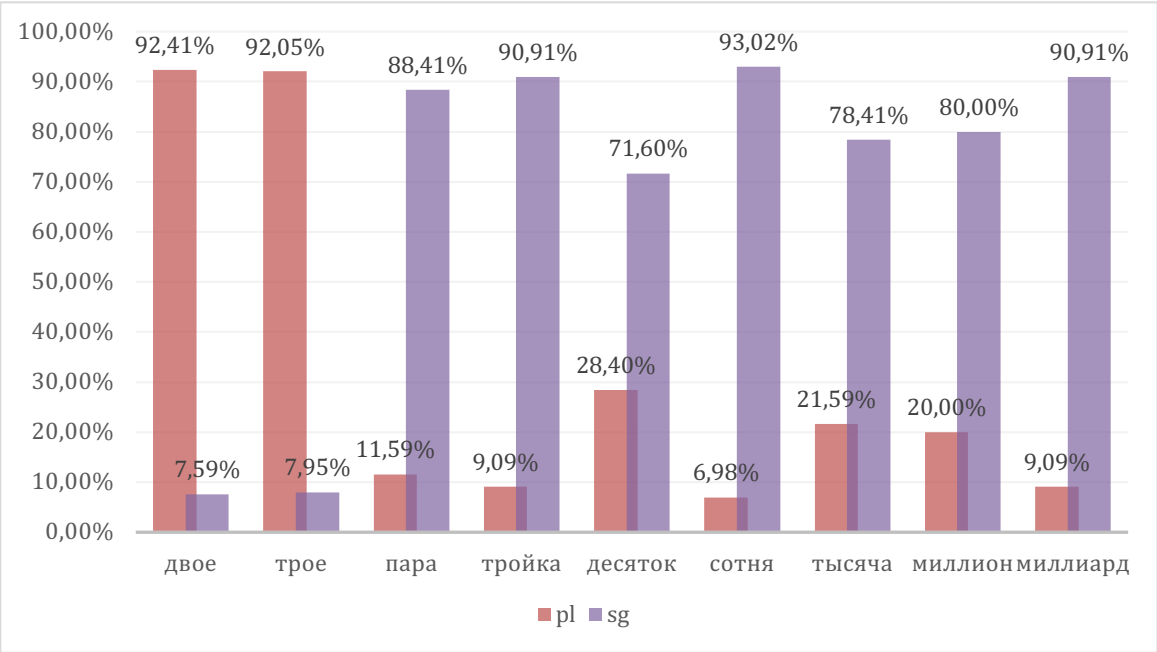

Figure 1. Verb number distribution for each keyword

In order to estimate the influence of all independent variables on the choice of agreement pattern, I ran a Random Forest model on all variables presented in the dataset. Random Forest (Strobl et al. 2009) is a machine learning algorithm that employs an ensemble of decision trees. The algorithm combines two main ideas: the bagging method and the random subspace method. The main idea is to use votes from a large ensemble of decision trees, each of which by itself may give a very low quality of classification, but due to their large number, the overall result is reliable. Random Forest is used for the investigation of the relative importance of the variables (Figure 2).

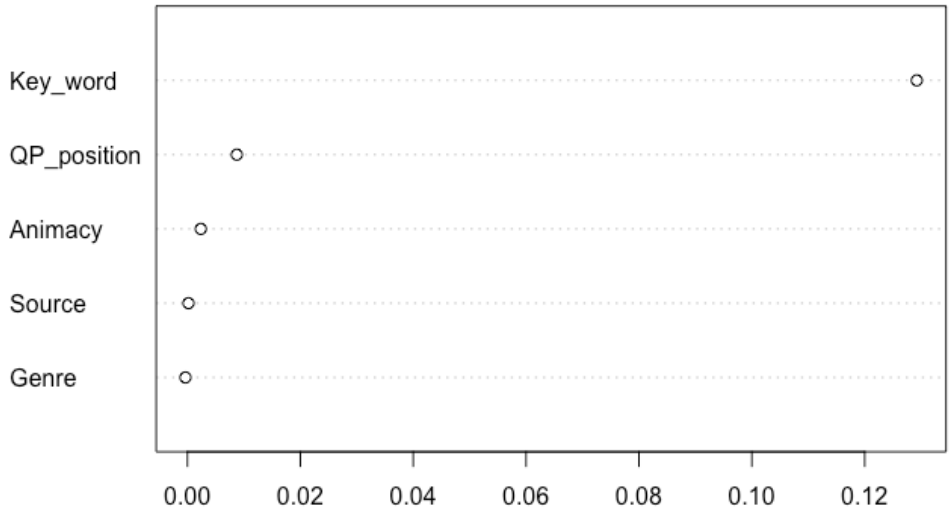

Figure 2. Conditional importance of variables for all quantifiers

Keyword appeared to be the most important factor as it was expected. Quantifier phase position is also quite influential, while Genre and Source compared to it did not make any difference. Based on the results of the Random Forest analysis, I decided not to take the last two factors into the account. Given that the Keyword is the most influential variable,

it is natural to ask where the split takes place, i.e., which of the quantifiers form groups. To address this question, I decided to run CIT analysis.

CIT or Conditional Inference Tree (Tagliamonte, Baayen 2012) is a statistical algorithm that is quite helpful for the analysis of the independent variables. The model builds binary decision trees that contain only two daughters for each node. While conducting the analysis, the examples of the training set are recursively divided into subsets, the records in which have the same values of the target variable. CIT facilitates analysis of the interaction of predictor variables (in this study Keyword, QP position, Animacy) by choosing such a partition from all possible at a given node and one factor so that the resulting daughter nodes are as homogeneous as possible. CIT is quite similar to regular regression models; however, the results are much easier to interpret. The tree diagram depicts the levels of importance for each factor and shows the values that are grouped (Figure 3).

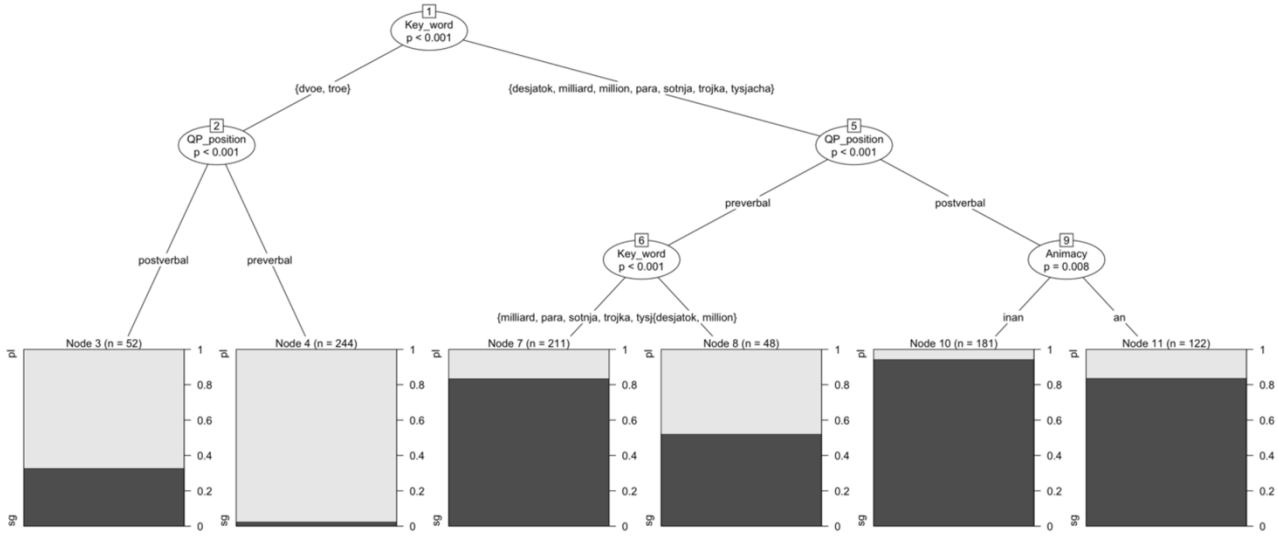

Figure 3. CIT algorithm output for all variables

The CIT model provides evidence that collective and non-collective quantifiers in Russian are different. The first (and the most significant) split factor is the keyword. The quantifiers are divided into two groups: двое, трое on the one hand and пара, тройка, десяток, сотня, тысяча, миллион, миллиард on the other.

Let us take a closer look at the collective numerals двое and трое. Compared to all the other words they trigger much more cases of plural agreement, and singular agreement seems exceptional. As shown by the Random Forest model, the quantifier phrase position is an important factor. All sentences with predicates in the singular form appeared when the subject followed the verb, which is itself not a neutral word order in Russian. It also supports the idea that singular subject-verb agreement is not characteristic of collective numerals.

Although двое and трое behave similarly according to the CIT analysis, the contexts of their use are quite different. Moreover, they have different IPM (item per million) score in Russian National Corpus: двое occurs almost 4 times more often than трое. To check if

there are any differences in the agreement patterns within collective quantifiers, I decided to run a chi-square test.

A chi-square test is a statistical method to compare the distribution of two variables in a two-dimensional matrix. In other words, it shows whether two categorical variables behave differently. As a null hypothesis, I suggest the following statement: двое and трое are used similarly. Then the alternative hypothesis will state that the distribution of the quantifiers is significantly different.

As it turns out, one cannot reject the null hypothesis since the X-squared is almost zero (1.6651e-30) and the p-value is 1, which is higher than the critical value. Since the distributions are not significantly different, it is not necessary to measure the effect size, so it is unnecessary to take Cramer's V into the account.

As the collective numerals appeared to be almost identical, I suggest focusing on the analysis of the non-collective ones, since there is a certain variation in their use according to the CIT that I plotted before.

To start with, I performed a Random Forest analysis to see whether the importance of independent variables is different when the collective numerals are excluded from the analysis.

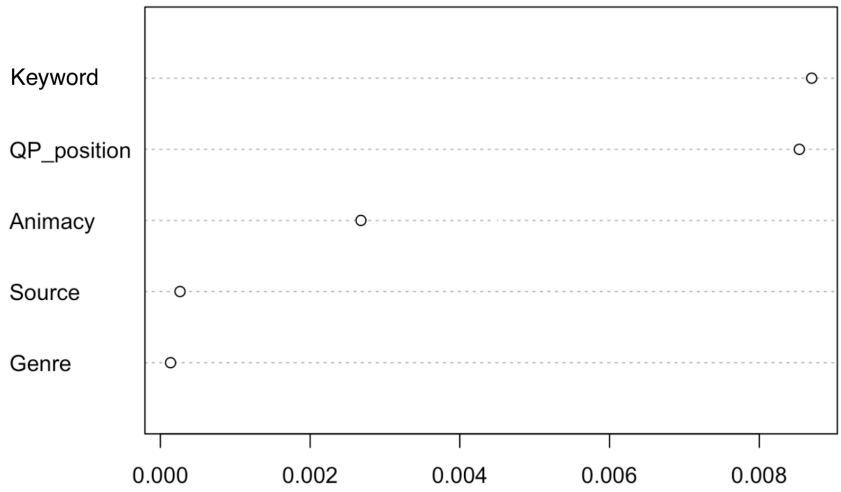

Figure 4. Conditional importance of variables for non-collective quantifiers

The Random Forest output graph (Figure 4) compares the conditional importance of five factors: Keyword, QP position, Animacy, Genre and Source. The lexeme itself still seems to be very important for the choice of the agreement pattern within the sentences with the non-collective quantifiers. The subject position remains the second important factor, however, the relative level of influence is much higher compared to the analysis I conducted on the whole dataset. Animacy of the nominal argument of the construction is also influential, while the Source and Genre of the text do not affect the choice of the verb number.

To visualize the split of the data, I carried out a CIT analysis on the subsample.

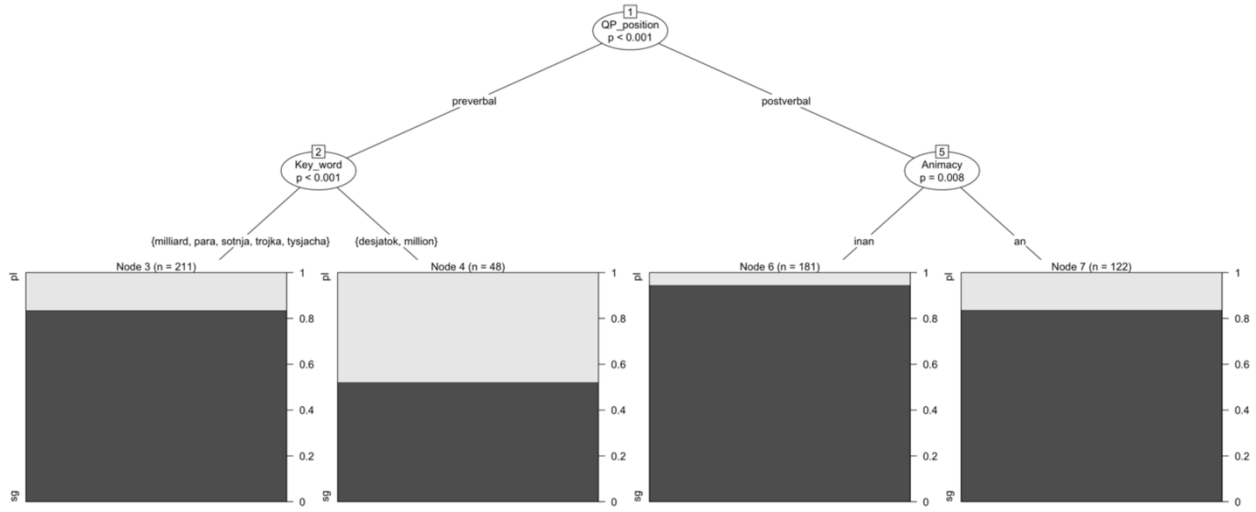

Figure 5. CIT algorithm output for non-collective quantifiers

Surprisingly, the CIT model contradicts with Random Forest. According to the tree plot above, it can be seen that the first split is based on the quantifier phrase position factor, which means that it is considered the most important independent variable. I assume that the reason behind the different analysis the models provide is the way they treat the data: for CIT the level of the split is important, while for Random Forest the overall influence of the factor is the key factor. To begin with, the difference in values between QP position and Keyword is very small, both of the variables are very important according to both models. Notice that the difference between the distributions in nodes 4 and 5 is much bigger than elsewhere. From the point of view of the first split, there is a very small difference between the two, and QP was chosen, but in general Keyword is doing more separation in the model because when Keyword is applied in preverbal position, it makes the biggest difference. Therefore, Keyword is more influential as a variable, but it does its work lower down in the tree as it is applied after choosing a preverbal position, not at the first split.

The second node of the CIT tree suggests splitting the non-collective numerals into two groups: two, three, hundred, thousand, and billion as opposed to ten and million. Going back to the hypothesis I suggested while introducing the factors, I must admit that this division is quite unexpected. Based on the noun vs. numeral oppositions (Corbett 1978), I assumed that the quantifiers that describe big numbers would behave differently from those which stand for small numbers. However, for our data, this is not true. Десяток and миллион are located in different parts of the scale, but nevertheless belong to the same group according to the CIT tree.

For now, the nature of the difference remains unclear, but I can hypothesize that it might be related to the morphological features of the numerals. Most of the numerals in the first group belong to the second declension (i.e., have -a ending in the Nominative case), whereas десяток and миллион behave like masculine nouns of the first declension type with zero endings. The only exception, in this case, is the word миллиард, which has the same declension pattern as десяток and миллион but belongs to another group. This marginal behavior can be explained by two facts. First, the lexeme миллиард has the least

number of occurrences in our dataset and the data might not be representative enough. Second, the IPM score for the word for billion in the Russian National Corpus is the lowest among all the other quantifiers (2,5 compared to 5 and higher for other lexemes).

The next step in our analysis is to figure out to what extent each of the independent variables influences the target variable. CIT does not show whether the factor influences the target value as the main effect or as an interaction. Moreover, its predictive power is not strong enough. To address those issues, I applied a Logistic Regression analysis. Since the dependent feature is binary, this model should be helpful to examine to which extent each of the factors influences the choice of verb form.

Taking into account the Random Forest analysis, I constructed 7 models with different combinations of three predictors: Key_word, QP_position and Animacy. Afterwards, I calculated the Akaike Information Criterion (AIC) to estimate the relative quality of the performance for each model on my data. Based on this criterion the optimal model included only main effects for all three independent variables.

To make sure that the results of the model are reportable I checked the factors for multicollinearity. The conventional value of the Variance Inflation Factor (VIF) that provides the evidence for multicollinearity is 10. However, for all variables VIF is slightly above 1, which means that the model is valid.

Furthermore, I studied the full output of the best model (Table 7–8). Although the R2 value is relatively low, the Dxy value is higher than 0.5 which means that the model is meaningful (Baayen 2008, 204).

Index	Value
R2	0.178
C	0.745
AIC	444.58
F-score of prediction	0.88

Table 7. Best model analysis parameters

	Coefficient	Z value	Pr (>|Z|)
Intercept	1.0026	3.269	0.001
Keyword = миллиард	1.4002	1.724	0.085
Keyword = миллион	0.5515	1.314	0.188
Keyword = пара	1.2567	3.379	0.0007
Keyword = сотня	2.0864	3.114	0.0018
Keyword = тройка	1.9745	4.317	1.58e-05
Keyword = тысяча	0.3730	0.978	0.328
QP_position = preverbal	-1.2525	-4.725	2.30e-06
Animacy = inanimate	0.9179	3.551	0.0003

Table 8. Summary of parameters for the best regression model

In our case, the intercept is the word десяток and predicted value is plural agreement. All factors, except for the preverbal position of the quantifier construction, trigger singular

number markers on the verb. Furthermore, Logistic Regression lends additional support to the CIT analysis, according to which десяток and миллион are not significantly different. However, it is also shown that миллиард has a similar agreement pattern, which supports my hypothesis about the importance of the type of declension. The F-score value is relatively high, so I can conclude that the model has quite good predictive power and describes the data well enough.

5. Conclusions

This paper deals with the subject-verb agreement of quantifier constructions in Russian. I discussed 9 quantifiers: двое, трое, пара, тройка, десяток, сотня, тысяча, миллион, миллиард. Statistical analysis (the CIT model) showed that collective numerals behave differently compared to other quantifiers. Collective numerals trigger plural agreement on the verb, while for non-collective numerals singular agreement is the more widespread option.

For the non-collective quantifiers, the phrase position is important: pre-verbal quantifier phrases tend to trigger plural agreement more than post-verbal ones. This corresponds to the fact that the semantic plural agreement is more likely to appear when the subject is located before the verb since it is essential to process the semantics of NP first to choose the agreement pattern of the verb based on semantic criterion. Syntactic agreement is applied at the lower level and thus does not correlate with surface word order.

Out of all the independent variables that were explored in this paper, the extralinguistic features of the text do not really influence the agreement. Although Animacy has some impact on the choice of agreement, it turned out to be less important than one might expect from previous research about the agreement of adjectives.

An interesting direction for future work would be the diachronic analysis of the quantifiers and taking into consideration more language-internal factors, e.g., verb type information.

6. References[1]

Baayen, R.H. (2008). Analyzing Linguistic Data. A Practical Introduction to Statistics Using R. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511801686

Bivon, R. (1971). Element order. Studies in the Modern Russian Language, vol. 7. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/S0022226700003601

Dyakonova, M. (2009). A phase-based approach to Russian free word order. Netherlands Graduate School of Linguistics.

Endresen, A., & Janda, L. A. (2020). Taking Construction Grammar One Step Further: Families, Clusters, and Networks of Evaluative Constructions in Russian. Frontiers in Psychology, 11, 2900. https://doi.org/10.3389/fpsyg.2020.574353

[1] All the translations of Russian titles were made by the author of this paper.

Mikaèljan, I. L. (2012). О категории одушевленности в конструкциях с числительными в русском языке. [Animacy in Constructions with Numerals in Russian Language] Смыслы, тексты и другие захватывающие сюжеты: сб. статей в честь, 429-447.

Nesset, T. (2020). A long birth: The development of gender-specific paucal constructions in Russian. Diachronica, 37(4), 514-539. https://doi.org/10.1075/dia.18057.nes

Slioussar, N. (2011). Russian and the EPP requirement in the Tense domain. Lingua, 121(14), 2048-2068. https://doi.org/10.1016/j.lingua.2011.07.009

Tagliamonte, S. A., & Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. Language variation and change, 24(2), 135-178. https://doi.org/10.1017/S0954394512000129

Баранова, В. В. (2016). Конструкции со счетными словами и семантическая одушевленность в русском языке. [Constructions with Counting Words and Semantic Animation in Russian] Русский язык в научном освещении, 1(31), 84-100.

Валгина, Н. С. (1991). Синтаксис современного русского языка: Учеб. для вузов. [Syntax of the Modern Russian Language: Textbook for Higher Education Institutions.] М.: ВШ

Добрушина, Н. Р., & Пантелеева, С. А. (2008). Собирательные числительные: коллектив как индивидуализация множественности. [Collective Numerals: The Collective as an Individualization of the Plural] В А. Мустайоки и др.(ред.), Инструментарий русистики: корпусные подходы, 107-124.

Розенталь, Д. Э., Джанджакова, Е. В., & Кабанова, Н. П. (2005). Справочник по русскому языку. [Handbook of Russian Language.] Айрис-пресс.

author: Anna Aksenova
affiliation: Higher School of Economics, Moscow
email: aksanna_a@mail.ru