Tore Nesset*

Big data in Russian linguistics?

Another look at paucal constructions

https://doi.org/10.1515/slaw-2019-0012

Summary: With the advent of large web-based corpora, Russian linguistics steps into the era of "big data". But how useful are large datasets in our field? What are the advantages? Which problems arise? The present study seeks to shed light on these questions based on an investigation of the Russian paucal construction in the RuTenTen corpus, a web-based corpus with more than ten billion words. The focus is on the choice between adjectives in the nominative (*dve*/*tri*/*četyre starye knigi*) and genitive (*dve*/*tri*/*četyre staryx knigi*) in paucal constructions with the numerals *dve*, *tri* or *četyre* and a feminine noun. Three generalizations emerge. First, the large RuTenTen dataset enables us to identify predictors that could not be explored in smaller corpora. In particular, it is shown that predicates, modifiers, prepositions and word-order affect the case of the adjective. Second, we identify situations where the RuTenTen data cannot be straightforwardly reconciled with findings from earlier studies or there appear to be discrepancies between different statistical models. In such cases, further research is called for. The effect of the numeral (*dve*, *tri* vs. *četyre*) and verbal government are relevant examples. Third, it is shown that adjectives in the nominative have more easily learnable predictors that cover larger classes of examples and show clearer preferences for the relevant case. It is therefore suggested that nominative adjectives have the potential to outcompete adjectives in the genitive over time. Although these three generalizations are valuable additions to our knowledge of Russian paucal constructions, three problems arise. Large internet-based corpora like the RuTenTen corpus (a) are not balanced, (b) involve a certain amount of "noise", and (c) do not provide metadata. As a consequence of this, it is argued, it may be wise to exercise some caution with regard to conclusions based on "big data".

Keywords: Big data, corpus linguistics, Russian, numeral, paucal

***Corresponding author: Professor Dr. Tore Nesset,** UiT The Arctic University of Norway, Department of Language and Culture, NO-9037 Tromsø, Norway, E-Mail: tore.nesset@uit.no

1 Problem and data

Since the term "big data" was coined in Silicon Valley in the 1990s (Lohr 2013), the term has spread from computer science to other fields and is now used pervasively in popular science (e.g. Stephens-Davidowitz 2017) and everyday speech. The term has also made its appearance in linguistics, and some linguistics programs now offer special courses devoted to "big data".[1] As is not unexpected for a recent term relating to rapidly developing fields such as computer science and data analysis, there is some uncertainty as to what exactly qualifies as "big data". However, most definitions seem to be similar to the one provided in the Merriam-Webster dictionary, where "big data" is defined as "an accumulation of data that is too large and complex for processing by traditional database management tools".[2]

To what extent is "big data" relevant for Slavic and Russian linguistics? In a few decades we have moved from a situation where we would read texts on paper and write out linguistic examples on index cards to a situation where large amounts of linguistic examples can be easily extracted from electronic corpora such as the Russian National Corpus (www.ruscorpora.ru). Even more recently, web-based corpora like the RuTenTen corpus (https://www.sketchengine.co.uk/rutenten-russian-corpus/) have appeared on the scene with more than ten billion words to offer. Clearly, Russian linguistics has entered the era of "big data".

The question is what consequences "big data" have for Slavic and Russian linguistics. In order to shed light on this question I will report on a case study of data from the RuTenTen corpus and compare the results to earlier studies, in particular a parallel investigation of data from the Russian National Corpus (Nesset submitted). On this basis, I will discuss advantages and disadvantages of "big data". My case study concerns a well-known issue in Russian linguistics, namely the form of the adjective in paucal constructions with the numerals *dve*, *tri* and *četyre* and a feminine noun:[3]

(1) V otkrytyx istočnikax upominajutsja **dve ser'eznye avarii**.
'**Two serious accidents** are mentioned in the open sources.'

1 An example is University of Pennsylvania where Mark Liberman offers the course "Big Data in Linguistics" (http://www.ling.upenn.edu/~myl/. Last accessed: September 19, 2018).

2 https://www.merriam-webster.com/dictionary/big%20data. Last accessed: March 28, 2018.

3 Throughout the article numbered examples are from the RuTenTen corpus. For the convenience of the reader, the relevant paucal constructions are boldfaced. The quantifiers *pol* 'half', *poltora* 'one and a half' and *obe* 'both' are not considered in the present study.

(2) V načale dekabrja proizošli srazu **dve ser'eznyx avarii**.
'**Two serious accidents** happened right in the beginning of December.'

In constructions of this type, there is rivalry between adjectives in the nominative as in (1) and the genitive as in (2).[4] Since, as mentioned, this rivalry has recently been investigated on the basis of the Russian National Corpus (Nesset submitted), we have a good standard of comparison for the data from the RuTenTen corpus to be analyzed in the present study. Data from the Russian National Corpus indicate that the use of adjectives in the nominative has increased in the twentieth century, and that it has reached a level in the beginning of the twenty-first century where more than 80 % of the examples have nominative adjectives (Nesset submitted, see also Pereltsvaig 2009 and Madariaga & Igartua 2017 and references therein). Nesset's (submitted) study suggests that the numeral itself is relevant for the choice of case in the adjective, insofar as *tri* and *četyre* have been more innovative and adopted nominative adjectives earlier than *dve*. However, beyond this, it has not been possible to establish robust support for other predictors on the basis of data from the Russian National Corpus.

The present study is an attempt to identify such predictors. First, I show that modifiers and predicates in the plural favor nominative adjectives, while modifiers and predicates in the singular are more likely to combine with adjectives in the genitive. Second, it is shown that genitive adjectives are favored by prepositions governing the numeral phrase. Third, contrary to the findings of Nesset (submitted), the RuTenTen data suggest that *dve* is more likely to combine with adjectives in the nominative than *tri* and *četyre*. Fourth, the RuTenTen data indicate that word-order is relevant, insofar as predictors to the left of the numeral phrase have more predictive power than predictors to the right. Finally, it is shown that the predictors of the nominative are more general (cover larger classes of examples) and display a stronger preference for the case in question. This arguably makes the pattern with nominative adjectives in (1) more easily learnable than the pattern with adjectives in the genitive in (2), and it is likely that the nominative may oust its competitor from the language in the future.

Although these points suggest that large datasets may help us detect new generalizations about paucal constructions, "big data" are not without problems. First, as opposed to the Russian National Corpus, the RuTenTen corpus is not

4 A question that is tangential to the present study and therefore will not be discussed in the following is whether the paucal numerals combine with nouns in the genitive (the traditional analysis defended in Andersen 2006), or whether the relevant nouns are in a different "numerative case" (Russian: *sčëtnaja forma*/*sčëtnyj padež*, Zaliznjak 2002 [1967]: 47; Mel'čuk 1985) or in a "paucal number" (Corbett 1993; Pereltsvaig 2009).

balanced, and it is therefore not clear to what extent it is representative of the Russian language. Second, since the RuTenTen corpus is based on data from the internet, we cannot be sure about the quality of the language in the examples, which may for instance involve machine translated text of poor quality and other "noise". Third, while the Russian National Corpus provides important metadata such as the time the example was created, the genre of the text, the name and gender of the author, etc., no such metadata are found in the RuTenTen corpus. In view of this, I argue that we should not accept generalizations based on "big data" uncritically, but instead use it as a supplement to data from balanced corpora or psycholinguistic experiments.

The dataset analyzed in the present study was extracted from the RuTenTen corpus in 2016. All examples with the paucal numerals *dve*, *tri* or *četyre* followed by an adjective and a noun were searched for. The searches yielded a number of irrelevant examples (e.g. with nouns in other cases), but these were weeded out manually. The database was then annotated manually for relevant predictors, such as pre- and postposed predicates and modifiers.[5] The resulting dataset, which consists of 93,261 examples (67,961 with adjective in the nominative as in (1), and 25,300 with genitive adjectives as in (2)), was subjected to statistical analysis. In section 2, we will explore the statistical model Random Forest, before we turn to Classification and Regression Trees (CART) in section 3. Section 4 summarizes the findings of the study.[6]

2 Interaction of predictors: Random Forest

A number of predictors have been mentioned in the literature on numeral constructions in Russian. In the following, we will see how the statistical model Random Forest enables us to assess the relative importance of predictors. I will show that the numeral itself (*dve* vs. *tri* vs. *četyre*) and predicates appear to have the strongest effect, followed by prepositional and verbal government and modifiers. Word-order is furthermore shown to be important, whereas the stress pattern of the quantified noun appears to be of no consequence.

5 Notice that the manual tagging was limited to the word immediately preceding the numeral and the word immediately following the quantified noun. In view of the large size of the dataset, manual annotation of words further away from the numeral construction was not feasible.

6 The dataset and the code for the statistical analysis are available at TROLLing (the Tromsø Repository of Language and Linguistics, https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/DG75YC).

As mentioned in the previous section, in Nesset (submitted) it is found that time is an important predictor, insofar as the use of nominative adjectives have increased dramatically since the middle of the twentieth century. Regrettably, since the RuTenTen corpus does not include information about the time the examples were created, it is not possible to consider development over time in the present study. However, it stands to reason that the data in RuTenTen represent Russian usage from recent years, since the data are taken from the internet. The proportion of genitive adjectives supports this. In Nesset's analysis of data from the Russian National Corpus, the proportion of genitive adjectives was over 80 % in the first half of the twentieth century, then dropped to 37 % for the period 1950–74 and to 32 % for the period 1975–99, while the period 2000–12 had 13 % adjectives in the genitive. The corresponding percentage for the RuTenTen data under scrutiny is 27 % (25,300 examples with the genitive out of a total of 93,261 examples), which suggests that the RuTenTen data are predominantly from the past two or three decades.

A predictor that has been discussed in the literature is the numeral itself. Gallis (1947: 70) found that the relative frequency of adjectives in the genitive is higher for *tri* and *četyre* than for *dve*, and the relevance of this variable has also been considered by *inter alia* Suprun (1957: 77), Mel'čuk (1985: 127), and Corbett (1993: 24–25). In order to test the importance of this variable, each example in the dataset was tagged for the relevant numeral. As we will see, the Random Forest analysis lends further support to the relevance of the numeral.

Another predictor included in the study by Nesset (submitted) is preposed modifiers, which can occur in the nominative (e.g., *èti* 'these' in (3) or the genitive, e.g., *celyx* 'as much as' in (4)):[7]

(3) **Èti tri magičeskie bukvy** vposledstvii stali sinonimičnymi ego imeni.
'**These three magic letters** later on became synonymous with his name.'

(4) **Celyx tri novyx audiosistemy** predusmotreny dlja XF 2012 goda.
'**Three whole new audio systems** were designed for the 2012 XF model.'

Modifiers in the nominative and genitive can also occur after the numeral phrase, frequently as participles:

7 Notice that I use the term "modifier" in a broad sense so as to cover both pronominal determiners like *èti* in (3) and quantifiers like *celyx* in (4).

(5) Skvoz' arku vidny **tri prjamye allei, obsaženny**e temno-zelenymi eljami.
'Through the arch one could see **three straight alleys, planted** with dark green fir trees.'

(6) V Irkutske bylo **dve ženskix gimnazii, nosivšix** imja svoego osnovatelja I.S. Xaminova.
'In Irkutsk, there were **two girls' schools, carrying** the name of their founding father I.S. Xaminov.'

The data from the Russian National Corpus did not show an effect of modifiers, but as we will see shortly, the dataset under scrutiny in the present study indicates an effect, albeit not a strong one.

When a numeral phrase is the grammatical subject of the sentence, the predicate can be in the singular or plural. In (6) we have the singular predicate *bylo* 'was', while plural predicates are found in (3)–(5). Again, the question arises as to whether singular vs. plural in the predicate has an impact on the choice of form in the adjective. As we will see, the dataset from the RuTenTen corpus suggests an effect, although no such effect was detected in the data from the Russian National Corpus analyzed in Nesset (submitted).

A predictor that is often mentioned in the literature on numeral phrases is the stress pattern of the quantified noun. For instance, Wade (2011: 215) proposes that a "genitive plural adjective is preferred with a feminine noun after 2–4 when there is a stress difference between the genitive singular and nominative plural of the noun" (see also Rozental' 1987: 276 and Graudina et al. 2001: 41). According to this view, one would expect a genitive adjective with e.g. *ruka* 'hand', which has a different locus of stress in the genitive singular (*rukí*) and nominative plural (*rúki*), but not for nouns with immobile stress, such as *škola* 'school'. However, other scholars (e.g. Pereltsvaig 2009: 426 and Šaronov 2014) have been critical to this idea, and Nesset's (submitted) study of the Russian National Corpus did not find an effect of stress. The data under scrutiny in the present study converges with the data from the Russian National Corpus in showing no effect of stress.

Yet another predictor that has received some attention in the literature is prepositional government (Suprun 1957: 79; Gorbačevič 1971: 261; Rozental' 1987: 277). Does it influence the choice of case in the adjective whether the numeral phrase is governed by a preposition or not? Here are two examples with the preposition *na* 'on':

(7) V konce III v. Provincija Panonija byla razdelena **na četyre administrativnye oblasti**.
'At the end of the third century, the province Panonia was divided **into four administrative areas**.'

(8) Ostrov Krit razdelen **na četyre administrativnyx oblasti**.
'Crete is divided **into four administrative areas**.'

As we will see below, the RuTenTen data confirms an effect of prepositional government, contrary to the findings reported in Nesset (submitted). Finally, I decided to investigate the possible effect of verbal government, although this is a factor that appears to have received little attention in the literature on paucal constructions in Russian:

(9) Sledstvie opredelilo **dve osnovnye versii** katastrofy.
'The investigation identified **two basic versions** of how the disaster happened.'

(10) Pravitel'stvo opredelilo **dve osnovnyx celi** programmy.
'The government identified **two basic goals** of the program.'

Here, the numeral phrase is the grammatical object of the verb *opredelit'* 'define, identify', which governs the accusative. As we will see, the present study indicates that verbal government has some predictive power, although the effect is weaker than that of prepositional government.

In order to analyze the relative importance of the predictors discussed above, a Random Forest analysis was carried out. Random Forest (Strobl et al. 2009) is a technique that creates a large number of random bootstrap samples and then for each of them makes a decision tree, which tries to predict the outcome based on the relevant predictors. The result is a "forest" of decision trees, where each tree "votes" on the relative importance of the predictors in question. Taken together, the trees in the forest give a reliable estimate of the relative importance of the relevant predictors (Baayen et al. 2013: 265).

The Random Forest analysis of the predictors discussed above returned the variable importance plot in Figure 1. The taller the bars in the bar diagram, the more important the relevant factor. As shown, the numeral is by far the most important factor, followed by predicate (in the database referred to as "PredicateLinear") and preposition. Verb government and modifiers (referred to as "ModifierLinear") appear to have a weaker effect, while the effect of noun stress is negligible.

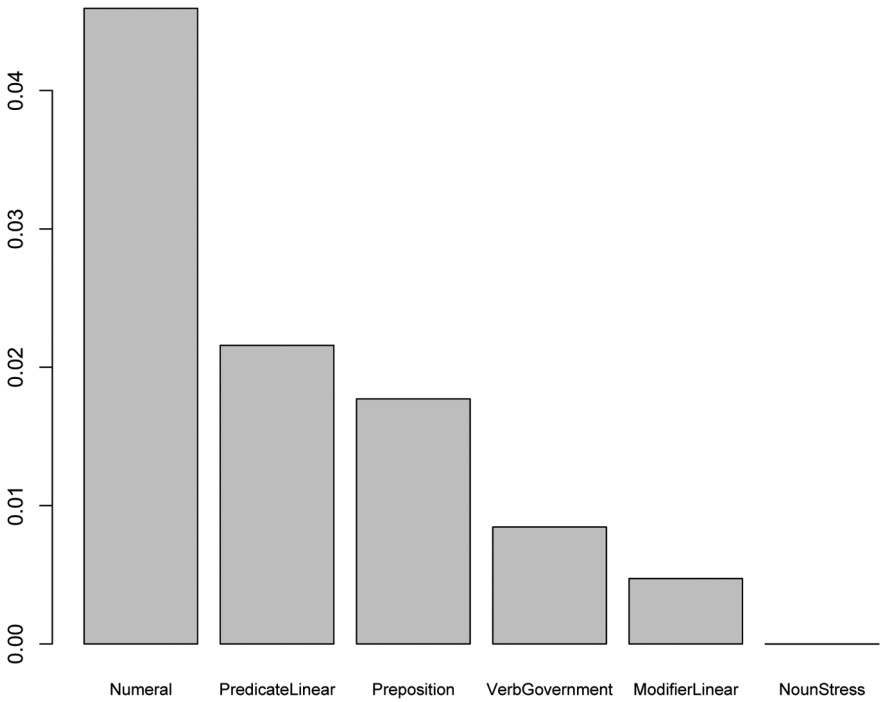

Figure 1: Variable importance plot for six potential predictors of nominative vs. genitive adjective in Russian paucal constructions

In order to find out whether word-order has an effect, another Random Forest analysis was performed. In this run, preposed predicates, modifiers, prepositions and governing verbs are included in the category "LEFTCAT", while the category "RIGHTCAT" subsumes postposed predicates, modifiers and governing verbs. (There are no postpositions in the data under analysis.) As shown in the variable importance plot in Figure 2, the analysis indicates that LEFTCAT is much more important than RIGHTCAT. In other words, preposed predictors appear more important than postposed predictors for the choice of case on adjectives in paucal constructions.

It is possible that this boils down to a frequency effect. The database contains 45,015 examples of preposed predictors, but only 8,394 postposed predictors. Since preposed predictors are more than five times as frequent in the database, it is not surprising that the Random Forest analysis found preposed predictors more valuable than postposed predictors. At the same time, in order for a predictor to be successful, it must also show a strong preference for either outcome, and it is possible that preposed predictors are not just more frequent, but also display a

clearer preference for one outcome. The Random Forest analysis leaves both options open.

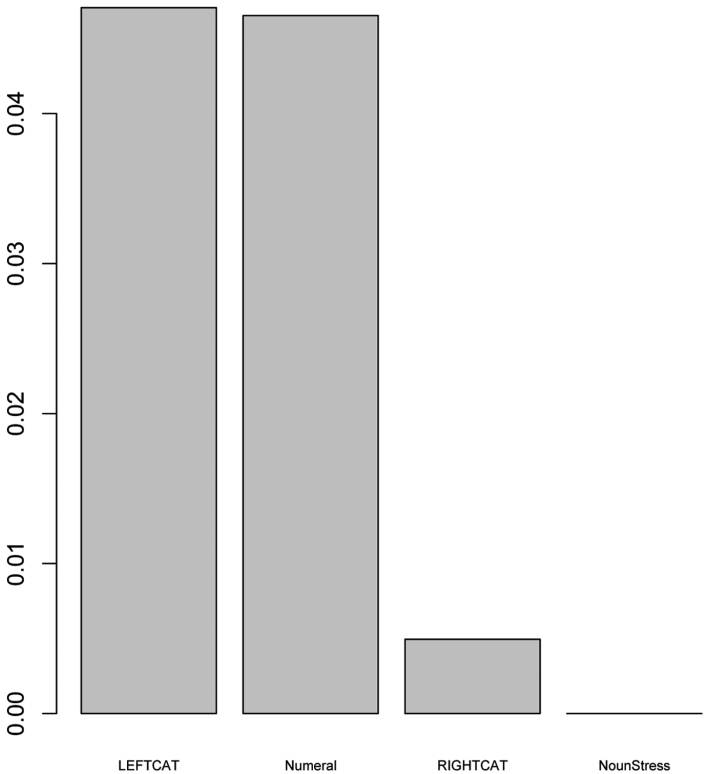

Figure 2: Variable importance plot showing the relevance of word-order for the choice between nominative and genitive adjectives in Russian paucal constructions

What can we learn from the Random Forest analysis of the RuTenTen data? First, the analysis confirms some of the conclusions from the literature on paucal numerals, e.g. Nesset's analysis of the Russian National Corpus. In particular, the two studies converge on numerals (which both studies find important) and noun stress (for which neither study indicates importance). Second, the study of the larger RuTenTen dataset suggests that more predictors have an impact, which is not surprising, since a larger dataset provides more statistical power. As we have seen, four predictors have been detected (mentioned in decreasing order of importance): predicates, prepositional government, verb government and modifiers. Finally, the analysis of the RuTenTen corpus has shown a strong word-order effect, whereby preposed predictors are much more important than predictors following the numeral phrase.

3 Interaction of predictors: CART

While the Random Forest analysis explored in the previous section enables us to assess the relative importance of predictors, it does not tell us whether the predictors favor adjectives in the genitive or nominative. In order to find out more, I carried out a CART (Classification And Regression Tree) analysis. We will see that for *dve* there is a strong preference for nominative adjectives, whereas *tri* and *četyre* display a more balanced situation where adjectives in both cases are widely used. Plural predicates and modifiers in the nominative plural favor the nominative case in the adjective, while numeral phrases governed by prepositions show a weak preference for genitive adjectives. It will furthermore be argued that nominative adjectives have better predictors, which suggests that adjectives in the nominative may outcompete genitive adjectives over time.

CART (Strobl et al. 2009) is designed for analyzing the interaction of predictors, i.e. situations where several independent variables (in the present study numeral, predicate, modifier, prepositional government, verbal government and noun stress) influence the choice between two or more dependent variables (in this study nominative vs. genitive form of the adjective). In linguistics, the use of CART was pioneered by Tagliamonte & Baayen (2012; see also Levshina 2015: 291). Baayen et al. (2013) have shown that CART performs nearly identically compared to traditional regression models, but CART has the advantage of providing tree diagrams, which makes it easy to interpret the results of the statistical analysis.

Consider the tree diagram in Figure 3, which conveys the result of a CART analysis of the RuTenTen dataset on the basis of the same six predictors that were considered in the Random Forest analysis reported on in Figure 2 in the previous section. The tree diagram contains fifteen numbered nodes. For each node, the model names the relevant predictor, and if the node is non-terminal a p-value indicating statistical significance is also provided. The model tries to predict the choice between nominative and genitive adjectives by making binary splits based on the relevant predictors. Each split represents the cleanest possible division between nominative and genitive adjectives permitted by the available information. Node 1 on the top of the diagram shows that the model first performs a split between constructions with the numeral *dve* (the left branch from node 1) on the one hand, and *tri* and *četyre* (the right branch from node 1) on the other. The fact that the first split is based on the numeral is not surprising, since the Random Forest analysis in the previous section singled out the numeral as the most important predictor.

The terminal nodes in the tree diagram are bar diagrams that represent the relative frequency of adjectives in the nominative (dark grey shading in the lower portion of the bar diagrams) and genitive (white color in the upper portion of the bar diagrams). Thus node 4 in the bottom left corner of the diagram indi-

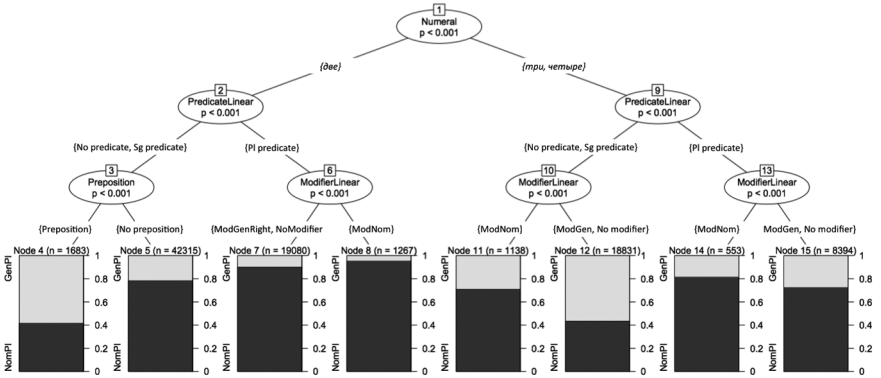

Figure 3: CART with six independent variables (numeral, predicate, modifier, prepositional government, verbal government, and noun stress)

cates that for the relevant configuration of predictors we have approximately 40 % nominative and 60 % genitive. In addition, each terminal node gives information about how many examples are covered by the node. In node 4, for instance, n = 1,683 indicates that this node covers 1,683 examples.

Let us take a closer look at the left portion of the diagram, which, as mentioned, concerns paucal constructions with the numeral *dve*. Node 2 involves the predicates and makes a split between examples with no predicate or a predicate in the singular on the one hand, and plural predicates on the other. Examples with no predicate or predicate in the singular take us to node 3, which performs a split between examples with and without a preposition. Node 4 thus portrays the situation where a construction with *dve* is governed by a preposition and involves a predicate in the singular or no predicate. For this configuration of predictors, genitive is used in 40 % of the examples in the dataset.

The neighboring terminal node, node 5, is more important – for two reasons. First, this node covers 42,315 examples, which is about 45 % of the entire dataset, and 66 % of the examples covered by the left part of the tree diagram (all nodes under node 2). The second reason why this node is particularly important is that it represents the situation where no particular predictor is available (no predicate and no preposition). In other words, node 5 is the default for constructions with *dve*.[8] As shown, in this situation nearly 80 % of the adjectives are in the nominative.

8 Notice that "default" is a notion that is used in different ways in scholarly literature (Fraser & Corbett 1997). For the purposes of the present study, the term is used in the meaning 'normal case', i.e. what applies in the absence of blocking information.

Terminal nodes 7 and 8 concern constructions with *dve* and a predicate in the plural. Both nodes have more than 80 % nominative, so it is clear that plural predicates increase the likelihood of having an adjective in the nominative. Strongest is the effect in node 8, where in addition to a plural predicate we also have a modifier in the nominative plural. Here, the proportion of adjectives in the nominative is close to 100 %. Notice, however, that this combination of predictors is not very common, since node 8 only covers 1,267 examples, whereas node 7 encompasses 19,080 examples.

When we turn to the right portion of the tree diagram, which concerns constructions with *tri* and *četyre*, it is striking that there is considerable parallelism between the left and right parts of the diagram. In both parts, the second branching is between no predicates and singular predicates to the left and plural predicates to the right. Further down, node 13 parallels node 6 insofar as both nodes involve splits between modifiers in the nominative on the one hand, and other situations on the other. The major difference between the left and the right portions of the diagram is that prepositions do not play a role in the right part; unlike node 3, node 10 involves modifiers, not prepositions.

Among the terminal nodes in the right part of the diagram, node 12 represents the default situation (neither predicate, nor modifier present) and the majority of the examples with *tri* and *četyre* (18,831 examples, i.e. 65 % of all examples under node 9). Unlike *dve*, where the default situation involves a strong preference for the nominative, *tri* and *četyre* are much closer to a "fifty-fifty situation" with no strong preference for either case in the adjective. In node 12 the genitive is used in nearly 60 % of the examples. This means that while *dve* by default has a strong preference for nominative adjectives, the situation for *tri* and *četyre* is much less clear.

Terminal node 11 shows that nominative modifiers increase the likelihood of encountering adjectives in the nominative, and terminal nodes 14 and 15 indicate that predicates in the plural have the same effect. The proportion of nominative adjectives is largest when the construction contains the combination of a plural predicate and a nominative plural modifier, as shown in node 14, which has a little more than 80 % adjectives in the nominative.

The following generalizations sum up the CART analysis of the RuTenTen dataset:

(11) a. For *dve*, the default is a strong preference for adjectives in the nominative.
b. This preference is strengthened by a predicate in the plural, in particular in the combination with a modifier in the plural.
c. When *dve* is preceded by a preposition, there is a weak preference for genitive adjectives.

(12) a. For *tri* and *četyre*, the default situation involves an almost equal likelihood of nominative and genitive adjectives.
b. When *tri* and *četyre* combine with plural predicates and/or modifiers in the nominative plural, adjectives in the nominative are preferred.

The results in (11a) and (12a) are unexpected, since earlier research has given what seems to be the opposite result. While in the present study *dve* has the strongest affinity for nominative adjectives, the investigation of data from the Russian National Corpus in Nesset (submitted) showed that *tri* and *četyre* were the leading edge of the change in the second half of the twentieth century, whereby the use of the nominative increased until it became the preferred option for all numerals around the turn of the century. A possible explanation for the apparent discrepancy between the two studies may be that the RuTenTen corpus reflects a very recent development where *∂ɐe* has surpassed *tri* and *četyre* and is now developing a stronger affinity to adjectives in the nominative. Analysis of data from the Russian National Corpus from 2000–2013 offers some support for this assumption.[9] As shown in Table 1 and Figure 4, *dve* displays a lower percentage of genitive adjectives than *tri* and *četyre* in these data, although the difference is smaller than the RuTenTen data summarized in (11) and (12). Clearly, more research is needed in order to clarify the exact role of the numeral as predictors of nominative vs. genitive adjectives in paucal constructions.

Table 1: Adjectives in the nominative vs. genitive in the Russian National Corpus from 2000 to 2013 (raw numbers, only one example per document included)

	Dve 'two'		*Tri* 'three'		*Četyre* 'four'	
	Nom	Gen	Nom	Gen	Nom	Gen
2000–2001	299	100	126	56	51	32
2002–2003	847	173	306	96	112	40
2004–2005	294	63	130	58	47	17
2006–2007	128	31	57	14	21	8
2008–2009	160	38	75	31	22	9
2010–2011	102	33	37	20	7	8
2012–2013	100	34	29	18	15	6

9 The corpus searches were carried out in April 2018. For each two-year period, I searched for a paucal numeral in the nominative followed by an adjective in the nominative or genitive followed by a feminine noun. The data were not subjected to further analysis.

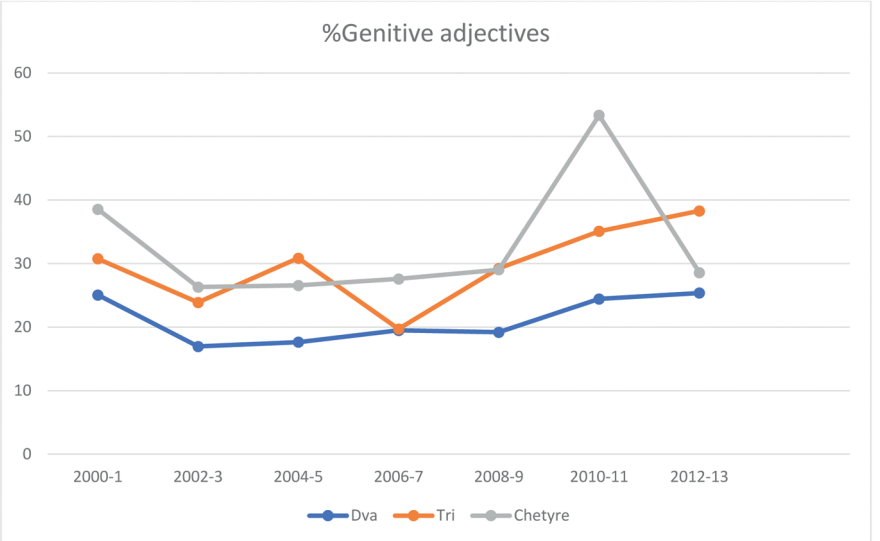

Figure 4: Adjectives in the nominative vs. genitive in the Russian National Corpus from 2000 to 2013 (percent, based on Table 1)

For the generalizations about predicates and modifiers in (11b) and (12b) the situation is that the data from the Russian National Corpus investigated in Nesset (submitted) did not give an effect for these predictors. It seems likely that the reason is simply that the dataset from the Russian National Corpus was not large enough to identify these generalizations. It is worth noting that (11b) and (12b) converge in that predicates have the same effect for all numerals. The generalization that plural predicates and nominative plural modifiers combine with adjectives in the nominative plural also makes conceptual sense, since the same grammatical features (nominative and plural) repeat themselves across the board.

Generalization (11c) concerning prepositions was not identified in Nesset's (submitted) investigation of the Russian National Corpus. Once again, the most likely explanation is that the dataset from the Russian National Corpus is too small. It is worth mentioning that earlier studies have argued that prepositions have an effect, as was discussed in section 2 above. In this sense, generalization (11c) replicates earlier findings about the relationship between prepositions and adjectives in paucal constructions.

Before we leave the CART analysis, it is interesting to compare the predictors of nominative and genitive adjectives, because the comparison illustrates the importance of two properties, which we may refer to as "generality" and "strength of prediction". Let us start with adjectives in the nominative. First, the predictors of

nominative adjectives are general in that they tend to cover large numbers of examples, and second these predictors have strong preferences for the relevant case. An example is node 5, which covers 42,315 examples and has an 80 % preference for nominative adjectives. It stands to reason that both generality and strength of preference will contribute to learnability; a strong preference for an outcome makes the predictor a good cue, while a high score on generality entails that language learners come across the relevant cue frequently (Ellis et al. 2015: 169; Ellis et al. 2016: 47). If we assume that L1 learners more readily pick up on predictors with a high score for generality and strength of preference, it follows that these predictors have better chances to survive and thrive diachronically.

Four terminal nodes in Figure 3 (nodes 5, 7, 8, and 14) display more than 80 % nominative adjectives. Taken together, these nodes cover 63,225 examples, i.e. about two thirds of the entire database. Since it is clear that predictors of nominative adjectives receive high scores for both generality and strength of prediction, the question arises as to whether genitive adjectives have as good predictors as nominative adjectives. The CART analysis carried out in the present study suggests that the answer is negative. The decision tree in Figure 3 has only two terminal nodes that show a majority of genitive adjectives, viz. nodes 4 and 12. With regard to strength of preference both nodes receive low scores, since for the relevant combinations of predictors adjectives in the genitive are only attested in about 60 % of the examples. Node 4 also receives a low score for generality since it covers only 1,683 examples (less than 2 % of the entire dataset). Even node 12, which represents 18,831 examples – more than ten times as many as node 4 – is not very general compared to the main predictor of the nominative, node 5, which covers 42,315 examples.

In sum, the CART analysis carried out in the present study indicates that adjectives in the nominative have better predictors than genitive adjectives. This entails better learnability for the nominative, which in turn suggests that the nominative is more likely to expand over time. Nesset (submitted, see also Pereltsvaig 2009) found that the use of the nominative has increased since the middle of the twentieth century. The findings of the present study suggest that this development is likely to continue in the future.

A final remark is in order. As shown in Figure 1 in section 2, the Random Forest analysis suggests that verbal government has a larger impact than modifiers on the choice between nominative and genitive on the adjective. However, verbal government is not mentioned in the decision tree of the CART analysis in Figure 3, although modifiers do play a role in the CART analysis, as mentioned in (11b) and (12b). This discrepancy is surprising and suggests that further analysis of verbal government is necessary. A possible explanation may be that verbal government scores high on generality, but low on strength of preference. I spec-

ulate that this would make Random Forest sensitive to verbal government, but that it would not be enough to include it in the decision tree in Figure 3.

4 Concluding remarks

In order to assess the value of "big data" in Russian and Slavic linguistics, I have carried out Random Forest and CART analyses of paucal constructions with feminine nouns in the web-based RuTenTen corpus, and compared the results with earlier studies, such as the parallel investigation of data from the Russian National Corpus reported in Nesset (submitted). The following three conclusions can be drawn.

First of all, the larger RuTenTen dataset enables us to identify additional generalizations that were not supported by the smaller dataset from the Russian National Corpus. In particular, the RuTenTen data indicate that plural predicates and nominative plural modifiers favor nominative adjectives, while prepositional government increases the use of the genitive case on the adjective. In addition, a word-order effect was identified, whereby preposed predictors have a stronger effect than postposed predictors. The fact that the larger RuTenTen supports additional generalizations is not surprising, since more data entails more statistical power.

A second conclusion concerns situations where the results of the present study seem to contradict earlier findings or there are discrepancies between the two statistical analyses carried out in the present study. Of particular interest are the numerals. While data from the Russian National Corpus (Nesset submitted) suggested that *tri* and *četyre* were most likely to combine with nominative adjectives, in the present study it is *dve* that shows the strongest affinity to nominative adjectives. However, this apparent discrepancy may have a diachronic explanation. As argued in section 3, although *tri* and *četyre* may have been more innovative in the second half of the twentieth century, *dve* may have surpassed them in recent years and may now be more likely to combine with nominative adjectives. However, further research on the role of numerals as predictors in paucal constructions is necessary. Further research is also required to clarify the effect of verbal government, since for this predictor there appears to be a discrepancy between the Random Forest and CART analyses carried out in the present study.

A third conclusion pertains to predictors in general. I have argued that a successful predictor must be general, i.e. cover a large class of examples, and show a strong preference for one particular outcome. On both parameters, we have seen that the predictors of nominative adjectives score higher than those of adjectives in the genitive. This suggests that the pattern with nominative adjectives is more

easily learnable, and a reasonable prediction is therefore that this pattern will continue spreading in the future and possibly oust genitive adjectives from paucal constructions with feminine nouns completely.

What can we learn about "big data" from this study? The fact that the large RuTenTen dataset made it possible to identify generalizations that could not be established in the smaller dataset from the Russian National Corpus studied in Nesset (submitted), strongly suggests that "big data" may become a valuable source of information in Russian and Slavic linguistics. Nevertheless, I will end this article with a note of caution. Large internet-based corpora like the RuTenTen corpus are not balanced, involve a certain amount of "noise", and do not provide metadata. Although "big data" are here to stay and may serve as a welcome supplement to other sources of data, web-based corpora should not be considered a replacement for balanced and deeply annotated corpora such as the Russian National Corpus.

References

Andersen, Henning. 2006. Some Thoughts on the History of Russian Numeral Syntax. In *Harvard Ukrainian Studies* 28 (1–4). 57–67.

Baayen, R. Harald, Anna Endresen, Laura A. Janda, Anastasia Makarova & Tore Nesset. 2013. Making choices in Russian: Pros and cons of statistical methods for rival forms. In *Russian Linguistics* 37 (3). 253–291.

Corbett, Greville G. 1993. The head of Russian numeral expressions. In Greville G. Corbett, Norman M. Fraser & Scott McGlashan (eds.), *Heads in Grammatical Theory*, 11–35. Cambridge: Cambridge University Press.

Ellis, Nick C., Matthew B. O'Donnell & Ute Römer. 2015. Usage-based language learning. In Brian MacWhinney & William O'Grady (eds.), *The Handbook of Language Emergence*, 163–180. Oxford: Wiley-Blackwell.

Ellis, Nick C., Ute Römer & Matthew B. O'Donnell. 2016. *Usage-based Approaches to Language Acquisition and Processing: Cognitive and Corpus Investigations of Construction Grammar*. Malden, MA: Wiley.

Fraser, Norman M. & Greville G. Corbett. 1997. Defaults in Arapesh. In *Lingua* 103. 25–57.

Gallis, Arne 1947. Tallordenes syntaks i russisk. In Erik Krag & Arne Gallis (eds.), *Festskrift til professor Olaf Broch på hans 80-årsdag: fra venner og elever*, 63–75. Oslo: Det norske videnskapsakademi.

Gorbačevič, Kirill S. 1971. *Izmenenie norm russkogo literaturnogo jazyka*. Leningrad: Prosveščenie.

Graudina, Ludmila K., Viktor A. Ickovič & Lija P. Katlinskaja. 2001. *Grammatičeskaja pravil'nost' russkoj reči*. Moskva: Astrel'.

Levshina, Natalia. 2015. *How to do linguistics with R*. Amsterdam: John Benjamins.

Lohr, Steve. 2013. The Origins of 'Big Data': An Etymological Detective Story. In *The New York Times*, February 1, 2013. (Retrieved 28 March 2018).

Madariaga, Nerea & Iván Igartua. 2017. Idiosyncratic (Dis)agreement Patterns: The Structure and Diachrony of Russian Paucal Subjects. In *Scando-Slavica* 63 (2). 99–132.

Mel'čuk, Igor A. 1985. *Poverxnostnyj sintaksis russkix čislovyx vyraženij* (= Wiener Slawistischer Almanach; Sonderband 16). Vienna: Institut für Slawistik der Universität Wien.

Nesset, Tore. submitted. *A cascading s-curves: the birth of a new paucal construction in Russian*.

Pereltsvaig, Asya. 2009. As Easy as Two, Three, Four. In *FASL* 18. 418–435.

Rozental', Detmar È. 1987. *Praktičeskaja stilistika russkogo jazyka*. 5th edition. Moskva: Vysšaja škola.

Šaronov, Igor. 2014. Narodnaja ètimologija i količestvennye sočetanija v russkom jazyke. In *Antropologičeskij forum* 21. 137–144.

Stephens-Davidowitz, Seth. 2017. *Everybody lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*. New York: HarperCollins.

Strobl, Carolin, James Malley & Gerhard Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. In *Psychological Methods* 14 (4). 323–348.

Suprun, Adam. 1957. K upotrebleniju roditel'nogo i imenitel'nogo padežej množestvennogo čisla prilagatel'nyx v sočetanijax s čislitel'nymi *dva*, *tri*, *četyre* v sovremennom russkom jazyke. In *Kirgizskij gosudarstvennyj zaočnyj pedagogičeskij institut, učenye zapiski* 3. 72–84.

Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests and trees of York English: *Was*/*were* variation as a case study for statistical practice. In *Language Variation and Change* 24 (2). 135–178.

Wade, Terence. 2011. *A comprehensive Russian grammar*. 3rd edition. Oxford: Blackwell.

Zaliznjak, Andrej A. 2002 [1967]. *Russkoe imennoe slovoizmenenie*. Moskva: Nauka.