

Rival forms of Comparatives in Russian*

Daria Kosheleva

1. Introduction

Rival forms, i.e. forms that appear to have the same function, have long been an important issue that many linguists are paying attention to. Rivalry is interesting because it has implications for inter alia variation and language change. Examples of rival forms are morphological doublets like *-ness* and *-ity* in English (Szymanek 2005, 441). In this paper I explore the rivalry between comparatives in the predicative position in Russian. My findings lend support to the “cocktail hypothesis” described in Nessel (2016), whereby the choice of one of the rival forms depends on the interplay of diverse factors. As argued in the present study, the most important factors that determine the choice of comparative forms are the length of the adjective, its frequency and the year when the text with the relevant example was published.

2. Theoretical background

In Russian, adjectives in the comparative degree can be formed in two ways: synthetic (*мягче*) and analytical (*более мягкий*). Normally synthetic comparatives are formed by adding the ending *-ee/-ей* to the stem of an adjective. To form an analytical comparative one has to use the word *более* ‘more’ plus an adjective in the positive degree.¹ There also exists a prefixed synthetic comparative formed by the means of the prefix *по-*, e.g. *подешевле* ‘cheaper’ and declined simple forms like *больший* ‘more’ and *меньший* ‘less’ (Knjazev 2007, 197). However, in the present paper I am going to focus only on the standard synthetic and analytical forms.

*The data upon which the conclusions of this work are based are available on the Tromsø Repository of Language and Linguistics (TROLLing: <http://hdl.handle.net/10037.1/10240>).

¹ Some grammars do not regard the analytical comparative as a comparative because *более* preserves its lexical meaning (Грамматика 1980(2): §1342), but I will not discuss this issue in the following.



The synthetic and analytical comparative forms are traditionally believed to be in complementary distribution. This is shown in the Table 1, which describes the distribution of the rival comparative forms in Russian according to their syntactic function. The “+” sign in the table means that the particular type of form can occur in the relevant function.

	Analytical	Synthetic
Attributive function	+	–
Appositive function	+	–
Predicative function	(+)	+

Table 1: the distribution of analytical and synthetic comparatives

Synthetic forms are used in the predicative function; analytical forms are widely used in attributive and apposition functions (Timberlake 2004, 214). Some researchers note that the distribution depend on up to three factors; in addition to syntactic function (attributive, predicative, adverbial), style of speech and frequency are claimed to be important (Graudina et al. 1976, 235). Mathiassen (1990, 105) provides the following illustrative examples:

- (1) Теперь мы живём в *более красивой* квартире, чем раньше (attributive).
- (2) Теперь мы живём в квартире, *более красивой*, чем раньше (appositive).
- (3) a. Ваша квартира *более красива*, чем наша (predicative)
 b. Ваша квартира *красивее*, чем наша (predicative)
 c. Книги *дороже* в Норвегии, чем в России (predicative).

Mathiassen’s examples suggest that the distribution is not quite complementary since analytical forms are beginning to appear in predicative position. Hence, I have included a plus sign in Table 1 in parentheses. According to Mathiassen (1990, 106), the analytical comparative is particularly likely to be used in sentences describing two properties of one person or object:

- (4) Он *более прилежен*, чем даровит.

A complicating factor is the fact that we come across the rivalry between short and long forms of adjectives in the comparative degree. Furthermore, the long form can appear in two cases: nominative and instrumental, which complicates the picture further since the number of options increases. In addition, certain adjectives have problems forming synthetic comparatives due to phonological reasons. This is the case with adjectives where the stem ends in a velar consonant, e. g. *вязче* – *более вязкий* (7 vs. 25 entries in RNC), *чутче* – *более чуткий* (19 vs. 112 entries in RNC) etc. (Graudina et al. 1976, 234). Another potentially relevant factor could be the length of the adjective:

- (5) Чем *дороже* чай, тем *более качественное* сырьё используется для его изготовления — будь он хоть в пакетике, хоть в баночке. [Ольга Цыбульская. Чай без церемоний // «Русский репортер», № 15 (143), 22–29 апреля 2010, 2010]

In example (5) taken from RNC the length of adjective may play a role in the choice of analytical or synthetic form. As can be seen from the example, the adjective *дорогой*, which consists of 3 syllables, is used in the synthetic form, whereas the adjective *качественный* has 4 syllables and is used in the analytical form.

Perhaps, frequency also plays a role in the distribution of forms since frequency has been shown to have a strong impact on the distribution of morphological categories in many languages (see Bybee 1985). As mentioned above, Graudina et al. (1976, 235) have suggested that frequency is relevant for the choice of comparative form in Russian.

My research questions are the following: what are the reasons for the analytical comparative to be used in the predicative function, and to what extent and in what cases is it actually used? A further question is whether it is more widely used in the past 20 years or whether the percentage is the same as in the beginning of the 20th century. The possible null hypothesis is that all the factors mentioned above matter, and that the analytical comparative becomes more and more widely used in the course of time

In this study I will use the terminology of Haspelmath and Buchholz (1998, 279). Let us consider an example:

- (6) Все вместе они были красивее, чем каждая в отдельности. [Алексей Варламов. Купавна // «Новый Мир», 2000]

In (6) *все вместе они* is a “compare” (CMP), *красивее* is a “parameter” (PARA), *чем* is a “standard marker” (STM) and *каждая в отдельности* is a “standard” (STAN).

According to Stolz (2013, 64), Russian is unique among the Slavic languages since in the standard marker position (in the appositive construction) the historical instrumental form *чем* of the interrogative pronoun *что* is used. Czech and Slovak use *než*; the other Slavic languages use *od*, *ot* or *vid* which mean ‘from’.

Except for Czech, Bulgarian and Macedonian, the Slavic languages can form more than one type of comparative constructions (Stolz 2013, 67). In Russian there is a rivalry between standard marker *чем* and morphological marking of the genitive case on the NP that is used as a standard of comparison (Stolz 2013, 68):

- (7) Больше всего понравился ему Кавказ — эта прекрасная южная местность была красивее и величественнее Швейцарии. [Э. Г. Казакевич. Звезда (1946)]

In this sentence, the standard *Швейцарии* occurs in the genitive case.

The type of standard marker is to some extent connected with the choice between analytic and synthetic comparative (Stolz 2013, 68). With analytic comparative one can use only the *чем*-construction. If one uses the synthetic form, on the other hand, both options (synthetic and analytical comparative) are possible. In the corpus of the Russian translation of *Harry Potter* most constructions with *чем* involved the synthetic form of comparatives (9 cases):

- (8) Гарри оказался быстрее, чем Хиггс.
 (9) более храбрый человек, чем дядя Вернон.

Apart from the standard marker *чем*, in Russian comparative constructions we can use *нежели*. It is claimed that in the nineteenth century there were also the double markers *нежели чем* and *чем нежели* (Rakhilina 2015, 316). See the following examples from the same article:

- (10) Что же, — говорю, — это гораздо разумнейше, *нежели чем* совсем плешкой остаться. [Н. С. Лесков. Заячий ремиз (1894)]
- (11) Поклоны «от головы до сырой земли», «жив и здоров, чего и вам желаю», а где-то невзначай, неуместно вплетется писакой: «Скучно, *чем нежели* допреже одному без семьи жить... » [К.С. Петров-Водкин. Моя повесть. Часть 1. Хлыновск (1930)]

Nowadays the double markers are not used anymore.

Rakhilina (2015, 314) identifies a total of five possible types of comparative constructions in modern Russian whose distribution seems not quite clear:

- (12) a. Прибыль от разрушения *заметнее прибыли* от строительства.
 b. Прибыль от разрушения *заметнее, чем* прибыль от строительства.
 c. Прибыль от разрушения *заметнее, нежели* прибыль от строительства.
 d. Прибыль от разрушения *более заметна, чем* прибыль от строительства
 e. Прибыль от разрушения *более заметна, нежели* прибыль от строительства.

However, it seems that the distribution of comparatives is even more complex. Not only is there a rivalry between analytic and synthetic forms, analytic comparative constructions have several options for the choice of the comparative: it can also be a long form like *более заметный* or a short one as in (11d). Furthermore, a long form can be used in two cases: the nominative — *более заметный* — and the instrumental — *более заметным*.

3. Data

For the purposes of this study I decided to look at constructions with a noun or a pronoun in the nominative case, the verb *быть* in the past tense followed by a synthetic or analytic form of an adjective in the comparative degree. This facilitates an empirical investigation of the comparative constructions in the predicative position, where we expect a rivalry between analytical and synthetic forms. I made 3 search queries in the Russian National Corpus:

- (13) a. Constructions with analytic form of comparatives, long form
 b. Constructions with analytic form of comparatives, short form
 c. Constructions with synthetic form of comparatives

For the first two queries I received 377 and 319 examples, respectively, so I analyzed all of them. However, it was not so easy with synthetic comparatives. According to the query “(S|SPRO), nom/быть praet/A, comp” there are 18,704 contexts in the main corpus. Since this sample is too large to annotate manually, I decided to estimate the amount of noise in

the sample. It seems that most of the noise is due to the adverb *уже* that is interpreted by the parser as a synthetic form of the adjective *узкий*. Removing *уже* (“-уже” in the query form) gives 8, 941 contexts or 47.8% of the initial sample. Another problem is that in many cases it is difficult to judge whether we are dealing with an adjective or an adverb. Most of these examples start with *это было: это было раньше; это было нужнее* etc. There are 773 examples (4.1%) of this construction in the corpora (the query is “это/было/ -уже A, comp”).

The last evident problem is infinitives as subjects like in the example *устраиваться на работу было трудней*. Unfortunately, there is no obvious way of getting rid of such examples. The total number of synthetic constructions in RNC can be estimated as roughly 8,000 showing that the synthetic construction is still the most widespread type of comparative in predicative position

Another method for estimating the remaining noise is to annotate a random sample of 100 examples manually and then extrapolate the result to the entire whole corpus. In this case, I found 42 examples with the adverb *уже*, 3 more of *это было* and 6 noisy examples of other types:

- (14) a. The adverbial particle *скорее*
- b. A descriptive adjective of another noun like *блестящей идеей, пиццущей машинкой*
- c. The adverb *раньше*;
- d. A misinterpretation of the adjective *общий* in the instrumental case

So, the percentage of synthetic constructions is 49%, which makes 9,165 examples in total.

In order to arrive at a manageable database with as little noise as possible I decided to look at the examples from the disambiguated part of RNC. There are expectedly far less examples — just 129 — but at least they are quite reliable: just 2 examples were considered noise.

Most examples in the first two samples (of long and short forms of analytic comparatives) match the queries; we had only few examples that can be considered noise

- (15) *Песни, марши, кинокадры — все это было более вещественной материей, нежели вещество еды, черепица крыши, ситец [...]*

There seem to be more examples with the instrumental case than nominative:

- (16) a. *Если бы они были более благоразумными, более сдержанными <...> [Л. И. Шестов. Шекспир и его критик Брандес (1898)]*
- b. *<...> у которых клеточки были более короткие и толстостенные. [К. А. Тимирязев и др. Жизнь растения (1878)].*

The examples I extracted were organised in a database. The database contains the following information:

- (17) a. Type of comparative (synthetic, analytical)
 b. Short or long form and case (nominative or instrumental) for analytical comparatives
 c. The year of publication of the document that contains the example
 d. The standard marker (*чем, нежели* or the standard in the genitive case)
 e. Whether the comparison is made for 2 different objects (*X более заметный, чем Y*) or 2 qualities of the same object (*X более заметный, чем яркий*)
 f. The length of the adjective in positive degree measured in syllables
 g. The frequency of the adjective taken from Lyashevskaya, Sharov (2009)
 h. Whether there are 2 or more comparatives in a row
 i. The example itself

	Synthetic	Analytical long	Analytical short
Total	129	374	316
Relevant	127	298	309
<i>Čem/Neželi/Gen</i>	15/1/70	62/9	81/13
Frequency (ipm)	36,5 – 944,4	14,8 – 483,3	14,9 – 483,3
Freq not attested	2	98	126
Sequences	6	53	37

Table 2: summary of the data analyzed

The summary of the data is presented in Table 2. The columns correspond to the different types of comparatives: synthetic, analytical with long form of adjective and analytical with short form of adjective. The row “Total” represents the number of examples for each of 3 queries; the row “Relevant” corresponds to the number of examples without noise. “*Čem/Neželi/Gen*” shows the number of examples with standard markers *чем, нежели* or the standard in the genitive case (if the standard of comparison is in the genitive, no standard marker is used) respectively. Frequency is the range of the frequencies of the adjectives taken from the frequency dictionary mentioned above. For some adjectives the frequency is not attested there, so the number of such adjectives is represented in the row “Freq not attested”. If there is more than 1 comparative in one example following each other, they are counted as sequences

3. Statistics

I decided to carry out a CART analysis based on the data described earlier. CART or conditional inference trees are “a method for regression and classification based on binary recursive partitioning” (Levshina 2015, 291). The predicted variable is the type of comparative (synthetic, analytical long or analytical short); the independent variables are the year, the standard marker, the length of the adjective, the frequency and the sequence. In the analysis some of the terms could demand more clarification. Length is a parameter determining the length of the adjective in positive degree in nominative case; it is a

number. Long and short (lng and shrt in the bar graphs respectively) forms refer to 2 possible forms of adjectives; it is a Boolean parameter.

The results are presented as a decision tree in Figure 1. As we can see, the most important factor is the length of the adjective, which appears in nodes 1 and 7. If the number of the syllables is 2 or less, then the year is important, as shown in nodes 2 and 3. For trisyllabic adjectives frequency plays, while for longer adjectives the only relevant factor is length.

The terminal nodes in the decision tree show the distribution of the three types of comparatives for different combinations of independent variables. As indicated in node 5, the synthetic comparative is likely to be chosen if the year is between 1959 and 1987 and the number of the syllables is not more than two. Long analytical forms have less special preferences: if the adjective is short and the text it has appeared in was created recently (after 1987) or if the adjective contains 3 syllables and is quite infrequent (less than 55 ipm), then the long analytical form is preferred. Short analytical forms tend to be used if the adjective has 3 or more syllables, but they are also quite widespread for shorter adjectives before 1959.

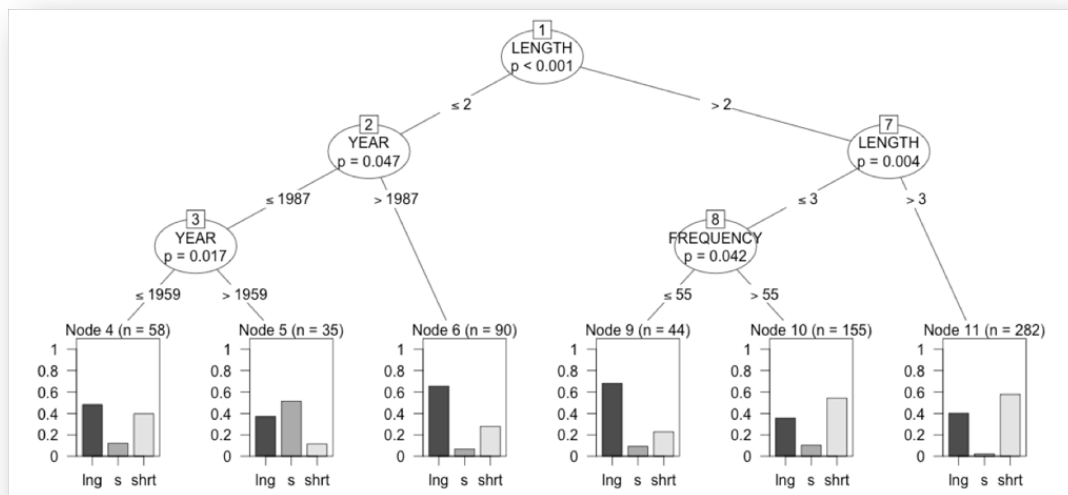


Figure 1: CART analysis

Although the decision tree in Figure 1 testifies to the importance to the length of the adjective, I also carried out a Random Forest analysis in order to investigate the relative importance of the factors under scrutiny. The plot in Figure 2 represents the scale of importance of the factors I used in the analysis. As expected, the most important is the length of the adjective (in syllables); frequency and year also still quite important, while the sequence and the standard marker are of little importance.

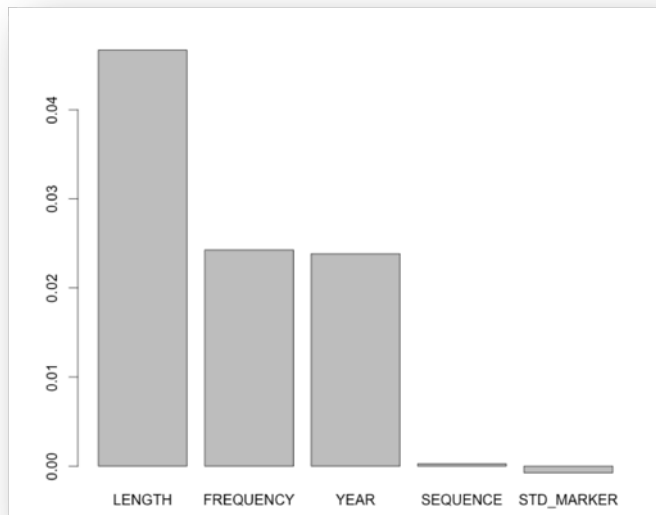


Figure 2: Random forest analysis

5. Conclusion

In this paper I have studied rival Russian comparatives in the predicative function. My analysis shows that the choice between analytical and synthetic comparatives is based on a number of factors such as length, frequency and year of publication. This is in line with the “cocktail hypothesis” (Nesset 2016), according to which rivalry tends to depend on the interplay of several factors. The most important factor in my study is the length of the adjective in positive degree in the nominative case. However, it is important to mention that there are some adjectives that show variation (e.g. adjectives of size like *высокий, глубокий, длинный, короткий, мелкий, низкий, тонкий, широкий*): they have no preference for just one construction. In order to shed light on such adjectives, we need to take into account a detailed classification of the semantic properties of individual adjectives. However, this task is left for future research

References

- Bybee, Joan L. 1985, *Morphology: A Study of the Relation between Meaning and Form*, vol. 9, Amsterdam/Philadelphia.
- Graudina, L. K., Ickovič, B. A., Katlinskaja, L. P. 1976, *Grammatičeskaja pravil'nost' russkoj reči: opyt častotno-stilističeskogo slovarja variantov*. Moscow.
- Haspelmath, Martin and Oda Buchholz 1998, *Equative and Similitive Constructions in the Languages of Europe*. Berlin and New York. <https://doi.org/10.1515/9783110802610.277>
- Knjazev, Ju. P. 2007, *Grammatičeskaja semantika: Russkij jazyk v tipologičeskoj perspective*, Moscow.
- Levshina, Natalia 2015, *How to do Linguistics with R: Data Exploration and Statistical Analysis*, Amsterdam/Philadelphia.
- Lyashevskaya, O. N., Šarov, S. A. 2009, *Novyj častotnyj slovar' russkoj leksiki*, URL: <http://dict.ruslang.ru/freq.php>.
- Mathiassen, Terje 1990, *Russisk grammatikk*, Oslo.

- Neset, Tore 2016, *Russiske rivaler: Cocktailhypotesen*, Festschrift article to appear.
- Raxilina, E. V. 2015, Stepeni sravnenija v svete russkoj grammatiki ošibok. *Trudy Instituta russkogo jazyka im. V. V. Vinogradov* 6, 310–333.
- Stolz, Thomas. 2013, *Competing Comparative Constructions in Europe*, vol. 13. Berlin and New York. <https://doi.org/10.1524/9783050064994>
- Švedova, N. Ju. 1980, *Russkaja grammatika: Fonetika. Fonologija. Udarenie. Intonacija. Slovoobrazovanie. Morfologija*, Moscow.
- Szymanek, Bogdan 2005, The Latest Trends in English Word-Formation, *Handbook of Word-Formation*. Dordrecht, 429-448.
- Timberlake, Alan 2004, *A Reference Grammar of Russian*, Cambridge.

author: Daria Kosheleva

affiliation: National Research University – Higher School of Economics

email: koshelevada@gmail.com