Russian Linguistics

The long and the short of it: Russian predicate adjectives with zero copula

--Manuscript Draft--

Manuscript Number:	
Full Title:	The long and the short of it: Russian predicate adjectives with zero copula
Article Type:	Original Paper
Funding Information:	
Abstract:	The present article presents an empirical investigation of the choice between so-called long (e.g., prostoj ‘simple’) and short forms (e.g., prost ‘simple’) of predicate adjectives in Russian based on data from the syntactic subcorpus of the Russian National Corpus. The data under scrutiny suggest that short forms represent the dominant option for predicate adjectives. It is proposed that long forms are descriptions of thematic participants in sentences with no complement, while short forms may take complements and describe both participants (thematic and rhematic) and situations. Within the “space of competition” where both long and short forms are well attested, it is argued that the choice of form to some extent depends on subject type, gender/number, and frequency. On the methodological level, the approach adopted in the present study may be extended to other cases of competition in morphosyntax. It is suggested that one should first “peel off” contexts where (nearly) categorical rules are at work, before one undertakes a statistical analysis of the “space of competition”.
Corresponding Author:	Tore Nesset UiT The Arctic University of Norway: UiT Norges arktiske universitet Tromso, NORWAY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	UiT The Arctic University of Norway: UiT Norges arktiske universitet
Corresponding Author's Secondary Institution:	
First Author:	Tore Nesset
First Author Secondary Information:	
Order of Authors:	Tore Nesset
	Laura A. Janda
Order of Authors Secondary Information:	
Author Comments:	A good reviewer for this manuscript might be Professor Masako Fidler, Brown University. She has published an important monograph on short/long forms in Russian under her earlier name Masako Ueda. This work is cited in the manuscript. We mention this since the editors might not be aware of the name change.

Title:
The long and the short of it: Russian predicate adjectives with zero copula

Authors:
Tore Nesset, UiT The Arctic University of Norway
Laura A. Janda, UiT The Arctic University of Norway

Corresponding author:
Tore Nesset
Tore.nesset@uit.no
+47 97641714 (cellphone)
+47 77645633 (office phone)

The long and the short of it: Russian predicate adjectives with zero copula

Anonymous authors

Abstract: The present article presents an empirical investigation of the choice between so-called long (e.g., *prostoj* 'simple') and short forms (e.g., *prost* 'simple') of predicate adjectives in Russian based on data from the syntactic subcorpus of the Russian National Corpus. The data under scrutiny suggest that short forms represent the dominant option for predicate adjectives. It is proposed that long forms are descriptions of thematic participants in sentences with no complement, while short forms may take complements and describe both participants (thematic and rhematic) and situations. Within the "space of competition" where both long and short forms are well attested, it is argued that the choice of form to some extent depends on subject type, gender/number, and frequency. On the methodological level, the approach adopted in the present study may be extended to other cases of competition in morphosyntax. It is suggested that one should first "peel off" contexts where (nearly) categorical rules are at work, before one undertakes a statistical analysis of the "space of competition".

Keywords: Russian adjectives, short forms, long forms, corpus data, regression analysis

1. The problem

A classic problem in Russian morphosyntax concerns the choice between so-called long forms and short forms of predicate adjectives. The examples in (1) and (2), which all involve the adjective *prostoj* 'simple', show that both forms occur in similar contexts:[1]

(1) a. Kriterii očen' **prostye**$_{\mathbf{LF}}$.
'The criteria are very simple.'
b. Kriterii očen' **prosty**$_{\mathbf{SF}}$, i ix vsego dva.
'The criteria are very simple, and there are only two of them.'

(2) a. Otvet **prostoj**$_{\mathbf{LF}}$: v obespečenii intellektual'nyx sverxvozmožnostej važnejšuju rol' igraet aktivacija opredelennyx, a verojatno, i mnogix mozgovyx struktur.
'The answer is simple: for intellectual superopportunities the activation of certain and, presumably, numerous brain structures play a major role.'
b. Otvet **prost**$_{\mathbf{SF}}$: zdanija-pamjatniki raspoloženy na samyx kommerčeski privlekatel'nyx territorijax goroda.
'The answer is simple: the historic buildings are situated in the commercially most attractive parts of town.'

In Russian, predicate adjectives occur in sentences with full verbs (e.g., verbs of motion) or copula verbs (e.g., *byt'* 'be' and *stat'* 'become'). A copula verb may be overt (e.g., *byl* 'was (masculine singular)' or covert (so-called zero copula), as in examples (1) and (2). In the present study, we

[1] Unless otherwise indicated, all numbered examples are from the syntactic subcorpus of the Russian National Corpus, which is available here: https://ruscorpora.ru/new/search-syntax.html. Examples are given in transliterated orthography. For the convenience of the reader, in each example the predicate adjective is boldfaced and supplied with a subscript LF for "long form" and SF for "short form".

limit ourselves to sentences with zero copula, where the rivalry is between nominative forms of the predicative adjective. In sentences with overt verbs, the predicative adjective may also occur in the instrumental.[2] We follow traditional terminology and refer to forms like *prost* 'simple' (feminine singular: *prosta*, neuter singular: *prosto*, plural: *prosty*) as "short forms" (SF), while forms like *prostoj* 'simple' (feminine singular: *prostaja*, neuter singular: *prostoe*, plural: *prostye*) are called "long forms" (LF), since their agreement endings are longer than those of the short forms.

The rivalry between long and short forms of Russian adjectives has received considerable attention in the scholarly literature. We may distinguish between two traditions. First, in a "semantic" or "theoretical" tradition, linguists of various theoretical persuasions have tried to pinpoint semantic differences between long and short forms (e.g., Peškovskij 1933, Vinogradov 1947, Švedova 1952, Timberlake 2004, and Fonnes 2013) and incorporated their ideas in various theoretical frameworks. For instance, in a number of seminal works Babby (1973, 2009, 2013) has analyzed the rivalry between long and short forms from the perspective of generative grammar. A second and more empirically oriented tradition involves scholars focusing on contextual factors that motivate the choice between the two forms, often with quantitative investigations of individual factors (e.g., Iversen 1978, Gustavsson 1976, Nichols 1981, and Ueda 1992). The two traditions can be brought together in one research question: is the choice between long and short forms motivated by the speakers' desire to express different meanings, or is the choice determined by the context in which the long and short forms occur?

In order to address this research question, we apply the following methodology in four steps:

(3) Methodology:
  a. Establish a database of examples culled from the Russian National Corpus (syntactic subcorpus)
  b. Identify (nearly) categorical rules whereby the choice of form is determined by the context
  c. Pinpoint a "space of competition", i.e., a set of contexts where both forms are possible
  d. Analyze statistically the interplay of contextual factors within the space of competition.

The idea behind this methodology is that semantic differences between long and short forms should preferably be investigated in contexts where both forms occur freely. We therefore propose to first "peel off" contexts that allow only one of the forms, and then explore the statistical competition between factors within the remaining "space of competition". If in this

[2] Predicate adjectives in the instrumental case are marginally attested in sentences without an overt copula verb, typically in elliptic constructions where an overt copula (e.g., the past tense form *bylo* 'was') can be reconstructed from the context:

(i) I želanie obresti vse èto bylo neobyknovenno **sil'nym**$_{\textbf{INS}}$. Takim že, navernoe, **sil'nym**$_{\textbf{INS}}$, kak želanie imet' kvartiru i krasivuju mašinu.
'And the desire to acquire all this was unusually strong. Apparently, as strong as the desire to have an apartment and a nice car.'

Here the instrumental form *sil'nym* 'strong' first occurs with the past tense copula *bylo* 'was' and then is repeated in the next sentence, where the copula is implicit. In our dataset, we have 16 examples of this type, which amounts to less than 0.5% of all examples with zero copula in our dataset. We will not discuss instrumental forms in the present study.

space there are strong statistical tendencies that motivate the choice of form in certain contexts, there is little room for the expression of different meanings by means of the two competing forms. If, on the other hand, there are no strong contextual factors at work, it is likely that speakers use the two forms to express different meanings.

The contribution of our study can be summarized as follows. First, our data indicate that short forms are the dominant option for predicate adjectives. In the dataset as a whole, short forms account for 89% of the data, while in the "space of competition" we have 67% short forms. Second, our investigation confirms that there are several contexts where the use of the short form is a (nearly) categorical rule: (a) sentences with complements, (b) impersonal constructions with clausal, infinitival, or no subject, (c) sentences with certain "non-personal" pronouns as subjects, and (d) sentences where the predicate adjective comes before the subject and the subject is realized as a noun phrase or a personal pronoun. Third, we advance the "Situation/Participant Hypothesis", whereby long forms are descriptions of thematic participants, while short forms are less restrictive and can describe both participants (thematic and rhematic) and situations. Fourth, we demonstrate that both long and short forms are well attested in a "space of competition" consisting of sentences where the predicative adjective (a) has no complement, and (b) describes a thematic participant. Fifth, we argue that within the "space of competition" the choice between long and short form depends on subject type, gender/number, and frequency. We identify the following statistical tendencies: sentences with (overt) nouns and pronouns as subjects favor short forms, a tendency that is stronger if the noun combines with an adjectival or similar modifier. In elliptic sentences without a subject, long forms are preferred. The likelihood of short forms increases with increasing frequency; for hapaxes, we predict free variation between long and short forms, while for high frequent adjectives we predict short forms in more than 80% of cases. Sixth, the statistical tendencies we identify do not involve particularly strong predictors. This, we argue, indicates that both long and short forms are used with few limitations within the "space of competition". This may suggest that the long and short forms are recruited to convey different meanings within this space, but our finding is also compatible with a situation whereby the two forms are stylistically different or are in the process of undergoing diachronic change. Last but not least, with regard to frequency, our study reveals a "locality effect" since the "local" measure of frequency in predicative position proves a better predictor than "global" frequency in the corpus as a whole.

Our argument is structured as follows. In section 2, we provide an overview of relevant scholarly literature and formulate a hypothesis about relevant predictors of the choice of adjective form. After a presentation of the data in section 3, we discuss (nearly) categorical rules in section 4, before we undertake a statistical analysis of the "space of competition" in section 5. Our findings are summarized in the concluding section 6.

2. Short forms: previous scholarship and hypothesis

2.1 Complement

Predicative adjectives combine with complements, which may be finite clauses (4), infinitive constructions (5), prepositional phrases (6), or noun phrases in the genitive (7), dative (8) or instrumental case (9):

(4) Ja, konečno, **vinovata**$_{SF}$, čto ne pošla v partizany...
‘I am, of course, guilty for not becoming a partisan.’

(5) Ona **sposobna**$_{SF}$ vyražat’ èmocii.
‘She is able to express emotions.’

(6) Oni ne **soglasny**$_{SF}$ so srokom […].
‘They don’t agree on the deadline […].’

(7) Opisanija Marko Polo **polny**$_{SF}$ netočnostej.
‘Marco Polo’s descriptions are full of inexactitudes.’

(8) Ja očen’ **blagodarna**$_{SF}$ Vladimiru Švarcmanu i Saše Čačko za iskrennee želanie pomoč’.
‘I am very grateful to Vladimir Švarcman and Saša Čačko for their sincere wish to help.’

(9) Čem on tak **gord**$_{SF}$?
‘What is he so proud of?’

A number of researchers have suggested that the presence of a complement promotes the use of the short form. For instance, Nichols (1981: 309) states that “it is well known that a complement or other dependent on the adjective is associated with the short form” (see also Benson 1959: 94, Børresen 1966: 150, Gustavsson 1976: 179, and Ueda 1992: 141–141 and 149). This generalization is also frequently mentioned in grammars and handbooks (e.g., Isačenko 1962: 150, Mathiassen 1996: 95–96, and Wade 2011: 191–192).

While the tendency for short forms to be used with complements is often mentioned, it is not clear whether this tendency is equally strong for all kinds of complements. For instance, Iversen (1978: 31), who studied data from texts from the 20th century (fiction and non-fiction), found that long forms occasionally occurred in sentences where the complement was a prepositional phrase. In order to find out more about the relationship between complements and the choice between long and short form, we included this factor in our analysis.

2.2 Subject

It is sometimes mentioned that different realizations of the subject may influence the choice between long and short form of the predicate adjective. In particular, pronominal subjects such as *èto* ‘this (neuter singular)’ and *vsë* ‘all (neuter singular)’ are said to promote the short form (e.g., Babby 2009: 105, 2013: 76, Benson 1959: 94, Gustavsson 1976: 310, Mathiassen 1996: 97, Ueda 1992: 149):

(10) a. Èto prosto **smešno**$_{SF}$.
‘This is simply ridiculous.’
b. Vsë **prozračno**$_{SF}$.
‘Everything is transparent.’

Some researchers have regarded the short form as the only option in sentences with *èto* and *vsë*, but Iversen (1978: 65) argued that we are instead dealing with a mere statistical tendency. Exactly which pronouns promote the use of the short form is not clear from the literature. For the purposes of the present study, we will draw a distinction between personal pronouns and other pronouns, such as *èto* and *vsë*, which we will refer to as “non-personal”. We return to this distinction in section 4.3.

Noun phrases as subjects have received less attention than pronouns. However, in a thorough empirical study, Ueda (1992: 149) found that the presence of an adjectival modifier in the subject made the use of the short form more likely in the predicate. Although Ueda's finding concerned sentences with an overt copula verb, for the purposes of our analysis we decided to distinguish between modified (11a) and unmodified nouns (11b) as subjects:[3]

(11) a. Svežij vozdux očen' **važen**$_{\text{SF}}$.
'Fresh air is very important.'
b. Situacija v zdravoxranenii **trudnaja**$_{\text{LF}}$.
'The situation in healthcare is difficult.'

Our database contains examples of the following types:

(12) a. **Estestvenno**$_{\text{SF}}$, čto èto rešenie k ožidaemomu rezul'tatu ne privelo.
'It is natural that this decision did not lead to the expected result.'
b. **Glupo**$_{\text{SF}}$ lgat', kogda tebja tak legko uličit'!
'Lying is stupid when it is so easy to catch you!'

In the corpus, examples like (12a) are classified as having finite clause as a subject, while the infinitive construction in (12b) is analyzed as the subject of the sentence. In traditional grammars, sentences of these types are often referred to as "impersonal constructions" without subjects, and some analysts classify the predicates as adverbs, "predicatives", or "category of state" (Russian: *kategorija sostojanija*), rather than as adjectives (Vinogradov 1947: 399–401, Isačenko 1962: 13 and 194–196, Švedova ed. 1980: 215). We will not pursue these questions here. The focus of our investigation is the distribution of forms (long or short) in sentences like (12a-b).

2.3 Word order

Word order, which in Russian is closely related to information structure, has been claimed to have an impact on the choice of adjective form. More specifically, many scholars have argued that the short form is preferred in sentences where the predicate adjective precedes the subject (e.g., Gustavsson 1976: 376, Isačenko 1962: 149, Mathiassen 1996: 97–98, Rozental' and Telenkova 1974: 42–43, Pereltsvaig 2001: 214–215 for discussion):

(13) **Izvestno**$_{\text{SF}}$ imja ego otca – Rejner Fos.
'The name of his father is well known – Reiner Foss.'

We included word order in our analysis in order to find out more about the relationship between word order and the choice between long and short forms.

2.4 Gender and number

We decided to include the gender and number of the predicate adjective as a factor in our analysis, although these categories have not received much attention in the scholarly literature on long and short forms of adjectives. However, Gustavsson (1976: 119) observed that the short form "seems to be avoided in the neuter". Iversen (1978: 59) reported a similar finding; in his dataset the short form was more likely to occur in masculine singular than in the feminine or

[3] Ueda (1992: 149) also made a distinction between proper names and common nouns as subjects, but we had too few examples of proper names to include this factor in our analysis. Fonnes (2013: 130–189) also considered a wide variety of subject types and found weak effects on the choice between long and short forms. His classification is too fine-grained for quantitative analysis of our relatively small dataset.

neuter singular. Fonnes (2013: 254) also found fewer short forms in combination with neuter nouns.

2.5 Frequency

To the best of our knowledge, no one has carried out a systematic investigation of the relationship between the frequency of the adjective and the choice between long and short form. However, frequency has been shown to be relevant for other cases where there is competition between (nearly) synonymous forms. For instance, in a recent study of the competition between singular and plural agreement in Russian sentences with quantified subjects, frequency was shown to play an important role (Nesset and Janda 2023).

Many researchers have claimed that the rivalry between long and short forms reflects language change, whereby the use of the short form is decreasing (see Pereltsvaig 2001: 224–225 for discussion and references). If we accept this view, it seems reasonable to expect high frequent adjectives to be more likely to occur in the short form, since high frequency tends to inhibit analogical change (Bybee 2007), such as the replacement of a form (in our case the short form) by a putatively more productive form (in our case the long form). In view of this, we decided to include frequency in our analysis.

2.6 Other factors not tested in the present study

A classic hypothesis about predicate adjectives in Russian is that the long form expresses permanent characteristics, whereas the short form is used about temporary properties (e.g., Peškovskij 1933: 221, Vinogradov 1947: 270). However, already Švedova (1952: 85) questioned the validity of this generalization as a strict rule in Contemporary Standard Russian, and most commentators following her have cited the idea with caution (see Fonnes 2013: 222–223 for critical discussion). For instance, Benson (1959: 90) first stated that the distinction between permanent and temporary properties "was found often not to be valid" by his informants, but then adds that "in certain instances such a distinction does exist". Commenting on the distinction, Timberlake (2004: 292) admitted that there "is a certain truth to this" and proposed a modified version of the hypothesis whereby the short form "means not so much that the state is literally temporary as that it is contingent and therefore potentially variable". Whatever its merits, the hypothesis about permanent vs. temporary properties is not suitable for test against a large dataset of corpus examples, since it cannot be established from corpus examples whether the language users had permanent or temporary properties in mind. Therefore, this hypothesis will not be explored in the present study.

Another influential idea is that the choice between long and short forms depends on stylistic factors. Gustavsson (1976: 89–133) devoted an entire chapter to this idea, arguing that the use of the short form increases "as the degree of bookishness of the lexeme increases". A similar conclusion is drawn by Fonnes (2013: 205), who also devoted a whole chapter to stylistic factors. This is a well-known hypothesis, but since the syntactic subcorpus of the Russian National Corpus does not include metadata about style or other socio-linguistic variables, this hypothesis cannot be tested in the present study.

Fonnes (2013: 250) advances the hypothesis that long forms of predicate adjectives have a meaning akin to so-called restrictive relative clauses, while short forms have a "non-restrictive function". This analysis appears somewhat similar to the analysis proposed by Babby (2009 and 2010) in a generative framework. Babby analyzes long forms of predicative adjectives as modifiers of an empty head. These subtle nuances depend on the intentions of the language users, which cannot be read out of corpus examples. Therefore, these ideas will not be pursued in the following.

As mentioned in the previous subsection, it has been claimed that the distribution of the long and short forms has changed over time. However, unlike other parts of the Russian National Corpus, the syntactic subcorpus does not include information about when the examples were created. It is therefore not possible to investigate diachronic change in the present study.

Although the factors of meaning, style and language change cannot be tested directly in the present study, our investigation nevertheless contributes indirectly to the study of these factors. Through our quantitative study of contextual factors, we establish a “space of competition”, where both forms are well attested. This space forms a good basis for the future study of the impact of meaning, style and language change on the choice between long and short forms.

2.7 Hypothesis

On the basis of the scholarly literature reviewed in sections 2.1 through 2.6, we propose the following hypothesis:

(14) The choice between long and short forms of predicative adjectives depends on the following factors (values expected to promote the use of the short form in parentheses):
  a. Complement (presence of complement)
  b. Subject (clauses, infinitives, non-personal pronouns, and modified nouns)
  c. Word order (Adjective before subject)
  d. Gender and number (Masculine singular)
  e. Frequency (High)

In sections 4 and 5, we test this hypothesis against data that will be presented in section 3.

3. Data

In order to test the hypothesis advanced in section 2.7, we created a database of 5,813 example sentences from the syntactic subcorpus of the Russian National Corpus.[4] This subcorpus was selected because it includes syntactic annotation, which makes it possible to unambiguously identify sentences with predicate adjectives in combinations with various types of subjects and complements.[5] A series of queries were carried out, and the resulting data were exported to one spreadsheet, where doublets were removed using the automatic tools in Excel. The database was lemmatized manually, and each example sentence was manually annotated for the factors listed in Table 1.

[4] Data and code are available here: URL TO BE ADDED.

[5] Corpora without syntactic annotation would return considerable amounts of irrelevant examples (“noise”), which would then have to be sorted out manually. For instance, a query for a noun in the nominative followed by an adjective in the nominative might return sentences like *Ivan molodoj student* ‘Ivan is a young student’, where *molodoj* ‘young’ is not in a predicate adjective although it comes after the subject of the sentence.

Factor	Values
A. Complement:	Finite clause
	Infinitive clause
	No complement
	Noun phrase
	Prepositional phrase
B. Subject:	Finite clause
	Infinitive clause
	Modified noun
	Unmodified noun
	Pronoun
	Other overt subject
	No subject
C. Word order:	Subject before predicate adjective
	Predicative adjective before subject
D. Gender and number of predicate adjective:	Masculine singular
	Feminine singular
	Neuter singular
	Plural
E. Frequency as a predicate adjective:	Total number of attestations as a predicative adjective in the syntactic subcorpus
F. Frequency of lemma:	Total number of attestations in the syntactic subcorpus
G. Lemma:	Lemmas of each adjective

Table 1: Factors and values included in the present study

As shown in Table 1, we included two measures of frequency in our investigation. “Frequency as a predicate adjective” represents the number of attestations of each lemma in this particular syntactic function. “Frequency of lemma” is the total number of attestations of each lemma in the syntactic subcorpus as a whole. This includes not only adjectives functioning as predicates, but also all other syntactic functions, e.g., as modifiers of nouns. We included both measures in order to find out which of them is the best predictor of the choice between long and short forms.

Before analyzing our dataset, we weeded out examples with lemmas that are irrelevant for our study, because they do not have a contrast between a short and a long form. An overview is provided in Table 2, which lists the relevant items and gives numbers of attestations in our dataset.

Category	Excluded lexical items	# Attestations
A. *Odin* ‘one’	*odin* ‘one’	20
B. Ordinal numerals	*pervyj* ‘first, *vtoroj* ‘second’, etc.	73
C. Pronouns with adjectival declension	*čej* ‘whose’, *drugoj* ‘another’, *ètot* ‘this’, *inoj* ‘another’, *kakoj* ‘which’, *kakoj-to* ‘some’, *kakov* ‘what sort of’, *moj* ‘my’, *naš* ‘our’, *sam* ‘by oneself, alone’, *takoj* ‘such’, *takov* ‘such’, *tot* ‘that’, *ves’* ‘all’,	223
D. Past active participles	*byvšij* ‘former’, *sumašedšij* ‘crazy’	2

E. Adjectives that lack long forms	*dolžen* ‘indebted’, *rad* ‘glad’	561
F. Adjectives that lack short forms		
i. Adjectives in *-skij*/*-ckij*	*amerikanskij* ‘American’, *nemeckij* ‘German’, etc. See appendix for full list.	51
ii. Deverbal adjectives in *-l*	*otstalyj* ‘backwards’, *zagorelyj* ‘sunburnt’	2
iii. Superlatives in *-šij*	*lučšij* ‘best’, *men’šij* ‘smallest’, *prostejšij* ‘simplest’, *starejšij* ‘oldest’, *xudšij* ‘worst’	9
iv. Relative adjectives in *-ovoj*/ *-ovyj*	*darmovoj* ‘free of charge’, *kadrovyj* ‘personnel, career’, *mirovoj* ‘worldwide’, *podrostkovyj* ‘adolescent’, *rakovyj* ‘cancer’, *razovyj* ‘one-time’, *rozovyj* ‘pink’, *šelkovyj* ‘silk’, *teplovoj* ‘heat, thermal’	9
v. Possessive adjectives in *-ov*	*pirrov* ‘pyrrhic’	1
G. Superlatives with *samyj* + LF	*samyj* followed by a long form	54
Total for all categories		1005

Table 2: Lexical items excluded from the database because they do not have a contrast between long and short forms

The syntactic subcorpus of the Russian National Corpus classifies the cardinal numeral *odin* ‘one’ and all ordinal numerals and as adjectives, because they have adjectival declension. However, since these items do not have a contrast between long and short forms, they are not relevant for the present study, and they were therefore excluded from the database. For the same reason, we excluded a number of lexical items that traditional grammars often analyze as pronouns, although they have adjectival declension. A full list of lexical items is provided under C in Table 2.

Past active participles have adjectival declension and are classified as adjectives in the corpus.[6] However, since they do not have short forms, they are not relevant for our study. The relevant lexical items are listed in D in Table 2. It is worth mention that our dataset contains a few items like *neobxodimyj* ‘indispensable’ that historically are passive participles. However, since the relevant lexical items seem to have become full-fledged adjectives with both long and short forms in Contemporary Standard Russian, these items were *not* excluded.

Grammars and handbooks contain lists of adjectives that lack the long form. Of these, only three were attested in our dataset, as shown under E in Table 2. The most important of these adjectives is *dolžen* ‘required’, for which we have 547 attestations in our dataset.

A number of adjectives are said not to have short forms, typically possessive and relative adjectives. It is notoriously hard to draw a clear line between relative and qualitative adjectives, since many adjectives may be compatible with both qualitative and relative readings. As pointed out by Townsend (1975: 210), for instance, *serdečnyj* can be analyzed as relative in *serdečnaja bolezn’* ‘heart disease’, but as qualitative in *serdečnyj čelovek* ‘warmhearted person’. In view of this, we decided to exclude the relative adjectives with the formal features listed in F (i-iv) in Table 2 but to keep other putative relative adjectives in the dataset.

Finally, the corpus searches returned a number of superlative constructions, where a long form is preceded by the superlative marker *samyj*. In superlatives of this type, the short form is not attested, and we therefore excluded such examples from the statistical analysis.

[6] As opposed to past active participles, present active participles in *-ščij* are attested as both long and short forms in our dataset and therefore not excluded. Some of the relevant examples may be analyzed as adjectives in present-day Russian, e.g., *blestjaščij* ‘brilliant’ and *podxodjaščij* ‘suitable’.

After weeding out the irrelevant lexical items in Table 2 we were left with a dataset of 4,808 example sentences. These examples will be analyzed in the following sections.

4. Quantitative analysis: (nearly) categorical rules

Before we turn to the statistical analysis in section 5, it is necessary to discuss some cases where our data show that we are dealing with (nearly) categorical rules. Since the relevant contexts allow only short forms, it does not make sense to include these constructions in the statistical model.

4.1 Adjectives with complements

As mentioned in section 2.7, we hypothesize that adjectives with complements favor the short form. Our data confirm this hypothesis and suggest that this is a (nearly) categorical rule.

As shown in Table 3, long forms are not attested at all in examples with finite clauses or infinitive constructions as complements. Complements realized as noun phrases and prepositional phrases also strongly prefer short forms, although we have five counterexamples in our database. Two of them involve noun phrases:

(15) **Dostojnye**$_{\mathbf{LF}}$ uvaženija.
‘They are worthy of admiration.’

(16) Oni mne teper’ počti **rodnye**$_{\mathbf{LF}}$, kak i Èstoncy.
‘For me, they are almost relatives, like the Estonians.’

Example (15) involves the adjective *dostojnyj* ‘worthy’, which governs the genitive case. We have six attestations of this adjective in our database, five of which contain short forms as predicted by our hypothesis. Sentence (16) includes the dative experiencer *mne* ‘me’, which the corpus classifies as a complement of the adjective. We anticipate that not all scholars will share this analysis.

The three attestations we have of long forms with prepositional phrases all involve the preposition *dlja* ‘for’, which is semantically close to a dative experiencer:

(17) Odnako jazyk ètot dlja nego soveršenno **čužoj**$_{\mathbf{LF}}$.
‘However, for him this language is completely foreign.’

(18) Smešannyj princip komplektovanija – naibolee **priemlemyj**$_{\mathbf{LF}}$ dlja našix Vooružennyx Sil.
‘Mixed recruitment is most acceptable for our armed forces.’

(19) **Važnyj**$_{\mathbf{LF}}$ dlja každogo iz nas i dlja vsego obščestva.
‘He is important for each of us and for society as a whole.’

Arguably, prepositional phrases with *dlja* and dative experiencers are more appropriately analyzed as adjuncts than as complements, in which case they would not be counterexamples to the rule that adjectives with complements are in the short form. However, we will not pursue this issue here, since in any case our data show that complements (in a wide sense) strongly favor short forms. This also includes examples with *dlja*, for which 92 out of a total of 95 examples display short forms. Given that we are dealing with a (nearly) categorical rule, it does not make sense to subject the relevant examples to statistical analysis. In the following, we will therefore exclude examples with complements from our analysis, and limit our analysis to the 3,999 examples without complements.

Complement	# Long forms	# Short forms	# Total	% Short forms
Finite clause	0	95	95	100%
Infinitive clause	0	247	247	100%
Noun phrase	2	153	155	99%
Prepositional phrase	3	309	312	99%
No complement	542	3,457	3,999	86%
Total	547	4,261	4,808	

Table 3: Distribution of short and long forms for predicate adjectives with and without complements

4.2 Impersonal constructions

After excluding sentences with complements, we consider various types of subjects in the remaining part of the database. As mentioned in section 2.2, our database includes a number of examples with clausal or infinitival subjects, which traditionally are known as "impersonal constructions". Traditional grammars expect short forms in such sentences (although they may be classified as adverbs or "predicatives" rather than adjectives, e.g., Švedova ed. 1980: 215). As shown in Table 4, this expectation is borne out by the facts. Moreover, our data suggest that this is a (nearly) categorical rule. In view of this, we will exclude impersonal constructions from the statistical analysis, which we turn to in section 5.

Subject	# Long forms	# Short forms	# Total	% Short forms
Finite clause	2	451	453	99.6%
Infinitive	0	962	962	100%
No subject: Neuter	17	297	314	95%
No subject: Non-neuter	84	45	129	35%
Noun phrase	360	1196	1556	77%
Pronoun	71	459	530	87%
Other	8	47	55	85%
Total	542	3457	3999	

Table 4: Distribution of long and short forms for different kinds of subjects in sentences without a complement

Some of the examples with no subject may be analyzed as impersonal constructions. This applies to sentences where the predicative adjective is in the neuter gender:

(20) Mne tak **grustno**$_{\textbf{SF}}$ ...
'I am so sad ...'

In such constructions, the neuter gender is required, and the adjective is in the short form.

Examples with no subject and predicate adjectives in other forms than the neuter singular show a different distribution of short and long forms, as shown in Table 4. Relevant examples are:

(21) a. Očen' **xrupkaja**$_{\textbf{LF}}$.
'She is very fragile.'
b. Očen' **krasiv**$_{\textbf{SF}}$.
'He is very handsome.'

This construction may be referred to as "elliptic sentences" since an implicit subject can be restored from the context. In (21a-b), for instance, a personal pronoun such as *on* 'he' or *ona* 'she' may be inserted without affecting the grammaticality of the sentence.

Notice that in sentences with no overt subject it is difficult to draw the line between the impersonal and the elliptic constructions. While adjectives in the masculine singular, feminine singular and the plural are always of the elliptic type, adjectives in the neuter singular may in principle be of both types. In order to decide, one would have to inspect the wider context of each example. Unfortunately, the syntactic subcorpus of the Russian National Corpus does not provide sufficient access to context. Instead, we decided to use gender and number as a proxy, treating all examples in the neuter singular as impersonal constructions. This is good enough for our purposes, since it allows us to correctly predict the use of the short form in 95% of the examples in the neuter singular.

To summarize, we will exclude the three types of impersonal constructions we have identified above: examples with clausal subjects, examples with infinitival subjects, and examples with no subject and predicative adjective in the neuter singular. For these sentence types, the short form is a (nearly) categorical rule. After excluding examples with these subjects, we are left with 2,270 examples to analyze.

4.3 Non-personal pronouns as subjects

After having removed examples with complements and impersonal constructions, we are left with elliptic constructions, as well as examples where the subject is a noun phrase or a pronoun. We now take a closer look at pronominal subjects. As mentioned in section 2.2, certain types of pronouns are expected to combine with short forms.

In Table 5, we draw a distinction between personal pronouns and other pronouns, which we will refer to as “non-personal”. Personal pronouns are *ja* ‘I’, *ty* ‘you (sg)’, *on* ‘he’, *ona* ‘she’, *ono* ‘it’, *my* ‘we’, *vy* ‘you (pl)’, and *oni* ‘they’. All other pronouns are analyzed as “non-personal”, but this group is strongly dominated by the neuter forms *èto* ‘this’, *to* ‘that’ and *vsë* ‘all’, which account for 280 (86%) of the 325 examples with non-personal pronouns as subjects. As shown in Table 5, the distribution of long and short forms is different for personal and non-personal pronouns. While for personal pronouns both long and short forms are well attested, the short form appears to be a nearly categorical rule for non-personal pronouns. For the purposes of the statistical analysis, we will therefore exclude examples with non-personal pronouns as subjects.

Pronominal subject	# Long forms	# Short forms	# Total	% Short forms
Personal pronouns	58	147	205	72%
Non-personal pronouns	13	312	325	96%
Total	71	459	530	

Table 5: Distribution of long and short forms for pronominal subjects in sentences without a complement

4.4 Word order: predicate adjective before subject

As mentioned in section 2.3, there is a well-known hypothesis that a predicate adjective that precedes the subject tends to be in the short form. Is this a categorical rule? In Table 6, we take into account all the 1,816 examples that are left after removing the constructions discussed in sections 4.1 through 4.3. As the table shows, the short form is so close to being a categorical rule that it does not make sense to include examples with predicate adjectives before the subject.

	# Long forms	# Short forms	# Total	% Short forms
AS	10	419	429	98%
SA	416	971	1387	70%
Total	426	1390	1816	

Table 6: Distribution of long and short forms for different word orders. AS = predicate adjective before subject, SA = subject before predicate adjective. The table includes sentences with no complements and with subjects that are nouns and pronouns, as well as a small residual category of "other subjects" (e.g., nominalized adjectives)

4.5 Discussion and summary: the "space of competition" and the Situation/Participant Hypothesis

Our analysis of (nearly) categorical rules enables us to state some generalizations, which we explore in the following. In particular, we are now in a position to identify the "space of competition", where both long and short forms are well attested. We furthermore advance the "Situation/Participant Hypothesis" and a decision tree that captures the distribution of predicate adjectives in simple terms.

The environments discussed in the previous sections are summarized in Table 7. We first consider complements, then take into account the subject of the sentence, before we turn to gender/number and word order. This enables us to identify six environments where the short form is a (nearly) categorical rule. The two remaining environments, which are shaded in the table, represent the "space of competition", where both forms are well attested. As shown in the table, this space is defined by three factors, viz. complement, subject and word order, in the following way:

(22) "Space of competition":
Both long and short forms are used in sentences where:
a. the predicative adjective has no complement,
b. the subject is an elliptic or overt noun phrase or personal pronoun, and
c. the subject (if overt) precedes the predicate adjective.

Environment	#Long forms	#Short forms	# Total	% Short forms
1. With complement	5	804	809	99%
2. Without complement				
a. Subject = Clause	2	451	453	99.6%
b. Subject = Infinitive	0	962	962	100%
c. Subject = non-pers. pronoun	13	312	325	96%
d. No subject				
i. Neuter (impersonal)	17	297	314	95%
ii. Non-neuter (elliptic)	84	45	129	35%
e. Subject = pers. pronoun/NP				
i. Word order = SA	416	971	1387	70%
ii. Word order = AS	10	419	429	98%

Table 7: Summary of distribution of long and short forms. Shaded rows represent the "space of competition", where both long and short forms are well attested.

At this point it makes sense to ask if there are any properties that unite the contexts where short forms are preferred and set them apart from the contexts where both forms are used. First, consider subjects. It is instructive to compare sentences with the personal *ono* 'it' and the non-personal pronoun *èto* 'this', which both combine with predicate adjectives in the neuter singular:[7]

(23) a. Èto mesto ne uznat': **ono zolotoe**$_{\textbf{LF}}$, a za nim soveršenno čërnaja reka. (Lunin 2016)
'This place is unrecognizable: it is golden, and behind it there is a completely black river.'
b. Každyj akter privnes v fil'm svoju nepodražaemuju auru: Svetlana Nemoljaeva, Lija Axedžakova, Ljudmila Ivanova … Kak imi možno ne vosxiščat'sja i ne ljubit'? **Èto nevozmožno**$_{\textbf{SF}}$! (Internet forum 2006–2010)
'Each actor contributed their inimitable aura to the movie: Svetlana Nemojaeva, Axedžakova, Ljudmila Ivanova … How is it possible not to be excited about them and not to love them? It's impossible!'

While *ono* in (23a) refers back to the noun *mesto* 'place' in the left context, *èto* in (23b) does not pick out a single noun as its antecedent. Instead, *èto* refers to the whole situation described in the previous sentence, which concerns the aura that certain actresses contributed to a movie. Simplifying somewhat, we may say that *ono* refers to participants, while *èto* refers to situations. Thus, predicative adjectives that combine with *ono* are descriptions of participants, while those combining with *èto* are descriptions of situations.

Is it possible to extend this analysis to other sentence types? As shown in Table 7, short forms are obligatory in sentences with clausal (24a) and infinitival subjects (24b), as well as impersonal constructions with no subject (24c):

(24) a. **Interesno**$_{\textbf{SF}}$, čto genial'nyj Ejnštejn v detstve stradal autizmom.
'It is interesting, that as a child the genius Einstein suffered from autism.
b. **Interesno**$_{\textbf{SF}}$ li zanimat'sja geologiej segodnja, kogda mnogoe uže izvestno?
'Is it interesting to study geology today when so much is already known?'
c. –Valjuša, milaja, skaži, èto ne… Ja kivnula. – Bože kak **Interesno**$_{\textbf{SF}}$. U tebja budet malen'kij bèbi! (Grekova 1962)
'Valjuša, my dear, say it isn't … I nodded. God how interesting. You are going to have a little baby!'

In (24a), the clausal subject represents the situation in Einstein's childhood whereby the future great scientist suffered from autism, and the short form *interesno* 'interesting' describes this situation. In (24b), the short form describes an activity (to study geology). In (24c), the context establishes a situation where the interlocutor is pregnant, and this situation is then described by the predicative adjective as "interesting".

We may contrast these examples with sentences where long forms are possible. As pointed out above, such sentences may have personal pronouns like *ono* as subjects (see 23a), or noun phrases as in (25a-b):

(25) a. Glaza u nix očen' **strannye**$_{\textbf{LF}}$.
'Their eyes are very strange.'

[7] These examples are from the main subcorpus of the Russian National Corpus, since the syntactic subcorpus does not contain examples that illustrate the relevant point.

b. Procedura vyborov očen' **blizkaja**$_{LF}$.
‘The election procedure is very similar.’

Although the subject nouns in (25) have different meanings, they both represent participants in a situation. The same holds for elliptic sentences, such as those in (21) above, where, as pointed out in section 4.2, a personal pronoun can be added without changing the grammaticality of the sentence.

While more research is needed about the relationship between various types of subjects and predicate adjectives, we propose “situation” as a broad label for impersonal sentences and sentences with non-personal pronouns and “participant” as an umbrella term for sentences with personal pronouns or noun phrases as subject, as well as for elliptic sentences. We suggest that long forms can only be descriptions of participants.

However, we may make the generalization more precise if we take word order into account. As shown in Table 7, long forms are found in sentences where the subject comes before the predicate adjective. In Russian, word order is closely related to information structure. While this relationship is complex, for present purposes the traditional insight that the subject in (26a) represents given information, while the subject in (26b) is new information, is sufficiently precise:

(26) a. Mesto **izvestnoe**$_{LF}$, xotja i ne každyj žitel' severnoj stolicy bez razdumij nazovet vse ulicy, sxodjaščiesja na ètom pjatačke.
‘The place is well known, although not every person in the northern capital would be able to name all the streets that come together in this small patch of land without thinking twice.’
b. **Izvestno**$_{SF}$ imja ego otca – Rejner Fos.
‘The name of his father is well known – Reiner Foss.’

In (26a), the place (*mesto*) is given information, while the new information is conveyed by the predicate adjective. In (26b), on the other hand, the subject represents new information – the name of the relevant person’s father. Using the traditional terms “theme” for given information and “rheme” for new information, we may say that long forms, which come after the subject, are descriptions of thematic participants. Short forms, on the other hand, are less selective for their subjects, since they also occur in sentences where the subject is rhematic and comes after the predicate adjective.

To sum up this discussion, we propose the following hypothesis:

(27) The Situation/Participant Hypothesis:
In sentences without a complement
a. Long forms are descriptions of thematic participants.
b. Short forms are descriptions of situations or participants.

While this hypothesis probably does not capture all the complexities of long and short forms, we maintain that it accommodates prototypical patterns of use.

Based on the hypothesis in (27), we can formulate a simpler description of the “space of competition”:

(28) "Space of competition" (revised):
Both long and short forms are used in sentences where the predicative adjective:
a. has no complement, and
b. describes a thematic participant.

The distribution of long and short forms can be represented as a simple decision tree, as shown in Figure 1. The decision tree represents the "space of competition" as the shaded terminal node in the lower right portion of the figure. The three white terminal nodes clarify that short forms (as opposed to long forms) may combine with complements and describe situations and rhematic participants.

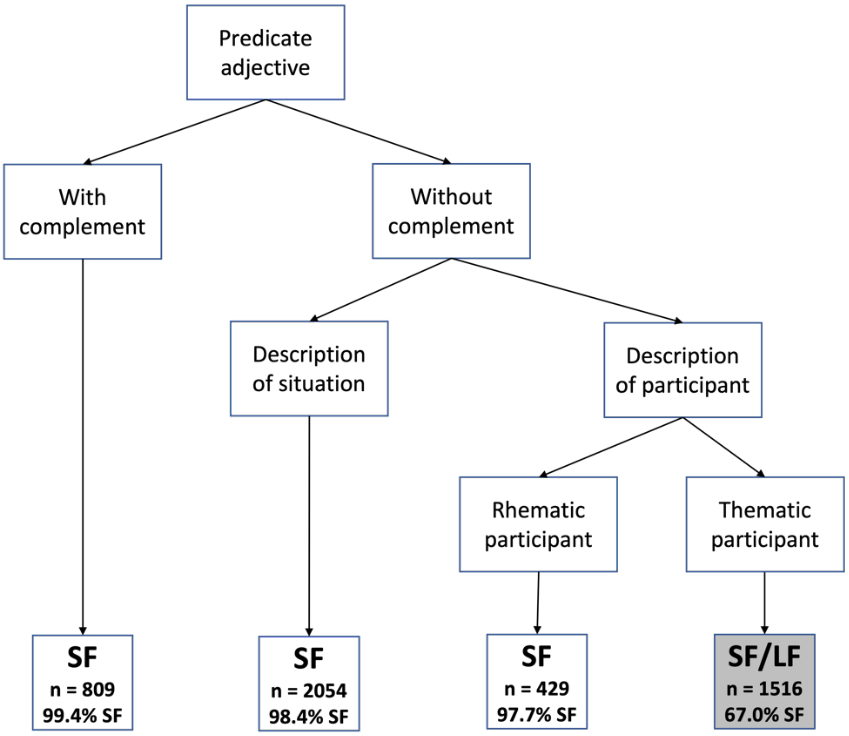

Figure 1: Decision tree summarizing the distribution of long forms (LF) and short forms (SF). The shaded terminal node represents the "space of competition", where both forms are well attested. In each terminal node, n represents the total number of examples in the relevant category. The percentages represent the proportion of short forms in each category.

5. Quantitative analysis: statistical analysis of the remaining factors

Now that we have considered the contexts where the short form is a (nearly) categorical rule, we turn to a statistical investigation of the "space of competition" where both long and short forms occur. In section 5.1, we present a regression model, before we discuss individual factors in sections 5.2 through 5.4.

5.1 A regression model

In section 2.7, we stated a hypothesis involving five factors: complement, subject, word order, gender and number, and frequency. We have already established that complements always combine with short forms, and that short forms are a (nearly) categorical rule for impersonal constructions with clausal, infinitival or no subject. It has furthermore been shown that word order involves a (nearly) categorical rule. We are therefore left with three factors: (a) subjects (except those mentioned above), (b) gender and number, and (c) frequency. In order to investigate these factors, we fitted a logistic regression model for predicting long vs. short form using the glmer() function in R (R Core Team 2022) with the following formula:

(29) Adjective_form ~ 1 + Log_Frequency_predicative[8] + Subject + Gender_Number_adj

The model can be interpreted as follows: "The form of the predicate adjective is predicted in relation to an overall intercept (1) with main effects of frequency, subject, and gender/number."

A drop1() function was applied that confirmed that none of the factors should be removed from the formula. Short forms are observed in 67.02% of this dataset. This means that a null model that merely guesses "short form" every time will be correct in 67.02% of cases, so our regression model must be judged against that baseline.

The results of the regression model are summarized in Table 8 and visualized in Figure 2. The estimate represents an effect size, where positive values indicate a relative preference for short forms, whereas a relative preference for long forms is represented as negative values. The intercept is the situation where the value of Log Frequency predicative is zero, the subject is a noun phrase, word order is canonical (subject before adjective or only an adjective), and gender/number is neuter singular. As shown, at the intercept the estimate has a positive value. This indicates a preference for the short form. The p-values in the fifth column of the table show that all of the predictors reach significance.

Predictors	Estimate	Std. error	z value	Pr(>|z|)	
Intercept	0.03269	0.14180	0.231	0.81769	
Log Frequency predicative	0.30291	0.03439	8.807	< 2e-16	***
Subject = No subject	-1.33142	0.20838	-6.389	1.66e-10	***
Subject = Noun modified	0.87744	0.15938	5.505	3.68e-08	***
Subject = Other	0.61348	0.42808	1.433	0.15184	
Subject = Pronoun	0.31726	0.18322	1.732	0.08335	
Gender Number adj=F	-0.39640	0.15165	-2.614	0.00895	**
Gender Number adj=N	-0.49291	0.19472	-2.531	0.01136	*
Gender Number adj=Pl	0.14257	0.15611	0.913	0.36108	

Table 8: Results of logistic regression model. Three stars indicate significance below 0.001, two stars indicate significance below 0.001, and one star indicates significance below 0.05.

[8] Regression analysis was carried out for logarithmic transformations of two frequency measures, one based on the total frequency of each adjective lemma, and another based on the frequency of each adjective lemma in the predicative position in the corpus. The measure of frequency specifically in predicative position proved to be more significant, and this is the measure that is used in the analysis presented here.

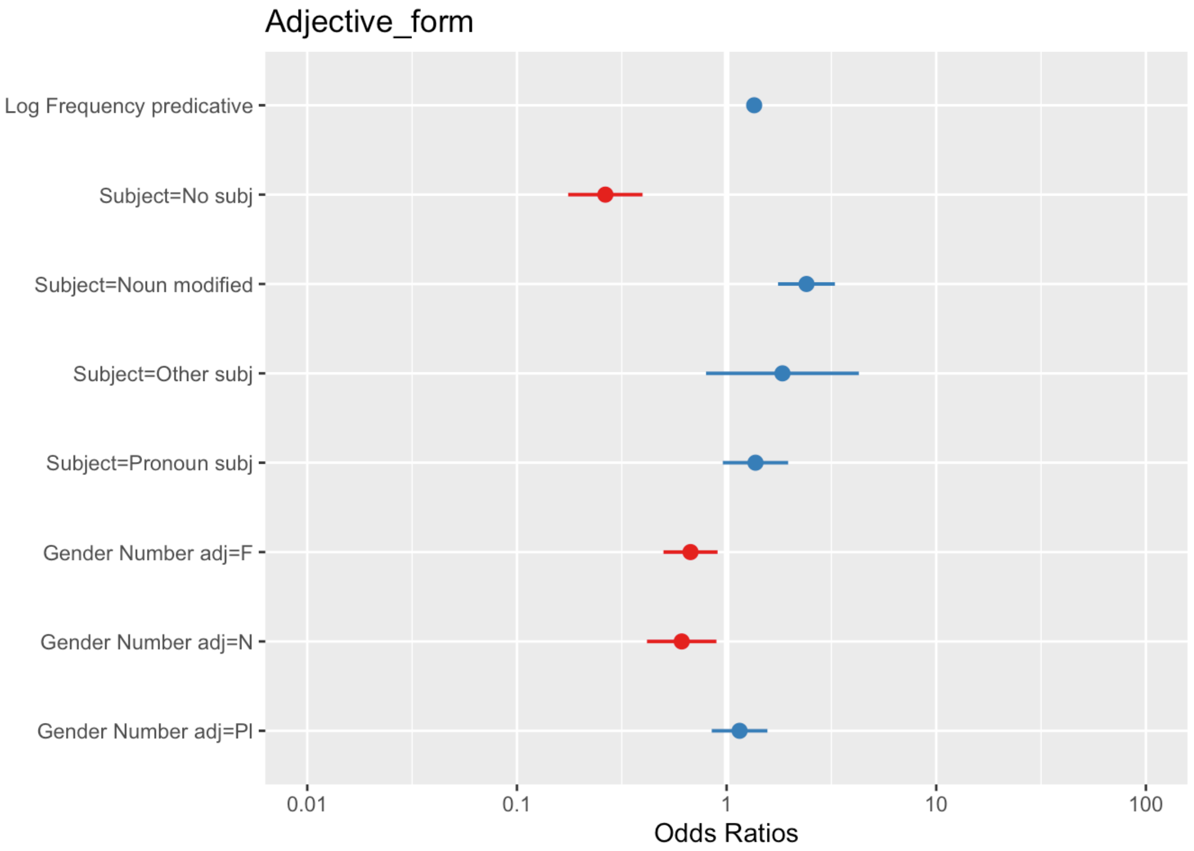

Figure 2: Results of logistic regression model: Odds ratios for all predictors. Negative values (given in red) indicate a relative preference for long forms, while positive values (given in blue) are indications of a relative preference for short forms. The lines represent the 95% confidence intervals.

As shown in Table 8 and Figure 2, where positive values for the estimate indicate a relative preference for short forms, high frequency and modified nouns as subjects are factors that promote short forms. No subject (in elliptic sentences) and feminine and neuter gender received negative values for the estimate, indicating a relative preference for long forms.

Our model yields correct predictions in 70.05% of all cases. While this is not a lot better than the baseline of 67.02%, a statistical comparison of our model's level of successful prediction with the baseline shows that the difference is highly significant and the probability that this difference could occur by chance is 0.006. The C-score measure of goodness of fit for the model is 0.704, which indicates a good model. The R-squared value indicates that the model accounts for 16.5% of the variance in the data.

We take these facts to indicate that the predictors under scrutiny do not constrain the use of predicate adjectives to a large extent. Differently stated, within the "space of competition" the long vs. short forms vary relatively freely, with some variation explained by the factors in our model. This suggests that the long vs. short forms may be recruited to convey different meanings, or that there are stylistic or diachronic differences between the two forms that cannot be revealed in the present study. There may, of course, also be other factors at work that we have not been able to identify.

5.2 Subject types

Figure 3 offers a visualization of the model predictions for various types of subjects. The *y*-axis represents the prediction of the value for predicate adjective, which is long form for values below 0.5 and short form for higher values. The horizontal dotted line at 0.5 shows the point where the prediction goes from a preference for long form (under 0.5) to short form (over 0.5). In the diagram, bar width indicates the relative amount of data for each subject type and the brackets show the 95% confidence interval.

As shown, unmodified nouns, modified nouns and pronouns favor short forms. This tendency is strongest for modified nouns, which is significantly higher than unmodified nouns. As mentioned in the previous section, pronouns are not significantly different from unmodified nouns. We will not discuss the small category of “other subjects”.

Sentences with no subject represent the only subject type that favors long forms, and in fact this is the only factor we have identified in the present study that shows a preference for long forms. Recall from section 4.2 that the subjectless sentences we are concerned with here are of the elliptic type. Impersonal constructions with no subject were removed from the statistical analysis, since for these constructions the short form is a (nearly) categorical rule.

Summarizing, we suggest the following hierarchy for the effect of subject types:

(30) The subject hierarchy:
Modified noun > unmodified noun, pronoun > no subject (where > means “is more likely to combine with a short form than”)

However, it is important to keep in mind that the differences between the subject types are relatively modest, thus suggesting a relatively free choice between long and short forms within the “space of competition”.

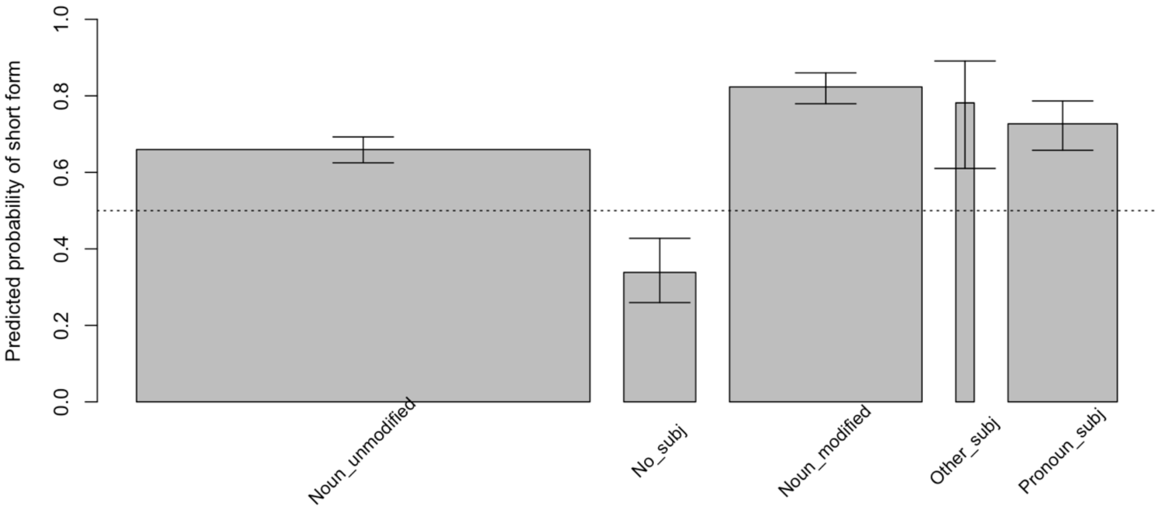

Figure 3: Effect of subject type on prediction of short forms

5.3 Gender and number

Figure 4 is a visualization of the effect of gender and number on the choice between long and short forms. The diagram is of the same type as Figure 3 in the previous section. As shown, the distribution of long and short forms for the four genders/numbers is relatively even. However, as pointed out in section 5.1, our statistical model finds a statistically significant difference between masculine singular on the one hand, and feminine and neuter singular on the other hand. There is no statistically significant difference between the plural and the masculine singular. On the basis of this, we propose the following generalization:

(31) The gender/number hierarchy:
Masculine singular, plural > feminine singular, neuter singular (where > means "is more likely to combine with a short form than")

Once again, we are dealing with small differences, which suggests that the choice between long and short forms is relatively free within the "space of competition".

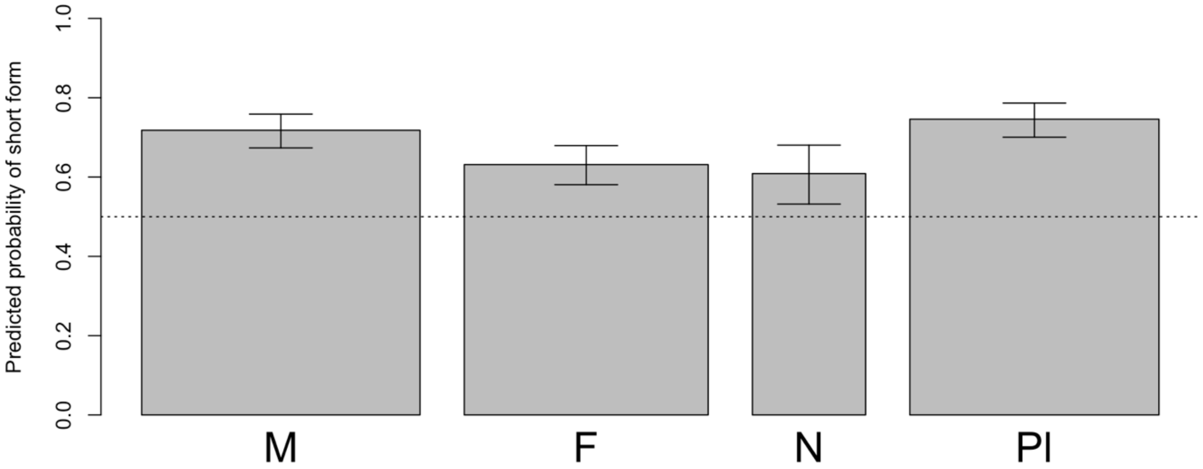

Figure 4: Effect of gender and number on prediction of short forms. M = masculine singular, F = feminine singular, N = neuter singular, Pl = plural

5.4 Frequency

In Figure 5, we offer a visualization of the effect of frequency on the choice between long and short forms. In the same way as in the previous sections, the *y*-axis represents the prediction of the value for the predicate adjective, viz. long form for values below 0.5 and short form for values above 0.5. The *x*-axis presents frequency of the adjective lemmas in predicative function with low frequencies to the left and high frequencies to the right. Recall that the frequency values have been logarithmically transformed. The "rug" at the bottom indicates the presence of observations, showing that most observations are for adjective lemmas of middle to high frequency. The blue line shows the model predictions of short forms, and the blue ribbon represents the 95% confidence interval.

The figure shows that for hapaxes (lemmas occurring once in the dataset) the prediction is about 50% short forms, and that the likelihood of short forms increases gradually to more than 80% for adjective lemmas of high frequency. In other words, while long and short forms seem to vary freely for low frequent adjectives, high frequent adjectives show a strong preference for short forms. Table 9 lists the ten most frequent lemmas as well as ten examples of hapaxes. In total, our dataset contains 319 hapaxes.

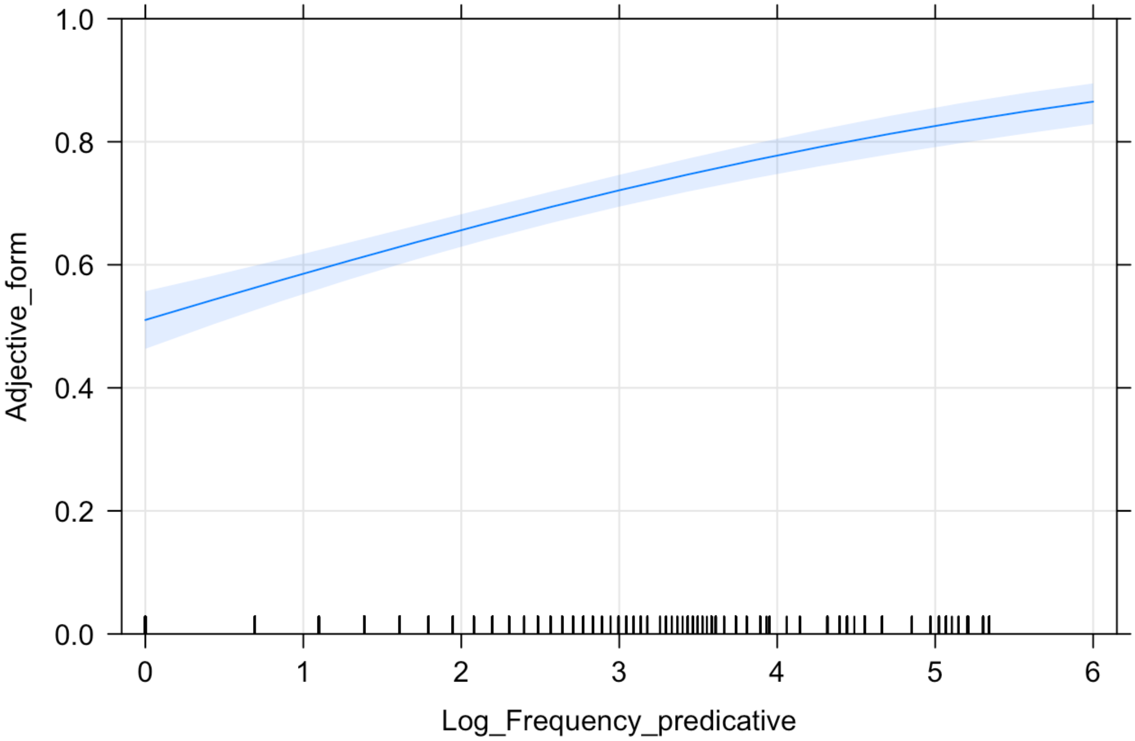

Figure 5: Effect of frequency on prediction of short forms

a. Lemmas of high frequency:			
nužnyj 'necessary'	613 (6.42)	*gotovyj* 'ready'	172 (5.15)
izvestnyj 'well-known'	209 (5.34)	*trudnyj* 'difficult'	165 (5.11)
važnyj 'important'	201 (5.30)	*xorošij* 'good'	159 (5.07)
neobxodimyj 'indispensable'	183 (5.21)	*ponjatnyj* 'understandable'	152 (5.02)
nevozmožnyj 'impossible'	182 (5.20)	*interesnyj* 'interesting'	144 (4.97)
b. Hapaxes:			
adekvatnyj 'adequate'	1 (0)	*mudryj* 'wise'	1 (0)
antinaučnyj 'anti-scientific'	1 (0)	*osobyj* 'special'	1 (0)
bezzaščitnyj 'helpless'	1 (0)	*slabovol'nyj* 'infirm'	1 (0)
iskusstvennyj 'artificial'	1 (0)	*zdravyj* 'healthy'	1 (0)
jadovityj 'poisonous'	1 (0)	*xmuryj* 'gloomy'	1 (0)

Table 9: The ten most frequent adjectives and ten hapaxes in our dataset. The numbers represent the total number of occurrences as predicate adjectives in the syntactic subcorpus of the Russian National Corpus and the numbers in parentheses represent the logarithmic transformations of these frequencies corresponding to the x-axis of Figure 5. Note that the highest frequency item is an outlier not visualized in Figure 5.

6. Conclusion

We have carried out an empirical investigation of the choice between long and short forms of predicative adjectives in Russian. We have limited ourselves to sentences with no overt copula verb ("zero copula").

Our findings can be summarized as follows. First, our data suggest that the short form is the dominant option for predicate adjectives in modern Russian. Short forms represent 89% of the dataset as a whole and 67% in the "space of competition", i.e., the contexts where both forms occur freely. Second, we have confirmed that there are contexts where the use of the short form is a (nearly) categorical rule. These contexts are: (a) sentences with complements, (b) impersonal constructions with clausal, infinitival, or no subject, (c) sentences with certain "non-personal" pronouns as subjects, and (d) sentences where the predicate adjective comes before the subject and the subject is realized as a noun phrase or a personal pronoun. Third, we have advanced a hypothesis whereby long forms are descriptions of thematic participants, while short forms are less restrictive and can describe both participants (thematic and rhematic) and situations. Fourth, we have shown that both long and short forms are well attested in a "space of competition" consisting of sentences where the predicative adjective (a) has no complement, and (b) describes a thematic participant. Fifth, within the "space of competition" the choice between long and short form depends to some extent on subject type, gender/number, and frequency. The following statistical tendencies were identified: sentences with (overt) nouns and pronouns as subjects favor short forms, a tendency that is stronger if the noun combines with an adjectival or similar modifier. In elliptic sentences without a subject, long forms are preferred. The likelihood of short forms increases with increasing frequency; for hapaxes, we predict free variation between long and short forms, while for high frequent adjectives we predict short forms in more than 80% of the cases. Sixth, the statistical tendencies we have identified do not involve particularly strong predictors. Although we cannot exclude that other unseen factors are at work, we take this to indicate that both long and short forms are used with few limitations within the "space of competition". This may indicate that the long and short forms are recruited to convey different meanings within this space, but our finding is also compatible with a situation whereby the two forms are stylistically different or are in the process of undergoing diachronic change. Finally, our investigation of frequency has revealed a "locality effect" since the "local" measure of frequency in predicative position has proven a better predictor than "global" frequency in the corpus as a whole.

Beyond the study of predicate adjectives, our analysis has contributed a methodology which may be extended to other cases of competing constructions in Russian and other languages. We have demonstrated the value of first "peeling off" contexts where (nearly) categorical rules can be formulated before one establishes as a "space of competition", where the interplay of factors may be investigated by means of statistical modeling.

The present study opens up four alleys for future research. First, it is necessary to undertake a systematic investigation of potential stylistic differences between long and short forms within the "space of competition". Second, a thorough investigation of potential semantic differences between long and short forms within the "space of competition" may be rewarding. Third, it would be beneficial to carry out a diachronic study in order to find out if the distribution of long and short forms has changed over time. Fourth, the methodology established in the present study should be extended to sentences with overt copula verbs, where the long and short nominative forms also compete with predicative adjectives in the instrumental. However, since the data available in the present study cannot shed light on these issues, they will be left open for future research.

References

Babby, Leonard H. 1973. The Deep Structure of Adjectives and Participles in Russian. *Language* 49.2: 349–360.

Babby, Leonard H. 2009. *The Syntax of Argument Structure*. Cambridge: Cambridge University Press.

Babby, Leonard H. 2013. The syntactic differences between long and short forms of Russian adjectives. In Patricia Cabredo Hofherr and Ora Matushansky (Eds.): *Adjectives: Formal analyses in syntax and semantics*. Amsterdam and Philadelphia: John Benjamins Publishing Company, 53–84.

Benson, Morton. 1959. Predicate Adjective Usage in Standard Russian. *Word* 15.1: 89–100.

Bybee, Joan B. 2007. Diachronic linguistics. In Dirk Geeraerts and Hubert Cuyckens (Eds.): The Oxford Handbook of Cognitive Linguistics. Oxford: Oxford University Press, 945–987.

Børresen, Tore. 1966. *Sootnošenie kratkix i polnyx form imeni prilagatel'nogo v sostave skazuemogo v russkom jazyke.* MA thesis: University of Oslo.

Fonnes, Malvin. 2013. *Det predikative adjektivet i presentiske kopulasetninger i russisk*. PhD dissertation: University of Bergen.

Gustavsson, Sven. 1976. *Predicative adjectives with the copula* byt' *in modern Russian*. = *Acta Universitatis Stockholmiensis. Stockholm Slavic studies* 10. Stockholm: University of Stockholm.

Isačenko, Aleksandr V. 1962. *Die russische Sprache der Gegenwart*. Halle (Saale): Max Niemeyer Verlag.

Iversen, Malvin. 1978. *Predikativt adjektiv ved preteritum av kopula* byt' *i moderne russisk*. = *Skrifter* 2. Bergen: University of Bergen.

Mathiassen, Terje. 1996. *Russisk grammatikk*. Oslo: Universitetsforlaget.

Nesset, Tore and Laura A. Janda. A network of allostructions: quantified subject constructions in Russian. *Cognitive Linguistics* 34.1: 67–97.

Nichols, Johanna. 1981. *Predicate Nominals: A Partial Surface Syntax of Russian*. Berkeley and Los Angeles: University of California Press.

Peškovskij, Aleksandr M. 1933. *Russkij sintaksis v naučnom osveščenii*. 6th edition. Moscow: Gosudarstvennoe učebno-pedagogičeskoe izdatel'stvo.

Pereltsvaig, Asya. 2001. Syntactic categories are not primitive: evidence from short and long adjectives in Russian. In Steven Franks et al. (Eds.), *Formal Approaches to Slavic Linguistics (FASL-9). The Bloomington meeting 2000.* Ann Arbor: Michigan Slavic Publications, 209–227.

R Core Team. 2022. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Rozental', Ditmar È. and Margarita A. Telenkova. 1974. Stilističeskoe ispol'zovanie form imen prilagatel'nyx. *Russkij jazyk za rubežom* 2: 42–47.

Švedova, Natalija Ju (ed.). 1952. Polnye i kratkie formy imen prilagatel'nyx v sostave skazuemogo v sovremennom russkom literaturnom jazyke. *Učenye zapiski MGU* 150: 73–132.

Švedova, Natalija Ju (ed.). 1980. *Russkaja grammatika*, vol. 2. Moscow: Nauka.

Timberlake, Alan. 2004. A reference grammar of Russian. Cambridge: Cambridge University Press.

Townsend, Charles E. 1975. *Russian word-formation*. Columbus: Slavica Publishers.

Ueda, Masako. 1992. *The Interaction between Clause-Level Parameters and Context in Russian Morphosyntax*. Bern: Peter Lang International Academic Publishers.

Vinogradov, Vladimir V. 1947. *Russkij jazyk. Grammatičeskoe učenie o slove*. Moscow and Leningrad: Učpedgiz.

Conflict of interest statement

We are not aware of any conflict of interest regarding this article.

Appendix: adjectives in *-skij*/*-ckij*

Here is a complete list of adjectives in *-skij*/*-ckij* that were excluded from the analysis (see Table 2 in Section 3):

Amerikanskij ‘American’, *analitičeskij* ‘analytical’, *antiputinskij* ‘anti-Putin’, *avtorskij* ‘authorial’, *bjurokratičeskij* ‘bureaucratic’, *carskij* ‘czar’, *demokratičeskij* ‘democratic’, *èkonomičeskij* ‘economic’, *èlektričeskij* ‘electrical’, *ènergetičeskij* ‘energy’, *fantastičeskij* ‘fantastic’, *fevral’skij* ‘February’, *gensekovskij* ‘general secretary’, *gigantskij* ‘giant’, *gipotetičeskij* ‘hypothetical’, *gospodskij* ‘Lord’, *imperskij* ‘imperial’, *istoričeskij* ‘historical’, *kitajskij* ‘Chinese’, *kritičeskij* ‘critical’, *lourensovskij* ‘Lawrence‘, *nemeckij* ‘German’, *nevažneckij* ‘insignificant’, *nevskij* ‘Neva’, *optimističeskij* ‘optimistic’, *peterburgskij* ‘Petersburg’, *piterskij* ‘Petersburg’, *političeskij* ‘political’, *proputinskij* ‘pro-Putin’, *ritoričeskij* ‘rhetorical’, *rossijskij* ‘Russian’, *sel’skij* ‘arcadian’, *šumerskij* ‘Sumerian’, *sovetskij* ‘Soviet’, *specifičeskij* ‘specific’, *stereofoničeskij* ‘stereophonic’, *texnologičeskij* ‘technological’, *unificirovanno-evropejskij* ‘Unified European’, *vserossijskij* ‘all-Russian’, *ženskij* ‘female’