Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2020"

Moscow, June 17–20, 2020

ZERO FORMS IN MORPHOLOGICAL PARADIGMS: THE VERB "BE" IN RUSSIAN[1]

Zimmerling A. V. (fagraey64@hotmail.com)

Pushkin state Russian language institute; Institute of Linguistics, Russian Academy of science; Moscow pedagogical state university, Moscow, Russia

This paper offers a corpus analysis of the Russian verb *быть* 'be' which has an abnormal present tense paradigm including a zero form Ø$^{BE.PRES}$ and overt forms *есть*$^{BE.PRES}$ and *суть*$^{BE.PRES}$ which do not discriminate person and number and are distributed syntactically. I discuss different approaches to the grammar of *быть* and argue that Apresjan's model which recognizes Ø$^{BE.PRES}$, *есть*$^{BE.PRES}$ and *суть*$^{BE.PRES}$ as parts of one and the same lemma is superior to alternative models splitting *быть* split into two lemmas representing copula vs content verb 'be'. The peripheral status of overt present BE-forms compared with Ø$^{BE.PRES}$ in the Russian National Corpus is confirmed by three measures: 1) dispersion of texts where a BE-form occurs; 2) uneven coverage in different persons and numbers; 3) ratio of copular uses vs content verb uses. 1–2 person present tense BE-forms attested in RNC are internal borrowings from Old Russian and Old Church Slavonic, while *есть*$^{BE.PRES}$ and *суть*$^{BE.PRES}$ are inherited 3rd person elements which take over 1–2 person uses. The historical 3Pl *суть* is redundant in a system, where a more frequent 3rd person form *есть* is licensed in the plural: it survives by a minority of speakers either as an optional 3Pl copula in formal discourse or as an emphatic copula in oral discourse. The form *есть*$^{BE.PRES}$ occurs in all persons and numbers both as content verb and as copula but is underrepresented as 3Pl copula: this gap is filled by Ø$^{BE.PRES}$. The frequency of the zero copula Ø$^{BE.PRES}$ can be measured in corpora without syntactic annotation on the basis of systemic proportion between present vs past tense uses of *быть* and on the basis of approximation samples for contexts where overt copulas alternate with Ø$^{BE.PRES}$.

Keywords: corpus linguistics, Russian, parametric grammar, morphology, agreement, lemmatization, copula, zero syntactic elements

DOI: 10.28995/2075-7182-2020-19-795-810

[1] This paper is written with financial support from the Russian science foundation, project RSCI 18-18-00462. I am grateful to the anonymous reviewers for the valuable comments.

НУЛЕВЫЕ ФОРМЫ В МОРФОЛОГИЧЕСКИХ ПАРАДИГМАХ: ГЛАГОЛ «БЫТЬ» В РУССКОМ ЯЗЫКЕ

Циммерлинг А. В. (fagraey64@hotmail.com)
Государственный институт русского языка им. А. С. Пушкина; Институт языкознания РАН; Московский педагогический государственный университет, Москва, Россия

1. The verb *быть* in Russian: grammar, lexicography and frequency

The Russian verb *быть* 'be' has an abnormal present tense paradigm consisting of 3 elements not distributed according to the principle of person-and-number agreement. A salient part of its uses is realized by the zero copula $Ø^{BE.PRES}$, which reduces the frequency of the overt present forms *есть* и *суть*. The lemma *быть* has a lower frequency than comparable lexemes in Standard Average European (SAE) languages. *Есть* and *суть* are historically linked with 3Sg and 3Pl respectively but their usage in Modern Russian does not follow these tags. It is generally acknowledged that *есть* spread over all persons and numbers. Basing on corpus data, I argue that *суть* underwent a similar development. A number of authors [Ščerba 1928]; [Jevgenjeva 1999] suggest that the distribution of $Ø^{BE.PRES} \sim$ *есть* follows the distinction of the copular 'be' vs content verb 'be'. However, the theory that copular *быть* and content word *быть* are different lemmas must be rejected, since both $Ø^{BE.PRES}$ and *есть* are used both as copula and content verb. I measure the ratio of copular vs content verb for each person and number form and argue that the ratio of overt copular sentences gives a key to the part covered by $Ø^{BE.PRES}$. This study is based on Russian National Corpus (RNC). The method requires partial or complete syntactic analysis of contexts involving the present tense forms of *быть* in order to identify them as part of the existential, copular or perfect construction. Direct measurement is possible only for forms with the lowest frequency, in other cases I implement approximation samples based on the next-neighbor method: the adjacent elements often diagnose the type of *быть* construction without look-up of the entire syntactic structure. An advantage of the chosen approach is that it minimizes the role of the text meta-data in a balanced corpus.

1.1. *БЫТЬ* vs *ЕСТЬ* in Russian lexicography

Vladimir Dal's (1880) dictionary claims that *есть* is the 3Sg form of the verb 'be', which is "dropped where other languages use it" [Dal 1880 I: 523]. The 3Pl *суть* is not mentioned. Dmitry Ušakov's dictionary (1935) has two entries— *БЫТЬ* and *ЕСТЬ* in the first volume. The first one claims that *быть* "lacks present tense except

for the 3Sg *есть* and outdated 3Pl *суть* in some meanings" [Ušakov 1935 I: 214]. The second entry tells that *есть* is used in all persons "due to the loss of the old forms of the present tense of *быть*" [ibid., 838]. The fourth volume adds *СУТЬ* introduced as a "bookish and outdated 3Pl of *быть*, primarily used in contexts of enumeration" [Ušakov 1940 IV: 599]. This description hints that *суть* is optional but does not specify, whether it is a variant of *есть*.

Ušakov's description is influenced by Lev Ščerba's theory that copular *быть* and content verb *быть* are different lexemes [Ščerba 1928]. He starts listing the uses of *БЫТЬ* from contexts where "the copula is dropped in the present tense" as in the "position between a subject and a nominal predicate" and in the participle passive [Ušakov 1935 I: 218]. Ščerba's program is implemented in the Minor Academic Dictionary (1957–1961) edited by Anastasia Jevgenjeva. Her description is close to Ušakov, but the entry *БЫТЬ* starts from contexts for the content verb. She claims that *БЫТЬ* lacks present forms "except for the 3Sg *есть* and the outdated 3Pl *суть*" [Jevgenjeva 1999 I: 130–131]. The entry *ЕСТЬ* however admits that *есть* is used in all persons and numbers "due to the loss of the present forms of *быть*" [ibid., 468]. The fourth volume has a short entry *СУТЬ*[2] defined as a bookish 3Pl form occasionally used in 3Sg [Jevgenjeva 1999 IV, 310]: this statement is based on examples like ***Cue***$_{SG}$ *не суть*$^{BE.PRES}$ ***угроза***$_{SG}$ "**This** is not a **threat**" (M. Gorki, 1912), which lack an agreement controller in the plural form.

Andrej Zaliznjak's Grammatical Dictionary tells that *есть* stands for all persons and numbers of *БЫТЬ*, while *суть* is a 3Pl form rooted in scientific or archaic discourse [Zaliznjak 1977: 133]. The loss of the present tense forms is not mentioned. Sergej Ožegov's dictionary revised by Natalia Švedova is close to Ušakov but less consistent. The entry *БЫТЬ* states that this verb lacks present tense "except for 3Sg *есть* and outdated bookish 3Pl *суть*" [Ožegov, Švedova 1992: 64]. The entry *ЕСТЬ* tells that this form spread over all persons and numbers "due to the loss of the old present tense forms" [ibid., 191]. The entry *БЫТЬ* starts from content verb contexts, while the entry *ЕСТЬ* starts from copular contexts. The entry *СУТЬ* claims that this bookish form of 3Pl is now primarily used as a copula, if both arguments are expressed by substantives [ibid., 808].

The author of the most detailed lexicographical description of *БЫТЬ*, [Jurij Apresjan 1996] rejects Ščerba's and Jevgenjeva's theory on two separate BE-lemmas and reinstates one paradigm consisting of 3 present forms: Ø$^{BE.PRES}$, *есть* and *суть*. Ø$^{BE.PRES}$ and *есть* lack person-and-number features, while *суть* is an optional variant of the copular BE but not the content verb BE in 3Pl [Apresjan 1996: 518, 528]. Apresjan shows that both Ø$^{BE.PRES}$ and *есть* have parallel uses as a copula and as a content verb, so that the identification of the copular BE with the hypothetical lexeme selecting Ø$^{BE.PRES}$ in the present tense and the content BE with a different lexeme selecting *есть* is impossible. This description has three advantages: 1) it recognizes Ø$^{BE.PRES}$ as part of the paradigm; 2) it does not stick to historical notions; 3) it does not identify *есть* as a content verb in all its uses. I adopt Apresjan's approach, but argue that his tag for *суть* must be fixed.

1.2. The verb ‘be’ in SAE languages and in Russian: frequency and grammar

The verb ‘be’ is a high frequent word in SAE languages, with a rank comparable to the ranks of the definite article and the conjunction ‘and’. The high rank of the SAE ‘BE’ in the top 5–10 lemmas is due to the fact that it is widely used in three types of contexts:

- Type I contexts: ‘BE’ as a content verb expressing a variety of existential, locative and possessive meanings;
- Type II contexts: ‘BE’ as a copula with different types of nominal predicates;
- Type III contexts: ‘BE’ as an auxiliary element in analytical verb forms and constructions like Germanic, Romance or Slavic BE-perfect or BE-progressive in English etc [2]

In Slovenian [Gigafida], the lemma ‘BE’ heads the list of the most frequent lemmas. In both British English [BNC] and American English [COCA] the lemma ‘BE’ holds the 2nd rank after the definite article. In German it holds the 3rd rank. In Russian [RNC], the lemma ‘BE’ is only the 6th from above, behind *и* ‘and’, *не* ‘not’, *в* ‘in’, *на* ‘on’ and *я* ‘1Sg’ [Lyaševskaja, Šarov 2009]. This results from two deviant features of *быть*. For the first, overt present forms of *быть* lack person-and-number specification which is unusual for SAE languages: English retains full-fledged person-and-number agreement exactly where Russian gives it up— the present tense of *be*. For the second, the most frequent present form of *быть* is $Ø^{BE.PRES}$. The status of $Ø^{BE.PRES}$ as part of the *быть* paradigm in Russian is acknowledged in linguistic typology [Stassen 1994]; [Pustet 2003]. Frequency lists normally ignore zero forms, since taking them into account would require processing uniform syntactic annotations for a family of corpora. Overt present forms of *быть* do not match the frequencies of the non-present forms. *Есть* (393,200 raw hits in RNC) is almost 6 times less frequent than the past tense forms *был, была, было, были* (2,267,476 raw hits). This is predictable since the past tense forms of *быть* correspond both to Ø in Type II contexts and to *есть* in Type I contexts.

Tab. 1: Present vs non-present forms of ***быть*** in Russian

Type of context	Present tense of *быть*	Past tense of *быть*
Type I. The overt present form *есть* is optional or obligatory: ***есть***/*был, -а, -о, -и*.	(1) a. У Ивана **есть** машина$_{SG.F}$. ‘John has a car.’ (2) a. У Ивана **есть** книги$_{PL}$. ‘John **has** books.’ (3) a. Ты и **есть** доктор. ‘You$_{2SG}$ **are** indeed a doctor.’	(1) b. У Ивана **была**$_{SG.F}$ машина$_{SG.F}$. ‘John had a car.’ (2) b. У Ивана **были**$_{PL}$ книги$_{PL}$. ‘John **had** books.’ (3) b. Ты и **был**$_{SG.M}$ доктором. ‘You$_{2SG}$ **were** indeed a doctor.’

[2] Type III contexts must be kept apart from Type II contexts, since Type II sentences always refer to present events, while Type III sentences with present tense BE-auxiliaries in such complex verbal forms as perfect and plusperfect refer to past events.

Type of context	Present tense of *быть*	Past tense of *быть*
Type II. Overt present forms are excluded, the silent form Ø$^{BE.PRES}$ is obligatory: **Ø**$^{BE.PRES}$*/был, -а, -о, -и.*	(4) a. Иван **Ø** $^{BE.PRES}$ умен$_{SG.M}$. 'John **is** intelligent'. (5) a. Ты **Ø** $^{BE.PRES}$ умен$_{SG.M}$. 'You$_{2SG}$ **are** intelligent.' (6) a. Иван и Марья **Ø**$^{BE.PRES}$ умны$_{PL}$. 'John and Mary **are** intelligent.'	(4) b. Иван **был**$_{SG.M}$ умен$_{SG.M}$. 'John **was** intelligent'. (5) b. Ты **был**$_{SG.M}$ умен$_{SG.M}$. 'You$_{2SG}$ **were** intelligent.' (6) b. Иван и Марья **были**$_{PL}$ умны$_{PL}$. 'John and Mary **were** intelligent.'

Neither Ø$^{BE.PRES}$ nor *есть* discriminate number and gender, while past tense forms do. Ø$^{BE.PRES}$ and *есть* do not discriminate person either, cf. (3a) and (5a)[3]. A general prediction for SAE languages is that present forms of BE are more frequent than the non-present ones given that corpora display the same proportion of present and non-present events. This holds both for languages with person-and-number agreement (English, German, Slovenian[4]) and for languages with a single present form (Danish, Swedish). In Danish, the only present tense BE-form, Da. *er* heads the frequency list for all word forms, while in Swedish, the only present tense BE-form, Sw. *är* holds the third rank. Ru. *есть* with its 66th rank in the list of frequent word forms is far behind, which is due to the fact that overt present forms of BE are excluded from all Type II and Type III contexts:

(7) a. Он **Ø**$^{BE.PRES}$ болен.
'He is ill.'
b. *Он **есть**$^{BE.PRES}$ болен.

(8) a. Он Ø$^{BE.PRES}$ арестован.
'He is arrested.'
b. *Он **есть**$^{BE.PRES}$ арестован.

[3] This feature however does not make a contrast with the past tense.

[4] In Slovenian [Gigafida] the lemma *biti* 'be' heads the list of most frequent lemmas with 91,522,113 uses (https://www.clarin.si/noske/run.cgi/view?corpname=gfida20_dedup;usesubcorp=;q=q%5Blemma%3D%3D%22biti%22%5D), whereby 66,247,726 of *biti* sentences (72.38%) have present tense BE-forms, incl. perfect and plusperfect auxiliaries in Type III contexts. Slovenian lacks simple past forms: therefore, the lexical form *bil, bila, bilo, bili* total only 7,750,160 forms (8.46%).

2. The present tense BE-paradigm in Modern Russian: grammar and corpus tags

2.1. Parametric grammar and present tense agreement

The definitional feature of the Russian present tense BE-paradigm is the absence of person agreement[5]. The key for what is recognized as the Modern Russian period is furnished by the extinct Old Russian construction of the *л*-perfect, which required person agreement and overt BE-auxiliaries in the 1–2 p.: *пришел*$_{\text{PART.SG.M}}$ *есмь*$_{\text{1SG}}$ 'I (male) came', *пришла*$_{\text{PART.SG.F}}$ есмь$_{\text{1SG}}$ 'I (female) came', *пришел*$_{\text{PART.SG.M}}$ *еси*$_{\text{2SG}}$ 'you (sg, male) came', *пришла*$_{\text{PART.SG.F}}$ *еси* 'you (sg, female) came', *пришли*$_{\text{PART.PL}}$ *есмѣ*$_{\text{1PL}}$ 'we came', *пришли*$_{\text{PART.PL}}$ *есте*$_{\text{1PL}}$ 'you (pl) came', *пришла*$_{\text{PART.DU}}$ *есвѣ*$_{\text{1PL}}$ 'we two came', *пришла*$_{\text{PART.DU}}$ *еста*$_{\text{2DU}}$ 'you two came'. This construction is incompatible with Russian grammar, since *л*-participles changed their morphological status from nominal to purely verbal forms which do not combine with BE-auxiliaries. Consequently, phrases like *пришел есмь* diagnose borrowed grammar in a language, where *л*-forms are verbal. The *л*-perfect is a Type III structure i.e. an analytical verb form with an auxiliary. For Type II structures with nominal predicates and copular BE the diagnostics is less clear, since the corresponding contexts survive in contemporary Russian.

2.2. Borrowed agreement in the Russian National Corpus: the 1–2 p. of *быть*

The occurrences of historical present 1–2 p. BE-forms in the main corpus of RNC must be explained as borrowings either from Old Russian or Old Church Slavonic. The form *суть* despite the tag 'archaic' assigned by Russian lexicographers is an inherited part of the paradigm. The main corpus of RNC (ca. 1700–) includes some historical present forms of *быти*, which come from dated texts, quotations, parodies or philological commentary. This is confirmed by the limited number of texts where these forms occur: the search for 1Sg *есмь* returns 442 documents and 951 hits, for 2Sg *еси*—538 documents and 1645 hits, for 1Pl *есмы*—85 documents and 129 hits, for 2Pl *есте*—92 documents and 180 hits, for 1Du *есвѣ* and *есва*—just 2 hits in 1 document, for 2Du *еста*—6 hits in 6 documents. These figures are low compared to 3rd p. forms: 3Pl *суть* occurs in 6,329 documents and 3Sg *есть*—in 41,160 documents. In this period, the *л*-perfect is extinct. Russian authors which tried to emulate the Church Slavonic usage occasionally attached agreement markers not to the *л*-participle, which is not specified for person but to the verbal forms that already had inflexional person markers, e.g. to present tense as in (9) or the aorist[6] in (10). Such fail-

[5] The identification of Russian and Hungarian as 'languages with a zero copula' in [Benvenist 1960] does not capture an essential difference between their BE-paradigms. Hungarian just as Old Russian has zero copula in the 3rd p., but overt copulas in 1–2 p. [Bánhidi, Jókay, Szabó 1965: 67–69], while Modern Russian has a 1–3 p. zero present BE-from both in the contexts for a copula and for a content verb [Apresjan 1996: 528; Testelets 2008: 784].

[6] The traditional estimate for the elimination of the aorist in Russian is late XV century [Borkovskij, Kuznetsov 1963: 279].

ures prove that the *л*-перфект did not correspond to the speakers' own idiom. They treat *л*-forms as verbs and combine the dated agreement forms of the auxiliary with verbs on the basis of a wrong analogy: ***наказал***$_{\text{PART.SG.M}}$ ***еси***$_{\text{2SG}}$ *мя*$_{\text{ACC.SG}}$ 'you punished me' → ****наказуешь***$_{\text{PRES.2SG}}$ ***еси мя*** 'you punish me'.

(9) Сосет под ложечкой неимоверно. Господи, за что **наказуешь**$_{\text{PRES.2SG}}$ **еси**$_{\text{2SG}}$ мя? [Влад. Азов. Маленький фельетон. Из дневника дипломата Уступчивого (1908.10.17) // «Русское слово», 1908].

(10) да будут Очи Твои отверсты на Дом сей день и нощь, на Место сие, о нем же **глаголах**$_{\text{AOR.1.SG}}$ **еси**$_{\text{2SG}}$, будет Имя твое тамо, еже услышати молитву [А. И. Богданов. Описание Санкт-Петербурга (1751)].

Unequal distribution of the *л*-perfect confirms that this construction is a borrowing. I checked all forms of 1–2 p. including the dual, which died out in Old Russian ca. 1600. Sequences like *был*$_{\text{PART.SG.M}}$ *еси*$_{\text{2SG}}$ i.e. combinations of a present tense auxiliary with a lexical form of *быть* were excluded.

Tab. 2. The *л*-перфект with 1–2 p. BE-forms in the main corpus of RNC from 1700 A.D.

	1700–1799		1800–1899		1900–1999		2000–...	
	All	*л*-perfect	All	*л*-perfect	All	*л*-perfect	All	*л*-perfect
1Sg: *есмь*	69	0	202	5 (2.5%)	336	102 (97%)	56	3 (5.35%)
2Sg: *еси*	291	169 (58%)	252	103 (40.9%)	487	201 (41.3%)	98	44 (44.9%)
1Pl: *есмы*	57	2 (0.35%)	20	2 (10%)	37	2 (0.54%)	3	0
2Pl: *есте*	57	1 (0.18%)	30	7 (23.3%)	20	1 (5%)	1	0
1Du: *есвѣ, -а*	2	0	0	0	0	0	0	0
2Du: *еста*	1	0	1	0	1	0	1	1

Tab. 2 shows that the *л*-perfect is more or less regularly reproduced in 2Sg, where it makes up 47.3% of the sample. Other combinations are sporadic: 32 hits from total 995 uses (3.21%). There is no substantial increase or decline of frequency in the use of 1–2 p. forms. I conclude that they are lexical borrowings that do not revive the lost mechanism of the person-and-number agreement. The variety of vernacular Old Russian described in [Zaliznjak 2008: 236] lacked overt 3rd p. auxiliaries in the *л*-перфект. Phrases like *пришел*$_{\text{PART.SG.M}}$ ***есть***$_{\text{3SG}}$, *пришли*$_{\text{PART.PL}}$ ***суть***$_{\text{3PL}}$ must be extremely rare in Modern Russian, since the speakers lack inherited grammar for such combinations. This prediction is born out: we found just 4 examples with *л*-perfect in the sample of total 5,040 uses from 1700 A.D. on.

Tab. 3. The *л*-перфект with 3Pl *суть* in the main corpus of RNC from 1700 A.D.

	1700–1799		1800–1899		1900–1999		2000–...	
	All	*л*-perfect	All	*л*-perfect	All	*л*-perfect	All	*л*-perfect
3Pl: *суть*	1.433	2	1.889	0	1.519	2	199	0

The negligable percentage of the *л*-перфект (0.08%) is expected if *суть* and *есть* are part of the BE-paradigm both in the source language(s) and in the target language, but the *л*-перфект is lacking from the target language. The survived uses of *есть* and *суть* correspond not to the *л*-perfect but to Type II contexts (copular BE) and Type I contexts (BE as a content verb).

Tab. 4. The *л*-perfect in Old Russian vs past tense in Modern Russian

	Source languages		Target language
	Old Russian	Old Church Slavonic	Modern Russian
л-forms as past tense markers	part of the analytical construction		single word form
BE-auxiliary in the past tense construction with *л*-forms	agreement marker		absent
Combination of an *л*-form with a 1–2 p. BE-auxiliary, type *пришел есмь*	required		ungrammatical
Combination of an *л*-form with a 3rd p. BE-auxiliary, type *пришел есть, пришли суть*	optional	required	ungrammatical

2.3. Modern Russian *суть*: residual agreement or a redundant present marker?

The form суть is more than 20 times less frequent (16,088 raw hits in the main corpus of RNC) than есть (393,200 raw hits[7]). The verb form *суть*$_2$ must be distinguished from the homonymic noun *суть*$_2$ 'essence' and from the collocation *не суть важн-о, -ое, -а, -ы,-ые* 'does not matter'. Preliminary observations show that *суть*$_1$ and *суть*$_2$ have comparable frequency, but the frequency of *суть*$_1$ increases towards the end of the period, while *суть*$_2$ displays the opposite tendency. As stated above,

[7] The vast majority of the occurrences feature the presence BE-form *есть*$_2$ and not the imperfective infinitive *есть*$_1$ 'to eat'. The search for the parallel perfective infinitive *съесть* 'to eat up' returns only 2,268 hits. There is also a third candidate for the disambiguation—the military response *есть*$_3$! 'I obey', which is an infrequent word.

Russian lexicographers link *суть*$_2$ to scientific or archaizing discourse. This practice is confirmed by the stats: *суть* occurs in the main corpus of RNC only in 6,329 documents vs 41,160 documents for *есть*. The search gives back both *суть*$_1$ 'essence' and *суть*$_2$: texts containing *суть*$_1$ can lack *суть*$_2$ and vice versa. The majority of hits for the search < *суть*$_1$ ∨ *суть*$_2$ > come from non-fiction texts (5,024 documents, 12,703 hits), most of them are from the groups 'journalism' (3,346 documents, 6,454 hits) and 'academic/pedagogical texts' (718 documents, 4,522 hits). Meanwhile, the group 'church and theology' adds only 217 documents and 709 hits.

[Jevgenjeva 1999: IV: 305] treats *суть*$_2$ as an optional form of the 3rd p. primarily used in 3Pl, while [Apresjan 1996] disapproves *суть*$_2$ in 3Sg and treats it is an optional variant of the copular BE in 3Pl [Apresjan 1996: 518, 528]. This model is rendered in **Tab. 5**:

Tab. 5. The present tense paradigm of ***быть*** in Russian, after [Apresjan 1996]

	Ø$^{BE.PRES}$		ЕСТЬ		СУТЬ	
	Content verb	Copula	Content verb	Copula	Content verb	Copula
1Sg.	+	+	+	+	*	*
2Sg.	+	+	+	+	*	*
3Sg.	+	+	+	+	*	*
1Pl.	+	+	+	+	*	*
2Pl.	+	+	+	+	*	*
3Pl.	+	+	+	+	*	+

I checked the main corpus of RNC for contact sequences of the type subject pronoun + суть in the window <–1; 1>. The search was limited by the period 1800–2015 in order to exclude doubts about the grammar of the XVIII century texts. The sample for *суть*$_2$ totals 239 sentences. 3Pl prevail (89.1%), but all other combinations are attested. In the second group, the most frequent combination is 1Pl *мы* 'we' + *суть* (12 examples). In the first group, 26 sentences (12.2% from all 3Pl uses) show *суть* as a content verb, therefore, Apresjan's statement that this option is out must be softened.

Tab. 6. The distribution of ***суть***$_2$. The figures show the number of contact sequences with subject pronouns in the main corpus of RNC from 1800 A.D.

	SG		PL	
	Content verb	Copula	Content verb	Copula
1p.	0	3 (1.25%)	0	12 (5%)
2p.	1 (0.42%)	2 (0.83%)	0	4 (1.67%)
3p.	0	4 (1.67%)	26 (10.87%)	187 (78.24%)

The examination of the RNC examples with *суть*$_2$ in 1–3Sg and 1–2Pl shows that such uses are rooted in the Russian language of 1800–1950. The list of authors includes Ivan Turgenev, Maxim Gorki, Sergei Bulgakov, Ivan Šmelev, Alexander Kuprin, Vyačeslav Šiškov, Konstantin Fedin. In 1950–2000, the list of authors who license *суть*$_2$ in 1–2 p. and in 3Sg includes Nina Berberova, Vladimir Makanin, cf. (11), Strugacki brothers and Iosif Brodskij, cf. (12). This prompts a hypothesis that for a group of speakers *суть*$_2$ survived as part of oral discourse, where it loses the person-and-number specification and assumes the status of an emphatic copula in the meaning 'X is in essence Y'.

(11) Вроде как **все**$_{PL}$ **мы**$_{1PL}$ суть$^{BE.PRES}$ брежневские инвалиды [Владимир Маканин. Андеграунд, или герой нашего времени (1996–1997)]
'It looks like **all of us** are in essence invalids from Brežnev's time.'

(12) Ибо **война**$_{SG}$ суть$^{BE.PRES}$ **эхо**$_{SG}$ кочевого инстинкта.
[И. А. Бродский. Путешествие в Стамбул // «Континент», 1985].
'Since **war** is in essence **an echo** of the nomadic instinct.'

The spreading of the more frequent form *есть*$_2$ over the plural makes a special form of the 3Pl redundant. That the latter survived is due to the tendency towards using *есть*$_2$ and *суть*$_2$ in different contexts. This tendency is captured by Apresjan's model, but the distribution in Tab. 6 has never been achieved because of the opposite tendency towards expanding the coverage of *суть*. This begs an alternative model outlined in Tab. 7 below.

Tab. 7. The present tense paradigm of *быть* in Russian: a corpus alternative

	Ø$^{BE.PRES}$		ЕСТЬ		СУТЬ	
	Content verb	Copula	Content verb	Copula	Content verb	Copula
1Sg.	+	+	+	+	(*)	
2Sg.						
3Sg.						
1Pl.						
2Pl.						
3Pl.					(+)	+

2.4. Ø$^{BE.PRES}$ vs *есть*: syntax and semantics

There is a consensus that Ø$^{BE.PRES}$ is a separate element in syntax but not an elided form of *есть* [Peškovskij 1928: 303]; [Testelets 2008]; [Letučiy 2018]. The correlations between the distribution of Ø$^{BE.PRES}$ vs *есть* and the taxonomic semantic type (existence, possession, location, characterization, identification etc.) are shown in [Arutyunova, Širyaev 1983]. I adopt this analysis with the additions proposed in [Yanko 2000]; [Dymarskij 2018]. [Letučiy 2018] argues that Ø$^{BE.PRES}$ and *есть* are always non-synonymic, so that (13a) presumably means 'John's flat is big', while (13b) means 'John has a big flat'. However, the shift from alienable possession

to characterization is not induced by Ø$^{BE.PRES}$, cf. the conjoined structure 'X has Y and Z' in (14), where the possessive reading is required.

(13) a. У Ивана **Ø$^{BE.PRES}$** большая квартира.
'John has a big flat.' [alienable possession].
Or: 'John's flat is big' [characterization]
b. У Ивана **есть** большая квартира.
'John **has** big flat.' [alienable possession], # 'John's flat is big'

(14) У Ивана **Ø$^{BE.PRES}$** большая квартира в городе и уютный дом в деревне.
'John **has** a big flat in the downtown and a nice house in the village.'
#'John's flat in the downtown is big and his house in the village is nice.'

3. The silent head: measuring the impact of the zero present form

The distribution of the *есть*$_2$ and *суть*$_2$ is constrained by the expansion of Ø$^{BE.PRES}$. It ousted the overt forms from a number of contexts and made them optional elsewhere. The ratio of the Ø$^{BE.PRES}$ vs *есть*$_2$ uses cannot be measured directly in corpora without syntactic annotation, but there are indirect estimates. I measure the distribution of *есть*$_2$ for different persons and numbers in the same context as with *суть*. The search was reduced to the sequences of the type subject pronoun + *есть*$_2$ in the window <–1; 1>. The uses of the content verb *есть*$_2$ vs copula *есть*$_2$ were measured on a separate basis. The default hypothesis is that unequal distribution of *есть*$_2$ reflects the impact of Ø$^{BE.PRES}$ which fills in the gap in certain persons and numbers.

3.1. The proportion of *есть* & *суть* vs Ø$^{BE.PRES}$

The sample for *есть*$_2$ with a contact subject pronoun totals 7,458 sentences. This is ca. 30 times larger than the sample for *суть* in the same context (239). Tab. 8 shows the ratio of copular uses in each combination subject pronoun + *есть*$_2$. A separate line shows how this ratio changes if measured for the pair *есть*$_2$ & *суть*$_2$.

Tab. 8 shows a big increase (>1%) with the adding of *суть*$_2$ only in 3Pl and in 1Pl. The percentage of copular *есть*$_2$ is abnormally low in 3Pl (4.24%), therefore adding 187 sentences with *суть*$_2$ is relevant. The combined ratio for 3Pl (16.01%) is nevertheless low compared to other persons and numbers. This confirms that *суть*$_2$, retains a systemically important status mainly as a 3Pl copula, where *есть*$_2$ is underrepresented. Since *суть*$_2$ is a low frequent word, it does not fully compensate this gap which must be filled by Ø$^{BE.PRES}$. The expectancy of an overt copula is higher in 1–3 Sg. (combined ratio 32.83%) than in the 1–3 Pl. (combined ratio 20.68%). This indicates that copular Ø$^{BE.PRES}$ is especially salient in the plural. The positions of *есть*$_2$ as a content verb are stable both in Sg and in Pl. The share of all uses in the 1–2 p. (both content verb and copula) is ca. 5 times less compared to the 3rd p.[8]: the figures are almost identical for Sg (21.03%) and Pl (20.64%). The ratio of the copular uses in 1Sg and

[8] In a sample including non-pronominal subjects, the contrast is even sharper.

2Sg is nevertheless high. I interpret this as a proof that *есть* is stable in these person-and-number forms.

Tab. 8. The distribution of ***есть*** and ***суть***. The figures show the number of contact sequences with subject pronouns in the main corpus of RNC from 1800 A.D. The percentage shows the ratio of content verb vs copular uses.

	SG		PL	
	Content verb	Copula	Content verb	Copula
1p.	440 (56.85%)	334 (43.15%)	191 (84.1%)	67 (25.9%)
w. *суть*	440 (56.13%)	337 (43.77%)	191 (70.75%)	79 (29.25%)
2p.	345 (48%)	375 (52%)	147 (61%)	94 (39%)
w. *суть*	346 (47.76%)	377 (52.14%)	147 (60%)	98 (40%)
3p.	All gender forms: 3,023 (72.68%) 3Sg.M 1194 (79.51%) 3Sg.F 1111 (75.48%) 3Sg.N 718 (61.27%)	All gender forms: 1,144 (27.32%) 3Sg.M 327 (21.49%) 3Sg.F 363 (24.62%) 3Sg.N 454 (38.73%)	1,243 (95.76%)	55 (4.24%)
w. *суть*	3,023 (72.48%)	1,148 (27.52%)	1,269 (83.99%)	242 (16.01%)
Total: 7,458	3,808 (67.27%)	1,853 (32.73%)	1,581 (77.98%)	216 (12.02%)
w. *суть*: 7,697	+ 1 3,809 (67.17%)	+ 9 1,862 (32. 83%)	+ 26 1,607 (79.32%)	+ 203 419 (20.68%)

The approximation does not provide absolute figures for $Ø^{BE.PRES}$, but heuristic estimates can be given. One of them is based on the next-neighbor method, which requires a lookup of the left and right context for the pivotal subject element …X… in order to check whether the right or left neighbor of X is its complement in the verb phrase $[Ø^{BE.PRES}—Y]_i$ linked with X_i. If the search is oriented to identifying the right neighbor as complement of the silent head $Ø^{BE.PRES}$ and X is the 1Sg subject pronoun $я_{1SG}$ 'I', sentences like *Это* $Ø^{BE.PRES}$ $я1_{SG}$, ***Иван***$_{NOM}$ 'That is me, **John**' will return 'false', while sentences like *Все-таки* $я_{1SG}$—$Ø^{BE.PRES}$ *дурак* 'Still, I am a fool' return 'true'. If one takes the context subject pronoun + noun in the nominative case in the window <1; 1>, the expectancy that these elements are part of the structure S_{pron}—$Ø^{BE.PRES}$—S_{NOM}, where the pronoun is the subject and its right neigbour it is part of its nominal complement can be measured. I checked sequences of the type 1Sg subject personal pronoun ***я***$_{1SG}$ + noun in the nominative case: the RNC search returns 78,676 raw hits. A test sample of 2,000 sentences dated with 1987–2015 was processed. The input had

wrong morphological tags fixed by the annotator manually. The lexical-grammatical search in RNC returns all elements which can be analyzed as nouns in the nominative case, cf. the adjective *рада* 'glad' (cf. the noun *рада* 'Ukrainian parlament'), adverb *дома* 'at home', cf. the noun *дома* 'houses', preposition *перед* 'in front of', cf. the noun *перед* 'the front end' as well all syncretic forms that can either stand for nominative or some other case. The sample also included sentences where *я* and its right neighbor belong to different clauses and other structures that do not match the pattern S_{pron}—$Ø^{BE.PRES}$—S_{NOM}. Sentences where the entire structure [$Ø^{BE.PRES}$—S_{NOM}] was located to the left from *я* were filtered out, since the right neighbor of *я* is not a complement of $Ø^{BE.PRES}$. At the same time, blind hits with expressions wrongly tagged as S_{NOM}, cf. ***Я рада***$_{ADJ.SG.F}$ 'I am **glad**' or ***Я дома***$_{ADV}$ 'I am **at home**' were rendered positive, if they matched the pattern with the proviso that the predicate complement is not a noun but an non-verbal element of different morphology. The trimmed sample returns 49.25% positive examples (985 from 2,000). If this ratio holds for the whole RNC collection in the searched context, it should include 38,748 sentences with the subject in 1Sg, zero copula $Ø^{BE.PRES}$ and the word order $Я_{1SG}$—$Ø^{BE.PRES}$ … S_{NOM}/PRED.

One more estimate is based on the proportional usage of past and present tense forms of *быть*. As stated in **1.2.**, overt past tense forms *был, была, было, были* partly correspond to overt present forms *есть*$_2$ and *суть*$_2$ (Type I contexts) partly—to $Ø^{BE.PRES}$ (Type II contexts). Let us assume that RNC has at least as much Type I sentences in the past tense as in the present tense. Let us also assume that all uses of *есть*$_2$ and *суть*$_2$ pattern with Type I structures and all uses of $Ø^{BE.PRES}$ pattern with Type II structures: this simplification maximizes the number of sentences with *есть*$_2$ and *суть*$_2$. If there are m sentences with *есть*$_2$ and *суть*$_2$ and n sentences with *был, была, было, были*, the number of sentences with $Ø^{BE.PRES}$ is n—m = k. The 4 forms *был, была, было, были* return 2,267,476 raw hits. These verb forms have two homonyms—the particle *было*$_2$ 'marker for a canceled event' and the nominal form *были* from the lemma *быль* 'legend'. Both are low frequent words: let us assume that they take maximum 7,476 hits, which is actually above than their frequency. Then we get 2,600,000 uses of the past tense forms of *быть* after the disambiguation. The present form *есть*$_2$ (393,200 raw hits) has homonyms *есть*$_1$ 'to eat' and *есть*$_3$ 'I obey' [military command]: the exact figures are not available, since the search returns the homonyms, but one can assume that the frequency of *есть*$_1$ corresponds to the frequency of its perfective correlate *съесть* 'to eat up' (2,268 hits) and *есть*$_1$ and *есть*$_3$ total maximum 3,200 hits. Then *есть*$_2$ gives 390,000 hits after the disambiguation. The present form *суть*$_2$ (16,088 raw hits) has a homonymic noun *суть*$_1$ 'essence', they have a comparable frequency. Let us assume that *суть*$_1$ takes maximum 8,088 hits. Then we are left with 8,000 hits of *суть*$_2$. With these stipulations, RNC should feature at least k = 1,862,000 sentences with $Ø^{BE.PRES}$, since n = 2,260,000 and m = 398,000. With the stipulations made, all these RNC sentences with $Ø^{BE.PRES}$ will be interpreted as copular, though in reality a minor part from 1,862,000 sentences are structures with a zero content verb in contexts like *У него* $Ø^{BE.PRES}$ *много книг* 'He has many books'.

Finally, the estimates for sentences with $Ø^{BE.PRES}$ in large corpora can be derived on the basis of tree banks with syntactic annotation. Such estimates however reflect the architecture of the parser. Apresjan's model of *быть* adopted in this paper

is implemented in the ETAP-3 parser [Apresjan et alii 2003]. The present forms Ø and *есть* are recognized here as separate elements, but both of them belong to the lemma *ЕСТЬ*, while all non-present forms of BE are linked to a different lemma—*БЫТЬ*[9]. In a parser based on Ščerba's model, $Ø^{BE.PRES}$ and *есть* will be linked to different lemmas. Since the notion of the zero element is non-neutral, any technical decision has impact on processing the coverage of $Ø^{BE.PRES}$.

3.2. Morphological paradigms and historical corpora

The present tense paradigm of *быть* 'be'—{$Ø^{BE.PRES}$, *есть*, *суть*} is historically a transition from an agreement system characteristic of Old Russian to a system without overt present BE-forms. It is surprisingly stable: the overt forms *суть* and *есть* did not disappear during the last 300 years. The historical 1–2 p. forms of *быть* behave as borrowed elements already in the XVIII century. The loss of number agreement in the 3rd p. is not a new phenomenon either. The XVII century traveler Pjotr Tolstoj (b. 1645) in his diary included in the Historical corpus of RNC uses $суть_2$ 4 times with plural nouns and 6 times with singular nouns, cf. (15).

(15) **Варшава**$_{SG}$ <u>суть</u>$^{BE.PRES}$ **место**$_{SG}$ **великое**$_{SG}$,
на левом берегу реки Вислы положенное.
[Путешествие стольника П. А. Толстого по Европе. 1697–1699 (1699)]
'**Warsaw** <u>is</u> [lit: <u>are</u>] **a big city** founded on the left bank of the Wisla-river.'

P. Tolstoy's treatment of $суть_2$ as an emphatic copula does not differ from the XIX–XX century examples (11) and (12). It is plausible that an idiom of Russian with such settings for $суть_2$ existed during a long time but was suppressed by Church Slavonic which only approved $суть_2$ in 3Pl.

The history of Russian 'BE' can be modeled on the basis of its usage in the Modern Russian period, if one takes into account three blocks of input data for each present BE-form: 1) frequency and number of texts, where this BE-form is attested; 2) even vs uneven distribution of BE-forms for different persons and numbers; 3) even vs uneven distribution of the copular vs content verb uses for each BE-form. If one adopts the hypothesis that the Russian present tense BE-paradigm {$Ø^{BE.PRES}$, *есть*, *суть*} originates from a paradigm where all elements were genuine agreement markers, its restructuring follows three steps:

- I. The 1–2 p. forms disappear first;
- II. The uses of *суть* get restricted by the pair {$Ø^{BE.PRES}$, *есть*};
- III. The uses of *есть* get restricted by $Ø^{BE.PRES}$.

This model allows making two predictions concerning the past and the future of the BE-paradigm:

(i) $Ø^{BE.PRES}$ is historically a 3rd person form and an inherited part of the BE-paradigm.
(ii) The form $есть_2$ will disappear from the paradigm of BE in the future.

[9] This decision is commented in [Apresjan 1996: 528].

The hypothesis (i) is in line with historical linguistics: the latter confirms that Old Russian, unlike Old Church Slavonic lacked overt copulas in the 3rd p. in Type III contexts (*л*-перфект) [Borkovskij, Kusnetsov 1963: 203] and partly also in the Type II contexts (copular structures with nominal predicates) [Zaliznjak 2008: 259–261]. The prediction (ii) is in line with Russian dictionaries, which claim that *быть* has no present forms except for the *есть*$_2$ which probably is a separate verb. This description does not hold for the present-day Russian BE-paradigm but anticipates its future.

4. Conclusions

The undertaken study has shown that in a language where zero syntactic forms gradually replace overt forms the status of endangered forms is revealed by two measures: 1) low frequency and uneven distribution in the texts; 2) uneven distribution in different persons and numbers. The history of the Russian BE-paradigm requires a third one and a more specific measure—3) uneven distribution of copular vs non-copular uses in each person and number form. The coverage of the zero copula $\emptyset^{\text{BE.PRES}}$ in Modern Russian can be processed in corpora without syntactic annotation on the basis of systemic proportions between different types of syntactic contexts.

References

1. *Apresjan Ju. D.* (1996), Lexicographical portraits. On the basis of the verb byt’ [Lexikografičeskie portrety (na materiale glagola byt’], Integral description of language and systemic lexicography [Integral’noe opisanie jazyka i sistemnaja leksikografia]. LRC publishing house, Moscow, 1996, pp. 518–537.
2. *Apresjan Ju. D., Boguslavsky I., Iomdin L., Lazurskij A. V., Sannikov V. Z., Sizov V. G., Tsinman L. L.* (2003), ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT, First International Conference on Meaning-Text Theory (MTT’2003). June 16–18, 2003. Paris: Ecole Normale Superieure, 2003, pp. 279–288.
3. *Arutyunova N. D., Širyaev E. N.* (1983) Russian sentence. The existential type [Russkoe predloženie. Bytijnyj tip], Nauka, Moscow.
4. *Bánhidi Z., Jókay Z., Szabó D.* (1965). Learn Hungarian. 3rd ed. Budapest.
5. *Benvenist E.* (1960). ‘Etre’ et ‘avoir’ dans leur fonctions linguistiques, Bulletin de la Société de linguistique, 1960, tome LV.
6. *Borkovskij V. I., Kuznetsov P. S.* (1963), Historical Russian grammar [Istoričeskaja grammatika russkogo jazyka], Nauka, Moscow.
7. *Dal V. I.* (1880), Explanatory dictionary of the living Great Russian language [Tolkovyj slovar’ živogo velikorusskogo jazyka]. Vols. I–IV. 2nd ed. Moscow: M. O. Wolf publishing house, Moscow.
8. *Dymarskij M. Ja.* (2018), A u menja v karmane gvozd’. Zero copula or predicate ellipsis? [A u menja v karmane gvozd’. Nulevaja svjazka ili ellipsis skazuemogo?], World of the Russian word [Mir russkogo slova], 2018, № 3, pp. 5–12.
9. *Jevgenjeva A. P.* (1999), A dictionary of Russian [Slovar’ russkogo jazyka], in 4 vols. 4th ed. Russkij jazyk, Moscow.

10. *Letučij A. B.* (2018), Zero copula [Nulevaja svjazka], Russian corpus grammar [Russkaja korpusnaja grammatika] http://rusgram.ru/.
11. *Lyaševskaja O. N., Šarov S. A.* (2009). Frequency dictionary of Modern Russian (on the basis of Russian national corpus) [Častotnyj slovar' sovremennogo russkogo jazyka (na material National'nogo korpusa russkogo jazyka)], Azbukovnik press, Moscow.
12. *Ožegov S. I., Švedova N. Ju.* (1992). Explanatory dictionary of Russian [Tolkovyj slovar' russkogo jazyka], Moscow.
13. *Peškovskij A. M.* (1928), Russian syntax in a scientific perspective [Russkij sintaksis v naučnom osveščenii].3rd ed., Moscow-Leningrad.
14. *Pustet R.* (2003), Copulas. Universals in the Categorization of the Lexicon. OUP, Oxford.
15. *Ščerba L. V.* (1928), On parts of speech in Russian [O častjah reči v russkom jazyke], Russian speech, [Russkaja reč], New series, II. Academia publishing house, Leningrad, pp, 5–27.
16. *Stassen L.* (1994), Typology versus mythology: the case of the zero copula, Nordic Journal of Linguistics, 17, pp. 105–126.
17. *Testelets Ya. G.* (2008), Syntactic structures with a covert copula in Russian [Struktura predloženij s nevyražennoj svjazkoj v russkom jazyke], Dynamic models: Sentence and text. In honour of E. V. Padučeva [Dinamičeskie modeli: Predloženie. Tekst. Sbornik statej v čest' E. V. Padučevoj], LRC publishing house, Moscow, pp. 773–789.
18. *Ušakov D. N.* (1935–1940), Explanatory dictionary of Russian [Tolkovyj slovar' russkogo jazyka], Vols. I- IV. Moscow.
19. *Yanko T. J.* (2000), Existence and possession: constructions with the verb byt' [Bytovanie i obladanie: konstrukcii s glagolom byt'], N. D. Arutynova, I. B. Levontina (eds.). Logical analysis of language. Languages of space. [Logičeskij analiz jazyka. Jazyki prostranstv], Indrik publishing house, Moscow, pp. 198–211.
20. *Zaliznjak A. A.* (1977), Grammatical dictionary of Russian [Grammatičeskij slovarj russkogo jazyka]. Nauka publishing house, Moscow.
21. *Zaliznjak A. A.* (2004), Old Novgorod dialect [Drevnenovgorodskij dialect], 2nd ed. MLRC publishing house, Moscow.
22. *Zaliznjak A. A.* (2008), Old Russian enclitics [Drevnerusskie enklitiki], LRC publishing house, Moscow.
23. *BNC*—British National Corpus: https://www.english-corpora.org/bnc/. Accessed at 31.03.2020.
24. *COCA*—Corpus of Contemporary American English: https://www.english-corpora.org/coca/.
25. *Gigafida*—http://www.gigafida.net/. Accessed at 31.03.2020.
26. *ETAP-3* —Linguistic processor ETAP-3: http://proling.iitp.ru/ru/etap3.
27. *RNC*—Russian National Corpus: http://www.ruscorpora.ru/. Accessed at 31.03.2020.