JOURNAL of SLAVIC LINGUISTICS

VOLUME 30 · ISSUE NO. FASL 29 EXTRA ISSUE

SOV in Russian: a corpus study

NATALIA SLIOUSSAR AND ILYA MAKARCHUK
HSE Moscow & Saint Petersburg State University

ABSTRACT

This paper analyzes the SOV order in Russian. Various hypotheses concerning its distribution have been proposed in previous functional and formal studies, but none of them became widely accepted. We tested these hypotheses on the large "Taiga" corpus and found that the main factor that triggers SOV is pronominalization: if the object is pronominal, it is highly likely to be preverbal. The absolute majority of non-pronominal objects follow the verb, although both giveness and contrastive, emphatic or narrow focus increase their (altogether very small) chances to be preverbal. Thus, the factors discussed in many previous studies play a role, but this role is extremely small. We propose a syntactic account to capture different information-structural properties of preverbal objects and the optionality of this construction.

KEYWORDS word order · information structure · Russian · corpus · SOV

1 INTRODUCTION

Word order alternations in Russian have been extensively analyzed in the literature. Most studies agree that information structure (IS) is the primary cause of these alternations. However, it is not always clear which IS distinctions or other interpretational constraints can be associated with a particular word order. The present paper focuses on the SOV order, which is especially interesting in this respect.

As we show in §2, SOV sentences have sparked fierce controversy. Some authors argue that they have particular IS properties (e.g. King 1995, Kovtunova 1976, Titov 2020). Others claim that they are characteristic of texts of a particular style or genre (Sirotinina 1965, Kodzasov 1989, 1996). However, the role of these factors should be limited at best: it is impossible to come up with a context in which the SOV order would be obligatory. Interpretational nuances associated with it are very subtle.

We believe that while some simpler grammatical phenomena can be analyzed based on individual examples, to study this complex picture, a large-scale corpus study is necessary. In §3, we present a study based on the "Taiga" corpus (Shavrina & Shapovalova 2017). The corpus contains several large, stylistically diverse subcorpora and has syntactic annotation in the Universal Dependencies framework. Despite some limitations, this annotation proved to be very effective for the purposes of our study. We integrate the results of this study into a syntactic analysis in §4 and draw conclusions in §5.

2 PREVIOUS STUDIES

Sirotinina (1965) studied Russian word order, relying on the Prague school information-structural tradition. Her approach to syntax was different from what we find in modern formal or functional studies: rather than analyzing when SVO, SOV, or other orders are used, she analyzed how often the object precedes or follows the verb and which factors affect this probability. In other words, in her data, SOV, OSV and OVS orders, which have distinct syntactic and IS properties, are combined. Nevertheless, she conducted one of the first corpus studies of written and spoken Russian (recording many monologues and

dialogues for this purpose) and made many important observations that were confirmed in subsequent work.

Firstly, Sirotinina noted that OV orders are more frequent in spoken Russian than in written Russian. Secondly, the dataset of spoken Russian that she had could be divided into more formal and less formal texts, and the share of OV orders was larger in the latter. Thirdly, pronominal objects were preverbal more frequently than non-pronominal ones.

At the same time, Sirotinina demonstrated that both given and new objects can precede and follow the verb. Several examples are provided in (1)–(5) (examples (1)–(3) are from Sirotinina 1965, 49–50, (4) is from Kodzasov 1996, 184, and (5) from Zemskaya 1978, 43). A new object follows the verb in (1-a), but precedes it in (3)–(5). A given object is before the verb in (1-b), but after it in (2). Capital letters are used to mark the main, or sentential, stress.

(1) a. A: A u nas segodnja načali vyvoziť urny.
and at us today started to-remove litter-bins.ACC
'At our place, they started to remove litter bins today.'
b. B: Interesno, kuda oni ėti urny denut?
interesting where they these.ACC litter-bins.ACC will-put
'I wonder, where are they going to put these bins?'

(2) Po-moemu, u menja tože estʹ. Ili ja videla ėtu knigu.
in-my-opinion at me also is or I saw this.ACC book.ACC
'It seems to me, I also have [it] (looking through a list of books). Or I saw this book.'

(3) Vesʹ denʹ odno morožennoe ela.
whole day only ice-cream.ACC ate
'I ate only ice-cream the whole day.'

(4) Ty RUBAŠKI stiratʹ sobiraešsja?
you shirts.ACC to-wash intend
'Are you going to wash (any) shirts?'

(5) Torgovcy, vladelʹcy larʹkov... Potom... DAČI sdavali.
merchants owners kiosks.GEN then summer-houses.ACC rented-out
'(They were) merchants, owners of kiosks... Then... (They) rented out summer houses.'

The share of OV orders in Sirotinina's spoken Russian data was very high: 60.9% for given objects and 40.3% for new objects. This led Slioussar (2007) to conclude that colloquial Russian is undergoing a diachronic shift from VO to OV. She hypothesized that it is becoming one of the languages in which communicatively prominent or salient information, both given and new, precedes the verb, while less salient information follows it.

Let us note that despite word order variation, new objects always bear the sentential stress, while given objects are destressed, i.e. this aspect of information structure is encoded unambiguously. Following Szendrői (2001) and other authors, one can view shifting the sentential stress from its default sentence-final position in (2) and moving the object out of this position in (1-b) as two operations with the same goal: destressing the object to mark its giveness. In the present study, we will not try to develop this approach or to refute it, focusing on a different problem that it does not address: which factors influence the choice of syntactic movement?

In (3)–(5), new objects retain the main stress after movement, so the nature of this movement may be different. As we will show below, some factors discussed in the literature are applicable to OV orders only with given or only with new objects, while others cover both groups. Therefore, in our corpus study we will first consider the latter and then the former, and will come back to the distinction between OV orders with given and new objects in §4, where a syntactic analysis is suggested.

Kovtunova (1976) formulated the rules of stress placement in different syntactic constructions depending on their word order and IS. In particular, she described clauses in which the main stress does not fall on the final word, like (3)–(5), as expressive. A similar distinction was adopted by Yokoyama (1986). Among formal syntacticians, King (1995) was the first to analyze the SOV order in Russian. Following Yokoyama (1986), she assumed that so-called emotive foci tend to be preverbal. In subsequent studies, this position was also associated with contrastive focus. In particular, Titov (2020) discussed movement of contrastive and emphatic foci to different syntactic positions, resulting in SOV, OSV and other word orders. She argued that contrastive or emphatic foci can move or remain in situ (i.e. their movement is always optional), while new information foci that lack these properties never move. She used the ¾ signature principle originally formulated by Bobaljik & Wurmbrand (2012) in her account: the neutral word order is compatible with different interpretations, but movement serves to exclude some options.

In the Russian functionalist linguistic tradition, Kovtunova's (1976) observation on the expressive nature of the sentences in which the main stress is non-final was disputed in many subsequent studies, including Bonnot & Fougeron (1982), Kodzasov (1989, 1996), and Yanko (2001). Their authors showed that in spoken Russian, OV orders do not sound expressive. Similarly, several formal syntacticians have argued that preverbal foci do not have any special properties: they are not necessarily contrastive, narrow or exhaustive (e.g. Dyakonova 2006).

To appreciate their point, let us look at examples (3)–(5) above. The terms contrastive and emphatic are well defined, and we can test whether the foci in different examples satisfy these definitions. A contrastive focus must be associated with a set of alternatives, which may be explicitly mentioned in the text or be salient in the discourse context. The notion of emphasis also presupposes that the referent of the focused constituent belongs to a contextually salient set of alternatives. It must occupy an extreme scalar position with respect to all other members: it must be either the least or the most expected alternative in this set (Titov 2020).

In (3), the whole sentence is in focus, but the moved object inside this wide focus is definitely associated with a set of alternatives, which is signaled by the word odno meaning 'only'. But no contextually salient set of alternatives seems to be associated with the objects in (4) or (5). Judging from the dialogue in which (4) was used, shirts are not the most or the least likely, the most or the least difficult thing to wash, washing shirts does not stand out in any way among other household chores etc. In (5), the speaker enumerates how local people of the Karelian isthmus could earn money before the October Revolution. The list is clearly not exhaustive, many items in this list can overlap (for example, a kiosk owner could also rent out a summer house), the VP with a fronted object does not appear to be juxtaposed to other alternatives. The object itself also does not seem to be associated with a contrastively salient set of alternatives. We would say that this OV sentence has a garden variety new information focus. Unlike the terms contrastive and emphatic, the term expressive does not have a strict definition. But, intuitively judging, (4) and (5) sound absolutely neutral in their contexts.

Why is the idea that sentences with new preverbal objects have some special properties so pervasive in the literature, coming up in different linguistic traditions? As far as we can judge, when SOV examples are presented out of context in linguistic papers, they indeed seem expressive, emphatic or otherwise marked. Among other things, this is confirmed by the results of a questionnaire reported by Titov (2020). But when naturally occurring examples like (4) or (5) are considered in their contexts, this impression disappears.[1] To explain this, some authors have argued that OV orders are more characteristic of

[1] Let us add that a similar discussion has taken place in other languages, for example, in Italian. Several authors claim that sentence-final focused constituents are interpreted as presentational, essentially conveying new information in a neutral way, while fronted foci are contrastive (e.g. Zubizarreta 1998, Belletti 2002). At the same time, Brunetti (2004) presents many examples in which this generalization does not hold. Of course, the conclusions reached for Italian cannot be blindly applied to Russian, so we mention this discussion only to show that this is a complicated question that cannot be decided by analyzing individual examples.

colloquial Russian (e.g. Slioussar 2007) or of dialogues as opposed to narratives (see below) and therefore look marked in the written formal non-dialogic text of a linguistic paper. After presenting different ideas from earlier studies in this section, we will come back to this question in §4.

Completing the discussion of different types of foci, let us note that several authors have claimed that all affirmative sentences with non-final main stress have expressive, contrastive, emphatic or otherwise marked interpretations in Russian (Kovtunova 1976, Yokoyama 1986, Titov 2020, e.g.). Kodzasov (1989, 1996), Yanko (2001) and other linguists working with Russian corpora dispute this generalization not only for SOV sentences. Starting with Sirotinina (1965), many authors have noted that in colloquial Russian stress shift is much more widespread than in written Russian, which reorders constituents to place new information at the end of the sentence and preserves the sentence-final main stress in the majority of cases. These observations are summarized in a collective monograph on colloquial Russian (Zemskaya 1973) and elsewhere. For example, in a sentence with a new subject and a given object, the OVS order will be used in written Russian, while colloquial Russian will prefer SVO with a stress shift. As a result, in many sentences, including those with new information focus, the main stress is not in the final position.

Other previous studies that are relevant for our paper include Mykhaylyk (2010, 2011). Working on Ukrainian, which is a close relative of Russian, she considers such factors as object definiteness, referentiality and specificity (analyzing both given and new objects). Mykhaylyk argues that specificity is the most relevant property: preverbal objects are interpreted as specific. If we try to extend this approach to Russian, we see that it does not cover many cases in which the preverbal object is in focus: focused elements may be specific, but in (3), the object is definitely non-specific. At the same time, many specific objects follow the verb, as in (2).[2]

Finally, several authors, primarily Kodzasov (1989, 1996), rely on the distinction between dialogues and narratives rather than written vs. spoken or formal vs. informal texts. This distinction was introduced by Benveniste (1974) to explain the choice between two past tenses in French: passé simple and passé composé. Kodzasov assumes that for narratives, in which it is more difficult for the speaker and the addressee to establish the common ground, text coherence is crucial. Therefore, it is preferable to place given information before new and, if the predicate and the object have no IS distinctions, to place the predicate before the object. For dialogues, coherence is also important, but the speaker has more freedom and can communicate the most important information (new, less predictable) first. As a result, dialogues have a larger share of OV orders.

Summarizing the overview of previous studies, we can conclude that most authors have viewed SOV as optional, e.g. arguing that contrastive or emphatic foci may be associated with syntactic movement (but may also stay in situ), that given objects may be moved out of the sentence-final position (but may also be destressed as a result of stress shift), or that the share of OV orders is larger in colloquial, informal or dialogical texts (but VO orders abound in them as well). In the present paper, we wanted to test at least some of these factors in a corpus study. We already demonstrated above why turning from analyzing singular examples to large datasets of naturally occurring sentences is important. In addition, all previous corpus studies of Russian word order analyzed only the incidence of postverbal and preverbal objects, i.e. such word orders as SOV, OSV or OVS were counted together, while it is generally assumed that they have different syntactic, prosodic and IS properties. Our corpus study is the first to estimate their frequency separately.

[2]The present paper discusses only Russian data, but as far as we can judge from consulting several Ukrainian speakers, Ukrainian also allows for examples like (2) and (3). Nevertheless, judging from informal observations, the overall distribution of OV orders seems to be different in the two languages, at least to a certain extent, which could be an interesting topic for further investigation.

3 CORPUS STUDY

3.1 METHOD

To shed new light on the nature of SOV sentences, we conducted a large-scale corpus study. We used the Taiga corpus (Shavrina & Shapovalova 2017). Firstly, it contains many stylistically diverse subcorpora, about 5 billion words in total. For our study, we selected the News subcorpus (92 million words), the Social Media subcorpus (80 million words) and the Subtitles subcorpus (101 million words). These datasets are large enough to trace even subtle tendencies in word order distribution.

Secondly, the Taiga corpus has been automatically annotated using the Universal Dependencies framework (https://universaldependencies.org/). This framework has well-known pluses and minuses: it has been adapted for many typologically diverse languages, which makes it a great tool for cross-linguistic comparisons, but it is often criticized for being oversimplistic. Moreover, automatic annotation produces many errors. To circumvent this problem, we decided to limit our dataset to the examples that could be identified accurately: from all subcorpora, we extracted clauses that contained only a finite verb, a nominative subject DP and a direct object DP. Even with these constraints, the resulting dataset was very large. It contained 1,050,900 clauses, and 98,835 (9.4%) of them had the SOV order.

As we noted in the previous section, sentences with the same word order may have different syntactic structure, and this cannot be controlled for in a corpus study. Being aware of this limitation, we will use our dataset to estimate how frequent the SOV order is in general and which factors affect its frequency. Then we will rely on the conclusions we reached in discussing a syntactic analysis.

In our study, we examined the factors (i)–(vii). The dependent variable was the word order. For the sake of simplicity, we focused on comparing SOV vs. SVO orders (whether a particular factor significantly affects their distribution).

(i) formal vs. informal style;

(ii) written vs. spoken texts;

(iii) prominalization of the subject and the object (nouns vs. pronouns);

(iv) grammatical features of the object and the verb (number, animacy, tense and aspect, finiteness);

(v) syntactic complexity of the object DP (whether the head noun is modified by adjectives, dependent DPs, relative clauses etc.);

(vi) syntactic properties of the clause (matrix vs. subordinate, presence of negation);

(vii) IS properties of the object (although in general a corpus study is not well suited to study them, we will explain in §3.2 how they could be determined at least in some cases).

The factors in (iv) and (vi) have not been mentioned in previous studies, and we had no hypotheses about their role — we analyzed them because our dataset allowed us to do so, and did not find any effects. In all other cases, some significant results were obtained. For the statistical analysis we used the $\chi 2$ test when individual lexical items were considered (e.g. object DPs with and without the modifier *ėtot* 'this'). When larger datasets were analyzed (e.g. the differences between the three selected subcorpora), even minor differences reached significance according to the $\chi 2$ test. So we complemented it with Cramer's V used to estimate the effect size. For the datasets discussed in the present study, $V \geq 0.5$ means a large effect, 0.3–0.5 is medium, 0.1–0.3 is small, and $V < 0.1$ means that the difference is negligible (Mangiafico 2016).

News		Social Media		Subtitles	
SVO	82.9%	SVO	65.5%	SVO	63.3%
OVS	7.1%	SOV	14.4%	SOV	18.2%
OSV	3.6%	OSV	9.2%	OSV	14.4%
SOV	2.7%	OVS	6.3%	OVS	2.5%
VOS	2.1%	VOS	2.4%	VOS	1.0%
VSO	1.6%	VSO	2.2%	VSO	0.6%
Total	540,531	Total	234,535	Total	275,834

Table 1: Clauses containing a subject, a direct object and a verb in different subcorpora.

	SVO and SOV		SVO		SOV		% SOV
Nominal S and O	425,497	49%	420,807	54%	4,690	5%	1%
Pronominal S, nominal O	276,261	32%	265,118	34%	11,143	11%	4%
Nominal S, pronominal O	35,299	4%	23,306	3%	11,993	12%	34%
Pronominal S and O	138,063	16%	67,054	9%	71,009	72%	51%
Total	875,120		776,285		98,835		

Table 2: SVO and SOV orders with nominal and pronominal arguments.

3.2 RESULTS

Formal vs. informal and written vs. spoken texts. To study the role of these factors, we compared the three selected subcorpora. None of them contains fiction. The News subcorpus represents formal written Russian, the Social Media subcorpus represents informal written Russian, and the Subtitles subcorpus including subtitles from TV shows brings us as close as we can get to spoken Russian (no large Russian corpora contain transcriptions of spontaneous everyday speech or similar data). Unfortunately, we could not automatically divide texts into dialogues and narratives, but it can be reasonably assumed that the Social Media and especially Subtitles subcorpora contain more dialogues than the News subcorpus. The distribution of different orders in the three subcorpora is shown in Table 1.

Table 1 shows various differences between the three subcorpora. In the Social Media and Subtitles, SOV order is the second most frequent, although it is still much less widespread than SVO, while in the News, OVS and OSV are more frequent than SOV.[3] In the News vs. Subtitles comparison, the effect was large ($\chi 2>100$, $p<0.01$, $V=0.30$), in the News vs. Social Media comparison, there was a medium effect ($\chi 2>100$, $p<0.01$, $V=0.25$), while the differences between Subtitles and Social Media were negligible ($\chi 2>100$, $p<0.01$, $V=0.05$).

Pronominalization. If SVO and SOV clauses are analyzed together, about half of the sentences have no pronominal arguments, and in another one third of them, the subject is a pronoun. As Table 2 shows, the distribution is relatively similar in SVO clauses alone, but in SOV clauses, it is dramatically different: in most of them, both arguments are pronouns. If we take a different perspective, when both arguments are nominal, the probability of SOV as opposed to SVO is 1%. It increases slightly when the subject is pronominal and surges dramatically when the object is pronominal, being maximal with two pronominal arguments. Accordingly, the pronominalisation of the subject and especially the object are significant factors ($\chi 2>100$, $p<0.01$, $V=0.27$; $\chi 2>100$, $p<0.01$, $V=0.58$).

The fact that OV is more frequent with pronouns was already noted by Sirotinina

[3]The higher frequency of OVS in the News subcorpus (especially compared to Subtitles) may be explained by the observation presented in the previous section: when the subject is new, and the object is given, written Russian prefers reordering, while colloquial Russian prefers stress shift. OSV that is usually used with topical objects is actually more frequent in the Social Media and Subtitles subcorpora than in the News.

	News		Social Media		Subtitles	
SOV	2,106	0.6%	2,313	3.2%	271	1.4%
SVO	330,083		70,998		19,726	

Table 3: SVO and SOV orders with nominal arguments in different subcorpora.

	SOV		SVO		% SOV
syntactically complex object	2,732	58.3%	327,304	77.8%	0.8%
single noun object	1,958	41.7%	93,503	22.2%	2.1%

Table 4: Syntactic complexity of the object DP in SOV and SVO clauses.

(1965), although in her data, the effects were less pronounced. Pronouns have distinctive IS and phonological properties that make them avoid the sentence-final position, the default position of the main stress.[4] For Russian, this has been discussed in detail by Kholodilova (2013), who conducted a corpus study based on the Russian National Corpus (www.ruscorpora.ru). In some other languages, like French, Italian or Spanish, there is an absolute rule: nominal objects are always postverbal, unless they are topicalized, while direct object pronouns must precede the verb.

With the pronominalisation factor being so influential, we reanalyzed the differences among the three selected subcorpora. The share of sentences in which at least one of the arguments is pronominal differed significantly among them: 28.2% in the News, 60.9% in the Social Media and 91.1% in the Subtitles ($\chi 2>100$, $p<0.01$, $V>0.30$ for all pairwise comparisons). If only nominal arguments are taken into account (see Table 3), the differences between the three subcorpora become negligibly small ($\chi 2>100$, $p<0.01$, $V=0.02$ for the News vs. Subtitles comparison; $\chi 2>100$, $p<0.01$, $V=0.09$ for the News vs. Social Media comparison; $\chi 2>100$, $p<0.01$, $V=0.04$ for the Subtitles vs. Social Media comparison). This shows that our initial observations in Table 1 were primarily due to the different share of pronouns in these subcorpora, not to some other factors.

Syntactic complexity of the object. While light elements, such as pronouns, tend to avoid the sentence-final position, heavy elements have a greater chance of being found there. For example, since Ross (1967), the phenomenon of heavy NP shift has been studied in English. The role of heaviness for word order has been discussed in numerous studies (e.g. Arnold et al. 2000, Faghiri & Samvelian 2014, Wasow 1997). As is well known (e.g. Ariel 1990), heaviness also correlates with certain IS properties, namely, with low accessibility. Therefore, we decided to test whether the prevalence of SOV in our dataset was affected by the syntactic complexity of the object. Here and below, we analyzed only SVO and SOV clauses with nominal arguments.

We analyzed object DPs in which the head noun was modified by an adjective, a pronominal adjective (indefinite, demonstrative etc.), a dependent DP, a numeral, or a relative clause, as well as object DPs containing appositives and conjuncts. The SOV order is viewed as ungrammatical with relative clauses in Russian and indeed was never found with them in our dataset. All other types of syntactic complexity were attested with both SVO and SOV orders. Table 4 summarizes information about all these types, basically comparing objects consisting of a single noun and all other objects. In the first two columns, we see the distribution of these objects within SOV and SVO subsets, while the third column shows the share of SOV examples with simplex or complex objects in the dataset containing SVO and SOV sentences with non-pronominal arguments.

Table 4 shows that SOV is definitely not limited to syntactically simplex objects, and they do not even constitute a majority in SOV sentences. The share of SOV sentences

[4]Pronouns must have highly accessible referents, and giveness is associated with destressing. In addition to that, personal pronouns tend to cliticize to the preceding or the following word, while other pronouns are prosodically heavier.

	SOV		SVO		% SOV
objects with a dependent DP	735	15.7%	183,409	43.6%	0.4%
objects without a dependent DP	3,955	84.3%	237,398	56.4%	1.6%

Table 5: Objects with dependent DPs and without them in SVO and SOV orders.

	SOV		SVO		% SOV
objects with an adjectival modifier	1,437	30.6%	137,393	32.6%	1.0%
objects without an adjectival modifier	3,253	69.4%	283,414	67.4%	1.1%

Table 6: Objects with adjectival modifiers and without them in SVO and SOV orders.

with one-word objects is higher than the share of similar SVO sentences, but the effect is negligibly weak ($\chi 2>100$, $p<0.01$, $V=0.03$). When the object is not pronominal, whether it is a single noun or not, its chances to precede the verb are extremely low: 2.1% vs. 0.8%.

Tables 5, 6, and 7 illustrate that the probability of different modifiers in SOV and SVO are not equal (but, according to Cramer's V, the relevant differences are negligible: $\chi 2>100$, $p<0.01$, $V<0.01$ for all comparisons). For example, adjectival modifiers are more frequent in SOV than dependent DPs. The only type of modifiers whose share is higher in SOV than in SVO are pronouns, and we will analyze them in more detail below.

Information structure. Unfortunately, no Russian corpora are annotated for IS, and annotating a dataset as large as ours would require many days of meticulous work. Therefore, we came up with the following solution. In our dataset, we searched for clauses in which the object was modified by demonstrative, possessive or indefinite pronouns or was immediately preceded by a focus particle. Some of these words tend to be associated with given information; others with new or contrastively focused information. Analyzing whether and how the presence of these words changes the distribution of SVO and SOV clauses, we can draw certain conclusions about the IS-related connotations of the SOV order.

We already saw in Tables 4–7 that the probability of SOV is higher with pronominal modifiers than without them, while with non-pronominal modifiers it is invariably lower than without them. Pronouns tend to refer to highly accessible referents. On the contrary, low accessibility correlates with longer descriptions: non-pronominal DPs with various modifiers are often necessary to introduce a new entity or to reintroduce it after it has not been discussed for a while (e.g. Ariel 1990). So these results already tell us something about the IS properties of SOV orders as opposed to SVO.

Table 8 analyzes different pronominal modifiers one by one. It shows that the probability of SOV increases significantly with pronouns associated with the highest accessibility and proximity to the deictic center, such as *ėtot* 'this', *moj* 'my', *tvoj* 'yours', but not with pronouns like *tot* 'that', *ego* 'his', *ee* 'her'. The specific indefinite pronoun *kakoj-to* 'some' also triggers a significant increase in SOV probability, while non-specific indefinite pronouns *kakoj-nibud'* and *kakoj-libo* 'some, any' decrease it.

Table 9 shows that focus particles *tol'ko* 'only' and *daže* 'even' significantly increase the probability of SOV. The results for *imenno* 'exactly' do not reach significance. All these particles are associated with a set of alternatives, which may be scaled, i.e. signal

	SOV		SVO		% SOV
objects with a pronominal modifier	1,065	22.7%	18,095	4.3%	5.6%
objects without a pronominal modifier	3,625	77.3%	402,712	95.7%	0.9%

Table 7: Objects with pronominal modifiers and without them in SVO and SOV orders.

	N of objects with X	% SOV with X	% SVO with X	% SOV without X	significance
ėtot 'this'	6,162	9.8%	90.2%	1.0%	$\chi2>100$, $p<0.01$
moj 'my'	947	2.2%	97.8%	1.1%	$\chi2=10.8$, $p<0.01$
tvoj 'yours'	350	3.8%	96.2%	1.1%	$\chi2=21.9$, $p<0.01$
tot 'that'	1,154	1.5%	98.5%	1.1%	$\chi2=2.2$, $p=0.14$
ego/ee 'his/hers'	2,089	1.3%	98.7%	1.1%	$\chi2=0.7$, $p=0.40$
kakoj-to 'some'	563	2.8%	97.2%	1.1%	$\chi2=15.7$, $p<0.01$
kakoj-nibud' 'any'	202	1.0%	99.0%	1.1%	$\chi2<0.1$, $p=0.88$
kakoj-libo 'any'	169	0.6%	99.4%	1.1%	$\chi2=0.4$, $p=0.52$

Table 8: Different pronominal modifiers and the probability of SOV.

	N of objects with X	% SOV with X	% SVO with X	% SOV without X	significance
tol'ko 'only'	911	3.0%	97.0%	1.1%	$\chi2=29.0$, $p<0.01$
daže 'even'	165	26.7%	73.3%	1.1%	$\chi2>100$, $p<0.01$
imenno 'exactly'	60	1.6%	98.4%	1.1%	$\chi2=0.2$, $p=0.68$

Table 9: Different focus particles and the probability of SOV.

contrastive or emphatic focus.

Table 10 presents the results for different quantifiers. *Každyj* 'every, each' and *nikakoj* 'no' significantly increase the probability of SOV. The results for *vse* 'all' (marginally significant) and *mnogo* 'many' show the same tendency. *Malo* 'few, little', *neskol'ko* 'several, some' and *nekotoryj* 'some' demonstrate the opposite pattern, but only the results for *nekotoryj* are marginally significant.

In general, we can conclude that modifiers associated with giveness and specificity on the one hand and with contrastive or emphatic focus on the other hand increase the probability of SOV. At the same time, one should keep in mind that this probability nevertheless remains very low. In the case of given objects, the increase is most significant with *ėtot* 'this' (see Table 8), but even in this case, it is only 9.8%. In the case of focused objects, *daže* 'even' boosts the share of SOV sentences to a dramatic 26.7% (see Table 9), but with all other modifiers, the increase is less than 5%.

Let us briefly summarize the results of our corpus study. It showed that many factors identified in the previous studies indeed influence the choice of SOV order, but, with only one exception, their role is very small. Only pronominal objects precede the verb regularly, while non-pronominal objects of any kind in different types of texts are postverbal in the absolute majority of cases. Thus, Slioussar's (2007) hypothesis that spoken Russian is undergoing a diachronic transition to an OV language is definitely wrong.

	N of objects with X	% SOV with X	% SVO with X	% SOV without X	significance
každyj 'every'	279	3.5%	96.5%	1.1%	$\chi2=15.8$, $p<0.01$
nikakoj 'no'	1338	4.7%	95.3%	1.1%	$\chi2>100$, $p<0.01$
vse 'all'	148	2.8%	97.2%	1.1%	$\chi2=3.50$, $p=0.06$
mnogo 'many'	1410	1.5%	98.5%	1.1%	$\chi2=1.94$, $p=0.16$
malo 'few'	337	0.6%	99.4%	1.1%	$\chi2=0.80$, $p=0.37$
neskol'ko 'several'	1571	0.7%	99.3%	1.1%	$\chi2=2.38$, $p=0.13$
nekotoryj 'some'	1027	0.1%	99.4%	1.1%	$\chi2=3.58$, $p=0.06$

Table 10: Different quantificational modifiers and the probability of SOV.

Most functional and formal studies, starting with Sirotinina (1965), have stressed that preverbal objects may be given or new. Unfortunately, our study does not allow estimating the share of these two groups in SOV sentences, but shows that they are both well represented and provides some interesting observations on their properties. Let us come back to the most debated question that we discussed in §2: whether SOV sentences with new objects are necessarily expressive, emphatic or contrastive. Our study shows that being contrastive or emphatic does indeed increase the chances that the object will end up before the verb, but it cannot be used to decide whether all sentences with new preverbal objects necessarily have these properties. However, we believe that this question may be solved based on other sources: various examples analyzed by Kodzasov (1989, 1996), Yanko (2001) and other authors, as well as in §2 of the present paper, show that this is not the case.

4 SYNTACTIC ANALYSIS

In this section, we try to outline a syntactic analysis compatible with the observed picture. It must capture several crucial observations listed in the last two paragraphs of §3, most importantly, that both given and new objects can move and the optionality of their movement. Let us start with the distinction between given and new objects.

Neeleman & van de Koot (2015) demonstrate that cross-linguistically, given constituents undergo A-scrambling, whereas focus movement is an A′-operation. Titov (2020) adopts this idea in discussing IS-related movement in Russian, but does not provide any syntactic details. However, the distinctions described by Neeleman & van de Koot (2015) cannot be found in Russian SOV sentences. The main difference between given and new objects in these sentences is prosodic: the former are destressed, while the latter bear the sentential stress.

Can we conclude based on this difference alone that they occupy different syntactic positions? No existing studies provide a comprehensive discussion of this question.[5] Therefore, the analysis we develop below assumes that given and new objects can target the same positions, but further research is necessary to address this problem. To estimate where these positions can be in the syntactic tree, let us consider several constructed examples with adverbs: (6-b)–(6-d) with a given object and (7-b)–(7-d) with a contrastively focused object. In their analysis, we rely on our native speaker judgments as well as on the informally collected judgments of five other Russian speakers.

(6) a. Ponedel'nik nacinalsja s matematiki.
Monday started with math
'Monday started with math.'

b. Petja ėtot urok často progulival.
Petja.NOM this.ACC lesson.ACC often missed
'Petja often missed this lesson.'

c. Petja často ėtot urok progulival.
Petja.NOM often this.ACC lesson.ACC missed
'Petja often missed this lesson.'

d. Petja často progulival ėtot urok.
Petja.NOM often missed this.ACC lesson.ACC
'Petja often missed this lesson.'

(7) a. Detjam bylo trudno učit'sja.
kids been difficult study
'It was difficult for the kids to study.'

[5] One thing we definitely know is that there is no correlation between syntactic positions and the availability of stress shift in Russian. A non-final main stress can fall on constituents occupying an A-position: for example, on the focused subject in an SVO sentence (as we noted above, when the subject is in narrow focus, Russian allows using OVS order or shifting the main stress). At the same time, a topicalized object undergoes A′-movement, but does not retain the main stress.

b. Petja tol'ko matematiku xorošo znal.
Petja.NOM only math.ACC well knew
'Peter only knew math well.'
c. Petja xorošo tol'ko matematiku znal.
Petja.NOM well only math.ACC knew
'Peter only knew math well.'
d. Petja xorošo znal tol'ko matematiku.
Petja.NOM well knew only math.ACC
'Peter only knew math well.'

SVO, as in (6-d) and (7-d), is the preferred word order in these cases, but SOV is also possible.[6] In (6-c) and (7-c), the object is on the edge of the *v*P, and this is what most researchers have in mind discussing SOV sentences in Russian. In (6-b) and (7-b) it is in a higher position, but still below the subject, which can be reasonably assumed to occupy the specifier of TP in such sentences. The nature of this position depends on the view of adverbs that we adopt. For example, if we assume that adverbs adjoin to the *v*P, the objects can target this larger constituent. The interpretational differences between the (6-c)/(7-b) and (6-c)/(7-c) sentences are rather subtle. In (6-b), the object will be interpreted as topical (together with the subject). (7-b) presupposes a more contrastive interpretation than (7-c).

The situation when the chances of being preverbal are higher for contrastive foci and at the same time for highly accessible constituents may seem paradoxical at first. But this is exactly the case in languages like Hungarian, Basque or Malay (Szendrői 2001, 2005, Ortiz de Urbina 1999, Jayaseelan 2001). However, in these languages the relevant movements are obligatory, while in Russian, they are optional.

Syntactic theory has a long tradition of dealing with optional movement, and many authors agree that it cannot be triggered by traditional syntactic features, opting for 'free movement' instead, like Titov (2020) in her discussion of moved focused constituents in Russian. However, many technical questions are usually left open: which syntactic positions are available for such movement, how exactly it takes place, etc. We suggest that using Chomsky's (2008) edge features or a similar mechanism may solve these problems. Edge features of phase heads can attract any constituent; this movement is essentially free, i.e. it is optional and does not have any prerequisites.[7] The only requirement is that the moved element receives a new interpretation based on the final position it reaches.

We hypothesize that moving the object to the edge of the *v*P marks that the object and the verb are not homogeneous with respect to IS. Either the object is given or specific while the verb is new, or the object is the most salient part of the new information. The latter is definitely true for contrastive and emphatic foci, as well as for the cases when the object is in narrow focus. However, we have argued that some moved new objects are not associated with such interpretations, as in the examples (4) and (5) ((5) is repeated below as (8)). We hypothesize that in such cases, the verb is interpreted as a less salient, or backgrounded, part of new information. As soon as the addressee hears the word *dači* 'summer houses', they can already guess the general meaning of the predicate (although not necessarily the exact word: in the next sentence of this text, the speaker uses the word *imeli* 'owned' to note that some people had several summer houses for rent).

(8) Torgovcy, vladel'cy lar'kov... Potom... DAČI sdavali.
merchants owners kiosks.GEN then summer-houses.ACC rented-out
'(They were) merchants, owners of kiosks... Then... (They) rented out summer houses.'

[6]Other orders may be used as well: for example, we can place the subject or the adverb in narrow focus at the end of the clause, or put the object that is a topic switch or a contrastive topic at the beginning.

[7]The details of the analysis depend on the position of the objects in (6-b) and (7-b): if they are in the specifiers of some aspectual projections, we will have to assume that not only phase heads possess edge features.

Notably, in all these cases movement does not create new interpretations. All interpretations discussed above are available without movement, but movement makes at least one of them obligatory. That is, we see a ¾ signature effect which was originally described for scopal phenomena (Bobaljik & Wurmbrand 2012) and subsequently used in studies of IS (e.g. Titov 2020): the neutral word order is compatible with different interpretations, but movement serves to exclude some options. From the edge of the *v*P the object can be moved further with more subtle interpretational effects.

5 CONCLUSIONS

We presented a corpus study of the Russian word order focusing on SOV sentences. All previous corpus studies analyzed only whether the object precedes or follows the verb, so our study is the first to tease apart SOV, OSV and OVS orders on the one hand and SVO, VSO and VOS on the other, which is important because they are all known to have distinct syntactic, prosodic and IS properties. Various suggestions have been made in the literature concerning the factors that may influence the choice of SOV, and our primary goal was to test their relative importance. We used the "Taiga" corpus (Shavrina & Shapovalova 2017), which contains several large, stylistically diverse subcorpora.

We conclude that only one factor, pronominalization, substantially influences the frequency of SOV. If we consider all the SVO and SOV sentences in our dataset, SOV order is found in 48% of those in which the object is a pronoun and only in 2% of those in which it is not (summarizing the data presented in Table 2). As for all the other factors that have been discussed in previous studies, both on Russian and on other languages, there is good news and bad news. Their influence is statistically significant, but extremely small (thanks to a very large dataset, we could detect significance even in case of subtle differences).

In other words, all non-pronominal objects, given or new, specific or not, in a narrow, emphatic or contrastive focus or not, with or without various modifiers, follow the verb in the absolute majority of cases. This is true for written and spoken, predominantly narrative and predominantly dialogic, more and less formal texts. As far as we can judge, this is an unexpected finding that could not be predicted by any previous study of Russian word order.

In reviewing the previous studies, we focused on the question of whether all SOV sentences with new objects have contrastive or emphatic foci or are stylistically marked ('expressive' or 'emotive'), and came to a negative answer. At the same time, our corpus study shows that being in a contrastive or emphatic focus increases the chances of the object preceding the verb. Being highly accessible has the same effect. We drafted a syntactic analysis relying on edge features to capture the observed picture.

ABBREVIATIONS

ACC	accusative	OSV	Object–Subject–Verb
GEN	genitive	OVS	Object–Verb–Subject
IS	Information Structure	SOV	Subject–Object–Verb
NOM	nominative	SVO	Subject–Verb–Object

ACKNOWLEGMENTS

The work on this project was carried out in the framework of the Basic Research Program at the National Research University Higher School of Economics, Russia.

CONTACT

Natalia Slioussar <slioussar@gmail.com>

REFERENCES

Ariel, Mira. 1990. *Accessing noun-phrase antecedents.* London: Routledge.

Arnold, Jennifer E., Thomas Wasow, Anthony Losongco & Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of complexity and information structure on constituent ordering. *Language* 76(1). 28–55.

Belletti, Ariadna. 2002. Aspects of the low IP area. In Luigi Rizzi (ed.), *The structure of IP and CP*, 16–51. Oxford: Oxford University Press.

Benveniste, Émile. 1974. *Problèmes de linguistique générale [Problems of general linguistics]. Vol. 2.* Paris: Gallimard.

Bobaljik, Jonathan D. & Susi Wurmbrand. 2012. Word order and scope: transparent interfaces and the ¾ signature. *Linguistic inquiry* 43(3). 371–421.

Bonnot, Christine I. & Irina Fougeron. 1982. L'accent de phrase initial en russe est-il toujours un signe d'expressivité ou de familiarité? [Is a phrase-initial accent always a marker of expressiveness or familiarity in Russian?]. *Bulletin de la Société de Linguistique de Paris* 87. 309–330.

Chomsky, Noam. 2008. On phases. In Robert Freidin, Carlos P. Otero & Maria-Luisa Zubizarreta (eds.), *Foundational issues in linguistic theory*, 133–166. Cambridge, MA: MIT Press.

Dyakonova, Marina. 2006. A unified analysis of ex situ and in situ focus in Russian. Paper presented at the 1st Meeting of Slavic Linguistic Society. September 8–10, 2006, Bloomington, IN.

Faghiri, Pegah & Pollet Samvelian. 2014. Constituent ordering in Persian and the weight factor. *Empirical issues in syntax and semantics* 10. 215–232.

Jayaseelan, Karattuparambil A. 2001. IP-internal topic and focus phrases. *Studia linguistica* 55(1). 39–75.

Kholodilova, Maria A. 2013. Pozicionnye svojstva mestoimenij v russkom jazyke [Positional properties of pronouns in Russian]. MA thesis: Saint Petersburg State University.

King, Tracy Holloway. 1995. *Configuring topic and focus in Russian.* Stanford, CA: CSLI Publications.

Kodzasov, Sandro V. 1989. Ob akcentnoj strukture sostavljajuščix [On the prosodic structure of constituents]. *Experimental phonetics speech analysis* 2. 122–127.

Kodzasov, Sandro V. 1996. Kombinatornaja model' frazovoj prosodii [A combinatorial model of phrasal prosody]. In *Prosodičeskiy stroj russkoj reči [Prosodic system of the Russian langauge]*, 85–123. Moscow: Institute of Russian Language RAS.

Kovtunova, Irina I. 1976. *Sovremennyj russkij jazyk. Porjadok slov i aktual'noe členenie predloženija [Modern Russian language. Word order and information structure].* Moscow: Prosveščenie.

Mangiafico, Salvatore S. 2016. Summary and analysis of extension program evaluation in R, version 1.18.1. Available at: http://rcompanion.org/handbook/.

Mykhaylyk, Roksolana. 2010. *Optional object scrambling in child and adult Ukrainian*. Stony Brook, NY: Stony Brook University dissertation.

Mykhaylyk, Roksolana. 2011. Middle object scrambling. *Journal of Slavic linguistics* 19(2). 231–272.

Neeleman, Ad & Hans van de Koot. 2015. Word order and information structure. In Caroline Féry & Shinichiro Ishihara (eds.), *The Oxford handbook of information structure*, 383–401. Oxford: Oxford University Press.

Ross, John R. 1967. *Constraints on variables in syntax*. Cambridge MA: MIT dissertation.

Shavrina, Tatiana & Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning: «Taiga» syntax tree corpus and parser. In *CORPORA 2017 Proceedings*, 78–84. Saint Petersburg: Saint Petersburg State University Press.

Sirotinina, Olga B. 1965. *Porjadok slov v russkom jazyke [Word order in Russian]*. Saratov: Saratov State University Press.

Slioussar, Natalia. 2007. *Grammar and information structure. A study with reference to Russian*. Utrecht: LOT. PhD thesis.

Szendrői, Kriszta. 2001. *Focus and the syntax-phonology interface*. London: University College London dissertation.

Szendrői, Kriszta. 2005. Focus movement (with special reference to Hungarian. In Martin Everaert & Henk van Riemsdijk (eds.), *The Blackwell companion to syntax, Vol. 2*, 272–337. Oxford: Blackwell.

Titov, Elena. 2020. Optionality of movement. *Syntax* 23(2). 347–374.

Ortiz de Urbina, Jon. 1999. Focus in Basque. In Georges Rebuschi & Laurice Tuller (eds.), *The grammar of focus*, 311–333. Amsterdam: John Benjamins.

Wasow, Thomas. 1997. Remarks on grammatical weight. *Language variation and change* 9(1). 81–105.

Yanko, Tatiana E. 2001. *Kommunikativnye strategii russkoi reči [Communicative strategies of the Russian language]*. Moscow: Jazyki Slavjanskoj Kul'tury.

Yokoyama, Olga. 1986. *Discourse and word order*. Amsterdam: John Benjamins.

Zemskaya, Elena A. (ed.). 1973. *Russkaja razgovornaja reč' ['Russian colloquial speech]*. Moscow: Nauka.

Zemskaya, Elena A. (ed.). 1978. *Russkaja razgovornaja reč'. Teksty ['Russian colloquial speech. Texts']*. Moscow: Nauka.

Zubizarreta, Maria-Luisa. 1998. *Prosody, focus, and word order*. Cambridge, MA: MIT Press.