

SOV in Russian

Ilya Makarchuk <ilya.makarchuk@gmail.com>

Higher School of Economics

16th Conference on Typology and Grammar for Young Scholars

Subject: SVO-to-SOV word alternation in Russian¹

(cf. e.g. Baylin 2012 for an overview of Russian word order alternations)

(1) **SVO:** Petja našel kluč. ‘Peter found a/the key’ (basic word order)

(2) **SOV:** Petja kluč našel. ‘Peter found a/the key’

Aim: To check explanations proposed literature on new corpus data.

Feedback, suggestions and extensive discussion is hugely encouraged.

Background

Several explanations of Russian SOV has been proposed:

- SOV is a stylistic marker of a more colloquial speech (Sirotnina 1965/2003; Zemskaya 1973)

Table 1. Position of object and style (adopted from (Sirotnina 1965/2003))

Word order	All parts of speech				Only nouns			
	Least formal		Most formal		Least formal		Most formal	
VO	38	33.6%	67	63.2%	32	45.7%	63	73.3%
OV	75	66.4%	39	36.8%	38	54.3%	23	26.7%

- Dialogue vs narrative mode of discourse (Kodzasov 1996). Sentences with initial sentential stress are characteristic of a dialogue, but not of a narrative, where the rheme is usually placed at the end of the sentence (Bonnot and Fougeron 1982). SOV, Kodzasov argues, is neutral in a dialogue.
- Russian is in the process of changing word order (Slioussar 2009). SOV seems to have no implications on information structure and also Russian develops a “tail” that is characteristic of SOV languages.
- According to (Mykhaylyk 2011), in Ukrainian, SOV is only compatible with a specific object, which might also be the case in Russian.

(3) Taras čytaje knyžku
Taras reads book_ACC
‘Taras reads a/some/a certain/the book.’

(4) Taras knyžku čytaje
Taras book_ACC reads
‘Taras reads #some/a certain/the book.’

Data and methods

Data source: three subcorpora of Taiga, a corpus with automatic syntactic annotation: news, social media, subtitles of TV shows. Only clauses with only two arguments — subject and object — were sampled.

Table 2. Sizes of subcorpora

Subcorpus	Tokens	Sample	SVOs & SOVs	SOVs
News	92 mln	540531	462836	14716
Social	80 mln	234535	187465	33770
Subtitles	101 mln	275834	224819	50349
Total	273 mln	1050900	875120	98835

Table 3. A sample without pronominal arguments

	News		Social		Subtitles		Total	
SOV	2106	0.63%	2313	3.16%	271	1.36%	4690	1.10%
SVO	330083	99.37%	70998	96.84%	19726	98.64%	420807	98.90%

Methods:

- Cramer’s V for grammatical features: 0.1 – 0.3 for small, 0.3 – 0.5 for medium, ≥ 0.5 for large effect with $k = 2$ (Mangiafico 2016)
- χ^2 -test for words: the threshold is $p < 0.05$

Style and part of speech

SOV is 2nd most frequent word order in Russian.

Table 4. Frequencies of word orders

Word order	SVO	SOV	OSV	OVS	VOS	VSO
Percentage	73.87%	9.40%	7.64%	5.72%	1.90%	1.47%

But there is a correlation between word order and genre of the text.

Table 5. Frequencies of word orders in different genres

Word order	News	Social	Subtitles
SOV	3.18%	18.01%	22.40%
SVO	96.82%	81.99%	77.60%

Cramer’s V = 0.276 (medium effect)

¹ Russian < East Slavic < Slavic < Balto-Slavic < Indo-European < Nostratic

But also word order correlates with part of speech of the arguments.

Table 6. Word order and part of speech of the object

	Noun	Wh- and C- words	Personal pronouns	Other pronouns
SOV	16.02%	2.74%	42.22%	39.02%
SVO	88.36%	0.78%	6.10%	4.76%

Cramer's V = 0.578 (large effect)

Table 7. Word order and part of speech of the subject

	Noun	Wh- and C- words	Personal pronouns	Other pronouns
SOV	16.88%	8.92%	69.86%	4.34%
SVO	57.21%	8.29%	31.94%	2.55%

Cramer's V = 0.268 (medium effect)

In fact, effect of style could be explained by frequency of (personal) pronouns. Genres with more dialogues have more 1- and 2-person pronouns: news are mostly narratives, subtitles are mostly dialogues and social media are in the middle. A dialogue is an exchange between two people, hence, a higher frequency of references to the speaker or the hearer i. e. personal pronouns. Since Russian pronouns tend to be proclitic, they prefer preverbal position (Kholodilova 2013). Thus, the more dialogical the text is, the bigger the percentage of SOV clauses.

Table 8. First and second person pronouns and genre

	News	Social	Subtitles
1- or 2-person pronoun as argument	7.26%	29.47%	67.99%
No 1- or 2-person pronouns as arguments	92.74%	70.53%	32.01%

Cramer's V = 0.565 (large effect)

That is why for other features we looked at a subsample of clauses with only nominal arguments to eliminate effects of phonology (see Table 2). On this subsample effects of genre are insignificant (see Table 9).

Table 9. Word order and style in a subsample with only nominal arguments

	News	Social	Subtitles
SOV	0.63%	3.16%	1.36%
SVO	99.37%	96.84%	98.64%

Cramer's V = 0.090 (no effect)

Morphosyntactic properties

Morphosyntactic features of the verb — polarity, tense, aspect, finiteness and mood — and the object — number, animacy, type of dependent — show no significant effect (Cramer's $V \leq 0.05$, details omitted for sake of space).

Determiners and quantifiers

Objects with determiners which denote “closeness” to a deictic centre and givenness like *ètot* ‘this’, *moj* ‘my’, *kakoj-to* ‘a, some kind’ prefer SOV. On the contrary, those which mark newness and objects away from the deictic centre do not prompt SOV (Tables 10 and 11).

As for quantifiers, only universal quantifiers show a preference to SOV while other quantifier words do not. It is because universal quantifiers have NP in presupposition, the NP should be somehow given, while it is not the case for non-universal quantifiers.

Table 10. Definite determiners of the object

Determiner	Det. wrt. SOV/SVO	SVO wrt. Det.	p-value	Prefers
<i>ètot</i> ‘this’	12.94%	9.86%	< 0.001	SOV
	1.32%	0.97%		
<i>takoj</i> ‘such’	2.62%	4.82%	< 0.001	SOV
	0.58%	1.08%		
<i>tot</i> ‘that’	0.38%	1.58%	0.117	None
	0.27%	1.10%		

Table 11. Definite determiners of the object

Determiner	Det. wrt. SOV/SVO	SVO wrt. Det.	p-value	Prefers
<i>moj</i> ‘my’	0.45%	2.23%	0.003	SOV
	0.22%	1.10%		
<i>tvoj</i> ‘your’	0.28%	3.77%	0.001	SOV
	0.08%	1.10%		
<i>ego/eë</i> ‘his/her’	0.58%	1.30%	0.397	None
	0.49%	1.10%		

Table 12. Quantifiers of the object

Determiner	Det. wrt. SOV/SVO	SVO wrt. Det.	p-value	Prefers
<i>ves</i> ‘whole’	4.46%	3.88%	< 0.001	SOV
	1.23%	1.07%		
<i>každyj</i> ‘every’	0.21%	3.53%	0.001	SOV
	0.06%	1.10%		

<i>nikakoj</i> ‘no’	1.34%	out of SOV	4.70%	with D	< 0.001	SOV
	0.30%	out of SVO	1.09%	w/o D		
<i>mnogo</i>	0.45%	out of SOV	1.50%	with D	0.156	None
‘a lot of’	0.33%	out of SVO	1.10%	w/o D		
<i>neskol’ko</i>	0.23%	out of SOV	0.69%	with D	0.119	None
‘several’	0.38%	out of SVO	1.10%	w/o D		
<i>malo</i>	0.00%	out of SOV	0.00%	with D	0.059	None
‘some’	0.08%	out of SVO	1.10%	w/o D		

Discourse words

A topic particle *-to* is always used with SOV. Focus particles seem to also prompt SOV. As for *liš*, in the modern language it is usually a part of a fixed expression and not a focus particle. *imenno* may not prefer SOV for reasons of processing.

Table 13. Topic- and focus-sensitive particles

Particle	P. wrt. SOV/SVO	SVO wrt. P.	p-value	Prefers		
<i>tol’ko</i> ‘only’	0.58%	out of SOV	2.99%	with P	< 0.001	SOV
	0.21%	out of SVO	1.10%	w/o P		
<i>-to</i>	0.51%	out of SOV	77.42%	with P	< 0.001	SOV
‘concerning..’	<0.01%	out of SVO	1.10%	w/o P		
<i>daže</i> ‘even’	0.94%	out of SOV	26.83%	with P	< 0.001	SOV
	0.03%	out of SVO	1.09%	w/o P		
<i>i</i> ‘also’	1.19%	out of SOV	3.78%	with P	< 0.001	SOV
	0.34%	out of SVO	1.09%	w/o P		
<i>vsě-taki</i>	0.15%	out of SOV	3.06%	with P	0.014	SOV
‘indeed’	0.05%	out of SVO	1.10%	w/o P		
<i>liš</i> ‘only’	<0.01%	out of SOV	<0.01%	with P	0.416	None
	0.03%	out of SVO	1.10%	w/o P		
<i>imenno</i>	0.02%	out of SOV	1.59%	with P	0.503	None
‘it is ... that’	0.01%	out of SVO	1.10%	w/o P		

As to other discourse words your ideas are welcome :)

Table 14. Other discourse words within a sentence and object position

Particle	w. wrt. SOV/SVO	SVO wrt. w.	p-value	Prefers		
<i>voobščē</i>	0.62%	of SOV	5.87%	with w.	< 0.001	SOV
‘in general’	0.11%	of SVO	1.10%	w/o w.		

<i>tut</i> ‘here’	0.49%	of SOV	4.13%	with w.	< 0.001	SOV
	0.13%	of SVO	1.10%	w/o w.		
<i>tam</i> ‘there’	0.83%	of SOV	3.12%	with w.	< 0.001	SOV
	0.29%	of SVO	1.10%	w/o w.		
<i>tak</i>	2.11%	of SOV	2.57%	with w.	< 0.001	SOV
‘this way’	0.89%	of SVO	1.09%	w/o w.		
<i>by</i>	1.43%	of SOV	4.06%	with w.	< 0.001	SOV
subjunctive	0.38%	of SVO	1.09%	w/o w.		
<i>ved’</i>	0.15%	of SOV	2.04%	with w.	0.110	None
‘after all’	0.08%	of SVO	1.10%	w/o w.		
<i>prostō</i>	0.09%	of SOV	0.55%	with w.	0.210	None
‘simply’	0.17%	of SVO	1.10%	w/o w.		
<i>uže</i>	0.96%	of SOV	1.18%	with w.	0.585	None
‘already’	0.89%	of SVO	1.10%	w/o w.		
<i>tože</i>	0.17%	of SOV	1.04%	with w.	1	None
‘as well’	0.18%	of SVO	1.10%	w/o w.		
<i>takže</i>	0.13%	of SOV	0.10%	with w.	< 0.001	SVO
‘in addition’	1.38%	of SVO	1.12%	w/o w.		

Conclusion

- The main driver of the SVO-to-SOV alternation is phonetic properties of pronouns. Genre effects found by previous researchers are only epiphenomena of this process.
- A given object prefers SOV. But it is not a hard rule, but rather a tendency.
- Given two previous points and the fact that SVO is at least 7 times more frequent than SOV, it seems at least too early to think about any diachronic change.

References

- Bonnot, Christine and Irina Fougeron (1982). “L’accent de phrase initial en russe est-il toujours un signe d’expressivité ou de familiarité?”. In: *Bulletin de la Société de Linguistique de Paris* 77 (1), pp. 309–330.
- Bailyn, John F. (2012) *The Syntax of Russian*. New York: Cambridge University Press

- Kholodilova, Maria (2013). “Pozitsionnye svoystva mestoimeniy v russkom yazyke”. Master thesis. Saint-Petersburg: SPbSU.
- Kodzasov, Sandro (1996). “Zakony frazovoy aktsentuatsii”. In: *Prosodicheskiy stroy russkoy rechi*. Ed. by Tatyana Nikolaeva. Moscow: Russian Language Institute of Russian Academy of Science, pp. 181–204.
- Mangiafico, Salvatore S. (2016). *Summary and Analysis of Extension Program Evaluation in R*, version 1.18.1. URL: <http://rcompanion.org/handbook/>
- Mykhaylyk, Roksolana (2011). “Middle Object Scrambling”. In: *Journal of Slavic Linguistics* 19.2, pp. 231–72.
- Sirotinina, Olga (1965/2003). *Poryadok slov v russkom yazyke*. Moscow: URSS.
- Slioussar, Natalia (2009). *Na styke teoriy: Grammatika i informatsionnaya strukturav russkom i drugikh yazykakh*. Moscow: Knizhnyy dom “Librokom”.
- Zenskaya, Elena, ed. (1973). *Russkaya razgovornaya rech*. Moscow: Nauka.

Sources

- Shavrina, Tatiana and Yana Kurmachova (n.d.). *Taiga Corpus. An open-source corpus for machine learning*. URL: https://tatianashavrina.github.io/taiga_site/