Eye of newt and toe of frog: Russian word order in between factors

Background. Russian is widely recognized as a language with significant word order flexibility, where the choice of word order is influenced by factors such as information structure (given > new), animacy (human > non-human > inanimate), definiteness (definite > indefinite), and phrase weight (light > heavy) (Titov, 2012; Slioussar and Makarchuk, 2022, among others). Corpus-based research on factors governing linear word order has typically focused on a specific word order or a subset of orders – OV/VO (Seržant et al., 2026; Sirotinina, 1965), SOV (Slioussar and Makarchuk, 2022). While valuable, these studies do not address what specific factors govern the choice between, e.g., OSV and SOV, or VOS and OVS – and to what extent each factor influences each word order. We aim to close this gap by constructing a quantitative profile of factors for each word order in Russian. We focus on the SPO (P for predicate) kernel – all six permutations of subject, verb, and object. The factors under investigation include animacy, phrase weight, information-structural status, definiteness, negation, clause type, and pronominality.

Method. We took the latest version of SynTagRus (Droganova et al., 2018) – a syntactically annotated (UD conventions) Russian corpus – and enriched it with additional annotation layers. Sentences were separated into clauses (root, coordinate, and subordinate), and for each clause we automatically annotated subjects, direct objects, predicates (verb, auxiliary, modal, or a combination thereof), indirect objects, prepositional phrases, adverbial phrases, etc. For each nominal phrase, we annotated the animacy status of its head. We also annotated whether the head of a nominal phrase has dependents (here and hereafter in UD terminology) with properties correlated with definiteness or indefiniteness. If a head had a direct dependent whose lemma belonged to a set correlated with definiteness in the literature – e.g., *ètot* this', *ego* his', *samyj* the most' – the phrase was tagged as DEFINITE. If the dependent's lemma belonged to a set correlated with indefiniteness – e.g., *nekij* a certain', *kakoj-nibud'* 'some' – the phrase was tagged as INDEFINITE. Nominal phrases without such dependents were left unmarked. Thus, we had 3 levels (definite, indefinite, and unknown) of this factor. Information-structural status was annotated via proxy: phrases whose heads had a direct dependent with a lemma belonging to a set of focus-sensitive items – e.g., *daže* even', *i* PRT', *tol'ko* only' – were tagged as HIGH_PROMINENCE. The unmarked elements were annotated as 'unknown'. Pronominality was annotated by marking whether the head of a phrase is a personal pronoun and, if so, of which person (1st, 2nd, or 3rd). Negation was annotated by checking whether the predicate had a direct dependent *ne* not', in which case the clause was tagged as NEGATED. Finally, phrase weight was operationalized by counting the number of syllables in each phrase, computing the mean syllable length across all phrases within the clause, and converting each phrase's length to a z-score relative to that mean – higher values indicating heavier phrases, lower values indicating lighter ones. We excluded sentences with coordinated core arguments (e.g., S *and* S, O *and* O), though other non-core elements (PPs, AdvPs) may be present in the clause but fall outside the scope of this study. Word order was determined by reducing each clause to its core arguments and predicate – e.g., PP S AdvP Predicate AdvP O PP" → X S X P X O X → SPO. We have also excluded questions and exclamations, as well as subordinate clauses where one of the core arguments (S or DO) is fixed.

Results. The general distribution of word orders in the corpus is presented in Table 1. The weight and direction of each factor (increasing or decreasing the probability of a given word order) are shown in Figures 1 and 2. First, pronominality emerges as one of the strongest factors across word orders: pronominal arguments avoid clause-final position. Accordingly, in word orders where the last element is an argument, the pronominality of that argument is among the main predictors. However, differences emerge in word orders where neither argument is final (SOP vs. OSP). While for SOP the pronominality of both arguments is one of the primary factors, for OSP only the status of the subject matters most. Second, information-structural factors are of near-universal relevance across word orders. Interestingly, however, the focal status of a non-final argument (and not predicate!) does not appear to decrease the probability of any word order. Moreover, focal status of the subject appears to be an increasing factor even for SPO. Another influential factor was negation – it significantly raisesthe probabilitiesitieoccurrencerrence of

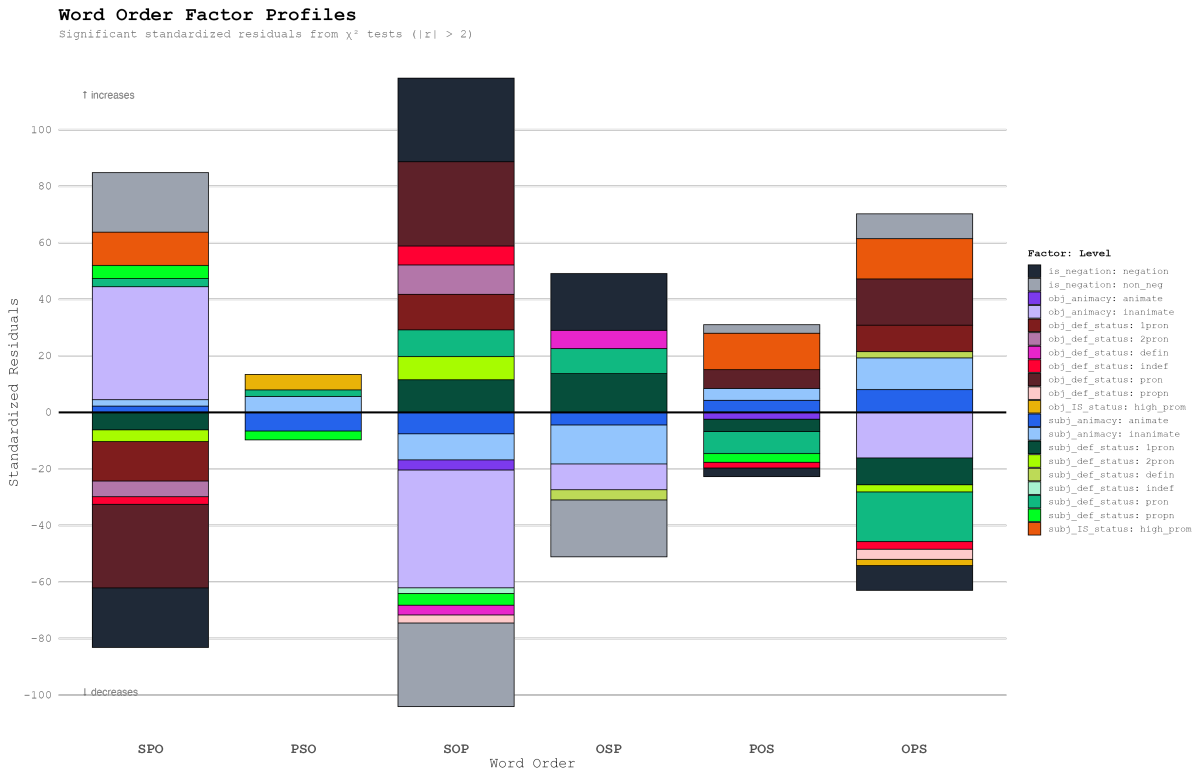

Figure 1: Word order profiles based on the standardized residuals for each level of a factor from X^2 tests

SOP and OSP. We hypothesize that such influence is at least partially connected with the increased level of prominence (focus) of such predicate. Third, while the primary driving factors are pronominality and focus, animacy hierarchies and differences in assumed definiteness status appear to play a role within these broader patterns. The first nominal element in a clause shows no strong preference for animacy or definiteness status, but the second nominal element tends to be equal to or lower than the first on the animacy and definiteness scales (see, for instance, "obj:animacy:animate" for SPO, or compare the increasing and decreasing factors for POS: while both animate and inanimate status of the subject are increasing factors, animate status of the object is a decreasing factor). Fourth, phrase weight proves to be a reliable predictor across all word orders with sufficient data: the clause-final element tends to be heavier than or equal to both preceding elements, regardless of grammatical function (argument or predicate). Finally, we fitted two binary logistic regression models for each word order (e.g., SPO = 1 vs. not SPO = 0): one with pronominality, phrase weight, negation, and animacy as fixed effects, and another with definiteness and phrase weight as fixed effects – in order to confirm that the effects of pronominality and the positional differences between definite and indefinite arguments are not epiphenomenal of phrase weight (e.g., Russian *kakoj-nibud' mal'čik* some boy' is heavier than *ètot mal'čik* this boy'). The effects of all factors, including pronominality, were confirmed after controlling for the remaining predictors.

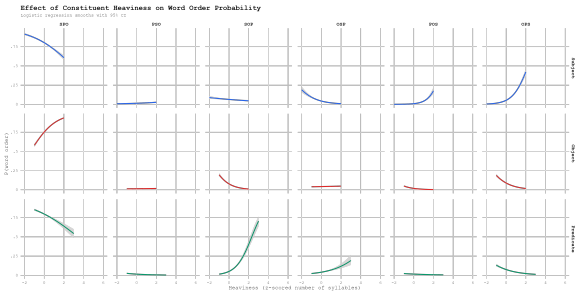

Figure 2: The Effect of Heaviness on WO Probs

Word order	%
SPO	79.09
OPS	7.34
SOP	6.29
OSP	4.31
POS	1.52
PSO	1.45

Table 1: % of WO in kernel

References

Droganova, K., Lyashevskaya, O., and Zeman, D. (2018). Data conversion and consistency of monolingual corpora: Russian ud treebanks. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, number 155, pages 52–65, Oslo, Norway. Linköping University Electronic Press.

Seržant, I. A., Alfimova, D., Biskup, P., and Seržants, I. (2026). Efficient sentence processing significantly affects the position of objects in russian. *Linguistics*, 64(1):189–226.

Sirotinina, O. B. (1965). *Poryadok slov v russkom yazyke [Word Order in Russian]*. Saratov State University, Saratov, 1st ed. edition.

Slioussar, N. and Makarchuk, I. (2022). SOV in Russian: A Corpus Study. *Journal of Slavic Linguistics*, 30(3):1–14.

Titov, E. (2012). *Information Structure of Argument Order Alternations*. PhD thesis, University College London, London.