Empirical Analysis of Russian Aspectual Prefixes: A Computational Approach to Productivity & Semantic Relatedness

Natalia Tyulina
CUNY Graduate Center
ntyulina@gradcenter.cuny.edu

Abstract

This work presents a computational analysis of the productivity of Russian aspectual prefixes. Using multiple complementary methods, including the Tolerance Principle (TP), Baayen's hapax-based measures (P and P*), and semantic similarity scores, we evaluate the extent to which different perfectivizing prefixes are synchronically productive. We construct a large-scale verb lexicon annotated for aspect, and leverage multiple corpora to identify novel prefixed word forms. Our findings reveal that productivity is not uniformly distributed across prefixes: some, like *za-* and *po-*, are frequent and semantically broad, while others, such as *niz-/nis-*, are rare and exhibit narrow unproductive usage, with most appearing productive. Finally, we examine the relationship between productivity and semantic transparency using cosine similarity, finding little evidence that meaning preservation drives rule productivity in the case of Russian prefixes.

1 Introduction

The morphological productivity of aspectual prefixation in Russian has been a subject of long-standing debate. In Slavic languages, the grammatical aspect is encoded in verbal morphology, distinguishing between perfective (PF) and imperfective (IMPF) actions. The IMPF aspect often correlates with atelicity, indicating events that do not have an inherent end-point or culmination, while the perfective aspect indicates completion. Perfectivizing prefixes, those that attach to IMPF base forms to derive PF verbs, are considered the most common morphological process for forming PF verbs in Russian (Forsyth, 1972). However, it remains unclear to what extent these prefixes function as productive processes. While the meanings of some derived verbs can be interpreted compositionally, others exhibit varying degrees of idiosyncracy. Frequency (Bauer, 2001), semantic coherence (Aronoff, 1976) and the ability to produce new forms (Hockett, 1954) are the three criteria for productivity that are often mentioned in the literature. Through a series of computational experiments, we measure the productivity of prefixation in forming PF verbs from simple IMPF verbs. Specifically, we assess:

- The productivity of perfectivization via prefixation.
- The semantic relatedness of PF verbs to their base forms.
- The correlation between productivity measures and semantic relatedness, as well as between productivity measures and a neologisms baseline.

2 Methodology

To quantify morphological productivity, we employ corpus-based and dictionary-based approaches. Specifically, we use two measures based on hapax-legomena, introduced by Baayen (Baayen, 1992): between-rule productivity P* and within-rule P productivity, along with the Tolerance Principle (TP) (Yang, 2005). Additionally, we propose a modified version of P* that incorporates dis-legomena (i.e., terms that occur exactly twice in the corpus), to capture potentially novel low-frequency forms. Furthermore, we compute TP using dictionary-based counts, while Baayen-style metrics rely on corpus statistics. This allows us to compare rule productivity from both a usage-driven and a lexicon-driven perspective.

2.1 Lexicon and Corpus Preparation

We begin by compiling a verb lexicon of 32,489 unique lemmas, each annotated for aspect. This lexicon is based on two sources: an online version of the Russian Morphological Dictionary [1] and a

[1] https://github.com/sshra/database-russian-morphology

precompiled Russian lemma lexicon based on the Grammatical Dictionary of the Russian Language (Zalizniak, 1977). The resulting combined lexicon is used to compute the TP threshold per prefix. Separately, we use a tagged corpus in CoNLL-U format (Nivre et al., 2016) to compute corpus-based derived statistics for Baayen-based measures. Due to computational expenses of processing the full Russian dataset, a subset of approximately 3,500,000 sentences was randomly selected. A variety of genres and styles are represented in the subset, from social media posts to literature passages and technical documents. Subsequently, we parse each sentence using the SynTagRus treebank of Russian model (Nivre et al., 2008) from DeepPavlov (Burtsev et al., 2018) trained on the UD corpora.

Prior to computing productivity measures, we preprocess the data by extracting simple verb forms and the prefixes they occur with, along with their aspects, for both approaches. We repeat the process of prefix extraction twice to account for prefix stacking. We also create a separate category for bi-aspectual verbs that have identical surface forms in PF and IMPF, and treat them as having a null prefix. We only use this category in TP computation for now. Additionally, we remove secondary IMPF verbs from our final set. As a result, we obtain valid prefix-aspect pairs, as defined in both implementations. We also map allomorphs of a given prefix to the same underlying form (e.g., *nis-* and *niz-* or *s-* and *so-*), giving us a total of 23 high-level prefixes.

2.2 Productivity Metrics

As a theory of rule learning, TP establishes a threshold for how many exceptions a productive rule can tolerate. Unlike frequency-based heuristics, TP models the cognitive plausibility of generalizations, explaining, for example, how minority rules, such as certain German plural patterns (Yang, 2016), can still be productive. The TP threshold is given by the formula:

$$\Theta_N = \frac{N}{\ln N}$$

where N is the total number of candidate items (in this case, all simple verb lemmas derived with a given prefix), and e is the number of exceptions (that is, IMPF verbs prefixed with the same prefix). A prefix is considered productive if and only if:

$$e \leq \Theta_N$$

In other words, a prefix that surpasses the threshold in forming PF verbs compared to IMPF verbs is likely productive under TP.

To complement the TP-based analysis, we calculate two metrics derived from the corpus-based statistics. The idea behind this approach is that, since productive affixes tend to give rise to novel words, their frequency distributions are likely to contain a large number of low-frequency forms. Therefore, a good estimate of the affix productivity might be computed as a proportion of low-frequency forms associated with the affix. Then P* is the proportion of all hapaxes in the corpus that are attributed to rule *r*, out of all hapaxes in the corpus. P, on the other hand, is the proportion of all words in the corpus that are attributable to *r* and appear only once, out of all words attributable to *r*. P* is used to compare the differential productivity of various affixes, while P measures the growth rate of the words derived via *r*.

To establish a baseline for productivity, we target neologisms chosen from the online dictionary of Russian neologisms of the 21st century.[2] Most reported neologisms are recent borrowings from English related to social media or technological concepts (e.g., *guglit'* 'to be googling' IMPF; *za-guglit'* 'to have googled' PF). We then compile a separate corpus for each neologism using the Russian web corpora database [3] to retrieve sentences containing the neologisms' base forms preceded by a given prefix. A full list of the neologisms used is provided in Appendix A. We also leverage the CC-100 dataset for Russian (Wenzek et al., 2020) to compute Baayen's productivity statistics and semantic similarity.

For semantic analysis, we use the pretrained neural embedding model DeepPavlov ruBERT (Kuratov and Arkhipov, 2019) to compute contextual embeddings for each token in a sentence, as well as embeddings for each verb from its subtoken components. We compute the average embedding score for each unique lemma and determine cosine similarity scores for each base verb – prefixed verb pair. Finally, we use Spearman's rank correlation to measure the relationship between cosine similarity and productivity metrics, as well as between productivity measures and the neologisms baseline.

[2] https://russkiiyazyk.ru/leksika/slovar-neologizmov.html
[3] https://int.webcorpora.ru/drake/

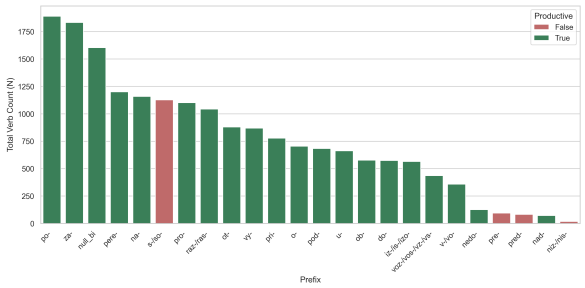

Figure 1: Prefix Productivity Based on the Tolerance Principle

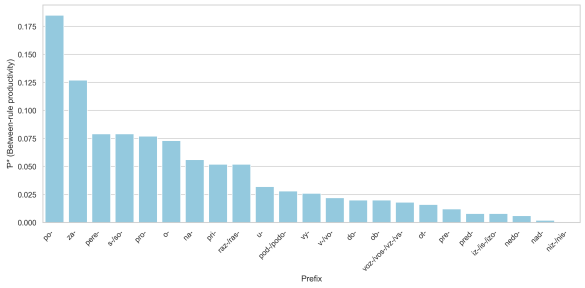

Figure 2: Between-Rule Productivity per Prefix

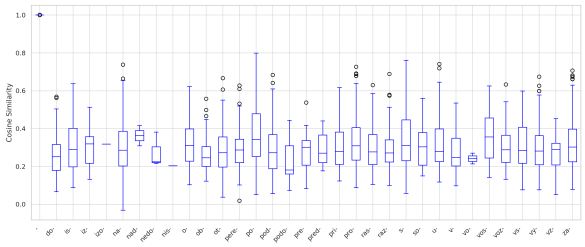

Figure 3: Cosine Similarity per Prefix

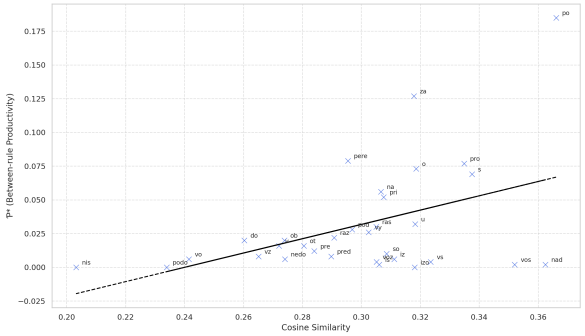

Figure 4: Cosine Similarity vs. Between-Rule Statistics

3 Results

Most prefixes passed the TP threshold, with the exceptions being *s-/so-*, *pre-*, *pred-*, and *niz-/nis-*. Figure 1 shows the TP results by prefix.

In contrast, P and P* yielded different rankings: *pere-*, *pre-*, and *ob-* scored highest under P, while *po-* and *za-* were most productive under P*. Both P and P* assigned *niz-/nis-* a score of zero. Results from between-rule P* measure are presented in Figure 2.

The neologism analysis showed strong alignment with P*, with *po-*, *za-*, and *s-* being dominant in recent coinages. Prefixes such as *niz-/nis-*, *nad-* and *pred-* were entirely absent from the neologism corpora, reinforcing their low productivity.

Cosine similarity between base and prefixed forms, computed using ruBERT, fell mostly between [0.2–0.37]. Prefixes like *po-* and *nad-* preserved meaning best, while *niz-/nis-* showed the largest shifts. Outlier-rich prefixes like *za-* and *pro-* suggest semantic variability within high-productivity classes. To get a clearer perspective on the distribution of prefixes, we plotted them in Figure 3 using the cosine similarity between each unprefixed–prefixed pair across all verb lemmas.

Finally, visualizing P* against cosine similarity, as shown in Figure 4, revealed no strong linear correlation, underscoring that semantic transparency and productivity do not always go hand in hand. However, both modified P* and P* exhibited the strongest correlation with neologism-based frequency, suggesting their utility in modeling current language trends. Table 1 presents correlation results.

Correlation Metrics	**Spearman's ρ**
Cosine Similarity & TP	.367
Cosine Similarity & P	−.065
Cosine Similarity & P*	.297
Cosine Similarity & Modified P*	.473
Cosine Similarity & Neologisms	.286
Neologisms & TP	.371
Neologisms & P*	.830
Neologisms & Modified P*	.743

Table 1: Spearman correlation coefficients between semantic similarity, productivity metrics, and neologism counts.

4 Discussion

Our results offer empirical evidence that Russian perfectivizing prefixes exist along a productivity continuum. The divergence in rankings across TP, P, and P* illustrates how dictionary-based and corpus-based models capture different nuances of rule behavior. The fact that most prefixes pass the TP threshold underscores their learnability, while

hapax-driven P* offers a stronger match to actual coinage trends. TP mainly diverged in its assessment of *s-/so-* and *pre-*, which emerged as unproductive. This warrants further investigation, as finer-grained semantic or phonological features may underlie these patterns and require more careful distinction.

We highlight *niz-/nis-* as a clear outlier across all metrics. It is not seen in any corpus of neologisms and is semantically the most divergent, suggesting that it is unproductive. Meanwhile, *za-* and *po-* illustrate how high productivity can coincide with semantic polysemy. These findings complicate the idea that productivity and semantic transparency necessarily align.

One interesting class of verbs that merits further attention is bi-aspectual verbs—forms that are compatible with both PF and IMPF contexts without overt morphological marking. In our analysis, we incorporated these verbs into the TP framework using a *null* prefix. Surprisingly, this class emerged as productive, supporting the idea that even prefix-less surface forms contribute to learnable morphological generalizations. A natural extension of this work would be to further refine how these verbs are integrated into prefixal paradigms.

We also acknowledge certain limitations, particularly around potential false decompositions. Some verbs exhibit suppletive forms or root alternations that can challenge prefix and suffix identification algorithms. While our pipeline attempts to minimize such cases, they could still influence prefix frequency or similarity metrics in subtle ways.

Importantly, our analysis offers a joint evaluation of productivity and semantic relatedness at scale using modern computational tools. While prior work has often assumed productivity as a binary feature or assessed it through intuition, our study quantitatively profiles prefixal behavior across thousands of verbs and aligns that with neologism usage and semantic drift.

5 Conclusions

Our findings suggest that almost all Russian PF prefixes are in fact productive, with one potential exception being both surface forms of the same underlying prefix niz-/nis-. This work provides a large-scale, computational account of aspectual prefix productivity in Russian. By combining Baayen's corpus-based productivity metrics, Yang's TP, and BERT-based semantic similarity, we show that:

- Prefixes differ in productivity, with P* best predicting real-world lexical innovation.
- Productivity does not always imply semantic transparency; highly productive prefixes like *za-* may exhibit broad or polysemous shifts.
- Discrepancies across metrics point to the need for multiple, complementary perspectives on morphological productivity.

Looking ahead, a deeper investigation into the semantic properties of base verbs—particularly those compatible with *s-/so-* and *pro-*, which were deemed unproductive by TP—may uncover finer-grained subregularities that the current TP-based approach classifies as exceptions. As Yang (Yang, 2023) observes, productivity does not always align with statistical dominance: minority patterns can be highly productive when conditioned by specific features. In our case, such conditioning is likely tied to semantic and phonological factors. Pursuing more granular semantic distinctions within each prefix class may therefore reveal minor but genuinely productive subrules that are obscured in aggregate analyses.

References

Mark Aronoff. 1976. *Word Formation in Generative Grammar*. Number 1 in Linguistic Inquiry Monographs. MIT Press, Cambridge, MA.

Harald Baayen. 1992. *Quantitative aspects of morphological productivity*, pages 109–149. Springer Netherlands, Dordrecht.

Laurie Bauer. 2001. *Morphological Productivity*. Cambridge Studies in Linguistics. Cambridge University Press.

Mikhail Burtsev, Artem Seliverstov, Yuri Kuratov, Dmitry Ermilov, Alexey Gureenkov, Denis Svirchev, Dmitry Kruchinin, Mikhail Shuvalov, Mikhail Aseev, Dmitry Ignatyev, and 1 others. 2018. Deeppavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127.

James Forsyth. 1972. The nature and development of the aspectual opposition in the russian verb. *The Slavonic and East European Review*, 50(121):493–506.

Charles F. Hockett. 1954. Two models of grammatical description. *WORD*, 10(3):210–234.

Yurii Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language.

Joakim Nivre, Igor M. Boguslavsky, and Leonid L. Iomdin. 2008. Parsing the SynTagRus treebank of Russian. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 641–648, Manchester, UK. Coling 2008 Organizing Committee.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Charles Yang. 2005. On productivity. *Linguistic Variation Yearbook*, 5:265–302.

Charles Yang. 2016. *The Price of Linguistic Productivity: How Children Learn to Break the Rules of LanguageHow Children Learn to Break the Rules of Language*.

Charles Yang. 2023. A user's defense of the tolerance principle: Reply to enger (2022). *Beiträge zur Geschichte der deutschen Sprache und Literatur*, 145:563–579.

A.A. Zalizniak. 1977. *Grammatical Dictionary of the Russian Language*. Firebird Publications, Incorporated.

A Neologisms

Neologism	**Gloss**	**# Sents**
spamit'	to spam	13,239
frendit'	to add as a friend	153
guglit'	to google	74,075
yuzat'	to use	37,430
donatit'	to donate	1,642
trollit'	to troll	2,469
čatit'	to chat	3,529
fotkat'	to take pictures	20,583
kserit'	to take a photo copy	155,328
skanit'	to scan	50,803
skrinit'	to screen	6,137
mejkapit'	to do make-up	32
piarit'	to promote	46,034
startapit'	to start-up	70
kopipastit'	to copy-paste	4,487
follovit'	to follow	2,663
lajkat'	to like	27,007
tegat'	to tag	1,576
šedulit'	to schedule	34
bathertit'	to be talked down to	36
hedlainit'	to make into a headline	11
monitorit'	to monitor	12,625
spoilerit'	to spoil (as in spoiler alert)	5,433
kreativit'	to be creative	7,632
brifit'	to brief	17
loginit'(s'a)	to log in	12,175
čekinit'(s'a)	to check in	2,171
tvitit'	to tweet	1,611
kommitit'	to commit	451
instagrammit'	to post on instagram	261