9-2021

Computational Representation of Russian Aspectual Morphology with a Focus on Perfective Prefixation

Natalia Tyulina
The Graduate Center, City University of New York

COMPUTATIONAL REPRESENTATION OF RUSSIAN ASPECTUAL MORPHOLOGY WITH A FOCUS ON PERFECTIVE PREFIXATION

by

NATALIA TYULINA

A master's thesis submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Master of Arts, The City University of New York

2021

Computational Representation of Russian Aspectual Morphology with a Focus on Perfective Prefixation

by

Natalia Tyulina

This manuscript has been read and accepted for the Graduate Faculty in Linguistics in satisfaction of the thesis requirement for the degree of Master of Arts.

Date	Kyle Gorman Thesis Advisor
Date	Cecelia Cutler Executive Officer

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

Computational Representation of Russian Aspectual

Morphology with a Focus on Perfective Prefixation

by

Natalia Tyulina

Advisor: Kyle Gorman

This work performs an empirical analysis of Russian aspectual morphology focusing on perfective derivation via prefixation. We present a number of computational experiments measuring productivity of morphological processes of prefixation that form perfective verbs from simple imperfective verbs. Several hypotheses related to the argument structure of perfective verbs vs. their prefixed derivatives are tested statistically. Furthermore, we investigate semantic relatedness by computing cosine similarities of unprefixed verbs vs. their prefixed versions. Finally, we analyze the correlation between productivity, frequency, argument structure and semantic similarity across both simple imperfective – prefixed perfective verb forms, and various perfectivizing verbal prefixes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction: Peculiarities of Russian Aspect

1.1. Problem Statement

The Slavic aspectual system has been a topic of ongoing interest and debate since at least the 1950s. Aspect in Russian, unlike many other non-Slavic languages, is encoded in verbal morphology, and exhibits a number of peculiar characteristics. It has been approached from syntactic (e.g., its role in the hierarchy of the syntactic domain of a clause), lexical (e.g., Vendlerian lexical aspectual classification), semantic (e.g., telicity and temporal relations), morphological (e.g., issues of word-formation and productivity), pedagogical (e.g., facilitating cross-linguistic comparisons), and semantico-pragmatic (e.g., Discourse Representation Theory and event semantics) perspectives. This is by no means a full list, as it has also been studied within psycholinguistic and other related frameworks, as well as at crossroads of various subfields and perspectives.

The focus of this research lies on the derivational and inflectional distinction in respect to the category of aspect and how it is expressed in Russian. This work is an attempt to analyze the Russian aspectual system empirically, as opposed to discussing the interaction between grammatical and lexical aspects. The main goal of this research is to perform a computational analysis of the perfective (PF) and imperfective (IMPF) verb forms by assessing the following criteria:

1. Productivity of perfectivization via prefixation.
2. Differences in the argument structure of PF and IMPF verbs.
3. The semantic relatedness of the PF verbs derived via prefixation.
4. Correlation between morphological processes of prefixation and semantic relatedness.

The main contribution of this work is performing an empirical computational analysis of contemporary Russian corpora and measuring productivity of morphological processes of affixation involved in formation of the Russian aspect. In section 1 a general literature overview is presented and the derivational vs. inflectional distinction along with the notion of morphological productivity are discussed in more detail. The next section reviews the corpora used and the preparation done to the corpora. Sections 3, 4 and 5 lay out the key points of conducted experiments, their design and the results obtained. In section 6 we look specifically at anything else that could be discussed and the future work objectives. Final conclusions are summarized in section 7, and additional figures are presented in section 8.

1.2. An Overview of the Russian Aspectual System

A lot of work has been done on Russian aspect within both theoretical and descriptive frameworks. Even though it is not the focus of this research, it is important to mention that interaction of lexical and grammatical aspects in Russian has been a major subject of disagreement in the literature. The traditional view on Russian aspect claims that aspect is lexically pre-attached to the verb, and its aspectual value is determined as soon as the morphosyntactic derivation of a verb stem is completed (Maslov 1959, Flier 1988). Therefore, it is assumed that Russian verbs originate in the lexicon as either perfective or imperfective (Isačenko 1960, Janda 2007). Other linguists have argued that as a lexical item verb is aspectless, and aspect appears as a part of the functional domain of a clause (Paslawska & von Stechow 2003). On either account, however, there seems to be a consensus at least in regard to the PF prefixes being derivational in nature. The main difference is that under a strictly lexicalist account, all derivational affixation happens before a verb enters

syntactic derivation. On an anti-lexicalist view, contrarily, derivational affixes are considered to be syntactically represented.

Furthermore, much of the recent work focuses on the role of prefixes in the formation of the perfective. Perfectivization is closely connected with prefixation. While various authors differ in their exact interpretation of this relationship, most of them agree that prefixation creates PF verbs from IMPF ones. Most lexicalist and anti-lexalist theories accept the derivational approach to the formation of PF verbs from simple[1] IMPF ones. It is basically assumed without much further analysis that prefixes should not be treated as instances of inflectional morphology. However, there has been a number of proposals that treat suffixes like ива-, that participate in a so-called 'secondary imperfectivization' of already prefixed PF verbs, as an inflectional morpheme rendering the IMPF semantic aspect (Filip 2005, Manova 2005). Tatevosov (2015) points out that such proposals face a number of empirical challenges. He mentions that the number of lexical items that hypothetically can be inflected is relatively small for ива-, as it does not exceed 30% of verb stems. Additionally, certain prefixes can attach in addition to ива-, which seems to be problematic in terms of admitting derivational morphology outside of inflectional morphology, according to Tatevosov.

1.3. Derivational vs. Inflectional Morphology

Outside of the lexicalist and anti-lexicalist debates that primarily focus on whether aspect belongs in the functional domain (after the verb projects) or is a part of the verb itself, the inflectional vs. derivational distinction of the Russian aspect has been another subject of disagreement. This distinction is not straightforward and can be discussed in terms of its category-changing

[1] By simple verbs we mean verbs that do not have any additional affixes added to the base.

characteristics, paradigm, productivity, position and frequency, syntactic restrictions, formation patterns and semantic irregularity, among others.

Generally speaking, inflection is viewed as a change in grammatical or morphosyntactic form, while derivation typically changes the syntactic class membership of the word. Assuming that aspect in Russian is grammaticalized, it is still an open question whether it belongs to the inflectional or derivational category. Proponents of the inflectional analysis of aspect (Isačenko 1962) claim that since inflection is always paradigmatic, the fact that most verbs form aspectual pairs speaks in favor of the inflectional analysis. Aspectual pairs (APs), according to the *Academy of Russian Grammar* (henceforth, RG), represent a systematic characteristic of Russian verbs. RG further classifies them into the following types:

- APs formed by means of perfectivization:
    - Simplex IMPF verb - prefixed PF verb, e.g.: IMPF писать ‘write’ - PF написать ‘have written’[2]
- APs formed by means of imperfectivization:
    - Prefixed PF verb - suffixed IMPF verb (with an IMPF suffix), e.g.: PF переписать ‘have rewritten’ - IMPF переписывать ‘rewrite’
    - Prefix-less PF verb - suffixed IMPF verb (with an IMPF suffix), e.g.: PF решить ‘have decided’ - IMPF решать ‘decide’
- APs with different roots or suppletive APs (the most rare and irregular type), e.g.: IMPF брать ‘take’ - PF взять ‘have taken’.

RG further states that such prefixes as в-, до-, над-, недо-, низ-, пред-, пре-, со- and the prefixes borrowed from Latin, such as де- (as in deduplicate), дис- (as in disallow), ре- (as in remake)

[2] English glosses are largely approximated since it is not possible to capture the exact contrast between the members of a Russian AP in English.

never form APs. According to RG, these prefixes either do not perfectivize verbs, or when they do, they also change its lexical meaning (RG, §1393).

The existence of suppletive forms (1) has been frequently mentioned as instances of inflectional morphology. However, besides the fact that these forms are extremely rare, most analyses do not explain the motivation behind the pairing of such forms. It is not exactly clear on what basis the members of suppletive APs are paired up, and why such verbs cannot be analysed as not having a corresponding AP (such gaps do exist, examples are given in (4) and (5) below).

(1) IMPF говорить 'to speak' - PF сказать 'to say' or 'to tell'

It might seem that IMPF and PF verbs in (1) convey the same lexical meaning. However, subtle differences can be detected, in terms of both semantics and argument structure. Semantic or lexical meaning patterns deserve special emphasis. In traditional Russian grammars, much work done on analyzing individual meanings of prefixes. Subsequently, these prefixes were grouped and regrouped according to certain patterns of the shift in meaning. We will not go into the details of such groupings here, since they tend to overcomplicate empirical analysis.

Another argument in favor of the inflectional analysis concerns the paradigm of grammatical forms. While both PF and IMPF verbs form infinitives, past tense, imperatives, conditionals and past active participles, other forms are limited to one aspect. Neither PF nor IMPF verbs have a complete paradigm of grammatical forms: PF verbs lack present tense forms whereas IMPF verbs lack synthetic future tense forms. The example given in (2) is ungrammatical since an analytic future form is never available for PF verbs. Furthermore, the present tense inflection of PF verbs gives rise to a future reading (3).

(2) *Я буду по-смотреть[3] (PF) 'I will be see'

[3] Dashes are purposely added after prefixes and before suffixes to illustrate morphological boundaries in some of the examples. They are not a part of standard Russian orthography.

(3) Он пишет письмо (IMPF) ‘He is writing a letter’

Он на-пишет письмо (PF) ‘He will write a letter’

RG, however, somewhat ambiguously classifies the category of aspect as “non-inflectional”, rather than either inflectional or derivational. One of the arguments in favor of the derivational approach to Russian aspect, is that traditionally it is prefixes and suffixes that form PF and IMPF forms, but not inflection, which is known to be a typical mechanism of inflectional morphology. Additionally, not all verbs come in aspectual pairs: certain IMPF verbs do not have corresponding PF forms (4).

(4) знать ‘to know’,

находиться ‘to reside’,

стоить ‘to cost’,

соответствовать ‘to correspond’,

противоречить ‘to contradict’, etc.

The same is true for some IMPF verbs that do not have analogous PF pairs (5).

(5) очнуться ‘to regain consciousness’,

хлынуть ‘to flow’,

состояться ‘to take place’, etc.

While in some cases it is possible to retrieve the meaning of a prefixed simple verb compositionally (e.g., по- often stands for reduction in action expressed by the original simple verb as illustrated in 6), in other cases the meaning is partly or fully lexicalized. A crucial point of inflectional morphology is that it does not trigger semantic changes. Russian verbs, however, frequently take prefixes that result in modifying their original meaning in addition to perfectivizing them. Prefixes

can add a wide range of extra interpretations, including diminutive, completive, spatial, cumulative, inchoative, and distributive. Examples of such prefixed PF forms are listed in (6):

(6) по-делать ‘do something for a short period of time’ (diminutive)

до-делать ‘finish doing something’ (completive)

за-плакать ‘burst into tears’ (inchoative)

пере-йти ‘cross’ (spatial)

у-делать ‘out-do; spoil something’ (completive)

раз-делать ‘carve; lay out’ (distributive)

на-делать ‘do something wrong; inflict’ (inchoative)

In some cases, the resulting meaning can obviously be categorized as mostly completive or spatial, for example, but in many other cases, the categorization of the newly formed verb remains unclear. Another puzzling question concerns double or even triple prefixation (or prefix stacking). In (7-9) some multi-prefixed verbs are illustrated.

(7) PF пред-по-читать ‘prefer’

(8) PF по-на-делать ‘make a large number of something; do something at a larger scale’

(9) PF пред-рас-по-ложить ‘predispose’

In some cases, like (8), the new meaning is still largely based on the original simple verb (in this case it is читать ‘read’), whereas in other cases like (7) and (9), the new meaning has shifted considerably from the original prefix-less verb. In the subsequent sections, we present a number of experiments to further challenge the aforementioned arguments for both inflectional and derivational approaches.

1.3.1. Perfective Derivation

Prefixation is considered the most common morphological process to form PF verbs. In traditional Russian grammars, there exists a further classification into lexical (Babko-Malaya 2003, Romanova 2004) and superlexical (Isačenko 1960) prefixes. Other linguists use different terms to differentiate between these two classes, such as *semantic* versus *semantically vacuous* prefixes, *qualifying* versus *modifying*, *internal* versus *external*, *specific* versus *natural*. Despite such variance in terminology, they all seem to point to the same distinction. Lexical prefixes, supposedly, alter the original meaning of a verb, and can derive so-called secondary imperfective forms (i.e., when a PF form undergoes another affixation via suffixation, producing an IMPF form), and are positioned closer to the root in the case of prefix stacking.

Furthermore, there appear to be various degrees of semantic shift, which makes it even more challenging to categorize prefixes. Most prefixes in Russian can convey more than one meaning, depending on the stem they join, and the same prefix could be classified as both lexical (10) and superlexical (11). Lexical prefixes affect the core lexical meaning of the verb, resulting in idiosyncratic and not necessarily compositional meanings. Moreover, they may change the verb's argument structure. Superlexical prefixes, on the other hand, do not change either the lexical meaning or the argument structure of the verb. Additionally, they are normally considered incompatible with secondary imperfectivization (Babko-Malaya 1999). However, at least as far as the meaning is concerned, it is not always clear where to draw a line between lexical and superlexical prefixes.

(10) PF по-любить 'fall in love'

PF за-бегать 'start running'

(11) PF по-писать 'write for a short period of time'

PF за-строить ‘build up’

Besides prefixation, there are a few other mechanisms of forming PF verbs, such as -ну suffixation (12) or a thematic vowel change triggered by stress alternation (13).

(12) IMPF прыгать ‘jump’ - PF прыг**ну**ть ‘have jumped’

(13) IMPF брос**а**ть ‘throw’ - PF брос**и**ть ‘have thrown’

1.3.2. Imperfective Derivation

Secondary imperfectivization (SI) is a term that is sometimes used to refer to IMPF derivation. Some linguists consider such derivatives a third element added to the simple IMPF - prefixed PF pair (Tatevosov 2015). SI from either prefixed or non-prefixed PF verbs has been assumed to be fairly regular via suffixation, but certain subsequent changes still occur. As far as the lexical meaning is concerned, such pairs are considered more identical than a simple IMPF - prefixed PF verb pair. Additionally, not all verbs have this derivation available. In (14), for example, the secondary IMPF cannot be derived, while in (15) it is perfectly grammatical.

(14) IMPF делать ‘to do’ - PF с-делать ‘have done’ - IMPF *сделывать

(15) IMPF делать ‘to do’ - PF пере-делать ‘have redone’ - IMPF пере-дел-ыва-ть

A traditional approach used in Russian grammars employs a single suffix *(-ива)* analysis. This suffix can be represented by the following morphs[4] (depending on a verb’s morphological characteristics): *-а, -ва, -ива (-ыва)*. A number of predictable and unpredictable changes can be triggered by IMPF derivation. The most common ones are basic rules of combination (Levin 1978), suffix truncation, and changes to the root vowel. Basic rules of combination are triggered when suffixes are added to base forms, resulting in one of four possible combinations: V+V, C+C,

[4] By morph we mean a single shape of a morpheme, as opposed to allomorph, which stands for various shapes of a morpheme (Hockett 1954)

V+C, C+V. Whereas unlike elements combine without any changes, the combination of like elements normally leads to the truncation of the first one (~~V~~ + V or ~~C~~ + C). These are considered regular patterns that work the same way for all verb types. In addition to the basic rules of combination, there are a few other changes that occur in IMPF derivation, such as V+V mutation, suffix truncation, и / ы insertion before the final consonant of the root in all non-syllabic stems, and root vowel change (from o to a).

It is important to note that, according to RG, an aspectual pair is defined as a pair of lexically identical PF and IMPF verbs that only differ in terms of the grammatical semantics of aspect, e.g.:

(16) IMPF делать 'to do' - PF сделать 'have done'

(17) PF пере-делать 'have redone' - IMPF пере-делывать 'to redo (regularly)'

Therefore, the category of aspect is viewed as 'non-inflectional', according to RG, and the aspectual pairs are motivated by derivational relations (with the rare exception of suppletive forms). To illustrate a somewhat opposing perspective, we should mention that Maslov and Bondarenko, among others, agree that (16) is an AP, but do not consider verbs in (17) a valid AP due to the lack of regularities and subsequent semantic changes associated with prefixation compared to suffixation.

1.3.3. Bi-aspectual Verbs

RG defines bi-aspectual verbs as "verbs whose aspectual meaning for the majority of forms is not expressed by special means." Bi-aspectual verbs present an intriguing phenomenon within the Russian aspectual system. They behave differently from most verbs, and do not fit into the standard aspectual model with its main focus on APs. Even though they lack any aspectual markers, they

are perfectly capable of expressing both PF and IMPF aspects. Below we illustrate a bi-aspectual verb исследовать ‘explore; study; analyze’ used as IMPF (18) and PF (19).

(18) Я при вас его допрашивал, IMPF **исследовал** все обстоятельства и не нашел его виновным.

Right in front of you I was interrogating him, studying all the circumstances, and I found him innocent.

(19) Будда PF **исследовал** этот вопрос и пришел к выводу, что страданием являются все движения ума без исключения.

Buddha has studied this subject and came to a conclusion that all of the brain’s activity leads to suffering without exception.

Both forms are phonologically and morphologically identical but can belong to different aspectual classes. Isačenko (1960), for instance, referred to these verbs as an anomaly that should be eliminated. This view, however, lacks merit. The aspectual meaning of such verbs can be easily disambiguated within the context. Besides, they are not the only class of verbs that do not fit into the model of APs (as we have seen earlier in the verbs that do not have aspectual correlates, as well as verbs that have more than one aspectual correlate). According to the research by Zalizniak and Šmelyov (2000), bi-aspectual verbs need to be incorporated into the aspectual system. They suggested having three classes of verbs in this system: aspectually paired, aspectually unpaired, and bi-aspectual. This conception of the aspectual system seems more satisfactory, since it can actually account for various phenomena, instead of considering them anomalies. It does not seem to be the case that bi-aspectual verbs do not fit into the aspectual system, but rather that the system itself requires a broader understanding and reevaluation. Without going too much further into the specifics of bi-aspectual verbs, we will mention that they can be generally divided into two groups:

etymologically Russian (considered not productive anymore, according to RG) and Western European borrowed verbs (considered productive). The latter group is considerably larger, and includes verbs like презентовать ‘present’, резервировать ‘reserve’, спонсировать ‘sponsor’, финишировать ‘finish’, etc.

1.4. The Notion of Morphological Productivity

The term ‘productivity’ has been frequently used in various linguistic contexts, but there is still no unified definition to be found. Some linguists have made clear distinctions between productivity and creativity, while others consider the two concepts to be either fully or partially overlapping. Hockett (1958: 307), for example, broadly characterized it this way: “The productivity of any pattern – derivational, inflectional or syntactical – is the relative freedom with which speakers coin new grammatical forms by it,” while Di Sciullo and Williams (1987) mentioned that the fact that an affix can be used to make new words makes it “productive in the most basic sense of the word.” Although it may seem straightforward, productivity has been associated with different concepts in the literature. Various researchers have discussed this notion on different levels, from affixes and morphological processes to rules, complete grammar modules, or even features of language systems. However, according to Bauer, those views do not necessarily have to contradict one another. If we assume that a process, such as derivational morphology, is productive, then we can further break it down into individual derivational processes like affixation. The same logic can be applied to the language system as a whole. What matters here is that there are certain patterns of making new word forms. Bauer raises a question of the productivity of the paradigm slot, stating that in order to fill such a slot for a given case, some morphological process is expected to be in place. Another question is whether there exist various degrees of productivity or a lack thereof.

For the time being, we will assume that, even if the various degrees exist, productivity can still be regarded as a feature of certain morphological processes rather than as the slots in a paradigm. It is also important to mention that some linguists use the term 'semi-productivity.' Pinker and Prince (1991) use this term in a relation to a process that can be extended to new forms only to some degree. Frequency (Bauer 2001), semantic coherence (Aronoff 1976), and the ability to produce new forms are the three criteria for productivity that are often mentioned. This work will not examine the details of synchrony and diachrony but will focus on observations made from the analyzed corpora. According to RG, certain verbal affixes are labeled as productive, since they have been continuously recorded in oral or colloquial speech before making their entry into the dictionary, and in some cases never even became a part of the norm (labeled as "просторечье"). Furthermore, for native speakers, it seems perfectly satisfactory to coin new word forms by attaching such affixes to the existing verb forms. Most affixes that are labeled as productive come in a combination of prefix – base – reflexive[5] suffix.

Another important issue to address in this regard is what mechanisms can be appropriate for establishing the productivity of certain processes. An interesting consideration is the notion of existing words (in the speech community) and the way they determine potential new word formations. Bauer gives a working definition of the productivity of a morphological process as "its potential for repetitive non-creative morphological coining." Thus, here it is distinguished from creativity such that productivity is seen as rule-governed, and creativity is viewed as rule-changing. Some researchers have developed quantitative approaches to measuring the productivity of morphological processes. Aronoff (1976) developed one of the first methods for quantifying morphological productivity. His main idea was to estimate the number of all possible words that

[5] Reflective verbs in Russian take either the suffix -сь or the suffix -ся.

were produced by a certain word formation rule (WFR), as opposed to the ones that occurred in a sampled text. The concept behind WFR basically revolved around the ratio of all possible words to actual words formed by a given rule. Another interesting quantitative approach was introduced by Baayen (1992). He formalized a way to calculate the probability of coming across new words in a sampled text. O'Donnell (2015) has developed further the idea of creativity being possible by computation. In his proposal of productivity and reuse as probabilistic inference, he introduced the model of Fragment Grammars, and compared it to other probabilistic models.

2. Materials Used

Two main sources of data are used in the experiments, which are described in more detail in sections 2.1 and 2.2. Across all experiments, the `spaCy` library was used for the Russian specific sentence segmentation and tokenization. To avoid false decompositions, we consistently sort the prefix removal starting with the longest prefixes. Furthermore, in order to account for cases with stacked prefixes, we run the removal function multiple times to ensure that an actual simple verb is the result of the prefix stripping.

2.1. Verb Lexicon

For the first set of experiments (section 3), a lexicon of verbs was compiled for further analyses. The data was taken from two sources: an online version[6] of the *Russian Morphological Dictionary* (Hagen 2014) and a pre-compiled Russian lemma lexicon[7] based on the *Russian Grammar Dictionary* by Zalizniak (Zalizniak 1980). For the purposes of computing frequency and productivity metrics, only verb lemmas with their aspectual categories were extracted. The first lexicon consisted of 23,395 unique lemma-aspect pairs, while the second one had 27,552. Even though most of the vocabulary was overlapping between these two sources, we still decided to combine both lexicons to obtain a fuller and more representative vocabulary sample. This combined lexicon consisted of 32,489 unique lemma-aspect pairs. We should also note that some labels for bi-aspectual[8] verbs from the *Russian Morphological Dictionary* had to be corrected. We checked them against a different Zalizniak-based online dictionary,[9] as well as another dictionary

[6] https://github.com/sshra/database-russian-morphology
[7] https://github.com/undrits/ruslex
[8] Verbs that have identical PF and IMPF forms (see section 1.3.3).
[9] https://starling.rinet.ru/

by Ushakov.[10] All of the mislabeled verbs were originally labeled as "NULL" (standing for bi-aspectual verbs); examples of corrections are illustrated in Table 1.

Table 1: Examples of corrected aspect labels

verb	**source label**	**final label**
быть 'be'	NULL	IMPF
регулировать 'regulate'	NULL	IMPF
публиковать 'publish'	NULL	IMPF
раз-минировать 'demine'	NULL	PF
рушить 'destroy'	NULL	IMPF
об-народовать 'make public'	NULL	PF

Furthermore, some ambiguous cases were discovered when comparing aspectual labels between the Ushakov (UD) and Zalizniak dictionaries (ZD). In such cases, we consulted a third dictionary, an online version of *Encyclopedic Dictionary of the Russian Language*[11] (ED). A few examples of such cases are presented in Table 2.

Table 2: Examples of ambiguous aspect labels

verb	**source label**	**UD aspect label**	**ZD aspect label**	**ED aspect label**	**final label**
рекламировать 'advertise'	NULL	IMPF & PF	IMPF	IMPF & PF	IMPF & PF
ре-организовать 'reorganize'	NULL	Not found	PF, IMPF in present & future forms	IMPF & PF	IMPF & PF
характеризовать	NULL	IMPF & PF	not found	IMPF	IMPF & PF

[10] https://slovar.cc/rus/ushakov
[11] https://slovar.cc/enc/slovar

‘characterize’					
шампанизировать ‘champagnize’	NULL	IMPF & PF	IMPF	IMPF	IMPF

2.2. CC-100 Dataset

For the second set of experiments (sections 4 and 5), the CC-100 dataset was used.[12] The corpus consists of separate monolingual datasets for over one hundred languages, including Russian. It was generated using the Common Crawl open net repository.[13] More information about the original extraction and compilation of these datasets can be found in the paper by Wenzek et al. (2019). To bypass the computational expense of processing the full Russian dataset, a subset of approximately 3,500,000 sentences was randomly generated. The data has document and paragraph delimiters. A variety of genres and styles are represented in the subset, from social media posts to literature passages to technical documents.

[12] http://data.statmt.org/cc-100
[13] https://github.com/facebookresearch/cc_net

3. Measuring Morphological Productivity

Keeping in mind the argument about whether the aspectual category is inflectional or derivational, we evaluate the productivity of various morphological processes. While inflectional processes are typically assumed to be productive, derivational ones could be either productive or not. In this section, we measure morphological productivity of prefixation as a perfectivizing process when applied to a simple verb. Then we evaluate the results using a number of methods.

3.1. Experiment Design

Several metrics have been proposed in the literature to measure morphological productivity. In this research we adopt the Yang style analysis by applying the Tolerance Principle (Yang 2018) to Russian perfectivizing prefixes. The Tolerance Principle (TP) is formulated as follows:

Let a rule R be defined over a set of N items. R is productive if and only if e, the number of items not supporting R, does not exceed Θ_N :

$$e \leq \theta N = \frac{N}{lnN}$$

Thus, productivity is determined by the relationship between *e* and and *N*. If *e* exceeds *θN*, then the learner will "lexicalize" only these "exceptions" and not generalize beyond them: that is, a given rule *R* is unproductive. In other words, in order to be considered productive, a rule must have relatively few exceptions. In our case, the rule R is perfectivization of simple verbs via prefixation. IMPF verbs containing additional affixes are not being considered here. For this analysis we used all of the possible verbal prefixes (except for the three borrowed ones due to their low frequency as perfectivizors), and not just the ones typically characterized as "pure perfectivizors" (i.e., prefixes that change a verb's aspectual category without altering its lexical meaning in any way).

This was motivated by the definition of the rule, as well as by our goal to compare the productivity of a wider range of perfectivizing prefixes. Thus, we do not draw any additional distinctions as far as the resulting lexical meaning is concerned. N corresponds to the number of simple verbs that take a given prefix.

First, we removed prefixes and suffixes[14] from all verbs in the combined lexicon. To avoid false decompositions, we also checked whether resulting prefix-less verbs (called 'p1less_verb' since only the outermost prefix is removed at this stage) occur in the lexicon. A preliminary data frame of the following format was created (Table 3).

Table 3: Sample data frame of morphological features

index	**verb**	**aspect**	**p1less_verb**	**prefix**	**s-less_verb**	**suffix**
0	абсолютировать	impf	NaN	NaN	абсолютир	-овать
1	абсолютировать	perf	NaN	NaN	абсолютир	-овать
2	агглютинировать	impf	NaN	NaN	агглютинир	-овать
3	агглютинировать	perf	NaN	NaN	агглютинир	-овать
4	агрегатировать	impf	NaN	NaN	агрегатир	-овать
...	...	...	...	...	...	...
32484	всходить	impf	ходить	вс-	NaN	NaN
32485	обреветься	perf	NaN	NaN	NaN	NaN
32486	снаряжать	impf	наряжать	с-	NaN	NaN
32487	подичиться	perf	дичиться	по-	NaN	NaN
32488	вытравляться	impf	NaN	NaN	NaN	NaN

[14] Technically, those are not suffixes, but groups of morphs.

Subsequently, we removed additional prefixes and generated simple verb forms. In Table 4, some of the rows for the verb читать ‘read’ are illustrated. We should also note that in the case of multiple prefixes, the outermost prefix was considered to be perfectivizor.

Table 4: Sample data frame for the verb читать ‘read’

verb	**aspect**	**p1less_verb**	**prefix**	**s-less_verb**	**suffix**	**simple**	**p2less_verb**	**prefix2**
почитать	perf	читать	по-	NaN	NaN	читать	читать	NaN
почитать	impf	читать	по-	NaN	NaN	читать	читать	NaN
считать	perf	читать	с-	NaN	NaN	читать	читать	NaN
дочитать	perf	читать	до-	NaN	NaN	читать	читать	NaN
засчитать	perf	считать	за-	NaN	NaN	считать	читать	с-
вычитать	perf	читать	вы-	NaN	NaN	читать	читать	NaN

Table 5 below reflects distribution of verbal prefixes with PF and IMPF aspects (some prefixes with lower frequencies were omitted in the table).

Table 5: Prefix frequency by aspect

	в	**вы**	**до**	**за**	**из**	**на**	**о**	**об**	**от**	**пере**	**по**	**под**	**при**	**про**	**раз**	**рас**	**с**	**у**
impf	34	38	35	56	19	46	26	31	39	90	33	43	88	40	39	19	94	46
perf	195	609	375	**1308**	186	**830**	509	389	557	**771**	**1537**	466	526	**761**	295	338	568	459

As shown in Table 5, all of the PF prefix values are higher than the IMPF ones. The most frequent perfectivizing prefixes turned out to be по-, за-, на-, про-, пере-.

3.2. Results

We formulated and tested a few rules by applying the TP formula of productivity to the data described in 2.1 and 3.1. The first two rules measure the productivity of each PF prefix individually, while the last one measures the productivity of all PF prefixes as one concept.

Rule 1: A PF prefix is productive if and only if e, a number of simple verbs that can be perfectivized by other prefixes, but not by this prefix, does not exceed Θ_N, where N is the number of simple verbs that occur with this PF prefix. As a result, every single prefix was unproductive by this method. For example, even for the two most frequent PF prefixes, the following presented in Table 6 values were obtained.

Table 6: Productivity Results Method 1

prefix	**N**	**Θ_N**	**e**	**result**
по-	1289	180	1262	unproductive
за-	1278	178	1273	unproductive

It could be the case that a more appropriate comparison between e and Θ_N should be formulated. One of the hypotheses as to why this method might not be reliable centers on its inability to account for potential new formations that have either not taken place yet (for the sake of clarity, we can refer to them as the not-yet-used examples), or that have not been registered in dictionaries as a norm. For example, one of the relatively recent borrowings from English was the verb френдить ‘to add as a friend (normally on social media).’ There also exist prefixed versions of this verb that have been widely used, first in the internet community, and then in colloquial speech in general, such as, PF за-френдить ‘add as a friend,’ PF от-френдить ‘not accept as a friend’ or ‘remove

from friends,' PF рас-френдить 'remove from friends.' Unfortunately, those words have not yet been added to the dictionaries we used.

Rule 2: A PF prefix is productive if and only if e, the number of occurrences of this prefix as an IMPF prefix, does not exceed Θ_N, where N is the number of simple verbs that occur with this prefix as a PF one. Most prefixes were productive by this method, contrary to the previous one. Θ_N and e values are presented in Table 7 (section 8) for each prefix. These results, contrary to Rule 1, seem to offer better general intuitions about certain prefix productivity. Those prefixes categorized by RG as never participating in APs (в-, до-, над-, недо-, низ-, пред-, пре-, со-), i.e., prefixes that always alter the underlying lexical meaning of the verb, received the lowest productivity scores by this method. This also goes along with assumptions about derivational morphology being less productive and more meaning-changing than inflectional morphology.

Rule 3: A PF prefix is productive if and only if e, the number of occurrences of PF verbs not formed via prefixation does not exceed Θ_N, where N is the number of simple verbs that are formed via prefixation. We identified 4,194 verbs in our data frame that were PFs not formed via prefixation, while 12,969 verbs were categorized as formed via prefixation. Then, the following values were used to compute the productivity score:

e = 4,194

N = 12,969

Θ_N = 1,369

According to this method, $e > \Theta_N$; therefore, PF prefixes in general were reported as unproductive.

4. Argument Structure of Perfective and Imperfective Verbs

It has been claimed in the literature that lexical prefixes may change the argument structure of the verb, whereas purely perfectivizing prefixes never do (Kagan 2012). However, according to Bailyn (2012: 32), morphological aspect "does not directly interact with argument structure." We decided to test this further, since not much empirical evidence has been reported in the literature to determine whether aspectual properties can actually affect argument structure. Additionally, we should note that the terms argument structure and valence are used interchangeably here.

4.1. Experiment Design

The hypothesis tested in this section is that perfectivized PF verbs statistically significantly differ from the original simple verbs in terms of their argument structure. To implement this computationally, the `syntagrus`[15] model from `deeppavlov`[16] trained on UD corpora[17] was used. Then the sentences were parsed in CONLL-U format, containing part of speech tags, dependency relations, and other features such as aspect. Subsequently, the valence data frame was generated with dependency counts for every verb in the dataset. Sentential subjects and optional arguments were included in valence counts. A sample of the data frame is presented below in Table 8.

Table 8: Sample Valence Data Frame

word_form	lemma	valence	deprel	aspect	prefix	stem	suffix	has_prefix	multiprefix

[15] https://aclanthology.org/C08-1081.pdf
[16] https://www.researchgate.net/publication/326649360_DeepPavlov_Open-Source_Library_for_Dialogue_Systems
[17] https://universaldependencies.org/treebanks/ru_syntagrus/index.html

указать	указать	1	csubj	Perf	у-	указа	NaN	TRUE	TRUE
скажут	сказать	5	root	Perf	с-	сказа	NaN	TRUE	TRUE
скажите	сказать	2	root	Perf	с-	сказа	NaN	TRUE	TRUE
скажут	сказать	3	root	Perf	с-	сказа	NaN	TRUE	TRUE
сказать	сказать	1	csubj	Perf	с-	сказа	NaN	TRUE	TRUE
показать	показать	1	xcomp	Perf	по-	показа	NaN	TRUE	TRUE
покажем	показать	2	root	Perf	по-	показа	NaN	TRUE	TRUE
сказанного	сказать	2	xcomp	Perf	с-	сказа	NaN	TRUE	TRUE
показали	показать	4	parataxis	Perf	по-	показа	NaN	TRUE	TRUE
указать	указать	2	advcl	Perf	у-	указа	NaN	TRUE	TRUE

4.2. Results

We performed several two-sided t-tests for two independent samples. The mean valence count was used for each unique verb lemma. We set $\alpha = .05$ for all tests, and the detailed results for each test are reported below.

The first test was performed to determine whether valences of prefixed PF verbs are statistically significantly different from valences of non-prefixed IMPF verbs. The first sample consisted of unprefixed IMPF verbs, while the second sample consisted of prefixed PF verbs. The sample mean for non-prefixed IMPF verbs was 2.325, and the sample mean for prefixed PF verbs was 2.245. The results are significant at $\alpha = .05$ ($t = 53.565$, d.f. = 3,405,251, $p < .001$; 95% CIs: .076, .082), leading us to reject the null hypothesis that two samples have equal valences.

IMPF:$\underline{x}$= 2.325, std = 1.375, *s = 1,773,168*

PF: $\underline{x}$ = 2.245, std = 1.366, *s = 1,632,085*

Therefore, the alternative hypothesis is accepted, namely, that valence of non-prefixed IMPF verbs is statistically significantly higher than the valence of prefixed PF verbs.

The second test was conducted to determine whether valences of verbs that never occur with prefixes are significantly different from valences of verbs that occur with one or more prefixes. The first sample included cases where prefix was empty irrespective of aspect, while the second sample included instances with non-empty prefixes. The sample mean for the first group was 2.232, and the sample mean for the second group was 2.333. The results are significant at α = .05 *(t = -48.723,* d.f. = 2,673,826, *p < .001;* 95% CIs: -.108, -.1), leading us to reject the null hypothesis.

IMPF:$\underline{x}$= 2.232, std = 1.315, *s = 484,143*

PF: $\underline{x}$ = 2.333, std = 1.360, *s = 2,189,685*

Considering that we obtained negative statistics results, the valence of non-prefixed verbs is statistically significantly lower than the valence of prefixed verbs (irrespective of aspect).

In the last test, our goal was to compare the statistical significance of valences between non-prefixable PF verbs and prefixable PF verbs. The first group represented cases where the prefix was empty with the PF aspect, and the second group included cases where prefix was not empty with the PF aspect. The sample mean for the first group was 2.184, and the sample mean for the second group was 2.254. The results are significant at α = .05 (*t = -24.291,* d.f. = 1,632,083, *p < .001*; 95% CIs: -.085, -.072), leading us to reject the null hypothesis.

IMPF:$\underline{x}$= 2.184, std = 1.341, *s = 199,886*

PF: $\underline{x}$ = 2.254, std = 1.370, *s = 1,432,199*

Considering that we obtained negative statistics results, the valence of non-prefixed PF verbs is statistically significantly lower than the valence of prefixed PF verbs.

Since all of the groups exhibited statistically significant changes in valence, we should take note of this as being characteristic of derivational morphology. One of the future goals is to perform similar valence tests for each prefix individually.

5. Semantic Relatedness

5.1. Experiment Design

Finally, we computed cosine similarities[18] of simple verbs vs. their prefixed derivatives. The `Pymorphy2`[19] library was used for part-of-speech tagging to extract verbs and to generate normal forms (i.e., lemmas). A context window of 5 tokens (two tokens preceding a target word and two tokens following it) was implemented to incorporate some of the contextual information. The `scikit-learn`[20] Hashing Vectorizer was used to generate embeddings for all verbs as well as other words within the context window. We called a `fit_transform` function to generate a sparse array for each token. For the first sample set, we selected verbs with no prefixes or IMPF suffixes. The second sample included verbs with prefixes and no IMPF suffixes. The dimensionality in the output matrices was set to 16, which we are going to tune in future work. Cosine similarity was computed for every verb pair, and was averaged across the verb lemmas generated for every inflectional variant. Additionally, we decided to keep the real full-range angles in the range of -1.0 to 1.0 (default).

5.2. Results

The main goal of this experiment was to assess semantic relatedness between simple verbs and all their prefixed derivatives, and then evaluate PF prefixes and their potential semantic contributions. In Table 9 (section 8), we present average similarity scores for each prefix. As we can see from Table 9, there is not much variation between prefixes in terms of their average cosine similarity

[18] https://nlp.stanford.edu/IR-book/html/htmledition/the-vector-space-model-for-scoring-1.html
[19] https://pymorphy2.readthedocs.io/en/stable/
[20] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html#sklearn.metrics.pairwise.cosine_similarity

scores. Although most of the values are positive, and we need to keep in mind that the range here is [-1, 1], the highest average score per prefix was 0.16. However, there was a lot more variation in terms of individual scores per prefix. In Table 9 (section 8), we illustrate one example per prefix with the highest similarity score across all other verbs with that prefix. The highest recorded score was .89, which is considerably above the cumulative average.

While analyzing the correlation of this experiment with prefix frequency and productivity results, we observed certain tendencies. The most frequent prefixes - по- (.14), за- (.08), на- (.12), про- (.09), пере- (.07) - were spread across the range of scores, but without going too far below the overall average of .09. According to the productivity results based on the application of rule 2, the top four prefixes overlapped with the most frequent ones.

Subsequently, we tested the claim stated in RG that prefixes в-, до-, над-, недо-, низ-, пре-, пред-, со-, де-, дис-, ре- do not form aspectual pairs since they alter the lexical meaning of the verb. The first group reflected cosine similarities of unprefixed simple verbs vs. prefixed verbs with one of the prefixes listed above. The second group included cases of cosine similarities of unprefixed simple verbs vs. verbs prefixed with all other PF prefixes. The sample mean for the first group was 0.091, and the sample mean for the second group was 0.098. The results are significant at $\alpha = .05$ *(t = -2.105,* d.f. = 5,638, *p =0.035*), leading us to reject the null hypothesis.

Prefixes not forming APs: $\underline{x} = 0.091$, std = 0.076, *s = 533*

Prefixes forming APs: $\underline{x} = 0.098$, std = 0.071, *s = 5,107*

Considering that we obtained negative statistics results, cosine similarities of the second group are statistically significantly higher than cosine similarities of the first group. This contradicts the claim made in RG that prefixes like в-, до-, над-, недо-, низ-, пре-, пред-, со-, де-, дис-, ре-

trigger more semantic alterations than other verbal PF prefixes. Figure 1 (section 8) illustrates the distribution of prefixes in respect to their similarity scores. In future work we are going to examine the effect size in more depth.

Lastly, we looked at frequencies of prefixed verbs and their semantic relatedness to the original unprefixed simple verbs. Results are presented in Figure 2 (section 8). We can clearly see from the plot that frequency is directly correlated with semantic similarity. More semantic differences are happening in groups of lower size. Thus, we can conclude that more frequent types of prefixation do not influence semantics of the original verb as considerably.

6. Discussion

Undoubtedly, many open questions remain in relation to verbal aspect in Russian. This research attempted an empirical computational analysis of contemporary Russian corpora and computed a number of productivity, valence, and similarity metrics. Most traditional Russian grammars label morphological processes as productive, semi-productive or unproductive without a clear definition of methods used for such analyses. We should clearly state that incorporating other parameters, such as rate of additions over a certain period of time (which fell beyond the scope of this research), would contribute even more towards the accuracy of such computation.

Another issue we are going to address in the future concerns certain irregularities of aspectual derivation. Special rules need to be added to account for changes including truncation, consonant mutation, root vowel alternations, and the basic rules of combination. We decided not to focus our attention on such changes at this stage, since they represent special cases and are not likely to interfere with the overall results of the methods proposed here.

Another goal for future study is to incorporate BERT contextual embeddings and compute the productivity of morphological processes that form IMPF from morphologically complex PF verbs (so-called secondary imperfectivization) with a special focus on more granular degrees of productivity and semantic relatedness.

Additionally, we would like to further investigate the effect size for every t-test we conducted. Furthermore, it would be beneficial for our analysis to examine the effect size separately for every verbal prefix.

7. Conclusions

We discussed several defining properties of inflectional and derivational morphology. Considering those properties in relation to a broad classification of PF prefixes into lexical (i.e., resulting in idiosyncratic, unpredictable meanings) vs. superlexical (i.e., purely perfectivizing), we can generally map lexical prefixes to the instances of derivational morphology (DM), and superlexical prefixes to inflectional morphology (IM). Table 11 summarizes assumptions from the literature and results of the experiments conducted in this research.

Table 11: DM & IM Feature Summary[21]

	DM	**IM**
Productivity	⫛	⫙
Frequency	⫛	⫙
Valence	⫙	⫛
Semantic Relatedness	⫙	⫛
Paradigm	⫙	⫛

Now, we can evaluate every verbal prefix according to the parameters from Table 11 and their characteristic tendencies of being comparatively higher or lower between DM and IM. This evaluation is shown in Table 12 (section 8).

[21] Upward and downward arrows stand for higher and lower indicators per group that are expected for each category.

Based on the analyses and experiments presented in this research, several intermediate conclusions can be drawn. The most essential one relates to the binary classification of verbs into aspectual pairs. Such an account cannot explain multiple existing phenomena, such as bi-aspectual verbs, verbs without aspectual correlates, or verbs with more than one aspectual correlate. Another important question to raise is how morphological productivity can be accurately measured. We observed certain patterns with respect to methods used for productivity computation and semantic similarities between simple IMPF verbs and their PF prefixed derivatives. The results obtained for both the argument structure and semantic relatedness showed that there are significant differences between simple verbs and prefixed PF verbs. Existing classifications of prefixes do not seem to be sufficient to account for semantic granularity that is triggered by both the aspectual change and the degree of variation within every unique prefix and verb combination. Finally, based on the results of the experiments, PF prefixes as a class are likely to be a part of derivational morphology based on their significant impact on argument structure, inconsistent productivity and frequency patterns, gaps in grammatical paradigm, and triggered changes in semantic relatedness.

8. Appendix: Extra Tables and Figures

Table 7: Prefix productivity (sorted by most productive)

Prefix	**0N**	**e**	**Dif**	**Result**
по-	209.47	33	176.47	productive
за-	182.27	56	126.27	productive
на-	123.49	46	77.49	productive
про-	114.7	40	74.7	productive
вы-	94.98	38	56.98	productive
о-	81.67	26	55.67	productive
от-	88.1	39	49.1	productive
рас-	58.05	19	39.05	productive
об-	65.23	31	34.23	productive
под-	75.84	43	32.84	productive
у-	74.89	46	28.89	productive
до-	63.27	35	28.27	productive
ис-	34.04	7	27.04	productive
пере-	115.98	90	25.98	productive
вс-	21.2	4	17.2	productive
из-	35.59	19	16.59	productive
раз-	51.87	39	12.87	productive
вз-	26.22	14	12.22	productive
над-	12.4	5	7.4	productive

недо-	17.01	12	5.01	productive
подо-	7.77	3	4.77	productive
изо-	4.59	0	4.59	productive
в-	36.98	34	2.98	productive
вос-	10.65	9	1.65	productive
во-	6.23	6	0.23	productive
воз-	13.54	17	-3.46	unproductive
при-	83.95	88	-4.05	unproductive
с-	89.56	94	-4.44	unproductive
со-	23.4	28	-4.6	unproductive
пре-	12.21	24	-11.79	unproductive
пред-	9.64	32	-22.36	unproductive

Table 9: Average similarity scores by prefix

Prefix	**Average Score**
вос-	0.16
пред-	0.16
со-	0.16
над-	0.14
по-	0.14
из-	0.13
пре-	0.13
вз-	0.12

вы-	0.12
на-	0.12
воз-	0.11
раз-	0.11
с-	0.11
вс-	0.10
под-	0.10
в-	0.09
от-	0.09
при-	0.09
про-	0.09
у-	0.09
де-	0.08
за-	0.08
рас-	0.08
пере-	0.07
до-	0.06
об-	0.06
ис-	0.05
недо-	0.05
о-	0.05
подо-	0.01
ре-	0.00

во-	-0.04
изо-	-0.07

Table 10: Highest cosine similarity scores

Unprefixed	**Prefixed**	**Score**	**Prefix**
гнаться	угнаться	0.89	у-
врать	соврать	0.73	со-
купиться	скупиться	0.79	с-
теряться	растеряться	0.66	рас-
будить	разбудить	0.68	раз-
браться	пробраться	0.71	про-
тащить	притащить	0.79	при-
полагать	предполагать	0.86	пред-
обретать	преобретать	0.75	пре-
двинуть	пододвинуть	0.53	подо-
разделяться	подразделяться	0.67	под-
думать	подумать	0.85	по-
думать	передумать	0.81	пере-
благодарить	отблагодарить	0.79	от-
думать	обдумать	0.57	об-
знакомить	ознакомить	0.7	о-
понимать	недопонимать	0.57	недо-
ломать	надломать	0.69	над-

гонять	нагонять	0.8	на-
портить	испортить	0.72	ис-
гнуться	изогнуться	0.25	изо-
брать	избрать	0.69	из-
говорить	заговорить	0.81	за-
пускать	допускать	0.65	до-
кинуть	выкинуть	0.77	вы-
крикнуть	вскрикнуть	0.62	вс-
хотеть	восхотеть	0.73	вос-
гордиться	возгордиться	0.65	воз-
брать	вобрать	0.29	во-
алкать	взалкать	0.72	вз-
вязаться	ввязаться	0.71	в-

Table 12: Prefix quantitative summary

Prefix	**Prod**	**Frequency**	**Similarity**
вос-	1.65	1.83%	0.16
пред-	-22.3	1.13%	0.16
со-	-4.6	0.19%	0.16
над-	7.4	0.57%	0.14
по-	176.47	0.38%	0.14
из-	16.59	0.81%	0.13
пре-	-11.7	5.17%	0.13

вз-	12.22	3.28%	0.12
вы-	56.98	10.90%	0.12
на-	77.49	1.64%	0.12
воз-	-3.46	0.09%	0.11
раз-	12.87	1.46%	0.11
с-	-4.44	7.00%	0.11
вс-	17.2	0.42%	0.10
под-	32.84	0.68%	0.10
в-	2.98	4.27%	0.09
от-	49.1	3.36%	0.09
при-	-4.05	4.76%	0.09
про-	74.7	6.88%	0.09
у-	28.89	12.54%	0.09
за-	126.27	4.07%	0.08
рас-	39.05	0.22%	0.08
пере-	25.98	0.57%	0.07
до-	28.27	0.53%	0.06
об-	34.23	4.91%	0.06
ис-	27.04	6.40%	0.05
недо-	5.01	2.67%	0.05
о-	55.67	2.85%	0.05
подо-	4.77	5.29%	0.01
во-	0.23	1.10%	-0.04

изо-	4.59	4.04%	-0.07

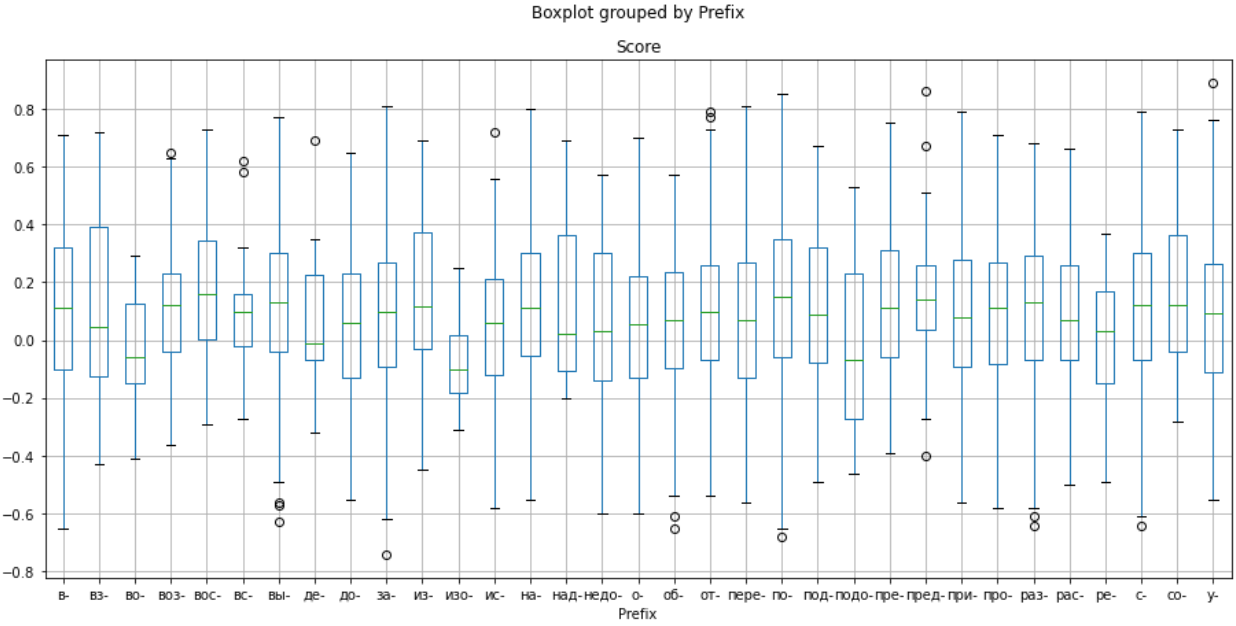

Figure 1: Distribution of Prefixes According to Cosine Similarity Scores

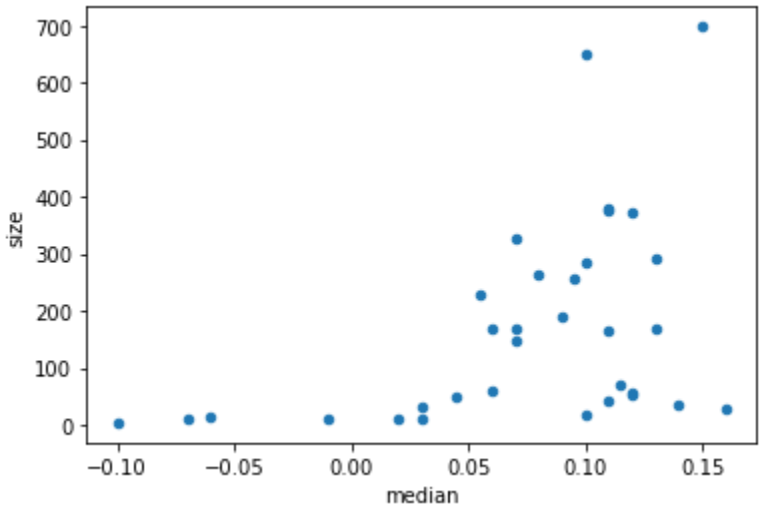

Figure 2: Prefix frequency (size) vs. cosine similarity (median)

9. References

Anderson, C. (2002). Biaspectual Verbs in Russian and their Implications on the Category of Aspect. Honors Thesis, University of North Carolina-Chapel Hill.

Aronoff, M. (1976). *Word Formation in Generative Grammar*. MIT Press.

Baayen, H. (1992). Quantitative Aspects of Morphological Productivity. In Geert Booij and Jaap van Marle (eds), *Yearbook of Morphology*. Kluwer, (109–149).

Babko-Malaya, O. (1999). Resultatives and Zero Morphology. In Shahin, Kimary, N., Blake, S., Eun-Sook, K. (eds.), *Proceedings of the Seventeenth West Coast Conference on Formal Linguistics.* CSLI Publications.

Babko-Malaya, O. (2003). Perfectivity and Prefixation in Russian. *Journal of Slavic Linguistics*, 11/1, (5-36).

Bailyn, J. F. (2012). *The Syntax of Russian*. Cambridge University Press.

Bauer, L. (2001). *Morphological Productivity*. Cambridge University Press.

Di Sciullo, A. M., Williams, E. (1987). *On the Definition of Word.* MIT Press.

Filip, H. (2005). On accumulating and having it all. Perfectivity, prefixes and bare arguments. In Verkuyl, H.; de Swart, H. & van Hout, A. (eds.), *Perspectives on Aspect.* Springer, (125-148).

Flier, M. S., Timberlake, A. (1988). *The scope of Slavic aspect. Russian Linguistics*, 12, (303-306).

Grønn, A. (2003). The Semantics and Pragmatics of the Russian Factual Imperfective. PhD Thesis. University of Oslo.

Guiraud-Weber M., Zalizniak, A. A., Šmelev, A.D. (1999). Lectures on Russian Aspect Studies. *Russian Linguistics*, 23, (297-301).

Hockett, C. F. (1954). Two Models of Grammatical Description. *WORD*, 10/3, (210-234).

Hockett, C. F. (1958). *A Course in Modern Linguistics.* Macmillan.

Isačenko, A. V. (1960). Grammatičeskij stroj russkogo jazyka v sopostavlenii s slovackim – Čast' vtoraja: morfologija. *Izdatelstvo akademii nauk.*

Isačenko, A. V. (1962). *Die russische Sprache der Gegenwart*. Halle.

Janda, L. A. (2007). Aspectual clusters of Russian verbs. *Studies in Language*, 31/3, (607-648).

Klein, W. (1995). A time-relational analysis of Russian aspect. *Language*, 71/4, (669-695).

Kagan, O. (2012). Degree Semantics for Russian Verbal Prefixes: The Case of pod- and do-. *Oslo Studies in Language*, 4/1.

Levin, M. I. (1978). *Russian Declension and Conjugation. A Structural Description with Exercises*. Slavica Publishers.

Manova, S. (2005). Derivation versus Inflection in Three Inflecting Languages. In W. Dressler, D. Kastovsky, O. Pfeiffer, F. Rainer (eds.), *Morphology and its Demarcations.* (233-252).

Maslov, Ju. S. (1959). Glagol'nyj vid v sovremennom bolgarskom jazyke. In S. B. Bernštejn (ed.), *Voprosy grammatiki bolgarskogo literaturnogo jazyka*. (157-312).

O'Donnell, T. (2011*).* Productivity and Reuse in Language. PhD Thesis. Harvard University.

O'Donnell, T. (2015). *Productivity and Reuse in Language*: *A Theory of Linguistic Computation and Storage*. MIT Press.

Padučeva, E. V. (1996). Sematičeskie issledovanija. Semantika vremeni i vida v russkom jazyke. Semantika narrativa. *Jazyki russkoj kultury*.

Paslawska, A., von Stechow, A. (2003). Perfect Readings in Russian. In Alexiadou, Artemis, Rathert, Monica & von Stechow, Arnim (eds.). *Perfect Explorations*. Mouton de Gruyter, (307-362).

Pinker, S., Prince, A. (1991). Regular and Irregular Morphology and the Psychological Status of

Rules of Grammar. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society.* 17, (230-251).

Romanova, E. (2004). Superlexical vs. Lexical Prefixes. *Norlyd*, 32/2, (255-278).

Spencer, A. (1991). *Morphological Theory. An Introduction to Word Structure in Generative Grammar*. Wiley.

Tatevosov, S. (2015). Severing imperfectivity from the verb. In Zybatow, Gerhild; Biskup, Petr; Guhl, Marcel; Hurtig, Claudia; Mueller-Reichau, Olav & Yastrebova, Maria (eds.). *Slavic Grammar from a Formal Perspective*, (465-494).

Šmelev, A., Zalizniak, A. (2000). *Vvedenie v russkuju aspektologiju*. Jazyki russkoj kul'tury.

Šmelev, A., Zalizniak, A. (2006). Aspect, modality, and closely-related categories in Russian.

Švedova (ed.) (1980). *Academy of Russian Grammar, Volume I.* The Academy of Sciences of USSR.

Weznek, G., Lachaux M., Conneau A., Chaudhary, V., Guzman, F., Joulin, A., Grave, E. (2019). CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. *Proceedings of the 12th Language Resources and Evaluation Conference (LREC),* (4003-4012).

Yang (2018). *A User's Guide to the Tolerance Principle*. Ms., University of Pennsylvania.

Zalizniak, A. (1980). *The Grammatic Dictionary of the Russian Language.* Moscow: Russian Language Publishers.