Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2023»

14–16 июня 2023 г.

RUSSIAN PREDICATIVES AND FREQUENCY METRICS

Anton Zimmerling
Pushkin State Russian Language
Institute / Institute of Linguistics,
Russian Academy of Science
fagraey64@hotmail.com

Abstract

This paper introduces five metrics for measuring the frequencies of dative predicatives in Russian.A dative predicative is a word or multiword expression licensing the dative-predicative-structure, where the semantic subject of the non-agreeing non-verbal predicate is marked by the dative case. I measure the frequencies of the predicatives in the contact position <-1;1> with the same-clause dative subject pronouns in 1Sg (*m*-metrics) and 3Sg (*e*-metrics). The *m*-metrics is applied for retrieving a list of dative predicatives from a corpus. I argue that for each large text collection there is a minimal *m*-value confirming that an item belongs to the core of the dative-predicative structure. The *m/e* score makes up the third metrics that shows whether an element is oriented towards the use in the 1st person or not. Basing on the *m*-metrics, I retrieved 3 lists of predicatives in the subcorpus of 2000–2021 texts included in the Russian National Corpus. The A list includes 87 items with $m \geq 10$, the B list includes 44 items with $m \geq 50$, the C list includes 24 items with $m \geq 100$. 72-79% of items in each list have an *m/e* value ≥ 1,25. A linguistic interpretation of this result is that for each list of dative predicatives it is true that the majority of its elements are autoreferential expressions oriented towards the use in the 1st person present indicative tense in the direct speech. The fourth metrics shows the total number of occurrences of a word or multiword expression in the corpus (*N*). I argue that the *N* score must be measured before POS tagging, and lemmatization. The fifth and the last metrics is the *m/N* score. The RNC data suggest an inverse correlation between the score of an item in the context specific for dative-predicative structures (*m*) and its overall frequency in the corpus (*N*). This effect is explained by the regular homonymy of high frequent predicatives with high frequent adverbials and parenthetical expressions.

Keywords: corpus grammar, frequency dictionaries, lexicon, dative predicatives
DOI: 10.28995/2075-7182-2023-22-579-589

РУССКИЕ ПРЕДИКАТИВЫ В ЗЕРКАЛЕ СТАТИСТИКИ

Циммерлинг А. В.
Государственный институт русского
языка им. А.С.Пушкина / Институт
языкознания РАН
fagraey64@hotmail.com

Аннотация

В статье предлагается пять метрик для создания частотного словаря дативных предикативов в русском языке. Дативный предикатив определяется как элемент, допускающий дативно-предикативную структуру, где семантический субъект несогласуемого неглагольного предиката оформляется дат.п. Ранжирование предикативов производится по числу предложений дативно-предикативной структуры в выборке по запросу предикатив + субъектное местоимение 1 л.ед.ч. *мне* в контактной позиции на расстоянии <-1;1> (*m*-метрика) и предикатив + субъектные местоимения 3л. ед.ч. *ему/ей* в той же позиции (*e*-метрика). Словарь предикативов строится на основе *m*-метрики. Для каждой большой коллекции текстов имеется минимальное значение *m*, подтверждающее, что данный элемент принадлежит ядру класса дативных предикативов. Отношение *m/e* используется как третья метрика. Она указывает на то, ориентирован ли элемент на употребление в 1л. в

режиме речи. С помощью *m*-метрики было получено три списка в подкорпусе текстов 2000 – 2021 гг. в НКРЯ. Список А содержит 86 единиц с $m \geq 10$, список В — 44 единицы с $m \geq 50$, список С — 24 единицы с $m \geq 100$. 72-79% элементов каждого списка имеют значение $m/e \geq 1{,}25$. Этот результат подтверждает, что большинство элементов каждого списка ориентированы на употребление в 1 л. ед.ч. презенса индикатива в прямой речи. Четвертая метрика указывает общее число вхождений слова или словосочетания в корпус (*N*). Значение *N* подсчитывается до лемматизации и определения части речи. Отношение *m/N* является пятой метрикой. Данные НКРЯ указывают на обратную зависимость между числом употреблений в контексте, характерном для дативно-предикативной конструкции (*m*), и общим числом вхождений в корпус (*N*). Этот эффект объясняется тем, что наиболее частотные предикативы связаны отношениями регулярной омонимии с высокочастотными наречиями и вводными словами.

Ключевые слова: корпусная грамматика, словарь, дативные предикативы, конструкции

1. Introduction

I discuss the procedure of measuring the frequencies of a productive grammatical construction the elements of which do not make a single lexical class but represent special predicative uses of words from different parts of speech and multiword expressions linked with syntactic structures imposing non-trivial conditions on agreement and case-marking.

The baseline hypothesis is that the majority of Russian predicatives with the dative case-marking on the subject argument are autoreferential expressions including a link to the speaker, who is the source of information about the internal state experienced by him/her at the moment of speech. The aims of the study is to check this hypothesis and to establish, whether the autoreferentiality effects arise due to the inherent lexical features of Russian dative predicatives or are modeled in syntax.

2. Dative-predicative structures and their diagnostics

Russian has a productive class of predicatives licensing syntactic structures, where the animate semantic subject of a non-agreeing non-verbal element is marked with the dative case, hence — dative-predicative structures (DPS). The relation between DPS sentences and word classes is a puzzle. One the one hand, Russian grammar does not require that the dative slot of any predicative or verb is realized overtly. One the other hand, occasional combinations of a predicative with the dative argument do not prove that it is part of the DPS lexicon. The lexicon of a grammatical construction is a list of lexical items regularly used in this construction by all or most speakers. However, with Russian DPS predicatives one must measure the frequencies of the sentences with a filled dative slot, cf. *X-у было стыдно признавать ошибку* ‘X was ashamed to admit his/her mistake’, not just the hits of the lemma *стыдно* or the collocation *стыдно признавать* ‘ashamed to admit smth’. The word *стыдно* in contrast to *грустно* ‘sad’, ‘sadly’, *холодно* ‘cold’, ‘coldly’ belongs to the minority of predicatives that lack side-uses as adverbials. The preceding research provides no instructions how to get the ratio of the relevant DPS uses from the total number of hits of items like *стыдно* or *грустно*. Some DPS predicatives are idiomatic multiword expressions, cf. *X-у все равно* ‘X does not care’.

2.1. The syntax

The role of the dative element can be explained differently. According to [9: 151], most types of Russian sentences can be expanded by the position of the animate dative participant. On this account, it is a free ‘determinant’ or in conventional terms, adjunct, therefore the dative slot does not constrain any class of predicates. This prediction is wrong, since the DPS construction is selective and blocks the combinations that cannot be interpreted as standard designations of internal states experienced by an animate subject. Although Russian authors sporadically produce weird sentences like ??*Нам гневно делается* (Anthony of Sourozh, 1992) ‘we get angry’, lit. *‘to us becomes wrathfully’, ??*Морозно мне* (M.Ancharov, 1989) ‘I feel freezingly cold’, lit. *‘to me is chilly’, they are rejected by the majority of speakers according to [14] and have low frequency in text corpora[1]. Under the alternative ap-

[1] Note that *морозно* and *гневно* are equally marginal as DPS items, although *морозно* ‘It is frosty’, ‘It is chilly

proach, the dative element is semantic subject and the class of DPS predicatives consists of elements capable of describing internal states [8]. This analysis predicts that dative arguments switch the lexical meaning of the predicatives. This is likely for the physical sensations, cf. *Сегодня холодно* 'It is cold today' ⇒ *Мне холодно* 'I am cold', *здесь темно* 'It is dark here', *Мне темно здесь* ⇒ 'It is dark for me'. Without the dative argument *холодно* or *темно* normally describe ambient characteristics, while with the filled dative slot they describe the reactions of an experiential subject, cf. [5; 6]. With the predicatives of interpretation, which do not describe the sensations or affections directly but interpret them in some way, cf. *важно* 'important' the switch is less evident, cf. ***(Мне)*** *важно закончить работу сегодня* 'It is important (**for me**) to finish the work today'. If DPS predicatives make up a lexical class, one needs a list of non-verbal non-agreeing elements with a valency on the animate dative argument [2: 83]. However such lists can only be retrieved in the experiment or corpus study, where approval rates or frequency scores are measured.

2.2. Autoreferentiality

DPS sentences express the meaning of internal davidsonian states[2], i.e. spatiotemporal situations with an animate priority argument [10; 11: 273][3]. This meaning is not unique for Russian DPS sentences, cf. [13: 424-431]. However, the dative case-marking adds a special quality: DPS items are oriented towards the use in the 1Sg in the direct speech, while other types of Russian predicatives sharing the taxonomic meaning of davidsonian states with them normally cannot be used in this context. While it is standard to say *мне*$_{\text{DAT}}$ *грустно* 'I am sad', *мне*$_{\text{DAT}}$ *дурно* 'I feel bad' sentences like $^{??}$*я*$_{\text{NOM}}$ *сейчас навеселе*, int. 'I am tipsy now', **я*$_{\text{NOM}}$ *без чувств*, int. 'I am losing my senses', 'I faint' are awkward. A plausible explanation of this asymmetry is that the majority of Russian DPS predicatives are autoreferential expressions: the speaker himself/herself is the source of information about his/her internal state of feeling bad or sad in the interval including the moment of speech [18]. Meanwhile, Russian predicatives with nominative case-marking on the subject, cf. *навеселе*, *без чувств* are oriented towards describing the experience of other people. The autoreferentiality effect gives a clue for retrieving dative predicatives from a corpus. DPS sentences are copular structures with a slot for the BE-auxiliary or less frequent auxiliaries like *стать*, *сделаться* 'become'. The contact position of a predicative and the 1Sg subject dative pronoun *мне* roughly corresponds to the context of the present indicative, where the overt BE-auxiliary is missing in Russian. Although the search queries PRED + "мне" in the contact position <-1; 1> do not exclude the examples, where an overt auxiliary is found to the left or the right from the search window, cf. ***было***$_{\textbf{AUX.PST}}$ *<мне грустно>* ~ *<грустно мне>* ***было***$_{\textbf{AUX.PST}}$ 'I was sad', the preceding research indicates that the majority of hits retrieved by such queries indeed patterns with autoreferential contexts in the present indicative tense [16].

2.3. The lexicon

The DPS construction is characteristic of several European languages. The volume of the class of DPS predicatives was measured via a double sociolinguistic and corpus study for Russian [14] and Bulgarian [15]. These authors checked a set of 422 stimuli for Russian. They argue that most Russian

---

outdoors' is a standard impersonal predicative describing the state of weather. The Russian National Corpus (RNC) totals 2143 hits of *гневно*, 2135 of which represent the uses as a non-predicative adverbial and just 8 (0,38%) pattern with agreeing adjectives or predicatives. From 497 hits of *морозно,* 439 (88,4%) pattern with impersonal predicatives.

[2] The cover term *состояния* 'states' used in the Russian studies, is vague. The term 'davidsonian states' is a tribute to Donald Davidson, who defined states as static spatiotemporal situations that exist during a time interval [3]. Internal <davidsonian> states have a priority experiential argument [12; 13: 429 - 431].

[3] In Davidson's account, spatiotemporality is a definitional property: it is assumed that every process and every external or internal state, cf. *The sun is rising. X is in London. X is sad* takes place in some locus, irrespective of the fact, whether the predicate combines with a locative phrase or framing adverbial. An anonymous reviewer suggests that Russian sentences like *Я видел, как ему жаль птичку (*в темной комнате)* should be described as Kimian states, i.e. predicates lacking spatial features [7]. However, *X-у жаль птичку* 'X feels sorry for the bird' describes the feeling of *X* that holds during some time and not the result of *Y*-s observation. Moreover, internal states, e.g., the feeling of being sad, happy, sorry, etc. cannot be observed from outside, though *Y* via some kind of practical reasoning can reconstruct the situation, where X is sad or happy, basing on the external symptoms of sadness or happiness.

speakers have over 200 DPS predicatives in their active vocabulary, but only one part of it is shared. In the variable part, Russian speakers typically select quasi-synonymic DPS items corresponding to generalized lexical meanings like 'X does not care', 'X is delighted', 'X is disgusted', etc. The same test of stimuli was checked on RNC. The search was restricted with one dedicated context — the contact position of the predicative and the 1Sg dative subject pronoun *мне* in the window <-1;1>. The retrieved samples proved large enough to range 400 – 500 items. The authors conclude that high frequent DPS items always have a high approval rate, while DPS items with a high approval rate generally are high frequent, with the exception of some predicatives describing ontologically rare situations, cf. *X-у по колено* 'X is up to his knees', *X-у было по щиколотку* 'X was up to his ankles'. This effect was presumably due to the design of the experiment: the speakers had no difficulties with reconstructing the situations, where such DPS items were appropriate, but the corresponding contexts in the RNC were rare.

I adopt the method of retrieving DPS sentences by narrowing the search with the 1st person contexts and introduce several new metrics for ranging DPS predicatives. In order to eliminate the diachronic factor and make the input data homogeneous, I focus on 2000 – 2021 texts included in the RNC[4]. I also measure the scores of negative and non-negative DPS items on a separate basis and make other adjustments in the set of stimuli. The DPS lexicon in [13; 16: 248] was grouped into 15 thematic classes labeled 'physical sensations'(Class 1), 'modalities' (Class 2), 'affections' (Class 3), 'moral attitudes' (Class 4), '(in)convenience' (Class 5), '(im)pertinence' (Class 6), 'internal need' (Class 7), 'complience' (Class 8), 'difficulty of execution' (Class 9), '(in)disposition' (Class 10), 'general evaluations' (Class 11), '(ir)relevance' (Class 12), '(in)efficiency' (Class 13), 'sensory and intellectual responses' (Class 14), 'parametric features' (Class 15). I adopt this classification and add new items, where appropriate.

3. The frequency dictionary of Russian DPS predicatives

3.1. M-metrics

The lists of DPS predicatives are built by *m*-metrics, which tells the number of confirmed DPS clauses in the syntactic corpus assembled by the query "STIMULUS" + "*мне*" in the window <-1; 1>. The stimulus must be identified as a DPS predicative and the dative pronoun must be the same clause element acting as its semantic subject. The DPS sentences are copular structures that bring about several formal conditions, notably the absence of agreement and the nominative NP that could act as agreement controllers, see below **3.2**.

I take the list of DPS stimuli in [14; 17: 254-255] and adjust it to the tasks of present study. The set of 478 stimuli checked in the 2017 experiment included fillers and obsolete words that went into disuse in the second half of the XX century or earlier. I eliminate all low frequent items from the 2017 set and check the upper part of stimuli starting with $m \geq 10$. The main RNC corpus had 159 such items in 2017. The 2000 – 2021 corpus is smaller. Setting the lower limit at $m \geq 10$, we retrieved 87 DPS predicatives. By lifting the limit up to $m \geq 44$, we get a second list containing 44 DPS items. Setting the limit at $m \geq 100$ leaves us with 24 most frequent DPS items. These lists are referred to as A87, B44 and C24. The maximal *m* score is attested by НАДО ($m = 1402$). The syntactic corpus linked with A87 contains 9619 DPS sentences[5]. The mean expected score m_{87} is 9619/87 = 110, 56. The syntactic corpus linked with the shortest list, C24 contains 7322 DPS sentences. That means that the 24 most frequent DPS predicatives (27, 6%) give 76,1% of DPS sentences.

3.2. The stimuli

The combinations with the free negation *не* were treated as separate entries, if the non-negative expression is used as a DPS predicative: the examples with НЕ НАДО, НЕ НУЖНО, etc. were subtracted from the samples with НАДО, НУЖНО. We considered all spelling variants like НЕ ВАЖНО

[4] 43 928 texts, 98 023 229 words (11.2022).

[5] The requirement that the predicative and its subject are realized overtly and assume a contact position makes each sentence in the syntactic corpus unique. The duplication across samples is excluded. The duplication within a sample is only possible if the RNC search engine returns one and the same text fragment twice.

~ НЕВАЖНО. The A87 list contains 20 items with negation, the most frequent of them being НЕ НАДО (m=334), НЕ НУЖНО (125) and НЕ ЖАЛКО (64). Comparative forms were treated as separate entries, cf. ЛУЧШЕ (*m*=121), ЛЕГЧЕ (89), and ПРОЩЕ (53). The samples with the spelling variants –ЕЕ/-ЕЙ were merged, cf. ИНТЕРЕСН-ЕЕ/-ЕЙ (18). The optative combination ХОРОШО БЫ 'It would be nice' (10) was considered a separate entry different from ХОРОШО 'good' (176). The corresponding examples were subtracted from the scores of the positive forms.

The A87 list includes 12 multiword expressions, 5 of them are also contained in B44 and the upper 3 — in C24, cf. ВСЕ РАВНО (312), НЕ ДО Z-а (60), БЕЗ РАЗНИЦЫ (19), ТАК И НАДО (19) , НЕ ПО СЕБЕ (15), and НЕ ПОД СИЛУ (10). The idioms ВСЕ РАВНО 'X does not care' and ТАК И НАДО 'X deserved it' are treated as separate entries; the score of ТАК И НАДО is subtracted from the score of НАДО. The insertion of the subject dative pronoun into the idiom ТАК *мне* И НАДО was considered an idiosyncratic option equivalent to the contact position of the dative pronoun: otherwise this idiom should be excluded.

No filters were applied to sort out gross expressions. The colloquial words ПОФИГ (*m*=16) and ПО ФИГУ ~ ПОФИГУ (18) were considered separate entries. I substituted the predicate variable in the idiom *X-у* ***Z-ать*** *на Y-а* 'X **does not care** about Y' with the infinitives of physiological verbs: ПЛЕВАТЬ (*m*=135), НАПЛЕВАТЬ (76) и НАСРАТЬ (17) made it to the A87 list.

3.3. Syntactic disambiguation and nominative expressions

Russian DPS sentences are usually analyzed as structures blocking NPs in the nominative case both in the subject [8] and in the object position [15]. A different approach is outlined in [1: 305-308]. Non-adjectival predicates like *X-у не под силу* 'it is beyond X's reach' are an issue, since they license both DPS sentences, cf. *X-у не под силу решить эти задачи* 'To solve these tasks is beyond X's reach' and dative-nominative structures like *X-у* ***эти***$_{\text{NOM}}$ ***задачи***$_{\text{NOM}}$ *не под силу* '**These tasks** are beyond X's reach'. I adopt the mainstream approach and exclude the sentences with a nominative subject from the syntactic DPS corpus. This decision only has a minor effect on A87, since dative-nominative structures are infrequent in the samples derived by the *m*-metrics.

The sentences with a dative pronoun and a noun/NP from the class *лицо* 'face', *признание* 'confession' in the nominative-accusative are two-way ambiguous. If the nominative analysis is taken, the ambiguous predicate head is recognized as an agreeing short adjective in the neutrum singular form, cf. (1a-b). If the accusative analysis is taken, the predicate is recognized as a DPS item, cf. (2a-b).

(1) a. мне плохо видно$_{\text{ADJ.NOM.SG}}$ ее **лицо**$_{\text{NOM.SG.N}}$.
'I can't see her face clearly', lit. 'Her face is badly visible to me.'

b. Мне плохо видна$_{\text{ADJ.NOM.F}}$ ее **шея**$_{\text{NOM.SG.F}}$.
'I can't see her neck clearly', lit. 'Her neck is badly visible to me.'

(2) a.Мне плохо видно$_{\text{PRED}}$ ее **лицо**$_{\text{ACC.SG.N}}$.
'I can't see her face clearly.'

b. Мне плохо видно$_{\text{PRED}}$ их **лица**$_{\text{ACC.PL}}$.
'I can't see their faces clearly.'

Another kind of ambiguity is caused by the pronominal expressions *это* 'this', *все это* 'all this'. If they fill in the valency of an active or passive verb, they must be considered referential pronouns/DPs in the accusative or nominative case, cf. (3a). If they lack strong referential properties and refer to the situation as a whole without referring to any of its parts, they are caseless expressions that do not take the subject or object positions, cf. (4a).

(3) a. **Все это**$_{\text{NOM.SG.N}}$ мне куплено$_{\text{PRT.PASS.NOM.SG.N}}$
'All this has been bought for me.'

b. **Все эти вещи**$_{\text{NOM.PL}}$ мне куплены$_{\text{PRT.PASS.NOM.PL}}$.
All these things have been bought for me.'

(4) a. **Все это** мне грустно$_{\text{PRED}}$.
‘All this is sad to me’,

b. *Все эти вещи мне грустны.
int. * ‘All these things are sad for me.’

3.4. E-metrics

The same set of 87 stimuli was checked with the dative pronouns *ему* ‘3Sg.Dat.M’ and *ей* ‘3Sg.Dat.F’ in the contact position in the window <-1; 1>. The number of the confirmed DPS clauses is called *e*-metrics. The *e*-metrics provides a tool for checking autoreferentiality. The syntactic corpus built via the *e*-metrics for A87 contains 5434 DPS sentences and is ca. 1,8 times smaller compared to the corpus assembled by the *m*-metrics. The mean expected value e_{87} is 5434/87 = 61, 31. Another index showing the frequency drop in the *e*-corpus is the number of the DPS items fitting to the minimal values for C24, B44 and A87: there are only 11 predicatives in the C*11 list ($e \geq 100$), 31 predicatives in the B*31 list ($e \geq 50$) and 68 predicatives in the A*68 list ($e \geq 10$). The shrinking is most pronounced with high frequent DPS items, where C*11 exports 10 DPS items from C24 and lifts one item from B44, ДОСТАТОЧНО ($m = 79$, $e = 101$). All B*31 items, with the exception of УДОБНО$_1$ ($m = 34$, $e = 80$) are contained in B44 and all A*68 items are contained in A87. The last result is trivial, since A87 per definition lacks items with $m < 10$. The first two ones are not: they show that just 2 DPS items from 87 swap their positions in the mid-range and high-range lists.

3.5. Thematic classes

The thematic classes of the DPS lexicon are distributed evenly in our data. The largest list, A87 includes 12 classes from 15, only Classes 7 <‘internal need’>, 10 < ‘(in)disposition’> and 13 < ‘(in)efficiency’> are missing, since they lack frequent DPS predicatives with $m \geq 10$. B44 also lacks Classes 8 <complience> and 15 <‘parametric features’>. The shortest list, C24 retains 8 different classes but drops Classes 5 < ‘(in)convenience’> and 6 <‘<im>pertinence’ >.

Tab. 1. The coverage of the DPS construction in Russian (2000 – 2021).

List	m	Retained classes	Missing classes
A87	≥ 10	1, 2, 3, 4, 5, 6, 8, 9, 11, 12, 14, 15	*7, *10, *13
B44	≥ 50	1, 2, 3, 4, 5, 6, 9, 11, 14	*7, *10, *13, *, *8, *15
C24	≥ 100	1, 2, 3, 4, 9, 11, 12, 14	*7, *10, *13, *15,*8, *15, *5, *6

These figures confirm that Modern Russian has high frequent DPS predicatives in most thematic classes and uses them in diverse ontological situations.

3.6. Semantic disambiguation

A87 includes a pair of DPS items that are treated as homonyms, since they represent different thematic classes: *X-y* ПЛОХО$_1$ (Class 1, $m = 49$), cf. *Мне внезапно стало плохо* ‘I suddenly felt badly’ vs *X-y* ПЛОХО$_2$ (Class 11, $m= 149$), cf. *Ей было плохо жить со свекровью* ‘It was bad for her to live with her mother-in-law’. Their profiles can only be kept apart after semantic disambiguation. ПЛОХО$_2$, is also part of B44 and C24. Semantic disambiguation is relevant for *X-y* УДОБНО$_1$ (Class 5, $m = 34$), cf. *Я попыталась лечь, как мне удобно* ‘I tried to lie down as comfortably as I could’, НЕУДОБНО$_2$ (Class 4, $m = 40$), cf. *Неудобно мне как-то стало* ‘I felt kind of awkward’, НЕЛОВКО$_2$ ‘Class 4, $m = 39$’, cf. *Мне неловко об этом писать* ‘I am embarrassed to write about this’, where the homonymic predicatives are low frequent elements that do not make it to A87. The items (X-y) М*А*ЛО ‘X does not have enough’ ($m = 51$, $e= 96$) and (X-y) МАЛ*О* ‘Something is too small for X’ are pronounced differently but spelled in the same way, therefore the samples with М*А*ЛО must be checked for the casual hits of МАЛ*О*.

3.7. The m/e metrics and its application

The *m/e* score serves as the third metrics. It is applied after the lists of frequent DPS items are retrieved by the *m*-metrics. With low *m* scores > 10 and comparably low *e* scores, the fluctuations of the *m/e* score are not significant. With high and mid-frequent DPS items, it makes sense to measure both the individual profiles of DPS predicative and the general characteristics of the lists. Let us assume that a DPS predicative is autoreferential, if *m/e* ≥ 1, 25, i.e. if the uses in the 1st person singular are at least 25% more frequent compared to the uses of the 3rd person singular in the same position. The mean expected score for the A87 list m_{87}/e_{87} = 1,79 exceeds this level with a margin, but it is difficult to interpret this result without ranging the elements of each list on the basis of their individual *m/e* scores. Let us introduce a distinction of mildly non-autoreferential vs strictly non-autoreferential expressions. A DPS predicative is mildly non-autoreferential, if 1 ≤ *m/e* < 1, 25 and strictly non-autoreferential, if *m/e* < 1.

Tab. 2. Autoreferential DPS items in the Russian National Corpus (2000 – 2021).

	m/e	A87, *m* ≤ 10	B44, *m* ≤ 50	C24, *m* ≤ 100
+ Autoreferential	***m/e* ≥ 1,25**	**71,27%**	**72, 728%**	**79, 17%**
Mildly-non-autoreferential	1, 0 ≤ *m/e* < 1, 25	12,64%	13,636%	12,5%
-Autoreferential	***m/e* < 1**	**16,09%**	**13, 636%**	**8, 33%**

Tab. 2 shows that the share of the autoreferential DPS items increases with their frequency. More precisely, the C24 list containing the items with *m* ≤ 100 has just 2 strictly non-autoreferential items, ЛУЧШЕ (*m/e* = 0,85) and НЕОБХОДИМО (*m/e* = 0,87) and 19 autoreferential items (79,17%). Meanwhile, there is no contrast between A87 and B44: lifting the low *m* value from 10 to 50 leaves the percentage of the autoreferential items at the same level (71,3% — 72,7%). The *m/e* scores in A87 are in the range 0, 4 ≤ *m/e* ≤ 21. It makes sense to exclude the low frequent elements to get a more balanced picture[6].

3.8. The N-metrics and lemmatization

The *N*-metrics gives the number of hits of a word or multiword expression in a corpus. I argue that the *N* score must be measured before POS tagging and lemmatization. Almost all DPS items have regular homonyms predicted by their morphology. The largest group of homonyms is adjectival words with the – *o*-final, historically — short adjectives in Nom-Acc.Sg.N. Many of them, cf. *грустно* 'sad', 'sadly' are used in parallel as agreeing adjectives, adverbials and non-agreeing predicatives. Some items have a fourth side-use as parenthetical elements, cf. *видно* 'it is seen' ∨ or 'visible' ∨ 'apparently'. An *–o-* item can be tagged either as adverbial (ГРУСТНО$_{ADV}$) or as part of the adjectival paradigm (ГРУСТНО$_{ADJ}$). The latter decision depends on two factors: a) the existence of the adjectival lemma in the dictionary and/or the instruction confirming that the ГРУСТНО$_{ADJ}$ is used in the agreeing position; b) the (in)ability of the parser to recognize the agreement controller. The RNC parser occasionally fails to lemmatize *–o*-items correctly. I provide two illustrations. In (5) the parser failed to recognize the substantivized form *смешное* 'funny', 'what is funny' as the agreement controller and wrongly tagged *грустно* as an adverbial. In (6) the parser wrongly analyzed the non-argument expression *все это* 'all this' as an agreeing subject and tagged *грустно* as an adjective.

(5) Печальное$_{ADJ.SG.N}$ нам смешно$_{ADJ.SG,N}$, а **смешное$_{ADJ.SG.N}$** грустно$_{ADJ.SG.N}$ (A.Morozov, 1985-2001).
'What is sad is funny to us, and **what is funny** is sad.'

(6) Как-то грустно$_{PRED}$ мне$_{1SG.DAT}$ **все это** (A.Terekhov, 1997 – 2001)
'Somehow I feel sad about **all this**.'

[6] E.g., ДУРНО 'X feels badly' occurs in the 2000 – 2021 texts only 397 times but provides 15 autoreferential contexts (*m*=15) without a single example with the 3rd person singular subject pronoun in the contact position.

The deep syntactic annotation of DPS predicatives in the contact position with the subject dative pronoun makes the lemmatization of the *-o*-items in the remaining part of the corpus redundant. What matters is not the POS tags and lemmas of the elements homonymic to the DPS predicatives, but the share of the DPS hits in the sample derived by the *m*-metrics vs the raw data containing the total score of hits for the whole set of homonyms including the tested DPS item. RNC provides the ipm estimates for all words and collocations, but splits the data into different lemmas. This is unhappy with comparative forms. E.g., the search item *хуже* 'worse' returns back the lemmas ПЛОХОЙ, ПЛОХО, ХУЖЕ and even ХОРОШО (the antonym of ПЛОХО). The search item *лучше* 'better' returns back 7 lemmas, including exotic suggestions like ВСЕМИЛОСТИВШЕ (the second frequent lemma!). Similar issues arise in all cases, where the spelling varies.

3.9. The m/N metrics

The *m/N* score is the fifth metrics. It shows the proportion of the confirmed DPS hits in the syntactic sub-corpus built via the *m*-metrics vs the total score of all elements identic with or homonymic to the corresponding DPS predicative. I call this set 'quasi-homonymic list'. It is irrelevant for the *m/N* score whether the elements of this list are real homonyms, as, e.g. in the pair НАДО$_1$ 'necessary' vs НАДО$_2$ 'above', diverged uses of the same underlying morphological form, cf. *грустно* 'sad', 'sadly' or DPS uses outside the *m* context. A pair or tuple of quasi-homonymic lists is called 'quasi-homonymic hyperset'.

I checked two hypotheses: A) The number of DPS hits in the 1st person contexts feeds on the score of quasi-homonyms and increases proportionally; B) some elements are more specialized in the DPS construction than other elements. The hypothesis A) makes wrong predictions. The situation at the poles of the *N* scale resembles the inverse correlation between *N* and the *m/N* score. The highest frequent element, МОЖНО (N = 121490) has one of the lowest *m/N* scores (0,0022), despite a high *m* score (265). The second most frequent element, ЯСНО (N = 112008) has the lowest *m/N* score (0,0005). Meanwhile, the elements with the highest *m/N* scores, НАСРАТЬ (0,2394), ПО ФИГУ (0,2195) and ПОФИГ (0,1441) have the lowest *N* scores: НАСРАТЬ occurs only 71 times, ПО ФИГУ — 81 times and ПОФИГ — 111 times.

In the mid-range, there is neither a gradual decline nor a gradual increase of the *m/N* score with the rise of *N*. We dropped all low frequent elements with $N < 1000$, the two highest frequent elements with $N > 100000$, two elements with highest *m* score and set the *m* limit at $m \geq 30$. The trimmed list contains 48 items in the range $30 \leq m \leq 496$, $1025 \leq N \leq 46602$. The same or nearly the same *m* value is reached by the DPS items with very different *N* scores, cf. ХОРОШО (m = 176, N = 46602, m/N = 0,0038) with СТЫДНО (m = 175, N = 3076, m/N = 0,0568). This negative result hints that the hypothesis B) is correct. To explain the *m*/*N* scores, one has to consider the individual profiles of the items like ХОРОШО and СТЫДНО. In this pair, СТЫДНО is more specialized in the DPS construction and the expectancy of the 1st person use with a subject pronoun in the contact position for this item is almost 15 times higher compared to ХОРОШО.

The cross-comparison of negative and non-negative DPS items and their quasi-homonyms provides a tool for checking the hypothesis B). There are 13 such pairs in A87. In 3 of them the negation does not constrain the number of syntactic contexts: (НЕ) НАДО, (НЕ) ЖАЛЬ, and (НЕ) НУЖНО. These 6 items lack adverbial side-uses. The same holds for the pair (НЕ) ИЗВЕСТНО, but the non-negative member occurs here in a wider set of contexts. In 3 pairs — (НЕ) ТРУДНО, (НЕ) СТРАШНО and (НЕ) ЖАЛКО — the negative member lacks regular adverbial side-uses, while the non-negative member retains them. Finally, in 6 pairs adverbial uses are attested with both members of the quasi-synonymic hyperset. In all 13 pairs, the negative member is significantly less frequent. The baseline hypothesis is that the *m/N* score increases in the context of negation, since the negative members are expected to be less frequent and more specialized in the predicative function[7]. However, the absence or presence of adverbial uses does not predict that the negative member has an increased or decreased

[7] Almost all hits of НЕ СТРАШНО, НЕ ЖАЛКО and НЕТРУДНО tagged by the RNC engine as adverbials are actually non-agreeing predicatives. The sole example of the genuine adverbial use is weird: *Трудный, неприемлемый для нас человек, сыгранный с легкостью,* ***нетрудно****, ненапряженно, -- это и по-особому назидательный случай в практике сцены* (N.Berkovskij, 1990 – 2000).

m/N score: each subgroup includes both pairs of the type δ ($m/N_{NON\text{-}NEG} - m/N_{NEG}$) > 0 and pairs of the type δ ($m/N_{NON\text{-}NEG} - m/N_{NEG}$) < 0.

Tab. 3. Negative and non-negative DPS items in RNC, 2000-2021.

Without negation	*N*	*m/N*	With negation	*N*	*m/N*	*δ*
I. No adverbial side-uses with both members						
ЖАЛЬ	4606	*0,0486*	НЕ ЖАЛЬ	177	**0,1242**	0,0756
НАДО	78872	0,0192	НЕ НАДО	11828	**0,0282**	0,009
НУЖНО	35580	**0,0345**	НЕ НУЖНО	4145	0,03	-0,0045
ИЗВЕСТНО	15192	**0,0326**	НЕИЗВЕСТНО	4938	0,0141	-0,0185
II. No regular adverbial side-uses with the negative member						
СТРАШНО	7301	0,0036	НЕ СТРАШНО	778	**0,0411**	0,0375
ТРУДНО	14455	**0,0235**	НЕТРУДНО	1453	0,0151	-0,0084
ЖАЛКО	3482	**0,0459**	НЕ ЖАЛКО	711	0,09	-0,0441
III. Regular adverbial side-uses with both member						
ПОНЯТНО	12042	0,0053	НЕПОНЯТНО	4153	**0,0202**	0,0149
ИНТЕРЕСНО	11856	0,0231	НЕИНТЕРЕСНО	1230	**0,0349**	0,0118
ХОРОШО	46602	0,0036	НЕХОРОШО	1259	**0,015**	0,0114
ПРИЯТНО	5157	**0,0337**	НЕПРИЯТНО	1576	0,031	-0,0027
ВАЖНО	10792	***0,0093***	НЕВАЖНО	3616	0,0006	-0,0087
ЛЕГКО	14148	**0,0446**	НЕЛЕГКО	1229	0,0044	-0,0402

The pairs, where the *m/N* decreases in the context of negation, can have some hidden property, e.g. the high initial *m/N* score by the non-negative member. However, this does not explain the increase on НЕ ЖАЛЬ, despite ЖАЛЬ has a high *m/N* score (0,0486) and the slight decrease on НЕВАЖНО, despite ВАЖНО has a low *m/N* score (0,0486).

4. General discussion and conclusions

There are two kinds of data — the frequencies of specific elements associated with the described grammatical construction and general properties associated with the lists of DPS predicative representing the upper part of the frequency dictionary. The ranks of specific predicatives, with the possible exception of the 2-3 most frequent items (НАДО, НУЖНО, ИЗВЕСТНО) depend on the chosen corpus. Meanwhile, the orientation towards the 1st person contexts in the direct speech and the type of meaning indicating that the speaker himself/herself is the source of information about his/her internal state are general features of the Russian DPS construction and its lexicon. There are reasons to think that these features are only minimally text-dependent. One needs a corpus that is large enough to range a list of predicatives and has 1st person contexts. Since a vast majority of Russian DPS predicatives is autoreferential, the lists of the predicatives can be retrieved via the *m*-metrics, which serves two purposes: 1) it gives the number of confirmed DPS clauses with overt subject pronouns in the syntactically annotated corpus assembled by the search query "STIMULUS" + "*мне*" in the window <-1; 1>; 2) it provides a ranging of mid-frequent and high-frequent DPS items.

For each text collection, there is a minimal *m* value, which tells apart regular DPS items from occasional combinations with a dative pronoun. A control list can be retrieved via the *e*-metrics, which provides a second syntactic corpus with confirmed DPS hits in the 3rd person contexts with 3rd person singular subject pronouns in the contact position. The positive *m/e* score confirms that the predicative is entrenched in the DPS construction: ca. 71— 79% of mid- and high-frequent DPS items have the *m/e* scores ≥ 1, 25. The share of non-autoreferential predicatives with the *m/e* score < 1 is minimal in the list containing the most frequent items with *m* > 100.

Russian DPS predicatives always have homonyms. The score of all homonyms (*N*) provides the background for the frequency dictionary. The score *m/N* shows the expectation of finding a DPS construction in the 1st context with a subject pronoun. There is no general formula predicting the *m/N* ratio

for each item, at least in the RNC. This negative result is in accord with the baseline hypothesis that Russian DPS sentences represent a highly idiomatic grammatical construction that does not borrow its elements from the general lexicon but creates it in the dedicated syntactic contexts.

There are several ways of implementing the applied procedure in corpus studies, grammatical theory and cross-language comparison: 1) the retrieved dictionary can be checked on other corpora of Russian; 2) the frequency metrics can be applied for the description of other Russian constructions with an animate priority argument; 3) the statistic profile of the Russian DPS construction and the relevant features '± syntactic animacy', '± autoreferentiality'underlying it can be compared to the characteristics of similar dative constructions in the world's languages.

Acknowledgments

This research has been supported by the Russian Science Foundation, project no. 22-18-00528 "Clausal connectives in sentence and discourse: Semantics and grammaticalization paths".

References

[1] Apresjan Ju. D. Sintaksičeskie priznaki leksem [The syntactic features of lexemes], *Russian linguistics*. 1985. Vol. 19, 2-3. P. 289 – 317.

[2] Bonč-Osmolovskaja A. Kvantitativnye metody v diahroničeskih korpusnyh issledovanijah, Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference 'Dialogue 2015'. [Komp'juternaja lingvistika i intellektual'nye tehnologii. Trudy mezhdunarodnoj konferencii 'Dialog 2015']. Issue 14. [Vyp. 10]. Moscow: RGGU Publ. 2015. P. 80-94.

[3] Davidson D. The individuation of events, D.Davidson (ed.), Essays on actions and events. Oxford: Clarendon Press, 1980. P. 163-180.

[4] Ivanova E., Zimmerling A. Shared by All Speakers? Dative predicatives in Bulgarian and Russian, *Bulgarian Language and Literature*. 2019, LXI, 4. P. 353–363.

[5] Kustova G.I. Tipy infinitivnyx konstrukcij s predikativami (po dannym Nacional'nogo korpusa russkogo jazyka [The types of infinitive constructions with predicatives (according to the Russian National Corpus], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference 'Dialogue 2021'. [Komp'juternaja lingvistika i intellektual'nye tehnologii. Trudy mezhdunarodnoj konferencii 'Dialog 2021']. Issue 20. [Vyp. 20]. Moscow: RGGU Publ. 2021. P. 456-463.

[6] Kustova G.I. Semantičesjue tipy infinitivnyh konstrukcij russkih predikativov [The semantic types of infinitive constructions with Russian predicatives], S.Koeva, E.Yu.Ivanova, J.Tisheva and A.Zimmerling (eds.), Ontologija na situaciite za săstojanie – lingvistično modelirane. Săpostavitelno izsledvane za bălgarski i ruski [The ontology of stative situations – linguistic modeling. A contrastive study of Bulgarian and Russian]. Professor Marin Drinov publ, Sofia. 2022. P. 246–279.

[7] Maienborn C. On Davidsonian and Kimian states, Comorovski, I., K. von Heusinger (eds.). Existence. Semantics and Syntax. Dordrecht: Springer. 2007. P. 107–130.

[8] Pospelov N.S. V zaščitu kategorii sostojanija [In defence of the category of state], *Voprosy jazykoznanija* [*Issues in linguistics*]. 1955, 2. P. 55 – 65.

[9] Švedova N.Y. Russkaja grammatika [Russian grammar]. In 2 vols. Vol. 2. Nauka, Moscow, 1982.

[10] Yanko T.E. Kommunikativnyj status russkih benefaktivnyh konstrukcij [The communicative status of Russian benefactive constructions], *Moscowskij lingvističeskij žurnal* [Moscow linguistic journal], 1996.

[11] Yanko T.E. Kommunikativnye strategii russkoj reči [The communicative strategies in Russian speech . Yazyki slavyanskoi kul'tury, Moscow. 2021.

[12] Zaliznjak Anna A. Issledovanija po semantike predikatov vnutrennego sostojanija [Investigations in the semantics of inner state predicates]. Otto Sagner, München, 1992.

[13] Zaliznjak Anna A. Mnogoznačnost' v jazyke i sposoby ee predstavlenija [Polysemy and its representations]. Jazyki slavjanskoj kul'tury, Moscow. 2006.

[14] Zimmerling A. Russkie predikativy v zerkale eksperimenta i korpusnoj grammatiki [Russian predicatives in the perspective of the sociolinguistic experiment and corpus grammar], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference 'Dialogue 2017'. [Komp'juternaja lingvistika i intellektual'nye tehnologii. Trudy mezhdunarodnoj konferencii 'Dialog 2017']. Issue 16. [Vyp. 16]. Moscow: RGGU Publ. 2017. P. 466-481.

[15] Zimmerling A. Predikativy i predikaty sostojanija v russkom jazyke [Predicatives and the predicates of state in Russian], *Slavistična revija*. 2018, 1. P. 45–64.

[16] Zimmerling A. Avtoreferentnost’ i klassy predikativnyh slov [Autoreferentiality and predicative classes], V.V.Kazakovskaja, M.B.Voejkova (eds.), Problemy funkcional’noj grammatiki.Otnošenie k govorjaščemu v semantike grammatičeskih kategorij [The issues in functional grammar. The speaker-oriented grammatical categories]. Jazyki slavjanskoj kul’tury, Moscow. 2020. P. 23-58.
[17] Zimmerling A. Ot integral’nogo k aspektivnomu [From integral frameworks to aspective descriptions]. Aletheia, Sankt-Peterburg and Moscow. 2021a.
[18] Zimmerling A. Primary and secondary predication in Russian and the SLP: ILP distinction revisited, V. Warditz (ed.), Russian Grammar: System – Language Usage - Language Variation. Peter Lang, Frankfurt a.M. et al. 2021b. P. 543–560.

Russian National Corpus [Nacional'nyj korpus russkogo jazyka]: <www.ruscorpora.ru>.