NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Olga Lyashevskaya

FREQUENCY DICTIONARY OF INFLECTIONAL PARADIGMS: CORE RUSSIAN VOCABULARY

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: HUMANITIES
WP BRP 35/HUM/2013

Olga Lyashevskaya[1]

FREQUENCY DICTIONARY OF INFLECTIONAL PARADIGMS: CORE RUSSIAN VOCABULARY[2]

A new kind of frequency dictionary is a valuable reference for researchers and students of Russian. It shows the grammatical profiles of nouns, adjectives, and verbs, namely the distribution of grammatical forms in the inflectional paradigm. The dictionary is based on data from the Russian National Corpus (RNC) and covers a core vocabulary (5,000 most frequently used lexemes).

Russian is a morphologically rich language: its noun paradigms harbor two dozen case and number forms, while verb paradigms include up to 160 grammatical forms. The dictionary departs from traditional frequency lexicography in several ways: 1) word forms are arranged in paradigms, so their frequencies can be compared and ranked; 2) the dictionary is focused on the grammatical profiles of individual lexemes, rather than on the overall distribution of grammatical features (e.g., the fact that Future forms are used less frequently than Past forms); 3) the grammatical profiles of lexical units can be compared against the mean scores of their lexico-semantic class; 4) in each part of speech or semantic class, lexemes with certain biases in the grammatical profile can be easily detected (e.g. verbs used mostly in the Imperative, Past neutral, or nouns often used in the plural); and, 5) the distribution of homonymous word forms and grammatical variants can be followed over time and within certain genres and registers. The dictionary will be a source for research in the field of Russian grammar, paradigm structure, form acquisition, grammatical semantics, as well as variation of grammatical forms.

The main challenge for this initiative is the intra-paradigm and inter-paradigm homonymy of word forms in the corpus data. Manual disambiguation is accurate but covers approximately five million words in the RNC, so the data may be sparse and possibly unreliable. Automatic disambiguation yields slightly worse results. However, a larger corpus shows more reliable data for rare word forms. A user can switch between a ʽbasic' version, which is based on a smaller collection of manually disambiguated texts, and an ʽexpanded' version, which is based on the main corpus, a newspaper corpus, a corpus of poetry, and the spoken corpus (320 million words in total).

The article addresses some general issues, such as establishing the common basis of comparison, a level of granularity for the grammatical profile, and units of measurement. We suggest certain solutions related to the selection of data, corpus data processing, and maintaining the online version of the frequency dictionary.

JEL Classification: Z.

Key words: frequency dictionary, grammatical profile, inflection, grammatical homonymy, grammatical variation, Russian, Russian National Corpus.

[1] National Research University Higher School of Economics, Moscow / University of Helsinki; Phone: +7 906 798 60 21; email: olesar@gmail.com; www: http://olesar.narod.ru.

[2] This study comprises research findings from ʽA frequency dictionary of Russian grammar and lexical co-occurrence', a project carried out within The National Research University Higher School of Economics Academic Fund Program in 2012-2013, grant No. 11-01-0171.

1. Toward lexicon-oriented grammar

Some time ago, object-oriented programming revolutionized the world of computer technologies, the software industry, and the interfaces of web resources. A long list of step-by-step instructions that automated everything was replaced with sets of objects with shared attributes and behavior. It is now objects that monitor the environment for events and trigger assigned functions. The metaphor of rules and objects can be easily applied to the grammar of natural language. What if grammar is a self-organizing community of words rather than a general who leads the battle? What if the local grammar that these words evoke are more efficient and powerful? Could we prove that the local effects are still systematic and not random? This article reports on an experimental frequency dictionary that aims to provide evidence to answer these questions.

A pilot frequency dictionary shows the inflectional paradigm structure of the 5,000 most frequent Russian nouns, adjectives, and verbs. It follows a series of frequency dictionaries based on data from the Russian National Corpus.[3] Our lexico-grammatical dictionary presents a comprehensive account of how inflection works, thus filling the gap between data on lexical frequency and grammatical frequency.

As a general practice, most frequency dictionaries present a distribution of lexical data either at the level of tokens or at the lemma level (Leech et al. 2001, Davies and Gardner 2010 for English, Davies 2005 for Spanish, Čermák et al. 2010 for Czech, Sharoff et al. in press for Russian, etc.). In addition, the number of words in different parts of speech classes can be given. However, rarely does any frequency dictionary in the morphologically rich languages include information about the structure of the paradigm and grammatical form frequencies. The only exception we are aware of is Šteinfeldt (1963,1970), which calculates the frequency of 961 Russian nouns in each case and number form, as well as the distribution of some verbs in tense and mood.

As for grammatical frequency data, despite the fact that the task of constructing frequency in Russian grammar has long been recognized and promoted in the literature (Mustajoki 1973, Ilola and Mustajoki 1989, Baerman et al. 2010), the quantitative research is mostly focused on non-lexical units: part of speech classes, grammatical classes (hierarchies of case marking, agreement markers, etc.), as well as morphemes (Šajkevich et al. 2008).

Our project shifts the focus from the distribution of grammatical classes and categories to particular word forms as structured by the inflectional paradigm. Of particular concern are words with certain biases in the grammatical profile, such as:

- verbs used mostly in the imperative;
- verbs never used in the past neutral;
- nouns often used in the plural;
- nouns with low rates of usage in the nominative form.

The dictionary will be a source for much future research in the area of Russian grammar, paradigm structure, grammatical semantics, as well as the variation and alternation of grammatical forms (Graudina et al. 1976). It will provide a detailed account of the gradual nature of some important phenomena such as singularia and pluralia tantum. Data from the RNC provide a great opportunity to answer many research questions, taking into account current technologies of corpus linguistics.

[3] See Lyashevskaya and Sharoff (2009), who generate a general frequency dictionary of 50,000 words, and the online grammatical and collocational dictionaries found at http://dict.ruslang.ru

This article presents a short introduction to the structure of Russian paradigms in Section 2, examines the background of the lexico-centric approach to frequency grammar in Section 3, discusses various types of information available in the dictionary in Section 4, and mentions some issues associated with the processing and interpretation of the frequency data in Section 5 and in the Conclusion.

2. The structure of Russian paradigms

The data set in the dictionary is based on the morphological standard of the RNC (Lyashevskaya et al. 2005), which generally follows the paradigm inventory developed in the grammatical dictionary (Zaliznjak 1974). The dictionary takes into account only single-word forms, not paraphrastic forms, such as conditional forms with the particle *by*, imperfective future and passive participle constructions with the auxiliary *byt'* ('to be'), analytical forms of the comparative degree, etc.

The paradigm of Russian nouns has up to 17 cells: two grammatical numbers multiplied by six major cases (nominative, genitive, dative, accusative, instrumental, and locative), plus five minor forms that some words take in the singular: the so-called 'second' genitive, 'second' accusative, 'second locative', vocative, and adnumerative.

The adjectival paradigm has at least 32 cells: 26 inflected long forms (three genders in the singular and the plural multiplied by six cases; in addition, the masculine and the plural adjectives distinguish between two types – animate and inanimate – of the accusative forms), four short forms (three genders in the singular and the plural), and two comparative forms. Zaliznjak's grammatical dictionary (Zaliznjak 1974) and the morphological standard of the RNC exclude from the standard paradigm some archaic short forms with case endings and the superlative forms that have the same declension as the full forms. Russian pronouns and numerals function either as nouns or as adjectives.

There is a considerable syncretism within the nominal paradigm. Some case forms are fairly homonymous. For example, the accusative and the nominative forms of inanimate nouns, the accusative and the genitive forms of animate nouns (except for the feminine singular that has six distinct case forms), the genitive, dative, instrumental, and locative forms of many adjectives in the singular, etc.

The paradigm of verbal forms has less cohesion among its forms than the declension of nouns. Imperfective and perfective verbs are considered to be separate lexical units and have a slightly different paradigm structure. The non-past forms express the present tense for imperfective verbs and future tense for perfective verbs and distinguish three grammatical persons and two numbers. The past tense has four forms – three genders in the singular and the plural. Imperative forms distinguish between the first and second person, and singular and plural, plus a minor inclusive form. The non-past and past forms are usually used in the active-middle voice; passive forms formed with the reflexive *–sja* affix are almost non-existent but potentially double the number of indicative forms. The imperfective has two gerunds for present and past tense, whereas perfective verbs form only the past gerund. Imperfective verbs can have up to four participles (the present active, past active, present passive and past passive participles), whereas perfective verbs have two (the past active and past passive participles). Each participle can have a full complement of twenty-four adjectival full forms and four short forms. The infinitive has only one form and is a basic dictionary form of the verb.

3. Why mean frequencies may not always help

When linguists talk about frequency grammar, they presumably refer to a specific type of quantitative data like the frequency ratio for part of speech classes, and the frequency hierarchy of cases and other grammatical categories. The topic of case-frequency distribution is particularly popular in Russian linguistics: Kopotev (2008) cites three studies that were published during 1959-1961, and there are many more recent publications that report research results based on various digitized text samples.

Kopotev (2008) draws attention to the stability of frequency data in large and small corpora (see Table 1). The first two sets of data are based on the modern corpora (RNC and HANKO), while the two others are drawn from the earlier frequency dictionaries based on datasets smaller than 0.5 MW. Kopotev concludes that the modern corpora agree quite well in the assessment of the mean probability of case occurrences, and the differences lie in the structure of text collection in terms of genres.

	Nom	Gen	Dat	Acc	Ins	Loc
Russian National Corpus	27.06%	29.23%	5.98%	18.66%	8.44%	10.63%
Corpus HANKO	24.30%	32.62%	5.50%	17.73%	8.08%	11.78%
Josselson 1953	38.80%	16.80%	4.70%	26.30%	6.50%	6.90%
Šteinfeldt 1963	33.60%	24.60%	5.10%	19.50%	7.80%	9.40%

Table 1. Frequency distribution of six Russian cases in Kopotev (2008: 142).

However, the principle of 'choose genitive if unsure' may easily mislead a student of Russian as a foreign language if he or she has to choose an appropriate case for the word *shepot* (whisper). Table 2 shows that the frequency distribution of cases within some nominal paradigms deviates significantly from a typical pattern.

	Nom	Gen	Dat	Acc	Ins	Loc	Total (F.abs)
shepot 'whisper'	10.9%	3.7%	0.9%	8.3%	**75.6%**	0.6%	349
poza 'posture'	15.9%	6.3%	0.8%	19.0%	4.0%	**54.0%**	126
tropinka 'walking path'	27.6%	2.0%	**52.0%**	5.1%	5.1%	8.2%	98

Table 2. The grammatical profile of the nouns *shepot* (whisper), *poza* (posture),, and *tropinka* (walking path). (Case forms in singular.)

In as early as 1974 Greenberg proposed a hypothesis that different semantic groups may have a different distribution of cases – both with prepositions and without them (Greenberg 1974,1991). The choice of Russian as an object of study is not accidental: his hypothesis testing is carried out on data from the aforementioned frequency dictionary (Šteinfeldt 1963, source corpus 0.4 MW). Greenberg classifies one half of the nouns into twelve categories (animals, body parts, time periods, etc.), and calculates the average frequency of each group for each case. As expected, the place names are used mostly in the accusative and locative form, while the dative shows a higher value in personal pronouns (1st and 2nd person) (see Table 3). However, not every episode is explained that easily.

	No. of types	No. of tokens	Nom	Gen	Dat	Acc	Ins	Loc
1. All nouns	9.073*	102.173	33.6%	24.6%	5.1%	19.5%	7.8%	9.4%
2. Common nouns	9.073	89.384	28.3	26.0	5.0	21.8	8.6	10.3
3. Proper nouns	?	12.789	76.3	13.5	5.5	1.1	1.4	2.2
4. Personal common individual	119	11.769	54.1	22.5	6.9	7.0	8.0	1.5
5. Personal collective	29	2.565	23.9	48.0	4.2	9.6	6.2	8.3
6. Animal	9	746	35.6	28.4	3.8	21.6	6.0	4.6
7. Body parts	31	3.318	18.2	9.9	3.2	36.5	20.3	11.9
8. Concrete count	116	5.475	23.0	20.7	4.3	32.0	9.4	10.5
9. Concrete mass	25	1.565	21.3	31.6	2.2	24.3	13.6	6.9
10. Non-enduring objects	31	2.127	34.5	19.0	4.1	21.5	10.8	8.8
11. Abstract qualities	21	1.295	33.3	24.9	3.8	17.4	12.3	9.0
12. Place nouns	87	7.747	11.8	30.6	6.0	24.4	3.3	23.8
13. Place institutions	17	2.445	13.0	40.9	2.3	17.8	1.8	24.1
14. Time periods	8	2.998	12.8	37.5	2.0	36.0	3.4	8.3
15. Measures	7	480	2.7	85.4	0.8	5.2	1.2	4.6
16. First and second person pronouns	4	15.901	51.8	21.8*	14.5	9.5	2.2	0.2

Table 3. The relation of Russian case frequencies to semantic features, adopted from Greenberg (1974,1991: 210).

Rice and Newman (2005) and Newman (2005) make the observation that a distribution of grammatical variation may be present within lexical groups. They focus on the notion of inflectional islands (or the skewedness of the frequencies of certain inflectional forms according to the corpus data), and their source corpus is BNC. Rice and Newman notice that the frequencies of the English *think*, *know*, and *mean* at the inflectional level are very different, even though the meaning of the three verbs appear to be very similar. They claim that the large inflectional islands help to gain insight into what are believed to be semantic differences.

The concept of grammatical profile is introduced in Janda and Lyashevskaya (2011) as the relative frequency distribution of the inflected forms of a word in a corpus. On the one hand, it presents a more elaborated version of Rice and Newman's ideas for rich inflectional paradigms that involve many forms and complex grammatical oppositions[4]. On the other hand, this follows the behavioral profile approach proposed by Divjak and Gries (2006) as a tool to study the semantic and functional reasons for the discrepancies of corpus frequencies. Whereas Divjak and Gries indiscriminately take a vast range of morphological, syntactical, lexical, and semantic

[4] The grammatical profile approach was also applied to Old Slavonic data (Eckhoff Janda 2013). See also Janda et al. (2013) on constructional profiles and radial category profiles, Janda and Lyashevskaya (to appear) on semantic profiles, and Kuznetsova (2013) on collostructional profiles, and other quantitative approaches to explaining the RNC data developed by the CLEAR group at the University of Tromsø.

factors and build the hierarchical structure based on their contribution to the frequency patterns, the grammatical profile explores only inflectional data. Our study of aspect, tense, and mood forms in Janda and Lyashevskaya (2011) shows, among other things, the predictable frequency effects in the imperative form of the imperfective verbs that refer to a frame of being a guest. The imperatives do not provide new information, but rather express standard polite formulae such as *razdevajtes’* (‘take off your coat’), *sadites’* (‘sit down’), *zakusyvajte* (‘have some refreshments’), *zakurivajte* (‘have a smoke’), etc. The infinitive perfective forms are often used with verbs like *vospolnit’* and *sootnesti*, which in turn prefer the constructions with tentative verbs and adverbs describing the difficulty or importance of an achievement.

Kuznetsova (2013) studies the ratio of the masculine and feminine forms used in the past tense. Kuznetsova analyses the class of typically ‘feminine’ and typically ‘masculine’ verbs: whereas verbs of sewing, knitting, embroidery, darning and so on are associated with feminine forms, while verbs that describe being a leader or negatively evaluated behavior, like drinking or spitting, are found in the ‘masculine’ list (see Table 4).

lemma	fem	masc	neut	fem:masc	role
načal'stvovat’ ‘be chief’	2	124	0	0	leader
predvoditel’stvovat’ ‘chair’	3	77	0	0	leader
otrjadit’ ‘dispatch’	5	126	3	0	leader
pomilovat’ ‘pardon’	6	140	2	0	leader
knjažit’ ‘reign’	3	56	1	0.1	leader
predsedatel’stvovat’ ‘chair’	16	222	1	0.1	leader
kurirovat’ ‘supervise’	12	165	7	0.1	leader
…	…	…	…	…	…
ograbit’ ‘rob’	13	260	6	0.1	criminal
xuliganit’ ‘behave like a hooligan’	2	33	0	0.1	criminal
umyknut’ ‘walk away with’	2	29	0	0.1	criminal

Table 4. Verbs associated with masculine roles. Data from Kuznetsova (2013, Table 9).

What is most surprising about the Russian lexical system is that there are no nouns such that their grammatical profile would correspond to the ‘average’ profile of the substantives, verbs with the ‘average’ proportion of tense-person-number forms, and so on. In order to make Greenberg’s hypothesis work, we must assume that when we explore the lexical space, we deal with multiple overlapping classes, complex superposition of semantic features, and syntactic and structural properties. All these collectively affect the frequency output.

Greenberg was looking for the ‘magic’ ratio, which would allow the word to refer to a particular semantic class – no wonder he never succeeded in finding it. Now, from the viewpoint of cognitive semantics, his observation can be reinterpreted as a semantically motivated shift of grammatical forms frequencies. Let us take three examples. A high ratio of the instrumental case forms in the paradigm of the noun *shepot* (‘whisper’) can be explained as the overlap of the lexical semantics (whisper as a way of speaking) and the semantics of its grammatical form (the instrumental of manner). The overlap between the stative semantics of the noun *poza* ('posture') and the semantics of placement in the locative prepositional group *v* ('in') + S.loc explains the peak on the locative case forms: consider *v poze (lotosa)* ('in the lotus position'). The nouns that refer to a path or trajectory are most typical fillers of the prepositional group *po* + S.dat ('along (a path)'), and hence there is a prevalence of the dative forms in the noun *tropinka* ('walking path').

4. Frequency information in the dictionary

4.1. Small and expanded dataset

There are two sets of frequency data in use in the dictionary. The 'small' version is based on the manually disambiguated subcorpus and contains more accurate annotation from the point of view of grammatical homonymy. However, since the subcorpus is as small as 5 MW, the chances of coming across rare grammatical forms is very low or none at all. As a result, if a particular word form is not frequent, its distribution within the paradigm may not be reflected properly in the 'small' version.

The 'expanded' version is based on the collection of four RNC corpora, which is 60 times larger than the manually disambiguated subcorpus. The disambiguation of lemmas, parts of speech, and grammatical features is done by a computer, so the attribution of homonymous tokens (for example, the accusative form that mirrors the genitive animate form) is less reliable. Both versions are imperfect but the user can choose between them according to the task at hand.

4.2. The level of granularity for the grammatical profile

The most detailed level of granularity available in the data is not necessarily the most optimal one. Given the number and structure of forms, grammatical profiles can be shown at varying levels of resolution.

The user can evaluate certain grammatical distinctions for their relevance for his or her research task and collapse them within the grammatical profile. For example, there can be a cluster of full passive participle forms within the verbal paradigm (distinction in case, number, gender and animacy is collapsed), a cluster of four forms of the past tense (distinction in number and gender is collapsed), and a cluster of all singular forms of the noun as opposed to a cluster of all plural forms (distinction in case is collapsed).

Moving in the opposite direction, the grammatical profile of nouns can be extended with the distributional pattern of prepositions that are used with each case. This is done since the research practice suggests that the most fine-grained level of inflectional profiling is not enough: when the case forms are used alone, or used with different prepositions, their meaning and syntagmatics change dramatically.

4.3. Homonyms and grammatical variants

The user can restrict the output to the relevant subset of grammatical forms. Two look-up tables are pre-defined: a subset of inflectional forms that have a homonym within the paradigm or outside,[5] and a subset of graphically distinct variants of a grammatical form. The dictionary shows:

1) The ratio of graphically identical forms within one paradigm, e. g. *soldat* ('soldier'), Nom. sg. VS Gen. pl. VS Acc. pl.

[5] See Ventsov and Kasevich (2004) on intra-paradigm and inter-paradigm homonymy in Russian

2) The ratio of each homonymous form that belongs to the selected paradigm and those that belong to other paradigms, e. g. *zaplyv*, which is a form of the verb *zaplyt'* ('swim away' or 'be swollen shut') and the noun *zaplyv* ('a swim') (see Table 5).

Word form	Lemma	PoS	Classif.	Inflect.	F.abs	%
zaplyv	zaplyt' 'swim away; be swollen shut'	V	intr, pf	gerund, past, act	4	2.24%
zaplyv	zaplyv 'a swim'	S	inan, m	Nom. sg.	100	56.18%
zaplyv	zaplyv 'a swim'	S	inan, m	Acc. sg.	74	41.57%

Table 5. Ratio of the homonymous forms in the newspaper corpus: *zaplyv*.

3) The ratio of occurrence for grammatical variants, i.e. forms that have the same grammatical annotation (e. g. compar. *silnej* VS *silnee* ('stronger'), Ins. pl. *dverjami* VS *dver'mi* ('doors'), etc.).

4) Graphically distinct 'major' and 'secondary' substantive cases and adjectival comparative forms (these forms have a slightly different annotation according to the RNC standard. Cross reference Gen. sg. (*bez*) *tolka* and Gen 2. sg. (*bez*) *tolku* ('no use (doing smth.)'); compar. *silnej* and compar. 2 *posil'nej*, compar. *silnee* and compar. 2 *posilnee* ('stronger').

4.4. Grammatical profiles in time and genres

Information about the shifts in grammatical profiles over time is given in 10-year intervals, and in the corpus of modern newspapers these time spans are narrowed to 1 year. While the default dataset involves data from 1900 to 2010, the diachronic changes can be traced back to the 1800s.

The user can also follow the shift of distributional properties in different collections, including literary prose, poetry, periodicals, everyday communication, research, and teaching materials. The list also includes other genres of non-fiction, electronic communication, and oral non-public speech.

4.5. Units of measurement

The user can choose one or several units to measure occurrences in the corpus:

– a number of documents

– an absolute frequency of occurrences (F.abs) and the total corpus size (absolute scale is required for further use in statistical toolkits)

– a relative frequency in ipm (items per million)

– a hierarchy (or a ranked list) of grammatical features, like the following

Loc> Gen> Nom> Acc> Dat> Ins

– a percentage distribution (see Table 6)

	Nom	Gen	Dat	Acc	Ins	Loc	Total (F.abs)
sg	98	128	29	170	137	14	576
pl	4	9	3	7	2	2	27
	Nom	**Gen**	**Dat**	**Acc**	**Ins**	**Loc**	**Total (%)**
sg	17.0%	22.2%	5.0%	29.5%	23.8%	2.4%	100.0%
pl	14.8%	33.3%	11.1%	25.9%	7.4%	7.4%	100.0%

Table 6. Grammatical profile of the noun *vlijanie* (‘impact’). Case distribution: absolute frequencies and percentage.

What benchmark should be used as a basis for comparison (100%) is not a straightforward question. This can be the total of all word-form frequencies. However, the paradigm itself is sometimes not stable. For example, perfective and intransitive verbs systematically lack certain participles and gerunds, and the morphological shape of an imperfective verb determines whether it can form these peripheral forms, too. What is more, one can claim that participles and gerunds are not members of the verbal paradigm, but form two separate (adjective-like and adverb-like) classes of parts of speech. Thus, the stable part of the paradigm includes three basic moods: the infinitive, the indicative, and the imperative. Relative adjectives are prevented from forming short forms and comparative and superlative degrees, so the core part of the adjectival paradigm only includes long forms. Still, ‘secondary’ cases should not be excluded from the totals as they occupy roughly the same syntactic position as ‘major’ cases. The user determines which part of the paradigm is to be rated as 100%.

– form1:form2 ratio (the usage proportion of the two forms with respect to each other, see Table 7)

Lemma	Form1: partcp act praes	Form1: F.abs	Form2: partcp pass praet	Form2: F.abs	Form1 :form2 ratio	Total
trebovat’ ‘require, demand’	trebujuščij	83	trebovannyj	0	∞	83
vesti ‘lead, conduct’	veduščij	119	vedennyj	1	119.00	120
goret’ ‘burn’	gorjaščij	108	goretyj	1	108.00	109
znat’ ‘know’	znajuščij	88	znato	2	44.00	90
igrat’ ‘play’	igrajuščij	47	igran	3	15.67	50
govorit’ ‘speak, talk’	govorjaščij	60	govoreno	5	12.00	65
pisat’ ‘write’	pišuščij	32	pisan	35	0.91	67
bit’ ‘beat, hit’	b’juščij	20	bityj	42	0.48	62

Table 7. Ratio of active present participles to passive past participles in a sample of verbs.

– quintile scores of word forms (see Table 8)

Quintile scores are categorical values that allow users to track profile patterns compared to the activity of a part of speech class. Table 8 shows the quintile scores for a semantic group of vehicles. Quintiles divide scores of all nouns in a particular case into five equal groups ranked from the lowest (very rare, *a*) to the highest (very frequent, *e*). For example, the frequency of dative forms for the noun *parohod* (‘steamboat’) is found in the lowest group 0% to 20% (*a*), the score of the accusative forms falls into the group 20% to 40% (*b*), the score of instrumental forms falls into the group 40% to 60% (*c*), etc.

Lemma	Nom	Gen	Dat	Acc	Ins	Loc	Total (F.abs)
metro 'underground'	a	e	d	a	a	e	185
korabl' 'ship'	e	c	b	b	a	c	231
gruzovik 'truck'	e	d	c	b	b	c	134
parohod 'steamboat'	e	e	a	b	c	d	121
автомобиль 'car'	d	d	c	b	b	d	441
poezd 'train'	e	c	c	b	b	d	618
samolet 'plane'	d	c	d	c	c	d	385
tramvaj 'tram'	d	b	c	d	c	d	198
lodka 'boat'	d	c	b	d	b	d	280
vagon 'coach'	a	d	d	c	a	e	473
velosiped 'bicycle'	b	c	a	d	b	e	206
avtobus 'bus'	d	b	c	c	b	e	281
taksi 'taxi'	c	a	b	e	a	e	174

Table 8. Quintile scores of case forms in semantic groups of vehicles. The profile of each noun is compared against the case profile of all nouns.

4.6. Comparison of words and classes

The data in the dictionary can be magnified and reduced to get a better view of trends in lexical classes and in parts of speech. There are three levels of data representation. At the first level, the grammatical profiles of individual lexemes are presented. At the second level, the data are generalized into the profiles of lexico-semantic classes, such as verbs of motion, names of instruments, etc.[6] At the third level, information is given about major lexical categories, such as relative adjectives and transitive verbs, about parts of speech, and grammatical categories.

5. Corpus data processing

The main part of the dictionary is based on a corpus collection that spans from 1900 to 2010. The diachronic part involves data starting from the 1800s. Data for the 'small' dictionary were collected in the RNC gold standard collection (5.4 million entries), where lexico-grammatical homonymy was disambiguated manually. Data for the 'expanded' version of the dictionary were taken from the main corpus, newspaper corpus, the corpus of poetry, and spoken corpus of the RNC (320 million words in total).

[6] See http://ruscorpora.ru/en/corpora-sem.html.

First, the statistics of word forms were collected. The functional style and genre of each text, as well as the time of creation, was registered while processing the corpus. The database was indexed by lexical and grammar features (lemma, part of speech, inflectional attributes). We marked lexico-semantic classes according to the RNC classification, capitalized/non-capitalized, and other variants of spelling. Second, we collected 2- and 3-gram statistics that shows the prepositional-case preferences of nouns and pronouns.

The lists of tokens that have more than one grammatical tagset (grammatically ambiguous forms) and grammatical tagsets that correspond to more than one token (grammatical variants and distortions) were complied. Each case in the first group was marked as inter-paradigm or intra-paradigm homonymy. We assessed the reliability of grammatical annotation provided by the tagger. The disambiguated version of four corpora was created with the help of two programs – a light parser that relies on heuristics and the HMM module that is trained on the RNC gold standard collection. A subset of n-grams with conflicting annotation was further evaluated and corrected manually.

Inherently overlapping paradigms proved to be the major problem for disambiguation tools, such as the paradigms of the masculine and feminine variants of the noun *rojal'* ('grand piano') that share most forms and differ in the instrumental singular (consider *rojalem* m. and *rojal'ju* f.). The feminine variant was mostly used in the 19th century and the contemporary dictionary's built-in tagger does not include it. The form *rojal'ju* (f.) was detected as not-in-dictionary form and assigned the correct feminine tagset. At the same time, all plural forms received only masculine tagsets, and the dative and locative singular forms *rojali* got a masculine tagset with the wrong gender, case, and number attribution (nom. pl. or acc. pl.). Thus, the dictionary contains the defective paradigm of the feminine noun *rojal'* and the full paradigm of the masculine noun with an overweighed proportion of plural forms.

Another example is the adjectives *zapásnyj* ('emergency') – as in *zapasnyj vyhod* ('emergency exit') and *zapasnój* ('reserve') – as in *zapasnoj igrok* ('reserve player') – which share most of the forms and partly share contexts and meaning. Both paradigms are stored in the tagger dictionary, but their disambiguation is poor since the algorithm is not tuned to such a tiny difference.[7]

New approaches to disambiguation would solve the problem of overlapping paradigms in the future. In the pilot version of lexico-grammatical dictionary such cases are marked as 'weakly reliable'.

6. Conclusion

The dictionary is primarily addressed to those who study and learn Russian grammar. It will be valuable to researchers interested in inflection and grammatical semantics, as well as to editors and teachers who deal with grammatical variability. Nevertheless, it is noteworthy that the 'lexico-centric' approach, in spite of resource consumption and sparsity of data, proves to also be helpful in natural language processing. As some recent experiments have shown, estimating the lexical probabilities can improve the results of the HMM disambiguation module for Russian by 3% (Danilova et al. 2013). An online output of the dictionary can be freely downloaded in spreadsheet format for further use in any kind of theoretical and practical activity.

The digital form of the dictionary allows us to continuously improve and update it. First, we are planning to develop a functionality to meet the needs of the linguistic community: to visualize data graphically, to plug in external dictionaries, such as a dictionary of grammatical variants, morphemes, and word formations, etc. Second, we will keep improving data quality as we collect user feedback and error reports and improve the grammatical disambiguation of the RNC

[7] Note that it is not always possible to choose the tagset manually

data. Third, we will evaluate including new data and new distributional patterns (one of them is a set of author attributes). This results in data becoming sparser, so working with small frequencies requires special care and handling techniques to avoid skewedness and extreme ratios.

The main challenge, though, is to understand how to use the measures, how to interpret the data, how to transfer the obtained probabilities for other corpora and resources, and how to make accurate claims and predictions about the functioning of inflectional forms in general. The proposed project is the first experiment in making a large lexico-grammatical reference resource and, therefore, it will provide good material for analyzing the reliability of corpus data. There is no doubt that we need to better understand corpus data, the structure of corpora samples, and how it is related to the sustainability of statistical data. We will learn how to balance the samples of various time periods and genres, and the dictionary will give us a chance to experiment with rich lexical material. This will enable us to prove the adequacy of corpus data statistics.

References

Apresjan, Jury D. (1967). Eksperimental'noe issledovanie semantiki russkogo glagola [Experimental research on the semantics of the Russian verb], Moscow.

Baerman, Matthew, Dunstan Brown, Greville G. Corbett, Alexander Krasovitsky, and Peter Williams (2010). Predicate agreement in Russian: A corpus-base approach, Wiener Slawistischer Almanach, Sonderband 74, pp.109-121.

Čermák, František, Michal Křen, Renata Blatná et al. (2010). Frekvenční slovník češtiny. Vyd. 2. Praha: Lidové noviny.

Danilova, V., O. Volkov, A. Ladygina, D. Privoznov, I. Serbinova, and G. Sim (2013). Disambiguation with HMM [Snjatie omonimii metodom HMM]. Moscow (manuscript).

Davies, Mark (2005). A Frequency Dictionary of Spanish: Core Vocabulary for Learners. London–N.Y.: Routledge.

Davies, Mark, and Dee Gardner (2010). A Frequency Dictionary of American English: Word Sketches, Collocates, and Thematic Lists. London–N.Y.: Routledge.

Divjak, Dagmar, and Stefan Th. Gries (2006). Ways of trying in Russian: Clustering behavioral profiles. Corpus Linguistics and Linguistic Theory 2, pp. 23– 60.

Eckhoff, Hanne M., and Laura A. Janda (2013). Grammatical Profiles and Aspect in Old Church Slavonic. Transactions of the Philological Society, Vol. 111 (2).

Graudina, Ludmila K., Viktor A. Itskovich, and Lia P. Katlinskaya (1976). Correct Russian speech: Stylistical dictionary of grammatical choices [Grammaticheskaja pravil'nost' russkoy rechi. Stilisticheskiy slovar' variantov]. Moscow.

Greenberg, Joseph H. (1974/1990). The relation of frequency to semantic feature in a case language (Russian), in Denning, K., S. Kemmer (eds.), On Language, Selected Writings of Joseph H. Greenberg, Stanford, pp. 207-226.

Ilola, Eeva, and Arto Mustajoki (1989). Report on Russian Morphology as it Appears in Zaliznyak's Grammatical Dictionary, (Slavica Helsingiensia 7), Helsinki.

Janda, Laura A., Anna Endresen, Julia Kuznetsova, Olga Lyashevskaya, Anastasia Makarova, Tore Nesset, and Svetlana Sokolova (2013). Why Russian aspectual prefixes aren't empty: prefixes as verb classifiers. Bloomington, IN: Slavica.

Janda, Laura A., and Olga Lyashevskaya (2011). Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian, Cognitive Linguistics, 22 (4), pp. 719-763.

Janda, Laura A., and Olga Lyashevskaya. (To appear). Semantic Profiles of Five Russian Prefixes: po-, s-, za-, na-, pro-. Journal of Slavic Linguistics.

Josselson, Harry H. (1953). The Russian Word Count and Frequency Analysis of Grammatical Categories of Standard Literary Russian. Detroit (MI).

Kopotev, Mikhail (2008). Towards the frequency grammar of Russian: corpus evidence on the grammatical case system [K postroeniju chastotnoy grammatiki russkogo jazyka: padezhnaja sistema po korpusnym dannym] // Mustayoki A., Kopotev M.V., Birjulin L.A., Protasova E.Ju. (eds.), Instruments of Russian linguistics: corpus approaches [Instrumentariy rusistiki: korpusnye podkhody], Helsinki, pp. 136-150.

Kuznetsova, Julia (2013). Linguistic Profiles: Correlations between Form and Meaning. Ph.D. diss., University of Tromsø.

Leech, Geoffrey, Paul Rayson, and Andrew Wilson (2001). Word Frequencies in Written and Spoken English: based on the British National Corpus. Longman, London.

Lyashevskaya, Olga N., Vladimir A. Plungian and Dmitrij V. Sichinava (2005). O morfologicheskom standarte Korpusa sovremennogo russkogo jazyka [Morphological standard of the Corpus of contemporary Russian]. In: Nacionalʹnyj korpus russkogo jazyka: 2003-2005 [Russian National Corpus: 2003-2005]. Moscow, 2005, pp. 111-135.

Lyashevskaya, Olga N., and Serge A. Sharoff (2009). Frequency dictionary of modern Russian based on the Russian National Corpus [Chastotnyy slovar' sovremennogo russkogo jazyka (na materiale Nacional'nogo korpusa russkogo jazyka)], Azbukovnik, Moscow.

Mustajoki, Arto (1973). On compiling the frequency dictionary of Russian nouns [Opyt sostavlenija chastotnoy grammatiki russkikh suschestvitel'nykh], Hel'sinki, (manuscript).

Newman, John (2008). Aiming low in linguistics: Low-level generalizations in corpus based research. Proceedings of the 11th International Symposium on Chinese Languages and Linguistics, National Chiao Tung University, Hsinchu, Taiwan, May 24, 2008.

Rice, Sally, and John Newman (2005). Inflectional islands, ICLC-9, Yonsei University, Seoul, Korea.

Sharoff Serge, Elena Umanskaya, and James Wilson (in press). A Frequency Dictionary of Russian: core vocabulary for learners (Routledge Frequency Dictionaries). NY: Routledge.

Šajkevich, Anatoly Ja., Vladislav M. Andrjuščenko, and Natal'ja A. Rebetskaja (2008). Statističeskij slovar' jazyka russkoj gazety [A frequency dictionary of the Russian newspaper language]. Vol. 1. Moscow: Jazyki slavjanskih kul'tur.

Šteinfeldt, Evi (1963/1970). Russian Word Count, Moscow.

Ventsov, Anatoly V., and Vadim B. Kasevich (eds.) (2004). Dictionary of Russian homographs [Slovar' omografov russkogo jazyka], St.-Petersburg.

Zaliznjak, Andrej (1974). Grammatičeskij slovar' russkogo jazyka [The grammatical dictionary of Russian]. Moscow: Russkij jazyk.

Olga N. Lyashevskaya
Professor, Faculty of Philology, National Research University Higher School of Economics, Moscow, Russia;
PhD student, University of Helsinki, Finland
Phone: +7 906 798 60 21
E-mail: olesar@gmail.com
www: http://olesar.narod.ru

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.