ЧАСТОТНЫЙ ЛЕКСИКО-ГРАММАТИЧЕСКИЙ СЛОВАРЬ: ПРОСПЕКТ ПРОЕКТА[1]

Ляшевская О. Н. (olesar@gmail.com)
НИУ Высшая школа экономики, Москва, Россия

Обсуждается задача создания электронного частотного словаря, в котором будет отражено распределение грамматических форм в парадигме словоизменения русских имен существительных, прилагательных и глаголов, т. е. грамматический профиль индивидуальных лексем и лексических групп. В практике составления частотных словарей и квантитативных исследований стандартным объектом изучения является общая иерархия грамматических категорий, например, частотность частеречных классов или среднее соотношение частот именительного и творительного падежей. В данном проекте фокус переносится на распределение грамматических форм у конкретных лексем, выявление единиц с нестандартным перевесом тех или иных форм в парадигме. Словарь предназначен для исследований русской грамматики, грамматической семантики, а также изучения вариативности форм.

Ресурс строится на материалах Национального корпуса русского языка. В статье затрагиваются общие вопросы использования корпусов для создания частотных ресурсов подобного рода и технологии обработки данных. Предлагаются решения, связанные с отбором данных, уровнем дробности грамматических кластеров, параметрами мониторинга изменения грамматического профиля в зависимости от времени создания текста и жанрово-функционального регистра.

Ключевые слова: частотный словарь, грамматический профиль лексемы, словоизменение, грамматическая семантика, вариативность, русский язык, НКРЯ

LEXICO-GRAMMATICAL FREQUENCY DICTIONARY: A PRELIMINARY DESIGN

Lyashevskaya O. N. (olesar@gmail.com)
NRU Higher School of Economics, Moscow, Russia

A new electronic frequency dictionary shows the distribution of grammatical forms in the inflectional paradigm of Russian nouns, adjectives and verbs,

[1] В работе использованы результаты, полученные в рамках проекта № 11-01-0171, выполненного в рамках Программы «Научный фонд НИУ ВШЭ» в 2012–2013 гг.

i. e. the grammatical profile of individual lexemes and lexical groups. While the frequency hierarchy of grammatical categories (e. g. the frequency of part of speech classes or the average ratio of Nominative to Instrumental case forms) has long been the standard topic of research, the present project shifts the focus to the distribution of grammatical forms in particular lexical units. Of particular concern are words with certain biases in grammatical profile, e. g. verbs used mostly in Imperative, in past neutral or nouns used often in plural. The dictionary will be a source for many of the future research in the area of Russian grammar, paradigm structure, grammatical semantics, as well as variation of grammatical forms.

The resource is based on the data of the Russian National Corpus. The article addresses some general issues such as corpora use in compiling frequency resources and technology of corpus data processing. We suggest certain solutions related to the selection of data and the level of granularity of grammatical profile. Text creation time and language registers are discussed as parameters which may shape the grammatical profile fluctuations.

Key words: frequency dictionary, grammatical profile, inflection, semantics of grammar, form variation, Russian, Russian National Corpus

1. Введение

Частотный лексико-грамматический словарь продолжает серию частотных словарей, создаваемых на данных Национального корпуса русского языка, и является прямым продолжением частотного словаря (Ляшевская, Шаров 2009). В общем частотном словаре основная доля информации была представлена на уровне лексем. Из грамматической информации давались сведения о доле слов разных частей речи и о наиболее частотных словоформах русского языка. Вместе с тем, если смотреть с точки зрения конкретной лексемы, информации о частоте всех ее словоформ словарь не давал. Эту лакуну заполняет новый экспериментальный лексико-грамматический словарь. Он представляет грамматический профиль (т. е. распределение грамматических форм в парадигме словоизменения) 5000 наиболее частотных русских имен существительных, прилагательных и глаголов.

Далее в статье речь пойдет о задачах словаря, его структуре, а также о некоторых проблемных точках, связанных с обработкой и интерпретацией частотных данных.

2. Предназначение словаря

Квантитативныеисследованиянелексическихединицязыка—грамматических классов (например, иерархий падежного маркирования), грамматических форм внутри парадигмы конкретного слова, вариативности грамматических

и лексико-грамматических единиц, вариативности падежного и предложно-падежного оформления зависимых — были признаны необходимой составляющей лингвистического анализа еще в мировой лингвистике 50–70-х годов XX в. В русистике были получены замечательные результаты в классических работах Штейнфельдт 1963, Greenberg 1974, Граудина и др. 1976, Апресян 1967 и мн. др.). Однако именно появление представительных и сбалансированных лингвистических корпусов объемом от ста миллионов словоупотреблений и выше поставило эти исследования на принципиально новый уровень, как в плане используемых математических статистических моделей и компьютерных технологий, так и в плане осмысления частотных результатов и их устойчивости.

В теоретической лингвистике частотные исследования приобрели особую актуальность в связи с постулированием usage-based model — модели языка, предполагающей, что частота употребления языковых единиц оказывает непосредственное влияние на их конструктивные свойства, статус в системе, вариативность и изменение в истории языка (Kemmer & Barlow 2000). Еще одна гипотеза — о семантической мотивированности грамматических явлений — верифицируется в ходе исследований, изучающих сдвиги частот форм в разных лексико-семантических классах (см. об этом Janda, Lyashevskaya 2011): например, предполагается, что преобладание форм императива несовершенного вида связано с семантическими и функциональными особенностями лексических единиц. В когнитивных исследованиях изучается также гипотеза о том, что возможности языковой памяти таковы, что в частотных фрагментах человек оперирует единицами, большими чем слово (pre-fabricated units). Поднимается и вопрос, оперирует ли человек лексемами, т. е. единицами абстрактного уровня, или же это порождение грамматической схоластики, и человек опирается в своем языковом опыте исключительно на словоформы (Newman 2008). Наконец, изучение грамматических частотных профилей в разных языках помогло бы извлечь новые факты для лингвистической типологии и истории развития языков.

В грамматике русского языка, и теоретической, и практической, традиционно большую роль играет вопрос о дефектных парадигмах, а также о вариативных формах словоизменения. Несмотря на получившую общее признание точку зрения о градуальности таких явлений, как, например, singularia и pluralia tantum, выявление ассоциированных с ними лексических единиц и описание их функционирования все еще нуждается в эмпирических данных. То же можно сказать и о проблематике появления, со-существования и исчезновения вариативных форм типа род. мн. *помидор/помидоров,* прош. ед. *стыл/стынул*, статусе «вторых» падежей и т. п.

В преподавании родного и иностранных языков знание о частотных фактах грамматики позволяет выстроить правильную последовательность изучения грамматических тем (например, порядок изучения падежей), соотнести грамматические категории с теми лексемами, при которых они чаще всего встречаются, изучать лексику в контексте (знать самые частотные сочетания), выбирать для образца тексты, наиболее подходящие по жанрово-стилевому признаку к изучаемой грамматической теме и т. п.

И, конечно, неоценимую роль играют частотные данные в разработках систем автоматической обработки текста. Особенно это стало очевидно в эпоху стремительного развития алгоритмов машинного обучения, построенных на вероятностях. Грамматические и сочетаемостные преференции слов учитываются в синтаксических парсерах, системах разрешения неоднозначности, средствах исправления орфографии, распознавания текста, в онтологических расширениях поисковых систем и др.

Несмотря на то, что задача построения частотной русской грамматики и фронтального изучения грамматической вариативности осознана и ставится в литературе (Мустайоки 1973, Baerman et al. 2010), в настоящее время не существует ни одного сколько-нибудь полного лексикографического ресурса, приближающего нас к этой цели. Ресурс на материале НКРЯ дает уникальную возможность ответить на многие исследовательские вопросы, исходя из современных возможностей корпусной лингвистики.

3. Общая температура по больнице, или почему не всегда помогает статистика падежей

Когда говорят о частотной грамматике языка, в первую очередь, имеют в виду соотношения частот частеречных классов, падежей и других грамматических категорий. Особенно популярна тема частотного распределения падежей — в работе Копотев 2008 цитируются три исследования, появившихся только в 1959-1961 гг., что касается настоящего времени, то, как показывает веб-поиск, аналогичные работы, построенные на разных текстовых выборках, плодятся с невиданной скоростью. Работа самого М. Копотева привлекает внимание к устойчивости частотных данных на больших корпусах (см. табл. 1). Его вывод — в том, что современные корпуса довольно хорошо согласуются друг с другом в оценке средней вероятности появления падежей, а различия кроются в жанровой принадлежности текстов.

Табл. 1. Частотное распределение шести падежей по данным (Копотев 2008)

	И	Р	Д	В	Т	П
□ НКРЯ	27,06	29,23	5,98	18,66	8,44	10,63
■ ХАНКО	24,30	32,62	5,50	17,73	8,08	11,78
▨ J. 1953	38,80	16,80	4,70	26,30	6,50	6,90
▬ Št. 1963	33,60	24,60	5,10	19,50	7,80	9,40

Однако, легко видеть, что принцип «выбирай родительный, если забыл — не ошибешься» может сыграть злую шутку со студентом РКИ, в случае, если ему нужно употребить слово *шепот*. Как показывает табл. 2[2], распределение

[2] Здесь и далее в таблицах приведены данные по корпусу со снятой лексико-грамматической омонимией НКРЯ.

частот падежей у некоторых существительных может разительно отличаться от средней картины.

Табл. 2. Частотный грамматический профиль лексем *шепот, поза, тропинка* (падежные формы ед. числа)

	И	Р	Д	В	Т	П	Всего (F.abs)
шепот	10,9%	3,7%	0,9%	8,3%	**75,6%**	0,6%	349
поза	15,9%	6,3%	0,8%	19,0%	4,0%	**54,0%**	126
тропинка	27,6%	2,0%	**52,0%**	5,1%	5,1%	8,2%	98

Дж. Гринбергу принадлежит наблюдение, что разные семантические группы должны иметь разную дистрибуцию падежей (как в предложных, так и в беспредложных употреблениях), иными словами, средние значения падежных показателей в группе имен абстрактных качеств (или имен частей тела, или названий мер) должны отличаться от средних значений по всему массиву лексики (Greenberg 1974/1991). Выбор русского языка как объекта исследования Гринберга был не случаен — именно в тот момент русский язык, один из немногих, располагал частотным списком форм падежей и предложно-падежных сочетаний имен существительных, входившим в состав замечательного частотного словаря Э. Штейнфельдт (Šteinfeldt 1963). Гринберг искал «волшебное» соотношение, которое позволяло бы отнести слово к тому или иному семантическому классу — и, естественно, не нашел его. Позднее его наблюдение было реинтерпретировано как семантически мотивированный сдвиг частот грамматических форм. Например, большую долю форм творительного падежа *шепотом* легко объяснить пересечением в семантике грамматической формы (творительный способа) и семантике лексемы (*шепот* как способ произнесения); форм предложного падежа *(в) позе* — связью между стативной семантикой существительного и семантикой локативной группы *в + S.loc*, наиболее типичном контекстном варианте употребления этого слова. Аналогичным образом, преобладание форм датива у существительного *тропинка* объясняется тем, что слова со значением траектории — идеальный лексический наполнитель предложной группы *по + S.dat*.

В работе (Janda, Lyashevskaya 2011) мы ввели понятие грамматического профиля лексемы — как инструмента для изучения семантических и функциональных причин девиаций грамматических форм. Исследование поведения форм вида, времени и наклонения, частности, показало предсказуемые частотные эффекты в разных клетках парадигмы: в императиве несовершенного вида — для глаголов привлечения внимания, вежливой просьбы, лексики, относящейся к культурному фрейму встречи гостей и т. п., ср. *раздевайтесь, садитесь, присоединяйтесь, закусывайте, закуривайте, ступайте, прощайте*; в инфинитиве совершенного вида — для глаголов, в которых заложена презумпция труднодостижимого результата (вследствие этого они часто употребляются в контексте глаголов попытки, модальных предикативов, в целевых придаточных и т. п., ср. *попытался/тяжело было/чтобы восполнить*) и т. п.

В исследовании (Kuznetsova 2013) вводятся классы типичных «женских» и типичных «мужских» глаголов — соотношение форм мужского и женского рода у глаголов типа *вышивать* и глаголов типа *надвинуть* будет разным.

На материале BNC С. Райс и Дж. Ньюман (Rice, Newman 2005, Newman 2008) сделали наблюдение, что разброс грамматического распределения может присутствовать и внутри лексических групп. Они показали, что даже близкие по смыслу слова, английские *think, know* и *mean,* могут иметь значительную диспропорцию форм времени, лица и числа, и назвали это явление “inflectional islands”. Объяснение этого явления кроется в индивидуальных семантических особенностях каждого глагола, в способности присоединять разные типы субъектов и т. п. В (Janda, Lyashevskaya 2011) указывается также большой вклад устойчивых конструкций в формирование тех или иных грамматических «флюсов» у индивидуальных лексем, ср. *мне плевать, мне наплевать, на чужой каравай рот не разевай, хоть залейся, поминай, как звали*.

Однако, наиболее удивительный факт русской лексической системы состоит в том, что почти не существует существительных, грамматический профиль которых соответствовал бы «среднему» профилю нарицательной лексики, глаголов со «средней» пропорцией форм времени-лица-числа и т. п. По сути, мы имеем дело со сложным наслаиванием семантических особенностей, сочетаемостных и конструктивных свойств, которые суммарно влияют на частотный выход.

4. Обработка корпусных данных

Основная часть словаря строится на данных 1900–2010 гг., в диахронической части привлекаются данные, начиная с 1800 г. Данные для «малого» словаря были собраны по корпусу со снятой лексико-грамматической омонимией (5,4 млн словоупотреблений, стандартная коллекция), для «большого» словаря — по основному, газетному, поэтическому и устному корпусу. Сбор осуществлялся с учетом функциональных стилей и жанров текста, а также с учетом времени создания.

Прежде всего, была собрана статистика по словоформам с лексико-грамматическим разбором (лемма, часть речи, словоизменительные характеристики)[3], разметкой лексико-семантического класса капитализации написания. Были также собраны 2- и 3-грамы, отражающие статистику предложно-падежных сочетаний существительных и местоимений.

Для «борьбы» с грамматической омонимией словоформ внутри парадигм и между парадигмами использовалась автоматически дизамбигуированная версия основного, газетного, поэтического и устного корпуса. Она была создана с применением двух программ — модуля на эвристиках и HMM-модуля, обученного на текстах снятого вручную корпуса. Небольшая часть данных дополнительно корректировалась вручную.

Особо отметим, что большую проблему для дизамбигуации представляют ингерентно-пересеченные парадигмы, например, парадигмы мужского

[3] Использовались стандартные соглашения словаря (Ляшевская, Шаров 2009).

и женского рода имени *рояль* или парадигмы прилагательных вида *запасной* и *запасный*. Устаревший вариант женского рода существительного распознается словарем лишь в формах, не предусмотренных в парадигме мужского рода (*роялью*), и тем самым, в словаре отражается искусственно дефектная парадигма. Пересеченные парадигмы прилагательных, различающихся лишь в именительном падеже, также разводятся плохо, поскольку модели дизамбигуации не предусматривают столь тонкой настройки, да и вручную в письменном корпусе далеко не всегда удается однозначно определить лексему. Такие точечные места в словаре, где информация может быть недостоверна по причине несовершенной дизамбигуации, снабжаются специальной пометой.

5. Виды частотной информации в словаре

Пользователь имеет возможность пользоваться двумя наборами данных. «Малый» словарь представляет наиболее аккуратные результаты в смысле разведения омонимов. Однако в корпусе со снятой вручную омонимией многие многие грамматические формы частотных лексем могут быть либо не представлены вообще, либо встречаются редко, и следовательно, не могут показать достоверное распределение форм. «Большой» словарь строится на корпусах НКРЯ, в десятки раз превосходящих «снятник», однако следует учитывать, что в некоторых зонах (например, в зоне противопоставления родительного и винительного падежа одушевленных существительных) информация в нем менее достоверна.

5.1. Грамматические категории

Пользователь может выбрать данные как по всем грамматическим формам парадигмы, так и по более крупным кластерам форм. Например, могут быть приведены суммарные данные по формам полных пассивных причастий (без учета признаков падежа, числа и рода), по четырем формам прошедшего времени глагола, по всем формам единственного VS множественного числа существительного. Информация о падежных распределениях существительных и местоимений дополнена сведениями о распределении предложных конструкций, в которых задействован тот или иной падеж. Кроме того, можно получить сопоставительные данные для написаний с прописной VS строчной буквы.

5.2. Омонимия и вариативность

Из всей парадигмы пользователю могут быть выданы сведения только об омонимичных формах (в т. ч. внутрипарадигматическая омонимия, ср.

солдат — им. ед. и род. мн., омонимия форм, принадлежащих разным парадигмам, ср. *заплыв* — формы имени существительного и глагола, см. Венцов, Касевич 2004). Предоставляются сведения о соотношении частот вариантов грамматических форм (например, *сильней* и *сильнее*, *дверями* и *дверьми*), так наз. «основных» и «вторых» падежей, различающихся на письме (ср. *без толка* и *без толку*), и других секундарных форм (ср. *сильней* и *посильней*).

5.3. Распределение по годам и жанрам

Информация об изменении грамматических профилей во времени дается в 10-летних интервалах; в газетном корпусе учитываются интервалы в 1 год. Пользователь может увидеть распределения в художественной прозе, в поэзии, в периодике, в бытовой, учебно-научной и т. п. сферах нехудожественной литературы, в электронной коммуникации, а также в устной непубличной речи.

5.4. Единицы измерения

Пользователь может выбрать один или несколько вариантов представления частотной информации:

- количество текстов корпуса, в которых встретились формы;
- абсолютная частота вхождений и размер корпуса;
- частота в ipm;
- иерархия форм у рассматриваемой единицы/класса вида

  Loc > Gen > Nom > Acc > Dat > Ins;

- процентное распределение (см. табл. 3) и попарное соотношение форм;
- квинтильное распределение каждой из форм, например, положение формы предложного падежа единственного числа слова *велосипед* в первой, второй… пятой порции списка, в котором представлены формы предложного падежа единственного числа всех существительных (а — самые редкие, д — самые частые, см. табл. 4).

Табл. 3. Профиль падежных форм лексемы *влияние*: абсолютное и относительное распределение

	И	**Р**	**Д**	**В**	**Т**	**П**	**Всего (F.abs)**
sg	98	128	29	170	137	14	576
pl	4	9	3	7	2	2	27
	И	**Р**	**Д**	**В**	**Т**	**П**	**Всего (%)**
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%

Табл. 4. Квинтильное распределение падежных форм ед. числа в группе имен транспортных средств

Лемма	И	Р	Д	В	Т	П	Всего (F.abs)
метро	а	д	г	а	а	д	185
корабль	д	в	б	б	а	в	231
грузовик	д	г	в	б	б	в	134
пароход	д	д	а	б	в	г	121
автомобиль	г	г	в	б	б	г	441
поезд	д	в	в	б	б	г	618
самолет	г	в	г	в	в	г	385
трамвай	г	б	в	г	в	г	198
лодка	г	в	б	г	б	г	280
вагон	а	г	г	в	а	д	473
велосипед	б	в	а	г	б	д	206
автобус	г	б	в	в	б	д	281
такси	в	а	б	д	а	д	174

Оговорим, что пользователь может выбрать разные методики расчета соотношений частот в парадигме, известных из литературы. За основу сравнения (100%) может быть принята вся парадигма (т. е. сумма всех частот грамматических форм), некоторая базовая часть (например, парадигма глагола за вычетом форм причастий и деепричастий), приоритетная форма (например, сумма форм прошедшего времени), а также доля употреблений двух форм относительно друг друга (например, отношение частоты форм женского рода к частоте форм мужского рода).

5.5. Сравнение лексем. Классы

Информация в словаре разнесена на несколько уровней. Первый уровень — индивидуальные грамматические профили лексем. На втором уровне даются сведения для крупных лексико-семантических классов (в классификации НКРЯ), например, для глаголов движения, имен инструментов и т. п. Третий уровень — распределение грамматических частот на уровне частеречного класса (словарь также дает справочную информацию о встречаемости самих частеречных классов, а также именных и глагольных грамматических категорий).

Таким образом, информация об индивидуальных лексемах может быть сопоставлена с данными по их лексико-семантическому классу и, шире, со средним грамматическим профилем части речи. Возможно сопоставление грамматических профилей нескольких лексем между собой.

6. Заключение

Словарь адресован, в первую очередь, исследователям русского словоизменения, грамматической семантики, тем, кто изучает вариативность грамматической нормы. Вместе с тем, нужно заметить, что «лексикоцентричный» подход, несмотря на ресурсоемкость и неплотность данных, может оправдывать себя и в автоматической обработке текста. В частности, в экспериментах (Данилова и др. 2013) показано, что учет лексического фактора позволяет повысить качество автоматической дизамбигуации лексико-грамматической омонимии на 3 %.

Электронная форма словаря позволяет постоянно совершенствовать его. Во-первых, планируется развивать функционал с учетом пожеланий пользователей, в частности, дополнить словарь модулем графического представления результатов, подключить внешние словари (словарь вариантов, словообразовательный и т. п.) и др. Во-вторых, будет совершенствоваться качество даных за счет улучшения дизамбигуации корпусных данных и работы с сообщениями пользователей об ошибках. В-третьих, увеличение объема словаря: включение новых лексических данных, добавление информации об авторе и т. п., — требует дополнительных исследований, поскольку работа с малыми частотами (sparse data) требует особой осторожности и особых техник.

Главный вопрос — в том, как интерпретировать полученные данные, каким образом переносить сведения о статистических вероятностях на другие текстовые корпуса и как научиться делать аккуратные выводы о функционировании грамматических форм в целом. Предлагаемый словарь — лишь первый опыт составления большого лексико-грамматического ресурса, и, соответственно, станет благодатным материалом для исследования достоверности корпусных данных. Безусловно, мы должны лучше понимать структуру выборок, как она связана с устойчивостью статистических данных, научиться балансировать выборки для разных временных срезов, провести множество экспериментов с полученным лексическим материалом для того, чтобы достоверность интерпретации корпусных данных перестала вызывать вопросы.

Литература

1. *Baerman M., Brown D., Corbett G. G., Krasovitsky A., Williams P.* (2010), Predicate agreement in Russian: A corpus-base approach, Wiener Slawistischer Almanach, Sonderband 74, pp.109–121.
2. *Greenberg J. H.* (1974/1990), The relation of frequency to semantic feature in a case language (Russian), in Denning K., Kemmer S. (eds), On Language, Selected Writings of Joseph H. Greenberg, Stanford, pp. 207–226.
3. *Ilola E., Mustajoki A.* (1989), Report on Russian Morphology as it Appears in Zaliznyak's Grammatical Dictionary, (Slavica Helsingiensia 7), Helsinki.
4. *Janda L. A., Lyashevskaya O.* (2011), Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian, Cognitive Linguistics, 22 (4), pp. 719–763.
5. *Kemmer S., Barlow M.* (2000), A Usage-Based Conception of Language, Essen, 2000.
6. *Kuznetsova J.* (2013), Linguistic Profiles: Correlations between Form and Meaning. Ph.D. diss., University of Tromsø.
7. *Newman J.* (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.
8. *Rice S., Newman J.* (2005), Inflectional islands, ICLC-9, Yonsei University, Seoul, Korea.
9. *Šteinfeldt E.* (1963), Russian Word Count, Moscow.
10. *Апресян Ю. Д.* (1967), Экспериментальное исследование семантики русского глагола, М.
11. *Венцов А. В., Касевич В. Б.* (ред.) (2004), Словарь омографов русского языка, СПб.: Филологич. ф-т СПбГУ.
12. *Граудина Л. К., Ицкович В. А., Катлинская Л. П.* (1976), Грамматическая правильность русской речи. Стилистический словарь вариантов. М.
13. *Данилова В., Волков О., Ладыгина А., Привознов Д., Сербинова И., Сим Г.* (2013). Снятие омонимии методом НММ (рукопись).
14. *Копотев М.* (2008), К построению частотной грамматики русского языка: падежная система по корпусным данным // Мустайоки А., Копотев М. В., Бирюлин Л. А., Протасова Е. Ю. (ред.), Инструментарий русистики: корпусные подходы, Хельсинки.
15. *Ляшевская О. Н., Шаров С. А.* (2009), Частотный словарь современного русского языка (на материале Национального корпуса русского языка), М.: Азбуковник.
16. *Мустайоки А.* (1973), Опыт составления частотной грамматики русских существительных, Хельсинки, (рукопись).

References

1. *Apresjan Ju. D.* (1967), Experimental research on the semantics of the Russian verb [Eksperimental'noe issledovanie semantiki russkogo glagola], Moscow.
2. *Baerman M., Brown D., Corbett G. G., Krasovitsky A., Williams P.* (2010), Predicate agreement in Russian: A corpus-base approach, Wiener Slawistischer Almanach, Sonderband 74, pp. 109–121.
3. *Danilova V., Volkov O., Ladygina A., Privoznov D., Serbinova I., Sim G.* (2013). Disambiguation with HMM [Snjatie omonimii metodom HMM] (manuscript).
4. *Graudina L. K., Ickovich V. A., Katlinskaja L. P.* (1976), Correct Russian speech: Stylistical dictionary of grammatical choices [Grammaticheskaja pravil'nost' russkoy rechi. Stilisticheskiy slovar' variantov]. Moscow.
5. *Greenberg J. H.* (1974/1990), The relation of frequency to semantic feature in a case language (Russian), in Denning K., Kemmer S. (eds), On Language, Selected Writings of Joseph H. Greenberg, Stanford, pp. 207–226.
6. *Ilola E., Mustajoki A.* (1989), Report on Russian Morphology as it Appears in Zaliznyak's Grammatical Dictionary, (Slavica Helsingiensia 7), Helsinki.
7. *Janda L. A., Lyashevskaya O.* (2011), Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian, Cognitive Linguistics, 22 (4), pp. 719–763.
8. *Kemmer S., Barlow M.* (2000), A Usage-Based Conception of Language, Essen, 2000.
9. *Kopotev M.* (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.
10. *Kuznetsova J.* (2013), Linguistic Profiles: Correlations between Form and Meaning. Ph.D. diss., University of Tromsø.
11. *Lyashevskaya O. N., Sharoff S. A.* (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.
12. *Mustajoki A.* (1973), On compiling the frequency dictionary of Russian nouns [Opyt sostavlenija chastotnoy grammatiki russkikh suschestvitel'nykh], Hel'sinki, (manuscript).
13. *Newman J.* (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.
14. *Rice S., Newman J.* (2005), Inflectional islands, ICLC-9, Yonsei University, Seoul, Korea.
15. *Šteinfeldt E.* (1963), Russian Word Count, Moscow.
16. *Ventsov A. V., Kasevich V. B.* (eds.) (2004), Dictionary of Russian homographs [Slovar' omografov russkogo jazyka], St.-Petersburg.