Slavica Helsingiensia 34
Инструментарий русистики: корпусные подходы, Хельсинки, 2008
Под ред.: А. Мустайоки, М.В. Копотева, Л.А. Бирюлина, Е.Ю. Протасовой

М.В. Копотев
(Хельсинки)

К построению частотной грамматики русского языка: падежная система по корпусным данным[1]

Введение

Каждый новый этап в развитии машинной обработки языкового материала открывает новые возможности как для создателей корпусов, так и для лингвистов, осуществляющих корпусные исследования. Появление электронных собраний текстов открыло возможности к более точному описанию лексики, мониторингу словарного состава, к созданию частотных словарей и индексов. Последнее является одним из самых распространенных приложений компьютерных исследований и тесно связано с развитием обработки и представления языковых данных. Если говорить о русском языке, то уже довольно давно появились словари, представляющие частотные характеристики лексем (Šteinfeldt 1963; Засорина 1977; Лёнгрен 1999); существуют исследования, представляющие частотные распределения различных классов слов (см. Браславский 2001; Шаров). Все они так или иначе основаны на лемматизации (то есть на автоматическом сведении словоформ к начальной форме), и — реже — на автоматической частеречной разметке языковых единиц.

При безусловной ценности таких частотных словарей все же следует согласиться с мнением Н.В. Котовой:

> *Фактически лексемные частотные словари информируют не обо всей языковой материи, не обо всем ее многообразии, а о многообразии языковой материи некоего аморфного (аграмматического) трансформата, некоего «китаизированного» варианта описываемого языка («китаизированного» — в смысле традиционного мнения лингвистов не синологов о китайском языке как о типичном представителе не содержащих грамматических морфем «корневых» языков»)* (Котова 2004: 160).

В настоящее время, однако, появляется возможность для создания частотных индексов, основанным не только на лексемном уровне. Следуя за развитием методов автоматической обработки языка, русская корпусная лингвистика в настоящий момент предлагает достаточно надежные

морфологически аннотированные корпуса, появляются ресурсы, содержащие синтаксическую и семантическую разметку. Все это позволяет поставить задачу по созданию индекса, фиксирующего частотные характеристики грамматических категорий в текстах. Задача построения таких индексов для любого языка важна по целому ряду причин. Обозначу лишь некоторые из них.

1. Еще в начале 70-х годов Дж. Гринбергом была высказана гипотеза, согласно которой различные семантические классы предпочитают различные морфологические формы.

> *It is not merely in regard to the differential frequency, with which specific semantically defined subclasses of nouns occur in particular cases that the monolithic appearance of case system is illusory <…> the occurrence in one case rather that another may disambiguate several meanings which are listed by lexicographers as distinct but related and often such relations are recurrent and systematic in the language* (Greenberg 1974 / 1990: 208).

На материале словаря (Šteinfeldt 1963) Дж Гринберг показал, что русские существительные разных семантических классов с разной вероятностью появляются в том или ином падеже. Позже сходные выводы о поведении числовых форм существительного были сделаны Б.Ю. Норманом (2003) и другими исследователями (на материале других языков см., например, (Halliday 1993 / 2005, Arppe 2006)). В современной лингвистике эти наблюдения были обобщены в рамках модели языка, основанной на употреблении (*Usage-based model*), которая исходит, в частотности, из того, что частотность употребления языковой единицы имеет прямое влияние на ее статус в системе и на ее возможные модификации.

> *Because the system is largely an experience-driven one, frequency of instances is a prime factor in its structure and operation. Since frequency of a particular usage pattern is both the result and a shaping force of the system, frequency has an indispensable role in any explanatory account of language* (Kemmer & Barlow 2000: 4; ср. еще «Вероятностную грамматику» М. Халлидея (Halliday 1993 / 2005)).

2. Знание частотных характеристик морфологических категорий может быть использовано для создания учебных пособий, ориентированных на реальное языковое употребление, а не на изучение абстрактной языковой системы. Конечно, в большинстве современных учебных курсов этот

фактор так или иначе учитывается, однако чаще всего решения опираются на интуицию учителя или автора учебной грамматики[2].

> *<F>requency should also play a key role in the development of materials and in the choices that teachers make in language classrooms. With the recent availability of comprehensive frequency-based grammatical descriptions, such integration of pedagogy and research has become feasible* (Biber & Reppen 2002: 206-207).

3. Наконец, частотные характеристики грамматических категорий важны для автоматической обработки языка. Это позволяет использовать в программах-анализаторах вероятностные алгоритмы снятия лексической неоднозначности или выбора синонимов, основанные на учете частотности той или иной грамматической формы (Karlsson 1986; Arppe 2002).

Частотная грамматика русского языка

Создание полной частотной грамматики русского языка казалось до недавнего времени трудновыполнимой задачей, хотя существуют многочисленные исследования, решающие ее на ограниченном материале[3].

1950-60-ые годы можно охарактеризовать как время всплеска интереса к этой теме. По-видимому, первым опытом такого рода в русистике стало исследование Г. Йоссельсона (Josselson 1953). Это издание, являющиеся одним из первых частотных словарей русского языка, содержит сведения об употребительности морфологических категорий, подсчитанных на выборке в 46 896 текстоформ, извлеченных из текстов XIX-XX вв. В работе представлены данные по восьми грамматическим классам (части речи, падежи существительных, степени сравнения прилагательных, вид глагола, наклонения глаголы, временные формы индикатива, лицо в глагольных формах, окончания прошедшего времени индикатива) (Josselson 1953: 11-25).

Еще одним значительным для своего времени исследованием стала работа, выполненная эстонскими лингвистами под руководством Э. Штейнфельдт (Šteinfeldt 1963). В этой работе на материале значительной по тем временам выборки (400 000 текстоформ) даны не только сведения о частотности лексем, но и соответствующие характеристики частей речи; глагольных, субстантивных и местоименных категорий; сведения о предложных сочетаниях и глагольном управлении. Материалы исследований Г. Йоссельсона и Э. Штейнфельдт используются в эксперименте, описанном в настоящей статье.

В эти же годы появилось достаточно много работ, исследующих частотность частей речи и отдельных грамматических категорий (см., например, Никонов 1959; Николаев 1960; Волоцкая & Шелимова 1961;

Белоусова 1964). Большинство из них выполнено на небольшой и чаще всего тематически ограниченной выборке.

В начале 1970-х годов в работе А. Мустайоки (1973) была впервые теоретически обоснована (и частично решена) задача построения общей частотной грамматики русского языка. Прежде всего автор проводит разделение парадигматической и синтагматической вероятности появления языкового явления. Парадигматическая вероятность задана системой и легко определяется при наличии приемлемой классификации. Так, например, в рамках шестичленной падежной парадигмы вероятность выбора каждого падежа равна 1/6. В то же время очевидно, что в реальном тексте частотность появления именительного, например, падежа несопоставимо выше, чем предложного. Эту вероятность исследователь называет «синтагматической», и ее решение, безусловно, требует обращения к некоторому массиву текстов. В указанной работе эта задача решается на материале 5054 существительных, извлеченных из статей одного номера газеты «Комсомольская правда». Главным достоинством работы можно считать методологическое обоснование построения грамматики нового типа, которую А. Мустайоки называет «частотной грамматикой»:

> *Систематическое описание грамматических явлений какого-то (подъ)языка мы называем частотной грамматикой. Она отличается от традиционной грамматики тем, что в ней дается количественная информация о всех категориях и значениях* (Мустайоки 1973: 30).

Программа, заявленная в исследовании, была продолжена в монографии (Ilola & Mustajoki 1989). В этой работе представлены данные о лексической частотности — распределении словоизменительных категорий по лексемам русского языка. По сути, это исследование, опираясь на Грамматический словарь русского языка (Зализяк 1977), отвечает на вопросы типа «сколько лексем мужского рода в русском языке?» и т.п.[4].

В общем и целом, определение частотных характеристик грамматических категорий и классов русского языка является задачей, которая решается уже несколько десятилетий, однако до сих пор не решена. Среди главных проблем, с которыми сталкивались исследователи, можно назвать малый объем выборки (нерепрезентативность), трудоемкость и фрагментарность исследований. В настоящее время, при наличии современных морфологически размеченных корпусов эта задача может быть решена с большей эффективностью. Хорошо аннотированный корпус позволяет получать детализированные квантитативные данные, отражающие частотные распределения различных грамматических классов в различных группах текстов.

В настоящее статье анализируются результаты экспериментов, выполненных на материале НКРЯ и ХАНКО, а также делаются предварительные замечания о подготовке такой грамматики. Эксперименты, представленные в работе, позволяют ответить на следующие вопросы:

– Позволяют ли существующие корпуса (прежде всего НКРЯ) решить такую задачу?
– Насколько точны и представительны корпусные данные?

Для решения этих задач в качестве экспериментального объекта была выбрана падежная система русского языка. Приведенные в работе данные, конечно, имеют ценность и сами по себе, поскольку демонстрируют количественный срез одного фрагмента русской морфологии.

В следующей части статьи представлены результаты сравнения двух сопоставимых по объему выборок, извлеченных из двух корпусов — НКРЯ и ХАНКО. Кроме того, часть данных сравнивается с результатами, полученными другими исследователями. Все это позволяет сделать вывод о качестве используемого материала. Заключительная часть работы являет собой фрагмент русской частотной грамматики — падежной системы русского языка, как она представлена в НКРЯ.

Падежная система русского языка

Общепринятого списка русских падежей до сих пор не существует. После публикаций работ А.А. Зализняка (1967, 1977) частью исследователей признается восьмичленная система русских падежей (6 основных и два «неполных», по определению А.А. Зализняка, к которым часто добавляют «звательную форму», превращая систему в девятичленную). В то же время существуют и альтернативные классификации. Так, сам А.А. Зализняк в принципе признает 14 падежей (1967:52-55), считая, впрочем, их выделение нецелесообразным. Значительная часть исследователей вслед за Русской грамматикой (1980 I: 475-507) соглашается с традиционной шестичленной системой. В исследовании, основанном на корпусных данных, приходится опираться на классификации, положенные в основу соответствующих корпусов, если это не приводит к явному разрыву с традицией. Это тем более оправданно, когда речь идет о русской машинной морфологии, поскольку она в значительной степени базируются на Грамматическом словаре А.А. Зализняка[5].

В НКРЯ представлена одиннадцатичленная падежная система:

1) именительный падеж (*голова, сын, степь, сани, который*);
2) родительный падеж (*головы, сына, степи, саней, которого*);
3) дательный падеж (*голове, сыну, степи, саням, которому*);
4) винительный падеж (*голову, сына, степь, сани, который / которого*);
5) творительный падеж (*головой, сыном, степью, санями, которым*);

6) предложный падеж ((*о*) *голове, сыне, степи, санях, котором*);
7) второй родительный падеж (*чашка* ***чаю***);
8) второй винительный падеж (*постричься в* ***монахи****; по* ***два*** *человека*);
9) второй предложный падеж (*на* ***оси́****, в* ***лесу***);
10) звательная форма (*Господи, Серёж, ребят*);
11) счётная форма (*два* ***часа́****, три* ***шара́***)[6].

Разработчики ХАНКО исходили из девятичленной падежной системы, однако на практике она представлена восемью падежами, поскольку звательная форма в текстах корпуса ни разу не встретилась (И, Р, Д, В, Т, П, Р2, П2).

Для корректного сравнения результатов падежные системы НКРЯ и ХАНКО были сведены в девятичленную систему, про этом:

– два винительных в НКРЯ считались одним падежом;
– счетная форма в НКРЯ считалась родительным падежом;
– Звательная форма учитывалась, при этом считалось, что в ХАНКО количество употреблений равно нулю.

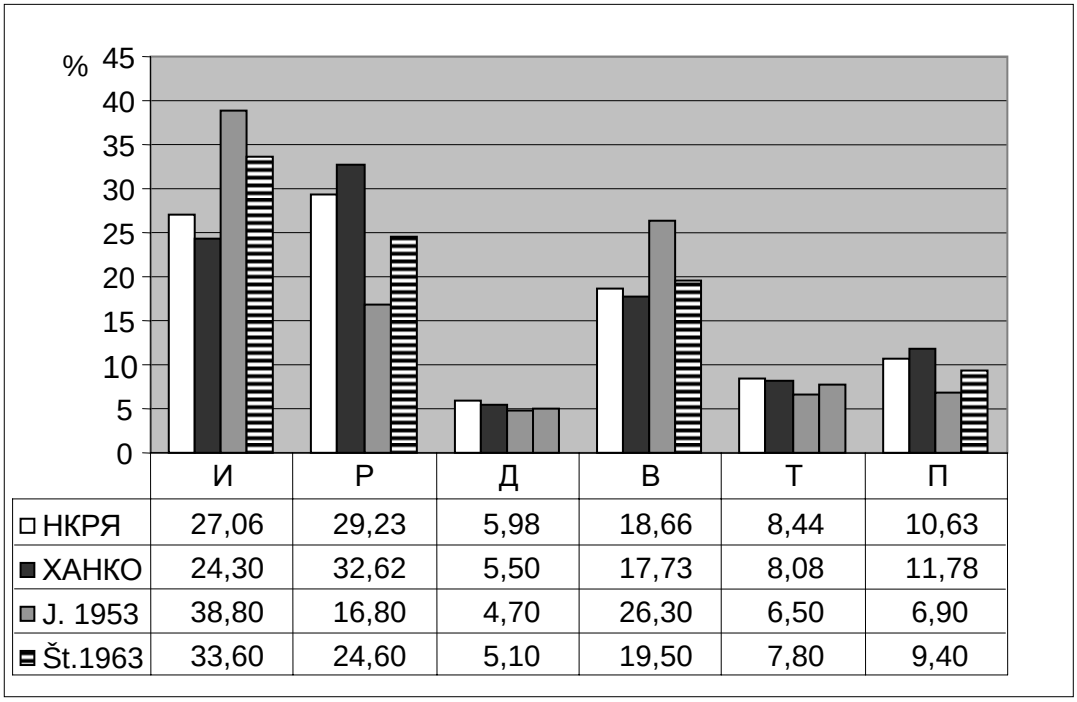
В то же время в заключительной части работы, представляющей данные только НКРЯ, схема аннотирования не модифицировалась, и числовые данные отражают, таким образом, употребление всех 11 падежей, выделенных в Национальном корпусе русского языка.

Для проведения эксперимента были составлены выборки, сопоставимые по объему (около 100 тыс. текстоформ), хронологии (1999-2001) и сфере функционирования (публицистика). Эксперимент позволил сравнить результаты и решить вопрос об аккуратности морфологического аннотирования и возможности использования результатов всего подкорпуса текстов со снятой омонимии в НКРЯ. Сравнивались доли падежей разных частей речи и общая доля падежей в выборке.

1. Доли падежных форм для имени существительного.

Поскольку существительное (и падежные формы в особенности) неоднократно служили объектом статистических подсчетов, данные корпусов сравниваются с результатами, полученными другими исследователями. Для этого был составлен график, отражающий помимо корпусных данных результаты Г. Йоссельсона (Josselson 1953: 18-20) и Э. Штейнфельдт (Šteinfeldt 1963: 35)[7].

Диаграмма 1 дает возможность сравнить результаты выборок[8].



	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

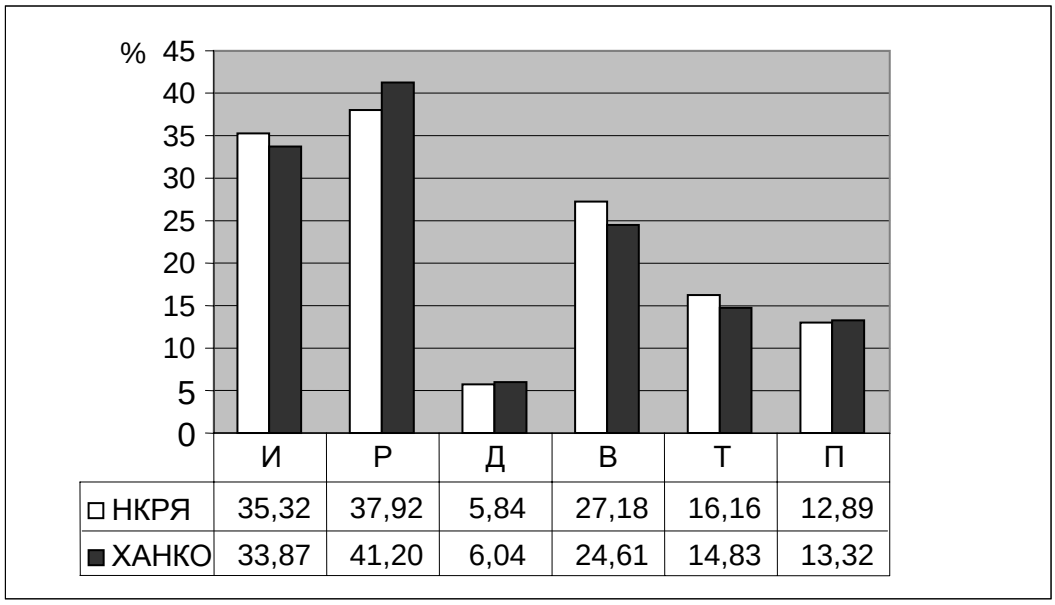
Диагр. 1

На графике видно, что результаты по разным выборкам существенно различаются. Для выявления возможной погрешности было подсчитано стандартное статистическое отклонение для данных по каждому падежу. При этом считалось, что каждая выборка так или иначе отражает генеральную совокупность текстов на русском языке. Величина стандартного отклонения позволяет определить, насколько данные конкретной выборки отличаются от средних совокупных[9]. В результате выяснилось, что данные из (Josselson 1953) не попадают в диапазон стандартного отклонения ни в одном случае[10]. Данные (Šteinfeldt 1963) полностью укладываются в диапазон статистических колебаний. Данные НКРЯ и ХАНКО входят в статистически оправданный диапазон колебаний в 5 из 6 случаев. При этом статистически некорректные данные лишь незначительно отличаются от допустимых:

– Дательный падеж в НКРЯ: на 1, 95 %;
– Именительный падеж в ХАНКО: на 0,55 %.

Обращает на себя внимание и тот факт, что данные корпусов в целом демонстрируют большую гомогенность при значительной разнице с данными других исследований. По всей видимости, на результаты влияет разница в объеме выборок, разножанровый набор источников, а возможно и точность обработки материала. В целом, можно сказать, что используемые корпуса представляют сопоставимые и взвешенные данные.

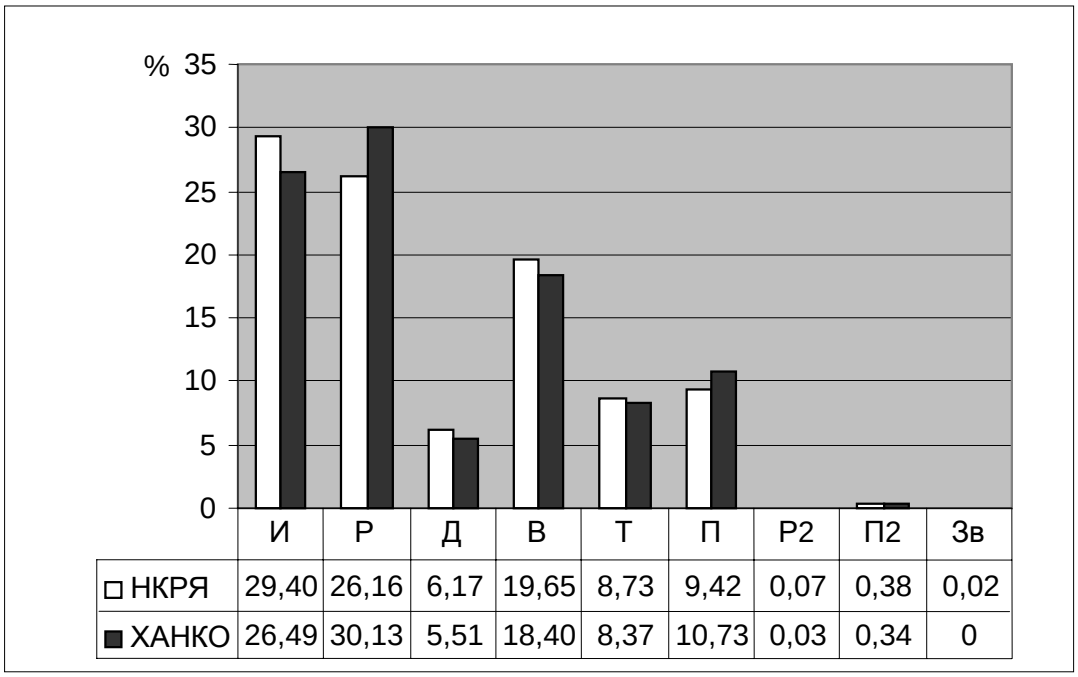
3. Доли падежных форм для имени прилагательного представлены на Диаграмме 2.



	И	Р	Д	В	Т	П
НКРЯ	35,32	37,92	5,84	27,18	16,16	12,89
ХАНКО	33,87	41,20	6,04	24,61	14,83	13,32

Диагр. 2

4. Общая доля падежей для всех частей речи отражена на Диаграмме 3.



	И	Р	Д	В	Т	П	Р2	П2	Зв
НКРЯ	29,40	26,16	6,17	19,65	8,73	9,42	0,07	0,38	0,02
ХАНКО	26,49	30,13	5,51	18,40	8,37	10,73	0,03	0,34	0

Диагр. 3

В целом, сопоставимые выборки НКРЯ и ХАНКО дают сопоставимые результаты. Разница, не превышающая 4 %, наблюдается только в данных для именительного и родительного падежей[11]. Вероятным объяснением этого является различие в обработке омонимичных пар И—В и

В—Р падежей. По данным А. Пайле (2003: 42-43), основные группы грамматических омонимов образуют именно эти формы. Так, наиболее часто встречающаяся морфологическая многозначность — это омонимия И—В и в единственном, и во множественном числе (*стол*, *столы*), на втором месте по частоте встречаемости многозначность В—Р (*котов*).

Решая вопрос о степени точности разметки в обоих корпусах, необходимо признать, что ошибки морфологической разметки встречаются и в ХАНКО, и в НКРЯ. И все же НКРЯ содержит большее количество неточностей, что объясняется и большим объемом обработанного материала в НКРЯ, и большим вниманием к ручному снятию омонимии в ХАНКО.

Так, в НКРЯ уже второе предложение в результатах по запросу «именительный падеж существительного» содержит ошибку аннотирования:

> *Вновь обрётший политическое* ***могущество*** *Анатолий Чубайс, самым активным образом участвовавший в интриге по снятию Евгения Примакова и ещё более активно — в интриге по назначению на его место Сергея Степашина, не был приглашён к столу, за которым делили наследство уволенного премьера* (Александр Андреев, «Как обманули Чубайса», Коммерсантъ Власть, № 20, 1999.05.25).

В то же время неточность обработки материала в морфологической части ХАНКО не превышает 4%[12]. Насколько мне известно, не существует оценок точности НКРЯ, однако надо признать, что процент ошибок в нем выше, чем в ХАНКО[13].

Таким образом, эксперимент показал, что несмотря на некоторый процент ошибок в «зонах повышенной омонимии», подкорпус со снятой омонимией в НКРЯ заслуживает доверия. Представляется, что при всех возможных неточностях составленная на его основе частотная грамматика может служить опорой в исследованиях. При этом нелишними в частотной грамматике, составленной на его основе, будут указания на данные «контрольной выборки», за которую можно условно признать ХАНКО. Это позволит исследователям самостоятельно оценивать вероятность статистической погрешности.

В следующей части статьи будет представлен фрагмент такой грамматики и обсуждены связанные с этим проблемы.

Распределение падежей по частям речи в НКРЯ

В настоящее время на сайте НКРЯ существует возможность получать точные данные о количестве найденных единиц в выборке любого размера. Это позволяет в полной мере использовать статистические методы, так необходимые при работе с большим массивом текстов. Ниже

представлены диаграммы с указанием доли падежных форм в НКРЯ, рассчитанные для:

– распределения падежей по частям речи;

– распределения падежей по жанрам.

Диаграмма 4 (стр. 146) представляет распределение падежей по частями речи. При этом за 100 процентов принимается общее количество словоформ соответствующей части речи. На диаграмме видно, что частотное распределение падежей для разных частей речи несколько различается. Более того, И, Р, В не всегда входят в тройку самых частотных падежей.

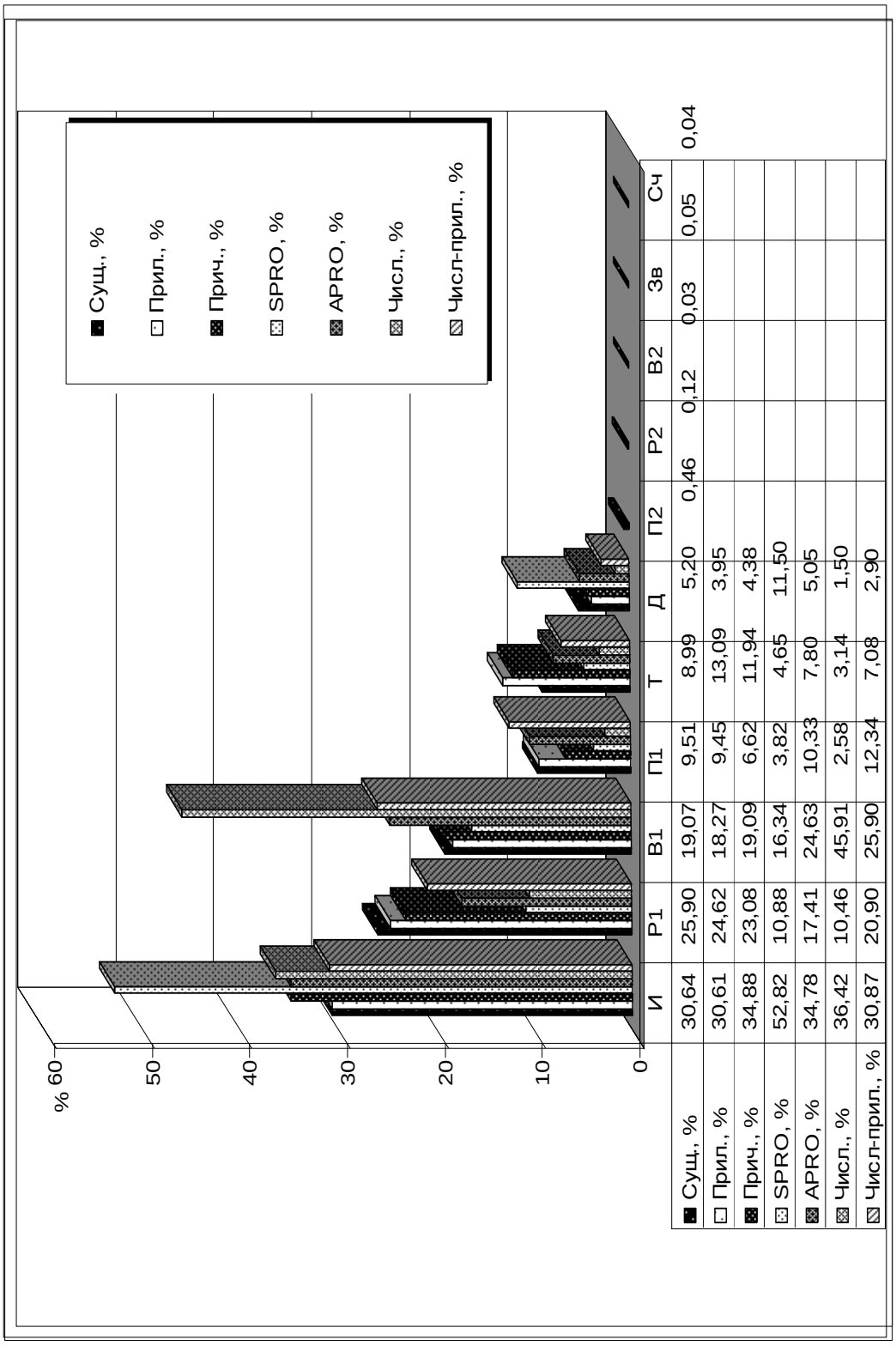
Существительные:	И—Р—В—Т—П—Д—П2—Р2—В2—Зв—Сч
Прилагательные:	И—Р—В—Т—П—Д
Причастия:	И—Р—В—Т—П—Д
Местоимения-сущ.:	И—В—Д—Р—Т—П
Местоимения-прил.:	И—В—Р—П—Т—Д
Числительные:	В—И—Р—Т—П—Д
Числительные-прил.:	И—В—Р—П—Т—Д

Распределения падежей по жанрам в НКРЯ

На диаграмме 5 (стр. 147) отражено распределение падежей по некоторым жанрам, представленным в корпусе. Для подсчета были выбраны жанровые группы, хорошо представленные в корпусе, и этот материал позволяет сделать выводы о разнице в распределении грамматических форм в разных типах текстов. Анализ такого рода выходит за рамки настоящей статьи, но совершенно очевидно, почему в учебно-научных текстах так велика доля родительного. Ясно и то, чем вызван рост частотности звательной формы в церковно-богословских текстах. Уже эти сведения позволяют, среди прочего, точнее описывать особенности грамматической стилистики, аккуратнее подбирать материал для грамматических занятий для различных групп студентов и т.п.

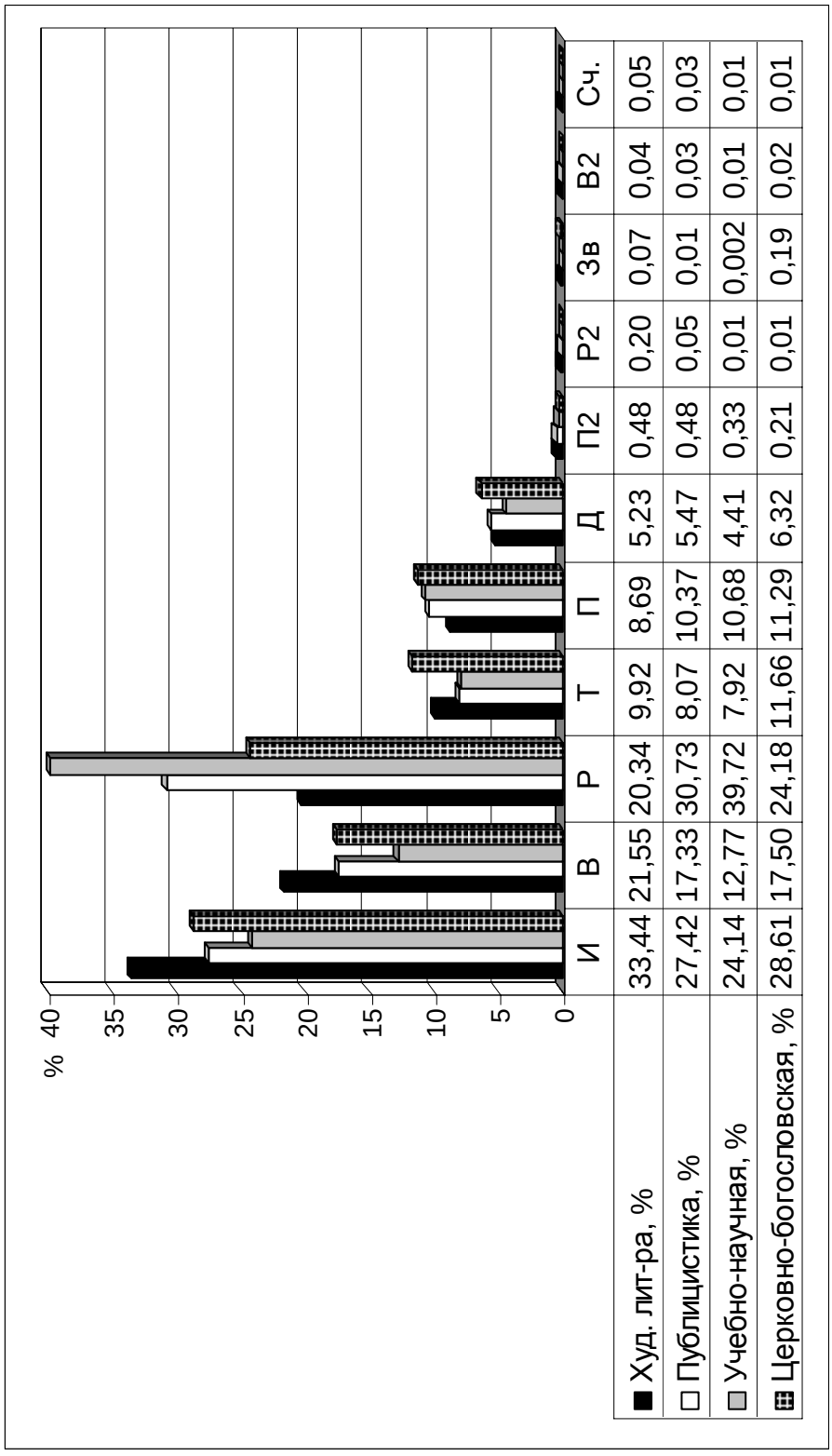
Порядок падежей для жанров таков:

Художественная литература:	И—В—Р—Т—П—Д—П2—Р2—Зв—В2—Сч
Публицистика:	И—Р—В—П—Т—Д—П2—Р2—В2—Сч—Зв
Учебно-научная:	Р—И—В—П—Т—Д—П2—Р2—В2 / Сч—Зв
Церковно-богословская:	И—Р—В—Т—П—Д—П2—Зв—В2—Сч—Р2



	И	Р1	В1	П1	Т	Д	П2	Р2	В2	Зв	Сч	0,04
Сущ., %	30,64	25,90	19,07	9,51	8,99	5,20	0,46	0,12	0,03	0,05		
Прил., %	30,61	24,62	18,27	9,45	13,09	3,95						
Прич., %	34,88	23,08	19,09	6,62	11,94	4,38						
SPRO, %	52,82	10,88	16,34	3,82	4,65	11,50						
APRO, %	34,78	17,41	24,63	10,33	7,80	5,05						
Числ., %	36,42	10,46	45,91	2,58	3,14	1,50						
Числ-прил., %	30,87	20,90	25,90	12,34	7,08	2,90						

Диагр. 4. Распределение падежей по частям речи в НКРЯ



	И	В	Р	Т	П	Д	П2	Р2	Зв	В2	Сч.
Худ. лит-ра, %	33,44	21,55	20,34	9,92	8,69	5,23	0,48	0,20	0,07	0,04	0,05
Публицистика, %	27,42	17,33	30,73	8,07	10,37	5,47	0,48	0,05	0,01	0,03	0,03
Учебно-научная, %	24,14	12,77	39,72	7,92	10,68	4,41	0,33	0,01	0,002	0,01	0,01
Церковно-богословская, %	28,61	17,50	24,18	11,66	11,29	6,32	0,21	0,01	0,19	0,02	0,01

Диагр. 5. Распределение падежей по жанрам в НКРЯ

Заключение

1. Традиционный порядок русских падежей складывался путем постепенной модификации латинской шестичленной системы в грамматиках М. Смотрицкого, М. В. Ломоносова и др. Неоднократно предлагались другие формы подачи русских падежей как с точки зрения структурных принципов, так и с точки зрения методики преподавания (Chvany 1996; Распопова 2001). Однако если исходить из представленного в статье материала, универсального порядка падежей не существует. Разные классы слов в разных жанрах имеют разные ранги, так что общую частотную иерархию падежной системы невозможно составить. Тем не менее, можно отметить, что порядок **И—Р—В—Т—П—Д—П2—Р2—Зв—Сч—В2** является наиболее нейтральным, с точки зрения статистки. При этом именительный падеж является самым частотным практически во всех выборках. Р—В падежи образуют пару, чаще всего занимающую второю-третью позиции. Д—Т—П образуют третью группу, непосредственно примыкающую ко второй, порядок внутри которой существенно зависит от части речи и, в меньшей степени, от жанра. При этом дательный падеж местоимений-существительных даже опережает винительный. Остальные падежи заметно отстают по частотности употребления даже среди существительных. В этой группе однозначного порядка распределения установить не удается, хотя чаще всего порядок таков: П2—Р2—В2—Сч—Зв. В целом, значительная разница в частотности между шестью «основными» падежами и остальными позволяет считать последние периферийными и в ряде случаев полностью игнорировать их (например, в преподавании русского в школе или в иностранной аудитории).

2. Сопоставление выборок из двух независимых корпусов показало, что полученные данные в целом корректны и позволяют с достаточной точностью получать сведения о частотности грамматических категорий. Сравнение со сделанными ранее статистическими подсчетами демонстрируют совпадение корпусных данных при серьезной разнице с данными, полученными представленными в некоторых исследованиях 1950-60-ых годов. Особенно значительно отличаются от средних данные Г. Йоссельсона.

3. Представляется, что и эксперимент, и исследование на большой выборке демонстрируют возможности создания на основе существующих корпусов частотной грамматики русского языка. При создании такой грамматики целесообразно учитывать следующее.

В основу классификации частотной грамматики должны быть положены принципы, считающиеся общепринятыми в научном сообществе. В то же время необходимо учесть существующую в корпусе разметку. Классификация морфологической системы должна представлять собой компромисс между лингвистической корректностью и возможностями

автоматического поиска. Так например, формы сослагательного наклонения должны быть каким-то образом учтены, несмотря на то, что в разметке НКРЯ они не учитываются и все глагольные формы на *-л* размечены как формы прошедшего времени; то же касается и аналитических форм и лексем (см подробнее Мустайоки & Копотев 2004).

Объем ХАНКО мал и непредставителен для того, чтобы делать выводы о современном русском языке во всем многообразии жанров и типов текстов. Подсчет целесообразно проводить на основе самого представительного на сегодняшний день корпуса — НКРЯ. Процент ошибок в этом корпусе высок только в «зонах повышенной омонимии», то есть в тех частях грамматической системы, которые наиболее трудны для автоматического анализа. Для оценки ошибок аннотирования целесообразно проводить сравнение сопоставимых выборок НКРЯ и ХАНКО, что даст исследователям возможность самостоятельно оценить разницу и решить вопрос о приемлемости результатов.

Будет ли частотная грамматика русского языка в таком виде востребована исследователями и преподавателями? По всей видимости, да. Во-первых, значительная часть материала НКРЯ содержит в целом корректно представленную морфологическую информацию. Кроме того, сравнение данных НКРЯ и ХАНКО даст возможность оценить степень точности материала. Во-вторых, корпусная лингвистика не стоит на месте, и разработка методики составления частотной грамматики, эксперименты в этой области позволят избежать ошибок в будущем.

Примечания

[1] За ценные комментарии к черновой версии этой статьи благодарю проф. Б.Ю. Нормана и рецензентов сборника.

[2] В то же время стоит напомнить, что частотность лексем не может являться единственным фактором отбора материала в учебных целях (см. Мустайоки 1986).

[3] Отмечу для сравнения, что для английского языка (имеющего наиболее богатую корпусную традицию) такие исследования давно существуют, см. (Francis & Kucera 1982; Johansson & Hofland 1989).

[4] Например, в современном русском языке большинство существительных принадлежит к мужскому роду (45.8%), тогда как на женский и средний приходится 38,5% и 13,5% соответственно (Ilola & Mustajoki 1989: 9).

[5] О русской компьютерной морфологии см. (Коваль 2005).

[6] Далее соответственно И, Р, Д, В, Т, П, Р2, В2, П2, Зв, Сч.

[7] В следующей таблице кратко представлены характеристики материала.

J 1953	(Josselson 1953)	46 896 форм	Художественная литература и литературная критика, XIX, XX вв.
Št 1963	(Šteinfeldt 1963)	400 000 форм	Художественная литература, публицистика, транскрипты радиопередач, XX в.

В работе не учитывались данные других исследований, прежде всего (Николаев 1960; Волоцкая 1961; Мустайоки 1973), поскольку эти выборки ограничены и жанрово, и по объему.

[8] Для адекватного сравнения классификации все данные были сведены к шестичленной системе, без разделения форм множественного / единственного чисел. При этом вторые Р2 и Д2 были включены в Р и Д соответственно, Сч. форма включена в В.

[9] Величина стандартного отклонения рассчитывалась по формуле $\sqrt{\frac{\sum (x-\bar{x})^2}{(n-1)}}$, где $(x-\bar{x})^2$, выборочное среднее, а n — число выборок (см. подробнее (Головин 1971: 19-28)).

[10] Исследование Г. Йоссельсона вызывало сомнения и в точности подсчета частотности частей речи (Марков 1960).

[11] Из рассмотрения исключены малочастотные Р2 и Зв.

[12] За проведенную оценку качества морфологической разметки благодарим участников семинара по корпусной лингвистике под руководством Г.Б. Гурина (ПетрГУ).

[13] Вместе с тем напомню, что данные этих корпусов по сравнению с данными (Josselson 1953) сопоставимы по большему числу параметров, что служит косвенным подтверждением репрезентативности этих корпусов.

Литература

Arppe, A.: 2002, 'The usage patterns and selectional preferences of synonyms in a morphologically rich language', Morin, A. & Sébillot, P. (eds.), *JADT—2002. 6th International Conference on Textual Data Statistical* Analysis, March 13-15, 2002, vol. 1, 21-32.

Arppe, A.: 2006, 'Frequency considerations in morphology, revisited — finnish verbs differ, too', *A Man of Measure. Festschrift in Honour of Fred Karlsson in his 60th Birthday*, 175-189.

Biber, D., Reppen, R.: 2002, 'What does frequency have to do with grammar teaching?', *Studies in Second Language Acquisition*, 24/2, 199-208.

Chvany, C., V.: 1996, 'Hierarchies in the Russian case system: for N—A—G—L—D—I, against N—G—D—A—I—L', O.T. Yokoyama & E. Kleinin (eds.), *Selected essays of Catherine V. Chvany*, Columbus (OH), 175-187.

Francis, W., N. & Kucera, H.: 1992, *Frequency analysis of English usage: Lexicon and grammar*, Boston, 1982.

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

Halliday, M.: 1993 / 2005, 'Quantitative studies and probabilities in grammar', *Computational and Quantitative Studies*, London / NewYork, 130-156.

Ilola, E. & Mustajoki, A.: 1989, *Report on Russian Morphology as it Appears in Zaliznyak's Grammatical Dictionary*, Helsinki (=Slavica Helsingiensia, 7).

Johansson, S. & Hofland K.: 1989, *Frequency Analysis of English Vocabulary and Grammar Based on the LOB Corpus. Vol. 2: Tag Combinations and Word Combinations*, Oxford.

Josselson, H., H.: 1953, *Подсчет ходовых слов русского языка*, Detroit (MI).

Karlsson, F.: 1986, 'Frequency considerations in morphology', *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, Berlin, 39/1, 19-28.

Kemmer, S. & Barlow, M.: 2000, *A Usage-Based Conception of Language*, Essen.

Šteinfeldt, E.: 1963, *Russian Word Count*, Moscow.

Белоусова, Е.А.: 1964, 'Статистический анализ глагольных форм (на материале русского языка)', *Актуальные вопросы современного языкознания и лингвистическое наследие Е. Д. Поливанова*, т. 1, Самарканд.

Браславский, П.И.: 2001, 'Морфологический строй функциональных стилей (на материале документов Internet)', *Известия Уральского государственного университета*, № 21, 9-17.

Волоцкая, З.М., Шелимова, И.Н., Шумилина, А.Л.: 1961, 'Некоторые количественные данные о формах существительных и глаголов русского языка', *Лингвистические исследования по машинному переводу*, Москва, 254-261.

Головин, Б.Н.: 1971, *Язык и статистика*, Москва.

Зализняк, А.А.: 1967, *Русское именное словоизменение*, Москва.

Зализняк, А.А.: 1977, *Грамматический словарь русского языка*, Москва.

Засорина, Л.Н. (ред.): 1977, *Частотный словарь русского языка*, Москва.

Коваль, С.А.: 2005, *Лингвистические проблемы компьютерной морфологии*. Санкт-Петербург.

Котова, Н.В.: 2004, 'К проблематике частотного морфемиария болгарского языка' *Славянский вестник*, вып. 2, 158-171.

Лённгрен, Л.: 1993, *Частотный словарь современного русского языка*, Uppsala.

Марков, Ю.: 1960, 'К вопросу о частотности грамматических категорий', *Русский язык в национальной школе*, № 4, 19-20.

Мустайоки, А.: 1973, *Опыт составления частотной грамматики русских существительных*, Хельсинки, (рукопись).

Мустайоки, А.: 1986, 'О минимизации учебного материала', *The Teaching of Russian Language and Literature in Europe*, Brussels, 84-98.

Мустайоки, А. & Копотев, М.В.: 2004, 'К вопросу о статусе эквивалентов слова типа *потому что, в зависимости от, к сожалению*', *Вопросы языкознания*, № 3, 88-107.

Николаев, В.: 1960, 'Некоторые данные о частотности употребления падежных форм в современном русском литературном языке', *Русский язык в национальной школе*, № 5, 19-26.

Никонов, В.А.: 1959, 'Статистика падежей', *Машинный перевод и прикладная лингвистика*, 3(10), Москва, 45-65.

Норман, Б.Ю.: 2003, 'Грамматическая информация в словаре vs. лексическая информация в грамматике', *Труды по русской и славянской филологии. Лингвистика*, VIII (новая серия), Тарту, 148-162.

Пайле, А.: 2003, *Автоматический анализ русского текста*, Хельсинки (рукопись).

Распопова, Т.И.: 2001, 'Иван родил девчонку, велел тащить пеленку или...?', *Мир русского слова*, № 1, 39-42.

Русская грамматика: 1980, Т. 1-2, Москва.

Шаров, С.А.: 'Леммы, отсортированные по частоте', [электронный ресурс], www.comp.leeds.ac.uk/ssharoff/frqlist/frqlist.html.