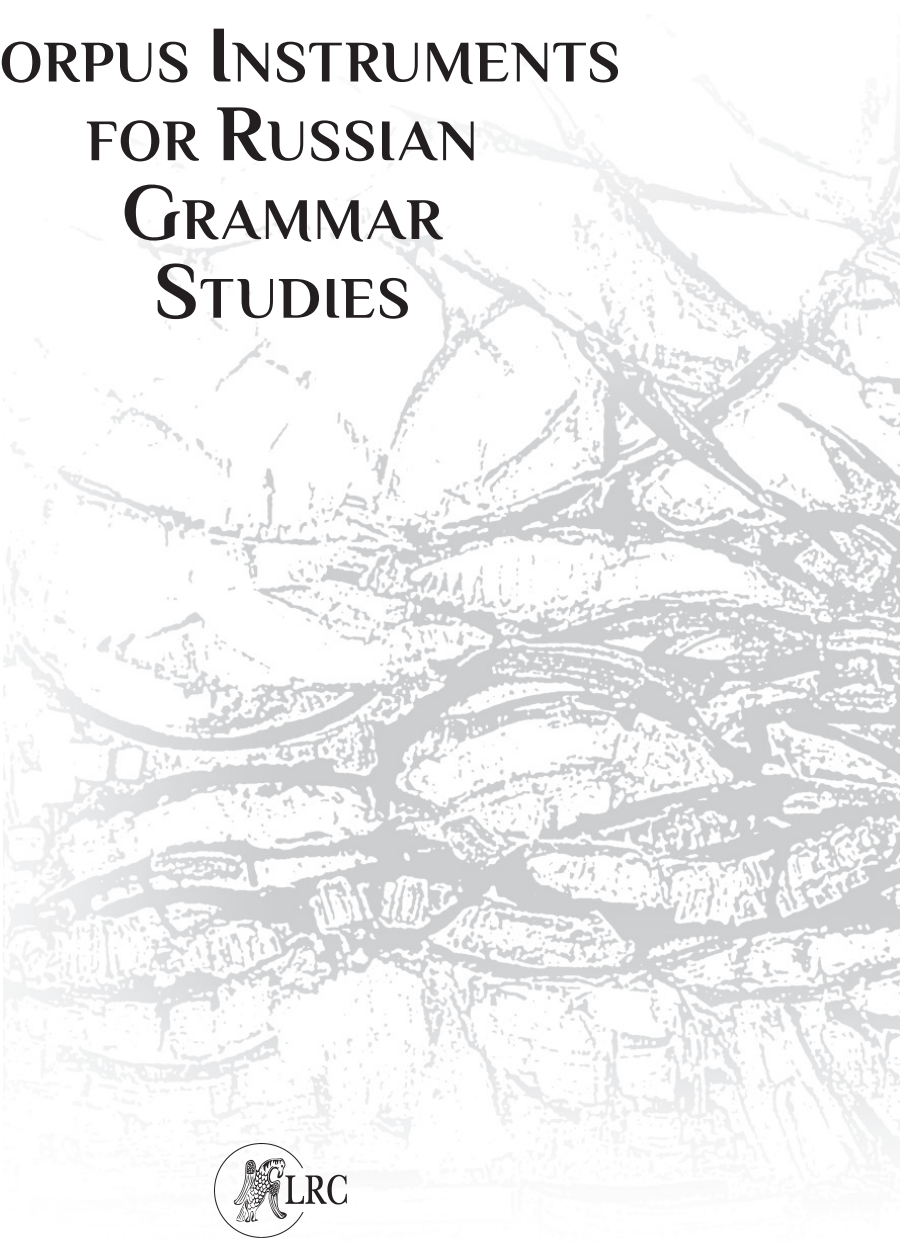
Olga Lyashevskaya

Corpus Instruments for Russian Grammar Studies



LRC PUBLISHING HOUSE
MOSCOW 2016

О. Н. Ляшевская

Корпусные инструменты в грамматических исследованиях русского языка

ИЗДАТЕЛЬСКИЙ ДОМ ЯСК
МОСКВА 2016

ББК 81.1
УДК 80/81
Л 99

Издание осуществлено при финансовой поддержке
Фонда фундаментальных лингвистических исследований
проект № В-28-2014

Утверждено к печати Ученым советом
Института русского языка имени В. В. Виноградова РАН

Рецензенты:
д. ф-м. н. М. Р. Пентус, к. филол. н. И. В. Азарова

Ляшевская О. Н.

Л 99 Корпусные инструменты в грамматических исследованиях русского языка. — М.: Издательский Дом ЯСК, 2016. — 520 с.

ISBN 978-5-9907947-8-8

Русская корпусная лингвистика представлена в книге двумя направлениями. Первая часть содержит описание подходов и методов аннотации Национального корпуса русского языка (http://ruscorpora.ru), включая разметку лексико-грамматической, лексико-семантической, семантико-синтаксической и словообразовательной информации. Кроме того, описываются процедуры оценки инструментов автоматической разметки текстов (морфологических и синтаксических парсеров) и идеология создания двух частотных корпусных словарей, общего и лексико-грамматического. Во второй части представлены результаты исследований грамматики и лексики русского языка с применением квантитативных корпусных методов: изучение грамматических, конструкционных и семантических профилей языковых единиц, в том числе глаголов и глагольных приставок, имен существительных и пространственных конструкций.

УДК 80/81
ББК 81.1

В оформлении переплета использована картина
Пита Мондриана «Серое дерево», 1911

ISBN 978-5-9907947-8-8

СОДЕРЖАНИЕ

Приложения

Предисловие

Корпусная лингвистика — довольно молодое направление лингвистической науки. Национальному корпусу русского языка исполнилось 10 лет, а самому старому представительному корпусу объемом более 100 миллионов словоупотреблений, Британскому Национальному, — всего 25 лет. Прежде всего уточним, что термин «корпусная лингвистика» предполагает два понимания: это и наука о том, как создавать лингвистические корпуса, и методы исследования языка с привлечением корпусных данных. Обычно считается, что созданием корпусов занимаются инженеры и программисты, а исследованиями на данных корпуса — собственно лингвисты. В случае Национального корпуса русского языка это не так: корпус создавался лингвистами и для лингвистов (хотя и с помощью «инженеров»). Мне повезло несколько раз: в начале двухтысячных оказаться в отделе Лингвистических исследований ВИНИТИ РАН, когда только появилась и начала реализовываться идея Национального корпуса; затем в отделе корпусной лингвистики и лингвистической поэтики Института русского языка им. В. В. Виноградова, где ведется основная работа над корпусом; после этого в Институте лингвистики Университета Тромсё, где были начаты первые исследования Национального корпуса с помощью квантитативных методов; и наконец в НИУ «Высшая школа экономики», где собралась замечательная команда исследователей русского языка. Так и получилось, что я работаю в обоих направлениях корпусной лингвистики.

Соответственно, книга, которую вы держите перед собой, тоже имеет две части. Первая часть посвящена лингвистической аннотации текстов Национального корпуса русского языка (ruscorpora.ru) на разных уровнях: словоизменения, словообразования, синтаксиса и семантико-синтаксического интерфейса, лексико-семантических классов. Мы обсуждаем исходные теоретические установки, связанные с системой аннотации, разработку вспомогательных лингвистических ресурсов (словарей и баз данных), компьютерных инструментов разметки и самое интересное — то, что я бы назвала «сопротивлением материала», — описание сложных случаев языкового материала, которые могут вызвать трудности как при автоматической аннотации, так и при ручной разметке. Чуть выходя за рамки задач непосредственно Национального корпуса, мы обращаемся к вопросам стандарта оценки автоматической разметки текстов и рассказываем о двух инициативах в области компьютерной лингвистики — о соревнованиях морфологических и синтаксических парсеров. В конце первой части описываются производные корпуса — частотные словари, которые можно построить на корпусных данных.

Во вторую часть входят работы по исследованию грамматики и лексики русского языка квантитативными корпусными методами. Понятие грамматического «поведения» языковых единиц в применении к корпусу видится как распределение разного рода элементов в контексте. Это грамматический профиль (распределение форм словоизменения), конструкционный профиль (распределение конструкций некоторой «целевой» лексемы), лексический, лексико-семантический профиль (распределение лексем или лексико-семантических классов в контексте другой лексемы или конструкции), радиальный профиль значения (распределение значений / частных употреблений языковой единицы). С помощью методов грамматического, конструкционного, семантического профилирования мы анализируем грамматическую специализацию русских глаголов по формам вида, времени и наклонения; вариативность образования приставочных видовых пар с разными приставками; ограничения на заполнение слотов и связанные с этим вариации значения в генитивной конструкции формы и в пространственной конструкции с предлогом *поверх*. Квантитативные методы, привлекаемые для анализа, разнообразны: от чисто описательных частот и процентных долей до теста Фишера и регрессии.

Создание корпусов и квантитативные исследования, требующие масштабной доразметки корпусных данных, — дело чрезвычайно трудоемкое, и его приятнее делать в коллективе. Поэтому в этом предисловии я бы хотела поблагодарить моих соавторов, с которыми мне посчастливилось работать в наших многочисленных корпусных проектах: В. А. Плунгяна и Д. В. Сичинаву (морфологическая разметка корпуса, см. Ляшевская и др. 2005в) пополнение грамматического словаря, см. (Ляшевская и др. 2007), Е. В. Рахилину, Г. И. Кустову, Е. В. Падучеву, О. Ю. Шеманаеву, Б. П. Кобрицова, Т. И. Резникову (лексико-семантическая разметка корпуса и разрешение неоднозначности, см. Kustova et. al. 2009; Шеманаева и др. 2007; Рахилина и др. 2006), С. Ю. Толдову (синтаксическая разметка корпуса), Ю. Л. Кузнецову, М. С. Кудинова и Е. В. Кашкина (проект ФреймБанк, см. Кузнецова, Ляшевская 2009; Кашкин, Ляшевская 2013; Lyashevskaya, Kashkin 2014), Е. А. Гришину, М. Г. Тагабилеву, И. Б. Иткина, Е. К. Павлову (словообразовательная разметка корпуса, см. Гришина и др. 2009), А. А. Бонч-Осмоловскую, Е. Г. Соколову, С. О. Савчук, С. А. Коваля, еще раз С. Ю. Толдову и команду студентов МГУ (И. Астафьева, А. Королева. М. Ионов, М. Кудринский, Д. Привознов, Евг. Сидорова и мн. др.), с которыми мы организовывали соревнования парсеров (см. Ляшевская и др. 2010; Толдова и др. 2012; Gareyshina et al. 2012; Bonch-Osmolovskaya et al. 2013), С. А. Шарова, моего соавтора по частотному словарю (Ляшевская, Шаров 2009). Вместе с А. В. Десятовой и А. А. Маховой мы делали проект по топологической классификации лексики и исследованию пространственных конструкций (см. Махова и др. 2009; Десятова и др. 2008), с О. А. Митрофановой, П. В. Паничевой, С. В. Романовым, Н. С. Кузнецовой, М. А. Грачковой, А. С. Шиморина и А. С. Шурыгиной — проект по автоматическому разрешению лексико-семантической омонимии, а с В. Г. Сибирцевой и Н. В. Карповым — проекты по использованию материалов корпуса в учебных целях. Наконец, самые большие слова благодарности — основа-

телям исследовательской лаборатории CLEAR group Университета Тромсё Л. Янде, Т. Нессету, С. В. Соколовой, (снова) Ю. Л. Кузнецовой, А. Б. Макаровой и А. В. Эндресен (Байдимировой), вместе с которыми мы учились применять квантитативные корпусные инструменты к данным Национального корпуса русского языка. Я еще раз благодарю своих соавторов за любезное разрешение использовать материалы наших совместных статей в этой книге. Первоначальные варианты многих глав были опубликованы в материалах конференции «Диалог» — и мы бесконечно благодарны ее организаторам и слушателям за многолетний интерес к публикациям разработчиков Национального корпуса.

Особенные слова должны быть посвящены светлой памяти безвременно ушедшего И. В. Сегаловича. Илья одним из первых поддержал идею Национального корпуса, щедро делясь своей позитивной энергией и креативными идеями на семинарах разработчиков корпуса. По инициативе Ильи «Яндекс» стал основным техническим партнером корпуса и инициировал исследовательские гранты, с помощью которых были проведены первые математические исследования на материалах корпуса. Тут же мы должны произнести много теплых слов благодарности в адрес других сотрудников компании «Яндекс», которые на протяжении более десятка лет обеспечивают техническую поддержку корпуса и терпят все капризы лингвистов-разработчиков: А. И. Зобнина, И. Е. Шалыминова, Н. В. Григорьева, А. В. Сокирко, А. А. Аброскина, В. А. Титова, С. А. Григорьеву, Е. С. Грунтову и др. И еще: огромное спасибо студентам трех московских вузов, МГУ, РГГУ и НИУ ВШЭ, принимавших участие в наших проектах в качестве разметчиков. Корпус не был бы таким, какой он есть, без ваших усилий.

В европейской традиции принято благодарить не только научных руководителей, начальников, учителей и коллег, но и тех, с кем просто пил чай. Я бы хотела поддержать эту прекрасную традицию и назвать тех, кто был рядом, помогал, спасал, создавал хорошее творческое настроение и беседовал за чаем о лингвистике и не только: Ю. Родина, М. Пост, Д. Пинеда, П. Иосад, М. Панчева, Х. Андреассен, Л. Антонсен, Р. Михайлык, М. Нордрум, Д. Папрот, Т. Горностай, А. Недолужко, А. Бердичевский, Х. Экхофф, А. Рубин, О. Урюпина, М. Кронгауз, М. Даниэль, Н. Добрушина, Е. Добрушина, В. Апресян, Б. Орехов, Т. Архангельский, Ю. Ландер, А. Летучий, Я. Ахапкина, Д. Алексеевский, О. Виноградова, А. Марушкина, Т. Никитина, Н. Слюсарь, В. Файер, Ю. Галямина, Ю. Кувшинская, М. Худякова, Т. Ряпина, Н. Зевахина, С. Князев, Б. Иомдин, Н. Стойнова, П. Браславский, П. Аркадьев, С. Сай, М. Овсянникова, А. и Л. Ландманы, И. Микулинская, Л. Кацман, В. Гусев, Н. Галицкая, С. Бурлак, В. Степанов, Т. Михайлова, Е. Марголис, Б. Кротов, Е. Калинина, В. Цуканова, Г. Дурново, Н. и А. Горовые, Н. и О. Сидоренковы, Е. Шаульский, А. Занадворова, Е. Ягунова, Л. Пивоварова, М. Копотев, М. и А. Беловы, И. и Ю. Ребриковы, Е. и А. Ребриковы и многие, многие другие. В заключение я хочу произнести слова признательности моим родителям Н. С. и Н. Ф. Ляшевским, моему мужу Саше и сыновьям Егору и Степе. Спасибо вам за терпение, сочувствие и поддержку.

Текст всей книги внимательно прочитали А. Ч. Пиперски, Е. В. Ягунова, А. Я. Шайкевич и официальные рецензенты М. Р. Пентус и И. В. Азарова. Я бесконечно благодарна им за вдумчивые замечания и уточнение ряда формулировок. Безусловно, все оставшиеся несообразности — недоработка автора. Моя глубокая благодарность В. В. Столяровой, Е. Г. Сметанниковой, И. В. Богатыревой, осуществившим техническую подготовку издания к печати.

* * *

Рукопись монографии подготовлена при поддержке Научного фонда НИУ ВШЭ, индивидуальный исследовательский проект № 14-01-0069, 2014-2015. Издание осуществлено с помощью издательского гранта Фонда фундаментальных лингвистических исследований, грант № В-28, 2014/2015 гг.

2.2. Грамматические профили

2.2.1. Грамматическая специализация глаголов в формах времени и наклонения*

Грамматический профиль — это соотношение вхождений форм словоизменения в корпусе. Прежде всего, имеется в виду картина распределения грамматических форм у конкретной лексемы (Janda, Lyashevskaya 2011a; 2011b), однако можно говорить и о грамматическом профиле некоторой семантической группы (например, у глаголов движения) и в целом частеречного класса (грамматический профиль глаголов, имен существительных и т. п.). Немаловажно, что метод грамматического профилирования предполагает сопоставление грамматических профилей языковых единиц между собой и противопоставление «усредненному» профилю класса, к которому они относятся, иными словами, позволяет увидеть среднее и максимальный разброс поведения в изучаемом классе языковых единиц.

В этой главе мы собираемся применить метод грамматического профилирования, чтобы выяснить, как распределение грамматических форм взаимосвязано с лексическим наполнением глаголов, а также с их видовыми характеристиками. Грамматические профили дают ценный материал для исследования двух важных вопросов: а) какова взаимосвязь между классами совершенного и несовершенного вида глаголов и б) каково взаимодействие категорий вида, времени и наклонения (ТАМ[1]) с лексическими классами. Уже много десятилетий в русистике ведется дискуссия по поводу того, формируются ли аспектуальные «пары» только с помощью суффиксов (гипотеза А. В. Исаченко (1960), поддержанная А. А. Зализняком) или же они образуются и через суффиксацию и через префиксацию (гипотеза «школьной» русистики). Мы собираемся проверить гипотезу Исаченко, используя корпусные данные о частоте финитных форм глаголов.

---

* Раздел представляет собой переработанный вариант статьи: *Janda L. A., Lyashevskaya O.* Grammatical profiles and the interaction of the lexicon with aspect, tense, and mood in Russian (Janda, Lyashevskaya 2011b).

[1] ТАМ — традиционное английское сокращение для связки tense, aspect и mood. Мы будем использовать композит «ТАМ-формы» как сокращенный вариант для термина «формы вида, времени и наклонения».

Интроспективные, докорпусные описания русского вида часто включают информацию о конкретных значениях форм времени и наклонения в совершенном и несовершенном виде. Например, употребление императивных форм НСВ связывают с выражением вежливости, а императивных форм СВ — с категорическим приказом. Сопоставление грамматических профилей позволяет обнаружить «выбросы» — такие глаголы, у которых обнаруживается необычно большая доля форм в какой-то конкретной «клетке» парадигмы. Интересно посмотреть, какова семантика таких лексем. В принципе, не исключено, что среди глаголов с неожиданно большой долей форм императива НСВ окажутся слова, у которых смысл вежливости или даже вежливой просьбы «вшит» в лексическое значение, а среди глаголов с большой долей императивных форм СВ — слова с компонентом «категоричности». Если это так, то это будет обозначать, что лексические и грамматические значения «притягиваются» друг к другу, что отражается в частотных распределениях словоформ на данных корпуса (гипотеза лексико-грамматического притягивания). Вместе с тем мы предполагаем, что грамматические профили дадут некоторые новые сведения о значениях форм времени и наклонения у НСВ и СВ.

Еще одна тема, важная для методологии корпусной лингвистики, — это дискуссия о статусе «лексемы», о том, является ли лексема единицей (корпусного) лингвистического описания или она — плод воображения лингвистов (точнее, тех из них, кого можно назвать «интуитивистами»). Дело в том, что определение исходной формы слова (леммы) при аннотации корпусов — довольно трудоемкая задача, а для некоторых языков почти неподъемная, если решать ее автоматическими методами. Но в какой-то момент лингвисты и инженеры, обрабатывающие очень большие массивы текстов (например, делающие классификацию новостей), заметили, что словоформы одной лексемы образуют кластеры «сами по себе», безо всякой предварительной аннотации — просто вследствие сходства их контекстного окружения. Так была выдвинута провокационная для традиционной лингвистики идея: раз аннотация на уровне лексемы — это дорого, долго и сопряжено со множеством ошибок, то, может быть, лексем не существует вовсе? Шт. Грис (Gries 2011) утверждает, что анализ употребления словоформ в корпусе может давать даже более качественный результат, нежели анализ поведения на уровне лексемы. Кроме того, различные корпусные данные могут в сумме давать тот же результат, что и данные по распределению лексем. Дж. Ньюман (Newman 2008) не соглашается с Грисом — он показывает, что распределение словоформ внутри лексемы уже само по себе несет ценную информацию. Таким образом, вопрос ставится так: потеряют ли теоретические исследования на базе корпуса и разработка компьютерных приложений, задействующих корпусные ресурсы, в качестве, если из корпуса будет изъята информация о лемме / о кластере словоформ, представляющих одну лексему? Наше исследование поддерживает точку зрения Дж. Ньюмана — в нашем случае словоизменительная и лексическая информация дает много ценных данных. В то же время мы согласны, что эффект может быть неодинаков для разных языков: для высокофлективных языков эта информация

может быть важнее, чем, скажем, для английского языка с его минимальным набором противопоставленных форм.

Структура грамматического профиля русского глагола

В типологической перспективе русский язык известен как язык, инвестирующий в морфологию: значительная часть информации в нем кодируется в грамматических словоизменительных формах. Соответственно, дистрибуция грамматических форм в корпусе, как самих по себе, так и в сочетании с лексической информацией, потенциально способна сказать лингвистам не меньше, чем распределение функционально нагруженных лексем и конструкций в корпусах языков типа английского. Парадигма русского глагола несовершенного вида содержит порядка 120 форм, парадигма глагола совершенного вида — 68 форм (см. табл. 57). Если рассматривать глаголы в видовых парах, это дает нам в максимуме 188 форм (если, конечно, существуют такие пары глаголов, у которых заполнены все клетки парадигмы).

Таблица 57

Формы словоизменения русских глаголов (синтетические формы)[2]

Подпарадигмы	Категории в подпарадигме	Количество форм в НСВ	Количество форм в СВ
индикатив			
непрошедшее время	лицо, число	6	6
прошедшее время	род, число	4	4
инфинитив		1	1
императив	лицо, число	4	4
деепричастие		1	1
причастие			
полные формы причастия	время, залог, падеж, род, число	96	48
краткие формы причастия	время, род, число	8	4
Всего		120	68

[2] Формы непрошедшего времени выражают настоящее время у глаголов НСВ и будущее время у глаголов СВ. Формы императива различают 1-е и 2-е лицо, во 2-м лице ед. и мн. число, в 1-м лице мн. число и формы совместного действия на *-мте* (ср. *пойдемте*). Глаголы НСВ имеют два деепричастия, настоящего и прошедшего времени; глаголы СВ — одно деепричастие прошедшего времени. Глаголы НСВ имеют до четырех причастий: активное причастие настоящего и прошедшего времени, пассивное причастие настоящего и прошедшего времени (пассивные причастия в основном у переходных глаголов). Глаголы СВ имеют два причастия (активное и пассивное причастия прошедшего времени). Каждое причастие имеет полный набор адъективных форм склонения (изменение по падежу, роду, числу) в полных формах. Кроме того, каждое пассивное причастие имеет 4 кратких формы.

В принципе, корпус может дать нам гигантский объем материала по распределению каждой грамматической формы у каждого русского глагола. Однако в первую очередь следует решить, насколько детальными должны быть различия между группами данных, то есть, последовав совету Шт. Гриса о выборе оптимального уровня грануляции для анализа, откалибровать профиль.

С учетом разнообразия грамматических форм, профили могут быть представлены на разных уровнях разрешения. Самое высокое разрешение, в котором представлены все 120 или 68 форм для каждого глагола, порождают громоздкие многомерные матрицы, где большинство значений равно или приближается к 0. Игнорируя противопоставления форм по падежу в адъективных подпарадигмах причастий (эти различия имеют отношение скорее к согласованию, чем к семантике собственно глагольных категорий и глагольных лексем), мы уменьшим количество грамматических измерений для каждого вида до 40 и 28 соответственно. На рис. 46 представлен профиль в еще более низком разрешении, так что формы лица, числа и рода складываются вместе как составляющие форм времени, наклонения и залога причастий.

Рис. 46 показывает «усредненный» грамматический профиль русских глаголов НСВ и СВ[3] по девяти формам: непрошедшее время, прошедшее время, инфинитив, императив, деепричастие и 4 формы причастий. Как и следовало ожидать, вклад форм в общую частоту глагола неравен. Мы видим, что в грамматическом профиле наиболее частотны формы непрошедшего и прошедшего времени, инфинитива, пассивного причастия прошедшего времени, что покрывает от 87 до 91 % употреблений глагола.

Однако причастия и деепричастия представляют проблему для дальнейшего сопоставительного анализа профилей видовых пар, образованных путем

Данный инвентарь не учитывает пассивные формы на *-ся* типа *договор заключается на два года*, а также аналитические формы словоизменения (формы сложного будущего времени типа *буду учиться*, формы условного наклонения типа *знал бы*, аналитические формы императива, выделяемые некоторыми грамматиками). Таблица дает максимальное число форм, которое может насчитывать парадигма синтетических форм, но не учитывает тот факт, что некоторые глаголы «дефектны» в некоторых зонах парадигмы, ср. семантическую дефектность у безличных глаголов типа **сплюсь*, морфологическую дефектность у глаголов типа *?побежу* / *?победю*, imperativa tantum *на* / *нате*. Вариативность форм, не противопоставленных по грамматическим категориям (ср. *поезжай* / *езжай* / *ехай* и т. п.; *знав* / *знавши* и т. п.; *проведенной* / *проведенною* и т. п.), в таблице не учтена.

[3] Рис. 46 представляет частоты по базе данных частотного словаря НКРЯ (Ляшевская, Шаров 2009), в которой учтены употребления в современной части (1950—2007) Основного корпуса (объем корпуса на момент включения данных в 2010 г. составлял 92 миллиона словоупотреблений; данные корпуса с неснятой лексико-грамматической омонимией были дизамбигуированы автоматически). Частота каждой формы была подсчитана по всем глаголам НСВ и всем глаголам СВ, а затем вычислено процентное распределение форм относительно частоты всех глаголов соответствующего вида. Таким образом, профиль показывает среднюю частоту грамматических форм у глаголов НСВ и СВ.

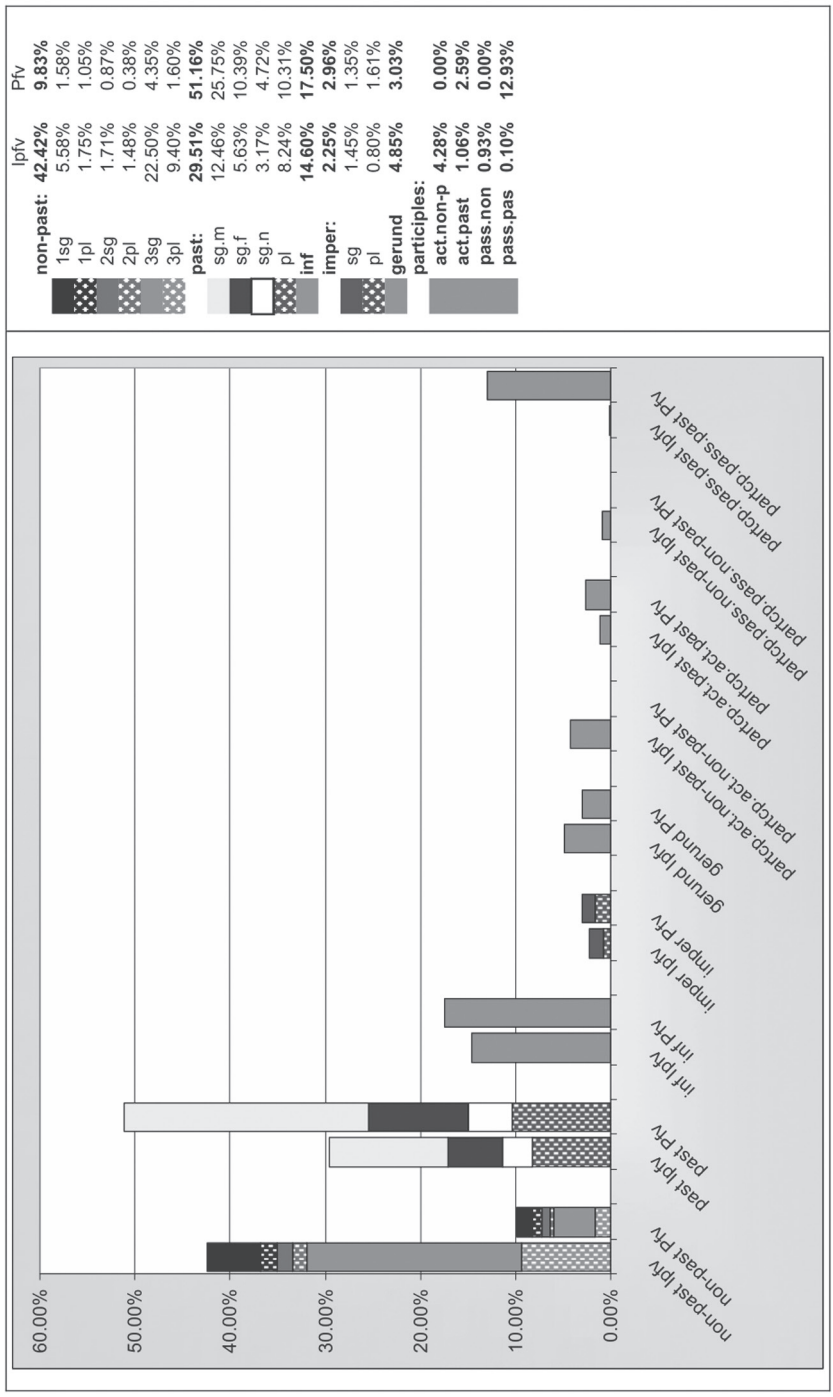


Рис. 46. Средняя частота форм словоизменения по данным НКРЯ
(паттерны внутри колонок показывают распределение форм лица и числа внутри подпарадигмы)

префиксации и суффиксации, так как системные ограничения на их образование у глаголов СВ и НСВ будут вызывать системные перекосы в распределении форм в целом. Так, глаголы СВ не образуют причастий настоящего времени; в целом только переходные глаголы образуют пассивные причастия; приставочным вторичным имперфективам запрещено образовывать пассивные причастия прошедшего времени. У многих морфологических классов глаголов нет возможности образовывать пассивные причастия и деепричастия (см. подробнее Сай 2011; 2014; Биккулова 2011). Если каких-то форм причастий и деепричастий нет, это значит, что доля остальных форм в процентном распределении будет выше. Кроме того, и с точки зрения языка в целом причастия и деепричастия образуют особую гибридную зону между глаголом и прилагательным или наречием (Пешковский 1956). Таким образом, в настоящем исследовании мы решили не включать в грамматический профиль частоты форм причастий и деепричастий.

Итак, мы ограничиваем грамматический профиль финитными формами времени и наклонения. Повелительное наклонение представлено менее чем в 5 % употреблений, однако мы включаем его в профиль наряду с инфинитивом и индикативом, чтобы все категории наклонения были представлены. В данной работе мы отвлекаемся от распределений по лицу, числу и роду в подпарадигмах индикатива и императива, так как они менее замечены во взаимодействии с категорией вида. В результате мы выбираем некоторый «средний» уровень разрешения для наших данных.

Далее мы обозначим некоторые известные в русистике вопросы грамматического описания вида, с оглядкой на которые будет производится анализ грамматических профилей, а именно вопрос о признании префиксальных пар чистовидовыми и отношения между категориями вида, времени и наклонения. После описания двух баз данных, собранных для нашего исследования, мы проверяем гипотезу А. В. Исаченко и исследуем глаголы, обнаруживающие слишком много или, напротив, слишком мало форм в грамматическом профиле относительно среднего поведения глаголов НСВ и СВ.

Видовые пары: словообразовательная перспектива и взаимодействие с категориями времени и наклонения

Литература по русскому виду поистине безбрежна. В этом разделе мы остановимся лишь на двух «вечных вопросах» русистики, важных для нашего исследования, а именно на гипотезе Исаченко и на особых значениях вида в формах времени и наклонения.

Гипотеза Исаченко

С точки зрения словообразования система русской глагольной лексики упрощенно выглядит следующим образом: большинство глаголов с простой основой[4]

[4] Структуры «корень + тематический суффикс + глагольная флексия» или «корень + глагольная флексия».

относятся к несовершенному виду (ср. *писать*, *колоть*), от них добавлением приставки или суффикса образуются глаголы совершенного вида (ср. *написать*, *выписать*, *заколоть*, *кольнуть*), а в свою очередь от приставочных глаголов с помощью суффиксов вторичной имперфективации образуются глаголы несовершенного вида (ср. *выписывать*, *закалывать*)[5].

Вторичная имперфективация представляет собой чистый тип аспектуальных отношений: оба глагола СВ и НСВ имеют одинаковое лексическое значение и различаются лишь семантикой вида. Напротив, глаголы, образованные от простых основ добавлением суффикса, никогда не образуют чистовидовых пар: суффикс *-ну-* добавляет к лексическому значению глагола НСВ семельфактивность (ср. *прыгать — прыгнуть*, подробное обсуждение см. в Makarova, Janda 2009; Dickey, Janda 2009; Горбова 2011[6]), а суффикс *-ива-* — итеративность (ср. *сиживать*, *видывать*, см. Danaher 2003: 31). Глаголы СВ, образованные от простых основ добавлением приставки, делятся на два типа: одни отличаются лексическим значением (ср. *вязать — отвязать*), другие похожи (ср. *делать — сделать*). В глаголах первого типа приставка специализирует значение, вслед за (Janda 2007) этот тип глаголов мы будем называть специализированными перфективами (Specialized Perfectives). Глаголы второго типа, отличающиеся только значением вида, образуют (условно, о чем ниже) чистовидовые пары, вслед за (Janda 2007) мы будем называть этот тип естественными перфективами (Natural Perfectives)[7].

---

[5] Возвратные глаголы образуют те же словообразовательные цепочки, а кроме того, некоторые возвратные перфективы образуются одновременным добавлением приставки и возвратного аффикса к простой основе (ср. *говорить — разговориться*). В нашу упрощенную классификацию также не входят: образование приставочных перфективов со сменой суффикса типа *ронять — уронить*, *менять — изменить*; образование имперфективов от простых глаголов совершенного вида *купить — покупать*, *дать — давать* или *сесть — садиться*, *лечь — ложиться*; образование способов действия приставочно-суффиксальным способом типа *прыгать — попрыгивать* и ряд других случаев (Грамматика 1980). Многие лексемы, прежде всего с именными корнями, начинают свою словообразовательную историю как глаголы со второго этапа (ср. *охладеть*). Закрытый класс глаголов движения устроен более сложно, включая ряды глаголов однонаправленного движения и неоднонаправленного движения; глаголы неоднонаправленного движения приставочным способом образуют и перфективы (см. *заходить* ‘начать ходить’) и имперфективы (ср. *проходить* ‘миновать’), см. (Janda 2010). Наконец, достаточно маргинально и явление полипрефиксации, или нанизывания приставок (prefix stacking, см. Розейзон 1970; Беляков, Гиро-Вебер 1997; Ramchand 2004; Svenonius 2004a; 2004b; Татевосов 2009; 2013), когда добавление приставки к приставочному глаголу дает, как правило, еще один перфективный глагол (ср. *переписывать*, которое имеет двойное прочтение, делимитативное и дистрибутивное).

[6] Исключением является употребление семельфактива в определенных типах контекстов, вынуждающих однократную интерпретацию, ср. *Потом прыгнул Иванов — Потом прыгает Иванов* (пример из (Горбова 2011: 28); см. также Перцов 2001: 128).

[7] Гнездовая классификации Янды дополнительно различает среди приставочных перфективов, специализирующих значение глагола НСВ, собственно Specialized Perfective

Таким образом, чистовидовые отношения можно усмотреть у двух типов пар, а именно:

а) суффиксальные пары (s-пары) «приставочный глагол — вторичный имперфектив» (*выписать — выписывать*);

а) префиксальные пары (p-пары) «глагол с простой основой — приставочный глагол» (*делать — сделать*)[8].

В то время как большинство описаний (Виноградов 1938; Шахматов 1941; Грамматика 1980; Бондарко 1983; Черткова 1996; Зализняк, Шмелев 2000) и практически все словари и учебники признают и суффиксальные, и префиксальные пары чистовидовыми, грамматика А. В. Исаченко (1960: 130—175) и вслед за ней самый авторитетный грамматический словарь русского языка (Зализняк 1977/2003) объявляют чистовидовой только суффиксальную имперфективацию. Позиция А. В. Исаченко основана на интроспективном наблюдении, что а) добавление приставки всегда привносит значение, ассоциированное с приставкой, в лексическое значение глагола, тем самым делая невозможным прямое сопоставление значений глаголов НСВ и СВ, и б) приставочные перфективы замещают свои корреляты-имперфективы не во всех контекстах, где это допустимо. Исаченко приводит контрпримеры, когда приставочные пары не проходят функциональные тесты на видовую парность. Отсутствие четкой границы противопоставления между приставочными лексемами, которые специализируют (Veyrenc 1980) лексическое значения глагола НСВ, и приставочными лексемами, которые не изменяют лексическое значение глагола СВ, вынуждают и А. А. Зализняка (1977/2003: 6, 136) принять позицию Исаченко. Автор последней по времени общей грамматики русского языка А. Тимберлейк (Timberlake 2004: 410—411) занимает промежуточную позицию. По мнению Тимберлейка, префиксальные перфективы и их вторичные имперфективы удовлетворяют критерию аспектуальной пары, а простые имперфективы и соответствующие приставочные перфективы являются квази-партнерами («near-partners»).

В целом все исследователи сходятся на идее расширения понятия видовой парности, где пары с суффиксальной вторичной имперфективацией видятся как прототипический случай, а пары с приставочными специализированными перфекти-

(ср. *доделать*, *разделать*, *выделать*), Complex Act Perfective, соответствующий приставочным «способам действия» (Akzionsart) (ср. сатуратив *наделать*, делимитатив *поделать*, инхоатив *закричать* и др.) и Specialized Single Act Perfective — приставочный перфектив, образованный от Single Act Perfective (семельфактивов, ср. *выпрыгнуть*, см. Makarova, Janda 2009).

[8] В этом случае также имеет место упрощение, так как суффиксальный и приставочный классы не учитывают пары с супплетивными основами типа *говорить — сказать*, *класть — положить*, имперфективацию от бесприставочного глагола совершенного вида типа *решить — решать*, а также двувидовые глаголы типа *жениться*, *миновать*, *пировать* (об образовании приставочных пар от таких глаголов см. Janda 2004: 523; 2007: 637—638). Эти глаголы исключены из нашего анализа.

вами — как заведомо «нечистовидовая» пара, хотя и соотносимая по виду. Вопрос в том, где проходит граница. Несмотря на то что в русском языке имеется порядка 20 перфективирующих приставок, почти все они используются и для образования специализированных перфективов, и для образования естественных перфективов (Кронгауз 1998)[9], т. е. этот критерий не помогает. Портит картину и отсутствие красивого правила, при каких глаголах какая приставка образует естественный перфектив (см. гл. 2.3.1 и 2.4), и множество конкретных спорных случаев, является приставочная пара «чистовидовой» или нет (например, пары типа *петь — пропеть* или *идти — пойти*). См. также широкую дискуссию по поводу выделения пар для отдельных значений глагола, отдельных аспектуальных интерпретаций его значения, необходимости и достаточности тестов на видовую парность, начиная с (Маслов 1948); последние обзоры в (Зализняк и др. 2010; Горбова 2011; Ясаи 2013). Таким образом, ощущается необходимость в новых объективных критериях[10] для определения «чистых» и «нечистых» видовых отношений, и, как кажется, корпусные данные здесь будут очень кстати.

Итак, мы имеем два противоположных взгляда на образование (чисто)видовых пар в русском языке:

(1) «традиционная» гипотеза: аспектуальные пары образуются либо с помощью префиксации от имперфективов с простой основой, либо с помощью суффиксации от приставочных перфективов;

(2) гипотеза Исаченко: аспектуальные пары образуются только суффиксальным способом от приставочных перфективов.

Логическим следствием из этих двух гипотез применительно к нашим данным будет следующее:

(1а) следствие «традиционной» гипотезы: оба типа пар (p-пары и s-пары) функционально тождественны и должны вести себя одинаково, в том числе в отношении словоизменения;

(2а) следствие гипотезы Исаченко: s-пары — единственно возможные аспектуальные пары; поскольку p-пары представляют другое отношение, они должны вести себя иначе, в том числе в отношении словоизменения.

Остается проверить эти два следствия, т. е. сопоставить наши данные о грамматических профилях p-пар и s-пар. Если поведение s-пар будет отличаться от поведения p-пар, это будет свидетельствовать в пользу гипотезы Исаченко. Если наши

---

[9] Не образуют естественные перфективы такие приставки, как *над-*, *недо-*, *до-* и нек. др. Приставка *до-*, впрочем, отмечается в составе глаголов типа *достигнуть*: будь у них бесприставочные корреляты типа *стичь* / *стигнуть*, они бы образовали естественную пару.

[10] Заметим, что, хотя критерии контекстной замены Маслова являются формальными, они оказываются бесполезными в практическом плане, например при изучении русского как иностранного или при создании систем машинного перевода, где знания о «хороших» парах очень нужны. Пока что трудно представить себе корпус, где бы были представлены пары употреблений прошедшего и настоящего исторического в идентичных или близких к идентичным контекстах.

данные покажут, что поведение p-пар и s-пар значимым образом не отличается, это будет свидетельствовать в поддержку «традиционной» гипотезы.

Вид и подпарадигмы времени и наклонения

Как мы уже указывали ранее, имеются рациональные основания для исключения деепричастий и причастий из нашего анализа. Однако мы намеренно опускаем грамматические противопоставления по лицу, числу и другим категориям, чтобы сфокусироваться на тех категориях, которые наиболее активно взаимодействуют с видом, а именно на категориях времени (прошедшее *vs.* непрошедшее) и наклонения (инфинитив, императив, индикатив[11]).

Особые взаимоотношения между аспектом, временем и наклонением хорошо известны в разных языках (Comrie 1976; Chung, Timberlake 1985; Binnick 1991; Bybee et al. 1994; Nuyts 2001; 2007) и занимают центральное место в русской глагольной системе. Заметим, что одна из последних по времени грамматик русского языка А. Тимберлейка (Timberlake 2004) включает всего семь глав, но лишь одна из них, под названием «Mood, tense, and aspect», занимает 73 страницы и включает детальное описание того, как время и наклонение взаимодействуют с видом. Тимберлейк (Ibid.: 373) выделяет для русского языка три наклонения, выражаемых морфологически: реалис (флективные формы непрошедшего и прошедшего времени), императив и инфинитив. Утверждается, что императивные формы более употребительны для глаголов СВ, однако в определенных контекстах (отрицание, вежливость, настойчивость) формы НСВ могут быть предпочтительнее (Ibid.: 374—376); см. также (Пулькина, Захава-Некрасова 1977: 284—285; Wade 1992: 303—304). В контекстах инфинитива характерны модальные маркеры (*нельзя*, *надо* и др.), но только инфинитивы НСВ разрешены в конструкциях с фазовыми глаголами (*начать*, *перестать* и др.) и в конструкциях сложного будущего со вспомогательным глаголом *буду*, *будешь* и др. (Timberlake 2004: 360—370; Пулькина, Захава-Некрасова 1977: 272—276; Грамматика 1980: 605; Wade 1992: 306—312). С точки зрения категории времени морфологически различаются только формы прошедшего и непрошедшего времени. Именно вид различает настоящеее время (непрошедшее НСВ) и будущее время (непрошедшее СВ), хотя возможны также другие интерпретации употреблений СВ и НСВ в непрошедшем времени (например, настоящее историческое, которое выражается главным образом формами непрошедшего времени НСВ, или кратно-цепная конструкция типа *бывает придет*, в которой непрошедшее СВ обозначает повторяющуюся последовательность событий; см. Бондарко 1971: 207; Comrie 1976: 73—78; Dickey 2000: 126—154, 52—68). В целом наблюдается связь имперфектива с настоящим

[11] Изучение дистрибуции форм сослагательного (условного) наклонения не входит в задачи нашего анализа, поскольку выражается не с помощью флексий, а перифрастически (Добрушина 2012; в печати). Однако нужно иметь в виду, что употребления условного наклонения учтены как употребления форм прошедшего времени.

(= непрошедшим) временем и перфектива с прошедшим временем (Comrie 1976: 83—84).

Область взаимодействия ТАМ-категорий еще не получила полномасштабного описания с точки зрения корпусного анализа. В этом исследовании мы сравниваем относительную частоту распределения форм в подпарадигмах совершенного и несовершенного вида, чтобы проверить гипотезу Исаченко о «неравноценности» префиксальных перфективирующих и суффиксальных видовых пар, а также исследуем отражение взаимодействия времени, наклонения и вида в корпусных данных. Далее мы сначала проанализируем поведение всех глаголов в совокупности, а затем сосредоточимся на поведении отдельных глаголов.

Данные

Для нашего исследования были созданы две базы данных, одна с данными о префиксальных видовых парах (имперфектив с простой основой и префиксальный перфектив), а другая с данными о суффиксальных видовых парах (префиксальный перфектив и образованный от него с помощью суффикса вторичный имперфектив). Частотные данные для обеих баз получены из Основного корпуса Национального корпуса русского языка (http://www.ruscorpora.ru), подкорпус современных текстов (1950—2007)[12]. Каждая база данных включает частотную информацию о следующих грамматических формах:

- Ipfv_NonPast: сумма вхождений форм лица и числа (1sg, 2sg, 3sg, 1pl, 2pl, 3pl) непрошедшего времени НСВ (т. е. настоящее время);
- Ipfv_Past: сумма вхождений форм рода и числа (m.sg, n.sg, f.sg, pl) прошедшего времени НСВ;
- Ipfv_Inf: вхождения форм инфинитива НСВ;
- Ipfv_Imper: сумма вхождений форм лица и числа (2sg, 2pl, 1pl, форма совместного действия на *-мте* типа *идемте*) императива НСВ;
- Pfv_NonPast: сумма вхождений форм лица и числа (1sg, 2sg, 3sg, 1pl, 2pl, 3pl) непрошедшего времени СВ (т. е. простое будущее время);
- Pfv_Past: сумма вхождений форм рода и числа (m.sg, n.sg, f.sg, pl) прошедшего времени СВ;
- Pfv_Inf: вхождения форм инфинитива СВ;
- Pfv_Imper: сумма вхождений форм лица и числа (2sg, 2pl, 1pl, форма совместного действия на *-мте* типа *пойдемте*) императива СВ.

Поскольку НКРЯ включает не только глаголы с высокой частотой, но и редкие глаголы, у которых распределение форм времени непоказательно, мы установили порог включения глаголов в наши базы данных. В них включались только те видовые пары, у которых финитные формы имели по 100 и более вхождений для каждого вида. Напротив, редкие пары типа *арканить — заарканить*, насчитывающей

[12] На момент создания баз данных подкорпус содержал 92 млн словоупотреблений. Все цитируемые в работе примеры взяты из этого подкорпуса.

в корпусе 2 формы НСВ и 21 форму СВ, из рассмотрения исключались[13]. Мы изъяли из баз еще некоторые пары, для того чтобы результаты анализа были, с одной стороны, аккуратными, а с другой — не требовалось бы сложной предобработки (дизамбигуации) данных. Детали будут изложены в двух следующих подразделах.

База данных префиксальных пар (простые имперфективы и префиксальные перфективы)

В первую очередь перед нами стояла задача составить список всех префиксальных пар, таких, что глагол НСВ с простой основой коррелирует с префиксальным глаголом СВ. Этот список был получен из базы исследовательского проекта Exploring Emptiness Университета Тромсе (http://emptyprefixes.uit.no). Список пар был скомпилирован по данным двух словарей — (МАС 1999) и (Ожегов, Шведова 2001) — и списка пар из работы П. Кабберли (Cubberly 1982), а затем дополнительно отфильтрован командой экспертов[14]; в итоге база включает 1981 видовую пару. Из этого множества мы исключили следующие пары, которые потенциально создавали «шум» в корпусных данных для нашего исследования:

a) глаголы, которые в НКРЯ либо не встречаются, либо встречаются с частотой ниже установленного порога;

b) глаголы, которые образуют более одной пары с простым глаголом НСВ (например, *валить — свалить* и *валить — повалить*; учет таких глаголов потребовал бы дизамбигуации пары для каждого контекста с глаголом НСВ в корпусе, что было бы неподъемной, а во многих случаях и принципиально неразрешимой задачей);

c) глаголы с омонимией обеих или какой-то одной видовой формы; к ним относятся двувидовые глаголы типа *арендовать* (формы СВ и НСВ совпадают), глаголы типа *сходить* (партнер СВ *сходить* ‘пойти и вернуться’ омонимичен глаголу НСВ *сходить*, ср. *Она медленно сходила с лестницы*), пары типа *жать — сжать* (различаются в той части парадигмы словоизменения, где используется основа настоящего времени, ср. *жму* и *жну*, *сожму* и *сожну*, но омонимичны в остальной части парадигмы)[15].

---

[13] Принимая во внимание общую статистику употреблений форм вида, времени и наклонения в НКРЯ, см. таблицы далее, мы стремились к тому, чтобы отсутствие в корпусе самой редкой формы, императива, объяснялось бы свойствами глаголов, а не низкой частотой лексемы (при частоте 100 в среднем ожидается 2—3 формы императива). Заметим, что в аналогичном исследовании (Eckhoff, Janda 2014) на данных очень небольшого корпуса старославянского языка частотный порог был вынужденно снижен до 20 вхождений, но тем не менее и это обеспечило достоверные данные о распределении форм в грамматическом профиле.

[14] Участники проекта Exploring Emptiness — носители языка Ю. Л. Кузнецова, О. Н. Лященевская, А. Б. Макарова и С. В. Соколова.

[15] Учитывая, что в корпусные данные включали порядка шести миллионов глагольных форм, ручная дизамбигуация в конкретных контекстах и тут была бы вряд ли возможна.

В результате база приставочных коррелятов включает только такие пары, где каждый глагол СВ и НСВ идентифицируется однозначным образом, что делает ее похожей на базу данных суффиксальных пар. В последней вероятность омонимии партнера НСВ с другим глаголом близка к нулю (см. ниже), а пары типа (b) и (c) по определению исключены. «Очищенная» база приставочных коррелятов включает 264 видовых пар, которые в целом встречаются в корпусе более 1,6 миллиона раз.

База данных суффиксальных пар (префиксальный перфектив и суффиксальный вторичный имперфектив)

Как и в предыдущем случае, исходной задачей было составить список всех потенциально возможных суффиксальных пар. Это было сделано на основе Грамматического словаря (Зализняк 1977/2003), в котором суффиксальные пары указываются особо, и по данным НКРЯ — всего 19 208 пар. Однако многие из входящих в эти видовые пары глаголов встречаются с частотой ниже установленного порога. Кроме того, отмечается несколько редких случаев, когда префиксальный глагол СВ имеет парный вторичный имперфектив с двумя вариантами форм, ср. *заготовить* и *заготовлять* / *заготавливать*. После удаления низкочастотных глаголов и глаголов с двойным вариантом формы НСВ в базе суффиксальных осталось 1311 пар.

Добавим также, что в двух базах данных наблюдалась небольшая зона перекрытия. Так, глагол СВ *вырасти* может значить либо ‘становиться выше или старше’, и тогда он входит в префиксальную пару *расти — вырасти*, либо ‘превращаться в кого-л.’, ‘становиться слишком большим для того, чтобы носить одежду’, и тогда он образует пару со вторичным имперфективом *вырасти — вырастать*[16]. Зона перекрытия охватывает 38 видовы пар.

Обе базы были дополнены информацией о частоте грамматических форм для каждого из видов, а сами глагольные лексемы были закодированы по признакам вида и типа пары.

Замечание о методе анализа

Данные, описанные ранее, включают в сумме 1 575 видовых пар и представляют 5 951 250 глагольных форм, встречающихся в НКРЯ. Огромный массив корпусной выборки представляет определенную проблему для статистического анализа. Перед тем как перейти к нему, требуется понять взаимосвязь между размером выборки и величиной эффекта (effect size). Имея большую выборку, легко ошибиться и найти эффект там, где его практически не существует, слишком уж велика «статистическая сила» больших массивов данных (см. Baayen 2008: 114—116; Tabachnik, Fidell 2007: 54—55). Модель хи-квадрата устроена таким образом, чтобы

[16] Соответствующие имперфективы, хотя и сходны по значению, отличаются особенностями употребления. Так, *расти* преимущественно ассоциируется с употреблениями в предметном значении, особенно когда речь идет о растениях, в то время как *вырастать* употребляется чаще метафорически.

показывать значимую разницу в дистрибуции групп данных. Но чем больше точек наблюдения мы имеем, тем легче модель хи-квадрата определяет все меньшие и меньшие расхождения. При стремлении размера выборки к бесконечности хи-квадрат идентифицирует бесконечно малые расхождения. Поскольку в корпусной лингвистике часто используются выборки размером в тысячи и миллионы примеров, меру хи-квадрата дополняют измерением величины эффекта. Мера Крамера (Cramer's V) нормирует значение хи-квадрата относительно количества точек наблюдения и ее значения варьируют в интервале от 0 до 1. Эмпирически установлено, что для таблиц данных, имеющих не более двух строк и/или столбцов, значение меры Крамера 0,5 соответствует большой величине эффекта, 0,3 соответствует среднему эффекту, а 0,1 — малому эффекту (Cohen 1988: 215—271; Cohen et al. 2003: 182; King et al. 2008: 327—330). Для таблиц большего размера Коэн вводит следующие поправки: для таблиц, имеющих не более трех строк и/или столбцов, большой эффект — 0,35; средний эффект — 0,21; малый эффект — 0,07; для таблиц, имеющих не более четырех строк и/или столбцов, большой эффект — 0,29; средний эффект — 0,17; малый эффект — 0,06.

Чем больше величина эффекта, тем лучше: мы с большей уверенностью можем утверждать, что наблюдаемые перевесы данных объясняются именно эффектом различительного признака, а не размером выборки. Поправки на количество строк и столбцов в таблице (определяется по меньшему из измерений) вводятся в связи с тем, что для таблиц размера 3 × 3 и более практически невозможно получить меру Крамера более 0,4 (эмпирически установлено Коэном).

Далее в анализе мы приводим меру Крамера после каждой метрики хи-квадрата, с тем чтобы показать, что статистическая значимость установлена надежно.

Грамматические профили видовых пар

Взятые вместе, базы данных по префиксальным и суффиксальным парам дают нам общую широкую панораму поведения глаголов НСВ и глаголов СВ. Табл. 58 позволяет сравнить грамматический профиль глаголов НСВ (из обеих баз данных) с грамматическим профилем глаголов СВ (из обеих баз данных). Каждый профиль включает четыре формы (непрошедшего времени, прошедшего времени, инфинитива, императива), приводятся абсолютные частоты и доля вхождений от суммы четырех форм (сумма процентов в четырех клетках для каждого вида дает 100 %).

Таблица 58

Грамматические профили ТАМ-форм совершенного и несовершенного вида

	Imperfective				Perfective			
	Ipfv_ NonPast	Ipfv_ Past	Ipfv_ Inf	Ipfv_ Imper	Pfv_ NonPast	Pfv_ Past	Pfv_ Inf	Pfv_ Imper
p- & s-пары	1 330 016 47,4 %	915 374 32,6 %	482 860 17,2 %	75 717 2,7 %	375 170 11,9 %	1 972 287 62,7 %	688 317 21,9 %	111 509 3,5 %

Данные в табл. 58 говорят о том, что распределение форм в двух видах совершенно разное. У глаголов НСВ доминируют формы непрошедшего (настоящего) вида, в то время как у глаголов СВ еще сильнее доминируют формы прошедшего времени. Метрики хи-квадрата (chi-squared = 947756, df = 3, p-value < 2,2e−16) показывают, что расхождения в поведении глаголов разного вида статистически значимы, а мера Крамера равна 0,399, что характеризует величину эффекта между «средней» (0,3) и «большой» (0,5). Таким образом, влияние признака вида на поведение грамматических форм в русском языке установлено с достаточной степенью уверенности. Наблюдение согласуется с теоретическими выкладками Б. Комри (Comrie 1976: 84) в отношении взаимодействия времени и вида, и наше эмпирическое исследование подтверждает гипотезу Комри на больших объемах корпусных данных.

Теперь мы предлагаем разделить точки, полученные из разных баз данных, с тем чтобы изучить, будет ли влиять характер видовой пары (префиксальный *vs.* суффиксальный) на грамматический профиль. В табл. 59 данные разведены по этому признаку: в верхней строке приводятся абсолютные и относительные частоты для глаголов, входящих в префиксальные пары, а в нижней строке — абсолютные и относительные частоты для глаголов, входящих в суффиксальные пары. Таким образом, показано 4 профиля, и по-прежнему сумма долей ТАМ-форм для каждого профиля дает 100 %. Как видно, фактор способа образования видовой пары оказывается гораздо слабее: верхний и нижний профили отличаются друг от друга меньше, чем отличаются левый и правый.

Таблица 59

Общие глагольные профили для глаголов НСВ и СВ по данным баз приставочных и суффиксальных пар

	Imperfective				Perfective			
	Ipfv_ NonPast	Ipfv_ Past	Ipfv_ Inf	Ipfv_ Imper	Pfv_ NonPast	Pfv_ Past	Pfv_ Inf	Pfv_ Imper
p-пары	475 893	397 409	195 926	36 427	72 439	317 570	114 460	24 280
	43 %	35,9 %	17,7 %	3,3 %	13,7 %	60,1 %	21,6 %	4,6 %
s-пары	854 123	517 965	286 934	39 290	302 731	1 654 717	573 857	87 229
	50,3 %	30,5 %	16,9 %	2,3 %	11,6 %	63,2 %	21,9 %	3,3 %

На рис. 47 процентные данные из табл. 59 представлены графически. Визуализация также подтверждает, что разница в поведении префиксальных пар (залиты темно-серым) и суффиксальных пар (светло-серый цвет) незначительна. Отсутствие эффекта способа образования пары на грамматический профиль подтверждается статистическим тестами.

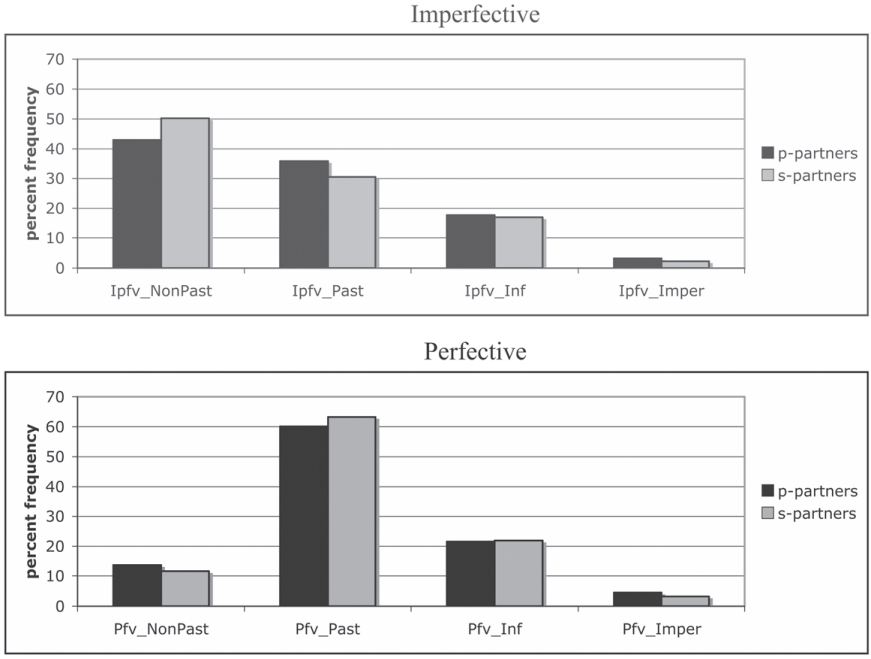


Рис. 47. Распределение грамматических форм в префиксальных парах (темно-серый цвет) и суффиксальных парах (светло-серый цвет)

Метрики хи-квадрата, сравнивающие профили глаголов НСВ в префиксальных *vs.* суффиксальных парах (ср. верхнюю часть рис. 47), показывают статистически значимое расхождение (chi-squared = 16 155,13, df = 3, p-value < 2,2e–16), однако значение меры Крамера (Cramer's V = 0,076) не достигает даже порога для малого эффекта. Аналогичные результаты дает сравнение профилей глаголов СВ в префиксальных *vs.* суффиксальных парах (ср. нижнюю часть рис. 47; метрики хи-квадрата chi-squared = 4365,078, df = 3, p-value < 2,2e–16 снова предсказывают статистически значимое расхождение, но величина эффекта слишком мала: Cramer's V = 0,037). Мы можем заключить, что некоторые наблюдаемые различия профилей НСВ и СВ объясняются большим количеством данных в выборке, а вовсе не фактором типа видовой пары.

Частотная иерархия грамматических форм соблюдается как в общей выборке, так и в выборках префиксальных и суффиксальных пар: в НСВ чаще всего встречаются формы непрошедшего времени, затем прошедшего, затем инфинитива и, наконец, императива. В парадигме СВ чаще всего встречаются формы прошедшего времени, затем формы инфинитива, затем формы непрошедшего времени, а замыкает иерархию снова императив.

Итак, общий вывод состоит в том, что в отношении грамматических профилей поведение префиксальных и суффиксальных видовых пар оказывается практиче-

ски идентичным. Этот вывод говорит в пользу «традиционной» гипотезы о том, что в русском языке видовые пары образуются как с помощью суффиксации, так и с помощью префиксации. Гипотеза Исаченко в нашем исследовании не подтверждается. Однако, безусловно, следует оговорить, что когда-нибудь в других корпусных исследованиях, возможно, будут обнаружены другие особенности поведения этих двух морфологических типов видовых пар, которые будут свидетельствовать, напротив, в пользу гипотезы Исаченко. Поиск таких факторов, впрочем, выходит за рамки нашей задачи.

В следующем разделе мы будем использовать всё те же корпусные данные для нахождения глаголов с необычной дистрибуцией форм. Поскольку противопоставление префиксальных и суффиксальных пар оказалось нерелевантным, мы объединяем данные из обеих баз данных вместе.

Выбросы: глаголы с необычной дистрибуцией форм

Основываясь на эмпирическом наблюдении Э. Штейнфельдт (1963) о том, что глаголы различаются между собой по дистрибуции форм в парадигме, мы ожидаем, что грамматические профили отдельных лексем будут значительно варьировать в нашей выборке. Поскольку рассматриваемые ТАМ-формы содержательно наполнены, мы также ожидаем, что эти различия могут быть связаны с семантикой и прагматикой глагола. В этой связи ожидается, что та или иная конкретная комбинация вида, времени и наклонения будет ассоциироваться с конкретными группами глаголов. Наша гипотеза состоит в том, что глаголы с максимальной концентрацией конкретной ТАМ-формы (например, имеющие 90 % форм императива вместо 3 %) обладают особым свойством, а именно, их семантическое наполнение особо благоприятно для данного грамматического значения. Цель этого раздела — проверить данную гипотезу.

Анализ разделен на восемь подразделов, в соответствии с числом комбинаций признаков вида, времени и наклонения. Естественно, мы собираемся сравнить новые эмпирические данные о значении ТАМ-форм и о лексической семантике глаголов с предшествующими наблюдениями аспектологов, а также по возможности предложить объяснение для новых фактов.

Статистический прием нахождения глаголов с необычной дистрибуцией ТАМ-форм строится на идентификации выбросов (outliers), которая производится следующим образом.

Во-первых, все глаголы ранжируются в порядке возрастания доли конкретной ТАМ-формы в грамматическом профиле.

Во-вторых, отсортированные данные делятся на четыре равных группы (четверти) и определяются значения долей ТАМ-формы (а) в точке первой квартили Q1, т. е. на границе 1-й и 2-й четверти данных; (б) в точке второй квартили Q2, или медианы, т. е. на границе 2-й и 3-й четверти; (в) в точке третьей квартили Q3, т. е. на границе 3-й и 4-й четверти. Разница между значениями в точках (а) и (в) составляет интерквартильное расстояние (inter-quartile range, IRC) и характеризует

разброс данных в 50 % выборки, находящихся в центре ранжированного списка (по четверти данных вниз и вверх от медианы).

В третьих, от значения в точке (а) отнимается, а к значению в точке (в) добавляется величина, равная полутора интерквартильным расстояниям. Точки, оказавшиеся за пределами этого порога, считаются выбросами (см. King et al. 2008: 71—72, 76—78), т. к. предполагается, что допустимый разброс данных в первой и четвертой четверти должен укладываться в границы полуторного интерквартильного расстояния[17]:

THRmin = Q1 – 1,5 IQR; THRmax = Q3 + 1,5 IQR.

Для графического представления центральных тенденций распределения данных и выбросов используют графики-боксплоты, см. рис. 48. Жирная горизонтальная линия в центре показывает медиану данных; «ящик» в центре показывает разброс значений во второй и третьей четверти (50 % данных вокруг медианы); «усы» — два отрезка пунктиром — соответствуют полутора интерквартильным расстояниям; мелкие кружки — это выбросы[18].

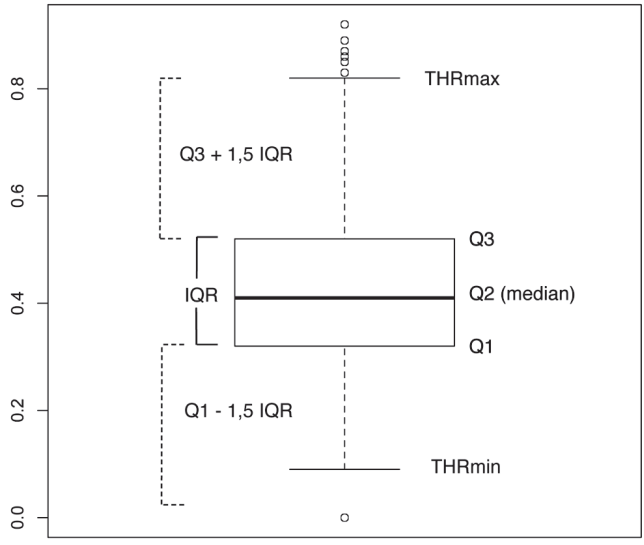


Рис. 48. График-боксплот, показывающий распределение глаголов НСВ по доле форм непрошедшего времени

[17] В статистической литературе оговаривается, что порог в 1,5 IQR довольно условен и диктуется традицией (установлен эмпирически). Поэтому «выбросы» более корректно называть «потенциальными выбросами», см., например, (Agresti, Finlay 1997). В качестве дополнительного критерия выделения «выбросов» часто предлагают обнаружить разрыв в расстояниях между соседними точками в районе THRmin и THRmax, однако это не было целью нашего анализа: мы пользовались расстоянием в 1,5 IQR как готовым внешним критерием.

[18] Для создания боксплотов использована функция языка R, см. (Baayen 2008: 30).

В старых пособиях по статистике выбросы предлагалось считать «плохими» данными, которые могут портить статистические метрики, особенно если они строятся на среднем арифметическом выборки (например, если среднее значение выборки {1, 2, 3, 3, 4, 5} — 3, то среднее значение от выборки {1, 2, 3, 3, 4, 5, 21} составит уже 7, то есть окажется уже за пределами основной группы). Поскольку выбросов обычно мало и они демонстрируют какое-то необычное для группы поведение, полагалось считать их случайным шумом, который не повлияет на выводы о поведении основной группы.

В нашем исследовании выбросы — это глаголы с необычно большой или, напротив, необычно малой концентрацией ТАМ-форм в грамматическом профиле. Таким образом, мы хотим показать, что выбросы могут представлять весьма ценные данные для лингвистического анализа. Заметим, что значения выбросов не влияют на значение медианы и интерквартильное расстояние, так как значения квартилей определяются с помощью ранжирования данных, а не через вычисление среднего арифметического.

Следующие восемь подразделов строятся по общему образцу. Сначала мы приводим сведения о значении той или иной ТАМ-формы, известные из долгой истории их изучения в русской грамматической традиции. Затем мы приводим боксплот, который показывает, сколько глаголов имеет низкую концентрацию данной ТАМ-формы, среднюю концентрацию и высокую концентрацию, — иными словами, он дает общую картины дистрибуции глаголов в плане частоты употребления изучаемой ТАМ-формы. После этого в таблице приводится список глаголов — выбросов (для каждого глагола указывается абсолютная частота изучаемой ТАМ-формы и доля употреблений относительно общей частоты глагола). В последней части раздела мы предлагаем содержательный анализ глаголов-выбросов: мы объединяем их в группы, характеризующиеся общими особенностями семантики и прагматики, и обсуждаем, насколько хорошо эти группы соответствуют известным описаниям. На всякий случай мы проверяем, что условие необычной концентрации ТАМ-форм необходимо и достаточно для выделения наших групп: выбирая случайным образом глаголы из разных частей ранжированного списка, мы показываем, что они по своим содержательным свойствам не могут быть отнесены к выделенным группам.

Порядок представления ТАМ-форм в следующих разделах определяется количеством выбросов. Больше всего выбросов наблюдается в императивных формах, на втором месте формы непрошедшего времени, на третьем месте инфинитивы, а меньше всего выбросов среди форм прошедшего времени. Мы по очереди рассматриваем соответствующие формы НСВ и СВ.

Императив НСВ

Известны три основных положения об особенностях употребления несовершенного вида в императивной форме: что имперфектив используется в контекстах категорического отрицания, что имперфектив используется для обозначения вежливости и что имперфектив используется, чтобы сигнализировать безотлагательность

или настойчивость императива (см. Бондарко, Буланин 1967: 127—128; Падучева 1996: 12—17; Грамматика 1980: 624; Timberlake 2004: 374—375). Последние два утверждения кажутся противоречащими друг другу. И. Б. Шатуновский (2002; 2009) предлагает следующее решение этой проблемы. В основе и вежливого и грубого употребления императива лежит общий функциональный мотив: слушатель должен понимать, что предлагаемое действие должно осуществиться. Кроме того, слушатель может быть заранее расположен к осуществлению определенных действий, в зависимости от ситуации, в которой он находится. В такой ситуации, как визит к знакомым, слушатель предполагает уже многое из того, что произойдет: он войдет, сядет, поест и т. д. Императивные указания говорящего в этих ситуациях будут интерпретированы как вежливые.

В иных ситуациях может получиться так, что слушатель не выполняет действия, несмотря на явные указания говорящего (*напишите заявление — пишите же!*). В этих случаях выбор говорящим несовершенного вида будет восприниматься как грубость, поскольку подразумевает настойчивость. Согласно Шатуновскому, имеется еще и третий тип ситуаций, в которых употребление императива нейтрально, не имея ни оттенка вежливости, ни оттенка настойчивости. Говорящий просто поддерживает слушателя в том, что они уже и так намеревается сделать.

Учитывая эти сведения, в корпусной выборке мы предполагаем найти среди глаголов НСВ с необычно большим содержанием форм императива такие, которые часто используются для выражения вежливого побуждения, грубой настойчивости, нейтральной поддержки намерений говорящего, а также те, которые используются в контекстах категорического отрицания.

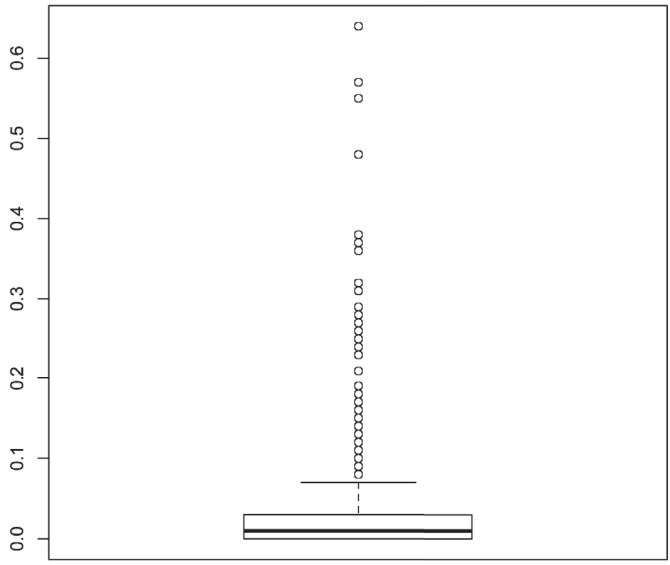


Рис. 49. Распределение глаголов НСВ по доле форм императива

Комбинация признаков несовершенного вида и императива дает более двухсот глаголов, которые ведут себя как выбросы, см. кружочки на рис. 49. Доля форм императива у этих глаголов составляет от 8 % до 64 %[19]. Список неоднороден, но в нем можно выделить группы, причем некоторые из них хорошо укладываются в традиционные описания, а другие пока не получили особого внимания русистов-аспектологов.

1. Глаголы, относящиеся к фреймовому сценарию приема гостей или визитеров. Сценарий включает вход в помещение (*входи*), снятие теплой одежды (*раздевайся*), занятие определенного положения тела в пространстве (*садись*), присоединение к другим людям за столом (*присоединяйся*), потребление чего-л. (*закусывай*, *закуривай*). К этому же классу можно добавить приглашение совершить визит (*заезжай*[20]) и присоединиться к поездке (*залезай <в автомобиль>*). В этом сценарии императивы не дают слушателю никакой новой информации, а скорее приглашают гостя совершить ожидаемое и гостем, и хозяином действие. Согласно Шатуновскому, эти императивные формы обозначают вежливость.

2. Другие вежливые побуждения, не получившие объяснения у Шатуновского, — это просьбы о помощи (*выручай*) и добрые пожелания (*выздоравливай*).

3. Многие императивные глаголы нейтральны или обладают оттенком «снисходительной фамильярности». Таковы просьбы покинуть говорящего с тем, чтобы совершить некое требуемое действие (*ступай*), сосредоточиться на новом объекте или новой теме разговора (*гляди*), взять нечто предлагаемое (*забирай*).

4. Грубые побуждения с оттенком настойчивости. В эту группу входят глаголы с указанием уехать, покинуть говорящего (*проваливай*, *отваливай*) или прекратить делать что-то (*кончай*, *бросай*).

5. Более половины «грубых» употреблений императивов связана с отрицанием и несет в себе указание на то, что собеседник неправильно себя ведет, особенно в коммуникации (*не перебивай*, *не прикидывайся*, *не передергивай*). Эта группа не учтена в анализе Шатуновского, который рассматривает отрицание отдельно и исходит из того, что под отрицанием все императивы обычно употребляются в несовершенном виде, за исключением тех случаев, когда они указывают на малоконтролируемые ситуации и ситуации непосредственной угрозы. Показательно, что глаголы с оттенком грубости чаще употребляются в единственном числе (которое ассоциируется с близостью и фамильярностью), а не во множественном числе (которое в случае императива ассоциируется либо с множественностью адресатов, либо с вежливостью). Например, *отваливай* имеет 95 % форм в единственном числе.

6. Прочие (нейтральные или фамильярные) употребления под отрицанием связаны с эмоциональной поддержкой собеседника (*не расстраивайся*, *не стесняйся*,

---

[19] Из-за большого размера полная таблица глаголов-выбросов приведена в Приложении 2А.

[20] Отдавая себе отчет в том, что некоторые глаголы могут иметь несколько значений, здесь и далее мы классифицируем глаголы по наиболее часто встречающемуся в контекстах корпуса значению формы.

не волнуйся). Отдельно стоит упомянуть два глагола, которые, по идее, должны были бы оказаться в классе «грубых» императивов, но все же входят в «нейтральный класс»: *не ленись* и *не забывай*. *Не ленись* похож на глаголы указания на неправильное поведение, однако, во-первых, употребляется в основном по отношению к детям или к тем, к кому снисходительно относятся, как к ребенку (это исключает грубость), а во-вторых, преимущественно употребляется как сопроводительная присказка к другим побуждениям, имеющая значение ‘делай тщательно’ или ‘делай, даже если тебе этого не хочется’, ср. *читай <доставай*, *напиши ему>, не ленись*; *не ленитесь перезванивать*. *Не забывай* — это нейтральный заместитель императива *помни*, который в современной культуре воспринимается как слишком категоричный и невежливый. *Не забывай* служит предупреждением о возможном наступлении неконтролируемого события (Зализняк Анна 2006б; Апресян 2008а) и как бы призывает адресата предпринимать постоянные усилия, чтобы это событие не наступило. В корпусе *не забывай* часто встречается во вторичной дискурсивной функции, вводя в рассмотрение новые обстоятельства и аргументы.

7. Оставшаяся часть глаголов обнаруживается в корпусе преимущественно в фиксированных грамматических или идиоматических выражениях. Императив *давай*(*те*) используется в качестве вспомогательного глагола при образовании описательных императивов, таких как *давай посмотрим* и *давайте я вам помогу*. Как указывает А. Барентсен (2006), это выражение определенно принимает во внимание перспективу говорящего и выполняет функцию вежливого предложения: глаголы, которые наиболее часто встречаются с *давай*, предполагают, что говорящий мотивирован скорее своим собственным желанием выполнить действие, нежели намерениями адресата, ср. *давай помогу*, *давай расскажу*, *давай покажу*, *давай сделаю.* Таким образом, данное идиоматизованное употребление *давай* совместимо с вежливыми употреблениями императивов НСВ в целом.

Императив *прощай* функционирует идиоматически как прощальная формула вежливости и, таким образом, несколько отделен от основного значения глагола *прощать*. У трех форм императива НСВ есть определенные «культурные корни»: *обогащайтесь* было модным лозунгом эпохи НЭПа 1920-х годов; *соединяйтесь* чаще всего употребляется в составе коммунистического лозунга *Пролетарии всех стран, соединяйтесь!*; *запевай* — армейская команда, связанная с тем, что солдаты в армии во время строевой подготовки обязаны петь в унисон. Три других императива НСВ часто наблюдаются в афоризмах: *не поминай лихом* и *поминай, как звали* (вместе оба контекста представляют 76 % всех употреблений императива в НКРЯ); *спасайся, кто может* (42 % вхождений в корпусе); *на чужой каравай рот не разевай* (53 % вхождений форм императива в корпусе).

Для сравнения мы взяли глаголы в нижней и средней частях ранжированного списка и проанализировали их (не)схожесть с глаголами в зоне выбросов. В нижней части списка находится 36 глаголов НСВ с 0 %-й долей императива. 32 из них содержат рефлексивный суффикс *-ся* и как группа в целом называют неконтролируемые действия с неодушевленным субъектом или употребляются безлич-

но, ср. *вспоминаться*, *начаться*, *приходиться*. В среднем диапазоне, где глаголы имеют от 2 % до 4 % императивных форм, обнаруживается 18 глаголов, которые описывают те или иные действия, никак не связанные с вежливостью или безотлагательностью, ср. *думать*, *решать*, *смеяться*. Таким образом, поведение глаголов в нижней и средней части списка не похоже на поведение глаголов-выбросов.

Императив СВ

Об императивах СВ в аспектологии написано относительно мало. Шатуновский (2002), описав почти на 30 страницах особенности императивов НСВ, ограничивается лишь несколькими поверхностными замечаниями в отношении императивов СВ (см. также Падучева 1996; Timberlake 2004). Доминантное употребление императивов СВ связано с намеренным побуждением адресата к действию (собственно, это и есть основная функция императива). Особого внимания аспектологов удостоились лишь две группы — грубые требования (противопоставляемые вежливым просьбам в НСВ, см. предыдущий раздел) и предупреждения (Пулькина, Захава-Некрасова 1977: 284—287; Грамматика 1980: 623—624; Wade 1992: 303—306).

Таким образом, в корпусной выборке глаголов-выбросов с очень большой долей форм императива СВ мы ожидаем увидеть нейтральные инструкции, грубые требования и предупреждения.

Рис. 50. Распределение глаголов СВ по доле форм императива

Рис. 50 показывает, что императивы СВ дают нам самое сдвинутое к нулю распределение и более трехсот точек-выбросов. Полный список глаголов-выбросов представлен в Приложении 2В.

Как и ожидалось, две наиболее многочисленные группы включают грубые требования и нейтральные указания-инструкции.

1. Примерами грубых требований служат формы *отстань*, *отвали* ‘оставь меня в покое’, *отпусти* ‘позволь мне уйти, покинуть тебя’, *перестань*, *уймись*. Здесь совершенный вид усиливает категоричность команды, однако заметим, что парный «вежливый» имперфектив к этим глаголам отсутствует.

2. Нейтральные инструкции чаще всего встречаются в текстах особого жанра (рецепты, полезные советы и т. п.) и относятся к характерным для таких текстов фреймам: кулинария и домохозяйство (*вскипяти*, *влей*, *завари*, *натри*, *высуши*), физические тренировки (*согни*, *расслабь*), учебные задачи (*запиши*, *начерти*, *умножь*, *перечисли*), заполнение официальных бумаг (*распишись*) и т. п. Инструкции представляют действие в перспективе результата (что будет, если инструктируемый совершит действие), поэтому совершенный вид в них вполне естественен. Сюда же можно отнести и текстовые инструкции, руководящие вниманием читателя, типа *рассмотрим <график x>*, см. группу 5 ниже.

3. Вопреки ожиданиям, группа глаголов-предупреждений типа *не упади* в выборке выбросов императива практически не представлена.

Вместе с тем найдено еще несколько новых групп.

4. Глаголы, обозначающие вежливую просьбу или сочувствие (*извини*, *потерпи*). К этой группе относятся и делимитативы, у которых ограничительная семантика в комбинации с императивной семантикой дает значение смягчения и вежливости (*погуляй*, *покури*, *поторопись*, *побойся*). Еще одна интересная подгруппа — «смещенные» императивы, ср. *постарайся не двигаться 15 секунд* (вместо *не двигайся*), *не поленитесь заглянуть под капот* (вместо *загляните под капот*), *проверьте, что дверь закрыта* (вместо *закройте дверь, если нужно*), *не забудьте добавить воды* (вместо *добавьте воды*), *запаситесь карандашами* (вуалирует трату ресурсов, которая может быть неприятна адресату).

5. Глаголы, обладающие дискурсивной функцией, используются для того, чтобы направлять внимание собеседника или подавать сигналы о переходе темы в беседе (Stefanowitsch, Gries 2003: 233—234 указывают на похожую функцию у английских императивов). Переключение внимания связано с направленным зрительным восприятием и слухом (*посмотри*, *вслушайся*), реже — с обонянием (*понюхай*) и включением канала воображения (*представь*, *угадай*, *вообрази*). Сигнал введения новой темы разговора подают глаголы *разреши*(*те*), *позволь*(*те*) (вводящие инфинитив или клаузу; заметим, что они же используются как сигнал для входа в помещение), *подскажи*(*те*) (приглашающий собеседника высказаться), *постой* (просьба не уходить от темы), *уволь*(*те*) (выражающий категорическое желание говорящего прекратить обсуждение темы). Дискурсивную функцию выполняет и идиоматизированное вводное употребление императива *пожалуй* (только в ед. числе).

Две меньшие, но все же показательные группы связаны с употреблением в религиозных контекстах и в составе устойчивых фразем.

6. Глаголы, связанные с религиозным дискурсом, используются в составе готовых формул, происходящих из литургических текстов и молитв, ср. *Господи помилуй* (91 % примеров императива в НКРЯ — это молитва или прямая цитата из нее), *благослови отче* <*Господи*, *Аллах*> (93 % употреблений в НКРЯ), *избави бог* (100 % употреблений варианта на *-и — избави*).

7. Императивы, которые преимущественно употребляются в составе устойчивых идиоматизированных выражений, включают: *залейся*, *завались* (ср. *хоть залейся* / *завались*, 68 % употреблений), *разлей* (ср. *не разлей вода*, 90 % употреблений), *раздери* (*черт тебя раздери*, 100 % употреблений). Наконец, императив *дай* используется как вспомогательный глагол в составе особой конструкции с перфективом будущего времени типа *дай посмотрю*, параллельно со своим аспектуальным партнером *давай* (см. выше). По мнению Барентсена (2006), наиболее часто встречающиеся в этой конструкции коллокаты — *дай поцелую*, *дай посмотрю*, *дай погляжу*, *дай взгляну* — указывают на мотивацию говорящего (а не адресата). Эти выражения вряд ли можно причислить к грубым, однако, поскольку с точки зрения этикета коммуникации они не принимают в расчет перспективу собеседника, они воспринимаются как нейтральные или фамильярные.

Сравним теперь точки-выбросы с теми данными, которые находятся в нижней и средней части ранжированного списка. В нижней части списка находятся 13 глаголов СВ с долей императивов в 0—1 %, которые ассоциируются с так называемой квазиимперативной конструкцией, ср. (1).

(1) *Начнись схватка — ее бы убили* [Сергей Лукьяненко. Ночной дозор (1998)].

В таких примерах описываются внезапно наступающие события, а не исполняемые человеком действия. Таким образом, глаголы из «подвала» списка явно отличаются от глаголов с необычно большой долей императивных форм.

Восемь глаголов в средней части ранжированного списка с долей форм 3,8 %—5,2 % обозначают в императиве типичные нейтральные сигналы к выполнению действия, ср. *попроси*, *покажи*, *принеси*, т. е. не представляют для нашего исследования ничего интересного.

Непрошедшее время НСВ

Грамматики русского языка согласно характеризуют непрошедшее (настоящее) время НСВ как имеющее функцию описания продолжающихся процессов, конкретных процессов, которые имеют длительность и/или одновременны с другим событием и повторяющихся действий (Пулькина, Захава-Некрасова 1977: 264—270; Грамматика 1980: 604; Wade 1992: 283—286). Лишь во вторую очередь упоминают употребление этой формы для выражения безвременных фактов и отношений (настоящее гномическое[21]) и для представления прошедших событий как

[21] Об использовании имперфективов для обозначения гномического времени см. (Janda 2004).

разворачивающихся в настоящем времени (настоящее историческое). Учитывая эти сведения, мы выдвинули гипотезу, что глаголы с необычно большой долей форм непрошедшего времени должны скорее всего обладать семантикой, предрасположенной к обозначению продолжающихся, длительных и одновременных действий. Однако эта гипотеза не подтверждается.

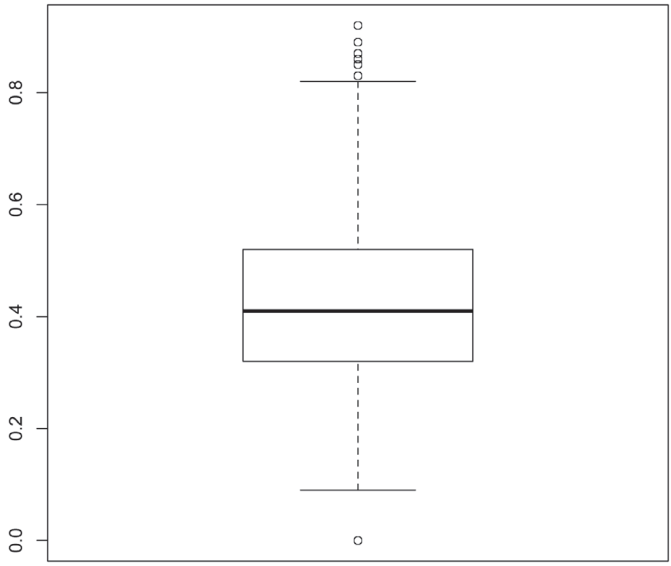


Рис. 51. Распределение глаголов НСВ по доле форм непрошедшего времени

Рис. 51 показывает, что имеется 10 глаголов с необычно высокой долей форм и один глагол с необычно низкой долей форм (0 %). Глаголы перечислены в табл. 60.

Все десять глаголов НСВ с необычно высокой способностью к употреблению в форме непрошедшего времени обладают общим свойством лексического значения, которое характерно и для значения видо-временной формы. Это — обозначение гномических отношений, т. е. вневременных истин. Примеры (2) и (3) иллюстрируют гномическое употребление таких глаголов.

(2) *Другими словами, я бы хотел спровоцировать дискуссию, что всегда является наиболее продуктивной формой научного обсуждения проблемы* [Вирусные гепатиты (2002) // «Вопросы вирусологии», 2002.12.02];

(3) *Как правило, данное обстоятельство влечет за собой негативные последствия для клиентов* [Т. Ливенкова. На всякий случай (2001) // «Туризм и образование», 2001.03.15].

Первые шесть глаголов в табл. 60 — это варианты гномических установок ‘X есть Y’. Еще четыре глагола с необычно высокой долей формы имеют каузальное значение ‘X влечет / является следствием Y’.

Таблица 60

Глаголы НСВ с необычно большой и малой долей форм непрошедшего времени

Глагол (инфинитив и 3sg)	Абс. частота	Доля форм, %
являться (*является*)	39 543	92 %
оказываться (*оказывается*)	10 869	85 %
подтверждаться (*подтверждается*)	677	83 %
выясняться (*выясняется*)	805	89 %
касаться (*касается*)	9719	87 %
исчерпывать (*исчерпывает*)	100	89 %
предопределяться (*предопределяется*)	34	85 %
обязываться (*обязывается*)	480	92 %
затрудняться (*затрудняется*)	275	86 %
влечь (*влечет*)	1555	85 %
слыхать (*слыхает*)	1	0 %

Гномические глаголы склонны употребляться в устойчивых речевых шаблонах. Следующие глаголы предпочитают вводные конструкции: *оказывается, P* (95 % употреблений в НКРЯ); *выясняется, что P* или *как выясняется, P* (87 % употреблений); *что касается X-а, P* (68 % употреблений в корпусе). *Обязываться* и *затрудняться* предпочитают жанр договора и описания результатов социологических опросов соответственно, ср. *стороны <партии> обязуются*, *представитель обязуется* и т. п. (75 % употреблений) и *затрудняюсь ответить*, *5 % затрудняются ответить* (27 % употреблений от числа форм непрошедшего времени).

Единственный выброс в нижней части ранжированного списка, у которого в корпусе засвидетельствована всего одна форма непрошедшего времени — *слыхать*, морфологически аномален. Эта аномалия мотивирована эвиденциальностью, которая «вшита» в значение глагола и которая связана с прошедшим временем (см. раздел «Прошедшее время НСВ», где этот же глагол цитируется как глагол с максимальным количеством форм прошедшего времени). В непрошедшем времени для обозначения слухового восприятия используется нейтральный глагол — *слышать*.

По традиции сравним теперь глаголы-выбросы с глаголами, которые находятся в нижней и средней части ранжированного списка. В нижней части (менее 20 % форм непрошедшего времени) мы находим глаголы типа *обедать* и *голосовать*, в центральной части находятся глаголы типа *работать* и *помогать*. Как видно, гномическими свойствами они не обладают.

Возвращаясь к нашей гипотезе, мы можем сделать вывод, что цитируемые грамматиками «основные» значения видо-временной формы, а также значение настоящего исторического времени не имеют лексической привязки, а значит, отсутствие соответствующих классов среди глаголов-выбросов вполне объяснимо.

Непрошедшее время СВ

Непрошедшее время СВ — морфологическая форма для выражения простого будущего времени — ассоциируется с конкретными простыми действиями, которые, предположительно, должны завершиться в будущем; значительно реже эта форма используется для обозначения хабитуальных повторяющихся действий и в «наглядно-примерной» функции (Пулькина, Захава-Некрасова 1977: 264—270; Грамматика 1980: 604; Wade 1992: 283—286). Мы предполагаем увидеть среди выбросов глаголы, описывающие предсказуемые ситуации и обещанные действия.

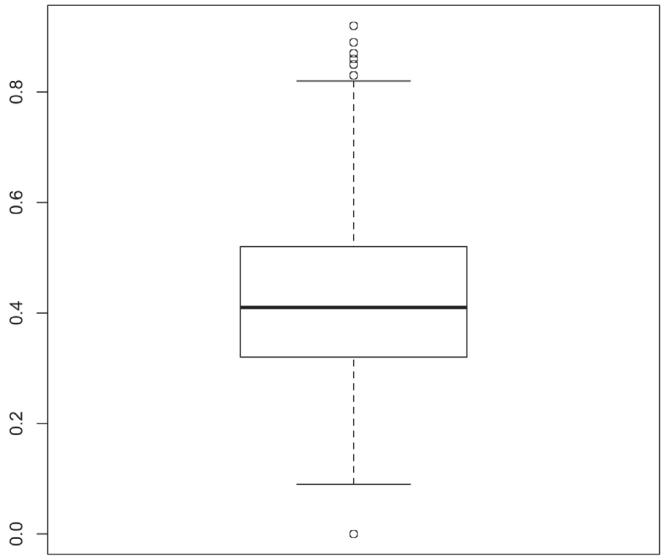


Рис. 52. Распределение глаголов СВ по доле форм непрошедшего времени

84 глагола СВ (перечислены в Приложении 2С) находятся в верхней части ранжированного списка из-за необычно большой представленности форм простого будущего в парадигме. В самом деле, многие из них обозначают предсказания и обещания. Предсказания могут быть параметрическими, например они указывают на длительность по времени (*продлится*) или на увеличение/уменьшение значения параметра (*превысит*, *уменьшится*). Предсказания также часто касаются ожидаемых улучшений (*наладится*, *выздоровеет*), возникновения проблем (*затруднит*, *разорится*) и нехватки ресурсов (ср. *потребуется*). Плохой ситуацией, которая часто служит предметом предсказания, является смерть; в списке мы находим пять глаголов с соответствующим значением типа *подохнет* и *загнется*. С ситуацией предсказанной нехватки ресурсов связаны два глагола, появляющихся в безличных модальных конструкциях: *придется* и *не обойдется* (*без чего-л.*).

Обещания — особый вид предсказания, ср. *управлюсь* и *постараюсь*. Угрозы — это обещания сделать что-то плохое, ср. *растерзаю*, *прокляну*. Некоторые

предсказания конца существования описываются с помощью метонимических и метафорических употреблений глаголов, таких как *сожрет*, *сгниет*.

К обещаниям можно отнести и перформативные употребления в форме будущего времени, ср. *осмелюсь* (*доложить*), *процитирую* (*новость*).

Наконец, несколько точек-выбросов объясняются тем, что соответствующие глаголы преимущественно употребляются в устойчивых выражениях в будущем времени, ср. *не придерешься* (98 % употреблений будущего времени), *остальное приложится* (74 % употреблений), *от тебя не убудет* (96 % употреблений), *врагу не пожелаешь* (с учетом вариантов, 65 % употреблений).

За пределами группы выбросов, в нижней и средней части ранжированного списка, глаголов предсказания, обещания и т. п. мало. Например, среди глаголов с низкой долей форм непрошедшего СВ (0—1 %) мы встречаем глаголы речи (ср. *пробормочет*, *взвизгнет*) и интерпретации (Апресян 2004в; ср. *извинит*, *недооценит*). Последние предпочитают прошедшее время, так как встроенная в них оценка обычно соотносится с прошедшими событиями, но вряд ли соответствующие действия можно считать предсказуемыми. В средней части (12—15,35 %) также не встречается ничего похожего на глаголы выделенных классов, ср. *услышит*, *покажет*, *пошлет*.

Инфинитив НСВ

Инфинитивы несовершенного вида используются в конструкции сложного будущего времени и с другими вспомогательными глаголами. Кроме того, инфинитив НСВ связывают с модальными оборотами (Пулькина, Захава-Некрасова 1977: 272—275; Wade 1992: 307—312). Вопреки типологическим тенденциям, дефолтный вид в модальных конструкциях в русском языке — СВ, хотя НСВ тоже возможен (см. (Divjak 2009), где ситуация в русском языке рассматривается в типологической перспективе). Анна А. Зализняк и А. Д. Шмелев (2000) указывают, что в таких модальных конструкциях СВ обозначает «алетическую» (или, иначе, «динамическую») модальность, т. е. физическую необходимость или возможность, в то время как НСВ выражает деонтическую модальность, связанную с социальными и моральными установками. Зализняк и Шмелев утверждают, что различие в виде объясняется контролируемостью, а именно: совершенный вид используется в контекстах, где событие находится вне контроля субъекта, тогда как несовершенный вид ассоциирован с контекстами, в которых субъект контролирует ситуацию. Этот интроспективный анализ хорошо работает в минимальных парах, которые приводят Зализняк и Шмелев, ср. их примеры *Нельзя разбудить отца* (физически) и *Нельзя будить отца* (неправильно так делать).

Д. Дивьяк (Divjak 2009), анализируя квантитативные корпусные данные, напротив, утверждает, что дело не в контролируемости, а в специфичности / обобщенности, которая предсказывает вид глагола в таких конструкциях. А именно: русский совершенный вид связан со специфическими, определенными ситуациями, и это влечет интерпретацию ситуации в перспективе личной (физической) способности.

В отличие от этого, несовершенный вид связан с универсальными ситуациями (ср., в частности, гномические употребления, обсуждавшиеся в разделе «Непрошедшее время НСВ»), и отсюда происходит его способность интерпретировать ситуацию с точки зрения постоянных норм социальной ответственности. Доказывая свою точку зрения, Дивьяк разметила базу данных корпусных примеров по нескольким признакам и затем использовала технику логистической регрессии со смешанным эффектом, для того чтобы выяснить, какие из выделенных факторов лучше предсказывают вид. Модель предсказала, что фактор специфичности/обобщенности сильнее фактора контролируемости.

Обратимся к нашему ранжированному списку. В распределении (см. рис. 53) имеется 12 выбросов, соответствующие глаголы перечислены в табл. 61. Выбросы с максимальной частотой форм инфинитива в грамматическом профиле по преимуществу употребляются в модальных конструкциях, поэтому нам только остается проверить две вышеизложенные гипотезы — Зализняк и Шмелева (2000) и Дивьяк (Divjak 2009). Это мы сделаем в следующем разделе, сравнивая списки НСВ и СВ.

Сравнение верхней, средней и нижней части общего распределения профилей в форме инфинитива НСВ показывает, что каждый из них связан с определенным типом употреблений глагола. Для выбросов, как уже было сказано, характерно употребление в модальных контекстах. В середине (доля форм инфинитива 16,4 — 20 %) мы находим глаголы, которые обычно употребляются в форме сложного будущего, ср. *будет демонстрировать*, *будет приветствовать*. В нижней части распределения располагаются глаголы, которые «недолюбливают» инфинитивные конструкции, такие как *ухитряться* и *переполнять*.

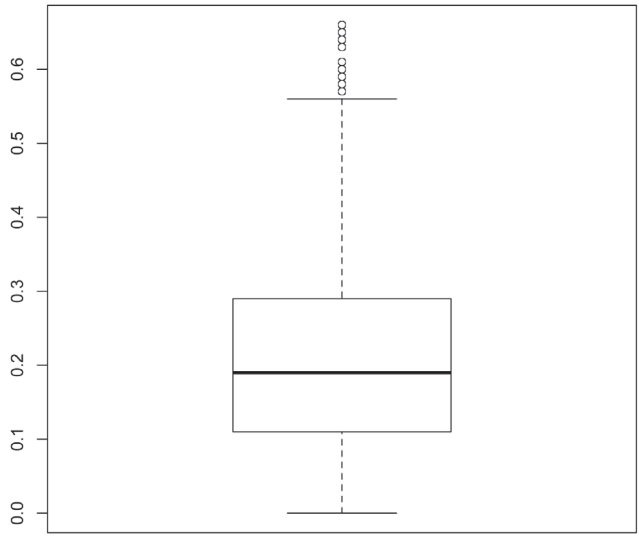


Рис. 53. Распределение глаголов НСВ по доле форм инфинитива

Таблица 61

Глаголы НСВ с необычно большой долей форм инфинитива

Глагол	Абс. частота формы	Доля форм, %
плевать	900	65 %
ввязываться	124	66 %
изыскивать	92	64 %
исправлять	283	61 %
переделывать	230	57 %
пересматривать	198	66 %
развивать	1363	57 %
размещать	272	58 %
распознавать	113	59 %
соблюдать	1013	60 %
согласовывать	176	63 %
учитывать	1850	66 %

Инфинитив СВ

В распределении глагольных профилей по доле форм инфинитива СВ (см. рис. 54) обнаруживается 12 выбросов. Соответствующие глаголы перечислены в табл. 62. В этом разделе мы хотим проверить гипотезу Зализняк и Шмелева и гипотезу Дивьяк одновременно на списках глаголов-выбросов НСВ и СВ.

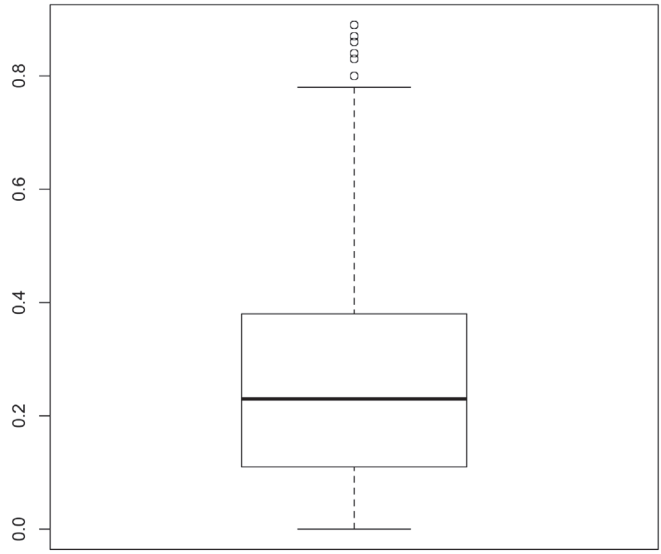


Рис. 54. Распределение глаголов СВ по доле форм инфинитива

Таблица 62

Глаголы СВ с необычно большой долей форм инфинитива

Глагол	Абс. частота формы	Доля форм, %
наплевать	860	89 %
совместить	385	87 %
предотвратить	792	86 %
воссоздать	248	84 %
помыслить	129	84 %
соблюсти	200	84 %
соотнести	118	84 %
возместить	304	83 %
восполнить	171	80 %
подработать	91	80 %
сразиться	108	80 %
устранить	686	80 %

Сравнивая списки, легко увидеть две пары видовых коррелятов: *плевать — наплевать* и *соблюдать — соблюсти*. Остальные глаголы в принципе тоже имеют видовые корреляты, но доля инфинитивов в их грамматическом профиле укладывается в стандартные рамки распределения, ограниченные полуторным интерквартильным расстоянием. Тем не менее мы видим, что некоторые глаголы в двух таблицах похожи по семантике, ср., например, *исправлять*, *переделывать* и *устранить*.

Пара *плевать — наплевать* единственная, у которой частота формы инфинитива объясняется употреблением в идиоматической инфинитивной конструкции, ср. *<мне> плевать!*, *да наплевать!* (90 % употреблений и 100 % употреблений в НКРЯ соответственно).

Остальные глаголы ассоциируются с модальными конструкциями вида *надо* / *нужно* / *должен* / *мочь* / *можно* / *нельзя* / *приходится* / *придется* / *следует* + инфинитив. Следуя гипотезе Зализняк и Шмелева, мы ожидали бы увидеть среди выбросов НСВ обозначения контролируемых действий, а среди выбросов СВ обозначения неконтролируемых действий, однако это не всегда так. Такие глаголы НСВ, как *ввязываться*, *согласовывать* и *соблюдать*, могут обозначать неконтролируемые ситуации. Напротив, глаголы СВ типа *совместить*, *сразиться*, *устранить* и др. обозначают контролируемые ситуации. Таким образом, гипотеза Зализняк и Шмелева не находит полного подтверждения на наших денных.

Что касается гипотезы Дивьяк, то, согласно ей, мы ожидали бы увидеть среди выбросов НСВ обозначения неспецифических, обобщенных ситуаций, а среди выбросов СВ — обозначения конкретных ситуаций. Противопоставляя специфические и обобщенные ситуации, Дивьяк в качестве одного из диагностических критериев рассматривает определенность участников и обстоятельств ситуации. Сравним следующие примеры:

(4) *По-моему, если ты действительно верующий человек, то, конечно, надо соблюдать*[Impf]*, как велит церковь* [Беременность: Планирование беременности (форум) (2005)];
(5) *Единственное правило, которое вы при этом должны соблюсти*[Pfv]*: стиль вашей одежды должен быть идентичен общему стилю, принятому на фирме* [Л. Стоцкая. Бой-баба или бизнес-леди? (2004) // «Бизнес-журнал», 2004.03.03].

Пример (4) имеет абсолютивное употребление. Поскольку другие детали ситуации также не названы, мы с уверенностью можем отнести его к обобщенным, неспецифическим употреблениям. Пример (5) содержит конкретные детали (определенная бизнес-ситуация, выбор одежды) и находится на шкале ближе к специфическим. Несмотря на то что в целом речь идет об общем совете, как одеваться деловой женщине, пример можно интерпретировать следующим образом: ‘(всегда) действуй так, как в этом конкретном случае’.

Сравним теперь глаголы исправления негативного результата в модальной конструкции *пришлось переделывать*, *исправлять* (НСВ) — *устранить* <*поломку*>, *возместить* <*потерю*>, *воссоздать* <*детали*>, *восполнить* <*недостающие подробности*> (СВ). В примерах (6—8) речь идет о специфической ситуации, об исправлении референциально определенного испорченного или утраченного объекта:

(6) *Операцию тогда сделали неудачно, пришлось переделывать*[Ipfv]*, правда, уже в другом месте и в конце концов я осталась довольна результатом* [Новое лицо, новые губы (2002) // «Домовой», 2002.10.04];
(7) *В результате ошибку Папы Урбана VIII, осудившего Галилея, пришлось исправлять*[Ipfv] *Иоанну Павлу II* [В. Быков, О. Деркач. Книга века (2000)];
(8) *Помимо вывесок и автомобилей пришлось воссоздать*[Pfv] *целый район города, куда можно было бы поместить весь проект* [«Витрина А»: ответный удар (2000) // «Наружная реклама России», 2000.08.17].

Представляется, что модальная инфинитивная конструкция с *пришлось/придется* нечувствительна к фактору специфичности/обобщенности; отличия (8) от (6—7) сводятся к прототипическому аспектуальному противопоставлению: (8) представляет ситуацию как достижение (achievement), а (6—7) — как деятельность. Подставив вместо инфинитивов парные им формы противоположного вида, ср. *пришлось переделывать — пришлось переделать*, *пришлось воссоздать — пришлось воссоздавать*, мы получили бы аналогичную разницу в значении. Таким образом, версия Д. Дивьяк на наших данных также выглядит довольно слабой. Более того, и сами по себе списки глаголов-выбросов СВ и НСВ не кажутся противопоставленными по свойству обозначать специфические/обобщенные ситуации.

Однако, возвращаясь к группе глаголов исправления негативного результата типа *переделывать*, *устранить* и др., обратим внимание, что глаголы НСВ и СВ «притягивают» к себе разные модальные конструкции (см. таблицу с мерами аттракции и репульсии в Приложении 2D). В конструкции с безличным *приходится* /

пришлось / *придется* формы СВ практически не представлены, и это же касается большинства других модальных конструкций, за исключением конструкций с *должен*, *мочь* и *можно*. Конструкции с *должен*, *мочь* и *можно*, с точностью до наоборот, предпочитают соответствующие глаголы СВ.

Отметим, что инфинитивы СВ притягивают не только модальные конструкции, ср. примеры (9), (10) и (11) с глаголом *восполнить*:

(9) *Поэтому мы попытаемся восполнить этот пробел, опираясь на факты и цифры, приведенные в работах современных историков*;
(10) *После занятия можно выпить воды, чтобы восполнить ее потерю*;
(11) *Фруктами истинный дефицит калия восполнить очень тяжело, практически невозможно*.

Это конструкции с глаголами попытки, целевые конструкции и предикативные конструкции. Корпусный анализ Дивьяк (Divjak 2004: 256) показывает, что глаголы попытки строго предпочитают совершенный вид инфинитива несовершенному, представляя ситуацию как достижение. Целевые конструкции с *чтобы* по умолчанию также сочетаются с достижениями, т. е. снова лоббируют употребление инфинитивов СВ.

Завершая раздел, сравним поведение группы выбросов с поведением глаголов в нижней (0—0,5 %) и средней (20—23 %) части распределения. Минимальное количество форм инфинитива СВ имеют глаголы изменения состояния, которые не замечены в тесной связи с модальными, предикативно-адвербиальными, целевыми и т. п. конструкциями, ср. *посерьезнеть*, *посинеть*. В средней части распределения находим ничем не примечательные глаголы типа *лишиться*, *открыть*.

Прошедшее время НСВ

В русских грамматиках утверждается, что прошедшее время НСВ употребляется в первую очередь для обозначения длительных и повторяющихся действий. Вторичные значения включают общефактическое, значение попытки действия и значение аннулированного результата (Пулькина, Захава-Некрасова 1977: 278; Грамматика 1980: 604—611; Wade 1992: 289—293). В этой связи мы хотим проверить гипотезу, что глаголы с необычно большой долей форм прошедшего времени НСВ будут либо дуративами, либо итеративами.

Группа выбросов (см. рис. 55 и табл. 63) включает 13 глаголов и неоднородна по своему составу. Высокая доля форм прошедшего времени у них может объясняться самыми разными факторами — морфологическими, лексическими и конструкционными.

1. Глаголы *слыхать* и *слыть* функционируют как эвиденциальные, а эвиденциальность в типологическом плане ассоциируются с прошедшим временем (Aikhenvald 2003). Как уже указывалось в разделе «Непрошедшее время НСВ», глагол *слыхать* морфологически дефектен, в его парадигме недостает форм не-

прошедшего времени. Вследствие этого он обладает самой большой долей форм прошедшего времени.

Большинство других глаголов в нашем списке не употребляются в формах императива, а значит, в их грамматическом профиле повышается доля остальных форм, в т. ч. и форм прошедшего времени, ср. *просиживал*, *прохаживался*, *белел*, *чернел*, *слыхал*, *слыл*, *унимался*, *надвигался*, *мрачнел*, *свешивался*.

2. *Просиживал* и *прохаживался* имеют хабитуальное значение. Хабитуалис по определению связан с несовершенным видом, а соответствующие глаголы предпочитают употребления в прошедшем времени, так как в их семантику входит перспектива наблюдения за серией (часто дискретных) событий. Тот факт, что глаголы с хабитуальным значением предпочитают прошедшее время, отмечен в (Danaher 2003).

3. Перспектива наблюдения присутствует и в лексических значениях глаголов-выбросов. *Белел*, *чернел*, *мрачнел* сообщают о воспринимаемом наблюдателем местоположении в пространстве объектов белого или черного цвета. Глагол *свешивался* обозначает воспринимаемую наблюдателем конфигурацию объекта (объектов) в пространстве, ср. *свешивалось белье*, *свешивались ноги.* И, наконец, *надвигался* обозначает поступательное приближение к наблюдателю грозы, дождя или, метафорически, войны. Описание таких наблюдений или даже серий наблюдений естественным образом связано с прошедшим временем и несовершенным видом.

4. (*Не*) *помышлял* и (*не*) *унимался* употребляются в основном под отрицанием, а о связи отрицания и несовершенного вида хорошо известно. Семантика этих глаголов включает ожидание наступления некоторого события или состояния, которое не выполняется в течение значительного периода времени, ср. *дождь всё не унимался.*

5. Глаголы *щурился*, *отшучивался*, а также часть контекстов с глаголом *мрачнел* дают интерпретацию наблюдаемого состояния говорящего (глаголы сопровождают прямую речь). О профиле глаголов, сопровождающих прямую речь, см. с. выше.

Бросая краткий взгляд за пределы зоны выбросов, на другом конце распределения (доля форм прошедшего времени ниже 10 %) мы находим либо гномические глаголы, ср. *являлся*, *касался* (см. о них подробнее в разделе «Непрошедшее время НСВ), либо те, которые встречаются по преимуществу в форме императива, ср. *прощался*.

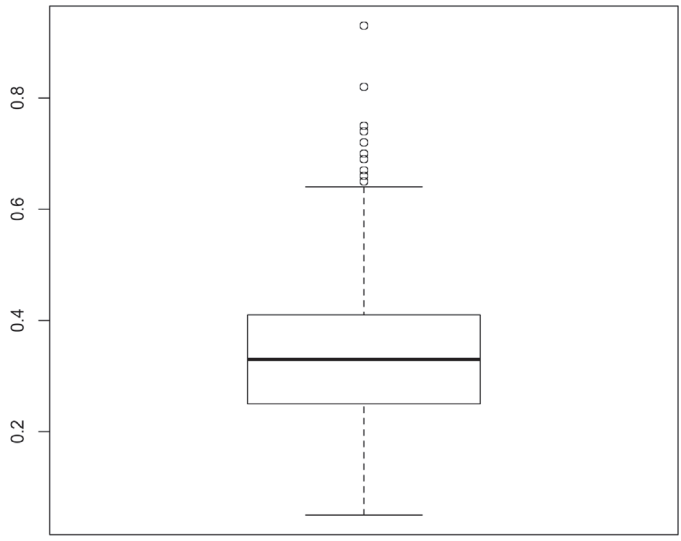


Рис. 55. Распределение глаголов СВ по доле форм прошедшего времени

Таблица 63

Глаголы НСВ с необычно большой долей форм прошедшего времени

Глагол (инфинитив и прош. вр.)	Абс. частота формы	Доля форм, %
слыхать (*слыхал*)	1161	93 %
слыть (*слыл*)	212	72 %
просиживать (*просиживал*)	123	67 %
прохаживаться (*прохаживался*)	207	69 %
белеть (*белел*)	366	70 %
мрачнеть (*мрачнел*)	99	75 %
чернеть (*чернел*)	348	75 %
свешиваться (*свешивался*)	105	74 %
надвигаться (*надвигался*)	260	66 %
помышлять (*помышлял*)	189	69 %
униматься (*унимался*)	381	82 %
щуриться (*щурился*)	196	67 %
отшучиваться (*отшучивался*)	80	74 %

Прошедшее время СВ

Вопреки высокой частотности, формы прошедшего времени СВ почти не удостаиваются в научной литературе описания своей грамматической семантики, помимо тривиального замечания о том, что они используются для обозначения простого единичного завершенного действия (Пулькина, Захава-Некрасова 1977: 279; Грамматика 1980: 604; Wade 1992: 289). В этой связи мы не будем заранее конструировать никакой гипотезы о глаголах-выбросах.

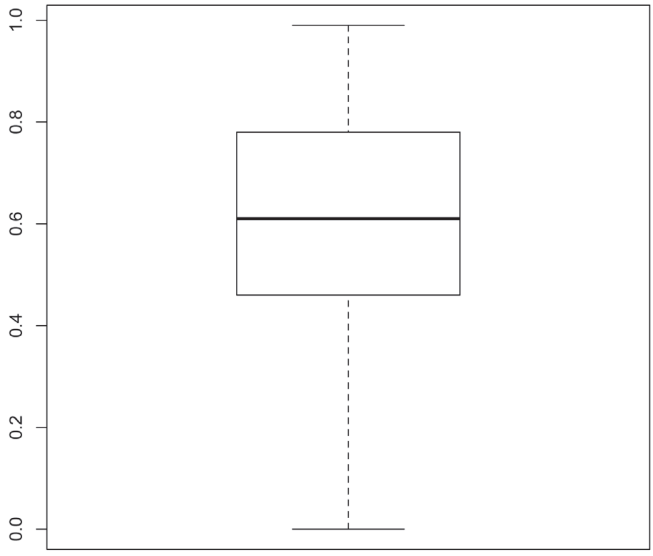


Рис. 56. Распределение глаголов СВ по доле форм инфинитива

Впрочем, в этой группе это и не нужно. В распределении, представленном на рис. 56, крайние точки укладываются в рамки полуторного интерквартильного расстояния, а точнее, средние 50 % точек занимают такой широкий диапазон, что глаголы с долей форм от 0 % до 100 % все еще находятся в пределах интерквартильного расстояния.

Обсуждение результатов

Итак, данные о распределении форм словоизменения в 6 млн контекстов в НКРЯ дают некоторые интересные наблюдения над поведением форм вида, времени и наклонения. Во-первых, это касается сходств и различий в поведении видовых коррелятов, образованных префиксальным способом (простой глагол НСВ — приставочный глагол СВ) и суффиксальным способом (приставочный глагол СВ — вторичный имперфектив). Мы обсуждаем две точки зрения на их счет. Согласно «традиционной» гипотезе и пары типа *писать — написать* и пары типа *переписать — переписывать* признаются чистовидовыми, если в семантику видового

партнера не добавляется никакого нового лексического значения. Согласно гипотезе Исаченко, чистовидовыми признаются только суффиксальные пары. Данные по распределению форм словоизменения не обнаруживают статистически значимых различий между грамматическими профилями префиксальных и суффиксальных пар, тем самым предлагая новый аргумент в поддержку «традиционной» гипотезы. (Анализ других аргументов в пользу первой и второй гипотезы, основанный на профилировании корпусных данных, представлен в (Janda et al. 2013).)

Во-вторых, сравнение индивидуальных грамматических профилей лексем позволило обнаружить «выбросы» — глаголы с необычно высоким или необычно низким содержанием форм словоизменения в той или иной части парадигмы, своеобразные грамматические «идиомы». Эти данные чрезвычайно интересны для активно продолжающейся в русистике дискуссии о взаимодействии лексического значения глаголов с семантикой форм вида, времени и наклонения (ТАМ-форм). В этой работе мы показали, что некоторые данные не покрываются семантическими формулировками, предложенными для тех или иных форм и конструкций И. Б. Шатуновским, Анной А. Зализняк и А. Д. Шмелевым, Д. Дивьяк, и надеемся, что они станут предметом более подробного качественного анализа.

Материал выбросов показывает, как важно разделять в грамматических описаниях лексикализованные и нелексикализованные особенности видо-временных форм. Лексикализованные особенности, как правило, находят параллели в лексическом значении отдельных подгрупп «выбросов». Нелексикализованные, высокопродуктивные значения ТАМ-форм никак себя в этом отношении не проявляют, так как они по определению должны быть равномерно представлены во всей глагольной выборке.

Идея лексикализации грамматического значения позволяет объяснить и то, почему в одной ТАМ-форме допускаются противоположные семантические эффекты, ср., например, эффекты вежливости, грубости / категоричности и нейтральности / фамильярности в значении императива. Несмотря на то что противопоставление видовых коррелятов в одном контексте позволяет однозначно определить тот или иной эффект, он неодинаков в разных лексических группах. Если пара *садитесь — сядьте* известна как классический пример противопоставления вежливого и настойчивого побуждения, то императивы *старайтесь — постарайтесь* оба воспринимаются как вежливые, а *отвали — отваливай* — оба как грубые.

Все группы «выбросов» обнаруживают некоторое число элементов, у которых либо а) ТАМ-форма ассоциируется с определенным сдвигом значения, либо б) ТАМ-форма ассоциируется с определенным устойчивым выражением или конструкцией, либо в) ТАМ-форма ассоциируется с определенным типажом текстов, например с юридическими договорами (глагол является «лексическим маркером» соответствующего типажа, и при этом имеется преференция определенной ТАМ-формы). Именно эти факторы повышают долю соответствующих ТАМ-форм в грамматическом профиле «выбросов». Право на специализацию индивидуальных грамматических форм — лексическую, семантическую, синтаксическую,

дискурсивную, жанровую — всё больше и больше признается в современной лингвистике (говорят об anchoring, или «укоренении» единиц на разных языковых уровнях). Переформулируя метафорически, индивидуальным грамматическим формам разрешается иметь собственные привычки, а не следовать единым правилам — и именно это позволяет языковым механизмам работать более эффективно. Квантитативные корпусные данные предлагают богатые возможности для исследования и документации такой специализации[22].

Методологически важный вопрос для лингвистического анализа корпусных данных — это калибровка уровня гранулярности в поведенческом профиле. Технически ничто не мешает собрать данные на самом дробном уровне, однако мелкие различия могут привнести в исследование много лишних факторов, которые окажутся бесполезными для целей исследования. Другая противоположность — если на выбранном уровень аннотации сжимается слишком много информации и часть структуры данных затеняется. Решение о степени детализации при исследовании грамматических профилей должно приниматься с тем расчетом, чтобы взаимодействие изучаемой грамматической категории с формальной структурой, лексико-семантическими классами, другими категориями и т. п. было бы видно как можно четче. В данном исследовании мы предложили средний уровень детализации: рассматривали противопоставления по времени-наклонению и виду, но не принимали во внимание противопоставления по лицу и числу. Не исключено, что другой дизайн данных в будущих исследованиях поможет обнаружить другие взаимодействующие силы в русской грамматике.

Понятие грамматического профиля, принятое в настоящем исследовании, было определено на субпарадигме финитных форм и инфинитива. Тем самым вне нашего поля зрения были оставлены субпарадигмы причастий и деепричастий, которые распределены в парадигме русского глагола заведомо неравномерно. Однако это не единственное возможное решение. В исследовании (Eckhoff, Janda 2014) все глагольные формы старославянских глаголов были включены в грамматический профиль. Помимо этого, в грамматический профиль можно включать или не включать аналитические формы (сложное будущее и условное наклонение). Для данных русского языка было бы интересно создать квантитативную модель, которая бы компенсировала известную дефектность парадигмы, выравнивая данные для разных классов.

Наконец, грамматические профили, ставшие предметом нашего исследования, строились на процентном распределении частоты грамматических форм относительно их общей суммы, принятой за 100 %. Другой возможный вариант — попарное сопоставление частоты форм (например, в работе (Kuznetsova 2013) изучалось соотношение глагольных форм мужского и женского рода). Обе альтернативы

[22] Заметим также, что ценность данных о специализации признается теперь и в педагогических технологиях, при изучении неродного языка: языковые единицы должны изучаться в наиболее привычных для них контекстах.

имеют свои плюсы и минусы. Если лексема имеет слишком много форм в одной части парадигмы, доля форм в других частях парадигмы автоматически снижается, даже если нет никаких семантических предпосылок к тому, чтобы сочетание этих форм с лексемой было затруднено; «процентный» профиль показывает нам это снижение. Кроме того, как уже было сказано, «процентный» профиль искаженно представляет данные дефектных парадигм. При использовании попарных соотношений профиль становится слишком сложным по структуре, если он включает 4 и более элемента.

Предлагаемый подход грамматического профилирования, безусловно, имеет свои ограничения. Во-первых, в зону «выбросов» часто попадают не самые частотные, но простые по значению глаголы. Объяснение тут состоит в том, что каждое значение полисемичной лексемы имеет свою долю употреблений и имеет свои семантически и т. п. мотивированные преференции в употреблении форм словоизменения. Накладываясь друг на друга, профили отдельных значений дают общую «нейтральную» картину грамматического поведения. Чтобы учесть этот фактор, требуется разметить употребления частных значений в корпусе, что пока представляется неподъемной задачей для широкомасштабного исследования с миллионами точек наблюдения на входе.

Во-вторых, мы не можем утверждать с точностью, что найденные взаимодействия лексического значения с грамматической формой всегда отражаются в частотах грамматического профиля. Иными словами, указывая на слова с некоторыми свойствами внутренней структуры в зоне «выбросов», мы не даем гарантии, что в других частях распределения не найдется слов с такими же свойствами. Пока что мы действовали чисто эмпирически, сравнивая случайную выборку глаголов с низкой и средней концентрацией форм с фокусной группой, — и опять же пока что не нашли слов с совпадающими свойствами. Тем не менее нельзя не согласиться, что требуется более строгая технология изучения силы этого метода.

В-третьих, предложенный подход строится на классификации с заранее известным ответом (up-down clustering, классификация с учителем), т. е. мы имеем заранее заданные классы грамматических противопоставлений (СВ — НСВ, настоящее время — прошедшее время) и анализируем, какие факторы вызывают попадание элемента в тот или иной класс (или, наоборот, работают против). Между тем нельзя с точностью утверждать, что расстояния между грамматическими классами одинаковы по всей ткани языка. В лингвистике допускаются переходные случаи, нечеткие границы категорий и прочие диффузности. Особенно интересен этот вопрос в применении к славянскому виду, как в лексикологическом ключе (см., например, наблюдения об ослабленных видовых противопоставлениях у отдельных русских глаголов в Иткин 2014), так и в исторической и типологической перспективе (генезис вида и сила противопоставления по виду в разных славянских языках). В нашем будущем исследовании (см. Eckhoff et al. 2014) мы предполагаем использовать грамматические профили для bottom-up классификации лексико-грамматических соответствий, с тем чтобы оценить

«расстояние» между видовыми формами в разных частях глагольной лексики (см. рис. 57).

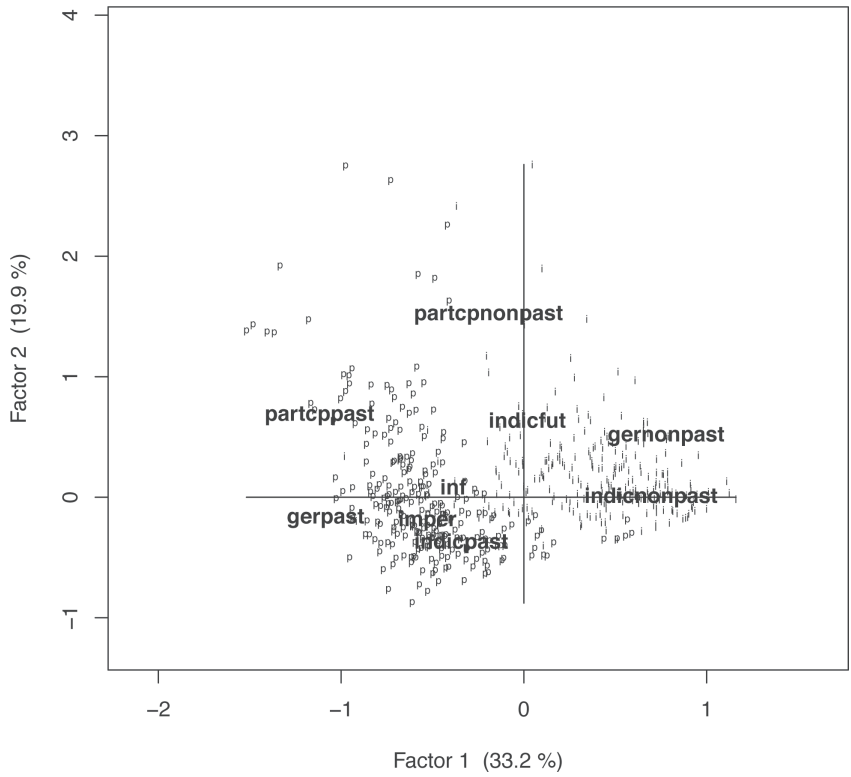


Рис. 57. Матрица correspondence analysis употребления видо-временных форм русских глаголов в публицистических текстах (предварительные данные проекта Eckhoff et al. 2014). i — точки, показывающие положение глаголов НСВ, p — точки, показывающие положение глаголов СВ

При bottom-up классификации (т. е. классификации от данных, с заранее неизвестным результатом) частоты форм словоизменения могут использоваться как факторы, предсказывающие попадание глагола в класс совершенного вида и класс несовершенного вида. Соответственно, кажется перспективным сопоставить пары глаголов, которые хорошо «разводятся» по виду с помощью грамматического профилирования, и пары глаголов с близким расстоянием между видами.

2.2.2. К описанию дистрибуции форм единственного и множественного числа имен существительных

В этой главе мы собираемся применить грамматическое профилирование для изучения русских имен существительных. Нашей задачей будет показать, что грамматический профиль форм числа является лакмусовой бумагой, которая проявляет

особенности лексического значения имени, включая исчисляемость и одушевленность, сочетаемостные особенности, формальные ограничения на образование форм числа, а также общую частоту лексемы.

Грамматический профиль числа у русских имен существительных как класса

В этом разделе рассматриваются общие частотные тенденции, которые будут представлены как доля вхождений форм мн. числа относительно вхождений обоих числовых форм, т. е. общей частоты лексемы (% PL)[23].

В целом в корпусе формы мн. числа употребляются примерно в 4 раза реже, чем формы ед. числа (%PL = 26%). Однако эта количественная оценка не имеет предсказательной силы для отдельно взятого существительного и даже группы слов, так как не принимает в расчет существование трех больших разрядов: существительных singularia tantum, pluralia tantum и имен с полной парадигмой. У имен singularia tantum %PL = 0%, хотя допустимо, что у некоторой небольшой подгруппы (потенциальных singularia tantum) доля мн. числа будет чуть выше. У имен pluralia tantum ситуация ровно противоположная: %PL=100 или около того.

Мы можем предположить, что средняя доля %PL у имен существительных в целом будет зависеть от соотношения классов singularia tantum и pluralia tantum: например, чем больше имен singularia tantum, тем ниже средний показатель %PL.

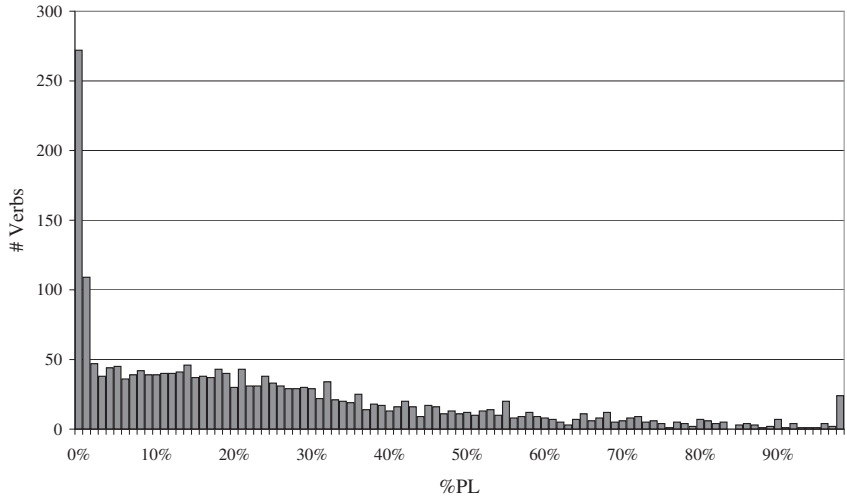


Рис. 58. Гистограмма распределения %PL, порог 100 и более вхождений, шаг 1 %

[23] Данные корпуса со снятой лексико-грамматической омонимией по состоянию на 1 ноября 2013 г.

У имен с полной числовой парадигмой ожидается нормальное распределение % PL. Однако если наложить гистограммы трех классов — существительных singularia tantum, pluralia tantum и имен с полной парадигмой, то можно предположить распределение с приподнятыми краями (так наз. fat-tailed distribution): чем больше слов singularia tantum и pluralia tantum, тем выше края. На рис. 58 показана гистограмма распределения %PL для имен нарицательных, встречающихся в корпусе более 100 раз, с шагом 1 %. На ней можно выделить четыре зоны:

0—3 % (убывание по закону Ципфа);

2—30 % (примерно равное количество глаголов в каждой группе);

30—99 % (постепенное убывание количества глаголов);

100 % (небольшой всплеск на правом краю).

Рис. 59 представляет те же данные, но с шагом в 10 %. Здесь лучше видна тенденция к убыванию по гиперболе, однако в зоне 10—39 % убывание происходит медленнее, чем это предполагает кривая Ципфа.

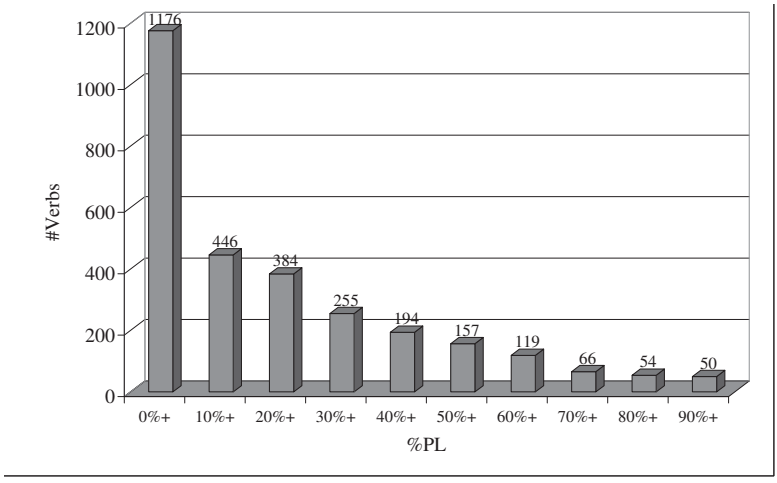


Рис. 59. Гистограмма распределения %PL, порог 100 и более вхождений, шаг 10 %

Однако распределение может несколько меняться, если изменить частотный порог. Если мы добавим в выборку имена существительные, встречающиеся от 25 до 100 раз, то существенно возрастет группа имен с %PL = 0 %, но группы имен с 0 % < %PL < 20 % и с 20 % < %PL < 40 % поменяются местами. На рис. 59 и 60 показаны гистограммы для выборок с порогом от 100 вхождений и более и с порогом от 25 вхождений и более. Мы показываем разбиение с шагом 20 %, причем вклад категорий 0 % и 100 % показан отдельно:

- доля вхождений мн. числа 0 % (singularia tantum),
- доля вхождений мн. числа менее 20 % (преобладают формы ед. числа),
- доля вхождений мн. числа от 20 % до 40 %,

- доля вхождений мн. числа от 40 % до 60 % (примерно равное соотношение),
- доля вхождений мн. числа от 60 % до 80 %,
- доля вхождений мн. числа от 80 % до 100 % (преобладают формы мн. числа),
- доля вхождений мн. числа 100 % (pluralia tantum).

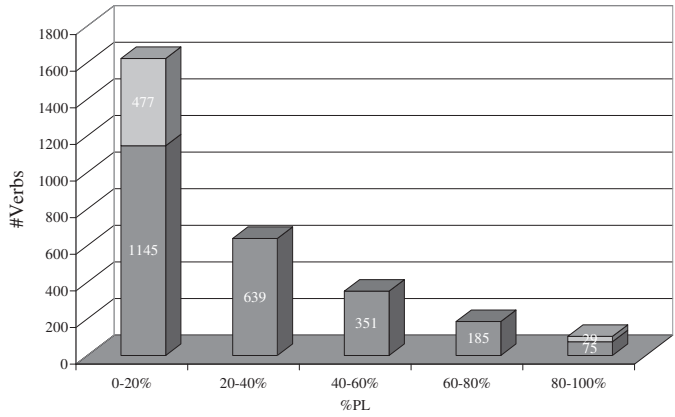


Рис. 60. Гистограмма распределения %PL, порог 100 и более вхождений

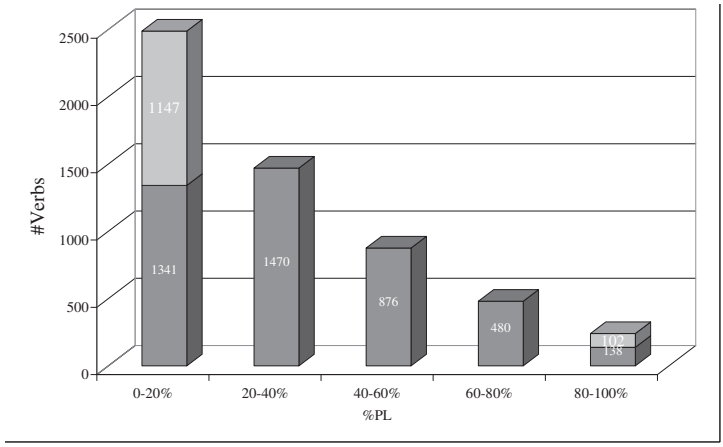


Рис. 61. Гистограмма распределения %PL, порог 25 и более вхождений

Из подсчетов исключены имена, в которых формы ед. числа или формы мн. числа встречаются от 1 до 4 раз, т. е. такие, у которых употребление этих форм, возможно, окказионально, ср.:

(1) <…> *у микрофона — Гарик Осипов, известный также как граф Хортица: писатель, переводчик, радиоведущий, знаток таких* ***музык*** *и таких слов, какие в наше время доводится слышать нечасто* [А. Мунипов. Какая боль, какая боль. Обзор CD (2002) // «Известия», 2002.01.22];

(2) *При газете этой я состоял давно, катал прямо на машинку юбилейные статьи, давал* ***информации*** *обо всех интересных приобретениях и находках нашего музея, консультировал, правил, знал всех и меня знали все* [Ю. О. Домбровский. Хранитель древностей, часть 1 (1964)];

(3) *Однако суды общей юрисдикции не наделены* ***полномочием*** *признавать закон субъекта Федерации, противоречащий федеральному закону, недействительным, т. е. утратившим юридическую силу* <…> [И. Петрухин. Исторический очерк деятельности прокуратуры // «Отечественные записки», 2003];

(4) *Четвертое —* ***взаимоотношение*** *законодательной и исполнительной ветвей власти* [В. Федоткин. Власть и оппозиция (2003) // «Советская Россия», 2003.07.03].

Табл. 63 показывает, как меняется распределение %PL в группах имен с частотой 1000 употреблений и более, от 100 до 999 употреблений, от 50 до 99 употреблений и от 25 до 49 употреблений. Присутствие имен singularia tantum (%PL = 0 %) и pluralia tantum (%PL = 100 %) среди имен с частотой более 1000 вхождений минимально, но эти классы увеличиваются с падением частоты, достигая 34 % и 3 % соответственно у имен с частотой от 25 до 49. Заметим, что среди имен с частотой менее 25 вхождений в корпусе класс singularia tantum продолжает расти, охватывая более половины имен, за счет абстрактных существительных; класс pluralia tantum также растет[24].

Таблица 63

Распределение имен с частотой 1000 и более, от 100 до 999, от 50 до 99, от 25 до 49 по категориям с разной долей PL%

Freq	0 %		0..20 %		20..40 %		40..60 %		60..80 %		80..100 %		100 %	
1000+	4	2 %	111	51 %	59	27 %	26	12 %	9	4 %	7	3 %	2	1 %
100+	217	10 %	703	34 %	574	28 %	321	15 %	179	9 %	62	3 %	25	1 %
50+	307	21 %	324	22 %	401	27 %	240	16 %	124	8 %	36	2 %	28	2 %
25+	619	34 %	203	11 %	436	24 %	289	16 %	168	9 %	33	2 %	47	3 %
Total	1147		1341		1470		876		480		138		102	

У имен с полной парадигмой наблюдается следующая тенденция: с падением общей частоты уменьшается доля имен с %PL от 0 до 20 % и увеличивается доля имен с %PL от 20 до 80 %[25] (доля имен с %PL более 80 % всегда порядка 3%),

[24] 1692 имени singularia tantum (58 %), 170 имен pluralia tantum (6 %) из 2935 имен с частотой от 10 до 24.

[25] Распределение значимо: $\chi^2 = 196{,}629$ при df = 18, $p < 0001$, величина эффекта средняя по (Cohen 1988) (Cramer's V = 0,12). Величина эффекта увеличивается вдвое, если противопоставлены только группы с %PL = 0..20 % и %PL = 40..60 % (Cramer's V = 0,24 (средняя величина эффекта), $\chi^2 = 131{,}018$ при df = 3, $p < 0001$).

см. рис. 62. Растет также медиана и среднее значение %PL с уменьшением общей частоты, см. табл. 64.

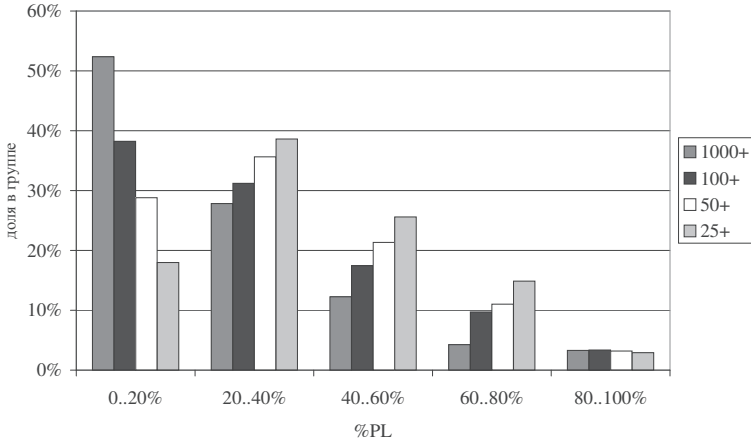


Рис. 62. Распределение %PL у имен с полной парадигмой

Таблица 64

Распределение %PL у имен с полной парадигмой

Freq	0..20 %	20..40 %	40..60 %	60..80 %	80..100 %	Total	Median	Mean
1000+	52 %	28 %	12 %	4 %	3 %	212	19 %	27 %
100+	38 %	31 %	17 %	10 %	3 %	1839	26 %	30 %
50+	29 %	36 %	21 %	11 %	3 %	1125	30 %	35 %
25+	18 %	39 %	26 %	15 %	3 %	1129	35 %	39 %

Объяснение указанной тенденции видится в следующем:

а) среди высокочастотных существительных много имен с «ситуативно дефолтным» единственным числом (см. Ляшевская 2004: 54): так, имена *спина*, *голова*, *сердце* обозначают часть тела, единственную относительно ранее упомянутого в контексте обладателя; *отец*, *жена*, *директор*, *президент* употребляются в контексте обозначения определенной группы или организации, в которой в нормальном случае возможен только один участник с соответствующей ролью; *море*, *квартира*, *сцена*, *дорога*, *кабинет* обозначают определенную и ситуативно единственную для наблюдателя локацию и т. п.;

б) среди высокочастотных существительных часто встречаются многозначные имена. Одно из значений у них является доминирующим и, как правило, предполагает преференцию формы ед. числа. Так, в группу с %PL = 0..20 % попадают имена, у которых наибольшая доля употреблений приходится на обозначение неисчисляемой или опять же дефолтно единственной сущности, ср. *вода*, *земля*, *начало*, *труд*, *опыт*, *уровень*, *война* (по умолчанию — Великая Отечественная), *рынок* (как эко-

номическое понятие). Напротив, среди низкочастотных существительных можно чаще ожидать имена, имеющие только одно значение;

в) среди высокочастотных многозначных существительных присутствуют имена с большой долей конструкционно связанных употреблений (иными словами, в большом числе случаев они употребляются в одной или нескольких определенных конструкциях), где конструкция вынуждает ту или иную форму числа. Например, у существительного «пора» %PL = 59,9 %, причем 45 % приходится на употребления *до сих пор*, *до тех пор* и *с тех пор*. У имени *мера* (%PL = 22,3 %) 39 % употреблений приходится на конструкции *по крайней мере* и *по мере чего-л.*;

г) среди низкочастотных слов несколько больше одушевленных существительных, нежели неодушевленных[26], см. табл. 65. В свою очередь, низкочастотные одушевленные имена чаще обозначают социальные группы в целом и «наборы» участников, ассоциированных с конкретным событием, ср. *испанцы*, *меньшевики*, *декабристы*, *спелеологи*, *устроители* (*праздника*), *очевидцы* (*события*), *россиянки*, *лыжники* (в обоих случаях — о членах спортивной команды), вследствие чего доля %PL у них ожидаемо высока.

Таблица 65

Распределение одушевленных и неодушевленных имен среди высокочастотных и низкочастотных существительных

	25+	50+	100+	1000+	Всего
одуш.	400	322	359	36	1117
неодуш.	1395	1138	1724	183	4440
проц.	+10,9 %	+9,7 %	–14,3 %	–18,2 %	
отклонение	–2,7 %	–2,4 %	+3,6 %	+4,6 %	

Соотношение форм ед. и мн. числа в разных лексических классах

Доля форм мн. числа в зависимости от одушевленности

Табл. 66 демонстрирует, что доля %PL у одушевленных существительных в среднем выше, чем у неодушевленных. Различаются также медиана (38,8 % против 26,9 %) и среднее значение (36,4 % против 27,7 %). Одушевленные существительные обладают большими индивидуализирующими свойствами (таким образом, у них можно ожидать формы обоих чисел), а среди неодушевленных существительных довольно много имен с вещественным и абстрактным значением, что обеспечивает превалирование форм ед. числа. Неодушевленные существительные значительно чаще также представлены в классах singularia и pluralia tantum.

[26] Величина эффекта ниже порога, позволяющего показать, что распределение интересно для нашего исследования (Cramer's V = 0,062, $\chi^2 = 21{,}28$ при df = 3, $p < 0{,}0001$).

Таблица 66

Распределение %PL у одушевленных и неодушевленных имен. Темным подсвечены процентные отклонения выше ожидаемого, светлым — ниже ожидаемого

	0..20 %	20..40 %	40..60 %	60..80 %	80..100 %	Всего ПЧ	0 %	100 %
одуш.	217	319	276	185	44	1041	64	12
неодуш.	1124	1151	599	295	94	3263	1085	90
одуш.	–33,1 %	–10,3 %	30,4 %	59,3 %	216,0 %	20,3 %	–72,3 %	–41,5 %
неодуш.	10,6 %	3,3 %	–9,7 %	–18,9 %	–258,0 %	–5,1 %	18,2 %	10,4 %

Разбиение имен на 3 класса — конкретных одушевленных, конкретных неодушевленных и абстрактных[27], см. табл. 67, показывает, что у конкретных неодушевленных существительных с полной парадигмой доля %PL несколько больше, чем у абстрактных неодушевленных. Вместе с тем медиана и среднее значение достаточно близки: 27,9 % и 29,4 % соответственно у конкретных неодушевленных против 25,5 % и 25,8 % у абстрактных. Если говорить об их представленности среди имен с полной парадигмой в целом, singularia tantum и pluralia tantum, то абстрактные чаще встречаются среди singularia tantum, тогда как конкретные неодушевленные — среди имен с полной парадигмой и имен pluralia tantum.

Таблица 67

Распределение %PL у конкретных одушевленных и неодушевленных и абстрактных имен. Темным подсвечены процентные отклонения выше ожидаемого, светлым — ниже ожидаемого

	0..20 %	20..40 %	40..60 %	60..80 %	80..100 %	Всего ПЧ	0 %	100 %
конкр. одуш.	217	319	274	184	44	1038	63	12
конкр. неод.	654	752	380	195	61	2042	350	66
абстр. неод.	442	372	205	95	33	1147	724	24
конкр. одуш.	–32,70 %	–10,00 %	29,90 %	58,10 %	213,60 %	20,60 %	–72,80 %	–42,20 %
конкр. неод.	3,10 %	7,90 %	–8,40 %	–14,80 %	–252,00 %	7,40 %	–31,50 %	43,90 %
абстр. неод.	24,10 %	–5 %	–12,10 %	–26,10 %	–255,40 %	–21,70 %	83,70 %	–32,10 %

[27] Использовалась лексико-семантическая классификация НКРЯ по разрядам существительных для первого значения имен; переходные и сомнительные случаи исключены из рассмотрения.

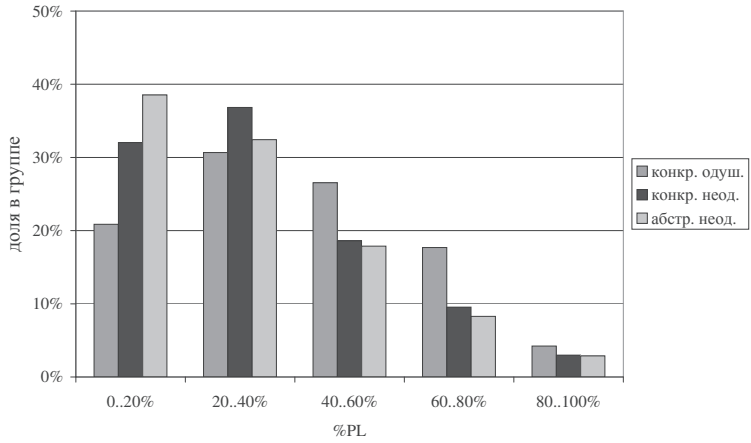


Рис. 63. Распределение %PL у одушевленных, неодушевленных конкретных и абстрактных имен

Следует заметить, что факторы одушевленности и конкретности / абстрактности объясняют далеко не все распределение[28], поэтому следует обратиться к более частным лексическим классам.

Доля форм мн. числа у имен лиц

Имена лиц чаще всего имеют долю %PL от 20 % до 60 % (медиана 39,1 %, среднее 36,4 %), т. е. у них примерно с равным успехом представлены формы ед. и мн. числа. Однако в отдельных лексических группах наблюдается большой разброс. Имена родства в основном встречаются в форме ед. числа, имена профессий показывают распределение, близкое к среднему, а у этнонимов %PL максимальна, см. рис. 64 и табл. 68.

[28] Распределение одушевленных и неодушевленных существительных в подклассах с полной парадигмой значимо (χ^2 = 136,04, df = 4, p < 0,0001), но имеет малую величину эффекта (Cramer's V = 0,178); то же касается их противопоставления в классах с полной парадигмой, sg.tt. и pl.tt. (χ^2 = 200,16, df = 2, p < 0,0001, Cramer's V = 0,1898). Распределения конкретных одушевленных, конкретных неодушевленных и абстрактных имен ведут себя похожим образом: χ^2 = 146,84, df = 8, p < 0,0001, Cramer's V = 0,1318 в пяти подклассах с полной парадигмой и χ^2 = 581,89, df = 4, p < 0,0001, Cramer's V = 0,2307 в классах с полной парадигмой, sg.tt. и pl.tt. В последнем случае можно наблюдать среднюю величину эффекта, согласно (Cohen 1988). В общем, величины эффекта от малой до средней говорят о том, что большая часть наблюдаемых распределений объясняется допустимым варьированием данных при таком большом числе наблюдений.

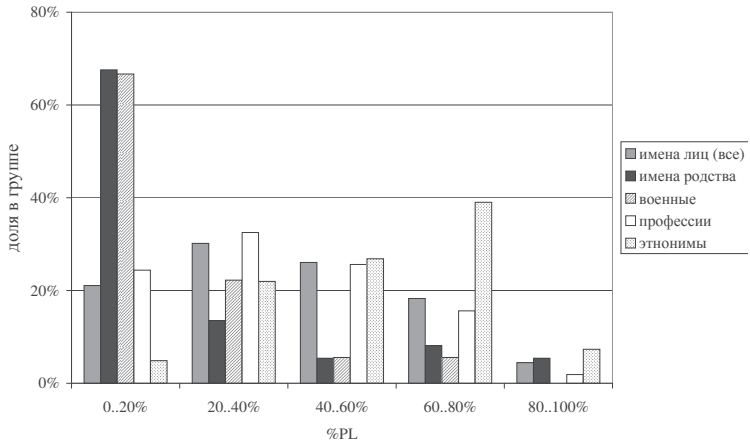


Рис. 64. Распределение %PL у имен лиц

Таблица 68

Распределение %PL у имен лиц

	0..20 %	20..40 %	40..60 %	60..80 %	80..100 %	Всего ПЧ	0 %	100 %
имена лиц (все)	99	195	176	116	32	618	46	12
в т. ч.:								
имена родства	25	5	2	3	2	37	5	
военные	12	4	1	1		18	2	
профессии	39	52	41	25	3	160	2	
этнонимы	2	9	11	16	3	41		
	0..20 %	20..40 %	40..60 %	60..80 %	80..100 %	Всего	Медиана	Среднее
имена лиц (все)	21 %	30 %	26 %	18 %	4 %	100 %	39,1 %	36,4 %
в т. ч.:								
имена родства	68 %	14 %	5 %	8 %	5 %	100 %	9,7 %	13,3 %
военные	67 %	22 %	6 %	6 %	0 %	100 %	10,6 %	21,4 %
профессии	24 %	33 %	26 %	16 %	2 %	100 %	35,0 %	41,2 %
этнонимы	5 %	22 %	27 %	39 %	7 %	100 %	55,3 %	62,2 %
ср. имена (все)	31 %	34 %	20 %	11 %	3 %			

Имена родства

Большинство имен родства употребляется в контексте ситуативной единственности, ср. *папа*, *мама*, *муж*, *жена*, *теща*, *отчим* и др. *Папы*, *мамы*, *мужья*, *жены* и т. д. не образуют «естественных» пар и множеств, т. к. в норме у человека один

родственник с такой ролью[29]. Даже если у персоны, от которой ведется отсчет родства, может быть два или более одного родственника, именуемого *сыновья*, *бабушки*, *внучки*, *невестки*, *племянники* и т. п., в большинстве контекстов обозначается уникальный, единственно определенный в ситуации родственник, например бабушка, живущая в семье, или единственный ребенок, ср.:

(5) «*Ну, мне мама и* ***бабушка*** *рассказывали и я кое-что читал, да и в церкви слышал*» [Митрополит Антоний (Блум). О христианстве (1995)];
(6) *Я, например, для* ***внучки*** *настегала своими руками лоскутное одеяло, зная, что оно будет ее оберегать, давать ей энергию* [Народный костюм: архаика или современность? // «Народное творчество», 2004].

Помимо культурных реалий, частота употребления формы ед. числа может отражать традицию номинации персонажа по имени родства, свойственную художественной литературе, публицистике и бытовой коммуникации:

(7) *Новый год обернулся двойным праздником: Катерина — родная* ***сестра*** *хозяина — приехала погостить из далекой Сибири, правда ненадолго, проездом. Корытин сам ездил на станцию, к поезду, ее встречать и привез прямо к накрытому столу.* ***Сестра*** *на родине не гостила давно. Было о чем поговорить. Вот и просидели у елки далеко за полночь, пели и даже танцевали под музыку. Но по привычке и обычаю людей немолодых* ***сестра*** *хозяина всё равно проснулась довольно рано* [Б. Екимов. Пиночет (1999)];
(8) *Вот, например, он очень любил маленького* ***сына****, но: «Наркотик был дай Бог! Вернее, не дай Бог. Потому и ломка оказалась страшной». Любовь к* ***сыну*** *оказалась наркотиком, а значит, ее нужно вырезать из сердца под корень. Отцовская привязанность принесла только зло:* ***сын*** *превратился в инфантильного жирного борова, который не может расстаться с беззаботным детством и не в силах взять на себя взрослую ответственность за свои поступки* [И. Новикова. Преодоление иллюзий (о романе Александра Мелихова «Любовь к отеческим гробам») // «Октябрь», 2003].

Можно заметить, что некоторые имена родства во мн. числе не обозначают «естественного набора», ср. *тести*, *свекрови*; форма *дедушки* обозначает менее «естественный» набор, чем номинация *дедушка и бабушка* (не имеющая, однако, однословного аналога).

Ряд имен родства не представлен в корпусе формами мн. числа вообще или представлен на уровне окказионализмов (частота <5). Технически, с точки зрения рассматриваемого корпуса, это имена singularia tantum (или «потенциальных sin-

[29] Ср. здесь также любопытное наблюдение С. В. Кодзасова (1987) о том, что в сочиненной конструкции при обозначении двух лиц вида **бывший и нынешний муж* выбирается форма ед. числа.

gularia tantum», согласно Чельцова 1976), ср. *тесть*, *теща*, *мачеха*, *кум*, *прабабка*, *батя*, *папенька*, *матушка*, *папаша*.

В то же время среди имен родства выделяется группа с долей %PL более 50 % — *родитель*, *предок*, *потомок*, *родственник*, *чадо*, *супруг*. Данные существительные — это особые номинации для обозначения класса объектов (ср. *родственники*, *потомки*, *предки*, *родители*) или пары (ср. *родители*, *супруги*). Формы ед. числа у этих слов часто вытеснены другими, более частотными номинациями, ср. *родители* и *мать* / *отец*, *супруг* и *муж* / *жена*, *чада* и *ребенок*, вследствие чего им достается лишь стилистически маркированное употребление, ср. (высок.) *чадо*, (высок. или офиц.-дел.) *родитель*, *супруг*, ср.:

(9) *Целеустремленность родителей, решивших вывести свое* ***чадо*** *в люди, поистине безгранична* [М. Давыдова. Кто в доме хозяин? (2003) // «100 % здоровья», 2003.01.15] — обыгрывание высокого стиля;

(10) ***Супруг****, которого дотоле баловали со всем пылом нерастраченной материнской любви и нежности, вдруг оказывается в собственных глазах «третьим лишним»* [М. Давыдова. Кто в доме хозяин? (2003) // «100 % здоровья», 2003.01.15].

Таблица 69

Имена родства

	Freq	%PL		Freq	%PL		Freq	%PL
матерь	46	0 %	дядюшка	70	6 %	невестка	73	18 %
папенька	62	0 %	зять	110	6 %	родственница	86	23 %
тесть	112	0 %	муж	61	7 %	золовка	99	24 %
прабабка	177	0 %	батюшка	51	8 %	сестра	290	26 %
батя	6445	0 %	тетушка	72	9 %	брат	854	27 %
папа	219	1 %	дочка	52	9 %	прадед	1149	29 %
тетя	124	1 %	жена	271	9 %	внук	68	48 %
дедушка	105	2 %	бабка	86	9 %	супруг	404	49 %
мама	67	2 %	мамаша	54	10 %	чадо	95	63 %
мать	48	3 %	прабабушка	121	10 %	родственник	885	74 %
дядя	48	4 %	сын	144	11 %	потомок	135	77 %
дед	48	4 %	племянник	67	11 %	предок	303	83 %
отец	49	4 %	внучка	80	12 %	родитель	1018	95 %
бабушка	108	4 %	дочь	115	12 %			

Следует заметить, что и у других имен родства, способных обозначать классы — уже не собственно родственников, а лиц одного поколения или социальной группы, доля форм мн. числа значительно выше средней (см. табл. 69), ср. *внуки*, *прадеды*, *братья* и т. п.

(11) *Настроим мы дач и наши* ***внуки*** *и* ***правнуки*** *увидят тут новую жизнь…* [А. П. Чехов. Вишневый сад (1904)];

(12) — ***Братья*** *и* ***сестры****, — проникновенно сказал он, — у меня только что… от нежности содрогнулась душа* [В. Шукшин. Калина красная (1973)].

Таким образом, семантическую иерархию, соответствующую росту доли форм мн. числа, можно представить следующим образом:

ситуативно дефолтная единственность (*мама*)	>	наборы (*сыновья*)	>	классы (*предки*)

Можно предположить и другие функционально-семантические факторы, а именно влияние доли апеллятивных и гипокористических употреблений. Примерно половина имен с долей %PL, равной или близкой нулю, употребляется преимущественно как обращения. Это еще один фактор, предопределяющий преобладание форм ед. числа, так как обращение, как правило, относится к одному лицу, ср. *батя*, *папенька*, *матушка*, *папаша*. Формы *батюшки* (реже *матушки*) во мн. числе утратили свое лексическое значение имени родства и употребляются как междометие.

(13) *Но когда некоторые просвещенные* ***мамаши*** *начинают всё стерилизовать, у ребенка возникают серьезные нарушения микрофлоры в организме, а затем и болезни* [Т. Батенева. Анатолий Воробьев: «Жизнь без микробов была бы невозможна» (2002) // «Известия», 2002.10.02].

Можно было бы ожидать, что уменьшительно-ласкательные номинации (*дочка*, *дедушка*, *папенька*, *матушка*, *мамаша*, *дядюшка* и др.) будут употребляться в ед. числе чаще, чем соответствующие «полные» имена (ср. *дочь*, *дед*, *отец*, *мать*, *дядя*), так как диминутивность и гипокористичность связаны с большей индивидуализацией. Однако корпусные данные не подтверждают эту гипотезу: наблюдаются незначительные расхождения в доле %PL как в ту, так и в другую сторону, которые лежат в пределах допустимого свободного варьирования. Вместе с тем следует заметить, что в целом гипокористические имена имеют более низкую частотность, и поэтому большая их часть оказывается за «порогом достоверности» в 25+ употреблений.

Формальные факторы — давление лексической системы. В парах *супруг — супруга*, *родственник — родственница*, *племянник — племянница* и др. формы мужского рода обозначают как множество лиц мужского пола, так и множество лиц обоего пола, ср. *супруги*, *родственники*, *мои племянники*. Вследствие этого область употребления формы мн. числа женского рода сужается, ограничиваясь достаточно экзотическими употреблениями:

(14) *Дело в том, что со всеми своими тремя* ***супругами*** *(в хронологическом порядке) я познакомился в одном и том же месте!* [С. Ткачева. День влюбленных... (2003) // «100 % здоровья», 2003.01.15], —

и, соответственно, доля форм мн. числа падает.

У имени *матерь* (*Божья*) формы ед. числа выделились из парадигмы имени *мать* / *матери* и в настоящее время противопоставлены ей как морфологически (суффикс *-ер*(*ь*)- в ед. числе), так и по лексическому значению. Таким образом, имя получило формальную дефектность, которая подкрепляется также и семантически: *матерь* обозначает ситуативно единственную (уникальную) персону. Употребления во мн. числе потенциально допустимы, ср.:

(15) *Очень ухоженная территория вокруг, множество скульптурных групп, в т. ч. Фатимская и Лурдская божьи* ***матери*** (Google, autotravel.ru/otklik.php/10072), —

но будут считаться формами имени *мать*.

Этнонимы

Для этнонимов характерно обозначение классов лиц, вследствие чего доля %PL у них чрезвычайно высока (согласно иерархии, приведенной на с. 333). Максимальную долю %PL имеет имя *славянин*, которое только в 4 текстах корпуса имеет референтом конкретную персону, а в остальных случаях употребляется как имя класса, прежде всего, в исторических и публицистических текстах.

Помимо собственно национальности, эти имена во мн. числе обозначают группы спортсменов, как правило членов сборных команд:

(16) *В воскресенье она даже была близка к общей победе в Евротуре, но в итоге заняла второе место, пропустив вперед только* ***финнов*** [А. Демин. Игры разума. Российские хоккеисты победили за явным преимуществом (2003) // «Известия», 2003.02.09].

Подобно именам родства, этнонимы используются и для обозначения отдельных лиц, прежде всего в качестве номинации персонажа литературного или публицистического сочинения:

(17) *Посвятив целые 2 часа на сие упражнение,* ***швед*** *разобрал свою флейту, вложил ее в ящик и стал раздеваться. В это время защелка двери его приподнялась, и красивый молодой человек высокого росту, в мундире, вошел в комнату. Удивленный* ***швед*** *встал испуганно* [А. С. Пушкин. Арап Петра Великого (1828)];

(18) *Но если у* ***немок***, ***голландок*** *или* ***американок*** *в такой ситуации лидером команды может стать кто-то другой, то наставникам нашей сборной оставалось корить судьбу-злодейку да посыпать голову пеплом* [С. Подушкин. Все мимо. Итоги выступлений в сезоне-2001/2002 женской сборной России по конькам (2002) // «Известия», 2002.04.24].

Номинации лиц женского пола (*цыганка*, *немка*, *француженка*, *американка*) имеют значительно меньшую долю %PL, так как для обозначения класса используются соответствующие номинации лиц мужского пола. Исключение — имя *россиянка*, которое в корпусе используется почти исключительно в текстах спортивной тематики для обозначения женщин-спортсменок.

Этноним *жид* расходится с этнонимами *еврей*, *иудей* по номинативной функции. Он употребляется — в литературных текстах преимущественно XIX века — как обозначение персонажа, а также — в текстах XX века — как пейоративное обращение (в обоих случаях задействована форма ед. числа). Для обозначения класса используются нейтрально окрашенные *еврей* и *иудей*, вследствие чего *жид* имеет низкую долю %PL (23,3 % форм мн. числа, ср. %PL = 63,4 % у *еврей* и 39,7 % *у иудей*[30]).

Таблица 70

Этнонимы

	Freq	%PL		Freq	%PL		Freq	%PL
цыганка	53	17,0 %	цыган	60	46,7 %	европеец	57	68,4 %
немка	46	19,6 %	итальянец	91	47,3 %	араб	61	68,9 %
жид	90	23,3 %	татарин	109	48,6 %	финн	26	69,2 %
француженка	30	26,7 %	китаец	71	50,7 %	россиянка	26	69,2 %
узбек	26	26,9 %	чех	35	51,4 %	чеченец	56	71,4 %
американка	33	27,3 %	грек	161	52,8 %	русский	391	75,2 %
грузин	68	29,4 %	француз	235	55,3 %	американец	290	76,2 %
швед	33	30,3 %	азербайджанец	84	58,3 %	испанец	45	77,8 %
казах	29	34,5 %	англичанин	133	60,9 %	японец	96	78,1 %
латыш	45	37,8 %	поляк	54	61,1 %	украинец	34	79,4 %
иудей	63	39,7 %	эстонец	46	63,0 %	турок	107	81,3 %
калмык	27	44,4 %	еврей	451	63,4 %	мусульманин	85	83,5 %
австралиец	39	46,2 %	немец	1056	67,2 %	славянин	144	91,0 %
кавказец	28	46,4 %	армянин	50	68,0 %			

Имена профессий, занятий, должностей и званий

В распределении имен профессий, занятий, должностей и званий особую роль играет социальный статус, а также атрибуция лица в обозначаемой ситуации как дефолтно единственного элемента (например, имя начальника) *vs.* множества. В группе обозначений воинских должностей и званий это проявляется следующим образом: «руководители» (*главнокомандующий*, *генерал*, *капитан*, *поручик* и т. п.) практически не имеют номинаций во мн. числе (%PL < 20 %), в то время как «рядовые» (*боец*, *солдат*) обозначаются чаще мн. числом (%PL > 60 %).

Названия профессий чаще используются как имя класса (*железнодорожники*, *предприниматели*, *военнослужащие*, *геологи*, *социологи* и т. п.) и, следовательно, имеют высокую долю %PL. Вместе с тем выделяются:

а) названия профессий и занятий в области сервиса (в широком смысле) с дефолтной единственностью — они, как правило, имеют референтом конкретное лицо, которое в обозначаемой ситуации обслуживает других персонажей,

[30] В реальности доля %PL у иудей 71,4 %, так как 28 вхождений содержится в тексте Л. Улицкой «Казус Кукоцкого», выполняя, по сути, функцию имени собственного.

ср. *домработница*, *няня*, *швейцар*, *почтальон*, *участковый*, *таксист*, *повар*, *переводчик*, *концертмейстер*;

б) названия «руководителей» — как и в других случаях, для них также характерна дефолтная единственность, ср. *режиссер*, *дирижер*, *заведующая*, *продюсер* и др.

Названия частей тела и органов

Распределение форм числа у наименований частей тела, как и других частей, подчиняется иерархии:

единственные части	>	парные части и наборы (Ляшевская 2004)

Кроме того, важен статус части тела как активной (подвижной) *vs.* пассивной (неподвижной).

Так, имена пар и наборов имеют %PL от 28 % до 90 %, тогда как имена единственных частей тела имеют %PL < 25 %[31]. У многих названий пар и наборов доля %PL всё же существенно ниже ожидаемой, так как они часто обозначают один (выделенный из пары/набора) элемент в фокусе внимания наблюдателя. Это либо активно двигающаяся часть тела, ср. *стукнуть кулаком*, *махнуть рукой*, *погрозить пальцем*, *чертить ногтем*, либо активно используемая локация, ср. *повесить на плечо*, *сказать на ухо*, *подставить щеку*. Таким образом, в соответствии с иерархией активности *рука* имеет больше форм ед. числа, чем *нога*, а *пальцы* — больше форм ед. числа, чем *зубы*. *Брови* и *губы* — тоже «активные» части тела, однако они активно задействованы именно как пара (ср. *поднять*, *нахмурить*, *сдвинуть брови*, *шевелить губами*, *сказать одними губами*) и поэтому не подчиняются иерархии активности.

Таблица 71

Имена частей тела и органов

Noun	PL, %	PL, abs.f.	Noun	PL, %	PL, abs.f.	Noun	PL, %	PL, abs.f.
подбородок	1 %	2	морда	13 %	28	око	62 %	72
горло	1 %	4	хвост	14 %	54	ноготь	67 %	136
желудок	2 %	3	череп	23 %	56	нога	70 %	2356
грудь	4 %	41	кулак	28 %	116	крыло	70 %	337
шея	5 %	33	ладонь	29 %	157	рог	71 %	93
лоб	5 %	29	локоть	34 %	93	колено	74 %	530
нос	5 %	51	бедро	51 %	62	ноздря	79 %	85
живот	6 %	21	рука	52 %	4588	легкое	80 %	86
рот	6 %	57	ухо	53 %	510	зуб	82 %	751
голова	7 %	387	щека	55 %	302	бровь	87 %	291
спина	8 %	87	висок	55 %	97	губа	87 %	887
сердце	8 %	172	плечо	57 %	1004	глаз	90 %	5860
душа	9 %	224	лапа	58 %	201	веко	90 %	99
лицо	13 %	612	палец	58 %	897	уста	100 %	128

[31] У многозначных имен приведена статистика по всем неодушевленным употреблениям, ср. *крыло*, *лапа*, *хвост*, *язык*.

Конструкционно связанные распределения

Существительное *глаз* не подчиняется иерархии активности, несмотря на то что *глаза* трудно признать пассивным органом. Большую долю занимают употребления этого существительного в конструкциях, обозначающих локацию как способ действия: (*сказать*) *в глаза*, (*выглядеть*) *в глазах* (*кого-л.*), (*произойти*) *на глазах* (*кого-л.*), (*пройти*) *перед глазами.* В этих конструкциях *глаза* конституируются как парный орган зрения, и замена формы мн. числа на форму ед. невозможна без изменения смысла. Конструкции, в которых *глаз* обозначается, прямо или метафорически, как один активный инструмент зрения или жестикуляции, употребляются значительно реже, ср. *одним глазом* (*взглянуть*), (*видно*) *невооруженным глазом*, (*моргнуть* / *подмигнуть одним*) *глазом.*

Наименования транспортных средств

Таблица 72

Наименования транспортных средств[32]

Noun	PL, %	PL, abs.f.	Noun	PL, %	PL, abs.f.	Noun	PL, %	PL, abs.f.
метро	1 %	1	локомотив	20 %	24	корабль	29 %	154
такси	1 %	1	трамвай	20 %	156	трактор	32 %	45
карета	10 %	19	мотоцикл	19 %	66	автомобиль	36 %	246
велосипед	12 %	12	поезд	18 %	63	грузовик	36 %	77
тележка	12 %	12	автобус	17 %	25	танк	75 %	170
пароход	15 %	16	лодка	16 %	39	сани	100 %	136
самолет	22 %	57	вагон	28 %	33			

С точки зрения описания перемещения на транспорте важно отметить, что, как правило, речь идет об одном лице или группе лиц, передвигающихся на одном транспортном средстве. Это согласуется с тем, что медиана распределения доли %PL у имен транспорта 28,3 %, среднее значение — также 28 %. Однако обозначения средств транспорта в тексте выполняют две функции — обозначают не только движущиеся объекты, собственно транспорт, но и местоположение или место действия.

Соответственно, важно, где находится наблюдатель — извне (т. е. наблюдатель смотрит на дорогу или другое пространство, через которое движется средство транспорта) или внутри (в этом случае средство транспорта будет дефолтно определенным и единственным, так как наблюдатель не может находиться внутри более чем одного пространства). Если наблюдатель смотрит на ситуацию извне, то появляется возможность для восприятия перемещения множества средств транспорта. Этим объясняется небольшой разброс в доле %PL у названий пассажирского, повседневного транспорта (есть большая вероятность, что это будет локация ситуации, ср. *в такси*, *трамвае*, *поезде*, *лодке*, *на пароходе*) и у других наименований, ср. (*мчатся*) *самолеты*, *экипажи*, *грузовики*, *трактора*, *танки.*

[32] Имя *локомотив* обозначает также название спортивного клуба.

Имя *метро* обозначает не только само средство транспорта (ср. *ехать в метро*), но также городскую систему и совокупность пространства под землей. Поскольку в одном городе — одно *метро*, это существительное лишь в 1 % контекстов не имеет дефолтной единственности, ср.:

(19) *Вот вы над нами и смеетесь в своих* ***метро*** *и автобусах и так самоутверждаетесь за наш счет, пока мы здесь коченеем и, не покладая рук, производим молоко, сливки, картошку и прочее, чтоб было чем вам наполнить брюхо…* [Б. Окуджава. Искусство кройки и житья (1985)].

Существительное *вагон* обозначает часть транспортного средства и во мн. числе обозначает набор частей, составляющих поезд. Доля %PL у него близка к средней для класса, хотя, по идее, должна была бы быть выше. Это связано с тем, что *вагон* обозначает чаще пространство — место действия, чем собственно транспортный объект.

Может ли грамматический профиль предсказать лексический класс?

Согласно общетипологической гипотезе Смита-Старка, доступность числа для разных лексических групп (т. е. возможность выражать у них числовые противопоставления) определяется иерархией одушевленности:

Speaker > Addressee > Kin > Non-human rational > Human rational > Human non-rational > Animate > Concrete inanimate > Abstract inanimate.

Она предсказывает, что не бывает языков, у которых выражено числовое противопоставление, например, у конкретных одушевленных, но нет противопоставления по числу у имен родства.

Эта гипотеза, однако, не оправдывает себя в объяснении статистического распределения употреблений форм ед. и мн. числа на корпусных данных (Brown et al. 2013). Напротив, данные показывают зависимость распределения от лексической группы: чаще всего формы мн. числа употребляются у названий лиц по «не-интеллектуальному» признаку (ср. *сестры*, *близнецы*), затем — у названий животных и «интеллектуальных» обозначений лиц (ср. *соавторы*); реже всего формы мн. числа употребляются у абстрактных имен, имен родства и местоимений 2-го лица.

Гораздо примечательнее опыт Дж. Гринберга (Greenberg 1974/1990), который наблюдал распределения в 16 группах существительных. Суть опыта Гринберга состояла в том, чтобы распределить имена по классам (при достаточно грубом делении на классы ему удалось расклассифицировать около 50 % имен) и попробовать по доле форм числа предсказать попадание имени в тот или иной класс. По сути, Гринберг искал тот философский камень, который помог бы провести семантическую классификацию по чисто грамматическим основаниям. Несмотря на то что опыт — в том чистом виде, как его понимал Гринберг, — не показал большой эффективности, Гринбергу удалось обнаружить некоторые относительные отклонения частот в отдельных классах.

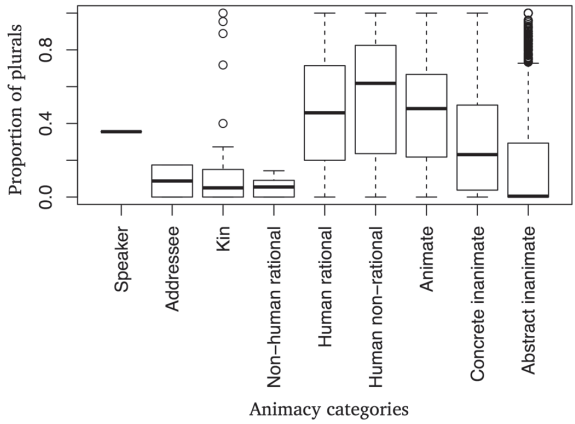


Figure 1: Box plot of proportion of plurals and animacy

Рис. 65. Пропорция форм мн. числа в разных лексических группах (Brown et al. 2013: 235)

Как кажется, наш подход позволяет объяснить эффекты, показанные Гринбергом. В (Greenberg 1974/1990) полагалось, что лексический класс состоит из достаточно однородных семантических элементов. Мы же предполагаем (и доказываем на эмпирическом материале), что, сильно огрубляя, почти каждый лексический класс имеет состав вида {A, A, A, A, B, B, C} (число элементов условно). Категории A, B, C, обладая семантикой, по-разному чувствительной к виду, образуют формы PL в x% × A, y% × B, z% × C случаев. В результате показатель %PL в каждом лексическом классе имеет достаточно большой разброс и классы перекрываются. Если воспроизвести гипотезу Гринберга в нашей формулировке, то он с помощью показателя %PL хочет получить кластер, в котором будут присутствовать четыре элемента {A, A, A, A} рассматриваемого кластера и еще достаточное число элементов вида {T, T, T} и {R} из других кластеров. Ожидание, что этот класс будет семантически однороден, таким образом, теряет смысл.

Corpus Instruments for Russian Grammar Studies

[Olga Lyashevskaya. Korpusnye instrumenty v grammaticheskikh issledovanijakh russkogo jazyka]. Moscow: LRC Publishing House, 2016. 520 pp.

Corpus linguistics can be broadly defined in terms of two partially overlapping research dimensions. On the one hand, corpus linguistics is knowledge of how to compile and annotate linguistic corpora. On the other hand, corpus linguistics is a family of qualitative and quantitative methods of language study based on corpus data. The book presents the first steps taken by Russian corpus linguistics toward the development of language corpora and corpus-based resources as well as their use in grammatical and lexical analysis.

The first part of the book focuses on the annotation of Russian texts at several levels: lemmas, part of speech and inflectional forms, word formation, lexical-semantic classes, syntactic dependencies, semantic roles, frames, and lexical constructions. We discuss various theoretical principles and practical considerations motivating the corpus markup design, provide details on the creation of lexical resources (electronic dictionaries and databases) and text processing software, and consider complicated cases that present challenges for the annotation of corpora both manually and automatically. In most cases we describe the annotation of the Russian National Corpus (RNC, ruscorpora.ru) and its affiliate project FrameBank (framebank.ru).

Frequency data depend not only on the representativeness and balance of texts in a corpus, but also on the rules and tools used for annotation. The book addresses the development of evaluation standards for Russian NLP resources, namely, morphological taggers and dependency parsers. In addition, the book presents several experiments on automatic annotation and disambiguation: lemmatization of word forms not in the dictionary; word sense disambiguation based on vectors formed by lexical, semantic and grammatical cues of context; and semantic role labeling.

The final chapters of the first part of the book outline two types of frequency dictionaries based on the RNC data: a general-purpose frequency dictionary and a lexico-grammatical one.

The second part of the book presents an analysis of corpus data and includes a number of case studies of Russian grammar and lexical-grammatical interaction using quantitative methods. The key concept underlying our analysis is the behavioral profile (Hanks 1996; Divjak, Gries 2006), which is the frequency distribution of variable elements in

a linguistic unit as attested in a corpus. This covers grammatical profiles (the frequency distribution of inflected forms of a word), constructional profiles (the frequency distribution of argument or any other constructions attested for a key predicate), lexical and semantic profiles (the frequency distribution of words and lexical-semantic classes in construction slots or, more generally, in the context of a word), and radial category profiles (the frequency distribution of word senses and word uses across the radial category network of a polysemous unit). We use grammatical, constructional, semantic, and radial category profiling to study tense, aspect and mood specialization of Russian verb forms; to identify singular-oriented and plural-oriented nouns; to investigate factors for prefix choice and prefix variation in natural perfectives (chistovidovye perfectivy); to analyze constraints on the filling of slots in a construction and how this affects the meaning of the construction, taking as an example the Genitive construction of shape and the spatial construction with the preposition *poverkh* 'up and over'.

The quantitative corpus-based techniques used for the analysis vary from simple descriptive statistics (e. g., absolute frequencies, percentages, measures of the central tendency and outliers) to exact Fisher test and logistic regression. We claim that the vector modeling approaches to quantitative grammatical studies in theoretical linguistics are no less effective than in computational linguistics, where they have become a standard tool.

Olga Lyashevskaya is Professor in the School of Linguistics, Higher School of Economics in Moscow, and Senior Researcher in Vinogradov Institute of the Russian Language, Russian Academy of Sciences. She is author of *Semantics of number in Russian* [Semantika russkogo chisla] (2004), *Frequency dictionary of contemporary Russian based on the Russian National Corpus data* [Chastotnyj slovar' sovremennogo russkogo jazyka (na materialakh Nacional'nogo korpusa russkogo jazyka)] (with Serge Sharoff, 2009), and *Why Russian aspectual prefixes aren't empty: Prefixes as verb classifiers* (with Laura Janda, Anna Endresen, Julia Kuznetsova, Anastasia Makarova, Tore Nesset, Svetlana Sokolova, 2013).

Научное издание

Ольга Николаевна Ляшевская

КОРПУСНЫЕ ИНСТРУМЕНТЫ В ГРАММАТИЧЕСКИХ ИССЛЕДОВАНИЯХ РУССКОГО ЯЗЫКА

Корректор Е. Сметанникова
Ведущий редактор В. Столярова
Оригинал-макет и художественное оформление переплета И. Богатыревой

Подписано в печать 01.04.2016. Формат 70×100/16.
Бумага офсетная № 1, печать офсетная. Гарнитура Times.
Усл. печ. л. 42. Тираж 600. Заказ №

Издательский Дом ЯСК
№ госрегистрации 1147746155325
Phone: 8 (495) 624-35-92 E-mail: Lrc.phouse@gmail.com
Site: http://www.lrc-press.ru, http://www.lrc-lib.ru

Оптовая и розничная реализация — магазин «Гнозис».
Тел.: +7 (499) 255-77-57, e-mail: gnosis@pochta.ru
Костюшин Павел Юрьевич (с 10 до 18 ч.).
Адрес: Москва, Турчанинов пер., д. 4