К построению частотной грамматики русского языка

Preliminaries to the Russian frequency grammar

М. В. Копотев, PhD, адъюнкт-профессор, старший научный сотрудник
mihail.kopotev@helsinki.fi
Хельсинкский университет, Хельсинки, Финляндия

Ключевые слова: *частотная грамматика, НКРЯ, ХАНКО*

1. Введение

В 1973 году в работе А. Мустайоки была поставлена задача создания построения грамматики нового типа, которую автор назвал «частотной грамматикой»:

> *Систематическое описание грамматических явлений какого–то (подъ)языка мы называем частотной грамматикой. Она отличается от традиционной грамматики тем, что в ней дается количественная информация о всех категориях и значениях* [Мустайоки 1973: 4].

Часть работы по созданию такой грамматики была проделана в исследовании [Ilola 1989], где была подсчитана *лексическая частотность,* то есть доля лексем, имеющих определенные морфологические характеристики. Например, какова доля существительных в русском языке, сколько существительных мужского рода и т.д. Вторая часть исследования была сформулирована так: «подсчет синтагматической частотности, то есть того, как часто слова с определенными морфологическим параметрами встречаются в тексте» [Ilola 1989: 2]. Эта работа не могла быть выполнена в то время по причине отсутствия соответствующих ресурсов, однако, современный этап лингвистики – развитие языковых корпусов – открывает новые возможности как для создателей корпусов, так и для лингвистов, осуществляющих корпусные исследования. Появление электронных собраний текстов открыло возможности к более точному описанию лексики, мониторингу словарного состава, к созданию частотных словарей и индексов. Последнее является одним из самых распространенных приложений компьютерных исследований и тесно связано с развитием обработки и представления языковых данных. Если говорить о русском языке, то уже довольно давно появились словари, представляющие частотные ха-

рактеристики лексем[1]; существуют исследования, представляющие частотные распределения различных классов слов [Браславский 2001; Ляшевская&Шаров]. Все они так или иначе основаны на лемматизации, и — реже — на автоматической частеречной разметке.

В настоящее время появляется возможность для создания частотных индексов, основанным не только на лексемном уровне. Следуя за развитием методов автоматической обработки языка, русская корпусная лингвистика в настоящий момент предлагает достаточно надежные морфологически аннотированные корпуса, появляются корпуса, содержащие синтаксическую и семантическую разметку. Среди прочего существующие ресурсы позволяют поставить задачу исследования связи частотные характеристик грамматических категорий и лексем [Норман 2003; Janda&Lyashevskaya 2011].

2. Частотная грамматика русского языка

Создание полного частотного грамматического словаря русского языка казалось до недавнего времени трудновыполнимой задачей, несмотря на то, что уже давно существуют исследования, решающие ее на ограниченном материале [Josselson 1963; Никонов 1959; Николаев 1960; Волоцкая и др. 1961; Белоусова 1964; Мустайоки 1973; Ilola&Mustajoki 1989]. С сожалением, все они охватывают лишь определенные грамматические зоны и большинство выполнено на небольшой и чаще всего тематически ограниченной выборке. В работе [Мустайоки 1973] была впервые теоретически обоснована (и частично решена) задача построения общей частотной грамматики русского языка. Прежде всего, автор проводит разделение парадигматической и синтагматической вероятности появления языкового явления. Парадигматическая вероятность задана системой и легко определяется при наличии приемлемой классификации.

Так, например, в рамках шестичленной падежной парадигмы вероятность выбора каждого падежа равна 1/6. В то же время очевидно, что в реальном тексте частотность появления именительного, например, падежа несопоставимо выше, чем предложного. Эту вероятность исследователь, как уже было

[1] См. [Šteinfeldt 1963; Засорина 1977; Лённгрен 1993].

отмечено, называет «синтагматической», и ее решение, безусловно, требует обращения к некоторому массиву текстов. В указанной работе эта задача решается на ограниченном материале одного номера одной газеты.

Среди главных проблем, с которыми сталкивались исследователи, можно назвать малый объем выборки (нерепрезентативность), трудоемкость и фрагментарность исследований. Однако в настоящее время, при наличии современных морфологически размеченных корпусов эта задача может быть решена с большей эффективностью. Хорошо аннотированный корпус позволяет получать детализированные квантитативные данные, отражающие частотные распределения различных грамматических классов в различных группах текстов.

В работе [Копотев 2008] были проанализированы результаты экспериментов, выполненных на материале НКРЯ и ХАНКО и некоторых предыдущих исследований, а также сделаны предварительные замечания о подготовке такой грамматики. Сопоставление выборок из двух независимых корпусов показало, что полученные данные в целом корректны и позволяют с достаточной точностью получать сведения о частотности грамматических категорий. Сравнение со сделанными ранее статистическими подсчетами демонстрируют совпадение корпусных данных при серьезной разнице с данными, полученными представленными в использованных исследованиях 1950–60–ых годов. Представляется, что проведенный эксперимент подтверждает возможность создания на основе существующих корпусов частотной грамматики русского языка. При создании такой грамматики целесообразно учитывать следующее.

1. В основу классификации частотного грамматического словаря должны быть положены принципы, считающиеся общепринятыми в научном сообществе. В то же время необходимо учесть существующую в корпусе разметку. Классификация морфологической системы должна представлять собой компромисс между лингвистической корректностью и возможностями автоматического поиска. В следующей таблице приведены отличия в морфологической разметке ХАНКО и НКРЯ.

Грамматический признак	**ХАНКО**	**НКРЯ**

нарицательные сущ-ые	—	+
сущ-ые pluralia tantum	+	—
звательная форма	—	+
счетная форма (два часа́)	—	+
безличные формы глагола	+	—
второй императив (*пойдемте*)	—	+
будущее аналитическое	+	—
средне-возвратный залог	—	+
двувидовые глаголы	+	—
возвратность глагола	+	—
сослагательное наклонение	+	—
разряды прилагательных	+	—
аналитические формы сравнительные формы прил.	+	—
возвратное местоимения	+	—
сравнительная степень наречия прилагательных	+	—
дробные числительные	+	—
собирательные числительные	+	—
составные числительные	+	—
предикативы	—	+
вводные слова	—	+
местоимение-сущ-ое	—	+
местоимение-прил-ое	—	+
местоимение-наречие	—	+
praedic-pro	—	+

Надо сказать, что отличий от традиционной грамматики в целом немного. В то же время ясно, что каждое отступление от традиции в частотной грамматике должно быть мотивировано. Так например, формы сослагательного наклонения должны быть, по-видимому, учтены, несмотря на то, что в разметке НКРЯ они не учитываются и все глагольные формы на *-л* размечены как формы прошедшего времени; то же касается и аналитических форм и лексем[2].

2. Подсчет целесообразно проводить на основе самого представительного на сегодняшний день корпуса — НКРЯ. Процент ошибок в этом корпусе высок только в «зонах повышенной омонимии», то есть в тех частях грамматиче-

[2] См подробнее Мустайоки&Копотев 2004.

ской системы, которые наиболее трудны для автоматического анализа. Для оценки ошибок аннотирования целесообразно проводить сравнение сопоставимых выборок НКРЯ и ХАНКО, что даст исследователям возможность самостоятельно оценить разницу и решить вопрос о приемлемости результатов.

Литература

Ilola E., Mustaioki A. *Report on Russian morphology as it appears in Zaliznyak's grammatical dictionary*. Helsinki. 1989.

Janda, L., Lyashevskaya O. Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian // Cognitive Linguistics 22:4 (2011), 719-763.

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

Šteinfeldt E. Russian Word Count. Москва. 1963.

Белоусова. Е. А. Статистический анализ глагольных форм (на материале русского языка) // Актуальные вопросы современного языкознания и лингвистическое наследие Е. Д. Поливанова. Т. 1. Самарканд. 1964.

Браславский П. И. Морфологический строй функциональных стилей (на материале документов Internet) // Известия Уральского государственного университета. 2001. № 21. 9–17.

Волоцкая. З.М. Шелимова. И.Н. Шумилина. А.Л. Некоторые количественные данные о формах существительных и глаголов русского языка // Лингвистические исследования по машинному переводу. Москва. 1961. С. 254–261.

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

Копотев М. В. К построению частотной грамматики: русские падежи по корпусным данным // Инструментарий русистики: корпусные подхода. Хельсинки. 2008. С. 136-151.

Лённгрен Л. Частотный словарь современного русского языка. Uppsala. 1993.

Ляшевская, О. Н., Шаров, С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.

Мустайоки А. Опыт составления частотной грамматики русских существительных. Хельсинки. 1973. С. 30 [рукопись]

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

Николаев В., Некоторые данные о частотности употребления падежных форм в современном русском литературном языке // Русский язык в национальной школе. 1960. № 5. С. 19–26.

Никонов. В.А. Статистика падежей // Машинный перевод и прикладная лингвистика. 3(10). Москва. 1959. С. 45–65.

Норман Б. Ю. Грамматическая информация в словаре vs. лексическая информация в грамматике. // Труды по русской и славянской филологии. Лингвистика. VIII (новая серия). Тарту. 2003. С. 148–162.