М.В. Копотев

К ПОСТРОЕНИЮ ЧАСТОТНОЙ ГРАММАТИКИ РУССКОГО ЯЗЫКА

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*[1].

Новый этап в развитии 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 характеристики лексем[2]; существуют исследования, представляющие частотные

---

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

[2] *Šteinfeldt E.* Russian Word Count. Москва. 1963; *Засорина Л.Н.* (ред.). Частотный словарь русского языка. Москва. 1977; *Лённгрен Л.* Частотный словарь современного русского языка. Uppsala. 1993.

распределения различных классов слов[1]. Все они так или иначе основаны на лемматизации (то есть на автоматическом сведении словоформ к начальной форме), и – реже – на автоматической частеречной разметке языковых единиц.

В настоящее время, однако, появляется возможность для создания частотных индексов, основанным не только на лексемном уровне. Следуя за развитием методов автоматической обработки языка, русская корпусная лингвистика в настоящий момент предлагает достаточно надежные морфологически аннотированные корпуса, появляются корпуса, содержащие синтаксическую и семантическую разметку. Среди прочего существующие ресурсы позволяют поставить задачу по созданию частотного индекса, фиксирующего частотные характеристики грамматических категорий в текстах. Задача построения таких частотных грамматик для любого языка важна по целому ряду причин. Обозначу лишь некоторые из них.

1. Еще в начале 70–х годов Дж. Гринбергом была высказана гипотеза, согласно которой различные семантические классы предпочитают различные морфологические формы[2].

На материале словаря Э. Штейнфельдт Дж. Гринберг показал, что русские существительные разных семантических классов с разной вероятностью появляются в том или ином падеже. Позже сходные выводы о поведении числовых форм существи-

---

[1] См.: *Браславский П.И.* Морфологический строй функциональных стилей (на материале документов Internet) // Известия Уральского государственного университета. 2001. № 21. С. 9–17; *Шаров С. А.* ‘Леммы, отсортированные по частоте’. **www.comp.leeds.ac.uk/ssharoff/frqlist/frqlist.html**.

[2] *Greenberg J.H.* The Relation of Frequency to Semantic Feature in a Case Language (Russian) // On language. Selected writings of Joseph H. Greenberg. Stanford. 1974/1990. P. 207–226.

тельного были сделаны Б.Ю. Норманом и другими исследователями[1].

2. Знание частотных характеристик морфологических категорий может быть использовано для создания учебных пособий, ориентированных на реальное языковое употребление, а не на изучение абстрактной языковой системы. Конечно, в большинстве современных учебных курсов этот фактор так или иначе учитывается, однако чаще всего решения опираются на интуицию учителя или автора учебной грамматики[2].

3. Наконец, частотные характеристики грамматических категорий важны для автоматической обработки языка. Это позволяет использовать в программах–анализаторах вероятностные алгоритмы снятия лексической неоднозначности или выбора синонимов, основанные на учете частотности той или иной грамматической формы[3].

---

[1] *Норман Б.Ю.* Грамматическая информация в словаре vs. лексическая информация в грамматике. Труды по русской и славянской филологии. Лингвистика. VIII (новая серия). Тарту. 2003. P. 148–162; *Halliday M.* Quantitative Studies and Probabilities in Grammar // Computational and Quantitative Studies. London–New–York. 1993/2005. P. 130–156; *Arppe A.* Frequency Considerations in Morphology. Revisited – Finnish Verbs Differ. Too // A Man of Measure. Festschrift in Honour of Fred Karlsson in his 60th Birthday. Helsinki. 2006. P. 175–189.

[2] См.: *Biber D., Reppen R.* What does frequency have to do with grammar teaching? // Studies in Second Language Acquisition. 2002. 24/2. P. 199–208. В то же время стоит напомнить. что частотность лексем не может являться единственным фактором отбора материала в учебных целях (см. *Мустайоки А.* О минимизации учебного материала // The teaching of Russian language and literature in Europe. Brussels. 1986. P. 84–98.

[3] *Karlsson F.* Frequency Considerations in Morphology // Zeitschrift für Phonetik. Sprachwissenschaft und Kommunikationsforschung. Berlin. 39/1. P. 19–28; *Arppe A.* The usage patterns and selectional preferences of synonyms in a morphologically rich language // JADT–2002. 6th International Con-

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[1]. С сожалением, надо признать, что все они охватывают лишь определенные грамматические зоны и большинство выполнено на небольшой и чаще всего тематически ограниченной выборке[2]. 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

---

ference on Textual Data Statistical Analysis. March 13–15. 2002. Vol. 1. P. 21–32.

[1] *Josselson H.H.* Подсчет ходовых слов русского языка. Detroit (MI). 1953; *Šteinfeldt E.* Russian Word Count. Москва. 1963; *Никонов В.А.* Статистика падежей // Машинный перевод и прикладная лингвистика. 3 (10). Москва. 1959. С. 45–65; *Николаев В.* Некоторые данные о частотности употребления падежных форм в современном русском литературном языке // Русский язык в национальной школе. 1960. № 5. С. 19–26; *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; *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; *Ilola E., Mustajoki A.* Report on Russian morphology as it appears in Zaliznyak's grammatical dictionary. Helsinki, 1989.

[2] Отмечу, для сравнения, что для английского языка (имеющего наиболее богатую корпусную традицию) такие исследования давно существуют, см.: *Francis W.N., Kucera H.* Frequency analysis of English usage: Lexicon and grammar. Boston, 1992; *Johansson S., Hofland K.* Frequency analysis of English vocabulary and grammar based on the LOB corpus. Vol. 2: Tag combinations and word combinations. Oxford, 1989.

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[1] 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 отличие машинной морфологии ХАНКО и НКРЯ.

---

[1] *Копотев М.В.* К построению частотной грамматики: русские падежи по корпусным данным // Инструментарий русистики: корпусные подхода. Хельсинки, 2008. С. 136–151.

Грам. параметр	**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 прилагательных	+	—
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[1].

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

3. Заключение

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

---

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