I.Bolshakov, A.Gelbukh

Separate Representation of Word Combinations for Singular and Plural Number of Nouns *(In Russian).*

// Proceedings of the International Conference on Applied Linguistics "Dialogue-96", Puschino, May 4 -- 9, 1996, Moscow, Russia.

Statistical prove is given of the necessity of separate representation of word combinations with singular and plural number of nouns when compiling combinatory dictionaries. Namely, with consistent separation of singular and plural forms of nouns in the CrossLexica dictionary system, as little as 30 to 40 per cent in average of included noun collocations are present with both numbers of the same noun.

РАЗДЕЛЬНОЕ ПРЕДСТАВЛЕНИЕ СОЧЕТАЕМОСТИ ЕДИНСТВЕННОГО И МНОЖЕСТВЕННОГО ЧИСЛА СУЩЕСТВИТЕЛЬНЫХ

И.А.Большаков, А.Ф.Гельбух

Приводятся статистические данные, обосновывающие необходимость раздельного представления словарных статей для существительных в единственном и множественном числе при составлении словарей словосочетаний. Именно, при последовательном разделении словарных статей разных чисел в словаре КроссЛексика (600 тыс. словосочетаний), только в 30-40% случаев в словарь вошли словосочетания существительного с другими словами в обоих числах.

В лингвистической литературе давно замечено, что текстуальная сочетаемость многих русских существительных заметно зависит от их грамматического числа. В одних случаях эта зависимость носит жесткий, запретительный характер. Например, можно "делать прыжки", но нельзя - "сделать прыжки". В других случаях эта зависимость более мягка и обусловлена несколькими семантическими факторами, например, можно "анализировать цены", но едва ли "анализировать цену".

Столкнувшись с указанными обстоятельствами, составители словарей словосочетаний используют в качестве элементов таких сочетаний оба грамматических числа. При этом либо приводится число, единственно допустимое или наиболее употребительное при данном связанном слове, либо "ради единообразия" одно из чисел, обычно единственное, выбирается представителем, а другое число используется лишь тогда, когда сочетание с выбранным числом недопустимо.

Ясно, однако, что любая неэксплицитность и непоследовательность при формировании корпуса словосочетаний, даже если она вызвана соображениями "единообразия" или экономии места, вызывает чувство неопределенности у пользователя, особенно - у иноязычного. На наш взгляд, здесь необходим более последовательный и обоснованный подход.

В системе КроссЛексика [1, 2], содержащей в настоящее время 600 тысяч словосочетаний различного характера, два грамматических числа одного существительного (не являющегося ни pluralia tantum, ни singularia tantum) рассматриваются как отдельные единицы представления при описании их сочетаемости (естественно, морфологическая парадигма и парадигматические свойства даются для лексемы в целом).

Необходимо сразу пояснить, что при что при отборе словосочетаний для включения в словарь автор руководствовался не только и не столько соображениями формальной нормативности, сколько рядом соображений типа частотности употребления, нормативности (естественности) описываемой словосочетанием ситуации, семантической простоты, степени идиоматичности, а порой просто свойственным носителю языка чувством "естественности".

Действительно, на периферии сочетаемости языка существует огромное число потенциально возможных и формально допустимых сочетаний (напр. "оштрафовать котенка"), вряд ли вообще когда-либо употреблявшихся. По-видимому, описание этого корпуса возможно скорее с помощью набора правил (вероятно, сложных: например, нельзя "доказать котенка"), а не путем явного перечисления в словаре.

Словарь же служит иной цели: в нем отражено ядро сочетаемости, описание границ которого в виде набора правил, вероятно, невозможно вовсе. Хотя отдельные закономерности, конечно, хорошо известны, - как, например, связь глагольного вида с числом существительного, феномен парных предметов типа "лыжи", тяготение пар существительных одного числа друг к другу ("случаи нарушений"), - они не дают возможности столь же ясно сформулировать правила выбора числа при формировании "естественного" словосочетания, как, например, модель управления глагола позволяет однозначно выбрать нужный падеж.

Обсуждение принятых в системе КроссЛексика критериев отнесения словосочетания к ядру (то есть включения в словарь) выходит за рамки настоящей статьи. Для данного обсуждения существенно, что даже в случае, когда оба числа формально могут сочетаться с каким-либо словом, может оказаться (и, как мы увидим, оказывается довольно часто), что одно из таких

словосочетаний преодолевает порог включения в ядро, а другое - нет. Например, "ходить в школу (школы?)", "купить спичек (спичку?)", "зажечь спичку (спички?)", "завести ребенка (детей?)", "завести кур (курицу?)". Информация о предпочтительности сочетания с одним из чисел во многих случаях необходима для правильного построения предложений, например, иностранцем: ср. "Он часто любовался закатом".

Необходимо подчеркнуть, что в подобных случаях речь идет не о формальной невозможности образования словосочетания и не об отсутствии соответствующего примера (напр. "Они обе решили завести детей."), а лишь о неупотребительности словосочетания или о семантической сложности или ненормативности описываемой ситуации (что связано и с употребительностью). Поскольку словосочетания для системы КроссЛексика отбирались в основном на основе анализа реальных текстов, употребительность можно считать одним из основных критериев отбора материала для данного конкретного словаря.

Доказательством правильности идеи раздельного словарного представления сочетаемости для разных чисел может служить статистика совпадения или несовпадения множеств слов, сочетающихся в текстах с данным словом в разных его числах. В качестве статистического показателя мы выбрали среднее (по словарю системы КроссЛексика) отношение I/U пересечения I к объединению U соответствующих множеств (то есть I есть число слов, сочетающихся с выбранным словом в обоих его числах, а U - хотя бы в одном). Такое отношение было бы равно 100% при полном совпадении этих множеств - в этом случае надобности в раздельном представлении сочетаемости не было бы, - и равно 0 при полном несовпадении - в этом случае сочетаемось разных чисел была бы абсолютно различна. Результаты отдельно по каждому классу связанных с данным существительным слов показаны в первой строке таблицы 1.

	Всего	Эпи-теты	Преди-каты	Упр-щие глаголы	Упр-щие существ
I/U, %	26	39	30	31	28
I/S, %	37	48	42	42	40
I/P, %	52	62	48	48	44
P/(S+P), %	40	41	45	43	44

Табл. 1

Для сравнения во второй строке таблицы указана доля числа всех слов S, сочетающихся с единственным числом данного существительного, которые также сочетаются и с его

множественным числом. Аналогичный показатель для множественного числа P приведен в третьей строке. Наконец, четвертая строка показывает отношение объема сочетаемости множественного числа P к общему объему словарной статьи; цифра, близкая к 50%, означает, что составителем уделялось примерно равное внимание обоим числам (все значения - средние).

Существительные pluralia tantum и singularia tantum были исключены из рассмотрения полностью; также исключались при подсчете средних значений слова, для которых сведения о сочетаемости одного из чисел в словаре отсутствовали.

В целях обеспечения чистоты эксперимента при подсчете статистики из рассмотрения были исключены и связи с теми словами, которые сами сочетаются только с существительными в определенном числе. В этом последенем случае сочетаемость с определенным числом следует считать скорее свойством модели управления такого слова, нежели свойством данного существительного. Например, тот факт, что возможно "люди разбежались", но невозможно "*человек разбежался", при подсчете статистики не влиял на показатели для слова "человек", поскольку это явление характеризует скорее слово "разбежаться". Без отделения таких слов все приведенные цифры получались еще примерно на четверть меньше.

Подобных слов было выделено 436, из них 374 с сочетаемостью только с единственным числом, и 62 - с множественным. При отборе этих слов был принят 95%-ный порог сочетаемости с одним числом в целях учета собирательных существительных (так, слово "многочисленный" входит в 196 словосочетаний с множественным числом и в 3 словосочетания с единственным, напр. со словом "коллектив").

Как видно из таблицы, совпадение сочетаемости в разных числах оказывается не просто далеким от 100%, а скорее близким к совершенно различному, в чем мы и видим доказательство правильности идеи о необходимости раздельного описания такой сочетаемости. Анализ литературы по этому вопросу [3, 4] показывает слабую его изученность и недостаточное внимание к нему со стороны лингвистов и лексикографов.

Мы осознаем, что приведенные нами цифры отражают не только объективные свойства русского языка, но и возможные ошибки и упущения при составлении словаря системы КроссЛексика, тем более что работа над словарем продолжается. Однако, насколько нам известно, статистические исследования подобного рода до сих пор не проводились (ввиду отсутствия соответствующих корпусов словосочетаний), и мы надеемся, что данная работа послужит началом более широкого обсуждения вопроса о возможной

структуре современного словаря словосочетаний (не ограниченного объемом бумажной книги), его назначении, критериях отбора материала и в особенности о природе и границах ядра сочетаемости в языке.

Литература

1. Большаков И.А. Многофункциональный словарь-тезаурус для автоматизированной подготовки русских текстов // НТИ, 1993

2. Большаков И.А., Кассиди П.Дж., Гельбух А.Ф. КроссЛексика - словарь словосочетаний и тезаурус общеупотребительной лексики русского языка // Труды Международного семинара Диалог'95, Казань, 1995

3. Прокопович Н.Н. и др. Именное и глагольное управление в современном русском языке. М.: Русский язык, 1975

4. Словарь сочетаемости слов русского языка. Под ред Н.П. Денисова и В.В. Морковкина. М.: Русский язык, 1983