«Несмотря на» «потому что», или Многокомпонентные единицы в аннотированном корпусе русских текстов

Михаил Копотев
(Хельсинки)

Доклад открывается описанием аннотированного корпуса русских текстов ХАНКО (**www.slav.helsinki.fi/hanco**), важнейшей особенностью которого является учет многокомпонентных единиц (аналитических форм и «эквивалентов слова», или служебных фразем типа «потому что», «несмотря на», «друг друга» и др.). С точки зрения машинной обработки текста разница между ними состоит в том, что аналитические формы сводятся к однокомпонентной лемме, тогда как «эквиваленты слова» предполагают многокомпонентность при лемматизации. В докладе обсуждаются возможности автоматической обработки исследуемых единиц, которые можно разделить на три группы: контактные неомонимичные (числительные, аналитический компаратив и суперлатив, некоторые фраземы); контактные омонимичные («в прошлом», «в общем» и др.); дистантные (аналитическое будущее, сослагательное наклонение, служебные фраземы типа «если…, то»). Относительно надежных результатов можно достичь при автоматическом выделении составных числительных, аналитических компаратива и суперлатива. Выделение сослагательного наклонения и будущего сложного времени требует разработки сложного алгоритма, который, вероятно, не избавит от ошибок. Автоматический анализ «эквивалентов слов» в ряде случаев дает надежные результаты, но большая часть этих единиц предполагает ручную корректировку. Подводя итог наблюдениям, автор напоминает, что точка зрения, согласно которой словоформа равна текстоформе, не соответствует языковой действительности. В то же время, поскольку общее количество таких единиц не превышает 3 %, в ряде случаев ими можно пренебречь.

I. Корпус ХАНКО

Предлагаемые ниже наблюдения проводились на материале аннотированного корпуса русских текстов ХАНКО, который создается на Отделении славянских и балтийских языков и литератур Хельсинкского университета. По этой причине представляется оправданным дать краткое описание корпуса (подробную информацию о нем можно найти в статьях [Копотев, Мустайоки 2003; Kopotev, Mustajoki в печати], а также на сайте **www.slav.helsinki.fi/hanco**). Основные особенности создаваемого корпуса сводятся к следующему. Предполагается, что корпусом будут пользоваться не только лингвисты, но и студенты, учителя, иностранцы, изучающие русский язык. Это, разумеется, не значит, что мы полностью избегаем употребления лингвистических терминов, но выбор параметров поиска осуществляется так, что их знание минимизируется. Из этого логично вытекает следующий принцип: направленность на максимальный охват грамматической информации, а не на объем

материала. Наша задача - предоставить для широкого использования аннотированный корпус, содержащий достаточно точную грамматическую информацию. Понятно, что при таком подходе корпус объемом в несколько миллионов текстоформ потребовал бы колоссальных усилий при ручной обработке материала.

Корпус ХАНКО будет содержать многостороннюю информацию, включающую морфологические, синтаксические и функциональные (семантические) характеристики (последние - в рамках теории функционального синтаксиса А. Мустайоки [Мустайоки в печати]). При этом мы опираемся на устоявшиеся теоретические концепции, которые приняты в известных лингвистических трудах и/или учебной литературе по русской грамматике. В то же время описание некоторых лингвистических явлений во многих, в том числе и фундаментальных исследованиях недостаточно последовательно и часто предполагает множественность квалификаций в силу того, что целый ряд сложных вопросов русской грамматики еще далек от окончательного решения. При подготовке корпуса ХАНКО авторы исходят из, так сказать, «расширенного» толкования грамматических единиц, которое предполагает максимально широкий подход при выделении и квалификации тех или иных языковых явлений. В качестве примера укажем на формы типа «более/менее красивый» или «самый красивый», вопрос о статусе которых остается открытым несмотря на то, что эти многокомпонентные единицы обладают необходимыми свойствами словоформы: отделимостью, дистрибутивной вариативностью и переместимостью, тогда как «более», «менее», «самый» в гораздо большей степени удовлетворяют условиям, характерным для морфем (ср. [Мельчук, 2001: 195-196, 103], ср. однако [Русская грамматика 1980: 547, 562]). В ХАНКО такие единицы интерпретируются как сравнительные аналитические и превосходные аналитические формы соответственно.

Еще одна особенность ХАНКО является следствием такой «расширенной» грамматики: в корпусе предусмотрена возможность более чем одной интерпретации языковых фактов. Не вдаваясь в подробное обсуждение причин собственно языковой и теоретической неопределенности (об этом подробнее см. в [Kopotev, Mustajoki в печати]), укажу только на один частный случай, имеющий непосредственное отношение к теме. Известно, что в русской лингвистической традиции многокомпонентные единицы типа «в продолжение», «потому что» и т. п. могут описываться и как одна единица (предложная, союзная фразема соответственно), и как две отдельные. В ХАНКО реализована и та, и другая возможность.

Наконец, последний и наиболее существенный в рамках настоящего доклада принцип связан с последовательным выделением многокомпонентных единиц. Известно, что при машинной обработке текстов за исходную единицу часто принимается текстоформа, или набор знаков от «от пробела до пробела». Однако очевидно, что такой «орфографический» критерий не может служить надежным основанием для выделения словоформы [Мельчук 2001: 198-199]. При таком подходе определенная часть важной лингвистической информации теряется, что признается неизбежным допущением при автоматическом или полуавтоматическом аннотировании. Если иметь в виду русскую морфологию, то при машинной обработке часто выпадают следующие формы: сослагательное наклонение глагола («прочитал бы»); сложное будущее глагола («буду читать»); аналитические формы прилагательных и наречий («самый быстрый», «более ясно»); составные числительные («сто сорок восемь», «две третьих»); аналитические формы местоимений («ни от кого», «кое у кого»). Кроме того, обычно не учитываются и так называемые «эквиваленты слов», или служебные фраземы типа «в течение», «так как», «несмотря на», «друг друга» и др. (которые, впрочем, не пользуются особым вниманием и в традиционной описательной русистике, см. [Мустайоки, Копотев 2004]). Все эти единицы получили отражение в создаваемом нами корпусе.

Итак, корпус ХАНКО представляет собой хоть и не идеальный (прежде всего, с точки зрения объема), но удобный инструмент для ответа на вопрос, как много мы теряем, не выделяя

многокомпонентные единицы. Учитывая расширенный подход к грамматическим явлениям, принятый создателями ХАНКО, можно попытаться определить максимальный список таких потерь и выявить возможные пути решения.

II. Типы многокомпонентных единиц

Если поставить перед собой цель создания минимально упорядоченной классификации, то естественным, как представляется, выглядит разделение многокомпонентных единиц на две группы.

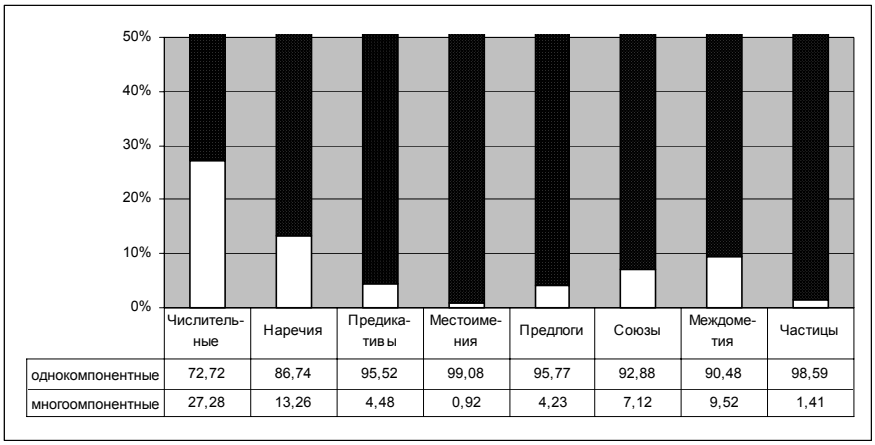
1. Многокомпонентные лексемы («друг друга», «двадцать три», «потому что») сохраняют сложность состава во всех морфологических формах (конечно, если морфологические формы есть вообще). Назовем их, используя термин, введенный В. В. Виноградовым, «эквивалентами слова».
1. Морфологические аналитические формы («буду читать», «самый быстрый», «менее ярко»), то есть отдельные морфологические варианты лексемы, в общем случае представленной одним компонентом.

С точки зрения машинной обработки текста разница между этим группами очевидна и сводится, грубо говоря, к тому, что «эквиваленты слова» состоят из нескольких компонентов во всех формах (и, следовательно, предполагают многокомпонентность при лемматизации), тогда как аналитические формы - только в некоторых (и сводятся к однокомпонентной лемме).

Необходимо указать, что единицы, о которых идет речь, даже при принятом в ХАНКО расширенном подходе составляют менее 3 процентов от общего числа текстоформ. Видимо, для решения определенных задач этими единицами можно пренебречь. В то же время при создании синтаксически или морфологически размеченного корпуса их роль существенно возрастает. Крайний случай искажений такого рода - отсутствие форм условного наклонения или сложного будущего в «машинной грамматике» русского языка. Ясно, что эти искажения влияют и на количество других глагольных форм. Так, в нашем корпусе 209 форм условного наклонения, которые при полностью автоматическом аннотировании добавились бы к 4214 настоящим формам прошедшего времени, что составит погрешность в 4,96 % для форм прошедшего времени и 100 процентов для форм условного наклонения.

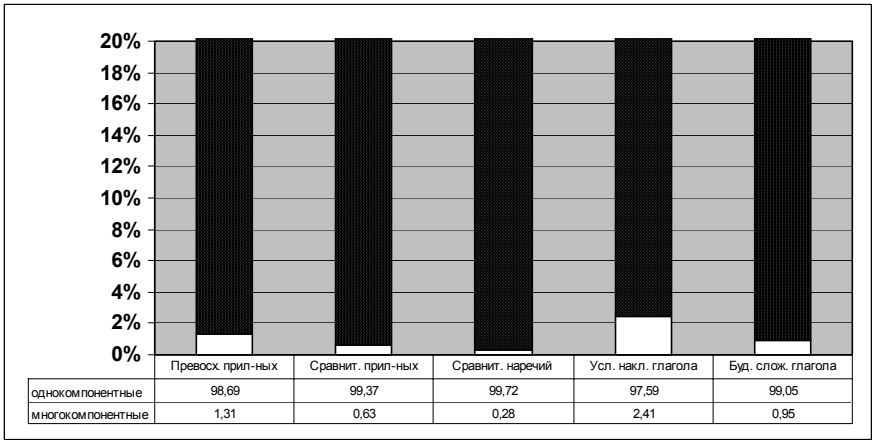
В приведенных ниже таблицах указаны доли многокомпонентных единиц по отношению к общему количеству текстоформ соответствующей части речи (деепричастные и причастные формы для глаголов не учитывались). Представляется, что они не нуждаются в дополнительных комментариях и наглядно демонстрируют, какая часть морфологической информации теряется при машинной обработке.

Табл. 1. «Эквиваленты слов»



	Числительные	Наречия	Предикативы	Местоимения	Предлоги	Союзы	Междометия	Частицы
однокомпонентные	72,72	86,74	95,52	99,08	95,77	92,88	90,48	98,59
многоомпонентные	27,28	13,26	4,48	0,92	4,23	7,12	9,52	1,41

Табл. 2. Аналитические формы слова



	Превосх. прил-ных	Сравнит. прил-ных	Сравнит. наречий	Усл. накл. глагола	Буд. слож. глагола
однокомпонентные	98,69	99,37	99,72	97,59	99,05
многокомпонентные	1,31	0,63	0,28	2,41	0,95

III. Возможные решения

Прежде всего, необходимо отметить, что поиск приемлемых решений для всех случаев, указанных выше, задача невыполнимая. Дело в том, что значительная их часть относится к сфере фразеологии, точнее входит в мало исследованную область служебной фразеологии; см. подробнее [Мустайоки, Копотев 2004]. А фразеологические единицы, как известно, плохо поддаются обобщению и требуют «точечного» анализа. Вторая проблема - различная степень фразеологизации. Если более или менее очевидные фраземы такие, как предлог «несмотря на» или местоимение-реципрок «друг друга» находят отражение в лингвистических описаниях, то многие единицы, являясь переходными по своей природе, еще не получили ясной лингвистической квалификации, отсутствуют точные списки подобных единиц. Понятно, что интуитивное вычленение служебных фразем не может обойтись без ошибок и неточностей. Об этом говорит хотя бы тот факт, что список наречий в [Рогожникова 2003] и [Ефремова 1991] разнится в два раза. Кроме того, ряд явных кандидатов на включение в число служебных фразем отсутствует в словарях вообще. Например, в [Рогожникова 2003: 330] при наличии частицы «само собой», наречий «сам по себе» и «само собой» отсутствует местоимение «сам себе.» Не выделено оно и в [Ефремова 1991; Богданов, Рыжова 1997].
С другой стороны, можно предположить, что сама процедура вычленения служебных фразем должна опираться на корпусные исследования, а именно на составление списка коллокаций, как это сделано, например, для английского языка в [Kjellmer 1994]. Только после этого можно приступать к составлению словаря служебной фразеологии, установив предварительно четкие критерии различение коллокаций и фразем.
Еще один вопрос, актуальный для корпусной лингвистики, связан с определением возможностей автома-тического анализа. Имея в виду именно это, многокомпонентные единицы можно разделить на три группы.

1. Контактные неомонимичные

К этой группе относятся единицы, в которых компоненты непосредственно располагаются друг за другом, и сочетание образующих их знаков (включая пробелы) однозначно идентифицирует определенную фразему. В эту группу входит, прежде всего, класс составных числительных («тридцать три», «33»), а также единицы типа «потому что», «несмотря на». Кажется, самый простой случай представляют собой составные числительные. Как показывает наш материал, значительная их часть представлена в текстах в цифровом виде. Оставляя в стороне вопрос о целесообразности детальной морфологической разметки числительных в цифровой записи («1999 год», «в 22-ом ряду» и т. д.), отмечу, что, строго

говоря, это тоже числительные, и автоматическая разметка их как составных не представляет особых сложностей. То же относится и к числительным, записанным словами («в двадцать втором ряду» и т. д.). Несложный алгоритм позволит проверять, является ли правый сосед числительным определенного типа, и в случае положительного ответа объединять их в сложную единицу. Естественно, что это возможно при условии предварительно проведенного морфологического аннотирования. Другой вариант предполагает обработку числительных по списку.

Относительно легко можно выделить и формы сравнительной/превосходной степени прилагательных и наречий: поиски по шаблонам типа «БОЛЕЕ/МЕНЕЕ + наречие/прилагательное»; «САМЫЙ + прилагательное» позволят вычленить соответственно формы аналитического компаратива наречий, прилагательных и формы аналитического суперлатива. Сведение их к соответствующей однокомпонентной лемме тоже не представляет большой труда.

Другие единицы этой группы не поддаются обобщению. Их выделение требует предварительного анализа каждой единицы в силу того, что возможны (а во многих случаях высоко вероятны) омонимичные «эквивалентам слова» свободные сочетания. И это значит, что такие единицы попадают в группу 2.

2. Контактные омонимичные

Эту, самую многочисленную группу образуют единицы с контактным расположением компонентов, но имеющие омонимичные сочетания двух (или более) полнозначных слов, связанных синтаксически или даже не имеющих непосредственной синтаксической связи.

В качестве иллюстрации приведем данные для наречий «в общем» и «в прошлом». На 8 случаев наречного употребления «в общем» («…рядовой беженский быт поначалу кажется **в общем** сносным») приходится 3 случая омонимичного свободного сочетания («Завтракали и обедали мы **в общем** ресторане нашей части гостиницы»). Для наречия «в прошлом» картина обратная: «эквивалент слова» («Золотой век остался **в прошлом**») употребляется только 6 раз против 13 случаев предложно-падежного сочетания («**В прошлом году** в Европе заговорили о создании собственных европейских сил быстрого реагирования»).

Очевидно, что если такая работа по автоматизации и целесообразна, то она требует предварительных исследований с целью выявить вероятностные характеристики для каждой пары омонимов. Некоторую помощь в этой работе может оказать словарь [Рогожникова 2003], в котором присутствует помета «!Не смешивать», указывающая на возможные трудности. Однако эта помета используется, к сожалению, далеко не всегда последовательно. Так, в статье предлога «в ходе» («в ходе операции») не указано возможное омонимичное предложно-падежное сочетание, ср. «Чувство меры мертвой точкой // обернулось **в ходе поршня**…» (Е. Сабуров).

3. Дистантные

В эту группу входят единицы, компоненты которые располагаются (или могут располагаться) дистантно. В принципе нет никаких ограничений для создания алгоритма поиска в пределах одного предложения, однако это сопряжено с вполне предсказуемой сложностью: поскольку компоненты могут быть разделены любым количеством единиц, велика вероятность совмещения в одной такой «единице» и случайно подпадающих под условия поиска компонентов. Например, можно предположить, что простой поиск по «БЫТЬ.БУДУЩЕЕ_ВРЕМЯ + инфинитив» в пределах одного предложения выделит все формы будущего (и только будущего) времени. Однако материал корпуса ХАНКО показывает, что это далеко не так. Предложение «Предполагается **оставить** лишь те подразделения, что **будут дислоцироваться** в Чечне постоянно» показывает, что необходимо

вводить барьеры, разграничивающие клаузы (очевидно, что запятая не является надежным барьером). Поиск по «БЫТЬ.БУДУЩЕЕ_ВРЕМЯ + ближайший инфинитив справа» тоже не покрывает всех вариантов: ошибка возникает в случаях типа «…мы **должны будем идти** на непозволительные политические уступки». Кроме того, выпадают и менее частотные случаи, в которых инфинитив располагается слева: «…сын и дочка с утра в свой лицей пошли, никто же не думал, что **бомбить будут»**. Наконец, еще одну трудность создают однородные ряды типа «**буду читать, писать»**.

С еще большими сложностями связана обработка сослагательного наклонения. В дополнение к проблемам, описанным выше, добавляется то, что показатель сослагательного наклонение (частица «бы») часто «встроен» в союз «чтобы», который в свою очередь далеко не всегда маркирует клаузу с формой сослагательного наклонения глагола. Данные корпуса ХАНКО показывают, что 192 предложения с «БЫ + *л/*ла/*ло/*ли» и 116 случаев «ЧТОБЫ + *л/*ла/*ло/*ли» далеко не точно соотносятся с реальными 209 случаями употребления условного наклонения.

При работе с «эквивалентами слова» возникают в общем те же проблемы, усугубляемые тем обстоятельством, что практически каждая фразема требует отдельной обработки. Небольшой оптимизм, впрочем, вселяют вероятностные характеристики случайного появления компонентов фраземы. Так например, для союза «если …, то» вероятность правильного автоматического выделения, по нашим наблюдениям, достаточно высока: в корпусе ХАНКО из 58 случаев предложений с «если» и «то», 48 (или 82,76 %) представляют союз «если …, то», а 10 случайное совмещение двух слов в пределах одного предложения. Сопоставимые данные получены и для сложных союзов «как …, так и», «не просто …, но и». Представляется, что, исследовав вероятностные характеристики единиц этой группы с учетом порядка следования компонентов, возможно создать механизм автоматического выделения, выдающий разумно малое количество ошибок.

Подводя итог наблюдениям, хотелось бы еще раз подчеркнуть, что точка зрения, согласно которой словоформа равна текстоформе, не соответствует языковой действительности. В то же время потери при таком подходе составляют не более 3 процентов от общего числа текстоформ, и в ряде случаев ими можно пренебречь. Относительно надежных результатов можно достичь при автоматическом выделении составных числительных (включая записанные цифрам), аналитических компаратива и суперлатива. Выделение сослагательного наклонения и будущего сложного времени глаголов требует разработки сложного алгоритма, который, вероятно, не избавит от всех ошибок. Автоматический анализ «эквивалентов слов» в ряде случаев дает вполне надежные результаты, но большая часть этих единиц все же предполагает корректировку вручную. Если иметь в виду некоторую последовательность при обработке материала, то целесообразно либо вообще отказаться от выделения этих единиц, либо провести серьезную предварительную работу по составлению списка и проверке на вероятность омонимии со свободными сочетаниями.

Литература

1. Kjellmer G. A Dictionary of English Collocations. Oxford: Clarendon Press, 1994.
2. Kopotev M., Mustajoki A. The Helsinki Annotated Corpus of Russian Texts (HANCO): an Attempt to Create a More Adequate Description of Linguistic Facts // The International Journal of Corpus Linguistics (в печати).
3. Богданов С. И., Рыжова Ю. В. Русская служебная лексика. Сводные таблицы. Санкт-Петербург: Издательство Санкт-Петербургского университета, 1997.
4. Ефремова Т. Ф. Толковый словарь служебных частей речи русского языка. М.: Русский язык, 2001.

5. Копотев М. Мустайоки А. Принципы создания Хельсинкского аннотированного корпуса русских текстов (ХАНКО) в сети интернет // Научно-техническая информация. Сер. 2: Информационные системы и процессы. № 6: Корпусная лингвистика в России. 2003. С. 33-37.
6. Мельчук И. А. Курс общей морфологии. Том I. М.: Языки русской культуры, 1997.
7. Мустайоки А., Копотев М. 2003: Эквивалент слова – необходимое понятие в описании языка? // Вопросы языкознания. 2004. № 3 (в печати).
8. Мустайоки А. Теория функционального синтаксиса: от семантических структур к языковым средствам. М.: Языки русской культуры, [в печати].
9. Рогожникова Р. П., Словарь эквивалентов слова. М.: Астрель, АСТ. 2003.
10. Русская грамматика / Под ред. Н. Ю. Шведовой. Тт. I-II. М.: Наука, 1980.