Микросинтаксическая разметка в корпусе русских текстов СинТагРус

Анна Маракасова
ИППИ РАН, НИУ ВШЭ

Леонид Иомдин
ИППИ РАН

Аннотация

Данная статья посвящена основным принципам разметки микросинтаксических единиц в синтаксически аннотированном корпусе СинТагРус. Разметка производится как вручную, так и автоматически, и предполагает два режима: 1) разметка всех микросинтаксических единиц связного текста и 2) разметка отдельных значений конкретной единицы (рассматривается на примере двух единиц: всё равно *и* как будто*).*

1. Введение[1]

Под микросинтаксическими единицами мы здесь понимаем широкий класс конструкций, которые находятся на стыке грамматики и лексики и характеризуются высокой степенью идиоматичности (см., в частности, Iomdin 2007, Иомдин 2013, 2014, 2015). Конструкции микросинтаксиса неоднородны; можно выделить два больших класса. Во-первых, это состоящие из нескольких элементов лексические единицы, или синтаксические фраземы, полноценное толкование которых часто отсутствует в толковых словарях: *как будто, будто бы, разве что, только что, не прочь, кто угодно, не до, что за* и др. Во-вторых, это слабо лексикализованные нестандартные синтаксические конструкции: инфинитивно-модальные конструкции с дативным субъектом (*мне скоро улетать*), различные конструкции с повторяющимися элементами (*знать не знаю, но …, люди как люди, гулять так гулять* и т.д.), конструкции с постоянной и переменной частью (например, *какой-никакой, а X*, как в *Здесь я какой-никакой, а Гельфанд*[2], или *X-у не до Y-а*, как в *Им не до сна*) и др. Разграничение на два класса, впрочем, условно, все зависит от степени лексикализации единицы, а ее не всегда легко определить.

В настоящее время в Лаборатории компьютерной лингвистики ИППИ РАН им. А.А.Харкевича в рамках проекта по микросинтаксису ведутся следующие работы:

1) Создание Микросинтаксического словаря русского языка – лексико-синтаксического ресурса, содержащего сведения о семантике и сочетаемости значительного числа микросинтаксических единиц[3];

2) Разработка микросинтаксической разметки на базе синтаксически аннотированного корпуса русских текстов СинТагРус. Сейчас объем корпуса составляет более 1 млн. словоупотреблений, каждому предложению сопоставлена его синтаксическая структура в виде дерева зависимостей [Апресян и др. 2005, Дьяченко и др. 2015]. Кроме того, в СинТагРус'е содержится лексико-функциональная разметка и предусмотрена возможность поиска по лексическим функциям[4], а часть корпуса снабжена эллиптической разметкой (в предложениях с эллипсисом в древесную структуру вставляется так называемый фантомный узел). В настоящее время ведется работа по созданию анафорически размеченного подкорпуса СинТагРус'а.

[1] Данная работа выполнена при поддержке Российского научного фонда (грант № 16-18-10422)

[2] *http://www.jewish.ru/style/science/2014/07/news994325499.php*

[3] Речь не идет о полном перечне, описать или даже заранее перечислить все единицы микросинтаксиса русского языка невозможно.

[4] Перечень и описание лексических функций см. http://ruscorpora.ru/instruction-syntax.html. Подробнее о лексических функциях, как они понимаются в модели «Смысл ⇔ Текст» и в работах нашей лаборатории, см. в Мельчук 1974, Апресян и Цинман 2002, Apresjan et al. 2007

Далее работа построена следующим образом. Во втором разделе дается краткое лексикографическое описание синтаксической фраземы *всё равно*, в третьем – описываются основные принципы разметки. Последние два раздела посвящены рассмотрению двух частных случаев разметки синтаксических фразем: *всё равно* и *как будто*.

2. Синтаксическая фразема *всё равно*: лингвистический аспект

Чтобы лучше представить, что такое микросинтаксические единицы, целесообразно разобрать какой-нибудь типичный пример. Мы рассмотрим с этой целью многозначную синтаксическую фразему *всё равно*.

Современный толковый словарь русского языка Т.Ф.Ефремовой (2000) выделяет три значения единицы *всё равно*:

1. Наречие:
    1. При любых обстоятельствах, в любом случае.
    2. Несмотря ни на что.
2. Частица:
    1. Усилении противопоставления ранее высказанному; всё-таки.
3. Предикатив:
    1. Оценка какой-л. ситуации как такой, которая не имеет значения, не играет роли; безразлично.

Однако корпусный анализ показал, что у *всё равно* целесообразно выделять следующие три значения (т.е. три лексемы):

1. *всё равно 1* ‘независимо ни от чего, в любом случае’ – *По сути, Хасимото своего добился: он может уйти в отставку уже сейчас и всё равно останется в истории*,
2. *всё равно 2* ‘безразлично, неважно’ – *Зрителям было всё равно, будет он лично выходить на манеж или нет*,
3. *всё равно 3* ‘равносильно’ – *Учить английский без учителя - все равно, что учиться танцевать со стулом*.

В некоторых контекстах лексема *всё равно 1* допускает две возможные интерпретации: ‘субъект хочет совершить некоторое действие и сделает это во что бы то ни стало’ и ‘независимо от того, хочет этого субъект или нет, он совершит некоторое действие’. Рассмотрим предложение, допускающее неоднозначную интерпретацию *всё равно 1*: *Он всё равно сдаст этот экзамен*. При добавлении придаточного, поясняющего ситуацию, неоднозначность снимается: *Он всё равно будет сдавать этот экзамен, каким бы сложным он ни был* и *Он всё равно будет сдавать этот экзамен, иначе не переведется на следующий курс*. Но это еще не даёт нам основания выделять два разных значения, потому что они различимы только в контексте ситуации или в речи, а не на уровне текста.

Лексема *всё равно 3* близка к лексеме *всё равно 2*, *однако всё равно 3* не допускает выражение субъекта дательным падежом; более того, элемент, выражающийся при *всё равно 3* предложно-падежной формой, характеризует не субъекта ситуации (кому все равно) а, скорее, субъекта оценки ситуации (кто считает, что нечто равносильно): *Для взрослого первый год жизни маленького человечка — всё равно что кинохроника, снятая в быстром режиме.*

Второе значение, выделяемое Т.Ф.Ефремовой, мы считаем разновидностью первого (ср. взаимозаменяемость *всё равно* и *всё-таки: Мишу просили остаться, но он все равно ушел* и *Мишу просили остаться, но он всё-таки ушел*). Т.Ф. Ефремова отдельно толкует частицу *всё равно что*. Мы не согласны с тем, что это целостная единица. *Всё равно что*, по нашему мнению, представляет собой сочетание предикатива *всё равно* и управляемого им союза *что*. Точно такой же по структуре является и единица *всё равно как* - здесь тот же предикатив управляет другим союзом: *Сравнивать жизнь с чем-нибудь - все равно как множить яблоки на груши, Это все равно как ловить рыбку в мутной реке, полной хищников, которые охотятся на нее сами*.

3. Общие принципы микросинтаксической разметки

По предварительным данным, в среднем тексте около 25% предложений содержит одну или несколько микросинтаксических единиц. Микросинтаксическая разметка предполагает обнаружение и выделение каждой микросинтаксической единицы в тексте, а в случае многозначности также указание конкретного значения, реализованного в данном контексте. Подготовительным этапом разметки является создания перечня единиц микросинтаксиса и их значений, поэтому процесс аннотирования тесно связан с работой над Микросинтаксическим словарем.

Непосредственно разметка может быть осуществлена как вручную, так и автоматически. Автоматически, с помощью правил или, в перспективе, с помощью алгоритмов машинного обучения, размечаются случаи, где, основываясь на морфологических характеристиках и синтаксической структуре, можно отличить:

(1) случайное соположение элементов микросинтаксической единицы от действительного употребления в качестве целостной единицы (ср., например, употребление микросинтаксической единицы *только что* в значении ‘совсем недавно’ в предложении *Это было название только что опубликованного в*

Нью-Йорке романа Одоевцевой и сочетание частицы *только* и союза *что* в предложении *Отмечу только, что мы писали записки и правительству, и президенту*);

(2) различные значения одной единицы (см. об этом в следующем разделе).

Специально разработанная среда для ручной разметки предполагает два режима.

Во-первых, это разметка связного текста. В данном случае аннотатор последовательно обращается к каждому предложению текста и в специальной форме отмечает микросинтаксические единицы (в том числе и такие, которые он находит впервые и тем самым вносит в создаваемый перечень), их значения, а также случаи ложно-положительного срабатывания (так называемые false positives) – в частности, случаи случайного соположения элементов той или иной единицы. В результате такой разметки мы получаем представление о распространенности тех или иных явлений микросинтаксиса в тексте.

Во-вторых, это аннотация предложений, содержащих определенную, заранее заданную, микросинтаксическую единицу. Предварительно с помощью поиска по корпусу составляется подкорпус предложений с выбранной единицей, а далее разметчик действует так же, как в первом случае. Такая разметка производится с целью выявления (или подтверждения) значений данной единицы и поиска случаев false positive. В следующем разделе представлены результаты, полученные как раз с помощью второго режима разметки.

4. Разметка единицы *всё равно*

Разметка ранее рассмотренной единицы *всё равно* может производиться полностью автоматически. Это возможно, во-первых, благодаря тому, что исключены случаи ложного распознавания местоимения *всё* и краткой формы прилагательного *равный* как целостной единицы (некоторые сильно лексикализованные многословные единицы уже в толково-комбинаторном словаре системы ЭТАП-3, на котором базируется синтаксический анализатор корпуса, представляют отдельную словарную статью, а в данном случае даже две: *всё равно* в значении ограничительного наречия и в значении предикатива).

Во-вторых, для всех трех значений характерна различная синтаксическая структура. Так, в первом значении *всё равно* представлено наречием, которое подчиняется управляющему слову по ограничительному отношению, как на рисунке 1 (*Но рано или поздно инструмент всё равно износится*).

Во втором значении *всё равно* представлено предикативом, являющимся вершиной

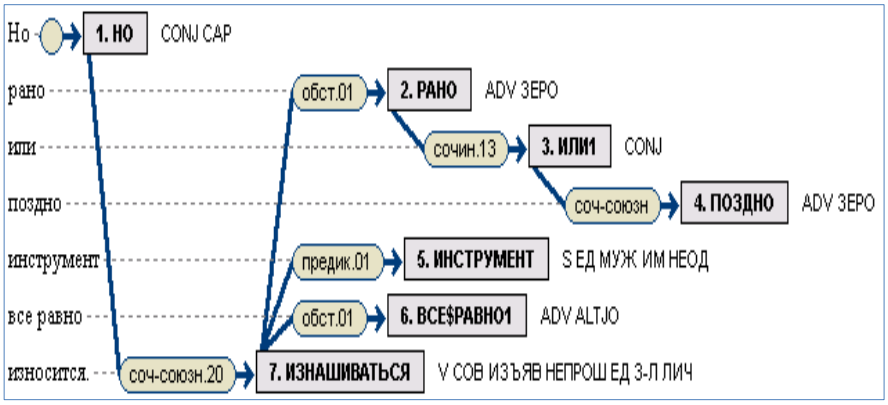

Рисунок 1.

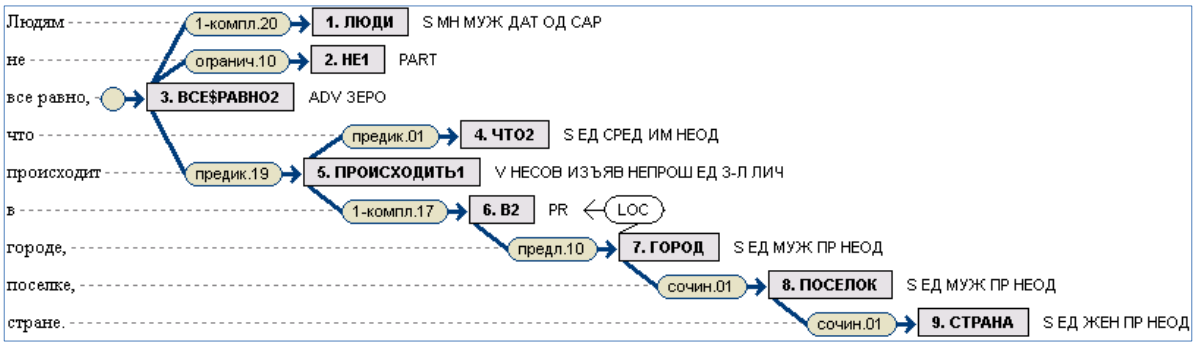

Рисунок 2.

синтаксического дерева (клаузы), которая подчиняет себе подлежащее по предикативному отношению и субъект состояния по 1-му комплетивному отношению (*Людям не все равно, что происходит в городе, поселке, стране*) – рисунок 2.

В своём последнем значении *всё равно* также представлено предикативом, но синтаксическая структура другая. В качестве подлежащего выступает инфинитив, существительное (*И ещё: я был уже популярным артистом, а популярный артист — это всё равно что счёт в банке*) или указательное местоимение-существительное *это* (*Это всё равно что посетить Союз и не увидеть мавзолея*), а в качестве 1-го комплетивного дополнения союз *что*, присоединяющий придаточную клаузу (*Назвать это формализмом – всё равно что назвать черное белым*) – рисунок 3.

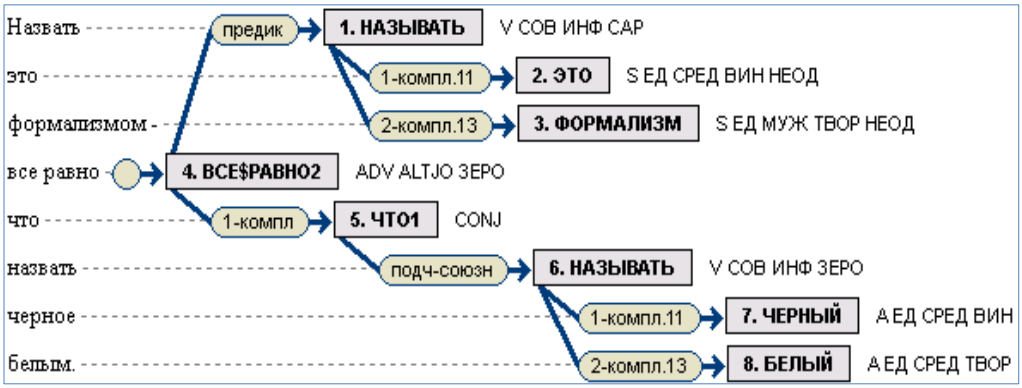

Рисунок 3.

Можно предположить, что случайное сочетание местоимения *всё* и краткой формы прилагательного *равный* или предикатива *равно* будет причиной случаев false positive. В Национальном корпусе русского языка[5] не обнаружено ни одного такого контекста, однако это не исключает возможности такого употребления – ср. *Это всё равно нулю*. Именно такие случаи представляют особенную сложность при автоматической разметке микросинтаксических единиц.

5. Разметка единицы *как будто*

Данная микросинтаксическая единица представлена двумя лексемами. Значение словарной единицы *как будто* – частицы традиционно описывается как выражающее предположительность, неуверенность, сомнение автора в сообщаемом: *Веселый голубой дымок вился над самокруткой и как будто согревал комнату*. Как нам кажется, *как будто* также может иметь оттенок сравнения: *И солнце равнодушно светило в окно - как будто сквозь тела небоскребов*, и выражать не предположительность и неуверенность, а несогласие и отрицание сообщаемого: *Как будто в нашей стране живут одни миллионеры, для которых потеря этой суммы может быть и не в тягость*. Но факт того, что в разных контекстах значение микросинтаксической единицы варьируется, не является достаточным для выделения отдельных лексем.

Во всех выше приведенных случаях *как будто* подчиняется управляющему слову по ограничительному отношению (в толково-комбинаторном словаре системы ЭТАП-3 данная единица представлена одной цельной лексемой).

Вторая лексема *как будто* – это сравнительный союз: *Только в стране народу поубавилось в таком количестве, как будто против нее воевали обе армии - и оккупанты, и защитники*. В таком случае союз *как будто* подчиняет вершину придаточной клаузы по сравнительно-союзному отношению и зависит от словоформы главной клаузы по сравнительному отношению. Поэтому определение, что именно представляют собой две подряд идущие словоформы *как* и *будто* – микросинтаксическую единицу или союз, может быть произведено автоматически с помощью правила (как и разметка различных значений единицы *всё равно*).

Кроме того, при разметке учитывается, что *как будто* может входить в состав другой микросинтаксической единицы – *как будто бы*, значение которой отличается от значения *как будто* лишь степенью неуверенности, предположительности: *И в самом деле, в магазине было как будто бы всё — и вместе с тем ничего не было*. В определенных контекстах *как будто бы* также является союзом: *По дороге мы изо всех сил старались шутить и острить, как будто бы ничего особенного не случилось*.

[5] http://ruscorpora.ru/search-main.html

Заметим, что при линейном следовании словоформ *как, будто* и *бы* всегда имеет место либо частица *как будто бы*, либо союз *как будто бы*. Это важно учитывать при разметке ещё одной синонимичной *как будто* микросинтаксической единицы – *будто бы*: *А в лесу, в горах, вблизи водоемов "дыхание" выравнивается, становится будто бы глубже и ритмичнее. Будто бы*, как и две рассмотренные выше единицы, может выступать сравнительным союзом: *Макроэкономические показатели у Грузии такие, будто бы она тоже добывает газ*.

Литература

Ю.Д. Апресян, И.М. Богуславский, Б.Л. Иомдин и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // НКРЯ: 2003—2005. М.:Индрик, 2005, с. 193—214.

Ю.Д. Апресян, Л.Л. Цинман. Формальная модель перифразирования предложений для систем переработки текстов на естественных языках // Русский язык в научном освещении. 2002 No 4.

П.В. Дяченко, Л.Л. Иомдин, А.В. Лазурский, Л.Г. Митюшин, О.Ю. Подлесская, В.Г. Сизов, Т.И. Фролова, Л.Л. Цинман. Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус) // Национальный корпус русского языка. 10 лет проекту. Труды Института русского языка им. В.В. Виноградова. М., 2015. Вып. 6. С. 272-299.

Т.Ф. Ефремова. Новый словарь русского языка. Толково-словообразовательный. В 2-х томах. М.: Русский язык, 2000

Л.Л. Иомдин. Некоторые микросинтаксические конструкции в русском языке с участием слова *что* в качестве составного элемента // Јужнословенски филолог. Белград: Институт сербского языка Сербской академии наук и искусств, 2013. Т. LXIX. С. 137-147

Л.Л. Иомдин. Хорошо меня там не было: синтаксис и семантика одного класса русских разговорных конструкций // Сб. статей «Grammaticalization and Lexicalization in the Slavic Languages». По материалам Международного симпозиума «Грамматикализация и лексикализация в славянских языках», 11-14 ноября 2011 г. München-Berlin-Washington/D.C.: Verlag Otto Sagner, 2014. Band 55. P. 423-436.

Л.Л. Иомдин. Конструкции микросинтаксиса, образованные русской лексемой раз), SLAVIA, časopis pro slovanskou filologii ročník 84, 2015, sešit 3, s. 291-306.

И.А. Мельчук. Опыт теории лингвистических моделей «Смысл ⇔ Текст». Семантика, синтаксис. М.: Наука, 1974.

Iomdin L.L. Russian Idioms Formed with Interrogative Pronouns and their Syntactic Properties // Meaning – Text Theory 2007. Proceedings of the 3rd International Conference on Meaning – Text Theory. Klagenfurt, Austria, May 21 - 24, 2007 / Wiener Slawistischer Almanach. Sonderband 69. München; Wien, 2007. P.. 179-189.