DOI:10.53921/itas2022_158

От безусловных оборотов к микросинтаксису[1]

А.В.Чага

ИППИ РАН им. А.А.Харкевича, г.Москва

chagachaga@gmail.com

Аннотация. Безусловные обороты – это техническое решение, принятое для сокращения количества единиц в тексте и упрощения лингвистического анализа предложений естественного языка. Поскольку многие из оборотов обладают неконвенциальной семантикой и нетривиальной сочетаемостью, они перерастают в гораздо более самостоятельные единицы, образуя микросинтаксические конструкции. Таким образом микросинтаксис являет собой более принципиальный уровень представления multiword entities, чем безусловные обороты. В работе представлены положительные и отрицательные стороны использования безусловных оборотов, а также область пересечения множеств безусловных оборотов и микросинтаксических единиц. Обсуждаются принципиальные особенности обеих категорий и способы их формального представления. За исключением имён собственных и специальных терминов безусловные обороты системы ЭТАП входят в список микросинтаксических единиц, хотя последних гораздо больше, и подавляющее большинство микроединиц не могут быть отнесены к безусловным оборотам.

Ключевые слова. Микросинтаксис, безусловные обороты, разметка, семантический анализ.

Вводные замечания

В процессе развития NLP многие разработчики пришли к необходимости сокращения количества слов в обрабатываемом тексте, чтобы получать синтаксические, а за ними и семантические структуры более оптимальным путём. Это, в частности, было одним из способов борьбы с комбинаторным взрывом, которым были чреваты многие алгоритмы обработки текста в условиях ограниченности компьютерных ресурсов, таких как память и быстродействие. Было введено понятие «неоднословные единицы» (multiword units), которые рассматривались как единое слово несмотря на наличие в них пробелов. Эти единицы в ряде случаев именовались также «безусловные обороты». Они довольно успешно решали свою задачу по сокращению числа слов в тексте. Сотни таких единиц вводились в компьютерные словари различных систем обработки текста, например, в систему ЭТАП. Такое решение давало значительную экономию при обработке текста, например, выражение ‘во что бы то ни стало’, объединённое в безусловный оборот, сокращало число слов в предложении на целых пять единиц, что заметно сокращало и число рассматриваемых алгоритмом гипотез.

Безусловные обороты

В системе ЭТАП безусловными оборотами называются неоднословные лексические единицы, состоящие из более, чем одного слова, имеющие в своём составе хотя бы один пробел. Порядок компонентов в безусловных оборотах фиксирован. Среди таких оборотов встречаются, в первую очередь:

1) составные предлоги (*в качестве, во главе*);
2) сложные союзы (*в случае, если; по мере того, как; потому что*);
3) наречные выражения (*без устали, где бы то ни было*);
4) частицы (*все же, все ж таки*);
5) заимствованные составные титулы (*ван ден, фон дер*);
6) существительные (*кто бы то ни было, персона нон грата*);
7) терминологические сокращения вроде *б/у* ‘бывший в употреблении’, *м/с* ‘метров в секунду’, *ч/б* ‘чёрно-белый’.

Безусловные обороты имеют применение в задачах, связанных с синтаксической и семантической обработкой текста. В связи с этим в практических целях инвентарь безусловных оборотов может расширяться в конкретных приложениях. Например, введение цельной единицы *красная карточка* или *положение вне игры* оказалось весьма

[1] Работа выполнена при поддержке гранта Министерства науки и высшего образования № 075-15-2020-793 «Компьютерно-лингвистическая платформа нового поколения для цифровой документации русского языка: инфраструктура, ресурсы, научные исследования».

удобным для лингвистического анализа текстов с футбольной тематикой. То же самое касается и часто встречающихся неоднословных имён (*Ким Ир Сен*), брендов (*Интернет Эксплорер*) или топонимов (*Карловы Вары*), а также именных и глагольных выражений типа *что бы то ни было, точка зрения, бросаться в глаза*. Возможность рассмотрения этих цепочек слов в виде единого целого существенно упрощает работу с текстом. Отметим при этом, что единицы, представляемые как безусловные обороты в расширительном смысле, не обладают некомпозициональной семантикой или нестандартным синтаксическим поведением, свойственным для составных предлогов, предложно-именных сочетаний, частиц, сложных союзов и т.д.

Безусловные обороты – это техническое решение, которое предполагает для неоднословной единицы в синтаксической структуре предложения один узел, а в толково-комбинаторном словаре отдельную статью, в которой указывается часть речи единицы, а также другие её характеристики и правила, по которым лингвистический анализатор обрабатывает данное выражение. В частности, такие единицы не обязаны быть неизменяемыми, так что их морфологические свойства приходится специально фиксировать. (В системе ЭТАП это делается достаточно громоздким способом в морфологическом словаре: в обороте *что бы то ни было* или *точка зрения* изменяется первый элемент, в обороте *Ким Ир Сен* – последний элемент, а в обороте *красная карточка* - оба элемента.) Это неудобство, однако, с лихвой компенсируется на синтаксическом и семантическом уровнях обработки текста.

На рис. 1 представлена статья толково-комбинаторного словаря для безусловного оборота *при условии, что*. Это подчинительный союз, располагающийся в препозиции или постпозиции к глаголу-хозяину и имеющий в своём составе запятую, управляющий зависимыми элементами по подчинительно-союзному отношению и зависящий от глагола по обстоятельственной связи. Управляющие свойства фиксируются ссылками на соответствующие правила в словарной статье.

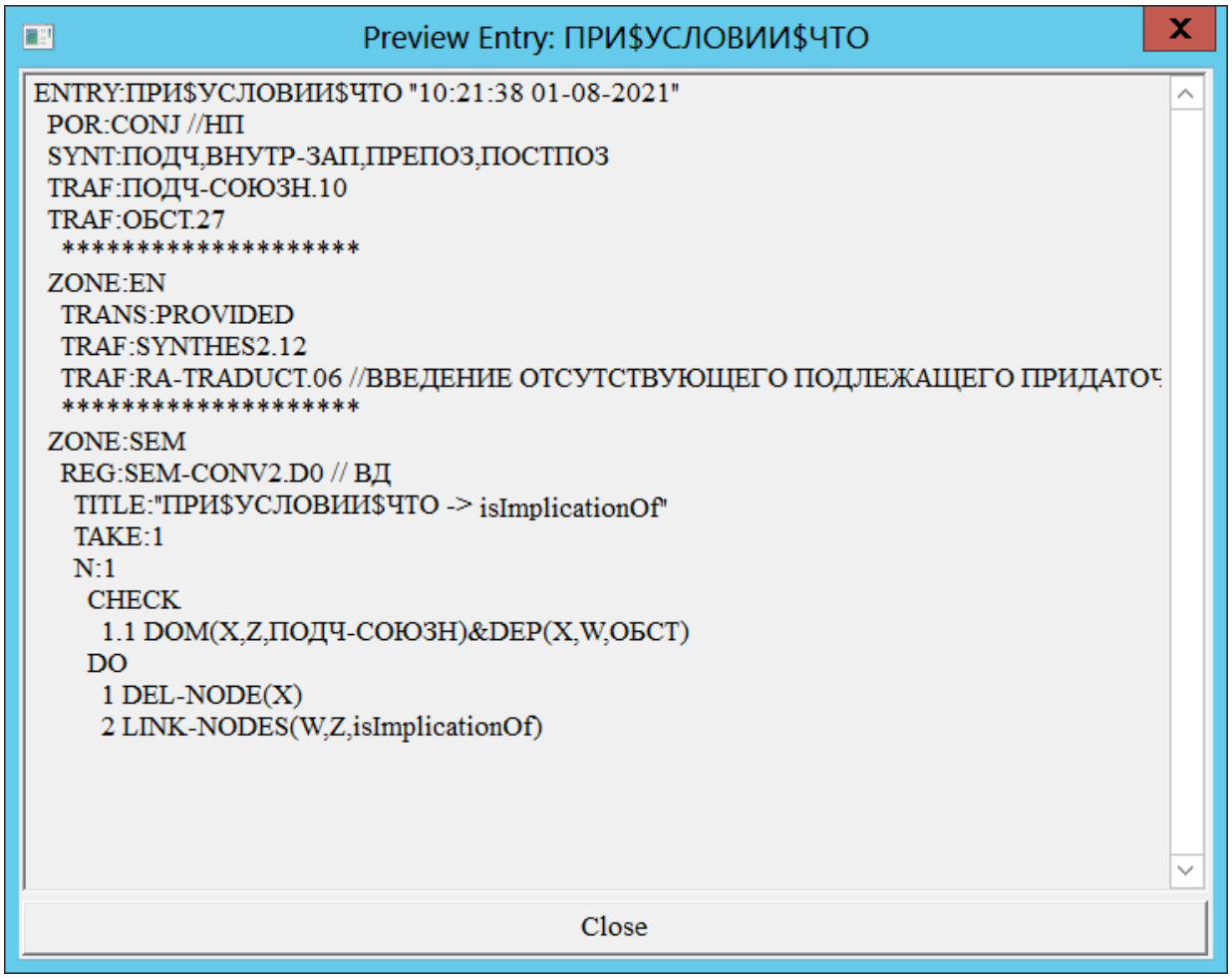

Рис. 1. Статья безусловного оборота в толково-комбинаторном словаре.

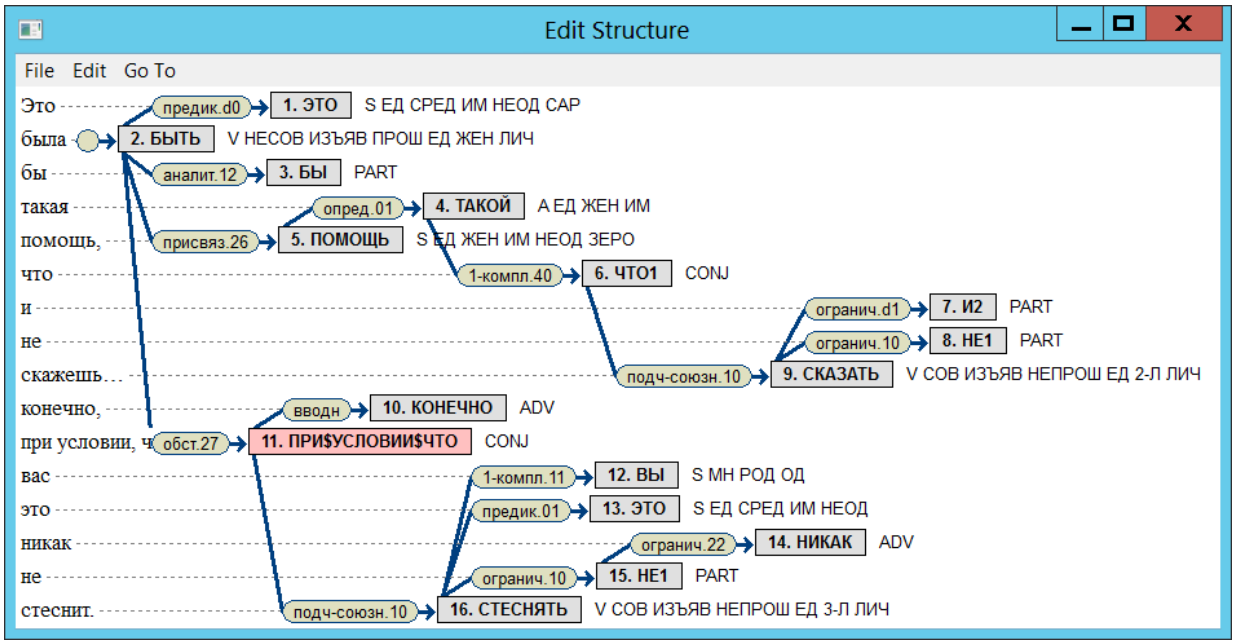

Рис. 2. Синтаксическая структура предложения с безусловным оборотом.

Со временем выяснилось, что, помимо уже отмеченных плюсов, широкое использование безусловных оборотов содержит и значительные издержки. Во-первых, многие обороты омонимичны цепочкам отдельных слов: *к примеру* vs. *нет претензий к примеру*, *то есть* vs. *если болит зуб, то есть трудно* и т.д. Во-вторых, среди оборотов преобладают такие, которые в некоторых случаях могут перемежаться другими словами и выражениями: *несмотря даже на; во что бы то, черт побери, ни стало*. В рамках понимания безусловного оборота как единого и неделимого целого такие явления побороть было нельзя. Таким образом постепенно выкристаллизовалось более широкое понимание неоднословных единиц – микросинтаксические единицы.

Единицы микросинтаксиса

В текущем списке безусловных оборотов, идентифицированных ЭТАПом, примерно 62% составляют микросинтаксические единицы. Фактически, все безусловные обороты, за исключением имён собственных и утилитарно введенных словосочетаний — это конструкции микросинтаксиса. Во многом они близки к фразеологизмам и имеют ряд общих с ними черт: неоднословность, цельность, семантическая некомпозициональность, высокая степень устойчивости, нерегулярность (выход за рамки общих грамматических правил), реинтерпретация грамматических характеристик (компонент выражения, или несколько таковых, меняют свой категориальный статус, как например, *лишь бы* – союз, в то время как по отдельности *лишь* и *бы* – частицы).

Не существует чёткого разделения между идиомами разных типов и единицами микросинтаксиса. Многие из перечисленных выше конструкций в равной степени относятся и к области классической фразеологии, и к микросинтаксису. Главным критерием для разграничения является нестандартное синтаксическое поведение единицы и множество нерегулярных способов выражения грамматических значений внутри неё.

Микросинтаксические конструкции занимают значительное место в русском языке. Эта область состоит не менее чем из нескольких тысяч разнородных единиц, каждая из которых имеет своё уникальное строение, свой неповторимый облик и набор параметров, которые отражают разные смыслы или обусловливают весьма неожиданные синтаксические свойства лексической единицы (Иомдин 2017, Чага 2021). Будучи периферийной частью грамматики, и в первую очередь, синтаксиса, данное языковое явление занимает хоть и меньшую, но отнюдь не малую часть языка. На данный момент словник выделенных конструкций состоит из примерно 2800 элементов, но можно с уверенностью сказать, что это далеко не исчерпывающий список того, что существует в русском языке.

Идентификация и лингвистическое описание единиц микросинтаксиса связаны со значительными трудностями. На текущий момент не существует надёжного способа их автоматической детекции. Создание полного перечня и тем более описания всех возможных микросинтаксических конструкций невозможно ввиду ряда причин: конструкций в языке тысячи, многие из них имеют более или менее свободное лексическое наполнение, каждая

единица имеет своё уникальное устройство и синтаксически неповторима. Более того, имея набор заданных показателей, не всегда легко даже вручную определить границу между свободным словосочетанием или иной регулярной конструкцией и единицей микросинтаксиса, поскольку встречаются случаи, когда выражения, по внешним признакам удовлетворяющие описанию конструкции, к ним не относятся.

Необходимо создание инструментов, позволяющих отделять случайное соположение элементов от действительных случаев употребления этих элементов в качестве единицы микросинтаксиса. Сложности также представляет то, что многие единицы разрывны, и порой одна входит в состав другой.

На данный момент микросинтаксические единицы размечаются вручную или на основании правил, записываемых в формализме системы ЭТАП. Автоматически собирать список единиц пока не удаётся, и эта задача ожидает своего решения. Корпус текстов с микросинтаксической разметкой, которая в скором времени станет доступна на сайте НКРЯ, предоставляет ценный материал для разработки и проверки работы алгоритмов машинного обучения с целью автоматической детекции единиц, входящих в область микросинтаксиса.

Тот факт, что перечень безусловных оборотов, выделяемых системой ЭТАП, практически на 100%, исключая лишь имена собственные и термины, входит в словарь микросинтаксических конструкций, говорит о том, что это первый серьёзный шаг на пути к их автоматической идентификации и анализу.

В целом, микросинтаксические единицы не могут быть сведены к безусловным оборотам, хотя некоторая их часть имеет с оборотами много общего. Отметим, что и те, и другие могут морфологически изменяться. И всё же большая часть микроединиц не может соответствовать одному слову и далеко не все из них могут занимать один узел в синтаксической структуре. Так, некоторые конструкции микросинтаксиса имеют устройство, характерное для полного предложения:

V + X-ы и X-ы ≈ ‘X-ы бывают разные’, где X – название явления, понятия или объекта:

(1) Он познакомился с ней тоже странно: пожав руку, неожиданно заявил: «*Бывают встречи и встречи*. [А. С. Грин. Тихие будни (1913)]

X X-у рознь:

(2) Вообще-то *неприятности неприятностям рознь*. Бывают неприятности разных уровней. [Аркадий Стругацкий, Борис Стругацкий. За миллиард лет до конца света (1974)]

X X-ом, а/но Y Y-ом ≈ ‘есть X, а есть Y, и они не зависят друг от друга’:

(3) И хотя мы жили с дедушкой под одной крышей, жили одной семьёй, но *крыша — крышей, семья — семьёй, а дело — делом*. [Анатолий Рыбаков. Тяжелый песок (1975-1977)]

На примере работы лингвистического анализатора, который определяет неоднословные лексические единицы, можно видеть, как многие из них, обладая неконвенциальной семантикой и нетривиальной сочетаемостью, перерастают в гораздо более самостоятельные единицы, образуя единицы микросинтаксиса. Таким образом, микроединицы – это более принципиальный уровень представления multiword units, чем безусловные обороты.

Неоднословные лексические единицы, автоматически выделяемые системой ЭТАП, делятся на два типа:

1) Неизменяемые и неразрывные конструкции, которые не допускают вставления других элементов внутрь. Такие единицы могут считаться эквивалентными обычным словам и представляются в синтаксической структуре как один узел. Например, предлог *по отношению к*, союз *коль скоро*, частицы *разве что, что ни на есть, нет-нет да и*, наречия *как бы то ни было, скрепя сердце, из рук вон плохо, стало быть, в обнимку*.

2) Выражения, которые естественно считать состоящими из нескольких слов (например, в тех случаях, когда эти слова могут изменяться или разделяться другими словами), но для которых не строится обычная синтаксическая структура. В этом случае все или некоторые слова в выражении соединяются вспомогательным СинтО. В следующих примерах устанавливается одна вспомогательная связь (от X к Y): и так [Y] далее [X]; как [Y] можно [X] быстрее и т.д.

Несмотря на явные преимущества, касающиеся лингвистического описания, наличие безусловных оборотов порождает ряд проблем. Учитывая, что в среднем от 30 до 40% предложений текста содержат хотя бы одну

микросинтаксическую единицу, а в некоторых предложениях встречается до пяти единиц микросинтаксиса, включая безусловные обороты, то числовые характеристики текста могут серьёзно искажаться, мешая расчётам, связанным, например, с частотностью слов, а также поиску конкретных лексем, входящих в состав неоднословных единиц. Таким образом, встаёт вопрос о том, как сохранить информацию о неоднословности лексической единицы, при этом не мешая поиску и подсчёту её отдельных компонентов.

Подход, разработанный нашей лабораторией, позволяет обнаруживать не только цельные и неразрывные отдельно стоящие обороты, но и те, которые разделены или имеют в своём составе какой-либо необязательный элемент: *всё равно* vs. *не всё ли ему было равно*, *в то время, как* vs. *в то самое время, как*. Значительная часть безусловных оборотов и подавляющая часть микроединиц разрывны. Так, довольно долго не идентифицировались случаи вроде ‘несмотря даже на’, ‘по этому ли, по другому поводу’, ‘в Серёжином же случае’ и т.д.

В настоящее время был принят ряд решений для создания правил, позволяющих сохранять информацию о наличии смыслового единства того или иного выражения при оформлении внутренней синтаксической структуры и добавлении морфологической информации о каждом компоненте.

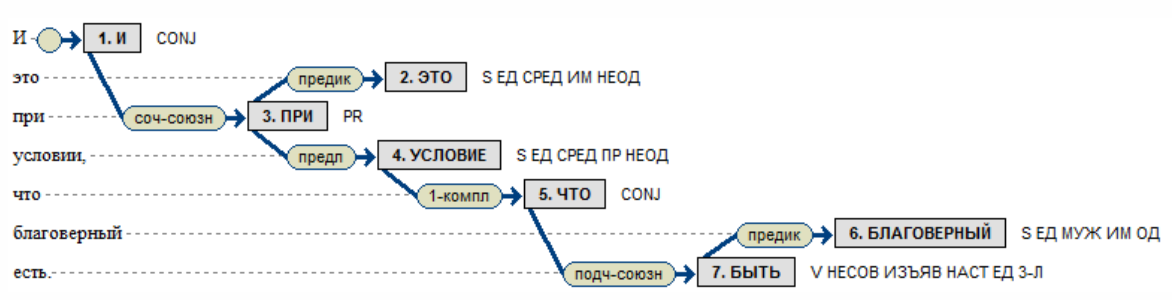

Рис. 3. Полная древесная синтаксическая структура фразы, содержащей неоднословную лексическую единицу ‘при условии, что’.

Благодарность

Выражаю искреннюю благодарность Л. Л. Иомдину за замечания и важные советы при подготовке этой статьи.

Литература

1. Богуславский И.М., Иомдин Л.Л. Безусловные обороты и фраземы в толково-комбинаторном словаре // Актуальные вопросы практической реализации систем автоматического перевода. Ч. 2. М.: Изд-во МГУ, 1982, с. 210–222.
2. Маракасова А.А., Иомдин Л.Л. Микросинтаксическая разметка в корпусе русских текстов СинТагРус // Информационные технологии и системы 2016 (ИТиС’2016). Сборник трудов 40-ой междисциплинарной школы-конференции ИППИ РАН. Репино, Санкт-Петербург. С. 445-449.
3. Иомдин Л.Л. Микросинтаксическая разметка в корпусе русских текстов. // Труды международной научной конференции «Корпусная лингвистика - 2017». Санкт-Петербург, Изд-во СпбГУ, 2017. С. 188-194. ISSN 2412-9623.
4. Иомдин Л. Л. Как нам быть с конструкциями типа как быть? // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 31 мая —3 июня 2017 г.). М.: Изд-во РГГУ, 2017. Вып. 16 (23). Т. 1. С. 161–176.
5. Чага А.В. О русских конструкциях типа не наХ-оваться. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Студенческая сессия (Москва, 16 - 19 июня 2021г.).
6. Avgustinova, T., Iomdin, L. Towards a Typology of Microsyntactic Constructions. In: Corpas Pastor, G., Mitkov, R. (eds) Computational and Corpus-Based Phraseology. EUROPHRAS 2019.
7. Leonid Iomdin (2016). Microsyntactic Phenomena as a Computational Linguistics Issue. // Grammar and Lexicon: Interactions and Interfaces. Proceedings of the Workshop. Osaka, Japan. 2016, pp. 8-18. (http://aclweb.org/anthology/W/W16/W16-38.pdf). ISBN 978-4-87974-706-8.