Л.Л.Иомдин
Leonid L. Iomdin

МИКРОСИНТАКСИЧЕСКАЯ РАЗМЕТКА В КОРПУСЕ РУССКИХ ТЕКСТОВ[1]

MICROSYNTACTIC TAGGING IN A RUSSIAN TEXT CORPUS

Аннотация. Представлен новый вид разметки в корпусе русского языка «СинТагРус». Эта разметка идентифицирует два типа фразеологических единиц, относящихся к области микросинтаксиса – нестандартные синтаксические конструкции (например, конструкции с повторяющимися лексическими элементами типа *Читать не читал*) и синтаксические фраземы (лексикализованные фразеологические единицы, отличающиеся синтаксической спецификой). При микросинтаксической разметке используются две стратегии: сплошной просмотр текста и целенаправленный поиск последовательностей слов или синтаксических поддеревьев, которые с большой вероятностью относятся к искомым элементам.
Ключевые слова. Микросинтаксис, размеченные корпуса, лексикография, семантический анализ, многозначность фразем.

Abstract. A new type of annotation in a Russian corpus, SynTagRus, is presented. The annotation identifies two types of idiomatic units belonging to the area of microsyntax: nonstandard syntactic constructions (e.g. those with recurring lexical elements like *Čitat' ne čital* ≈ 'I didn't exactly read it') and syntactic idioms (lexicalized idiomatic units characterized by considerable syntactic individuality). Microsyntactic tagging uses two strategies: continuous examination of the whole text and targeted search for word sequences or subtrees which are likely to contain the desired elements.
Keywords. Microsyntax, tagged corpora, lexicography, semantic analysis, ambiguity of idioms.

1. Вводные замечания

В настоящей работе представлен новый вид разметки, которая производится в глубоко аннотированном корпусе русского языка «СинТагРус» (о современном состоянии корпуса см. [Дяченко и др. 2015]). Эта разметка идентифицирует два присутствующих в тексте типа фразеологических единиц, относящихся к области микросинтаксиса[2].

[1] Автор выражает признательность за поддержку данной работы Российскому фонду фундаментальных исследований (грант № 15-04-00562).

[2] Микросинтаксис — раздел лингвистики, занимающий промежуточное положение между словарем и грамматикой, исследуемый автором и несколькими коллегами в течение последних полутора десятилетий и описывающий синтаксически мотивированную идиоматику языка (см., в частности, [Иомдин 2015, Iomdin 2016, Marakasova-Iomdin 2016]). Идейно микросинтаксис близок грамматике конструкций Ч.Филмора и его последователей.

Первый тип — это нестандартные синтаксические конструкции, обладающие нетривиальной семантической спецификой, например, конструкции с повторяющимися словесными элементами типа «X как X» (ср. ***Мальчик как мальчик****, таких много в каждом классе*), «X есть X» (ср. ***Дети есть дети****, они быстро устают*), «X-овать не X-овал» (ср. ***Видеть не видел,*** но *много слышал о нём*) и т. п.

Второй тип — это синтаксические фраземы, или лексикализованные фразеологические единицы, отличающиеся, помимо семантической некомпозициональности, той или иной синтаксической спецификой. Значительную их часть составляют многозначные единицы, такие как *всё равно* (ср. *Он всё равно[1] не послушается*, где имеет место сентенциальное наречие, означающее 'независимо ни от чего', *Мне всё равно[2], куда ехать*, где присутствует предикативное наречие со значением 'безразлично' или «*Сняться в плохом фильме — все равно[3] что плюнуть в вечность*» (Ф. Г. Раневская), где присутствует другое предикативное наречие со значением «равносильно») или *как бы* (ср. *Он как бы[1] предчувствовал опасность*, где присутствует дискурсивная частица со значением сравнения или осторожной номинации и *Мы боялись, как бы[2] он не заболел*, где выступает сильноуправляемый союз, встречающийся при предикатах лексического класса со значением опасения).

Очевидно, что корпус, оснащенный такой разметкой, весьма полезен как для теоретической и практической лексикографии и грамматики, поскольку он позволяет исследовать многообразные контексты, в которых выступают микросинтаксические единицы, так и для широкого класса компьютерно-лингвистических задач, в частности, для задач глубоко семантического анализа текста. До последнего времени в распоряжении исследователей не было корпусов с фразеологической разметкой, хотя нужда в них отчетливо осознается (см., например, [Grzybek & Jesenšek 2014]). По этой причине появление такой разметки в СинТагРус'е можно считать первым опытом частичной фразеологической аннотации — во всяком случае, для русского языка.

Микросинтаксическое аннотирование корпуса представляет собой достаточно сложную задачу. Одна из причин состоит в том, что сколько-нибудь полного списка микросинтаксических единиц русского языка еще не существует. Поэтому мы при разметке использовали две стратегии:

1) сплошной просмотр текста и отыскание в нем кандидатов в микросинтаксические элементы;

2) целенаправленный поиск в корпусе линейных последовательностей слов или же синтаксических поддеревьев, состоящих из таких слов, которые заведомо могут образовывать микросинтаксические единицы. Это, например, такие последовательности и поддеревья, как *всё равно*, *как бы*, *как будто, коль скоро, разве что, пока что, только лишь, мало ли, ни разу, черт знает* + вопросительное слово и многие другие.

В обеих стратегиях мы опираемся на материалы создаваемого автором Микросинтаксического словаря русского языка. Как в том, так и в другом случае мы получаем предварительный вариант разметки текста, с которым производится дальнейшая работа. В настоящее время вся такая разметка проводится вручную: у нас пока нет ни правиловых, ни статистических критериев, которые позволили бы автоматизировать эту работу хотя бы частично.

2. Первые результаты

Уже в начале работы над микросинтаксической разметкой корпуса выяснилось, что встречаемость в текстах интересующих нас единиц достаточно высока: в среднем тексте[3] почти четверть предложений содержит хотя бы одну такую единицу [Маракасова-Иомдин 2016].

С технической точки зрения микросинтаксическая разметка корпуса выглядит так: в XML-представление предложения, содержащего микросинтаксический элемент или элементы, вводится специальное поле, в котором отражается имя элемента (обычно это словосочетание или просто последовательность слов) и указываются его линейные границы. Как выбор имени для микросинтаксического элемента, так и идентификация его границ — не совсем элементарные задачи. Неясно, например, как правильно называть элементы, которые могут содержать переменные части (скажем, единицу типа *какого чёрта, в* котором в качестве второго слова вместо *чёрта* могут выступать существительные *дьявола, рожна, хрена* и нек. др. или единицу типа *вот* + вопросительное слово, которая может реализоваться вари-

[3] Для опытной разметки мы использовали типичные тексты СинТагРус'а — научно-популярные, новостные и публицистические. Можно предположить, что в художественных текстах встречаемость микросинтаксических единиц, как и идиоматики в целом, еще выше, а в научных и технических текстах таких единиц меньше, но количественными данными мы пока не располагаем.

антами *вот что, вот кто, вот где*, *вот какой* и т. д. (Мы называем эти единицы, соответственно, *какого черта* и *вот* + ВОПР). Что же касается неочевидных линейных границ конструкции, то в качестве примера можно привести уже упомянутое предикативное наречие *все равно*[2], между словесными элементами которого могут появиться посторонние вставки, ср. *Не все* ***ли вам*** *равно, когда он жил, если вы не знаете, кто он такой?* (Ю. Н. Тынянов) или конструкцию типа в (*чьих-либо*) *силах*, как в предложении *Оградить двери от взлома, письма от вскрытия не в* ***моих*** *силах.* (Л. К. Чуковская).

Снимок экрана, воспроизведенный ниже на рис. 1, дает представление о микросинтаксической разметке СинТагРус'а. Здесь представлены две микросинтаксические единицы — в *силах*[1] (ср. *Хирург был не в силах помочь ему* — с подлежащим, обозначающим агента) и в *силах*[2] (ср. *Помочь ему было не в силах хирурга* — с инфинитивным подлежащим, выражающим событие или действие). Вторая единица появляется реже, она встречается в корпусе лишь один раз (см. предложение 10). Первая единица выступает во всех остальных предложениях, за исключением 9 и 11, в которых встречаются выражения в *Военно-морских силах и уверенность в собственных силах,* очевидным образом не имеющих отношения ни к одной из двух синтаксических фразем.

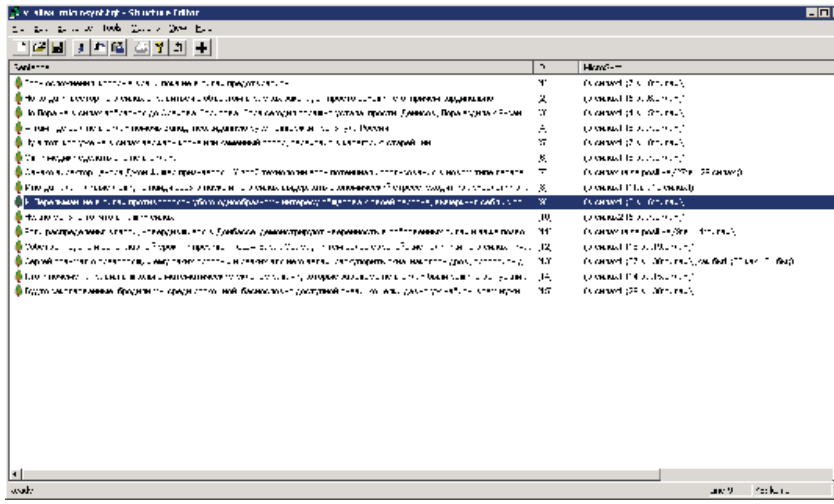

Рис. 1. **Микросинтаксическая разметка корпуса СинТагРус. Фразы с микросинтаксическими единицами в *силах*.**

На нынешней стадии создания корпуса мы считаем целесообразным оставлять в разметке и такие случаи, помечая их признаком false positive («ложное срабатывание», в расчете на то, что такую информацию впоследствии можно будет использовать для целей машинного обучения и автоматизированной идентификации фразеологических элементов и разрешения многозначности.

Принципиально, что микросинтаксическая разметка корпуса СинТагРус производится в дополнение к другим типам аннотации, в первую очередь, к синтаксической разметке, что дает исследователю возможность увидеть, как именно фразема встраивается в синтаксическую структуру предложения.

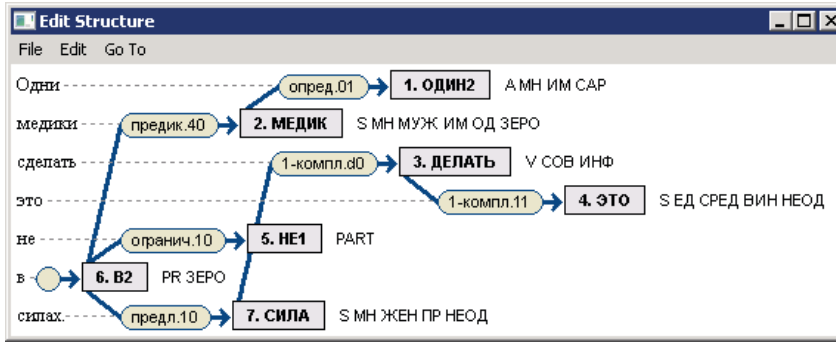

Рис. 2. **Полная древесная синтаксическая структура фразы, содержащей микросинтаксический элемент в *силах*[1].**

Рис. 2 иллюстрирует синтаксическую структуру предложения (9) из снимка экрана на рис. 1. Видно, что предлог в из фраземы в *силах* выступает в качестве вершины сказуемого, подлежащее при котором выражено одушевленным существительным *медик*. Что же касается инфинитива *сделать*, то он подчиняется второму элементу фраземы — слову *силах* по 1-му комплетивному синтаксическому отношению и по существу выражает вторую синтаксическую валентность этой фраземы.

3. Case study: конструкции типа *мало что*

Хотя корпус СинТагРус невелик по объему (он содержит около 1 млн. словоупотреблений (около 67 тыс. предложений), а микросинтаксическая разметка в нем затрагивает всего несколько тысяч пред-

ложений, с ее помощью уже можно получать нетривиальную лингвистическую информацию.

Показательным примером могут служить синтаксические фраземы типа *мало что*. Вхождений этих конструкций в корпусе немного, но и их достаточно, чтобы обнаружить интересную закономерность. Эти конструкции имеют три разновидности: первое слово в них — это *мало, много и редко, а* второе слово — вопросительное местоимение (*что, кто, где, какой* и др., причем ненаречные слова могут стоять в разных падежах и даже сопровождаться предлогами — *мало что, мало чего, мало чему, мало о чем*). Обратим внимание на две оппозиции: *мало чего — мало что и много чего — много что*. Материал корпуса обнаруживает несимметричность в их поведении: варианты *мало что и мало чего* встречаются с соизмеримой частотой, а вариант *много чего* заметно превышает по частоте вариант *много что*. Как кажется, этот факт можно объяснить переосмыслением в языке конструкции *много что* — вместо подчиненного форманта при вопросительном слове (ср. *мало что и кое-что*) слово *много* стало восприниматься как количественное наречие, подчиняющее местоимение *чего в* родительном падеже. Для слова *мало* этот процесс, возможно, начался, но еще не завершился.

Литература

1. *Дяченко П. В., Иомдин Л. Л., Лазурский А. В.* и др. (2015), 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. Вып. 6, с. 272–299.
2. *Иомдин Л. Л.* (2015), Конструкции микросинтаксиса, образованные русской лексемой раз // SLAVIA, časopis pro slovanskou filologii, ročník 84, 2015, sešit 3, s. 291–306.
3. *Маракасова А. А., Иомдин Л. Л.* (2016), 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.
4. *Iomdin L.* (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
5. Grzybek P., Jesenšek V. (2014), Phraseology in Dictionaries and Corpora. Introductory Remarks // Phraseologie im Wörterbuch und Korpus. ZORA 97. Maribor, Bielsko-Biała, Budapest, Kansas, Praha, pp. 19–25.

References

1. *Djachenko P. V., Iomdin L. L., Lazursky A. V.* et al. (2015), The State-of-the-Art of a deeply annotated corpus of Russian texts (SynTagRus) [Sovremennoe sostojanie gluboko annotirovannogo korpusa tekstov russkogo jazyka (SinTagRus)]. In: Nacional'nyj korpus russkogo jazyka. 10 let proektu. Trudy Instituta russkogo jazyka im. V. V. Vinogradova. M. Vyp. 6, pp. 272–299.
2. *Iomdin L. L.* (2015), Microsyntactic constructions formed by the Russian word RAZ [Konstrukcii mikrosintaksisa, obrazovannye russkoj leksemoj RAZ]. In: SLAVIA, časopis pro slovanskou filologii, ročník 84, 2015, sešit 3, s. 291–306.
3. *Iomdin L.* (2016). Microsyntactic Phenomena as a Computational Linguistics Issue. In: 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
4. Grzybek P., Jesenšek V. (2014), Phraseology in Dictionaries and Corpora. Introductory Remarks. In: Phraseologie im Wörterbuch und Korpus. ZORA 97. Maribor, Bielsko-Biała, Budapest, Kansas, Praha, pp. 19–25.
5. Marakasova A. A., Iomdin L. L. (2016), Microsyntactic Tagging in the Corpus of Russian Texts SynTagrus. [Mikrosintaksicheskaja razmetka v korpuse russkix tekstov SinTagRus]. In: Informacionnye texnologii i sistemy 2016 (ITiS'2016). Sbornik trudov 40-oj mezhdisciplinarnoj shkoly-konferencii IPPI RAN. Repino, St Petersburg, pp. 445–449.

Иомдин Леонид Лейбович
ИППИ РАН им. А. А. Харкевича (Москва, Россия)
Iomdin Leonid
Institute for Information Transmission Problems, Russian Academy of Sciences (Moscow, Russia)
E-mail: iomdin@gmail.com