РАЗМЕТКА СИНТАКСИЧЕСКОЙ НЕПОЛНОТЫ В КОРПУСЕ[1]
MARKING OF SYNTACTIC INCOMLETENESS IN A CORPUS

Копотев М. В. *(mihail.kopotev@helsinki.fi), Хельсинкский университет*
Г. Б. Гурин *(grigur@onego.ru), Петрозаводский государственный университет*

Доклад посвящен разметке синтаксических нулей и смежных явлений в русскоязычном корпусе текстов. Предлагается сводная классификация явлений, основанная на существующей научной литературе по теме, а также обсуждается два подхода к выделению таких единиц в корпусе.

1. Постановка задачи

Дж. Сова, отвечая на вопрос о возможных способах маркирования нулей, заданный авторами настоящей статьи в рассылке Corpora-List, написал: "What you are asking for is the Holy Grail of NLP". Похоже, это действительно так. Любой текст содержит большее или меньшее количество случаев неполноты, как смысловой, так и синтаксической [Леонтьева 1969]. В то же время идеология и процедурная основа компьютерной обработки языка вступает в очевидный конфликт с этим естественным свойством языка, поскольку компьютерный анализ языка возможен, только если объект существует материально – в виде звуковой или буквенной цепочки. Компьютерные лингвисты, по сути, становятся заложниками плана выражения языковой единицы: чем менее «материальны» языковые объекты, тем более затруднена их автоматическая обработка.

Этот конфликт теории и практики имеет прямое отношение и к корпусной лингвистике, очерчивая пределы победившей корпусной парадигмы. Подчиняясь «техническим» ограничениям, мы рискуем оставить за пределами сферы внимания лингвистов целые зоны языкового материала; большой массив данных словно перестанет существовать только потому, что корпус не будет давать возможности их искать. Одной из таких зон, безусловно, являются нулевые компоненты и другие виды значимого отсутствия.

Выделение разного рода «отсутствий» (нулей, эллипсисов, неполноты), без сомнения, является производным конкретной теории, а спектр подходов к решению проблемы может быть широк до крайностей: от отрицания необходимости постулирования нулей[2] до обоснования «полного» списка нулевых единиц[3]. Как нам представляется, для решения этой проблемы в корпусной лингвистике необходимо обсудить два круга вопросов: теоретические подходы к выделению синтаксических нулей и сложности практического представления нулевых знаков. В настоящем докладе обсуждаются эти вопросы, а также предлагаются возможные решения для корпуса ХАНКО.

2. Теоретические проблемы

В теоретической лингвистике проблема языковых знаков, лишенных означающего, обсуждается уже более ста лет (Ф. Ф. Фортунатов, Ф. де Соссюр, Ш. Балли, Р. Якобсон, Г. Майер, И. А. Мельчук, Д. Песецки, Л. Стассен и многие др.). Существуют и многочисленные исследования, посвященные выделению нулевых единиц в конкретных языках, в том числе в русском (А. М. Пешковский, М. В. Панов, Г. А. Золотова, А. П. Сковородников, П. А. Лекант, Е. Н. Ширяев, М. Макшейн, К. Чвани и др.).

В целом, можно выделить несколько релевантных для корпусной лингвистики общих проблем, которые должны быть решены при разработке схемы аннотирования, если не в теоретическом, то хотя бы в практическом отношении:

– различение синтаксических и семантических нулей и целесообразность их выделения в корпусе;
– разграничение эллипсиса и нулевого знака;
– разграничение эллиптических и неполных конструкций.

[1] Благодарим участников рассылки Corpora-list (Дж. Синклера†, Дж. Сову, А. Осенову и др.) за ценные комментарии.
[2] Характерно, что именно представители корпусной лингвистики обосновывают подходы, не предполагающие выделение нулей и смежных явлений, см [Sinclair, Mauranen 2006].
[3] The Penn Treebank (ftp://ftp.cis.upenn.edu/pub/treebank/doc/ tagguide.ps.gz); the Bulgarian Treebank (http://www.bultreebank.org/TechRep/BTB-TR05.pdf).

Нам представляется, что предварительным этапом решения этих задач в русской корпусной лингвистике может стать построение списка нулевых синтаксических единиц, в достаточной степени отраженного в учебной и научной литературе. На такой список, составленный на основе существующих исследований, и опираются авторы настоящей статьи. Этот список не является ни полным, ни теоретически однородным. В него попали не только «классические» нули, но и другие виды материально не выраженных знаков, по тем или иным причинам вводимых в описания русского языка (эллипсис, нулевые подлежащие нефинитных клауз и некоторые другие).

2.1 Нулевая лексема

Под нулевой лексемой понимается «одноэлементное множество лекс, содержащее только нулевую лексу; символ нулевой лексемы, наряду с символами прочих лексем, выступает в качестве пометы при соответствующем узле синтаксического дерева» [Мельчук 1974: 349]. На основе существующих исследований можно составить следующий список таких единиц.

1) Ø $_{\text{3-Л множ '(другие) люди'}}$: *Дорогу Ø засыпали песком*; нулевое подлежащее традиционных неопределенно-личных предложений[4];

2) Ø $_{\text{3-Л един сред 'стихии'}}$: *Дорогу Ø засыпало песком*; нулевое подлежащее части традиционных безличных глагольных предложений[5];

3) Ø $_{\text{ед (род, дат, вин) 'любой'}}$: *Своя ноша не тянет Ø; Нельзя Ø так говорить, Подобные поручения очень обременяют Ø*;

4) Ø $_{\text{един сред 'моя личность'}}$: *Мне Ø холодно;*

5) Ø $_{\text{един сред 'среда'}}$: *Здесь Ø холодно.*

6) Синтаксические агломераты: *Негде Ø спать*[6]*;*

7) Ø $_{\text{подлежащее инфинитивной клаузы}}$: *Он предложил Ø пройти в комнату*[7]*;*

8) Ø $_{\text{подлежащее причастной клаузы}}$: *Игрок, Ø нарушивший правила, должен быть удален с поля.*

9) Ø $_{\text{подлежащее деепричастной клаузы}}$: *Отвечая на вопросы, Ø будьте искренни.*

10) Ø $_{\text{подлежащее неглагольной клаузы}}$: *Ø Гордый своими успехами, Иван решил продолжить работу.*

2.2 Нулевая лекса

Под нулевой лексой (нулевой словоформой) понимается «символ в представлении фразы на n-местном синтаксическом уровне, который при движение «вверх», к тексту, всегда пропадает – переходит на уровне n+1 в пустую цепочку (во фразовом сегменте ему не отвечают никакие фонемы), а при движении «вниз», к смыслу, переходит в определенный комплекс символов уровня n-1» [Мельчук 1974: 349]. Список таких единиц представлен ниже.

Ø $_{\text{быть I 'являться'}}$: *Мой брат Ø боксер;* нулевая отвлеченная связка, выражающая комплекс модально-временных значений в предложениях характеризации и идентификации;

Ø $_{\text{быть II 'находиться'}}$: *Дядя Ø на улице*; нулевой глагол в локативном предложении[8];

Ø $_{\text{быть III 'существовать (в том числе в каком-нибудь фрагменте мира)'}}$: *На берегу Ø люди*; нулевой глагол бытийного предложения, способный употребляться и в ненулевой форме со значимыми различиями в семантике (*На берегу есть люди)*[9];

[4] [Мельчук 1974: 350, Булыгина, Шмелев 1995: 341]

[5] Типы 5-8 выделены в [Мельчук 1974: 350].

[6] [Апресян, Иомдин 1989], [Богуславский, Иомдин и др. 2000: 46].

[7] Группы нулевых местоимений (7-10) с синтаксическими функциями выделяется в теориях, оперирующих научным аппаратом составляющих, в качестве подлежащих нефинитных клауз. Использование этих «технических» нулевых знаков существенно облегчает теоретическое описание языка см. [Тестелец 2001: 269-280].

[8] [Мельчук 1974: 350].

[9] [Арутюнова, Ширяев 1983: 27].

2.3 Эллипсис

Признавая существование альтернативных определений эллипсиса, авторы настоящего доклада опираются на формулировку, данную И. А. Мельчуком: эллипсис – «это правило элиминирования определенных знаков в определенном контексте. <…> [Это] операция незнаменательная: ее применение не меняет смысла, но требуется для грамматической или стилистической корректности высказывания» [Мельчук 1974: 357]. Несмотря на попытки противопоставить синтаксический нуль и эллипсис, сделать это в значительном количестве случаев не удается. Поэтому разумно было бы в качестве периферии поля нулевых синтаксических знаков выделить случаи эллипсиса и контекстуальной / конситуативной синтаксической неполноты[10].

14) Сочинительный эллипсис и смежные явления: *Я купил рубашку, а он Ø галстук; Он купил красный галстук, а я синий Ø*[11];
15) Ø $_{\text{verbum finitum ‘движения, речи и др.’}}$: *Ты Ø куда? Это Вы Ø про Мейерхольда?;*
16) Ø $_{\text{инфинитив ‘движения, речи и др.’}}$: *Нам бы завтра в театр Ø? Я не мог об этом;*
17) Ø $_{\text{предикат}}$: *У нас машина Ø - мы совсем мало ездили.*

3. Возможные решения

3.1 Подход, основанный на введении тэгов

Этот подход основан на четких принципах технического выделения синтаксических нулей, которые «материализуются» в корпусе в виде специальных тэгов. По существу в корпус вводятся «текстоформы», обладающие особым статусом. При этом парсер может быть настроен на (полу-)автоматическое аннотирование таких единиц. Это самый распространенный способ решения проблемы, который имеет и свои достоинства и свои недостатки.

К его достоинствам можно отнести системность аннотирования, основанного на выбранной синтаксической теории (в идеале, «полное» представление нулей в рамках определенной теории). Кроме того, такое аннотирование позволяет пользователю работать с материальными объектами, четко описанными в инструкциях к корпусу, и производить обычные поисковые операции. Слабость такого подхода заключается в том, что, поскольку теоретически нейтрального русского синтаксиса не существует, корпус волей-неволей будет отражать представления составителей и в этом смысле неизбежно будет вызывать возражения пользователей как раз в теоретической части. Кроме того, сомнительной представляется возможность аккуратной разметки нулей при автоматической обработке достаточно больших массивов данных.

Говоря, о возможности введения «нулевых» тэгов в разметку корпуса, нам представляется, что программой-минимумом (если не считать таковой полное игнорирование нулей) должна стать задача введения тэгов для нулевых лекс (напр., нулевая связка БЫТЬ в настоящем времени), поскольку именно они являются наиболее признанными (и в этом смысле «теоретически нейтральными»). При желании разработчиков корпуса расширить список «нулевых» тэгов, следующей группой должны стать эллиптические предложение, как представляющие набольшую сложность при поисковом подходе. И, наконец, при максимально полном учете нулевых единиц и смежных явлений, этот список должен дополнится нулевыми лексемами, которые, во-первых, относительно легко ищутся на основе морфологических запросов, а во-вторых, – не являются общепринятыми.

3.2. Поисковый подход

Второй подход основан не на заранее введенных в корпус фантомных тэгах, а на использовании поисковых возможностей. Речь идет о том, что язык поисковых запросов (например SQL) позволяет искать элементы не только содержащие, но и *не* содержащие определенные элементы. Так, например, следующий запрос позволяет искать предложения без глагола: «найти все предложения, которые не содержат глагол».

Применение этого подхода, безусловно, зависит от аккуратности и полноты уже существующей в корпусе разметки. Однако его главное достоинство состоит в том, что он является действительно теоретически нейтральным (точнее, пользователь сам определяет поисковые запросы, основываясь на собственных предпочтениях). Главный недостаток такого подхода состоит в том, что запросы, по-видимому, будет возвращать достаточно большое количество нерелевантного материала. Второй недостаток связан с техническими ограничениями, поскольку выполнение запроса на поиск с оператором ‘не содержит’ требует значительного количества времени.

Как нам представляется, значительная часть случаев синтаксической неполноты, приведенных во второй

[10] Представительную классификацию таких структур можно найти в [Ширяев 1984: 18-20, Русская разговорная речь 1973: 299-317].
[11] [Падучева, Лященко 1973: 20-31, Богуславский, Иомдин и др. 2000: 46].

части доклада, может быть найдена при непрямом поиске: например, при поиске предложений, не содержащих спрягаемых форм глагола, можно получить близкий к полному список «безглагольных» предложений (экспериментальный поиск в ХАНКО выдал около 1000 предложений, и все, естественно, удовлетворяли условию поиска). Однако поиск предложений типа «Мой брат – учитель» (запрос «найти все предложения без глагола с двумя именительными падежами существительного/личного местоимения») представляет собой уже большую сложность. Однако и он возможен и выдает приемлемые результаты при наличии в разметке указания на границы клаузы. Гораздо острее эта проблема стоит для некоторых нулей, искать которые по косвенным признакам нельзя. Например, более–менее приемлемый косвенный поиск нулевого подлежащего инфинитивной клаузы (*Он пригласил Ø пройти в комнату*), как кажется, невозможен.

Таким образом, для пользователя корпуса поиск таких синтаксических единиц является серьезной проблемой, а для разработчиков их учет в схеме аннотирования может стать специальной теоретической и технической задачей. Кроме собственно лингвистического, теоретического обоснования, решение, как представляется, может быть найдено в двух независимых друг от друга направлениях: в разметке «фантомных» тэгов и разработке поисковых запросов. Само собой, оба способа могут быть объединены в рамках одного корпуса. В синтаксическом аннотировании ХАНКО такой подход может быть реализован и представлен в обоих типах синтаксической разметки, используемых в корпусе[12]. Оба подхода предполагают выделение определенного количества синтаксических нулей.

Список литературы

1. Sinclair J., McH., Mauranen A. Linear Unit Grammar. Integrating speech and writing. Amsterdam: John Benjamins. 2006.
2. Апресян Ю.Д., Иомдин Л.Л. Конструкции типа «негде спать»: синтаксис, семантика, лексикография // Семиотика и информатика. Вып. 29. 1989. С. 34-92.
3. Арутюнова Н.Д., Ширяев Е.Н. Русское предложение. Бытийный тип // М.: Русский язык, 1983.
4. Богуславский И. М., Иомдин Л. Л. и др. Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации // Труды Международного семинара Диалог’2000 по компьютерной лингвистике и ее приложениям. Т. 2. Протвино, 2000.
5. Богуславский И. М., Иомдин Л. Л. и др. Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации // Труды Международного семинара Диалог’2000 по компьютерной лингвистике и ее приложениям. Т. 2. Протвино, 2000.
6. Булыгина Т. В., Шмелев А. Д. Языковая концептуализация мира (на материале русской грамматики) // М.: Языки русской культуры, 1995.
7. Копотев М. В., Гурин Г. Б. Принципы синтаксической разметки Хельсинкского аннотированнго корпуса русских текстов ХАНКО // Компьютерная лингвистика и интеллектуалъные технологии. Труды международной конференции Диалог-2006 (Бекасово, 31 мая - 4 июня. 2006). М.: Изд-во РГГУ, 2006. С. 280-284.
8. Леонтьева Н. Н. О смысловой неполноте текста (в связи с семантическим анализом) // Машинный перевод и прикладная лингвистика. Вып. 12. М., 1969. С. 96-114.
9. Мельчук И. А. О синтаксическом нуле // Типология пассивных конструкций, диатезы и залоги. Л.: Наука, 1974. С. 343-360.
10. Падучева Е. В., Лященко Т. К. Эллипсис как нулевой анафорический знак // НТИ. Сер. 2, № 5, 1973. С. 20-31. Русская разговорная речь. М.: Наука, 1973.
11. Старостин А. С., Мальковский М. Г. Модель синтаксиса в системе морфосинтаксического анализа “Treeton” // Компьютерная лингвистика и интеллектуалъные технологии. Труды международной конференции Диалог-2006 (Бекасово, 31 мая - 4 июня. 2006). М.: Изд-во РГГУ, 2006. С. 481-492.
12. Тестелец Я. Г. Введение в общий синтаксис. // М.: Изд-во РГГУ, 2001.
13. Ширяев Е. Н. Основы системного описания незамещенных синтаксических позиций // Системный анализ значимых единиц русского языка. Синтаксические структуры. Красноярск: Изд-во Красноярского университета, 1984. С. 18-20.

[12] Напомним, что в ХАНКО будут использованы два типа синтаксической разметки: 1) традиционная (в терминах членов предложения, обособленных конструкций и т. д.); 2) интегрированная морфосинтаксическая, представляющая собой версию грамматики зависимостей с элементами анализа по непосредственным составляющим, см. [Копотев, Гурин 2006, Старостин, Мальковский 2006].