ПРИНЦИПЫ СИНТАКСИЧЕСКОЙ РАЗМЕТКИ ХЕЛЬСИНКСКОГО АННОТИРОВАННОГО КОРПУСА РУССКИХ ТЕКСТОВ ХАНКО

PRINCIPLES OF THE SYNTACTIC ANNOTATION IN THE HELSINKI ANNOTATED CORPUS HANCO

М. В. Копотев *(mihail.kopotev@helsinki.fi),*
Хельсинкский университет, Хельсинки

Г. Б. Гурин,
Петрозаводский государственный университет, Петрозаводск

В докладе обосновывается использование двух типов синтаксической разметки в корпусе русских текстов ХАНКО и описываются проблемы практического применения одной из них, основанной на традиционном учении о членах предложения.

I. Введение

Проект по созданию Хельсинкского аннотированного корпуса русских текстов ХАНКО рассчитан на несколько лет и в своем законченном виде предоставит пользователю информацию о трех языковых уровнях: морфологическом, синтаксическом и функциональном. В настоящее время на сайте ***www.slav.helsinki.fi/hanco*** *опубликована морфологическая часть корпуса. В настоящем докладе выносятся на обсуждение принципы синтаксической разметки ХАНКО.*

II. Типы синтаксического аннотирования в корпусе

Соблюдение принципов аннотирования, сформулированных в [Leech 1993], максимально расширяет круг потенциальных пользователей корпуса и существенно облегчает взаимодействие с информационным ресурсом, хотя может вызвать упреки в “ненаучности”. Однако, как кажется, подход к созданию корпуса, не принуждающий авторов нести всю ответственность за логичность и последовательность разметки, а опирающийся на существующие классификации, позволяет выявлять лакуны в описаниях языка, обнаруживать дефекты и противоречия в разных подходах к языку. Последний постулат Дж. Лича предупреждает критику как раз с этой стороны.

> There can be no claim that the annotation scheme represents ‘God’s truth’. Rather, the annotated corpus is made available to a research community on a caveat emptor principle. It is offered as a matter of convenience only, on the assumption that many users will find it useful to use a corpus with annotations already built in, rather than to devise and apply their own annotation schemes from scratch (a task which could take them years to accomplish) [Leech 1993: 275].

Таким образом, корпус − это несовершенный, но часто удобный инструмент исследования, пригодный для использования в самых разных областях лингвистики, доступный любому знакомому с базовой лингвистической терминологией пользователю: студенту, учителю, преподавателю, исследователю – и сам по себе не содержащий ответы на вопросы, но позволяющий их получать.

Говоря о синтаксической разметке, авторы полагают, что на сегодняшний день существует три теории, в рамках которых можно осуществить достаточно полное описание русского материала:

- грамматика зависимостей (И. А. Мельчук, И. М. Богуславский, Л. Л. Иомдин и др.);
- грамматика структурных схем (Н. Ю. Шведова, В. А. Белошапкова и др.);
- традиционные синтаксические учения (А. А. Шахматов, В. В. Виноградов, Н. С. Валгина и др.).

При частных совпадениях в описании все три теории претендуют на полное и независимое от других подходов описание языкового материала[1]. При этом степень подробности описания в рамках разных теорий

[1] Тремя вышеперечисленными подходами список синтаксических теорий, конечно, не ограничивается. Укажем еще описания русского языка с точки зрения функциональной грамматики (А. В. Бондарко, М. В. Всеволодова, А. Мустайоки и др.), семантического синтаксиса (Н. Д. Арутюнова, Е. В. Падучева, И. Б. Шатуновский и др.), «когнитивного» синтаксиса

различна. Так, грамматика зависимостей уделяет большее внимание типам синтаксических отношений (напр., в системе ЭТАП-3 и созданном на его основе корпусе количество поименованных отношений (ветвей) доходит до 80), с другой стороны, традиционное учение о членах предложения предлагает исследователю подробную классификацию синтаксических "узлов" (типы сказуемых, разряды обстоятельств и др.). В этом смысле синтаксис Русской грамматики 1980-го года выглядит, как кажется, самым неинформативным.

Наконец, следует отметить, что не все подходы одинаково приняты русистами. Самым "теоретически нейтральным" очевидно следует признать традиционный синтаксис, опирающийся на классификацию членов предложения: именно на его основе сформулированы пунктуационные правила русского языка, этой терминологической системой владеет и школьный учитель, и профессиональный лингвист. С другой стороны, распространение учебника под редакцией В. А. Белошапковой, долгое время считавшегося базовым во многих вузах России, привело к тому, что многие преподаватели опираются, в основном, на синтаксис структурных схем. В то же время грамматика зависимостей известна лингвистам сравнительно меньше и наиболее активно используется для решения прикладных задач. Таким образом, оказывается, что выбор синтаксической теории, которая бы и удовлетворяла "постулату теоретической нейтральности", и обладала бы достаточной полнотой, представляется нелегкой задачей.

После обсуждения всех возможных подходов создатели ХАНКО приняли решение использовать для синтаксической разметки две альтернативные синтаксические схемы разметки: грамматику зависимостей и традиционный синтаксис членов предложения. При очевидной эклектичности такого подхода, совмещение двух схем позволит решить следующие задачи:

- подробно описать и узлы, и связи синтаксических структур;
- удовлетворить нужды и преподавателей русского языка, и профессиональных лингвистов;
- в зависимости от желания пользователя представлять результаты альтернативных разметок как независимо, так и совместно.

Работа над созданием такого типа аннотирования логично разбивается на две части. В настоящее время идет работа по аннотированию в терминах членов предложения, именно эта схема обсуждается в докладе[2].

III. Традиционный синтаксис в ХАНКО

Как известно, основы традиционные подхода в общем и целом сложились в работах русских лингвистов еще в XIX веке. По-видимому, наиболее полным описанием русского синтаксиса с этой точки можно считать Академическую грамматику 1960-го года. Современная общеизвестная классификация отражена с небольшими вариациями в вузовских учебниках по современному русскому языку (см., напр., Валгина 2000, Кустова et al. 2005].

Плюсы этого подхода в следующем:

- общеизвестность и простота;
- возможность косвенным образом искать материал для исследований, даже опирающихся на другие синтаксические подходы (прежде всего, структурные схемы).

К минусам традиционного похода можно отнести следующее:

- очевидное несоответствие современным представлениям о природе синтаксических структур;
- описание синтаксических узлов и игнорирование синтаксических связей;
- непоследовательность в описании и неустранимые противоречия (отсутствие предложных групп, невозможность четко разграничить разные типы второстепенных членов и т. д.);
- сложность автоматической обработки.

Однако указанные достоинства и недостатки принятого подхода в целом не служат оправданием результатов работы; они, скорее, корректируют ожидания потенциального пользователя.

Создатели ХАНКО сознательно шли на серьезные компромиссы, отказываясь от тех вариантов разбора, которые им представлялись корректными, ради сохранения понятного простому пользователю уровня метаязыка. Ниже приводятся аргументы в пользу ряда частных и не связанных друг с другом решений, принятых в ХАНКО.

1. При решении тех или иных конкретных задач создатели ХАНКО всегда задавались вопросом, насколько ценной является та или иная синтаксическая информация и насколько трудно автоматизировать обработку данных. Прогнозируемый объем ручной работы и ценность результатов часто оказывались в противоречии:

(Г. А. Волохина, З. Д. Попова) и др. Будучи в отдельных частях глубокими и точными, они, однако, в настоящее время не могут служить основой для полного описания языкового материала.

[2] Полный список параметров доступен по адресу www.helsinki.fi/hum/slav/hanco/syntax.rtf. Автоматический поверхностно-синтаксический анализ в терминах деревьев зависимости давно применяется для аннотирования русского материала, гораздо проще автоматизируется и достаточно подробно описан (см. [Апресян и др. 1989, Ножов 2003]; см. еще публикации на сайтах **proling.iitp.ru**, **www.aot.ru**).

например, выделение в качестве единицы детерминанта привело бы к существенному увеличению ручной работы (автоматизировать поиск детерминантов невозможно), однако сколько-нибудь последовательно выполнить эту работу было бы трудно, так как объем понятия “детерминант” по-разному определяется в разных лингвистических работах, а многие синтаксисты обходятся вообще без этого понятия.

2. Необходимо было учитывать и удобство интерфейса. Синтаксическая информация приписывалась разным единицам, в том числе и текстоформам, которые уже содержат морфологическую информацию, в случае двойной разметки представленную в виде знака «+» в действующем корпусе. Естественно, синтаксическая разметка также нередко оказывается двойной (например, текстоформа одновременно может быть и дополнением, и обстоятельством, входить в состав обособленного оборота, выступать в роли союзного слова). Такая же проблема множественности описаний возникает при анализе клауз. На экране компьютера эта множественность будет представлена в виде серии специальных значков: однако чем их больше, тем труднее пользователю найти нужный. Поэтому в разметку не включаются те синтаксические единицы, поиск которых может быть легко осуществлен с помощью косвенных признаков: например, в корпусе не размечаются восклицательные и невосклицательные предложения, найти которые в корпусе можно по пунктуационному знаку.

3. Принципиальным решением разработчиков является выделение внутри осложняющих оборотов обычных второстепенных членов, то есть распространенное обособленное обстоятельство будет описано как целый комплекс под этой рубрикой, но его составные элементы как нормальные второстепенные члены. Это находящееся в противоречии с традиционной грамматикой решение необходимо для получения точной информации, например, на запрос “все прямые дополнения”. Было бы странно, если бы система выдавала дополнения в “*Иван читает **книгу***”, но игнорировала бы субстантивы в винительном падеже в “*Прочитав **книгу** до половины, Иван принялся за журнал*”.

4. Было принято решение отказаться от внутренней дифференциации типов сложноподчиненных, сложносочиненных и бессоюзных предложений. Причин несколько:

- отсутствие четких границ между типами бессоюзных предложений, традиционные классификации которых строятся на типологии сложносочиненных и сложноподчиненных предложений; конкретные решения при массовой обработке материала были бы открыты для семантической критики [Тестелец 2001: 264]. В то же время выделение непересекающихся типов союзных предложений часто просто невозможно.
- Традиционная классификация союзных сложных предложений в значительной степени соотносится с классификацией союзов (союзных слов – относительных местоимений). Таким образом, поиск, скажем, определительных связей может опираться на леммы “КОТОРЫЙ”, “ЧТО” и др. Определенный процент “шума” при этом неизбежен, но такой поиск окажется достаточно эффективным.
- Типы нерасчлененных сложноподчиненных предложений с коррелятивно-союзной и коррелятивно-местоименной структурой можно будет осуществлять с помощью разных типов скреп.

4. Конструкции с прямой речью особо не выделяются, поскольку она считается явлением текстового уровня: в частности, прямая речь может включать несколько пунктуационно оформленных автономных предложений.

5. Наконец, не используется классификация клауз по цели высказывания (повествовательные, вопросительные, побудительные). Автоматизировать разметку этих типов предложений трудно: формы выражения побуждения многообразны и несводимы к использованию морфологического императива, косвенные вопросы нельзя обнаружить при поиске по вопросительному знаку. К тому же, предлагаемая классификация позволяет обнаруживать некоторые типы, например, побудительных предложений (запрос “самостоятельная инфинитивная клауза + бы” будет выдавать побудительные предложения типа “*Почему бы тебе не помолчать?*”).

IV. Традиционный синтаксис в ХАНКО: проблемы применения

При ожидаемых сложностях применения выбранной схемы разметки к материалу корпуса непосредственная работа выявила ряд дополнительных проблем. Представляется, что их обсуждение имеет смысл и с точки зрения синтаксической теории, и с точки зрения целесообразности использования подобной схемы описания. Ниже перечислены проблемы, с которым столкнулись составители ХАНКО при применении разметки в терминах членов предложения.

1. Нечеткость критериев выделения определенной члена предложения. В ХАНКО изначально предусмотрена возможность двойной разметки, поскольку языковая система с одной стороны, и лингвистическая теория, - с другой, могут иметь своим следствием определенный процент многозначных единиц. В случае применения

разметки по членам предложения это приводит к появлению значительного числа случаев, которые невозможно определить однозначно[3].

a. Косвенное дополнение / несогласованное определение.
(1) *С мостов **через Сену** посрывало гирлянды иллюминации.*

b. Косвенное дополнение / разные виды обстоятельства.
(2) *У дешевых дубленок шкуры могут быть плохо подобраны по цвету и плотности, непрокрашены, и тогда они будут линять **при** влажной **погоде*** (косвенное дополнение / обстоятельство условия*).*

c. Разные виды обстоятельств.
(3) *Ничьей закончились и выборы в Сенат, который партии разделили **ровно** пополам* (обстоятельство образа действия / меры и степени).
(4)*...жестоко избивали хозяев **при попытках** возражать или жаловаться...* (обстоятельство времени / обстоятельство условия).

d. Нечеткость разделения прямого объекта и части сказуемого.
(5) *Российский лидер **соблюдает приличия**...*

2. Еще одной проблемой стал существенно больший список форм выражения разных членов предложения, чем тот, который извлекается из пособий.

a. Подлежащее и главный член односоставного предложения. Хотя в пособиях фиксируются аналитические формы выражение подлежащих (в частности, числовые выражения), все же их список нуждается в корректировке.
(6) *В стотысячной натовской группировке на Балканах уже выявилось почти два десятка смертельных случаев и **до 50 заболевших**.*
(7) *…на покупателя, даже просто пришедшего взглянуть на дубленки, тотчас накидывались сразу **с десяток продавцов***.

3. Серьезной теоретической и технической проблемой стала вложенность членов предложения, а именно плохо решенная в рамках этой теории ситуация, при которой многокомпонентный член предложения может быть разложен на компоненты, которым можно приписать определенную синтаксическую информацию.

a. Обстоятельства и определения, состоящие из нескольких лексем, в частности деепричастные и причастные обороты.
(8) *Холдинг NETBRIDGE заявил о $ 6 млн., **потраченных на проекты List.ru** <...>, и это **не считая собственных проектов**...*

b. Вводные единицы и обращения, которые не являются членам предложения, тем не менее, могут включать в себя зависимые элементы, которые тоже должны быть размечены.
(9) *Капиталисты **всех стран**, соединяйтесь!*
(10) *По* признанию ***менеджеров**, кофейни - дело выгодное, быстро окупаемое и перспективное.*

c. На практике далеко не всегда возможно провести границу между самостоятельными предложениями, вводными предложениями и вводными словами, как это сделано, например в [Валгина 2003: 266]:
(11) ***Известно**, что он хороший парень* (главная часть сложноподчиненного предложения*).*
(12) ***Известно**, он хороший парень (*вводное предложение).
(13) *Он,* **известно**, *хороший парень* (вводное слово*)*

Ситуация осложняется и тем, что на практике вводные слова часто не выделяются запятыми, что создает формальные основания для включения их в состав членов предложения. Однако определить синтаксическую функцию таких единиц представляется трудновыполнимой задачей.
(14) *...и **вообще** работать на комбинате - почти такое же везение, как жить в Москве.*

4. Еще одной проблемой стала разметка составного сказуемого. Это связано со следующим обстоятельствами.

a. Шаткость проведения границы между составным сказуемым и простым сказуемым, выраженным фразеологическим оборотом. Если считать фразеологический оборот составной лексемой, то сказуемое, выраженное такой лексемой должно размечаться как простое. Однако при решении такого рода вопросов приходится опираться на фразеологические словари, которые не отличаются последовательностью. Так

[3] В докладе не затрагиваются вопросы естественной синтаксической омонимии, о которой идет речь, напр., в [Дрейзин 1966, Иорданская 1967].

например, фразеологические словари выделяют фразеологизм «выводить из строя», но не выделяют «выводить из тени». При разметке это должно означать, что в первом предложении глагольное сказуемое «выводят из строя», а во втором – «выводят».

(15) *Они **выводят из строя** оптику.*

(16) *Они **выводят из тени** неизвестных артистов.*

5. Множество единиц «малого» синтаксиса вообще не поддаются никакой разумной синтаксической интерпретации в рамках выбранной классификации.

(18) *И работы Минкульту и Михаилу Швыдкому хватит **еще** надолго.*

(19) *Но **все** чаще и чаще президент проговаривается, обнаруживая истинные свои воззрения.*

(20) *Выходит, что посетители ГРМ **тоже** провинциалы.*

6. Наконец, отметим еще один незафиксированный в пособиях случай, который можно назвать «составным подлежащим», состоящим из инфинитива-связки и «присвязочного» имени.

(21) ***Стать археологом** было ее мечтой.*

V. Выводы

Опыт последовательного применения разметки по членам предложения показывает, что эта теория имеет существенные недостатки, главными из которых можно назвать вложенность однопорядковых компонентов (напр., определение в обстоятельстве) и пересекаемость классификационных признаков (количество случаев теоретической неоднозначности доходит до 30%). Возможным решением этих проблем могло бы стать частичная разметка корпуса, то есть выделение только хорошо определяемых, «чистых» случаев, однако это привело бы к отсутствию разметки для значительной части синтаксических единиц. С другой стороны, часть описанных проблем неотъемлема от диффузной природы языка и будет, следовательно, возникать и при применении альтернативных синтаксических теорий.

Кардинальным (и неосуществимым) решением было бы устранение этой устаревшей теории из практики преподавания (что, по-видимому, означает и перевод пунктуационных правил на другие теоретические основания). Однако в обозримом будущем общеизвестность схемы заставляет использовать ее при разметке корпуса, рассчитанного на самую широкую аудиторию.

Список литературы

Leech G. *Corpus annotation schemes // Literary and Linguistic Computing, 1993, № 8/4. Pp. 275-81.*

Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. *и др., Лингвистическое обеспечение системы ЭТАП-2. М.: Наука, 1989.*

Валгина Н.С., *Современный русский язык. Синтаксис. М.: Высшая школа, 2003.*

Иорданская Л.Н. *Синтаксическая омонимия в русском языке (с точки зрения автоматического анализа и синтеза) // НТИ, 1967, № 5. С. 9-17.*

Дрейзин Ф.А. *Синтаксическая омонимия // Машинный перевод и прикладная лингвистика. М., 1966. № 9. С. 38-43.*

Кустова Г.И., Мишина К.И., Федосеев В.А. *Синтаксис современного русского языка. М., 2005.*

Тестелец Я.Г. *Введение в общий синтаксис. М., 2001.*

Ножов И.М. *Морфологическая и синтаксическая обработка текста (модели и программы) сегментации русского предложения. /АКД, М. 2003.*