Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации.

Богуславский И.М., Григорьев Н.В., Григорьева С.А.,

Иомдин Л.Л., Крейдлин Л.Г., Санников В. З. , Фрид Н.Е.

ИППИ РАН

{bogus, grig, iomdin, lenya}@iitp.ru,

{san, nadya}@proling.iitp.ru

1. Вводные замечания

В настоящей работе описывается проект, посвященный разработке первого аннотированного корпуса русских текстов[1]. Большие текстовые корпуса давно и плодотворно используются в компьютерной лингвистике. Существует по крайней мере 20 крупных аннотированных корпусов для основных европейских языков, самый большой из них содержит несколько сотен миллионов словоупотреблений [Language Resources; Brill, Magerman, Marcus, Santorini 1990; Sadao Kurohashi, Makoto Nagao 1989]. Наиболее известен корпус английских текстов Penn Treebank, созданный в Пенсильванском университете в 1990 году [Marcus, Santorini, Marcinkiewicz 1993]. Для русского языка аннотированных корпусов фактически не существует. Отсутствие аннотированного корпуса русских текстов затрудняет для русистов, работающих в области компьютерной лингвистики, доступ к результатам современных исследований, использующих статистические методы в описании естественного языка. Понимая это, мы предприняли попытку заполнить существующий пробел.

Для разных задач требуются различные уровни аннотации текста, на которых вырабатывается различный объем дополнительных сведений. В рамках настоящего проекта разрабатывается корпус, состоящий из нескольких подкорпусов, тексты которых различаются уровнем аннотации. Выделяются следующие уровни:

- лемматизированные тексты, в которых для каждого слова указывается его основная форма и часть речи;
- тексты с морфологической информацией, в которых для каждого слова указываются его основная форма, часть речи и полный набор морфологических характеристик;
- тексты с синтаксической информацией, в которых для каждого слова указываются его основная форма, часть речи и морфологические характеристики, и для каждого предложения указывается его синтаксическая структура.

Особенностью нашего корпуса является то, что синтаксические структуры предложений изображаются в нем в виде деревьев зависимостей, а не деревьев составляющих, как в большинстве существующих корпусов, разработанных для других языков. Такая структура содержит сведения не только о наличии синтаксической зависимости между теми или иными словами предложения. Она относит каждую такую связь к одному из нескольких десятков синтаксических типов, обеспечивая тем самым наиболее полное и лингвистически содержательное представление синтаксической структуры.

Ближайшим аналогом нашего корпуса может служить корпус, подготавливаемый в настоящее время для чешского языка в Карловом университете в Праге [Hajičova, Panevova, Sgall 1998]. Он так же, как и наш корпус, предоставляет данные о синтаксических зависимостях, однако синтаксические функции слов представлены в нем с несколько меньшей детальностью. Так, в чешском корпусе используется набор из 23 синтаксических функций, а у нас — из 78.

Далее наша работа с корпусом будет охарактеризована в следующих направлениях: характер аннотируемых текстов (п. 2), формат представления данных (п. 3), ввод данных и инструменты разметки (п. 4), типы лингвистической информации (п. 5).

2. **Корпус текстов**

В качестве исходного материала для нашего аннотируемого корпуса послужили тексты машинного корпуса современного русского языка созданного в Уппсальском университете. Общий объем Уппсальского корпуса около 1 млн. словоупотреблений. В Уппсальском корпусе в равных объемах представлены тексты, относящиеся к жанрам художественной прозы и публицистики. Небольшую часть корпуса составляют публикации научного и научно-популярного характера. В корпусе представлено творчество современных русских прозаиков, а также материалы публикаций журналов и газет недавнего периода. Корпус представляет широкий спектр современного литературного русского языка в его письменной форме. Разговорная речь представлена весьма ограниченно в виде диалогов, присутствующих в художественных текстах. Синтаксическая разметка фрагментов корпуса, содержащих элементы разговорной речи, ставит перед лингвистом целый ряд технических и теоретических задач, основной из которых представляется описание эллипсиса. О том, как эта проблема решается в нашем корпусе, будет рассказано ниже.

3. **Формат представления данных**

Мы стремились сделать создаваемый аннотированный корпус применимым в максимально широком круге приложений. Для этой цели нам было необходимо подобрать формат записи аннотационной информации, отвечающий следующим условиям:

- наличие нескольких “слоев” информации, извлекаемых из разметки независимо друг от друга;

- потенциальная расширяемость на типы информации, не охватываемые аннотацией на настоящем этапе;

- возможность синтаксического разбора стандартными программными средствами.

Наиболее подходящим решением представляется использование формализма, построенного на базе SGML/XML. Такой подход принят в формализме TEI (Text Encoding for Interchange) — международном проекте стандартизации языков разметки [TEI Publication 1994]. В своей работе мы старались сделать язык разметки как можно более совместимым с TEI, вводя новые элементы только там, где предлагаемые в TEI решения не позволяют адекватно описать структуру текста в грамматике зависимостей.

Как известно, разметка в XML обозначается маркерами специального вида — тегами. Тегам могут быть приписаны атрибуты, представляющие собой пары “имя–значение” Существует два типа элементов разметки:

- пустые элементы передают “точечную” информацию о разметке; им соответствуют одиночные теги;

- элементы-контейнеры приписывают характеристики разметки сегменту текста. Они передаются парой из открывающего и закрывающего тега. Атрибуты приписываются открывающему тегу.

Ниже перечислены типы информации о структуре текста, которые должны быть отражены в разметке, и предлагаемые способы их кодирования с помощью тегов/атрибутов:

а) Разбивка текста на отдельные предложения. Имеется специальный элемент-контейнер <S> для выделения сегмента текста, составляющего единое предложение: (он есть и в TEI). У открывающего тега может быть (необязательный) атрибут ID — идентификатор предложения, уникальный в пределах текста; его можно использовать для записи информации об отношениях между предложениями в тексте. Другой необязательный атрибут предложения — COMMENT, в который лингвист может записать свои комментарии к синтаксическим явлениям, встретившимся в данном предложении.

б) Разбиение предложений на отдельные лексические элементы (слова). Слова выделяются специальным элементом-контейнером <W>. У слова также может быть атрибутID — идентификатор, уникальный в пределах предложения.

в) Приписывание словам морфологических характеристик. Морфологические характеристики (как категориальные, так и словоизменительные) приписываются словам при помощи набора атрибутов у элемента <W>:

LEMMA — нормализованная форма;

FEAT — морфологические характеристики.

г) Запись информации о синтаксической структуре предложения. Для записи информации о синтаксических связях между словами используются два других атрибута внутри элемента <W>:

DOM — идентификатор (ID) слова-хозяина;

LINK — тип синтаксического отношения.

Кроме того, в формализме предусмотрены специальные средства для записи промежуточных состояний размечаемого текста — множественные морфологические и синтаксические разборы, служебные пометы и пр. Предполагается, что они будут удалены из окончательного вида корпуса.

4. **Ввод данных и инструменты разметки**

Построение аннотированного корпуса осуществляется в полуавтоматическом режиме: аннотация вначале порождается системой автоматического морфологического и синтаксического анализа, а затем корректируется специалистом-лингвистом. Морфологический и синтаксический анализ производится системой машинного перевода ЭТАП-3 [Apresjan, Boguslavskij et al. 1992, 1993], на базе которой построен программный комплекс, состоящий из двух компонентов: а) программа разбиения неразмеченного текста на предложения — Chopper; б) программа построения и редактирования синтаксических структур — StructureEditor.

Степень участия лингвиста в процессе аннотации определяется самим лингвистом в зависимости от сложности структуры текста. Программа StructureEditor предлагает лингвисту несколько режимов работы. Большинство предложений может быть успешно обработано без участия лингвиста, в этом случае требуется лишь очень беглый просмотр и подтверждение правильности работы системы. В случае если анализатор построил структуру, содержащую ошибки, лингвист может сразу отредактировать ее. Если же ошибок слишком много лингвист может прибегнуть к режиму “split-and-run”. В этом случае он разрезает предложение на несколько фрагментов, структура которых представляется более простой. Синтаксический анализатор автоматически построит поддеревья для выделенных фрагментов, а лингвисту останется только связать эти поддеревья в единое дерево.

Если лингвист столкнулся с некоторой сложной синтаксической конструкцией, интерпретация которой требует от него более детального рассмотрения, он может воспользоваться функцией “I’m not certain” и специальным образом отметить узел, место которого в дереве не вполне ясно. После этого система сама отметит предложение, содержащее этот узел, знаком, сигнализирующим о том, что это предложение нуждается в дополнительном редактировании.

На рис. 1 изображено окно редактирования аннотационной информации. Лингвист может редактировать либо всю разметку в целом в специальном окне редактирования разметки, либо воспользоваться графическим интерфейсом для редактирования каждой отдельной сущности. Анализируется предложение *Хотя письмо не было подписано, я мгновенно догадался, кто его написал.* В нем слово *хотя,* с которого начинается это предложение, имеет идентификатор (ID=1), лемму (LEMMA=ХОТЯ), является союзом — (FEAT=CONJ) и зависит по обстоятельственному отношению (LINK=ОБСТ) от слова, имеющего идентификатор 8. Двойным щелчком мыши по любому слову в списке слов можно вызвать окно редактирования свойств отдельного слова.

Рис. 1.

Программа StructureEditor позволяет также отобразить разметку на экране в виде древесной структуры и редактировать ее, используя технику “drag-and-drop”. Как показала практика, такой интерфейс является наиболее удобным и естественным и значительно ускоряет работу над разметкой. Окно редактирования дерева показано на рис. 2.

Рис. 2

Слева написаны слова исходного предложения. Соответствующие им леммы записаны в серых прямоугольниках, а морфологические характеристики – справа от лемм. Синтаксические отношения изображаются в виде стрелок, идущих от подчиняющего слова к подчиненному; тип отношения записан в овальном картуше, расположенном слева от прямоугольника с леммой. Все текстовые элементы, кроме слов исходного предложения, могут редактироваться прямо в окне. Кроме того, картуши с именами отношений можно перемещать в технике “drag-and-drop”. Перетащив картуш на прямоугольник леммы некоторого слова, мы назначаем это слово синтаксическим хозяином того лова, около которого располагался картуш; опустив картуш на то же слово, около которого он находился, мы назначаем это слово вершиной предложения.

5. **Типы лингвистической информации на каждом уровне.**

5.1 Морфологическая информация.

Морфологический анализ приписывает каждому слову морфологические характеристики из следующего списка: часть речи, одушевленность, род, падеж, число, падеж, степень сравнения, краткость, репрезентация, вид, время, лицо, залог.

5.2 Синтаксическая информация.

Как уже было сказано, результатом синтаксического анализа предложения в нашем корпусе является дерево, где каждая стрелка идет из хозяина в слугу и помечена именем одного из синтаксических отношений. Все отношения бинарны, то есть связывают два слова, и ориентированы. В случае синтаксических групп один из членов группы выбирается в качестве представителя во внешних связях группы и подчиняет остальные члены группы.

В типовом случае, число узлов в синтаксической структуре равно числу слов в предложении. Однако в принципе число узлов в синтаксической структуре может быть меньше или, что особенно редко случается, больше числа слов в предложении. Исключения последнего типа сделаны для a) предложений с глаголом БЫТЬ в функции связки, для которого в русском языке возможна нулевая форма. Ср. *Он — учитель*. ОН<--БЫТЬàУЧИТЕЛЬ; b) так называемых синтаксических агломератов. Ср*. негде спать* НЕ à(СПАТЬß БЫТЬà ГДЕ); c) эллиптированных предложений. Ср. *Я купил рубашку, а он галстук.*

На предложениях последнего типа остановимся несколько подробнее. Как известно, эллипсис является одной из наиболее трудных проблем, связанных с формальным описанием синтаксиса естественных языков. Чтобы сделать представление синтаксических структур с опущенными элементами более адекватным и удобным для дальнейшей обработки, в нашем корпусе принято решение восстанавливать эти элементы в явном виде, приписывая им служебный признак "фантом". Так, в предложении с сочинительным эллипсисом *Я купил рубашку, а он галстук между* словами ОН и ГАЛСТУК будет вставлен новый узел синтаксической структуры — форма лексемы ПОКУПАТЬ с такими же морфологическими характеристиками, как реально присутствующая в предложении словоформа КУПИЛ, но имеющая дополнительный признак "фантом". В некоторых случаях может потребоваться модификация морфологических характеристик, как в предложении *Я купил рубашку, а она галстук*, где мужской род глагола заменяется женским. Такой способ представления опущенных слов применим к подавляющему большинству реальных эллиптичных предложений.

Инвентарь синтаксических отношений, автоматически порождаемых системой ЭТАП-3 достаточно обширен, на настоящий момент таких отношений насчитывается 78. Все отношения делятся на 6 больших групп: 1) актантные; 2) атрибутивные; 3) количественные; 4) обстоятельственные; 5) сочинительные; 6) служебные.

Актантные связи связывают предикатное слово с именами его аргументов. Приведем несколько примеров актантных связей, главное слово обозначается как [X], зависимое как [Y]. Предикативное отношение — *Петя* [Y] *читает* [X]; агентивное отношение — *прием* [X] *президентом* [Y] *представителей;* присвязочное — *Он был* [X]*учителем* [Y]; комплетивные отношения — *Транспортировка* [X] *грузов* [Y, 1-КОМПЛ] *от* [Y, 2-КОМПЛ] *причала к* [Y, 3-КОМПЛ] *складу;* предложное отношение — *в* [X]*столе* [Y].

Атрибутивные связи чаще всего связывают существительное с его согласованным или несогласованным определением, выраженным прилагательным, другим существительным, причастным оборотом и т.п. Например, атрибутивное отношение — *дом* [X] *Петра* [Y]; определительное отношение — *красивый* [Y] *дом* [X]; релятивное отношение — *Тот* [X], *кто приходил* [Y]; аппозитивное — *страны* [X]-*члены* [Y] *ООН.*

Отношения *количественной* группы обычно связывают существительное со словом количественной семантики или два таких слова между собой. Например, количественное отношение — *двадцать семь* [Y] *страниц* [X]; количественно-вспомогательное отношение — *двадцать семь* [Y] *страниц* [X]; аппроксимативно-количественное — *человек* [X] *десять* [Y];

Обстоятельственные связи связывают предикатное слово с различными сирконстантами. Например, обстоятельственное отношение — *приходит* [X] *вечером* [Y]; длительное отношение — *Он спит* [X] *по* [Y] *пять часов в сутки*; ограничительное отношение — *Он даже вскочил* [X2]; вводное отношение — *Проблема, конечно* [Y],*существует* [X].

Сочинительные связи обслуживают конструкции с сочинительными союзами. Например, сочинительное отношение — *собираем грибы* [X] *и* [Y] *ягоды;*сочинительно-союзное отношение — *собираем грибы и* [X] *ягоды* [Y]; сентенциально-сочинительное отношение — *Они не придут,* [X] *и* [Y] *мы останемся одни.*

Служебные СинтО обычно связывают два элемента, тесно связанные по смыслу (по сути, образующие неразрывное синтаксическое единство). Например, аналитическое отношение — *Будем* [X] *продолжать* [Y]; вспомогательное отношение — А. [Y] *Пушкин* [X]; *слева* [X] *направо* [Y]; соотносительное отношение — *Что* [X]*касается меня, то* [Y] *я приду.*

Список синтаксических отношений в нашей системе не закрыт. В процессе работы над корпусом мы постоянно сталкиваемся с редкими, малоизученными, не описанными в традиционных грамматиках синтаксическими конструкциями русского языка. В ряде случаев мы пришли к выводу о необходимости введения новых синтаксических отношений для того, чтобы наиболее адекватно отразить в синтаксисе смысловые отношения между отдельными словами и сделать синтаксическую структуру максимально однозначной. В частности было введено уточнительное отношение для описания примыкающих друг к другу семантически связанных адвербиальных групп. Это отношение возникает, когда в СинтС имеется два члена с одинаковой синтаксической функцией, при этом второй из них семантически уточняет первый, Ср. *Встретимся на [X] площади под [Y] часами.*

Приведем еще два примера. Мы ввели синтаксическое отношение для связывания компонентов бессоюзных сложноподчиненных предложений (сентенциально-подчинительное). Типичная ситуация, в которой реализуется это отношение, — предложения с тире, аналогичные по смыслу предложениям с союзом *если*: *Будешь спешить [X] — ошибешься [Y].*

Введено пролептическое синтаксическое отношение, используемое для включения в синтаксическую структуру предложения "предваряющих" элементов. Пролептическая связь от X к Y устанавливается, например, в предложении: *Школа [Y] — это [X] наш дом.*

6. **Заключение**.

Работа над корпусом еще не завершена. На настоящий момент синтаксическая разметка выполнена для 4 тысяч предложений или 55 тысяч словоупотреблений, что составляет примерно треть от планируемого объема. Описанный подход позволяет отразить в корпусе всю смысловую информацию, передаваемую в русском языке морфологическими и синтаксическими средствами. Поэтому мы надеемся, что корпус послужит материалом для широкого спектра исследований как фундаментального, так и прикладного характера

ЛИТЕРАТУРА

Language Resources. // Survey of the State of the Art in Human Language Technology. Eds. G.B.Varile, A.Zampolli. Linguistica Computazionale, vol. XII-XIII, 1997, pp.381-408.

Hajičova E., Panevova J., Sgall P. (1998) Language Resources Need Annotations To Make Them Really Reusable: The Prague Dependency Treebank. // Proceedings of the First International Conference on Language Resources & Evaluation, pp. 713-718.

TEI Publication — Guidelines for the Encoding and Interchange of Machine-Readable Texts. TEI Publication, May 1994. URL: http://www.tei-c.org.

Brill, Eric; Magerman, David; Marcus, Mitchell P.; Santorini, Beatrice, (1990) Deducing linguistic structure from the statistics of large corpora. In Proceedings of the DARPA Speech and Natural Language Workshop, pages 275--282.

Marcus Mitchell P., Santorini Beatrice, Marcinkiewicz Mary Ann (1993). *Building a large Annotated Corpus of English: The Penn Treebank.* Computational Linguistics, Volume 19, No. 2

Sadao Kurohashi, Makoto Nagao (1998). Building a Japanese Parsed Corpus while Improving the Parsing System // Proceedings of the First International Conference on Language Resources & Evaluation, pp.719-724

Apresjan Ju.D., Boguslavskij I.M., Iomdin L.L., Lazurskij A.V., Sannikov V.Z. and Tsinman L.L. (1992). The linguistics of a Machine Translation System. *Meta*, **37** (1): 97-112.

Apresjan Ju.D., Boguslavskij I.M., Iomdin L.L., Lazurskij A.V., Sannikov V.Z. and Tsinman L.L. (1993). Système de traduction automatique **ETAP**. *La Traductique.* P.Bouillon and A.Clas (eds). Les Presses de l'Université de Montréal, Montréal

[1] Работа выполнена при финансовой поддержке РФФИ, грант 98-0790072