Разработка синтаксически размеченного корпуса русского языка

И.М. Богуславский, Н.В. Григорьев, С.А. Григорьева,
Л.Л. Иомдин, Л.Г. Крейдлин, Н.Е. Фрид[1]
Институт проблем передачи информации РАН,
101447, Москва, Большой Каретный пер., 19
iomdin@cl.iitp.ru
тел. (095) 299 4927, факс (095) 209 0579

Dependency Treebank for Russian: Current State of Affairs

I.M. Boguslavskij, N.V. Grigoriev, S.A. Grigorieva, L.L. Iomdin, L.G. Kreidlin, N.E. Frid

Аннотация

В течение нескольких последних лет Лаборатория компьютерной лингвистики ИППИ РАН разрабатывает размеченный корпус русских текстов для последующего его использования в широком классе теоретических и прикладных задач. Значительную научную и практическую ценность корпусу придает глубина аннотации текста: в составляемом корпусе – первом в истории аннотированном корпусе для русского языка - тексты снабжаются детальной морфологической и синтаксической информацией. В настоящее время разрабатывается вторая очередь корпуса, по завершении которой общий его объем составит 12 000 синтаксически аннотированных предложений, или свыше 180 000 словоупотреблений. К обеим очередям корпуса после окончания работы будет обеспечен свободный телекоммуникационный доступ.

1. Вводные замечания

Большие корпусы текстов для разных естественных языков прочно вошли в научный обиход. В настоящее время для основных европейских языков создано около 20 таких корпусов разной глубины аннотации, самые крупные из которых насчитывают сотни миллионов словоупотреблений (см. Marcus, Santorini and Marcinkiewicz (1993); Language Resources 1997; Kurohashi and Nagao 1998). Однако для русского языка аннотированных корпусов текстов до 2000 года не существовало вовсе. Корпус, разрабатываемый в ИППИ (Богуславский и др. 2000, Boguslavsky *et al*. 2000), стал первым таким корпусом, после которого русская корпусная лингвистика стала развиваться значительно быстрее. Один из современных русских корпусных Интернет-проектов, корпус ЦЛД–МГУ (см. Сичинава 2001) представляется наиболее перспективным. Разработчики этого корпуса начинают с морфологической разметки, но предполагают в будущем обратиться и к синтаксической аннотации текстов.

Как известно, различные лингвистические задачи требуют разного объема и характера информации о структуре текста. Разрабатываемый в ИППИ корпус состоит из трех частей, различающихся глубиной разметки. 1) **лемматизированные тексты,** в которых для каждой присутствующей в нем словоформы указана словарная форма (лемма) и частеречная принадлежность; 2) **морфологически размеченные тексты,** в которых для каждой словоформы, помимо леммы и части речи, приводится полный набор словоизменительных морфологических характеристик; 3) **синтаксически размеченные тексты,** в которых, помимо полной морфологической разметки каждому предложению сопоставляется синтаксическая структура.

Ниже пойдет речь о последней, самой развитой части корпуса: цифры, характеризующие объем корпуса, касаются именно этой части. Русские тексты

[1] Данная статья выполнена при поддержке Российского фонда фундаментальных исследований (гранты №№ 00-15-98866 и 01-07-90405), которому авторы выражают искреннюю признательность.

аннотируются **пофразно** с помощью **деревьев зависимостей.** Каждому предложению сопоставляется древесная синтаксическая структура, узлы которой соответствуют словам предложения, а ветви помечены именами синтаксических отношений различных типов (в настоящее время используется 78 таких отношений, примерно половину из которых составляют синтаксические отношения, предложенные в традиционной теории «Смысл ⇔ Текст» И.А. Мельчука). Способ представления синтаксической структуры предложения отличает наш корпус от подавляющего большинства синтаксически аннотированных корпусов для других языков, которые используют синтаксис составляющих. Ближайшим аналогом разрабатываемого корпуса является аннотированный корпус чешских текстов Prague Dependency Treebank (PDT), созданный в Карловом университете Праги (Hajicova *et al*. 1998). В этом корпусе также используется грамматика зависимостей, хотя синтаксические отношения укрупнены по сравнению с нашим корпусом (их всего 23). Таким образом, русский корпус зависимостей предлагает более тонкое разграничение синтаксических конструкций; с другой стороны, в чешском корпусе представлена весьма интересная и перспективная для исследовательских целей информация о дискурсивной структуре (актуальном членении) предложения (Bemova *et al*, 1999). Из других корпусов текстов, в той или иной степени применяющих грамматику зависимостей, стоит отметить корпус NEGRA для немецкого языка (Brants *et al* 1999) and корпус Alpino для голландского языка (Van der Beek *et al*. 2001).

2. Содержание корпуса

Содержательно тексты корпуса подразделяются на две группы. Первая группа состоит из художественных произведений, входящих в хорошо известный русский Уппсальский корпус и содержит приблизительно 10000 синтаксически аннотированных предложений (около 15000 словоупотреблений). Вторая группа, активно разрабатывающаяся в настоящее время, содержит короткие (до 30 предложений) тексты общественно-политического, культурного, экономического и научно-технического характера, взятые с популярных новостных лент, доступных в сети Интернет в онлайновом режиме и характеризующихся высоким уровнем редакционной подготовки (yandex.ru, rbc.ru, polit.ru, lenta.ru. strana.ru, news.ru). Вторая группа текстов, таким образом, отражает самое современное состояние русского литературного языка. Было приложены значительные усилия, чтобы отбор текстов для корпуса сделать максимально репрезентативным. К настоящему моменту синтаксически аннотировано около 2000 предложений из второй группы текстов.

3. Формат разметки

С самого начала разработки мы стремились придерживаться следующих трех принципов разметки корпуса:

1) **Многоуровневость разметки**: любой элемент текста должен допускать несколько уровней аннотации разной глубины так, чтобы каждый уровень аннотации мог извлекаться из корпуса и обрабатываться отдельно;

2) **Наращиваемость разметки:** в любой момент глубина аннотации любого фрагмента корпуса может быть увеличена;

3) **Совместимость разметки** с стандартными компьютерными приложениями.

Естественным решением, обеспечивающим соблюдение этих принципов, представляется использование формализма на основе языка XML. Используемый нами формализм хорошо совместим с формализмом TEI (Text Encoding for Interchange) – признанным международным стандартом для языков разметки; правда, наш формализм

несколько расширен за счет элементов, предназначенных для задания древесной структуры.

Разметка текста осуществляется особыми маркерами – тегами. Различаются одиночные теги и теги-контейнеры: первые передают информацию о точечных элементах текста (словах), а вторые - о свойствах отрезков текста. Типы информации о структуре текста, которые отражаются в разметке, и способы их кодирования кратко перечисляются ниже.

1. **Разбивка текста на предложения** осуществляется с помощью парных тегов-контейнеров <S> : </S>. Открывающий тег может иметь параметр - идентификатор предложения: <S ID=идентификатор>, который можно использовать для записи информации об отношениях между предложениями в тексте.

2. **Разбивка текста на лексические элементы осуществляется** с помощью парных тегов-контейнеров <W> : </W>. Слово также может иметь идентификатор, уникальный в пределах предложения: <W ID=идентификатор>;

3. **Морфологические характеристики** приписываются словам с помощью одиночных тегов <HOM>, размещаемых внутри контейнеров <W> : </W>. У тега <HOM> имеются 4 поля: ID – идентификатор, LEMMA – словарная форма слова, (имя лексемы), POS - часть речи и FEAT - морфологические характеристики;

4. **Информация о синтаксической структуре предложения** записывается с помощью особой метки DOM внутри тега <HOM> : <HOM DOM= идентификатор / тип_связи >, где идентификатор указывает на слово, синтаксически подчиняющее данное; а тип_связи задает имя синтаксического отношения между подчиняющим и подчиненным словами.

Формализм обладает достаточной гибкостью и позволяет записывать не только готовые структуры, но и промежуточные состояния размечаемого текста. В частности, поместив несколько тегов <HOM> внутрь одного контейнера <W> : </W>, мы можем сохранить в разметке информацию о вариантах морфологического анализа словоформы; а допустив множественность полей DOM внутри тега <HOM>, мы можем хранить в разметке не только древесную структуру, но и ориентированный граф более общего вида.

4. Инструменты разметки и процедуры ввода данных

Построение аннотированного корпуса осуществляется в полуавтоматическом режиме, вначале структура автоматически порождается системой морфологического и синтаксического анализа, а затем корректируется специалистом-лингвистом. Порождение структуры выполняется системой машинного перевода ЭТАП-3 (Апресян и др. 1989, 1992; Apresjan *et al.* 1992, 1993), на базе которой построен специальный программный комплекс, состоящий из двух программ: сегментации неразмеченного текста на предложения – («Chopper») и построения и редактирования синтаксических структур – «Structure Editor» (рис. 1).

Степень участия лингвиста в процессе зависит от сложности обрабатываемого текста. Большинство предложений получают правильную структуру от системы: в этом случае от лингвиста требуется лишь беглый просмотр и подтверждение структуры. Если же построенная анализатором структура неверна, то лингвист может отредактировать ее, пользуясь простым и дружелюбным графическим редактором (рис.2). В единичных случаях, если фраза настолько сложна и громоздка, что системе «ЭТАП-3» совсем не удается построить ее структуру, лингвист может прибегнуть к аварийному режиму «split and run», при котором предложение вручную сегментируется на два или более фрагментов, поочередно подающиеся на анализ. Синтаксический анализатор

автоматически построит поддеревья для каждого из фрагментов, а лингвисту останется только соединить эти поддеревья в единое дерево.

Если лингвист столкнулся с нестандартной или спорной синтаксической конструкцией, которая требует более детального рассмотрения, он может специальным образом отметить всю фразу или конкретный узел, роль которого в древесной структуре не очевидна. вполне ясно. После этого система сама отметит предложение, содержащее этот узел значком, сигнализирующим о том, что разметка нуждается в дополнительном редактировании.

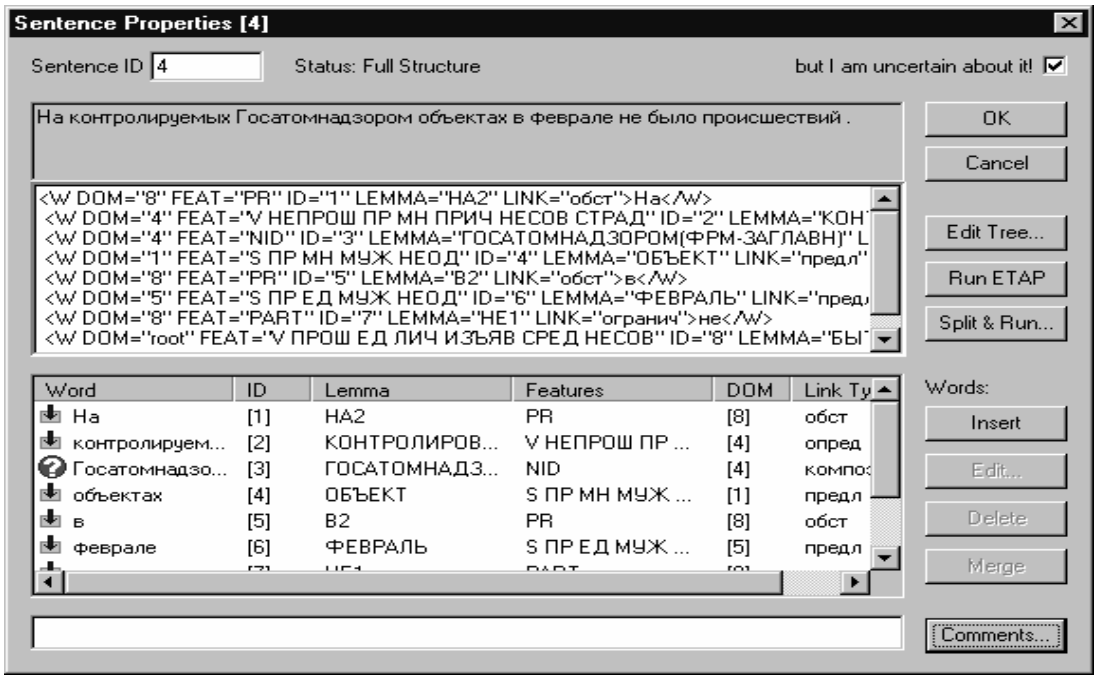

Рис.1. Рабочее поле разметки предложения.

На рис. 1 представлено рабочее поле редактирования разметки предложения. В верхнем светлом окне показано XML-представление обрабатываемого предложения, а в нижнем светлом окне та же информация дается в более наглядном табличном виде. Например, в первой строке окна представлена разметка для первой словоформы обрабатываемой фразы *На контролируемых Госатомнадзором объектах в феврале не было происшествий* – предлога НА2: как видно из последних двух колонок, это слово зависит от слова 8 (*было*) по обстоятельственной связи. Знак вопроса при слове 3 (*Госатомнадзором*) поставлен аннотатором, который к данному моменту не решил, как следует трактовать это не опознанное анализирующим компонентом слово.

Рис.2. Графический редактор структур. Ошибочная разметка

Один из узлов дерева содержит неопознанное слово *Госатомнадзором*, которое к тому же ошибочно подчинено существительному *объектах*, а не причастию *контролируемых*. Пользуясь простейшей технологией «drag and drop», лингвист

вручную заменяет информацию в этом узле и исправляет ошибку (см. рис. 3). Текст XML-разметки автоматически обновляется.

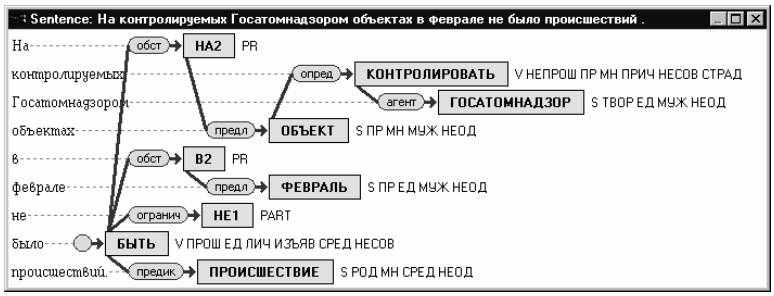

Рис.3. Графический редактор структур. Исправленная разметка.

На рис. 4 представлен фрагмент оглавления размеченного текста в окне структурного редактора. Символические деревья слева показывают, что структура предложения одобрена аннотатором. Вопросительные знаки напоминают, что над структурой еще предстоит поработать. Рис.3. Графический редактор структур. Исправленная разметка.

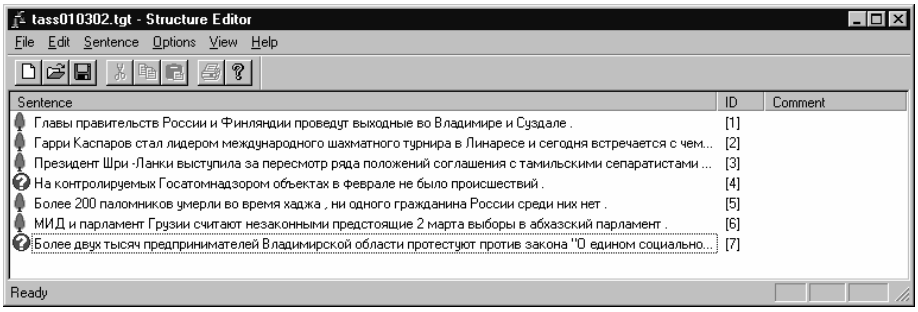

Рис.4. Статус обрабатываемого текста в обобщающем окне структурного редактора.

5. Типы лингвистической информации в размеченном тексте

Морфологическая информация. Морфологический анализ приписывает каждому слову часть речи, а также, в зависимости от частеречной принадлежности, граммемы следующих (словоизменительных и классифицирующих) морфологических категорий: одушевленность, род, падеж, число, степень сравнения, краткость, репрезентация, вид, время, лицо, залог.

Синтаксическая информация. Синтаксическая разметка предложения, как уже было сказано, представлена деревом зависимостей, где каждая стрелка направлена от «хозяина» к «слуге» и помечена именем одного из синтаксических отношений. Синтаксических групп и составляющих формально в структуре нет; фактически любой «куст» дерева можно считать группой, вершина которой выступает в ее качестве представителя во внешних связях «куста».

Как правило, число узлов древесной структуры равно числу слов в предложении. Исключение, с одной стороны, составляют случаи, когда некоторая цепочка словоформ представлена как цельная лексическая единица (*во чтобы то ни стало; ни один)* – тогда узлов в структуре меньше, чем слов, а с другой стороны, случаи, когда какое-то слово должно быть в структуре «расклеено» (таковы, в частности, синтаксические «агломераты» типа *негде, неоткуда*) или когда в структуру необходимо вставить слово, отсутствующее в реальном тексте (нулевая связка *быть* или особые узлы, восстанавливающие некоторые типы эллипсиса: *Отец работал врачом, а мать учителем* ⇒ *Отец работал врачом, а мать РАБОТАЛА учителем*.)

Следует особо подчеркнуть, что в процессе аннотирования полностью разрешается любая лексическая и синтаксическая омонимия обрабатываемого текста. Если этого не удается сделать автоматически (построенная системой структура, на взгляд аннотатора или редактора, не соответствует той, которая имеет место в предложении) она непременно исправляется вручную. В тех исключительных случаях, когда по каким-либо причинам разрешить омонимию не удается и аннотатору (например, в случае нарочитого каламбура, вроде *души прекрасные порывы*), предложение может снабжаться несколькими разметками – формализм, как уже отмечалось в п. 3, это допускает.

Поскольку аннотируемый корпус текстов не связан жестко с каким либо конкретным словарем русского языка, авторы отказались от идеи каким-либо образом помечать конкретные лексические значения слов - как в случае многозначности, так и в случае лексической омонимии. Тем самым в разметке слова *рак1* 'животное' и *рак2* 'болезнь', *топить1* (*печку*) и *топить2* (*подлодку*) не снабжаются никакими индексами. Исключения составляют некоторые служебные слова (в частности, предлоги), лексические значения которых описываются в документации к корпусу.

Двухлетний опыт создания корпуса показывает, что после краткого периода обучения и тренировки (две – три недели) с работой аннотатора легко справляются студенты-лингвисты. Научное редактирование готового корпуса, однако, должно осуществляться профессионалами, имеющими практический опыт работы с крупными системами автоматической обработки текстов, в особенности в области синтаксиса и прикладной лексикографии.

6. Применение аннотированного корпуса в задачах автоматической обработки текстов

Первый тип приложений, на которых мы начали тестировать синтаксически аннотированную часть корпуса, – это автоматическое разрешение синтаксической неоднозначности в ходе синтаксического анализа при машинном переводе с русского языка на английский, осуществляемом системой «ЭТАП-3». В рамках алгоритма синтаксического анализа был разработан дополнительный фильтр, приписывающий веса всем потенциальным поддеревьям обрабатываемого предложения, состоящих из двух – четырех узлов (так называемым N-граммам первого, второго или третьего порядка) в зависимости от их относительной частоты встречаемости в совокупности деревьев, составляющих корпус (подробнее об алгоритме см. Чардин 2001). Веса конкурирующих (т.е. не способных одновременно сосуществовать в готовом дереве) поддеревьев сравниваются с существующими значениями приоритетов индивидуальных синтаксических отношений в текущем рабочем пространстве синтаксических гипотез предложения и модифицируют эти приоритеты, что в конечном итоге способствуют первоочередному порождению наиболее адекватной древесной структуры. Первые результаты можно считать обнадеживающими. Добавим, что результаты такого эксперимента можно использовать и в построении очередных фрагментов самого корпуса, поскольку новые автоматически построенные разборы предложений, учитывающие веса поддеревьев, в целом должны лучше соответствовать отредактированной части корпуса и, как следствие, потребуют меньше ручного редактирования.

Литература

Апресян, Ю.Д., И.М. Богуславский, Л.Л. Иомдин, А.В. Лазурский, Н.В. Перцов, В.З. Санников, Л.Л. Цинман (1989). Лингвистическое обеспечение системы ЭТАП-2. М.: Наука.

Апресян, Ю.Д., И.М. Богуславский, Л.Л. Иомдин, А.В. Лазурский, Л.Г. Митюшин, В.З. Санников, Л.Л. Цинман (1992). Лингвистический процессор для сложных информационных систем. М.: Наука.

Богуславский, И.М., Н.В. Григорьев, С.А. Григорьева, Л.Г. Крейдлин, Н.Е. Фрид (2000). Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации. // Труды Международного семинара Диалог'2000 по компьютерной лингвистике и ее приложениям. Т. 2. Протвино.

Д. В. Сичинава (2001). К задаче создания корпусов русского языка в Интернете. URL: www.mccme.ru/ling/mitrius/article.html

И.С. Чардин (2001). Использование аннотированного корпуса для снятия синтаксической неоднозначности в лингвистическом процессоре ЭТАП-3. // Материалы 2-ой Всероссийской конференции "Теория и практика речевых исследований" (АРСО-2001). М., Филологический факультет МГУ им. М.В. Ломоносова.

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

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

Boguslavsky, I.M., S.A. Grigorieva, N.V Grigoriev, L.G. Kreidlin, N.E. Frid (2000). Dependency Treebank for Russian: Concepts, Tools, Types of Information. // Proceedings of the 18th Conference on Computational Linguistics. Vol 2, 987-991, Saarbrücken.

Brants, Th., W. Skut, and H. Uszkoreit H. (1999). Syntactic annotation of a German newspaper corpus. // Proceedings of the ATALA Treebank Workshop, pp. 69-76, Paris, France.

Van der Beek, L., G. Bouma, R. Malouf, G. van Noord. (2001). The Alpino Dependency Treebank. // Proceedings of LINC 2001.

Bemova, A., J. Hajic, B. Hladka, J. Panevova. (1999). Morphological and Syntactic Tagging of the Prague Dependency Treebank. URL: http://citeseer.nj.nec.com/296119.html

Hajicova E., J. Panevova, P. Sgall (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.

Kurohashi S., M. 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

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

Marcus M. P., B. Santorini and M.-A. Marcinkiewicz (1993). Building a large Annotated Corpus of English: The Penn Treebank. Computational Linguistics, Vol. 19, No. 2.

TEI Guidelines (1994). TEI Guidelines for Electronic Text Encoding and Interchange (P3). URL: http://etext.lib.virginia.edu/TEI.html