И.М. Богуславский, Л.Л. Иомдин, Д.Р. Валеев, В.Г. Сизов

СИНТАКСИЧЕСКИЙ АНАЛИЗАТОР СИСТЕМЫ ЭТАП И ЕГО ОЦЕНКА С ПОМОЩЬЮ ГЛУБОКО РАЗМЕЧЕННОГО КОРПУСА РУССКИХ ТЕКСТОВ[1]

1. Вводные замечания

Синтаксический анализатор, или парсер, разработанный группой ученых Института проблем передачи информации им. А.А. Харевича (ИППИ РАН) для многоцелевого лингвистического процессора ЭТАП-3[2] в значительной мере опирается на лингвистическую теорию «Смысл ⇔ Текст» И.А. Мельчука[3] и в первую очередь на синтаксический компонент этой теории.

В данной работе рассматривается синтаксический анализатор для русского языка. В рамках лингвистического процессора ЭТАП-3 он используется в ряде приложений, включая русско-английский машинный перевод, а также в системе разметки синтаксически аннотированного корпуса русских текстов СинТагРус.

[1] Данное исследование выполнено при частичной финансовой поддержке РФФИ (гранты №№ 07-06-00339, 08-06-00373). Авторы выражают Фонду искреннюю признательность.

[2] См. о нем в частности: *Apresian J., Boguslavsky I., Iomdin L., Lazursky A., Sannikov V., Sizov V., Tsinman L.* ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT // MTT 2003, First International Conference on Meaning – Text Theory. Paris, École Normale Supérieure. Paris, 2003. P. 279–288.

[3] *Мельчук И.А.* Опыт теории лингвистических моделей «Смысл ⇔ Текст». Москва: «Наука», 1974.

2. Основные компоненты парсера ЭТАП-3

2.1. Морфологический анализатор и предсинтаксический модуль

На стадии анализа текста наш парсер начинает работу после того, как **морфологический анализатор**, который обрабатывает каждое отдельное слово предложения, закончит обработку всего текста, предложение за предложением и выдаст морфологическую структуру (МорфС) для каждого предложения. МорфС – это упорядоченная последовательность всех слов предложения, причем каждый элемент такой последовательности содержит имя лексемы (лемму), атрибут части речи и набор словоизменительных морфологических характеристик. Если словоформа лексически и/или морфологически неоднозначна, она представлена в МорфС набором объектов, именуемых **омонимами**, каждый из которых состоит из леммы, атрибута части речи и набора морфологических характеристик.

Например, предложение

(1) *Иностранные рабочие часто плохо знают русский язык.*

получит следующую МорфС:

1.1 ИНОСТРАННЫЙ	A, ИМ, МН
1.2 ИНОСТРАННЫЙ	A, ВИН, МН, НЕОД
2.1 РАБОЧИЙ1	A, ИМ, МН
2.2 РАБОЧИЙ1	A, ВИН, МН, НЕОД
2.3 РАБОЧИЙ2	S, ИМ, МН, МУЖ, ОД
3.1 ЧАСТЫЙ	A, ЕД, КР, СРЕД
3.2 ЧАСТО	ADV
4.1 ПЛОХОЙ	A, ЕД, КР, СРЕД
4.2 ПЛОХО	ADV
5.1 ЗНАТЬ1	V, НЕПРОШ, МН, ИЗЪЯВ, 3-Л, НЕСОВ
6.1 РУССКИЙ1	A, ИМ, ЕД, МУЖ
6.2 РУССКИЙ1	A, ВИН, ЕД, МУЖ, НЕОД
6.3 РУССКИЙ2	S, ИМ, ЕД, МУЖ, ОД
7.1 ЯЗЫК1.	S, ИМ, ЕД, МУЖ, НЕОД
7.2 ЯЗЫК1.	S, ВИН, ЕД, МУЖ, НЕОД
7.3 ЯЗЫК2.	S, ИМ, ЕД, МУЖ, НЕОД
7.4 ЯЗЫК2.	S, ВИН, ЕД, МУЖ, НЕОД
7.5 ЯЗЫК3 .	S, ИМ, ЕД, МУЖ, ОД

Здесь S – существительное, A – прилагательное, V – глагол, ADV – наречие, ОД/НЕОД – характеристика одушевленности/неодушевленности, ИМ/ВИН – именительный/винительный падеж, ЕД/МН – единственное/множественное число, ИЗЪЯВ – изъявительное наклонение, 3-Л – 3-е лицо, НЕСОВ – несовершенный вид, КР – краткая форма.

Из данной МорфС видно, что все словоформы предложения (1), за исключением слова 5 (*знают*), неоднозначны. Так, слово 6 (*русский*) лексически неоднозначно и может выступать как в качестве прилагательного, так и в качестве существительного. Слова 3 (*часто*) и 4 (*плохо*) могут интерпретироваться либо как наречия, либо как прилагательные в краткой форме, в то время как слово 7 (*язык*) имеет три лексических значения: ‘речь’, ‘орган’ и ‘пленный’, где два первых, будучи неодушевленными, имеют совпадающие словоформы для именительного и винительного падежей.

Таким образом, предложение (1), состоящее из 7 слов, обладает МорфС, которая содержит 18 омонимов.

Парсер системы ЭТАП-3 не располагает отдельным тэггером, который бы определял части речи слов, входящих в предложение. Тем не менее в системе ЭТАП-3 имеется небольшой предсинтаксический модуль, который частично разрешает лексическую и морфологическую неоднозначность, учитывая близкий линейный контекст. При работе над предложением (1) данный модуль удалил лишь 2 омонима и уменьшил вес (≈ эмпирическую оценку вероятности того, что именно этот омоним окажется в результирующей синтаксической структуре) еще фразе) одного. В среднем предсинтаксический модуль осуществляет чистку порядка 20% омонимов предложения.

2.2. Парсер

Синтаксический анализатор получает на входе МорфС обрабатываемого предложения и строит для нее **дерево зависимостей**, используя набор синтаксических правил (синтагм). Каждая

синтагма соответствует некоторой минимальной синтаксической конструкции и устанавливает одну гипотетическую именованную направленную синтаксическую связь. В настоящее время в системе ЭТАП-3 используется для русского языка. 65 различных типов синтаксических отношений (СинтО), именами которых и помечаются синтаксические связи в дереве зависимостей. Например, предикативное СинтО в прототипическом случае эксплицирует **синтаксическую зависимость** подлежащего [Y] от личного глагола-сказуемого [X] (*художник* [Y] *видит* [X]); 1-ое комплетивное СинтО отображает отношение между предикатом [X] в качестве синтаксического хозяина и словом [Y], заполняющим его вторую валентность, в качестве слуги (*видит* [X] *свет* [Y]); определительное СинтО, опять-таки в прототипическом случае, характеризует зависимость прилагательного-определения [Y] от определяющего существительного [X] (*нежный* [Y_1] *голубой* [Y_2] *свет* [X]) и т.д. Алгоритм синтаксического анализа использует синтагмы для того, чтобы построить все возможные гипотетические связи между словами предложения, а затем применяет к полученному набору связей систему различных фильтров, которые удаляют лишние связи таким образом, чтобы оставшиеся связи вместе с узлами формировали дерево зависимостей (размеченный ориентированный связный граф без циклов).

В процессе анализа парсер обращается к так называемому комбинаторному словарю русского языка, который в настоящее время насчитывает свыше 100 тыс. статей, содержащих богатую лингвистическую информацию о слове (их синтаксические признаки, семантические признаки (дескрипторы), модель управления и др.

Заметная часть связей в деревьях зависимостей, соответствующих некоторому русскому тексту, является непроективной[1].

[1] В корпусе СинТагРус около 10% предложений содержит по крайней мере одну непроективную связь, однако доля таких связей по отношению к общему числу связей не превышает 1%.

На снимке экрана (рис. 1) представлено синтаксическое дерево зависимостей для предложения

(2) *Иностранные газеты можно купить в киоске на вокзале*,

построенное синтаксическим анализатором ЭТАП-3 (цифры при имени синтаксического отношения означают номера синтагм).

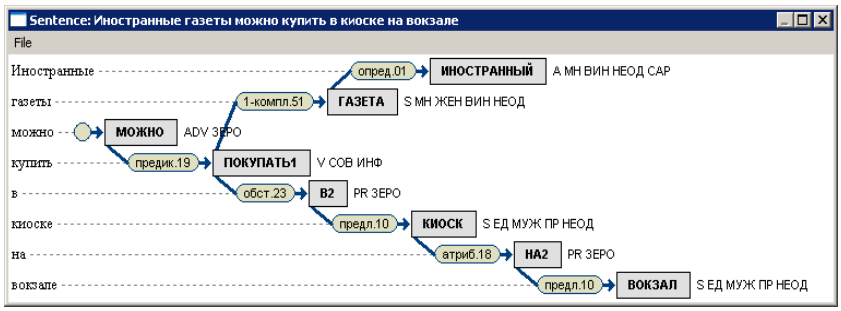

Рис.1. Дерево зависимостей для предложения (2)

Нетрудно заметить, что в этом дереве 1-ая комплетивная связь, идущая от глагола *покупать* к существительному *газета*, является непроективной.

Важной особенностью нашего парсера является его способность производить множество разборов для одного и того же предложения. Хотя разработчики парсера прикладывают все усилия к тому, чтобы первый порождаемый парсером синтаксический разбор предложения был правильным и максимально адекватным обрабатываемому предложению, это не всегда получается. Если первая из построенных парсером структур оказывается неудовлетворительной или если мы хотим получить другую структуру для реально неоднозначного предложения, мы можем снова обратиться к парсеру и поручить ему произвести альтернативный разбор предложения. Такая процедура может осуществляться как автоматически, так и в интерактивном режиме[1].

[1] *Boguslavsky I.M., Iomdin L.L. et al.* Interactive Resolution of Intrinsic and Translational Ambiguity in a Machine Translation System // CICLing

В целом работа парсера ЭТАП-3 может считаться робастной: в худшем случае, если для обрабатываемого предложения не удается получить адекватного дерева зависимостей, «безотказность» парсера достигается за счет того, что некоторые из слов предложения соединяются в аварийном порядке **фиктивным** СинтО. Словам, которые не были найдены в словаре, если их не удается надежно отнести к какой-либо конкретной части речи, приписывается атрибут части речи NID (неопознанное слово).

Как правило, в результирующем дереве зависимостей каждый узел соотносится со словом анализируемого предложения. Исключение составляют те случаи, когда слово образовано с помощью морфологического процесса словосложения и не имеет собственной словарной статьи в комбинаторном словаре (*восьмитомный, стодвадцатилетний* и т.д.). Для такого слова парсер, генерирует два или несколько узлов в дереве зависимостей.

3. Корпус «СинТагРус»

Парсер системы ЭТАП-3 активно используется для построения первого аннотированного корпуса русских текстов СинТагРус[1]. Этот корпус, насчитывающий свыше 35 тысяч статей (около

2005. Lecture notes in computer science / A. Gelbukh (ed.). Springer-Verlag Berlin Heidelberg, 2005. P. 383–394.

[1] *Boguslavsky I., Grigorieva S., Grigoriev N., Kreidlin L., Frid N.* Dependency Treebank for Russian: Concept, Tools, Types of Information // Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000). P. 987–991; *Апресян Ю.Д., Иомдин Л.Л., Санников А.В., Сизов В.Г.* Семантическая разметка в глубоко аннотированном корпусе русского языка // Труды международной конференции «Корпусная лингвистика – 2004». СПб: Изд-во Санкт-Петербургского университета, 2004. С. 41–54; *Апресян Ю.Д., Богуславский И.М., Иомдин Б.Л., Иомдин Л.Л., Санников А.В., Санников В.З., Сизов В.Г., Цинман Л.Л.* Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка 2003–

500 тыс. слов), строится в два этапа: сначала каждое предложение обрабатывается парсером, а затем полученные синтаксические структуры вручную редактируются лингвистами-экспертами. На стадии ручной правки в дерево зависимостей вносятся улучшения, которые не могут быть достигнуты автоматически. В частности, за счет введения дополнительных узлов в синтаксическую структуру более прозрачными оказываются представления предложений со сложным эллипсисом. Так, предложение *Я приехал из Москвы, а он из Мадрида* обзаведется еще одним глаголом *приехал* между словами *он* и *из* и тем самым результирующее дерево зависимостей примет более изящный вид. Такой дополнительный узел (в нашем случае глагол *приехал*) помечается как **фантом**.

В данном исследовании СинТагРус используется в качестве эталона («золотого стандарта») при оценке работы синтаксического анализатора).

4. Метрики оценки

Мы используем два типа оценки: общая оценка и оценка, основанная на штрафах. Обе оценки коротко рассматриваются ниже.

4.1. Общая оценка

4.1.1. Лексико-грамматическая оценка (ЛГ)

В разделе 2 уже отмечалось, что система ЭТАП-3 не имеет отдельного модуля для определения частеречной принадлежности слов обрабатываемого предложения. Дизамбигуация лексико-грамматических характеристик выполняется параллельно с уста-

---

2005 г. (результаты и перспективы). М: «Индрик», 2005. С. 193–214; *Apresjan J., Boguslavsky I., Iomdin B., Iomdin L., Sannikov A., Sizov V.* A Syntactically and Semantically Tagged Corpus of Russian: State of the Art and Prospects // Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, 2006. P. 1378–1381.

новлением синтаксических связей. Представляется поэтому полезным оценивать правильность установления атрибута части речи в процессе синтаксического анализа. Производятся следующие вычисления: сначала, для каждого опознанного слова L вычисляется лексико-грамматический коэффициент $KL = \frac{n_1}{n_2}$, где n_1 – число корректно установленных характеристик слова L, а n_2 – общее число характеристик слова L. Лексико-грамматическая оценка определяется как сумма всех лексико-грамматических коэффициентов KL предложения S, деленная на общее число слов в этом предложении.

4.1.2. Синтаксические оценки слов

1. **Оценка правильности установления хозяев**: доля слов, для которых синтаксический хозяин (или его отсутствие) был установлен корректно (=UAS)[1].

2. **Оценка правильности установления связей**: доля слов, для которых наличие корректной синтаксической связи или отсутствие хозяина было установлено верно, однако, связь может исходить из неверного хозяина.

3. **Оценка правильности установления хозяев и связей**: доля слов, для которых верно установлены как хозяин, так и синтаксическая связь (= LAS)[2].

4. **Статистика по разным типам связей**: для каждого типа связи вычисляются **точность, полнота** и ***F*-мера.**

---

[1] *Nivre J., Scholz M.* Deterministic dependency parsing of English text // Proceedings of the 20th International Conference on Computational Linguistics (COLING). 2004. P. 64–70; *Eisner J.M.* Three new probabilistic models for dependency parsing: An exploration // Proceedings of COLING. Copenhagen, Denmark, 1996.

[2] *Lin D.* A dependency-based method for evaluating broad coverage parsers // Natural Language Engineering 4, 1998. P. 97–114; *Nivre J., Scholz M.* Deterministic dependency parsing of English text…

4.1.3. Синтаксические оценки предложений

Данные оценки могут вычисляться как для всего корпуса, так и для предложений определенной длины, например, для предложений с длинами <10 слов, 10–20 слов, 20–30 слов и т.д.

1. **Оценка вершины**: доля предложений, для которых вершина была установлена правильно.

2. **Оценка правильности скелета**: доля предложений, для которых все связи были установлены верно, не учитывая их тип[1].

3. **Оценка строгой правильности структуры**: доля предложений, для который все связи были установлены верно.

4. **Оценка достижимости эталонной структуры**: доля предложений, для которых эталонная синтаксическая структура достигается на первых *N* альтернативных разборах.

Оценка достижимости эталонной структуры (ДЭС) является важной характеристикой парсера. Как отмечалось в разделе 2.2, парсер ЭТАП-3 может произвести все возможные разборы предложения, которые грамматически верны. Однако порядок выдачи таких альтернатив зависит от веса альтернативы, динамически устанавливаемого парсером. Иногда парсер может получить эталонную структуру, но соответствующий ей разбор предложения находится не в числе первых альтернатив. Таким образом, полезно было бы знать долю предложений, в разборах которых можно отыскать альтернативную структуру, эквивалентную эталонной. Эта оценка помогает определить, сколько некорректных разборов было получено из-за недостатков грамматики, а сколько – из-за несовершенства алгоритма выбора альтернатив. Оценка ДЭС позволяет также узнать, какова доля предложений, для которых можно получить альтернативную структуру, эквивалентную эталонной, если принимать во внимание только первые *N* альтернатив; а также другую дополнительную информацию.

---

[1] *Yamada H., Matsumotu Y.* Statistical dependency analysis with support vector machines // Proceedings of the 8th International Workshop on Parsing Technologies (IWPT). Nancy, France, 2003. P. 195–206.

4.2. Оценка, основанная на штрафах

Данная оценка основывается на разного рода возможных отклонениях разбора (Р) от эталонной структуры (ЭС). Ниже перечисляются эти типы расхождений:

4.2.1. Расхождение токенизации

1. В эталонной структуре присутствует **фантом**, для которого нет эквивалента в (Р) (см. раздел 2.2 выше).

2. Цепочка символов исходного текста по-разному членится на токены в ЭС и Р. Такая ситуация может возникнуть, когда неоднословное выражение (например, фразеологическая единица типа *всё равно*) представлено в корпусе как одно слово, а парсером разбирается как несколько слов. Здесь возможны два варианта событий: (а) словарная разница между предложениями может быть учтена автоматически, и соответствующий многословный узел в ЭС будет соответствовать нескольким узлам в Р (инъективное отображение ЭС в Р и сюръективное отображение Р в ЭС) и (б) когда отображение установить не удается. Приведем по примеру на обе ситуации. примеры: случай (а): слово *антитерроризм* представлено одним узлом в ЭС, но отображается двумя узлами в Р (*анти-* и *терроризм*). Парсер не нашел словарной статьи для слова *антитерроризм*, поскольку оно отсутствует в словаре, зато смог разложить это слово на два элемента, связанных (справа налево) композитным СинтО. В этом случае, *антитерроризм* в ЭС инъективно отображается в Р и соотносится со словом *терроризм* для дальнейшего сравнения. Случай (б): единицы типа *как бы*, которые по тем или иным соображениям были слиты в единые слова редактором корпуса, но которые не могут быть однозначно сопоставлены с какими либо из слов словаря.

4.2.2. Лексико-грамматические расхождения между узлами Р и ЭС с одинаковыми токенами

1. Слово в Р не распознано, для него нет словарной статьи (оно получает частеречную характеристику NID), и оно не может быть разложено на элементы с помощью словообразовательной морфологии.

2. Узлы в Р и ЭС имеют разные характеристики части речи, например, *что* в эталоне случае квалифицируется как союз, а в парсере – как (местоименное) существительное.

3. Узлы в Р и ЭС обладают несовпадающими наборами морфологических характеристик при совпадении атрибута части речи (см. раздел 2.1).

4. Узлы в Р и ЭС имеют разные имена лексемы (леммы), но их части речи совпадают. Например, словоформа *находится* может быть интерпретирована двояко – как глагол, *находиться,* обозначающий ситуацию местонахождения, и как форма страдательного залога глагола *находить*.

4.2.3. Синтаксические расхождения между узлами Р и ЭС с одинаковыми токенами

Все синтаксические расхождения являются взаимоисключающими:

1. В узел структуры в Р идет фиктивная связь (в эталоне фиктивных связей, разумеется, быть не может).

2. Узел является вершиной в ЭС и не является вершиной в Р, или наоборот.

3. В ЭС некоторый узел X подчиняется узлу Z по отношению R, отсутствующему в списке отношений парсера. Такое может произойти, поскольку корпус СинТагРус содержит определенное количество специфических конструкций, не предусмотренных даже весьма полной грамматикой ЭТАПа-3.

4. В ЭС узел X подчиняется узлу Z по отношению R, а в Р он подчиняется узлу Z, но не по отношению R.

5. В ЭС узел X подчиняется узлу Z по отношению R, а в P он подчиняется по отношению R, но не узлу Z.

6. В ЭС узел X подчиняется узлу Z по отношению R, а в P он не подчиняется ни узлу Z, ни по отношению R.

Для каждого типа расхождений предусмотрен свой уровень штрафных баллов. Таким образом, мы можем вычислять штрафы узлов, разборов предложений и корпусов. При этом могут быть использованы два типа оценки.

Ненормализованная оценка очень проста и удобна для сравнения результатов, полученных на одном и том же корпусе в разное время или при разных настройках алгоритма. В частности, такая оценка корпуса может вычисляется путем суммирования всех штрафов, приписанных его предложениям.

Нормализованная оценка позволяет сравнивать результаты, полученные на разных корпусах. Она вычисляется следующим образом: сначала для каждого узла вычисляется штраф, который делится на максимальный штраф, который этот узел может получить, затем все полученные таким образом нормализованные штрафы узлов предложения суммируются, а сумма делится на общее количество узлов предложения; наконец, вычисляется сумма всех нормализованных штрафов предложений корпуса, которая делится на общее число предложений в корпусе.

Кроме генерирования общего штрафа для узла, предложения и корпуса, можно также определить число конкретных типов ошибок, которые позволяют разработчикам парсера оценить точность обработки для специфических синтаксических конструкций. Приведем примеры таких ошибок:

- ошибка в установлении актанта существительного, прилагательного, наречия и глагола в личной/неличной форме;
- ошибка в установлении подлежащего в безглагольном предложении (*Он прав*, *Отец на море*, *Я замужем* и т.д);
- ошибка в установлении неактантной подчинительной связи;
- ошибка в установлении сочинительной конструкции;

- ошибка в установлении служебной связи.

Синтаксическая модель, которая лежит в основе парсера системы ЭТАП-3, содержит близкие, хотя и но тождественные типы синтаксических отношений. Например, для описания определительных конструкций используются два СинтО: определительное и описательно-определительное. Представляется естественным попытаться объединить близкие типы зависимости в одну гипергруппу для того, чтобы добиться увеличения параметра точности синтаксического анализа. Программа оценки позволяет оценить эффект объединения такого рода без внесения сложных изменений в правила. В частности, эта программа может быть настроена таким образом, чтобы игнорировать некоторые типы синтаксических расхождений. Например, возможна оценка разбора корпуса при условии, что отношения **R1** и **R2** рассматриваются как эквивалентные.

5. Оценка парсера ЭТАП-3

Ниже приводятся общие сведения об оценке парсера, полученные для фрагмента корпуса СинТагРус. Данный фрагмент представляет собой все размеченные тексты, вошедшие в корпус в 2007 г. Фрагмент содержит 66 401 слово в 4676 предложениях:

- оценка лексико-грамматической правильности: 0,971;
- оценка правильности установления вершины: 0,868;
- оценка правильности установления хозяина и связи: 0,847;
- оценка правильности скелета: 0,573;
- оценка правильности структуры: 0,253.

Информация об оценке достижимости эталонной структуры приводится в табл. 1.

Здесь МЧА означает заданное максимальное число альтернатив, которые структура может использовать при поиске эталонной структуры среди своих альтернативных разборов, ДЭС – достижимость эталонной структуры или доля структур, среди альтернативных разборов которой была найдена эталонная струк-

тура, ПМЧА – доля неудачных предложений, для которых был превышен лимит максимального числа альтернатив, указанный в столбце МЧА, ПВА – доля неудачных предложений, для которых были перебраны все возможные альтернативы и среди которых эталонной структуры не оказалось, СНА – средний номер альтернативы или, другими словами, средняя позиция успешно найденной эталонной структуры в стеке альтернативных разборов предложения.

Таблица 1. Достижимость ЭС

МЧА	**ДЭС**	**ПМЧА**	**ПВА**	**СНА**
5	0,395	0,659	0,341	2,430
10	0,424	0,596	0,404	3,317
100	0,481	0,169	0,831	10,688

Таблица 2. Достижимость эталонной структуры для разного числа слов в предложении

Длина предложения	**МЧА = 1**	**МЧА = 10**
1–10	0,495	0,721
10–20	0,179	0,371
20–30	0,034	0,100
>30	0,061	0,069

Насколько нам известно, на сегодняшний день не существует подобных исследований русского парсера с результатами которых мы бы могли сравнить полученные данные. Единственным исключением является автоматический (data-driven) MaltParser, созданный И. Нивре, обучающая и контрольная выборки для обучения и оценки которого были созданы на основе того же корпуса СинТагРус[1]. Ясно также, что результатом работы обоих парсеров является одна и та же эталонная структура, тем самым оба парсера находятся в одинаковых условиях.

[1] *Nivre, J., Boguslavsky I., Iomdin L.* Parsing the SynTagRus Treebank of Russian. 2008. [submitted].

Таблица 3. Статистика по синтаксическим отношениям, рассчитанная по первой альтернативе

СинтО	**полнота**	**точность**	***F*-мера**
1-ое комплетивное	0,869	0,887	0,878
2-ое комплетивное	0,799	0,809	0,804
3-ье комплетивное	0,738	0,752	0,745
4-ое комплетивное	0,571	0,667	0,615
аппозитивное	0,689	0,889	0,777
атрибутивное	0,752	0,541	0,629
вводное	0,688	0,860	0,764
длительное	0,667	0,459	0,544
инфинитивно-союзное	0,992	0,977	0,985
квазиагентивное	0,935	0,886	0,910
количественное	0,928	0,918	0,923
неактантно-комплетивное	0,679	0,809	0,738
обстоятельственное	0,764	0,833	0,797
ограничительное	0,890	0,840	0,864
определительное	0,955	0,979	0,967
пассивно-аналитическое	0,884	1,00	0,938
подчинительно-союзное	0,765	0,866	0,813
предикативное	0,862	0,910	0,885
предложное	0,979	0,988	0,984
присвязочное	0,816	0,819	0,818
пролептическое	0,309	0,850	0,453
разъяснительное	0,632	0,577	0,603
релятивное	0,756	0,905	0,824
сентенциально-сочинительное	0,705	0,509	0,591
сочинительно-союзное	0,779	0,885	0,829
сочинительное	0,816	0,874	0,844
сравнительно-союзное	0,765	0,771	0,768
сравнительное	0,813	0,700	0,788
эксплетивное	0,739	0,844	0,788
элективное	0,889	0,988	0,936

Однако прямое сопоставление результатов работы парсера системы ЭТАП-3 с результатами парсера MaltParser вряд ли будет корректным, поскольку исходные данные этих парсеров довольно сильно различаются. Парсер системы ЭТАП-3 обрабатывает непрепарированный текст, в то время как MaltParser начинает свою работу после того, как проработал тэггер части речи. Ввиду того, что у парсера системы ЭТАП-3 такого тэггера нет, исходными данными являются непосредственно эталонные предложения. Это означает, что для парсера MaltParser все расхождения в токенизации и расхождения лексико-грамматического характера между эталонной структурой и предложением (см. раздел 4.2) были уже устранены. Оценить влияние таких расхождений на работу парсера ЭТАП-3 довольно трудно. Тем не менее, результаты работы обоих парсеров можно сравнить, и это сравнение выглядят следующим образом:

- оценка правильности установления хозяина: 0,884 (ЭТАП) vs. 0,891 (MaltParser);
- оценка правильности установления хозяина и связи: 0,844 (ЭТАП) vs. 0,823 (MaltParser).

Что касается исследований парсеров для других языков, то можно сравнить оценку правильности скелета структуры нашего парсера для русского языка с данными, полученными для парсеров английского языка[1]:

[1] *Collins M.* Three generative, lexicalized models for statistical parsing // Proceedings of ACL. Madrid, 1997. P. 16–23; *Charniak E.* A maximum-entropy-inspired parser // Proceedings of NAACL. 2000; *Yamada H., Matsumotu Y.* Statistical dependency analysis with support vector machines…; *Nivre J., Scholz M.* Deterministic dependency parsing of English text…

Таблица 4. Оценка правильности скелета структуры

Charniak	0,452
Collins	0,433
Yamada & Matsumoto	0,384
Nivre & Scholz	0,304
ETAP-3	0,573

6. Анализ ошибок

Как явствует из табл. 3, из 30 синтаксических отношений представленных в корпусе, 7 обладают *F*-мерой превышающей 0,9, а 6 отношений обладают *F*-мерой ниже 0,7. Ниже на кратких примерах иллюстрируются обе группы отношений (хозяин обозначается через X, а слуга – через Y).

6.1. Отношения высокой точности:

- инфинитивно-союзное (*чтобы* [X] *встретить*[Y] *друга*)
- квазиагентивное (*придирки* [X] *со_стороны* [Y] *домовладельца*)
- количественное (*пять* [Y=им] *дней* [X=род])
- определительное (*три опытных* [Y=мн] *работника* [X=ед])
- пассивно-аналитическое (*был* [X] *исключен* [Y])
- предолжное (*в* [X] *длинном списке* [Y])
- элективное (*самая интересная* [X] *из* [Y] *книг*)

6.2. Отношения невысокой точности:

- 4-ое комплетивное (*арендовать* [X] *на* [Y] *три года*)
- атрибутивное (*дом* [X] *у* [Y] *дороги*)
- длительное (*он спит* [X] *пять часов* [Y] *в сутки*)
- пролептическое (*сомнения* [X]*, они* [Y] *никуда не исчезли*)
- разъяснительное (*мы купили все* [X] - *хлеб* [Y]*, сыр, зелень*)
- сентенциально-сочинительное (*они не придут* [X]*, и* [Y] *мы останемся одни*)

Детальный анализ ошибок мы оставляем за пределами данной статьи по соображениям места. Прокомментируем, однако, один тип конструкций, для которого число ошибок оказалось достаточно заметным. Это атрибутивные конструкции, состоящие

из существительного и несогласованного определения при нем. Известно, что атрибутивное СинтО сложно установить, поскольку формальные характеристики отсутствуют, а потенциальных хозяев может быть очень много. Большинство случаев, когда атрибутивная связь оказывается установленной неверно, относится к следующим ситуациям:

- атрибутивная связь установлена вместо обстоятельственной связи, идущей от глагола;
- атрибутивная связь установлена вместо атрибутивной связи, идущей от более далёкого существительного;
- атрибутивная связь установлена вместо аппозитивной связи, в случае, если подчинённый узел является неопознанным (NID) именем собственным, отсутствующем в словаре.

Последний случай неправильного установления атрибутивной связи стоит прокомментировать подробнее. Наличие неопознанных слов значительно сокращает полноту (recall) атрибутивной связи и точность (precision) аппозитивной связи. Такое положение дел можно улучшить, включив на стадии предсинтаксической обработки модуль распознавания имен собственных. Другой стратегией улучшения является улучшение правил угадывания, которые должны определять морфологическую форму слова, даже если оно отсутствует в словаре.

Еще один источник ошибок парсера – неравномерное словарное покрытие. Во многих случаях работает правило «незнание лучше полузнания», то есть лучше совсем не вводить целое семейство лексических единиц в словарь, чем вводить их частично. Представим, например, следующую типичную ситуацию, когда русское название города, скажем, *Красноярск*, присутствует в словаре, а соответствующее ему прилагательное *красноярский* в словаре отсутствует. Из-за специфического пересечения парадигм этих слов (они имеют совпадающие словоформы в структурно разных случаях: творительный падеж существительного совпадает с предложным падежом прилагательного) предложения типа
(3) *Он работает на красноярском заводе*

не будут правильно разобраны, поскольку прилагательное может быть интерпретировано как некое «отбившееся от рук» существительное в творительном падеже. Парсеру же довольно трудно догадаться, что словоформа, обеспеченная словарем, на самом деле представляет собой нечто совершенно другое, для чего в словаре нет эквивалента. Разбор для (3) был бы намного правильнее, если бы в словаре отсутствовало все семейство «красноярских» слов: в этом случае во время обработки предложения возникла бы лишь локальная ошибка разбора, в противном же случае, когда в словаре имеются не все элементы лексического класса, парсер просто не сможет построить сколько-нибудь адекватную структуру.

Заключение

В статье представлен парсер, который является частью многофункционального лингвистического процессора ЭТАП–3 и используется в ряде приложений. Парсер был разработан как для русского, так и для английского языков, но оценка работы парсера была осуществлена только на материале русского языка, в связи с наличием глубоко аннотированного корпуса русских текстов СинТагРус. Отличительной чертой парсера является наличие большого числа типов СинтО, которые позволяют формально представить предложение любой сложности в виде дерева зависимостей. Некоторые из этих СинтО встречаются довольно редко: так, в оценочном фрагменте корпуса обнаружилось лишь 30 СинтО. Для оценки используются различные метрики; одни (основанные на идее штрафов) применимы только для внутренней оценки, другие можно использовать для сравнения с другими парсерами.

В качестве основных направлений дальнейших исследований авторы видят:
1) улучшение правил для СинтО с низкими показателями,
2) разработку правил для обработки эллиптических конструкций,
3) улучшение алгоритма перебора альтернатив,
4) эксперименты, направленные на создание гибридного парсера, включающего, наряду с правилами, элементы машинного обучения.