Л. Л. Цинман

# РАНЖИРОВАНИЕ СИНТАКСИЧЕСКИХ ГИПОТЕЗ В СИНТАКСИЧЕСКОМ АНАЛИЗАТОРЕ ЛИНГВИСТИЧЕСКОГО ПРОЦЕССОРА ЭТАП-3[1]

## Вводные замечания

В предлагаемой статье речь пойдет о русском синтаксическом анализаторе многоцелевого лингвистического процессора ЭТАП-3, разработанном в лаборатории компьютерной лингвистики Института проблем передачи информации РАН (см. об этом процессоре и его приложениях, в частности, [Апресян и др. 1989; Apresjan et al. 2003; 2007]).

Эта статья — вторая публикация по данной теме. В первой статье [Дружкин, Цинман 2008] в основном обсуждались идеи ранжирования гипотез и только вскользь говорилось об экспериментальной проверке этих идей, поскольку эксперименты тогда только начинались. В этой работе предполагается более подробно обсудить результаты экспериментов и те уроки, которые мы извлекли из этих экспериментов.

В первом разделе кратко описывается устройство синтаксического анализатора в системе ЭТАП-3 и объясняется, почему полнота описания приводит к порождению большого числа неверных гипотез. Во втором разделе обсуждается идея ранжирования гипотез как один возможный способ отсеивания неверных гипотез. Третий раздел содержит описание проведенного эксперимента. В четвертом разделе подробно обсуждаются результаты эксперимента. Завершается статья краткими выводами из проведенной работы.

## 1.Проблема: перепорождение гипотез

### 1.1. Синтаксический анализ в системе ЭТАП

Синтаксическая структура фразы в виде дерева зависимостей строится в лингвистическом процессоре ЭТАП-3 с помощью специальных правил (синтагм). Этих правил для каждого из рабочих языков системы (русского и английского) насчи-

[1] Настоящая работа частично поддержана Российским фондом фундаментальных исследований (гранты № 10-06-00478 и 10-07-90001-Бел), которому автор выражает признательность.

тывается несколько сотен. Все они бинарны: это означает, что любая синтагма позволяет связать некоторым синтаксическим отношением два слова фразы, если все условия этой синтагмы, описывающие контекст данной пары слов во фразе, выполнены. Говоря более строго, синтагма связывает синтаксическим отношением не слова фразы, а некоторую пару омонимов этих слов, если они представлены в начале синтаксического анализа несколькими (морфологическими и/или лексическими) омонимами. Таким образом, омонимы слов фразы могут связываться синтаксическими отношениями независимо друг от друга.

В результате работы синтагм на первом этапе синтаксического анализа возникает граф гипотетических синтаксических связей (синтаксических гипотез). На дальнейших этапах работы синтаксического анализатора посторонние связи различными средствами отфильтровываются и из графа синтаксических гипотез выделяется дерево синтаксической структуры фразы. Иными словами, в основе алгоритма синтаксического анализа системы ЭТАП-3 лежит так называемый «фильтровый метод».

Успех работы синтаксического анализатора такого типа зависит от решения двух задач.

**Первая задача**: создание корпуса синтагм, максимально полно описывающих синтаксические явления в естественном языке.

**Вторая задача**: создание развитой совокупности фильтровых средств, позволяющих выделить из графа гипотетических связей искомую синтаксическую структуру.

## 1.2. Уроки массовой разметки текстов

Система ЭТАП-3 находится в экспериментальной эксплуатации уже довольно давно. В частности, в последние годы с помощью этой системы в рамках программы «Создание глубоко аннотированного корпуса русского языка» (подробнее о ней см. [Апресян и др. 2005]) были синтаксически размечены десятки тысяч фраз из разного рода текстов (сейчас в корпусе более 48000 фраз).

Все синтаксические структуры этих фраз сначала «начерно» строились системой ЭТАП-3, а затем при необходимости вручную редактировались специалистами-лингвистами. Массовая синтаксическая разметка текстов стала важнейшей проверкой возможностей нашего синтаксического анализатора. Анализ этой работы выявил два важных обстоятельства.

**Первое:** корпус русских синтагм обладает достаточной полнотой.

Действительно, для абсолютного большинства фраз граф гипотетических связей, построенный на основании синтагм, содержит все связи, которые должны составить правильную синтаксическую структуру этой фразы.

**Второе:** корпус русских синтагм в определенной степени избыточен.

Это означает, в частности, что для ряда фраз, содержательно не являющихся омонимичными, из графа гипотетических связей можно выделить несколько различных деревьев, все синтаксические связи в которых удовлетворяют нашему описанию синтаксиса.

### 1.3. Пример посторонних интерпретаций

Приведем простой пример: пусть на вход синтаксического анализатора поступает вопросительное предложение *Что делает правительство?* С точки зрения любого носителя русского языка, это предложение совершенно однозначно: слово *правительство* здесь является субъектом, подлежащим, а слово *что* — прямым дополнением глагола *делает*. С точки же зрения нашего парсера, это предложение допускает и другие интерпретации: в частности, 1) слово *что* может интерпретироваться как подлежащее, а *правительство* — как дополнение при глаголе *делает*; 2) слово *что* может интерпретироваться как союз, вводящий неполное предложение (такая интерпретация осмыслена, например, в контексте типа *Чем ты недоволен? Что ничего не делает президент? Что делает правительство?*). Очевидно, что вероятность того, что какая-либо из этих двух последних интерпретаций адекватно отражает структуру нашего предложения в каком-либо тексте, исчезающее мала.

### 1.4. Полнота описания приводит к избыточности

Эти особенности лингвистического процессора системы ЭТАП (достаточная полнота и определенная избыточность), на наш взгляд, взаимосвязаны. Так, если лингвист, обслуживающий систему, встречает в тексте синтаксическую конструкцию, не учтенную в синтагмах, то ему достаточно подправить одну из соответствующих синтагм или создать новую, чтобы возникло недостающее синтаксическое отношение. Однако часто бывает так, что некоторая языковая конфигурация, будучи погружена в другие контексты, образует иную синтаксическую конструкцию и должна анализироваться уже иначе. Предусмотреть все эти контексты при написании синтагм, по-видимому, невозможно в принципе. Из этого следует, что синтагмы неизбежно будут порождать лишние, неверные синтаксические гипотезы. К тому же синтагмы работают в тот момент, когда многие слова фразы представлены несколькими омонимами, которые, как говорилось выше, вступают в синтаксические связи независимо друг от друга.

Как показывает опыт эксплуатации парсера ЭТАПа-3, для некоторых фраз количество гипотез может достигать величины 15—20 n, где n — число слов фразы. Естественно, что в ряде случаев наряду с правильной синтаксической структурой из графа гипотетических связей могут быть выделены другие деревья зависимостей, которые хотя и удовлетворяют всем условиям нашего синтаксического описания, но фактически представляют неправильные интерпретации фразы. Таким образом, стремление к полноте описания синтаксиса увеличивает избыточность (в указанном выше смысле) этого описания.

## 2. Возможное решение: ранжирование гипотез

### 2.1. Интерактивный выбор гипотез: достоинства и недостатки

У пользователя, работающего с системой ЭТАП-3, есть возможность затребовать альтернативный синтаксический разбор. В силу упомянутой выше полноты корпуса синтагм для большинства фраз рано или поздно мы получим правильную структуру. Естественно, что такая форма работы с системой неудобна и неприменима при работе с массовым материалом.

### 2.2. Ранжирование синтаксических гипотез

Важной проблемой для нас является оптимизация процесса выделения правильной синтаксической структуры из графа гипотетических связей. Необходимо стремиться к тому, чтобы правильная структура выделялась первой или одной из первых.

Определенные надежды в связи с этим мы возлагаем на идею ранжирования синтаксических гипотез, порождаемых синтагмами.

Проиллюстрируем идею ранжирования синтаксических гипотез на примере всего одной, но весьма частотной русской синтагмы. Эта синтагма (в ЭТАПе она имеет имя *1-компл.11*) связывает 1-м комплетивным синтаксическим отношением переходный глагол (X) с прямым дополнением (Y). Она обслуживает, например, такие конструкции: (1) *купил* (X) *яблоко* (Y) — неосложненное прямое дополнение; (2) *купил* (X) *три яблока* (Y) — дополнение — количественная группа, «в целом» стоящая в винительном падеже (в которой существительное имеет родительный падеж); (3) *привезли* (X) *книг* (Y) *десять* — дополнение — аппроксимативно-количественная группа в винительном падеже (в которой также существительное стоит в родительном падеже); (4) *не купил* (X) *яблок* (Y) — дополнение в родительном падеже при глаголе с отрицанием; (5) *выпил* (X) *квасу* (Y), *выпил* (X) *пива* (Y) — дополнение в партитивном или родительном падеже с количественным значением.

Хотя в приведенных примерах дополнение выражено существительным, оно иногда может выражаться прилагательным, причастием или числительным. Далее, здесь дополнение располагается справа от глагола, но в ряде случаев может стоять и слева от него. Наконец, во всех рассмотренных фразах дополнение стоит рядом с глаголом-хозяином, но встречаются фразы, где оно расположено достаточно далеко от своего хозяина. Естественно, все эти возможности должны учитываться в синтагме.

Как следствие, если в большой фразе встретится переходный глагол, эта синтагма может породить большое число синтаксических гипотез, хотя известно, что глагол в принципе не может иметь более одного прямого дополнения.

Конечно, автор синтагмы, пытаясь сократить порождение посторонних гипотез, старается записать в условиях синтагмы всякого рода дополнительные сведения о словах X и Y и об их ближайшем контексте. Однако поскольку синтагма должна быть рассчитана даже на крайне редко встречающиеся синтаксические конструкции (требование полноты!), то такого рода дополнительных ограничений можно сформулировать не так много.

Рассмотрим теперь гипотетическую фразу, в которой имеется переходный глагол X, справа рядом с ним существительное Y1 в винительном падеже, а слева от X на некотором расстоянии находиться числительное Y2, также в винительном падеже. Пусть оба претендента на роль прямого дополнения X, т. е. Y1 и Y2, прошли все необходимые проверки, требуемые синтагмой. Тогда наша синтагма породит две гипотетические связи с именем *1-компл.11*: X → Y1 и X → Y2. В то же время очевидно, что вероятность вхождения в правильную синтаксическую структуру первой гипотезы неизмеримо выше, чем второй. Естественно, хотелось бы не потерять это знание и использовать его в предстоящем процессе фильтрации. Например, можно пометить эту гипотезу меткой «сильная».

Отметим, что средства для выделения «слабых» и «сильных» гипотез имеются и в действующем варианте синтаксического анализатора. «Сильные» гипотезы возникают, как правило, при описании устойчивых словосочетаний, а «слабыми» помечаются редко встречающиеся синтаксические конструкции. Однако эти средства используются лишь эпизодически.

## 3. Эксперимент: выделение «сильных» гипотез

Для проведения экспериментов мы пополнили парсер системы ЭТАП двумя группами правил, которые подключаются только в случае, когда заявлен экспериментальный режим работы синтаксического анализатора. Такой подход не создает помех работе системы ЭТАП в основном (штатном) режиме.

### 3.1. Первая группа правил

Назначение правил этой группы — ранжировать синтаксические гипотезы, построенные синтагмами. Эти правила работают сразу после построения с помощью синтагм графа гипотетических связей. Этот этап работы парсера мы называем INTERSYNTом. Предполагается, что для большинства основных синтагм в этой группе должно присутствовать соответствующее этой синтагме правило INTERSYNTа. Задача такого правила — просмотреть все гипотезы графа, созданные ассоциированной с этим правилом синтагмой, и приписать некоторым из них помету «сильная». Условия в этих правилах («сильных» INTERSYNTах) представляют собой условия ассоциированной с ним синтагмы, редуцированные таким образом, чтобы при их выполнении синтаксические связи, проводимые этими синтагмами, с высокой вероятностью вошли в синтаксическую структуру фразы.

Редукция условий синтагм сводится, в основном, к следующему:

1. ограничение расстояния во фразе между словами X и Y, которые связываются синтаксической связью (обычно это расстояние не превосходит 2—3);
2. указание на взаимное расположение этих слов (для некоторых синтагм обычным является расположение X левее Y, для других — наоборот);
3. возможно более подробное описание слов, которые могут находиться между X и Y;
4. информация о знаках препинания, которые могут располагаться между X и Y;
5. указание на необходимость или невозможность участия X и Y в некоторых других синтаксических связях в ближайшем контексте, и т. п.

Эти нетривиальные условия в «сильных» INTERSYNTax автор правила формулирует, с одной стороны, на основе **эмпирической статистики**, формирующейся у лингвиста-эксперта, на протяжении долгого времени эксплуатирующего действующую систему, и, с другой стороны, на основе **реальной статистики**, которую при необходимости можно получить с помощью разветвленной программы поиска по синтаксически размеченному корпусу текстов (см. об этом корпусе выше).

Гипотезы, для которых условия правила оказались выполненными, помечаются как «сильные» гипотезы. Следует отметить, что на этой стадии эксперимента мы занимались только усилением (но не ослаблением) гипотез. Кроме того, «сильные» гипотезы в настоящий момент не градуированы «по силе». Мы надеемся, что такое более тонкое описание условий правил станет возможным на следующем этапе эксперимента после накопления большого экспериментального материала.

### 3.2. Избыточность «сильных» гипотез

Поскольку в «сильных» INTERSYNTax рассматриваются локальные фрагменты фразы (обычно 2—3 близко расположенные слова), а слова в этих фрагментах к этому времени могут быть представлены несколькими омонимами, то, вообще говоря, при рассмотрении некоторых фрагментов могут возникнуть одновременно несколько противоречащих друг другу «сильных» гипотез.

Покажем это на примерах.

1. Если во фразе встретилось выражение *поступила установка*, то «сильный» INTERSYNT, ассоциированный с основной предикативной синтагмой, породит сразу несколько «сильных» предикативных гипотез, поскольку оба слова этой пары представлены в словарях ЭТАПа в нескольких значениях (*установка1* ‘механизм’, *установка2* ‘указание’, ⟨…⟩, *поступать1* ‘появляться’, *поступать2* ‘вести себя’,…). Какие значения этих слов войдут в синтаксическую структуру фразы, пока неизвестно. Это пример лексической омонимии.
2. Пусть во фразе встретилось выражение *автобус догоняет троллейбус*. Поскольку у обоих существительных этого выражения совпадают формы име-

нительного и винительного падежей, то каждое из них может выполнять роль как подлежащего, так и прямого дополнения. Поэтому для этого выражения строятся 4 «сильные» связи (две с именем *предик.01* и две с именем *1-компл.11*). Это пример синтаксической омонимии.

Итак, в результате работы «сильных» INTERSYNТов часть синтаксических гипотез получает помету «сильная». Как показали эксперименты, примерно половина правильных синтаксических связей на этапе INTERSYNТа получает помету «сильная». Но возникают и посторонние «сильные» гипотезы.

### 3.3. Вторая группа правил

Эта группа, состоящая из 6 правил, предназначена для чистки графа гипотетических синтаксических связей на основе информации о «силе» гипотез. Первое обращение к этим правилам происходит сразу после работы «сильных» INTERSYNТов. Другие обращения производятся на дальнейших этапах синтаксического анализа, когда прорабатывают правила, реализующие разного рода предпочтения одних гипотез над другими.

Поскольку в графе гипотетических связей могут оказаться посторонние «сильные» гипотезы, то производить чистку графа на основе информации о «силе» гипотезы (то есть удалить из графа все гипотезы, ей противоречащие) следует с большой осторожностью. Поясним эту ситуацию на самом простом примере. Пусть во фразе рядом стоят два слова. Одно из них — глагол в личной форме (обозначим его через X), а другое — существительное в именительном падеже (Y), согласованное по характеристикам с этим глаголом. Очевидно, что эти два слова с большой вероятностью выполняют во фразе роль сказуемого и подлежащего. Основная предикативная синтагма наверняка свяжет эти слова отношением с именем *предик.01*. Естественно, правило INTERSYNТа, ассоциированное с этой синтагмой, должно пометить эту гипотезу как «сильную». В абсолютном большинстве случаев наши действия оправданы. Но в реальных текстах можно встретить фразы, где усиление такой связи ошибочно и приводит к появлению посторонней «сильной» гипотезы.

Рассмотрим примеры таких фраз. Символом W в этих фразах помечено правильное подлежащее.

Валентин(W) Иванов(Y) сделал (X) заявление.
Подъезды (W), коридоры, переходы (Y) освещались (X) плохо.
Панель (W) нового типа позволяет (X) датчик (Y) устанавливать вертикально.
Люди (W) в ходе эволюции (Y) оставили (X) много следов.
Съемки(W) горы(Y) завершались(X) вечером.
От этого выиграет(X) все(Y) сообщество(W).
Волосы(W) женщины(Y) достигали(X) пояса.

Во всех этих примерах возникает «сильная» предикативная гипотеза X → Y, которая, однако, в структуру войти не должна. Поэтому чистить граф от гипотез, про-

тиворечащих этой «сильной» связи, нельзя. Такие ситуации необходимо отслеживать и учитывать в формулировках условий правил второй группы.

В данном случае, например, нетрудно заметить, что во всех приведенных примерах слово Y или какой-то омоним этого слова (омоним Y будем обозначать Y') участвует в качестве «слуги» в другой гипотезе, которую тоже естественно пометить как «сильная»:

*Валентин (W)* → *Иванов (Y)* (аппоз)
*коридоры* → *переходы (Y)* (сочин)
*датчик (Y')* ← *устанавливать* (1 компл)
*в ходе* → *эволюции (Y')* (предл)
*съемки (W)* → *горы (Y')* (1 компл)
*всё (Y')* ← *сообщество (W)* (опред)
*Волосы (W)* → *женщины (Y')* (квазиагент)

Обобщая это наблюдение, можно сформулировать одно из ограничений на применение правил этой группы следующим образом: чистить граф на основании «сильной» связи следует только в том случае, если слова, связанные этой гипотезой, или какие-либо их омонимы не участвуют в других «сильных» связях, противоречащих рассматриваемой.

Впрочем, это довольно жесткое ограничение в некоторых случаях может быть ослаблено. Например, для выражений типа *автобус догоняет троллейбус* (распространенный случай омонимии именительного и винительного падежей), как было указано выше, строятся 4 «сильные» связи (две с именем *предик.01* и две с именем *1-компл.11*), ни одной из которых на этом этапе нельзя отдать предпочтения. Но с большой долей уверенности можно удалить все другие не «сильные» гипотезы с этими именами (*предик* и *1-компл.*), исходящие из глагола *догонять*.

В системе ЭТАП-3 более 500 русских синтагм. Для некоторых типов синтаксических связей, возникающих достаточно свободно (атрибутивные, обстоятельственные), вряд ли возможно сформулировать условия усиления гипотезы. Но для остальных синтагм (а их около половины от общего числа) хотелось бы иметь соответствующие «сильные» INTERSYNTы.

В эксперименте в настоящий момент задействовано 111 «сильных» INTERSYNTов. Однако и эти «сильные» правила тоже находятся в процессе отладки. Дело в том, что написание такого правила — весьма непростой процесс. Если сформулировать условия правила достаточно жестко, то вероятность вхождения «сильной» гипотезы, построенной этим правилом, в правильную синтаксическую структуру станет более вероятной, но, с другой стороны, в реальных текстах условия такого правила будут выполняться редко, а значит, оно редко будет порождать «сильные» гипотезы, и ранжирование в таком случае будет малопродуктивным. Если же условия необходимого контекста, описанного в правиле, несколько ослабить, то оно чаще будет порождать «сильные» гипотезы, но среди этих гипотез будет много посторонних.

То же относится и к правилам второй группы нашего эксперимента, которые осуществляют чистку графа гипотетических связей по информации о «силе» гипотез. Более жесткие ограничения на применение этих правил замедлят процесс получения из графа правильной структуры, но уменьшат вероятность ошибочных действий, в то время как менее жесткие условия ускоряют процесс построения синтаксической структуры, но увеличивают возможность ошибки. Разумный компромисс может быть установлен только в процессе масштабных экспериментов на больших текстах.

Добавим, что и при экспериментальном режиме работы синтаксического анализатора остается возможность при построении «плохой» структуры продолжить работу и получать возможные альтернативные структуры. Если же в процессе эксперимента никакой структуры не удается построить, то синтаксический анализатор отключает правила эксперимента и переходит в штатный режим работы.

## 4. Результаты эксперимента

### 4.1. Сравнение работы синтаксического анализатора в двух режимах

Результаты эксперимента оценивались следующим образом. Выше говорилось, что в нашем распоряжении имеется представительный корпус текстов, синтаксически размеченных ЭТАПом и отредактированных экспертами-лингвистами. Этот корпус естественно рассматривать как **эталон**, с которым следует сравнивать работу синтаксического анализатора ЭТАПа в его текущем состоянии. Была написана программа, которая автоматически сравнивает синтаксические структуры фраз эталона с синтаксическими структурами, построенными ЭТАПом, и на основе развитой системы штрафов оценивает работу ЭТАПа. С помощью этой программы мы смогли сравнить работу синтаксического анализатора ЭТАПа в двух режимах: штатном (использовавшемся в лаборатории на тот момент) и экспериментальном (с учетом «сильных» гипотез). Уточним, что оцениваются в обоих режимах структуры анализируемых фраз, полученные первыми.

Различие в работе двух режимов синтаксического анализатора проявляется, главным образом, в тот момент, когда из двух конкурирующих друг с другом гипотез надо оставить одну, а другую стереть. Для этой цели в штатном режиме работы ЭТАПа есть целая группа «правил предпочтения», которые в ряде случаев помогают сделать выбор в пользу одной из гипотез. В остальных случаях выбор довольно случаен. Например, выбирается более короткая связь, а если обе связи одинаковой длины, то выбирается та, которая возникла раньше.

Экспериментальный режим синтаксического анализатора помимо «правил предпочтения» предоставляет еще одно мощное средство направленного выбора синтаксических связей для искомой структуры — информацию о «силе» синтаксических гипотез.

Сравнение двух режимов (штатного и экспериментального) работы синтаксического анализатора проводилась на текстах с общим объемом около 20000 фраз. Для этих текстов результат сравнения двух режимов таков: у 2150 фраз в экспериментальном режиме синтаксическая структура улучшилась (стала правильной или «ближе» к правильной), но для 633 фраз она стала хуже.

Рассмотрим несколько примеров работы синтаксического анализатора системы ЭТАП-3 в двух режимах.

## 4.2. Примеры успешной работы синтаксического анализатора в экспериментальном режиме

(1) *Он не был некомпетентным или коррумпированным руководителем.*

Синтаксическая структура, порожденная для (1) основным режимом ЭТАПа, имеет вид:

*(1а)*


Синтаксическая структура, полученная в экспериментальном режиме, выглядит так:

*(1б)*

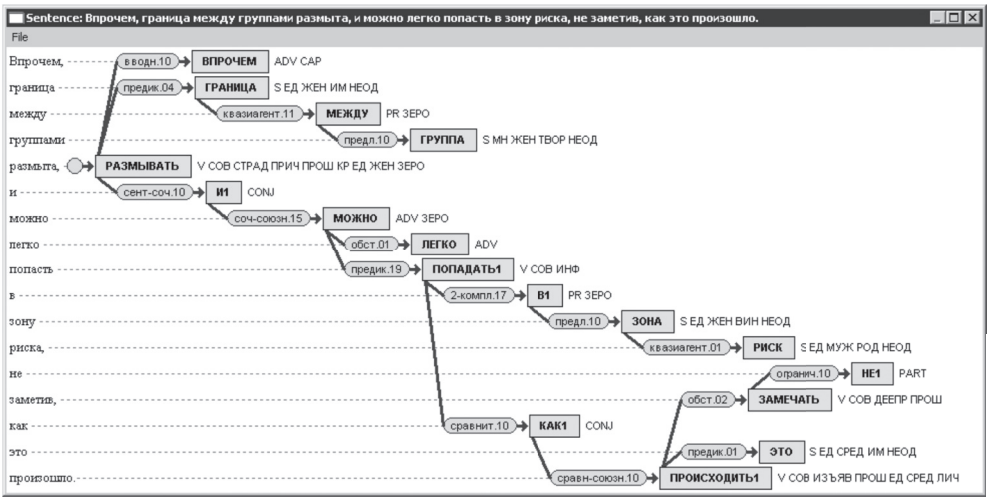

Нетрудно убедиться в том, что, хотя обе интерпретации этой фразы — (1а) и (1б) — с лингвистической точки зрения совершенно законны, (1б) выглядит предпочтительней. Лучший результат в эксперименте был достигнут за счет того, что «сильная» цепочка сочинительных связей сформирована для конфигурации «при-

лагательное + союз + прилагательное», но не для конфигурации пары «прилагательное + союз + причастие, управляющее чем-либо».

(2) *Впрочем, граница между группами размыта, и можно легко попасть в зону риска, не заметив, как это произошло.*

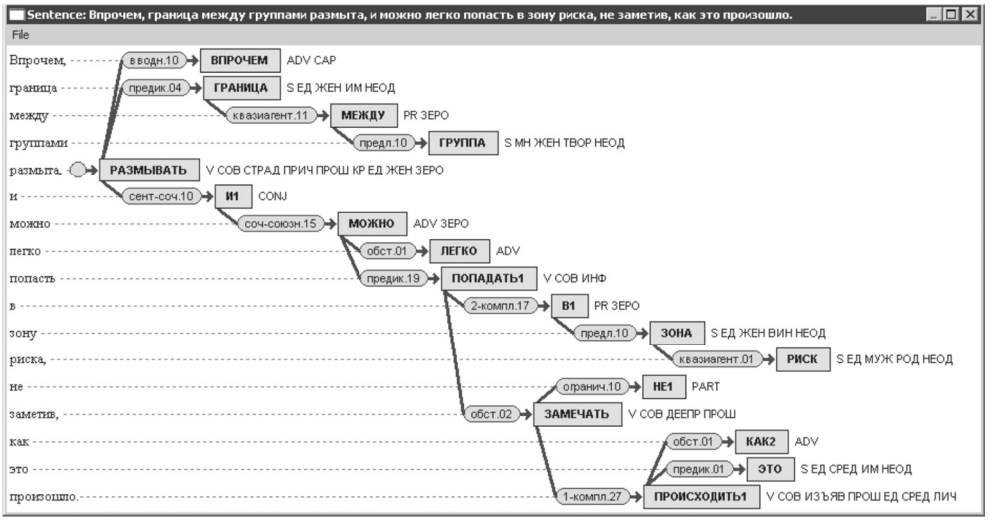
Основной режим ЭТАПа построил для (2) структуру:

*(2а)*




в то время как экспериментальный режим построил другую структуру:

*(2б)*

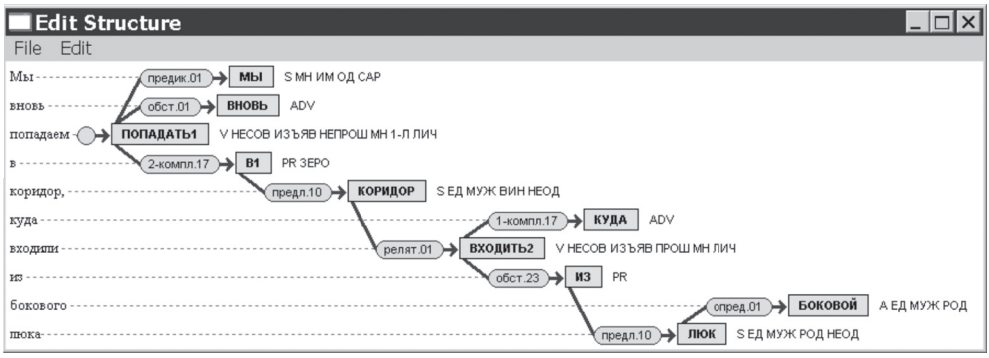


В (2б), в отличие от (2а), заключительный фрагмент предложения (*как это происходило*), правильно подчинен глаголу *замечать* и формирует при нем придаточное дополнительное, вводимое союзным словом *как*. Это произошло благодаря тому, что гипотеза *1-компл.27*, устанавливающая связь между глаголом *замечать* и вершиной этого придаточного *происходить*, была усилена. В основном же режиме структура (2а) интерпретирует этот фрагмент как сравнительный оборот при глаголе *попадать* — законно, но неестественно.

(3) *Мы вновь попадаем в коридор, куда входили из бокового люка*

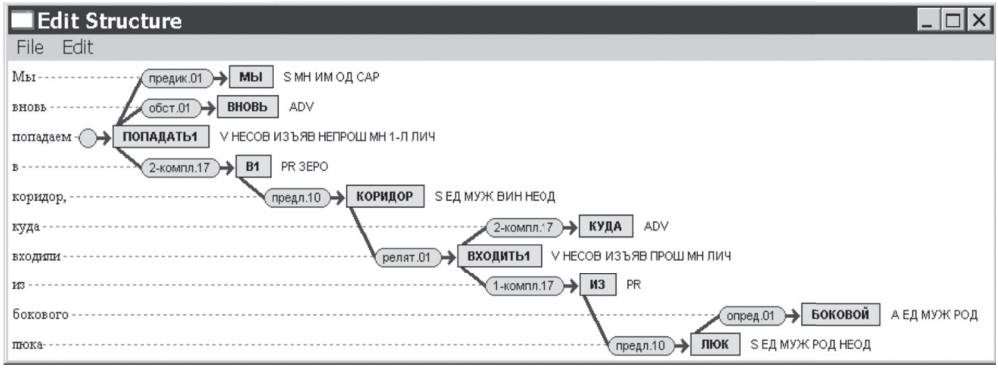
Результат работы штатного режима:

*(3a)*




При экспериментальном режиме была получена другая структура:

*(3b)*




В (3б) глагол *входить* использован в нужном значении *входить1* ‘начинать находиться внутри’, в отличие от (3а), где стоит другое значение этого глагола *входить2* ‘быть составной частью’. Правильное значение глагола в экспериментальном режиме выбирается потому, что из него исходят две «сильные» связи (*1-компл.17 и 2-компл.17*), а у *входить2* только одна такая связь.

### 4.3. Примеры неудачной работы синтаксического анализатора в экспериментальном режиме. Анализ причин этих неудач

Приведем теперь примеры фраз, при анализе которых усиление некоторых гипотез приводит к структуре, отличной от эталонной. При отладке экспериментального режима такие примеры мы изучали особенно внимательно. Именно на этом материале отлаживался эксперимент.

Пожалуй, можно выделить 4 основных типа фраз, которые при работе экспериментального режима оценивались штрафными баллами.

#### 1. Построенная структура синтаксически отличается от эталонной, но выражает тот же смысл

(4) *Банк проверяет остаток средств на счете клиента и разрешает (запрещает) операцию.*

Вместо синтаксической связи *средств* → *на (счете клиента)* (атриб.18) в эксперименте строится связь *проверяет* → *на (счете клиента)* (3-компл.11), для которой есть «сильный» INTERSYNT. Полученная синтаксическая структура отличается от эталонной, но выражает тот же смысл.

#### 2. Причина неправильного анализа — отсутствие «сильных» INTERSYNTов для некоторых синтагм

(5) *Мало этого.*

Конкурируют связи *мало* → *этого* (предик.12) и *мало* → *этого* (1-компл.20). Предпочтение отдается второй, которая усиливается соответствующим INTERSYNTом. Для первой (правильной) связи «сильного» INTERSYNTа у нас нет (довольно редкий случай предикативной связи, где сказуемое — наречие специального типа с подлежащим в родительном или партитивном падеже).

#### 3. «Сильный» INTERSYNT для синтагмы имеется, но мешают принятые в нем ограничения

(6) *Здесь тоже оставляется небольшой кусок проволоки, за который держат при лепке.*

Выбирается «сильная» связь *держат* → *при* (2-компл.17) вместо *за* ← *держат* (3-компл.11), для которой «сильный» INTERSYNT имеется, но в его условиях указано статистически оправданное ограничение — дополнение расположено правее своего хозяина. Заметим, что при этом ошибка усугубляется выбором неправильного значения глагола: *держать2* ‘держать за руку’ вместо *держать1* ‘держать при себе’.

(7) *К ним, под гусеницы, пулеметным огнем погонят немцы сбившихся в кучу людей.*

Выбирается связь *кучу* → *людей* (квазиагент.01) вместо *погонят* → *людей* (1-компл.11).

Первая гипотеза усиливается соответствующим INTERSYNTом, а вторая нет, так как в INTERSYNTе для 1-компл.11 ограничено расстояние между X и Y — не более 3.

(8) *Товарное изобилие рынка пока не более чем мираж, порожденный надеждой вконец изнуренного извечным дефицитом населения на любое экономическое чудо.*

В этой фразе гипотеза *дефицитом* → *населения* (квазиагент.01) усиливается, что приводит к стиранию правильной альтернативной гипотезы *надеждой* → *населения* (квазиагент.01), в которой ограничено расстояние между X и Y — не более 3.

## 4. Правильная гипотеза связывает далеко друг от друга расположенные слова с нетривиальным заполнением интервала между ними

(9) *Каковы, спросите вы, на сегодняшний день реалии?*

Конкурируют гипотезы *день* → *реалии* (квазиагент.01) и *каковы* → *реалии* (предик.04). Первая объявляется «сильной2 и стирает правильную вторую гипотезу (отдаленную предикативную связь, протянутую через вводную конструкцию).

(10) *Как только вы выдергиваете гвоздь, теннисный мяч, картошка или снежок улетают через двор.*

Здесь конкурируют гипотезы *гвоздь* → *мяч* (сочин.10) и *мяч (картошка или снежок)* ← *улетают* (предик.01). Первая объявляется «сильной» и стирает правильную вторую гипотезу (отдаленную предикативную связь, протянутую через сочинительную конструкцию).

Во всех фразах (4)—(10) усиление одной из конкурирующих гипотез приводит к стиранию правильной связи. Некоторых из этих ошибочных действий (4—6) можно избежать, написав недостающие «сильные» INTERSYNTы или дописав имеющиеся (рассмотрев, например, случаи, когда дополнение располагается слева от синтаксического хозяина). Далее, немного ослабив отдельные ограничения в условиях «сильных» INTERSYNTов, можно добиться правильного анализа фраз (7) и (8). Однако это ослабление ограничений может стать источником появления посторонних «сильных» связей в других фразах. Наконец, приходится признать, что с фразами (9)—(10), где контактной связи следует предпочесть связь весьма отдаленную, описанными средствами не справиться. Впрочем, массовые эксперименты на большом корпусе фраз показали, что таких трудных для процедуры ранжирования фраз немного и что для большинства этих фраз штатный ЭТАП также строит ошибочную структуру (то есть экспериментальный режим и на таких фразах практически не ухудшает общую картину). Наконец, всегда остается возмож-

ность затребовать от ЭТАПа построения альтернативных структур, среди которых должна оказаться и правильная.

## 5. Выводы

Мы отдавали себе отчет, что возможности улучшить синтаксический анализ задуманным нами способом ранжировать гипотезы — детальный ***синтаксический*** разбор локальных фрагментов анализируемой фразы — весьма ограничены. Все же 10—15-процентное улучшение результатов синтаксического анализа, которого, как показал эксперимент, можно ожидать от ранжирования гипотез, доказывает, на наш взгляд, целесообразность этой работы.

Не исключено, что чисто синтаксическими средствами добиться дальнейшего заметного улучшения результатов работы парсера уже невозможно. Для этой цели, вообще говоря, требуется семантический анализ фразы. В настоящее время в системы ЭТАП присутствуют лишь элементы семантики, которые требуют своего развития.

Л. Л. Цинман
Институт проблем передачи информации им. А. А. Харкевича РАН,
Москва
cinman@iitp.ru

## ЛИТЕРАТУРА

Апресян и др. 1989 — *Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л., Лазурский А. В., Перцов Н. В., Санников В. З., Цинман Л. Л.* Лингвистическое обеспечение системы ЭТАП-2. М.: Наука, 1989. С. 296.

Апресян и др. 2005 — *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—2005 г. (результаты и перспективы). М: Индрик, 2005. С. 193—214.

Дружкин, Цинман 2008 — *Дружкин К. Ю., Цинман Л. Л.* Синтаксический анализатор лингвистического процессора ЭТАП-3: эксперименты по ранжированию синтаксических гипотез // Компьютерная лингвистика и интеллектуальные технологии, выпуск 7 (14), труды междунар. конф. «Диалог 2008». 2008.

Apresjan et al. 2003 — *Apresjan J. D., Boguslavsky I. M., Iomdin L. L., Lazursky A. V., Sannikov V. Z., Tsinman L. L., Sizov V. G.* ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT // MTT 2003, First International Conference on Meaning — Text Theory (June 16—18 2003). Paris: Ecole Normale Superieure, 2003. P. 279—288.

Apresjan et al. 2007 — *Apresjan J. D., Boguslavsky I. M., Tsinman L. L.* Lexical Functions in Actual NLP-Applications // Selected Lexical and Grammatical Issues in the Meaning—Text Theory. In honour of I. Mel'čuk / Ed. by L. Wanner. Series 84. J. Benjamins, Studies in Language Companion. 2007. P. 199—230.