УДК 004.934

Н.В. Борисова[1], О.В. Канищева[2]
[1]НТУ «ХПИ», м. Харьков, Украина, borisova_nv@mail.ru;
[2]НТУ «ХПИ», м. Харьков, Украина, olya-kanisheva@rambler.ru

МОДЕЛИ И МЕТОДЫ СИНТАКСИЧЕСКОГО АНАЛИЗА

В статье рассмотрены формализмы, модели и методы, применяемые для синтаксического и семантического анализа, проанализированы их преимущества и недостатки, выделены наиболее перспективные для русского и украинского языков. Очерчены признаки для определения типов синтаксических отношений. Приведены лексико-грамматические типы словосочетаний, которые можно использовать в качестве основных синтаксических конструкций для решения различных задач в разных предметных областях.

ЕСТЕСТВЕННЫЙ ЯЗЫК, ГРАММАТИКИ, АВТОМАТИЗИРОВАННЫЙ СИНТАКСИЧЕСКИЙ АНАЛИЗ, АВТОМАТИЗИРОВАННЫЙ СЕМАНТИЧЕСКИЙ АНАЛИЗ

Введение

Автоматизированный синтаксический анализ является одним из наиболее сложных и актуальных направлений в теории компьютерной лингвистики. Синтаксические анализаторы широко применяются в таких областях как создание компиляторов, проектирование интерфейсов баз данных, искусственный интеллект (ИИ), автоматическая обработка текстов (АОТ), в том числе, для автоматизированных информационно-поисковых систем (АИПС), систем машинного перевода (МП), проверки правописания, составления аннотации документа, анализа химических формул и распознавания хромосом.

За несколько десятков лет, в течение которых разрабатывались разные алгоритмы автоматического синтаксического анализа, не было получено ни одного алгоритма, который бы удовлетворительно решал вопросы воссоздания синтактико-семантической структуры текста в формальных моделях для русского и украинского языков.

Ввиду стремительного роста объемов текстовой информации и сложной структурированности естественно-языковых (ЕЯ) текстов, анализ текстов представляет собой актуальную проблему.

Основной целью автоматического синтаксического анализа является получение с помощью алгоритмов в явном виде синтаксической структуры предложения, простого и сложного (составного). Под “предложением” понимается чаще всего простое предложение, в котором отсутствуют в качестве составляющих другие простые предложения, а сложное или составное предложение называют “фразой” [1].

В настоящее время все чаще используется еще одно определение синтаксического анализа – это процесс сопоставления линейной последовательности лексем (слов) языка с его формальной грамматикой. Результатом является дерево разбора или синтаксическое дерево. Такой процесс принято называть парсингом.

Цель работы: рассмотреть формализмы, модели и методы, применяемые для синтаксического и семантического анализа. Выделить среди рассмотренных наиболее перспективные.

1. Изображение синтаксической структуры

Существуют различные способы изображения синтаксической струк

туры предложения, среди которых можно выделить три основных и наиболее распространенных: скобочная запись; изображение зависимостей в виде стрелок, направленных от управляющего слова к управляемому; изображение синтаксической структуры в виде дерева.

Рассмотрим все три способа на примере предложения: *Синтаксический анализ – развивающаяся область прикладной лингвистики*.

Скобочная запись структуры этого предложения будет иметь вид:

((Синтаксический анализ) – (развивающаяся область) (прикладной лингвистики))

Рис. 1. Синтаксическая структура предложения в виде скобок

В данном случае в скобках заключены слова, непосредственно связанные друг с другом зависимостью синтаксического характера.

Изображение синтаксической структуры в виде стрелок от управляющего слова к управляемому можно представить следующим образом:

Рис. 2. Синтаксическая структура предложения в виде стрелок

Запись в виде дерева предполагает изображение структуры предложения в виде некоторого графа:

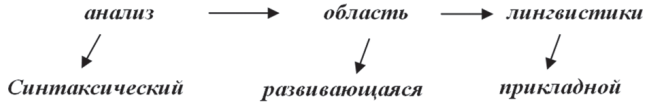

Рис. 3. Синтаксическая структура предложения в виде дерева

В русском, украинском и других языках деревья подчинения предложений “делового стиля”

подчиняются, как правило, закону проективности (рис. 2), состоящему в том, что все стрелки можно провести над прямой, на которой записано предложение, таким образом, что никакие две из них не пересекутся и корень не будет лежать ни под какой стрелкой [2]. В языке художественной литературы, особенно в поэзии, отклонения от закона проективности допустимы и чаще всего служат задаче создания определенного художественного эффекта (рис. 4).

Рис. 4. Нарушение закона проективности

2. Представление синтаксической структуры

В настоящее время можно отметить два основных способа представления синтаксической структуры, вариациями которых можно считать все другие способы.

Если использовать термин “грамматика” для обозначения системы связей, то два наиболее популярных способа называются “грамматикой непосредственно составляющих” (constituent grammars) и “грамматикой зависимостей” (dependency grammars).

Грамматика непосредственно составляющих отмечает в предложении наиболее связанные между собой по смыслу слова (рис. 5).

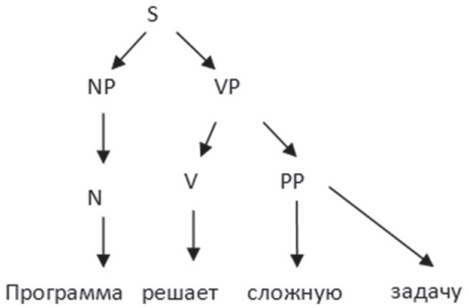

Рис. 5. Представление синтаксической структуры предложения в виде структуры составляющих

Грамматика зависимостей базируется на представлении об управлении, и в ее рамках наиболее популярен способ обозначения связей посредством стрелок.

Грамматика непосредственно составляющих построена на факте пространственного следования групп, составляющих предложение, – именных, глагольных, обстоятельственных и т.п. Здесь возможны разные виды связей. Предложение в целом рассматривается как совокупность групп, сводящихся к двум главным, образующим предложение: группе имени существительного (подлежащего) и группе глагола (сказуемого).

Грамматика составляющих используется преимущественно в описаниях языков с фиксированным порядком слов, грамматика зависимостей – в описаниях языков со свободным порядком слов.

3. Подходы, использующиеся при синтаксическом анализе

Для каждого языка имеется свой собственный синтаксис. Описать синтаксис – это значит описать структуру допустимых фраз. Для описания синтаксических структур наиболее часто используются такие формализмы [3]:

1. Деревья синтаксического подчинения.
2. Системы составляющих.
3. Расширенные сети переходов.
4. Различные виды грамматик.

Деревья синтаксического подчинения не лишены недостатков, например, таких как: 1) связь между словосочетаниями не передается; 2) трудно использовать для сложных предложений с сочинительными конструкциями, вводными оборотами и т.д. Наличие этих недостатков привело к тому, что деревья синтаксического разбора используются на этапе представления синтаксических структур предложения, а не при автоматическом синтаксическом анализе.

Системы составляющих также используются на этапе представления синтаксических структур предложения. Основной их недостаток – это отсутствие синтаксических связей между словами.

Расширенные сети переходов относятся к одному из методов синтаксического анализа, основанного на грамматиках. Этот подход хорошо себя зарекомендовал для таких языков как английский, немецкий, но для флективных языков, к которым относятся русский и украинский, его использование затруднительно.

Методы, основанные на грамматиках, используют две основные стратегии синтаксического анализа:

1. Нисходящая (top-down parsing), когда для данного предложения, исходя из начального символа грамматики, реализуют вывод.

2. Восходящая (bottom-up parsing), когда для данного предложения, исходя из символов самого предложения (терминалов), реализуют разбор.

Таким образом, различают нисходящие и восходящие анализаторы.

Отличительная черта нисходящего анализа – это целенаправленность. На каждом шаге анализа нисходящие распознаватели формируют цель – найти вывод, начинающийся с некоторого нетерминального символа и порождающий часть входной строки. Распознаватель пытается достичь этой цели путем направленного перебора различных возможностей. Основная идея нисходящего анализа в следующем: начиная процесс анализа входной строки, распознаватель исходит из предположения, что эта строка является предложением

входного языка. Отсюда вытекает главная цель анализа – найти вывод (построить дерево).

Если существует такой вывод, то существуют и промежуточные порождающие правила, но для каждого нетерминального символа в грамматике может быть несколько правил с разными правыми частями и какое именно правило следует применить, заранее неизвестно. При неудачном выборе правила, вспомогательная цель может оказаться недостижимой, тогда нужно попытаться применить другое правило. Возможны случаи, когда для какой-либо вспомогательной цели все правила приводят к неудаче. Описание процесса завершается, когда найден конечный вывод (цепочка) или когда установлено, что этого вывода не существует, т.е. входная строка не является предложением этого языка. Обычно нисходящий распознаватель просматривает символы входящей строки и символы правой части, применяя правило "слева-направо". Такие распознаватели называют левосторонними.

В памяти машины правила формальной грамматики могут храниться в виде синтаксических таблиц.

Трансляторы широко применяют комбинацию нисходящих и восходящих методов синтаксического анализа. Например, нисходящий анализ выделяет относительно крупные синтаксические конструкции (различные описания, операторы), каждый из которых затем анализируется подробнее методами восходящего анализа.

Методы восходящего анализа нашли широкое применение в действующих трансляторах. Общая идея восходящего анализа состоит в следующем: входная программа рассматривается как строка символов, распознаватель описывает часть строки, которую можно свести к нетерминальному символу, такую часть строки называют фразой. Фразу, прямо приводимую к нетерминальному символу, называют непосредственно приводимой. В большинстве восходящих распознавателей отыскивается самая левая непосредственно приводимая фраза, называемая основой. Основа заменяется нетерминальным символом, во вновь полученной строке опять отыскивается основа, которая также заменяется нетерминальным символом и т.д. Процесс продолжается либо до получения начального символа, либо до установления невозможности приведения строки к начальному символу. Последовательность промежуточных строк, которая заканчивается начальным символом, образует разбор. Если строка неприводима к начальному символу, то входная программа синтаксически некорректна, т.е. не является формой этого языка.

Существуют также грамматика Монтегю, служащая для одновременного описания синтаксических и семантических структур предложения; в них используется сложный математико-логический аппарат [4].

4. Виды грамматик, использующиеся при синтаксическом анализе

Большинство современных теорий формального синтаксиса (универсальная грамматика Н. Хомского [5], грамматики HPSG [6, 7], LFG [8], TAG [9] и др.) совмещают в себе рассмотренные выше способы описания (представления) синтаксической структуры.

Для формального описания синтаксиса ЕЯ для задач автоматического синтаксического анализа используются, как правило, формализмы одного из следующих классов [10].

Регулярные грамматики. В силу своей слабой описательной способности регулярные грамматики неприменимы в качестве базового формализма полноценного синтаксического анализа. Однако они используются для частичного парсинга (shallow parsing), когда не ставится задача полностью определить структуру предложения, а нужно лишь соотнести позиции одних лексических единиц относительно других.

Контекстно-свободные грамматики. Данные формализмы в чистом виде позволяют описывать большинство предложений ЕЯ, значительное количество которых может быть описано формализмами данного класса только в слабом смысле, т. е. с точностью до последовательности словоформ и без определения структуры предложения. Адекватное описание структуры предложения ЕЯ при помощи КС-грамматик невозможно из-за наличия в предложениях естественного языка так называемых разрывных составляющих, которые адекватно при помощи формализмов данного типа описаны быть не могут.

Контекстно-свободными правилами неудобно описывать согласование (например, в лице и числе между подлежащим и сказуемым). Аппарат КС-грамматик неудобен также для отображения разорванных зависимостей, вызванных передвижением слов во фразе, или для описания отсутствия составляющих.

Мягко контекстно-зависимые грамматики. Эти грамматики были предложены как средство описания структуры составляющих предложений естественных языков. Являясь формализмами более мощными, чем контекстно-свободные грамматики, они, в принципе, позволяют адекватно описывать синтаксические структуры предложений языков разрывными составляющими, допуская при этом полиномиальную вычислительную сложность анализа. Одно из современных направлений исследований по формальному синтаксису состоит в поисках формализмов, которые представляли бы собой точную модель синтаксиса естественных языков.

Вопрос о том, какой формализм лучше всего подходит для описания синтаксиса естественного языка, остаётся открытым.

Очевидно, что простых КС-грамматик недостаточно, однако, с другой стороны — существуют конструкции, структура которых может быть описана контекстно-свободными правилами, но которые, тем не менее, не встречаются в языках.

Для описания синтаксиса естественных языков используются также различные *унификационные формализмы*, в общем случае эквивалентные по мощности машинам Тьюринга, для которых, как известно, проблема определения принадлежности последовательности символов к данному языку неразрешима и парсинг невозможен. Вычислительная универсальность здесь не мотивируется лингвистически, но устройство самого формализма делает его удобной основой той или иной формально-синтаксической теории, которая не использует полную вычислительную мощность формализма.

Для практического создания систем синтаксического анализа естественного языка данный подход представляется достаточно удобным, так как позволяет иметь некоторый запас выразительных средств, заложенный в формализме и позволяющий, в случае необходимости, описывать различные синтаксические явления, не прибегая к контринтуитивным описаниям синтаксиса, в то время как для того, чтобы формализм мог быть адекватной *моделью синтаксиса*, необходимо ограничить его таким образом, чтобы он описывал только такие и только синтаксические структуры, которые возможны в естественных языках.

Существует много формальных синтаксических теорий, часть которых, например LFG, HPSG и TAG, ориентирована на представление синтаксиса естественных языков для автоматического синтаксического анализа, в то время как другие ставят целью объяснить принципы устройства синтаксиса ЕЯ, например такие, как универсальная грамматика Хомского.

В настоящее время существуют два направления в применении строгих математических формализмов в качестве основ синтаксических теорий. Теория может основываться либо на вычислительно неограниченном формализме (unrestricted formalisms), например, унификационные грамматики, либо на каком-то формализме с ограничениями (constrained formalisms). Особый теоретический интерес представляют ограниченные формализмы, так как они, в некотором приближении, могут рассматриваться как математические модели синтаксической способности человека.

Из ограниченных формализмов в синтаксических исследованиях используются в основном древовприсоединительные грамматики (Tree-adjoining grammars, TAGs [9]), а также минималистские грамматики (Minimalist grammars [11]), явившиеся результатом формализации основных аспектов “Минималистской программы” Н. Хомского. Ограниченные формализмы, применимые в синтаксисе, представлены двумя классами: локальными и нелокальными. Локальные формализмы в общем случае мощнее, чем линейные контекстно-свободные переписывающие системы (Linear context-free rewriting systems, LCFRS), однако свойства нелокальных формализмов изучены мало. Известно, что в общем случае проблема принадлежности структур этому классу формализмов неразрешима за полиномиальное время.

5. Определение синтаксических отношений в словосочетаниях

Для создания качественного синтаксического анализатора русского языка необходимо опираться на синтаксические отношения, которые существуют между словами и словосочетаниями.

Слова в предложении могут быть связаны между собой либо сочинительной, либо подчинительной связью, то есть различают два вида связи: сочинение и подчинение [12, 13].

Сочинение – это соединение синтаксически равноправных или независящих друг от друга элементов: однородных членов в простом предложении (кошки и собаки; медленно, но верно; пришёл, увидел, победил) или частей предложения (сложносочиненного и сложного бессоюзного).

Согласование может быть полным (зеленая трава, грамотный человек, жаркое лето) и неполным (наша врач, на озере Байкал, на семи ветрах).

Подчинение – это соединение синтаксически неравноправных элементов (слов в предложении, а также частей сложноподчиненного предложения). В словосочетании между словами существует только подчинительная связь.

Основными видами подчинительной связи являются: согласование, управление и примыкание.

Согласование – это вид связи, при котором зависимое слово уподобляется главному по своей форме, то есть ставится в том же роде, числе и падеже, что и главное слово – существительное или любая часть речи в функции существительного.

Управление – это вид подчинительной связи, при котором зависимое слово (имя существительное или любая часть речи в функции существительного: местоимение, субстантивированное слово, числительное (посмотреть на друга / на него / на сидящих / на обоих) ставится в той падежной форме (без предлога или с предлогом), которая обусловлена лексико-грамматическим значением главного слова (глагола, существительного, прилагательного, количественного числительного в именительном или винительном падеже, наречия или слова категории состояния). Иными словами, главное требует от зависимого определенной падежной формы.

Управление может быть сильным и слабым.

При сильном управлении главное слово своими лексико-грамматическими свойствами предопределяет обязательное появление при нем управляемой падежной формы, то есть связь является необходимой. Сильное управление обнаруживается регулярно при переходных глаголах, может встречаться при существительных и прилагательных определенной лексической семантики, например: послать письмо, слушать радио, уйма времени, тьма дел, верен долгу, предан другу и т.п.

При слабом управлении распространение господствующего слова является факультативным.

Обратим особое внимание и на то, что в некоторых словосочетаниях, несмотря на возможность постановки других, обстоятельственных, вопросов (досиживал (где?) на режиме, расположился (где?) в кузове, расположился (где?) у борта, показалось (где?) в дороге), перед нами управление, на что указывает наличие в этих сочетаниях предлогов. *Предлог – всегда признак того, что перед нами управление, а не примыкание.*

Примыкание – вид подчинительной связи, при котором зависимость подчиненного слова выражается не грамматически, а лексически (по смыслу), порядком слов и интонацией. Примыкают только неизменяемые знаменательные части речи: наречие, инфинитив, деепричастие, простая сравнительная степень прилагательного (дети постарше), неизменяемое прилагательное (цвет хаки), существительное – несогласованное приложение (в газете “Известия”), притяжательные местоимения его, её, их.

6. Лексико-грамматические типы словосочетаний или морфолого-синтаксическая классификация

В зависимости от принадлежности главного слова к той или иной части речи различаются лексико-грамматические типы словосочетаний: глагольные, именные, наречные [13].

Глагольные словосочетания имеют следующие модели:

1) глагол + существительное или местоимение (с предлогом или без предлога): купить хлеба, обратиться к нему;

2) глагол + инфинитив или деепричастие: просить приехать, сидеть задумавшись;

3) глагол + наречие: поступать правильно, повторять дважды.

Именные словосочетания делятся на субстантивные, адъективные, с главным словом числительным и с главным словом местоимением.

Основные модели субстантивных словосочетаний:

1) согласуемое слово + существительное: ясный день, мой мир;

2) существительное + существительное: город в огнях, отрывок из поэмы;

3) существительное + наречие: шаг вперед, лов зимой;

4) существительное + инфинитив: готовность помочь, повод поговорить.

Основные модели адъективных словосочетаний:

1) прилагательное + наречие: по-праздничному нарядный, едва слышный;

2) прилагательное + существительное (местоимение): широкий в плечах, равнодушный ко всему;

3) прилагательное + инфинитив: способный организовать, готовый сопротивляться.

Последние типы словосочетаний с главным словом числительным и с главным словом местоимением являются синтаксически несвободными и разнообразием моделей не отличаются: двое друзей, два товарища, некто в белом, что-нибудь особенное.

Словосочетания наречного типа (с предикативными и непредикативными наречиями) имеют 2 модели:

1) наречие + наречие: по-летнему жарко, весьма вкусно;

2) наречие + существительное: больно руку, высоко в горы, задолго до праздника.

Выводы

Опираясь на результаты морфологического анализа и признаки, присущие различным типам отношений, встречающихся в словосочетаниях, а также на лексико-грамматические типы словосочетаний можно реализовать автоматический синтаксический анализ для решения задачи в определенной предметной области.

Список літератури: 1. *Марчук Ю. Н.* Компьютерная лингвистика [Текст] / Ю. Н. Марчук. – М. : АСТ Восток – Запад, 2007. – 317 с. 2. Лингвистический энциклопедический словарь. Математическая лингвистика. [Электронный ресурс]: http://lingvisticheskiy-slovar.ru/ 3. *Бондаренко М. Ф.* Автоматическая обработка информации на естественном языке / М. Ф. Бондаренко, А. Ф. Осыка. – К. : УМК ВО, 1991. – 144 с. 4. *Парти Б.Х.* Грамматика Монтегю, мысленные представления и реальность [Текст]. – М. : Семиотика, 1983. 5. *Chomsky N., Lasnik H.* The Theory of Principles and Parameters // The Minimalist Program. N. Chomsky. – Cambridge: MIT Press, 1995. 6. *Pollard C., Sag I.* Head Driven Phrase Structure Grammar. – Chicago: University of Chicago Press, 1994. 7. *Rambow O.* Formal and Computational Aspects of Natural Language Syntax: PhD thesis. IRCS Technical Report, University of Pennsylvania, 1994. 8. *Bresnan J.* (ed.) The mental representation of grammatical relations. – MIT Press, 1982. 9. *G.M. Kobele, Michaelis J.* Two Type 0-Variants of Minimalist Grammars // Proc. the 10th conference on Formal Grammar and the 9th Meeting on Mathematics of Language. Edinburgh. 10. *Перекрестенко А. А.* Разработка и программная реализация системы автоматического выделения синтаксических групп для

естественных языков [Текст] / А.А. Перекрестенко// Системы и средства информатики. Вып. 14. – М.: Наука, 2005. 11. *Stabler E.* 1997. Derivational minimalism // Logical Aspects of Computational Linguistics / Ed. C. Retore. – Berlin: Springer Verlag, 1997. – P. 68-95. 12. *Беловольская Л. А.* Синтаксис словосочетания и простого предложения [Текст] / Л.А. Беловольская. – Таганрог, 2001. – 55 с. 13. *Чеснокова Л. Д.* Связи слов в современном русском языке. М., 1980.

Поступила до редколегії 14.06.2012

УДК 004.934

Моделі та методи синтаксичного аналізу / Н.В. Борисова, О.В. Каніщева // Біоніка інтелекту: наук.-техн. журнал. – 2012. – № 2 (79). – С. 89–94.

У статті розглянуто формалізми, моделі та методи, які використовуються для синтаксичного та семантичного аналізу, проаналізовано їх переваги та недоліки, виділено найперспективніші для російської та української мов. Наведено лексико-граматичні типи словосполучень, які можна використовувати у якості основних синтаксичних конструкцій для вирішення різноманітних задач у різних предметних областях.

Іл. 5. Бібліогр.: 13 найм.

UDK 004.934

Models and methods of syntactic analysis / N. V. Borisova, O.V. Kanishcheva // Bionics of Intelligense: Sci. Mag. – 2012. – № 2 (79). – P. 89–94.

This article deals with the formalisms, models and methods, which used for syntactic and semantic analysis, their main advantages and disadvantages has been analyzed, the most advanced for Russian and Ukrainian languages has been singled out. The lexico-grammatical types of word combination, which could be used like main syntactic constructions for solving different problems in the different fields of knowledge has been also stated.

Fig. 5. Ref.: 13 items.