ВВЕДЕНИЕ

Проблема автоматической обработки естественного языка сохраняет свою актуальность по прошествии более полувека со времени первых опытов в данном направлении. Сам факт говорит о высокой сложности задачи и отсутствии ясного представления о путях ее решения. В научной литературе последних десятилетий нашли отражение усилия многих исследовательских коллективов по моделированию частных аспектов анализа и синтеза естественно-языковых текстов, предложены различные методики и приемы, позволяющие добиваться приемлемых результатов для конкретных узких задач. В то же время можно буквально по пальцам пересчитать системы, предполагающие полный автоматический «проход» от текста к его смыслу и/или наоборот.

В настоящей монографии предлагается оригинальный подход к решению задачи автоматического анализа текста, написанного на русском языке, направленный на выявление его информационной структуры. Конечным результатом анализа является метаязыковая сеть, в узлах которой находятся «участники» описываемой в тексте ситуации (объекты), а на дугах — отношения между ними. Программное обеспечение, реализующее данный подход, охватывает все этапы обработки текста на естественном языке (ЕЯ) — от графематики до коммуникативного синтаксиса целого текста. В качестве материала анализа могут использоваться любые правильно построенные письменные тексты публицистического, научного, официально-делового стиля, а также художественной прозы. Это дает нам основание называть нашу модель лингвистическим процессором, понимая под этим термином то же, что и в монографии «Лингвистический процессор для сложных информационных систем» [Апресян и др. 1992: 3].

Книга носит поисковый характер, что отчетливо прослеживается в содержании. Авторы не ставили перед собой цели подробного описания лингвистической модели, включая исчерпывающий перечень правил, применяемых на каждом этапе анализа[1]. Основную ценность, на их взгляд, представляют фундаментальные принципы, на которых основывается предлагаемый подход к автоматической обработки текста. Эти принципы последовательно воплощаются на каждом этапе работы лингвистического процессора. Именно на них стоит остановиться здесь, перед тем как перейти к изложению частных аспектов, связанных с тем или иным видом анализа.

Описываемый подход отличается несколькими существенными чертами, выделяющими его на фоне современных разработок в области автоматической обработки текста. Прежде всего, он, как уже было сказано, является максимально широким, не ограниченным какой бы то ни было узкой задачей, отдельным видом лингвистического анализа или конкретным типом текста. Он носит комплексный характер и включает учет всех видов информации, содержащейся в вербальном компоненте текста на ЕЯ.

Важной отличительной характеристикой разрабатываемого лингвистического процессора является принципиальный отказ от применения вероятностно-статистических методов на этапе анализа текстовой информации. Признавая полезность аппарата математической статистики для ряда узко-прикладных задач в области языкознания, мы считаем, что при создании полномасштабной системы автоматического понимания текста на ЕЯ решающей является не количественная, а качественная сторона. Разнообразные логико-статистические приемы анализа текста, исходящие из примата средств программной реализации и сосредоточенные на поиске все более эффективного алгоритма, который позволил бы хотя бы на долю процента улучшить результат, представляются нам методологически ошибочными. С нашей точки зрения, адекватный анализ содержания может быть достигнут только при тщательном учете языкового контекста, поэтому приоритетной задачей является максимально полное и точное извлечение лингвистических данных из текста. Поскольку речь идет о текстах на русском языке, это объясняет заметное преобладание в библиографии русскоязычных источников.

[1] Описание алгоритмической составляющей модели тем более не входит в задачи книги.

В основе предлагаемого подхода лежат правила интерпретации различного рода информации, заключенной в вербальном компоненте текста. Итогом интерпретации является приписывание анализируемым единицам (символам, словоформам, словосочетаниям, синтаксическим конструкциям, риторическим структурам) значений признаков. Имеются в виду графематические признаки (ср. признак «тип» и его значения ‘буквенный’ / ‘небуквенный’, признак «алфавит» и его значения ‘кириллица’ / ‘латиница’ / ‘прочий’), морфологические признаки, включая часть речи (ср. «род» и его значения ‘мужской’ / ‘женский’ / ‘средний’), синтаксические (например, «тип синтаксической конструкции» со значениями ‘группа сказуемого’, ‘обстоятельственная группа’ и др.) и семантические признаки (ср. признак «фазовость» и его значения: ‘становление’ / ‘возникновение’ / ‘начало’ / ‘осуществление’ / ‘продолжение’ / ‘приостановка’ / ‘возобновление’ / ‘конец’ / ‘предел’ / ‘результат’).

Таким образом, вся извлекаемая из текста графическая, грамматическая и лексико-семантическая информация последовательно представляется в терминах признаков и их значений.

Подобная методологическая последовательность, хотя и делает описание несколько тяжеловесным, представляется нам принципиально необходимой. Следует отметить, что слово *значение* в данной работе используется в двух разных смыслах: как ‘языковое значение’ (ср. англ. *meaning*) и как ‘значение того или иного признака’ (англ. *value*). Что касается слова *признак*, его употребление также неоднозначно. Там, где речь идет о графематическом, морфологическом, лексико-семантическом и синтаксическом анализе, понятие «признак» используется в смысле ‘категории, которой присущи определенные значения’. Однако применительно к организации информационной структуры текста *признак* фактически обозначает информацию, приписываемую в тексте выделенным объектам (один из основных компонентов метаязыковой структуры текста), в том числе и соединяющую их друг с другом.

Переходя к другим особенностям описываемого лингвистического процессора, подчеркнем очевидный факт: традиционные лингвистические описания — словари и грамматики — не ориентированы на задачи автоматической обработки языка, а следовательно, нужные для этой задачи сведения не могут быть получены без дополнительных описаний значений слов и синтаксических структур. В частности, требуется тщательная работа с корпусами текстов, с тем чтобы необходимую информацию можно было выразить им-

пликативно («если… то…»). При отказе от математических методов выработка такого рода эвристик — путь довольно затратный, но способный давать эффективные решения. Мы стремились прежде всего к выработке правил, охватывающих не единичные факты, а их классы, хотя и частные фильтры, касающиеся отдельных значений слов и словосочетаний, представляют известный интерес[2]. На страницах книги читатель найдет немало эмпирически выведенных закономерностей, основанных на кропотливом анализе языковых данных. Собственно говоря, именно такой ракурс анализа текста наиболее точно отражает исследовательский интерес авторов.

Тема неоднозначности языковых выражений и, следовательно, принципов выбора правильной интерпретации проходит через всю книгу. Однако следует подчеркнуть, что, в отличие от подавляющего большинства лингвистов, работающих в области автоматической обработки ЕЯ, мы не ставили перед собой задачу во что бы то ни стало и как можно раньше снимать лексическую и грамматическую неоднозначность. Осознавая значимость этой проблемы, возникающей на всех этапах анализа текста, мы не считаем нужным решать ее сразу и любой ценой. В общем случае неоднозначность сохраняется вплоть до того момента, когда лингвистический анализ позволит ее надежно разрешить. Известно, что лексико-грамматическая омонимия, как правило, снимается на уровне синтаксиса, а для определения референции анафорического выражения нередко требуется выход за границы предложения. В то же время есть случаи, когда неоднозначность не может быть снята в принципе — например, если анафора отсылает к внешнему миру (так называемая «экзофора» [Halliday, Hasan 1976: 18]) или авторский текст намеренно двусмыслен. В подобных ситуациях сохраняются все допустимые варианты прочтения. Следует также добавить, что в задачи лингвистического процессора не входит выявление подтекста, понимание юмора и иносказаний, оценка истинности высказываний, определение художественной ценности текста и т. п.

Работа процессора состоит из нескольких этапов, соответствующих типам анализа (в тексте это отдельные главы). Процессор работает сугубо поступательно. На каждом этапе порождаются все

[2] Показателен в этом отношении пример из работы [Апресян и др. 1992: 76], посвященный снятию лексико-грамматической омонимии словоформы *механику*, которая в сочетании с предлогом *в* отсылает к лемме *механика*, а с предлогом *к* — к лемме *механик*.

возможные варианты интерпретации соответствующих единиц. Следующий этап может либо отменить какие-то из вариантов, либо сохранить их все. Возврат назад, к предшествующему этапу, для пересмотра результатов анализа не имеет смысла (поскольку все допустимые варианты были своевременно порождены).

Помимо алгоритмов, работа анализатора поддерживается специально созданными словарями, в том числе словарем аббревиатур и сокращений, словарем лексических комплексов (включающих составные предлоги, наречия и пр., а также коллокации и фразеологизмы) и словарем частотных имен собственных. Синтаксический анализ опирается на разработанный авторами семантико-синтаксический словарь, описанию которого посвящен специальный раздел.

Сначала осуществляется графематический анализ всего текста, алгоритм которого является авторской разработкой. Основной результат графематического анализатора состоит в выделении структурных единиц текста (от символа до абзаца) с приписанными им значениями графических признаков. Эти единицы, снабженные соответствующей информацией, используются на последующих этапах работы процессора. Например, такая единица, как токен, далее подвергается морфологической атрибуции. Синтаксический анализ опирается на выделенные ранее части предложения (ограниченные нетерминальными знаками препинания) и предложение. Для построения информационной структуры текста важны как связи между предложениями внутри абзаца, так и межабзацные связи.

После графематического анализа наступает этап атрибуции графических токенов, который в случае словоформ соответствует тому, что традиционно называется морфологическим анализом, хотя в действительности морфоанализаторы предлагают лишь возможные варианты морфологических характеристик словоформы.

Описываемый нами процессор способен обрабатывать более широкий круг токенов, регулярно встречающихся в тексте, а именно аббревиатуры, сокращения и прочие комплексы, которые наряду с буквенными символами могут содержать знаки иной природы. В основе собственно морфологического анализа лежит словарь словоформ, созданный в результате предварительных исследований, связанных с автоматической обработкой текстов, но существенно пополненный в ходе настоящей работы.

В связи с этим нам представляется, что было бы уместно называть этот этап морфографической атрибуцией. Во-первых, на вход поступают различные цепочки символов, в том числе и такие, кото-

рые содержат цифры, знаки препинания и другие символы; таким образом, дело не ограничивается собственно языковым материалом. Во-вторых, даже там, где процессор имеет дело с однородной последовательностью кириллических символов, он фактически производит не анализ, а механическое сопоставление этой последовательности (как некой целостной единицы) с массивом хранящихся в памяти цепочек символов (также целостных единиц) и, если находит совпадения, выдает их все в качестве результата. Это процедура сравнения, а не анализа.

Центральное место в описываемой модели принадлежит синтаксису. Все содержание глав 3 и 4, охватывающее формально-синтаксический анализ, семантическое описание лексики, снятие лексической и синтаксической неоднозначности, а также правила преобразования семантико-синтаксической структуры в информационную модель и сам проект последней, представляет собой исключительно авторские оригинальные разработки.

Принципиальный аспект синтаксического анализа состоит в том, что выявление структуры предложения осуществляется не через построение древесных структур «сверху вниз» (ср. англ. *top-down*), а путем последовательной циклической сборки «снизу вверх» (ср. англ. *bottom-up*), в соответствии с линейным порядком слов. Эта идеология, с одной стороны, отличает наш процессор от различных формальных моделей анализа естественного языка, а с другой — сближает его с современными когнитивными исследованиями, направленными на изучение процессов понимания дискурса.

Закономерно встает вопрос, что следует рассматривать в качестве *предложения*, если сугубо формальный (графический) подход не всегда дает удовлетворительные результаты. Во-первых, существуют неполные предложения, интерпретация которых требует обращения к предшествующему предложению (ср.: *Всегда он наталкивался на твердое сопротивление Лены, и с годами идея стала являться все реже. И* ***то лишь в моменты раздражения*** (Ю. Трифонов. Обмен)). Во-вторых, в случае прямой речи имеет смысл учитывать в качестве единого содержательного блока весь ее отрезок, который может состоять из сколь угодно большого числа графически оформленных предложений. Следует подчеркнуть, что в нашей модели предусмотрены возможности обрабатывать оба указанных случая (подробнее см. гл. 1), а также анализировать непроективные предложения (ср.: *Послушай: далеко, далеко на озере Чад* ***изысканный бродит жираф*** (Н. Гумилев. Жираф)).

Вообще, в работе по созданию лингвистического процессора всплывает целый ряд «вечных» теоретических проблем языкознания, таких как отдельность и тождество слова (ср. обработку сложных союзов, составных предлогов, наречий, числительных, сказуемых и пр.), композициональность семантики (проблема коллокаций и идиом), лексическая многозначность и способы ее описания и пр. В итоге ключевые термины лингвистического анализа — *слово* и *предложение* — оказываются довольно условными, допускающими вариативность в интерпретации. В своем изложении мы, стараясь сохранять ясность, не употребляем термин *слово* в главе о графематическом анализе, используем термин *словоформа* в морфологическом контексте и оперируем понятием семантико-синтаксический вариант (ССВ) слова там, где речь идет о лексической семантике (термина *лексема* ввиду его неоднозначности также приходится избегать). С термином *предложение* сложнее — заменить его нечем, поэтому мы здесь и позволили себе данное отступление.

Синтаксический анализ производится с учетом семантической информации, источником которой служит специально созданный семантико-синтаксический словарь, направленный на многоаспектное описание лексики в терминах семантических признаков, каждый из которых может иметь два и более значений (в итоге получается нечто наподобие семантических формул в лингвистическом процессоре В. А. Тузова [2003]). В этом заключается существенное отличие нашей модели от других семантических представлений, будь то электронные онтологии или словари (семантические, идеографические), составители которых приписывают фактически одному семантическому признаку многочисленные значения, обычно организованные либо в виде иерархического дерева (например, (Шведова 1998–2007)[3]), либо в виде общего списка отдельных морфологических, синтаксических, лексических свойств, характеризующих слово целиком или то его значение, которое, по мнению разработчиков, является наиболее частотным (например, [Апресян и др. 1992]).

Основное назначение семантического описания — способствовать снятию лексико-грамматической омонимии или, по крайней мере, сокращению числа вариантов, полученных в результате синтаксического анализа. Прежде всего это касается однозначности структурной схемы предложения. Так, с формальной точки зрения в предложении *Автомобиль сделал поворот* оба существительных

[3] Ссылки на словари и справочники даются в круглых скобках.

(*автомобиль* и *поворот*) могут быть как подлежащим, так и прямым дополнением. Благодаря семантическому описанию слов, составляющих данное предложение (*автомобиль* — транспортное средство, *сделать* — глагол действия, *поворот* — отглагольное существительное со значением движения), удается отсечь неверный вариант синтаксического разбора. Однако каким бы точным ни было семантическое описание, оно не всегда позволяет получить единственную структуру там, где предложение допускает множественную синтаксическую интерпретацию, ср. знаменитый пример *Мать любит дочь*, а также изречение *Бытие определяет сознание* (неоднозначность которого, впрочем, обычно не обращает на себя внимание). Синтаксическая омонимия, действительно, не такое уж редкое явление (ср. [Мельчук 1974: 31–32; Муравенко 2008: 159–162]).

Семантическое описание позволяет также снижать уровень лексической неоднозначности, а в идеале и вовсе ее снимать. Когда синтаксически связанные слова неоднозначны, алгоритм осуществляет последовательный перебор всех соответствующих комбинаторных вариантов с целью сопоставления приписанных им значений семантических признаков. В основе данной операции лежит известная идея о том, что такие слова не должны иметь противоречащих друг другу сем.

Рассмотрим в качестве примера предложение *Рабочий отстоял смену*. Допустим, что грамматическая омонимия словоформы *рабочий* уже снята, а в качестве существительного она имеет единственное значение. Остается разобраться со словами *отстоять* и *смена*, каждое из которых характеризуется довольно высоким индексом неоднозначности. Согласно словарю (БТС 2009), среди значений глагола *отстоять*[4] есть такие, которые имеют временной параметр: одно предполагает целенаправленную деятельность человека (*Отстоять два часа в очереди*), а другое — темпорально ограниченное существование предмета (*Храм отстоял пять веков*). А существительное *смена*, наряду с обозначением лица (или их совокупности), предмета (одежды) и выражением абстрактного понятия изменения, может использоваться для обозначения промежутка времени. Таким образом, имеет место согласование по семе времени между одним значением существительного и двумя значениями глагола. Число комбинаторных вариантов интерпретации, таким образом,

[4] Точнее, речь идет о значениях соответствующих омонимичных глаголов.

сократилось до двух. С учетом семантического типа подлежащего *рабочий* неоднозначность удается полностью снять.

Несмотря на большой объем используемого словаря, окказиональные употребления слов в расширительном или образном смысле могут, разумеется, приводить к такому «семантическому рассогласованию» [Гак 1972: 381] внутри отдельного словосочетания, что ни одна из комбинаций значений семантических признаков не дает удовлетворительной интерпретации. Подобная ситуация является важным сигналом, свидетельствующим о необходимости переосмысления одного из членов словосочетания; какого именно — решается на основе анализа последующего текста.

У семантического описания есть и другая важная функция, которая обращена не вспять (к синтаксическому анализу отдельного предложения), а, напротив, вперед — к интерпретации последовательности предложений, вплоть до целого текста. Дело в том, что комбинации значений семантических признаков у лексических единиц, связанных между собой синтаксическими отношениями, обусловливают категориальный тип описываемой ситуации. Например, сочетание кванторного выражения со значением всеобщности с глаголом ментального восприятия (типа *Каждый американец думает, что…*, *Все считают, что…*) обозначает вовсе не «массовую» интеллектуальную деятельность (что получается из буквального сложения смыслов), а наличие (в какой-то социальной группе) некоторой установки, нормы, верования и т. п., при отсутствии конкретного субъекта — носителя соответствующих представлений[5]. Когда за подобным предложением следует что-нибудь вроде: *Эта идея нашла поддержку / не выдерживает критики / является верной / полезной / тлетворной* / и т. п., — именно семантическое описание, основанное на корреляции семантических признаков и их значений, способно обеспечить когеренцию текста. В противном случае результат автоматического анализа будет напоминать известную поговорку «В огороде бузина, а в Киеве дядька».

Семантика, таким образом, перебрасывает мостик к следующему этапу — коммуникативно-синтаксическому анализу, распространяющему свое действие на более крупные, чем предложение, отрезки текста. Его организация связана с семантикой еще теснее, чем синтаксис и нижние уровни языковой структуры [Откупщикова 1982:

[5] Субъект в этих случаях имеет универсальный референциальный статус [Падучева 1985: 95–96].

21], что является существенным осложняющим фактором. Ведь из традиционных компонентов лингвистического процессора — морфологического, синтаксического и семантического — «достаточно высокой, если не исчерпывающей полноты можно добиться лишь в первых двух компонентах» [Апресян и др. 1992: 113].

Заключительный этап анализа носит отчетливо выраженный поисковый характер и потому описан в книге более схематично, чем предыдущие этапы. Большое внимание уделено принципам перехода от полученной на предыдущем этапе синтаксической структуры в виде структуры «подлежащее — сказуемое — дополнения — обстоятельства» (ПСДО) к субъектно-объектной метаязыковой информационной модели. Подробно рассматриваются различные варианты соответствия между компонентами синтаксической и информационной структур и проблемы, возникающие при их соотнесении.

При описании концепции информационной структуры, в которую преобразуется неструктурированный текст на ЕЯ, неизбежно затрагивается широкий круг задач, подлежащих решению на этом этапе (определение объектов и связей между ними, экспликация неявных компонентов смысла, развертывание пропозиций и выявление отношений между ними, определение кореференции именных групп, интерпретация дейксиса, установление хронологической последовательности событий и др.). В качестве материала привлекаются тексты разных стилей и жанров, что призвано наглядно продемонстрировать невозможность выработки единого подхода к указанным проблемам. Заметим, что имеющаяся литература весьма скупа по части их возможного решения. Более того, отсутствуют даже сколь бы то ни было развернутые описания различных типов текстов в соответствующих аспектах. За исключением, возможно, некоторых жанров научного дискурса, а также языка официальных документов, лингвистам очень мало известно о том, как устроены те или иные типы текстов, каковы их лексические и грамматические особенности, что собой представляют временные характеристики содержания, какие языковые средства используются для воплощения соответствующей темпоральной структуры и т. д.

Полезность информационной структуры для решения широкого круга прикладных задач не вызывает сомнений. На ее основе могут быть построены базы знаний, позволяющие хранить сведения (факты, утверждения), относящиеся как к уникальным объектам, имеющим собственное имя (*Петров, Москва, Украина*) или дескрипцию (*президент Молдавии; принадлежащий А. А. Сидорову автомобиль;*

памятник Минину и Пожарскому; день рождения А. В. Суворова), так и к классам объектов (*самолет Ту-134, дизельное топливо, курс валют, биржевые сделки*). Пополнение базы знаний должно происходить автоматически по мере анализа текстовых массивов, получаемых из Интернета или локальной сети. В результате может быть получена информационно-аналитическая система, способная производить мониторинг соответствующих текстов, интегрировать данные из разных документов, осуществлять семантический поиск по объектам, событиям, тематическую рубрикацию и пр.

Помимо этого, возможно применение модели для проверки текста на однозначность — это важно, в частности, для задач, связанных с лингвистической экспертизой.