Глава 1
ГРАФЕМАТИЧЕСКИЙ АНАЛИЗ

Автоматическая обработка текста начинается с предварительного этапа композиционного анализа, направленного на выделение абзацев и определение их типов — заголовков, подзаголовков, имен авторов, названий глав и разделов, сносок, примечаний, приложений, эпиграфов и т. д. Происходит также выделение в составе текста схем, таблиц, рисунков с подрисуночными надписями (ср.: [Леонтьева 2006: 49–50]).

Далее следует собственно графематический анализ, посредством которого осуществляется токенизация (англ. *tokenization*), — разбиение потока алфавитных и внеалфавитных графем на цепочки символов, такие как отдельные слова, аббревиатуры, буквенные сокращения, цифровые последовательности, комплексы смешанной природы (формулы и пр.). Используемый нами подход кардинальным образом отличается от тех, что обычно практикуются в существующих системах автоматической обработки языка[1].

Во-первых, наш графематический анализатор, как и лингвистический процессор в целом, не использует формальные модели и аппарат математической статистики (ср. [Уразлин 2005; Седунов 2007; Ле 2011]). В его основе лежит процедура последовательной сборки все более крупных сегментов текста (символы → атомы → токены → части предложения → предложение), предполагающая тщательный структурный анализ контекста и опирающаяся на разработанные авторами оригинальные правила интерпретации. Работа алгоритма построена на учете сугубо формальных свойств символов.

Во-вторых, для нас очевидна недостаточность простой «нарезки» текста по пробелам и знакам препинания, при которой точка, вопросительный и восклицательный знаки (так называемые терминальные знаки препинания) сигнализируют конец предложения, а пробел, запятая и др. — конец токена. При таком механическом подходе значительная часть информации может быть уте-

[1] См. также более раннюю публикацию [Клементьева, Скребцова, Суворов 2013].

ряна в силу неоднозначности знаков препинания, их способности выполнять разные функции (ср. [Безвербный 2000; Ровинская 2000; Кобзарева 2005]). Так, пробел и точка могут встречаться внутри чисел, больших тысячи, ср. *24 756* (= *24.756 = 24756*) или *31 245 678* (= *31.245.678 = 31245678*). В Великобритании и США, кстати, для этой цели обычно используется запятая, а в России запятая служит для отделения дробной части от целой в десятичных дробях (ср. *1,3 кг., 2,5 м.*)[2]. Помимо этого, точка используется в буквенных сокращениях (мер, весов и др.), которые могут встречаться в любом месте предложения, за исключением инициальной позиции. Точка и двоеточие применяются в обозначениях точного времени (*7.15* или *7:15*). Точка также встречается в обозначениях дат (*24.12.12*). Приведенные примеры, конечно, не исчерпывают все случаи, в которых знаки препинания не являются разделительным сигналом, но их достаточно, чтобы понять сложность задачи и неэффективность тривиальных решений.

Следует заметить, что формальный конец предложения не всегда знаменует законченность синтаксической структуры. В конструкциях с парцелляцией терминальные знаки препинания, по сути дела, выполняют функцию запятой, ср.:

> *Я требую амнистии. Я требую, чтобы она была полной и всесторонней. Без оговорок. Без ограничений* (В. Гюго. Речь об амнистии в Сенате).

> *Ребятам так и сказали: хотите, мол, нажраться, поезжайте. Там все есть. И хлеб там есть. И картошка. И даже фрукты, о существовании которых наши шакалы и не подозревают* (А. Приставкин. Ночевала тучка золотая).

> *Люда Голливуд мелькнула в кабинет начальника. Неслыханное дело: Чинков! Просит приема! У Фурдецкого!* (О. Куваев. Территория).

> *— Я? Вам? Дал телефон? Что за ерунда! — не понимая, сказал Никитин. — Никогда! Я вам не давал никакого телефона!* (Ю. Бондарев. Берег)[3].

---

[2] Все примеры здесь и далее приводятся в оригинальной орфографии.

[3] Это взятые из учебника, хрестоматийные примеры. Между тем парцелляция встречается не столь редко, как это может показаться на первый взгляд, и не только в художественной литературе. Следующий фрагмент заимствован из учебного пособия по стилистике русского языка (!), где он, как видно из текста, не используется в качестве иллюстрации, ср.: «В задачу этой книги не вхо-

При автоматической обработке текста соответствующие сегменты, несмотря на их пунктуационную самостоятельность, должны анализироваться вместе, как части одного предложения — только в этом случае они смогут получить адекватную интерпретацию. Сигналом к объединению членов парцеллированной конструкции служит выявляемая в ходе синтаксического анализа неполнота соответствующих парцеллятов. Их присоединение к базовой части конструкции происходит на этапе коммуникативного синтаксиса.

Графематический анализатор опирается на словарь символов, который создан авторами на основе ASCII-кодировки, т. е. включает буквы (русского и латинского алфавита, заглавные и строчные), знаки препинания, пробел, арабские цифры, дефис и некоторые другие часто используемые знаки компьютерной клавиатуры (знаки процента, доллара США, номера и т. д.). Использование символов ASCII-кода связано с тем, что разрабатываемый лингвистический процессор ориентирован на анализ текстов на русском языке, для которых в подавляющем большинстве случаев этой кодировки оказывается достаточно. В принципе, алгоритм может работать и с Unicode, правда, это потребует переделки словаря символов.

Важным преимуществом разработанного графематического анализатора является то, что он привязан к внешнему виду символа, а не к числовым кодам в компьютере. Это сохраняет возможность интерпретации при небрежном наборе текста или случайной опечатке (например, если вместо русского согласного *с* была напечатана английская буква *c*), а также в тех случаях, когда текст был отсканирован и распознан с ошибками.

Рассмотрим сказанное на примерах. Допустим, имеется последовательность из трех символов *CAT*. Каждый из них получит из словаря символов следующие варианты интерпретации[4]:

дит анализ разговорной речи, мы остановимся на трех функциональных стилях речи "книжной" — в первой части, и речи публичной, ораторской — во второй. На том, что в их современном состоянии есть "хорошего" или, наоборот, "плохого". Как они пришли к их современному состоянию, как развивались, как "становились на ноги" и куда "идут" теперь. Какими языковыми средствами мы можем пользоваться, а каких лучше избегать, чтобы не выглядеть смешным в глазах читателя или слушателя — обо всем этом мы поговорим на страницах нашей книги» [Дунаевская 2010: 8].

[4] Точнее, соответствующие значения по графическим признакам «тип символа», «алфавит», «регистр», «тип буквы».

С — латинский заглавный согласный[5], русский заглавный согласный, обозначение температуры по Цельсию, римская цифра;

А — латинский заглавный гласный, русский заглавный гласный;

Т — латинский заглавный согласный, русский заглавный согласный.

Работа алгоритма основана на соединении смежных символов одного типа, поэтому на выходе будет получено два варианта: последовательность из русских букв *САТ* и последовательность из латинских букв *CAT*, — а интерпретации, связанные с обозначениями температуры и римской цифры, будут отброшены.

Другой пример: слово *сахарный*. При последовательном анализе символов слева направо неоднозначность, связанная с одинаковым написанием русских и латинских букв, сохраняется вплоть до шестой позиции. При обработке символа *н* обнаруживается, что в латинском алфавите отсутствует буква с подобным начертанием, следовательно, данная цепочка символов распознается как принадлежащая русскому алфавиту, и графематическая неоднозначность снимается.

Фундаментальный принцип работы графематического анализатора (и описываемого лингвистического процессора в целом) состоит в поэтапном объединении знаков во все более крупные сегменты. Сначала происходит соединение смежных символов одного типа в однородные последовательности, которые мы называем атомами (таким образом, рассмотренные выше последовательности из русских букв *САТ* и *сахарный*, а также последовательность из латинских букв *CAT* являются атомами). Формирование однородных последовательностей базируется на учете таких признаков символов, как «алфавит» (значения ‘кириллица’ / ‘латиница’), «регистр» (значения ‘заглавный’ / ‘строчный’) и др. Соответствующую информацию анализатор получает из словаря символов, так что, по сути дела, работа идет не с символом как таковым, а со значениями его признаков. Подчеркнем снова, что это важная особенность лингвистического процессора, характеризующая процедуру обработки текста на разных этапах.

Буквенные атомы всегда однородны с точки зрения алфавита. Они могут представлять собой последовательность строчных букв, последовательность заглавных букв, последовательность строчных

[5] Более корректно: латинская заглавная буква, обозначающая согласный звук. См. также далее.

букв, открываемую заглавной буквой. При этом последовательность заглавных букв может означать либо аббревиатуру, либо шрифтовое выделение. Последовательность из начальной заглавной буквы и последующих строчных — это либо имя собственное, либо первое слово в предложении. Аббревиатуры, включающие строчные буквы (*СПбГУ, ЦПКиО*), представляются в виде нескольких атомов (в данных примерах — трех); их объединение происходит при следующих «проходах» графематического анализатора. Аналогично обстоит дело с графическим выделением отдельных букв в слове (ср. *рассы́пать* vs. *рассыпа́ть*) и использованием апострофа для обозначения ударного гласного *(рассы́пать*). Дефис обрабатывается как отдельный атом, поэтому слова типа *кое-где, кто-то* и т. п. на начальном этапе также предстают в «разорванном» виде, как и сокращения типа *г-жа, чл.-корр., д-р*.

Другие примеры атомов включают, например, последовательность арабских цифр, последовательность пробелов, несколько одинаковых знаков препинания подряд (многоточие и др.). Все смешанные цепочки символов выступают в виде последовательности однородных атомов, ср.: |*24*|-|*ое*|; |*5*| |*кг*||; |*ВУЗ*|*ы*|; |*И*|.|*И*|.|*Иванов*|; |*Он*| |*сказал*|:| |«|*Пошли*| |*домой*|»|. Как видно из примеров, атом может быть образован одним-единственным символом. Такая высокая степень детализации анализа необходима для дальнейшей правильной интерпретации последовательностей атомов.

Подобно тому, как символам в соответствующем словаре приписаны значения признаков, атомы также получают свои значения признаков на основе правил интерпретации. Например, атом *ВУЗ* будет охарактеризован как последовательность из русских заглавных букв, гласных и согласных вперемешку, с указанием типов символов, ограничивающих его с обеих сторон (это могут быть знаки препинания, пробелы, дефис; справа к данному атому может также примыкать строчная буква кириллицы). Также будет указано место данного атома в предложении (начальное / среднее / конечное).

Затем из атомов формируются токены. В общем случае токен — это сегмент, который выступает в качестве цельного блока (единицы) на дальнейших этапах анализа. Сборка токена нередко осуществляется в несколько «проходов», например в случае комплексов из фамилии с инициалами (*А. С. Пушкин, А. Конан-Дойль*), аббревиатур из заглавных и строчных букв (*СПбГУ*), графических выделений (*звони́т*), смешанных комплексов из цифр и знаков препинания (дат, обозначений точного времени, номеров телефонов и др.), циф-

ровых цепочек с буквенными сокращениями (*70 км., 20 млн. долл.*) или окончаниями (*12-ти, 5-му*), слов с дефисом и др. Токен может быть равен атому — ср. имена нарицательные, не имеющие в своем составе дефиса, или аббревиатуры, состоящие исключительно из заглавных букв.

Сборка токена осуществляется в соответствии с правилами интерпретации, предполагающими, в частности, его сравнение с набором шаблонов, таких как Ф.И.О. (с различными вариантами наполнения и расположения фамилии и инициалов, а также возможными сокращениями типа *г-н, тов.* в препозиции), аббревиатуры смешанного типа (состоящие из заглавных и строчных букв), пишущиеся через дефис слова, даты, цифровые последовательности (*25,2; 46 792*), цифровые последовательности с буквенными сокращениями (*45 мин.*) или окончаниями (*5-ая, 5-я, 5-ти, 125я, 125ти*), простые арифметические выражения (*2+2, 450:15 = 30*) и некоторые другие. В итоге токену приписывается значение графического признака «тип токена» (например, ‘словоформа’, ‘аббревиатура из заглавных букв кириллицы’, ‘дата’ и пр.), а также значения таких признаков, как «место в предложении» (‘начальное’, ‘среднее’ или ‘конечное’), «тип ограничивающего слева символа» и «тип ограничивающего справа символа».

Покажем на некоторых примерах, как определяется тип токена. Так, состоящий исключительно из заглавных согласных токен (он же атом) *СССР* по признаку «тип токена» получит значение ‘аббревиатура из заглавных букв латиницы или кириллицы’; дальнейший поиск в словаре аббревиатур осуществляется на этапе атрибуции токенов — см. ниже. А вот цепочка символов *МСХ*, помимо аббревиатуры латиницей или кириллицей, может обозначать записанное римскими цифрами число 1110. Графематический анализ не имеет никаких данных в пользу какого-либо из этих трех вариантов, и снятие данной неоднозначности будет происходить позднее — в ходе синтаксического анализа.

Токен ВУЗ, включающий как гласные, так и согласные, в соответствии с правилами интерпретации может оказаться как аббревиатурой, так и шрифтовым выделением. Но если в постпозиции к цепочке *ВУЗ* примыкает строчная буква, то она однозначно распознается как падежная форма аббревиатуры (ср. *ВУЗы, ВУЗе* и др.). Таким токенам, как *СПбГУ* или *УрГПУ*, приписывается значение ‘аббревиатура’ за счет вкрапления строчной буквы в последовательность прописных. Цифры, разделенные знаками препинания, автоматически

объединяются в один токен, который затем сопоставляется с шаблонами дат и точного времени, фиксирующими наиболее частотные варианты написания.

Токены типа *где-то* идентифицируются в качестве словоформ, так как атомы слева и справа от дефиса представляют собой последовательности из русских гласных и согласных вперемешку. В отличие от них, символьные последовательности *г-жа, г-н, д-ру* и т. п. характеризуются как 'буквенное сокращение'. Вообще, присоединение буквенных сокращений к соседним атомам с целью образования токена может осуществляться двумя способами: к предыдущему или последующему атому. Согласно правилам интерпретации, если буквенное сокращение стоит после цепочки цифр, но перед последовательностью строчных букв ('некапитализированной словоформой'), оно присоединяется к цифрам (ср. *20 кг яблок*). Если же перед сокращением нет числа, а за ним идет последовательность строчных букв, открываемая заглавной буквой ('капитализированная словоформа'), то сокращение относится к ней, ср. *г. Москва, тов. Иванов*. Данные правила охватывают наиболее частые случаи, однако не все последовательности атомов могут быть однозначно интерпретированы в рамках графематического анализа. Например, в цепочке *2012 г. Москва* буквенное сокращение *г.* может обозначать год и относиться к числу *2012*, а может обозначать город и примыкать к слову *Москва*.

Следует подчеркнуть, что приписывание токену его типа носит предварительный характер. Это гипотеза, которая проверяется на следующем этапе анализа, а именно при атрибутировании токенов. Именно там, в частности, происходит поиск буквенных цепочек в словарях (словаре общих имен, словаре имен собственных, словарях сокращений и аббревиатур). Так, токен *Токарев*, стоящий в середине предложения, с самого начала будет правильно охарактеризован как имя собственное, так что проверка этой гипотезы даст положительный результат. Но если тот же токен расположен в инициальной позиции, первоначально ему будет приписан тип 'словоформа', и только процедура атрибуции позволит по итогам поиска в соответствующих словарях исправить его на 'фамилия'.

Проверке подвергаются также и цифровые цепочки, и смешанные последовательности. К примеру, токены типа *24.12.2012* предположительно являются датами, но, чтобы подтвердить эту гипотезу, необходимо проверить составляющие его атомы с точки зрения соответствующих ограничений. Аналогичной процедуре подвергаются обозначения точного времени. Последовательностям заглавных

латинских букв (типа *XVIII*) при формировании токенов приписывается тип ‘число’, но далее осуществляется проверка их правильности с точки зрения допустимых символов и их сочетаний.

Как уже говорилось, помимо типа, для каждого токена, как и для атома, заполняются значения признаков, связанных с местом в предложении и типом ограничивающих символов. В отличие от атомов, токены не могут быть ограничены ни буквами, ни дефисами, ни цифрами — только знаками препинания и пробелами.

На следующем этапе работы графематического анализатора происходит сборка групп токенов, ограниченных знаками препинания. При этом следующие за знаком препинания пробелы удаляются, поскольку они являются сугубо конвенциональными и не несут функциональной нагрузки. Собранные таким образом группы токенов называются «частями предложения» (графическими, а не синтаксическими). Часть предложения может совпадать с одиночным токеном, ср. одиночное вводное слово или член однородного ряда и др.

В соответствии с общим принципом работы анализатора части предложения получают описания в терминах занимаемого места и ограничивающих символов. Так, пушкинское предложение *Мороз и солнце; день чудесный!* состоит из двух частей: первая стоит в начале предложения, ограничена слева началом текста и точкой с запятой справа, а вторая стоит в конце предложения и (после удаления конвенционального пробела) ограничена точкой с запятой слева и восклицательным знаком справа. Часть предложения также может быть охарактеризована с точки зрения ее типа: например, ‘скобочная структура’, ‘прямая речь’ и др.

Наконец, части предложения объединяются в предложение. В общем случае эта процедура основана на поиске терминального символа (точки, вопросительного или восклицательного знака), но следует упомянуть о двух исключениях из этого правила. Во-первых, как указывалось выше, точка может быть составной частью цифровых шаблонов и буквенных сокращений, следовательно, наличие точки не означает автоматически конец предложения. Во-вторых, терминальные знаки препинания нередко встречаются внутри прямой речи, при этом бывает целесообразно рассматривать в качестве предложения всю структуру, включающую как авторскую речь, так и речь персонажа, ср.:

> *«Да проститься же надо было!..» — понял он, когда крытая машина взбиралась уже на взвоз* (В. Шукшин. Осенью).

«Вдруг посеешь, — думал Семен, — а вырастет обыкновенный ячмень. Скорее всего, так и получится. Но попробовать надо. И, главное, поискать еще. Неужели в целом поле он был один?» (В. Солоухин. Счастливый колос)*;*

— Превосходно, Уотсон, — воскликнул Холмс, — мне кажется, говорю вам это искренне, что вы недалеки от истины. Вы видите сами, все карты у нас в руках, и теперь нам надо спешить, пока не случилось непоправимое. Если время позволит, мы захватим их непременно (А. Конан-Дойль. Случай с переводчиком).

В подобных случаях отдельные графические предложения внутри прямой речи рассматриваются нами в качестве специальной единицы, которую мы назвали подпредложением. И подпредложениям, и предложениям приписываются значения, связанные с их позицией в более крупной единице (соответственно предложении и абзаце) и типами ограничивающих символов.

Самым крупным сегментом графематического анализа является абзац, который, однако, иногда может быть равен предложению или даже еще меньшим сегментам. Расстановка символа «конец абзаца» происходит в рамках первичного композиционного анализа (см. выше). Однако следует заметить, что не всегда знак конца абзаца означает смысловую законченность: достаточно указать на списочные структуры после двоеточия, где каждый новый пункт (нумерованный или определенным образом маркированный) располагается с новой строки и начинается со строчной буквы. Фактически мы имеем здесь одно предложение, графически разбитое на несколько абзацев. (Можно усмотреть некоторую аналогию с членами парцеллированной конструкции, отделенными друг от друга терминальными знаками препинания.)

* * *

Основным источником проблем, возникающих при графематическом анализе текста, является неоднозначность знаков препинания, зависимость их функции от контекста. В этой связи следует прежде всего упомянуть кавычки, которые могут использоваться в разных целях, а именно: для выделения цитат, отдельных слов (устарелых, непривычных, употребляемых иронически, терминов и пр.), условных наименований (подробнее о семантике кавычек см. [Зализняк 2007]). Наши правила интерпретации охватывают следующие ситуации:

- Если в кавычки заключен одиночный токен, по признаку «тип» ему приписывается значение ‘название’ (если он написан с заглавной буквы и не расположен в начале предложения, ср. *издательство «Наука»*, *соус «Тартар»*) или ‘термин’[6] (если он написан со строчной буквы и не расположен в начале предложения, ср.: *Так и появилась болезнь под названием «арахнофобия»*). Написанный с заглавной буквы одиночный токен в кавычках в инициальной позиции может быть как тем, так и другим;
- Если новый абзац открывается группой токенов в кавычках (первый токен с заглавной буквы), за которой следует запятая или тире, а также если подобная группа токенов в кавычках расположена после двоеточия, она представляет собой прямую речь;
- если группа токенов в кавычках не занимает начальной позиции в предложении, не предваряется двоеточием и начинается со строчной буквы, она атрибутируется в качестве цитаты. Например: *Говоря о поэзии Пушкина, Н. А. Добролюбов писал, что «в его стихах впервые сказалась нам живая русская речь, впервые открылся нам действительный русский мир»*.

За пределами данных правил оказываются многословные наименования, например названия литературных произведений: «*Кому на Руси жить хорошо?*», «*Повесть о настоящем человеке*», фильмов: «*Москва слезам не верит*», «*Ирония судьбы, или С легким паром*», телевизионных игр: «*Кто хочет стать миллионером?*», «*Что? Где? Когда?*», газет и журналов: «*Аргументы и факты*», «*Русский язык за рубежом*», «*Знание — сила*», а также лозунги: «*Вся власть советам!*», «*Слава КПСС!*» и др. Подобным группам токенов графематический анализатор не может приписать тип; их атрибутирование происходит при синтаксическом анализе.

Графематический анализ предложений, содержащих кавычки, может быть дополнительно осложнен тем обстоятельством, что кавычки, как известно, бывают разного рисунка. В том числе есть и такой, в котором открывающие кавычки по внешнему виду совпадают с закрывающими. В ситуации, которую условно можно обозначить следующим образом: …."……"……"……", — возможны следующие две интерпретации:

---

[6] Данное обозначение следует понимать в расширенном смысле, а именно: термин или слово, к появлению которого в тексте автор в силу тех или иных причин хочет привлечь внимание (это может быть окказионализм, экзотизм, устаревшее слово, шутливое или ироническое наименование и т. д.).

- вложенные одна в другую структуры, ср.: *Пушкин писал Дельвигу: "Жду "Цыганов" и тотчас тисну"*;
- два фрагмента, следующие друг за другом, ср. *Читали ли вы статью в "Известиях" "Куда мы идем?".*

Автоматически распознать эти варианты в рамках графематического анализа невозможно.

Трудный случай представляет собой также использование в тексте непарных кавычек, ср. *ЗАО «Издательский дом «Комсомольская правда»*, *ОАО «НПП «Буревестник»*, *работа В. И. Ленина «О карикатуре на марксизм и об "империалистическом экономизме"»*.

Многозначность свойственна и другим знакам препинания. Особенно высок индекс неоднозначности у дефиса, который теоретически может трактоваться и как буква, и как знак препинания [Крылов, Старостин 2003: 354]. Кроме случаев его использования в составе слова и в качестве знака переноса, дефис может быть сигналом сокращенного написания слов (как простых, так и сложных, ср. *г-н, д-р, 150-летие, 5-го*), применяться при записи конструкций с однородными членами, имеющими общий конечный элемент (например, *теле- и видеофильмы; не двух-, а трехэтажный дом; как водо-, так и газоснабжение),* употребляться в выразительных целях, быть соединительным символом для составных существительных (типа *девочки-подростки*), числовых шаблонов (таких как даты, номера телефонов) и др. К тому же в компьютерной записи дефис (или два дефиса подряд) иногда применяется для обозначения тире. Графематический анализатор должен уметь распознавать все это многочисленные случаи и действовать сообразно: приводить написание слова к полной форме, строить слово по аналогии с соответствующим однородным членом, объединять последовательность знаков в единый смысловой блок и т. д.

Алгоритм стремится к максимально полному учету контекста любого символа, с тем чтобы снимать неоднозначность или, по крайней мере, ее сокращать. Это достигается за счет многоступенчатой процедуры анализа контекста на основе разработанных правил интерпретации. Помимо сокращения числа ошибочных интерпретаций, данная процедура позволяет осуществлять глубокий структурный анализ текста, в том числе распознавание вложенных структур, которые могут быть оформлены такими парными знаками препинания, как кавычки, скобки, тире и даже дефисы. В целом можно сказать, что мы стремимся «выжать максимум» из сугубо формального анализа на графематическом уровне.