Глава 2
АТРИБУЦИЯ ТОКЕНОВ

МОРФОЛОГИЧЕСКИЙ АНАЛИЗ БУКВЕННЫХ ТОКЕНОВ

Цель данного этапа работы процессора заключается в том, чтобы проверить сделанные на этапе графематического анализа гипотезы о типе токенов, а также приписать им значения релевантных морфологических признаков. Если токеном является словоформа русского языка, речь идет о процедуре, которую принято называть морфологическим анализом. Но, поскольку в текстах встречаются и другие виды токенов (см. выше), мы пользуемся более общим термином «атрибуция», подчеркивающим приоритетную задачу данного этапа — определение типа для всех токенов[1]. Решение данной задачи является необходимой предпосылкой успешного синтаксического анализа.

Рассмотрим процедуру атрибуции для токенов различной природы. Наиболее частотными в естественно-языковых текстах, несомненно, являются буквенные токены. Если не принимать во внимание регистр, буквенный токен может быть самостоятельной словоформой, аббревиатурой или частью устойчивого словосочетания. Соответственно, у нас имеются три различных словаря, в которых можно искать текущий токен. Заметим, что аббревиатуры также могут состоять из нескольких частей (ср. *МВД РФ*), поэтому есть смысл рассматривать подобные составные аббревиатуры также в качестве устойчивых сочетаний, игнорируя регистр. Тогда для всех буквенных токенов процедура атрибуции начинается одинаково: с обращения к словарю устойчивых неизменяемых словосочетаний[2].

Под устойчивым словосочетанием мы понимаем фиксированную последовательность словоформ или аббревиатур, выражающую еди-

---

[1] В некоторых системах автоматической обработки языка выделяется специальный этап постморфологии, посвященный обработке имен собственных, аббревиатур, сокращений, цифровых последовательностей и др.

[2] Так, независимо от того, как в тексте выглядит текущий токен — *МГУ, Мгу* или *мгу,* — в словаре устойчивых словосочетаний все равно ищется некапитализированная буквенная цепочка *мгу*, с целью дальнейшего выяснения, не является ли данный токен начальным компонентом устойчивого словосочетания.

ное понятие и функционирующую в качестве цельного блока[3]. Под это формальное определение подпадают такие содержательно разные вещи, как неизменяемые фразеологические единицы, вводные обороты, сложные союзы, составные наречия, местоимения, предлоги и частицы, а также аббревиатуры из двух и более компонентов (например, *ГУ МВД РФ*). Естественно, данный список не включает составные словоформы, образующиеся по правилам (аналитические формы сравнительной и превосходной степени наречий и прилагательных, аналитические формы глаголов, составные числительные), в силу того что они представляют собой открытый список изменяемых форм. На случай вариативности структуры устойчивого словосочетания (например, *не иначе/не иначе как*) в словаре предусмотрены все варианты его реализации.

Итак, алгоритм проверяет, не является ли текущий буквенный токен частью устойчивого словосочетания. Если найдено устойчивое словосочетание, начинающееся с данного токена, далее осуществляется сравнение остальной части этого словосочетания с токенами, которые в тексте непосредственно следуют за данным. (Поскольку процедура атрибуции последовательно осуществляется слева направо, от токена в инициальной позиции вплоть до токена в финальной позиции, достаточно всякий раз анализировать только правый контекст.) Когда устойчивое сочетание содержит более чем две словоформы/аббревиатуры, сопоставление требует нескольких шагов. Если последовательность словоформ найдена в словаре, ей приписывается соответствующее значение по грамматическому признаку «часть речи». Например:

на босу ногу — нар
в том числе — союз

Если найденное устойчивое сочетание представляет собой последовательность аббревиатур, словарь сопоставляет ей следующую информацию: морфологические характеристики составной аббревиатуры, определенные на основе анализа ее функционирования в тексте, лемму опорного слова всего сочетания и позицию ударного гласного в данной лемме. Морфологические характеристики аббревиатуры могут не совпадать с соответствующими характеристиками опорного слова. При вариативности морфологических характеристик приводятся все варианты, ср.:

[3] Ср. понятие безусловного оборота в работе [Апресян и др. 1992: 34].

гу мвд рф	сущ; неод; ср[4]	управление	3
мвд рф	сущ; неод; мр	министерство	3
мвд рф	сущ; неод; ср	министерство	3

Если рассматриваемый токен совокупно со своим правым контекстом не найден в словаре устойчивых словосочетаний, подключается предварительная (сделанная на этапе графематического анализа) гипотеза о его типе. Одиночный буквенный токен, в зависимости от своего состава, может быть либо словоформой (ср. *Мгу, мгу*), либо аббревиатурой (*МГУ*). Рассмотрим тот случай, когда токен предположительно является словоформой (имени нарицательного или собственного).

Для таких токенов процедура атрибуции продолжается поиском словоформ в словаре, который, наряду с нарицательными существительными, содержит частотные имена, фамилии, отчества, географические названия, наименования организаций и пр. в полном наборе их словоизменительных вариантов. В основу словаря словоформ первоначально был положен находящийся в открытом доступе морфологический словарь проекта АОТ[5]. В дальнейшем, однако, он был дополнен материалами других словарей и существенно переработан как в лингвистической, так и в алгоритмической части. Можно сказать, что из проекта АОТ заимствована главным образом сама идеология компактного хранения и быстрого поиска.

В словаре словоформ регистр игнорируется, и все заголовочные слова пишутся строчными буквами. Это обусловлено тем обстоятельством, что регистр не всегда позволяет однозначно определить, является ли текущий токен именем собственным или нарицательным. Так, слово в инициальной позиции в предложении (или в прямой речи) пишется с заглавной буквы независимо от того, собственное это имя или нарицательное, ср.: *мга* (‘туман, мгла’) и *Мга* (река и населенный пункт в Ленинградской области). В подобных случаях словарь словоформ выдает оба варианта:

мга 1 сущ; лок[6]; неод; жр, им; ед
мга 1 сущ; неод; жр, им; ед

[4] Неодушевленное существительное среднего рода. Ниже, соответственно, неодушевленное существительное мужского рода.

[5] URL: http://www.aot.ru (дата обращения: 04.04.2018).

[6] Локатив (населенный пункт, река, гора и др.).

Примеры подобной неоднозначности не так уж редки, ср. *пашу́* — *Пашу, мила́* — *Мила, вера* — *Вера, юля́* (деепр. от *юлить*) — *Юля, вена* — *Вена* и т. д.

Итак, словарь сопоставляет каждой словоформе ее лемму, порядковый номер ударного слога, частеречный разряд и грамматические характеристики (отдельно по словоизменительным и классифицирующим морфологическим категориям), а именам собственным дополнительно приписывает условный маркер (‘локатив’, ‘фамилия’, ‘имя’ и др.). Несклоняемые географические названия получают характеристику ‘неизм’, ср.:

тбилиси 2 сущ; неизм; лок; неод; ед; мр

Их род определяется по роду нарицательного слова, для которого имя собственное служит наименованием (в нашем примере — по слову *город*). Для географических названий, оканчивающихся на *-о* и *-е* и не имеющих в словаре указания на тип населенного пункта, указывается средний род (ср. *украинское Ровно, но сибирский Кемерово*). Кроме того, для многих географических названий на *-о* (типа *Останкино*) предусмотрено два варианта — изменяемое и неизменяемое существительное.

Неизменяемые фамилии получают описание, соответствующее их способности относиться к лицу мужского или женского рода, ср.:

петренко 2 сущ; неизм; фам; неод; мр/жр

Для подавляющего большинства словоформ словарь выдает несколько вариантов морфологического разбора. Например, для известного примера *стекло* будут получены следующие три варианта атрибуции:

стекло 2 сущ; неод; ср им; ед
стекло 2 сущ; неод; ср им; вин
стечь 2 гл; нп[7]; соверш прош; ср

Множественность полученных результатов, разумеется, усложняет дальнейший (синтаксический) анализ. Иногда число вариантов удается сократить, опираясь на значения графических признаков соответствующих токенов. Так, если токен занимает неинициальную позицию в предложении и написан с заглавной буквы, из списка вариантов выбирается тот, который является словоформой имени

[7] Непереходный глагол.

собственного (в примере со словом *мга* — первая строчка). Если результаты графематического анализа говорят о том, что токен имеет в своем составе одну неначальную капитализированную гласную, имеет смысл сопоставить порядковый номер соответствующего слога в токене с заданной в словаре позицией ударения (ср. *хло́пок* или *хлопо́к*): совпадение номеров подтвердит мысль о том, что таким образом помечен ударный слог, и позволит убрать лишние варианты атрибуции. Аналогичная процедура применяется, когда для указания ударного гласного в буквенных токенах используется апостроф (ср. *хло́пок* или *хлопо́к*).

Таким образом, если первоначально (при обращении к словарю словоформ) графематическая информация о текущем буквенном токене игнорируется (поскольку в словаре все заголовочные слова записаны строчными буквами и апострофы внутри токена удалены), то позднее, когда мы уже имеем результаты атрибуции, эта информация начинает работать для исключения заведомо неверных вариантов.

Обратимся теперь к рассмотрению работы алгоритма в случае, когда не найденный в словаре устойчивых словосочетаний токен предположительно является аббревиатурой. Происходит обращение к специальному словарю аббревиатур и буквенных сокращений (о последних см. ниже). Заголовочные единицы в нем записываются с соблюдением регистра, так как в ряде случаев он значим, ср. сокращения *Вт.* (ватт) и *вт.* (вторник). Если имеются различные способы записи одной и той же аббревиатуры (сокращения), в словаре приводятся все варианты, ср. *ВУЗ* и *вуз*.

В словаре аббревиатурам приписаны их морфологические характеристики, «вычисленные» с опорой на корпус текстов. В случае вариативности рода приводятся все варианты. Словарь также сопоставляет каждой аббревиатуре ее опорное (стержневое) слово и позицию ударного гласного в нем. Последнее необходимо для того, чтобы на этапе синтаксического анализа (см. ниже) получить семантические характеристики для данного токена по соответствующему словарю. Морфологические характеристики аббревиатуры и опорного слова зачастую не совпадают, ср.:

НИИ	сущ; неод; мр	институт	3
НИИ	сущ; неод; ср	институт	3
МИД	сущ; неод; мр	министерство	3
БАМ	сущ; неод; мр	магистраль	3

В некоторых случаях приписывание опорного слова осуществляется искусственно. Это касается, в частности, заимствованных из других языков звуковых аббревиатур: так, для *НАТО, ЮНЕСКО, ФИФА* опорным словом будет *организация*, а для *Би-Би-Си — радиостанция*. Для таких аббревиатур, как *СНиП* (строительные нормы и правила) делается выбор в пользу одного из членов однородного ряда.

Если искомого токена в словаре аббревиатур нет, предварительная гипотеза отменяется и производится его поиск в словаре словоформ. Ведь капитализация возможна не только в случае аббревиатур, но и текстовых выделений, а также в заголовках.

Различного рода слоговые аббревиатуры (они же сложносокращенные слова) типа *Минкульт, партком, Роспотребнадзор, телеинтервью, запчасть, завкафедрой* обрабатываются как обычные словоформы. Сначала происходит обращение к словарю словоформ (наиболее частотные из них туда занесены). Если сложносокращенное слово там не найдено, к работе подключается механизм морфологического предсказания (см. ниже).

Особо следует сказать о таких токенах, как *VIP-персона, SMS-сообщение, DVD-диск*: первая часть, оформленная латиницей, отбрасывается, и атрибуция токена производится по второй части, которая ищется в словаре словоформ.

АЛГОРИТМ МОРФОЛОГИЧЕСКОГО ПРЕДСКАЗАНИЯ

Рассмотрим теперь тот случай, когда буквенный токен не обнаружен ни в одном из трех упомянутых словарей, т. е. является так называемой несловарной словоформой [Ляшевская 2007; Клышинский 2007; Черненьков 2010]. В литературе описаны различные подходы к их морфологическому анализу (см., например, [Ермаков, Плешко 2004; Сокирко 2010]).

В нашем лингвистическом процессоре применяются следующие эвристические приемы определения грамматических характеристик неопознанных буквенных токенов. Если токен содержит дефис, прежде всего делается попытка удалить дефис (а вдруг это знак переноса?) и искать соответствующую цепочку символов без дефиса в словаре словоформ.

Если поиск не дал результата, алгоритм ищет части токена (до и после дефиса) по отдельности. Здесь возможны разные ситуации.

Если найдена только вторая часть, морфологическая атрибуция всего сложносоставного слова[8] производится по ней, ср. *итало-австрийский, физико-технический, социал-демократ*. Так же обстоит дело, если в словаре содержатся обе части сложносоставного слова, причем вторая часть является прилагательным (например, *светло-голубой, один-единственный*). Проблем не возникает и в случае такой редупликации, при которой компоненты полностью совпадают, ср. *еле-еле, чуть-чуть, белый-белый*.

Трудности появляются, когда в словаре мы находим только первую часть сложносоставного слова, ср. *рад-радехонек, старый-престарый, мало-помалу*. Казалось бы, можно осуществлять атрибуцию по этой части, однако в общем случае это неверно: например, в словах типа *белым-бело* это приведет к тому, что вместо неизменяемого наречия данное слово будет атрибутировано как прилагательное единственного числа, мужского или среднего рода в творительном падеже. Как поступать в такой ситуации, пока неясно.

Нами была подробно рассмотрена более частная проблема — определение морфологических характеристик сложносоставного существительного, обе части которого найдены в словаре[9]. Понятно, что когда оба компонента относятся к одному грамматическому роду (*вагон-ресторан, бал-маскарад, генерал-майор, девочка-украинка, старуха-процентщица*), род сложносоставного существительного будет таким же. Что касается словоизменительных характеристик, то их разумно определять по форме второго компонента, поскольку первый может быть неизменяемым или не склоняться в составе данного сложносоставного слова (например, *кофе-машины, генерал-майора,* но *вагона-ресторана, дому-музею*).

---

[8] Под сложносоставными словами мы, вслед за орфографическим справочником (Соловьев 1997: 740–741), понимаем сложные слова, пишущиеся через дефис; соответственно, собственно сложные слова — это те, которые имеют слитное написание. Иная точка зрения представлена в (Русская грамматика 1980: т. 1: 253), где разграничение между сложными и сложносоставными словами проводится исходя из особенностей их склонения: сложные слова содержат несклоняемый первый компонент, в то время как сложносоставные представляют собой словосочетания, в которых склоняются обе части. В то же время сами авторы цитируемого источника признают, что граница между этими двумя типами слов не вполне жесткая (Русская грамматика 1980: т. 1: 253). В связи с этим опора на материальный признак представляется нам более обоснованной и, вдобавок к тому, более удобной для изложения, поскольку мы рассматриваем здесь исключительно слова с дефисным написанием.

[9] См. [Скребцова 2014].

Задача перестает быть тривиальной, когда компоненты сложносоставного существительного имеют разный грамматический род. Род определяет согласование существительного с определением и сказуемым (когда такое существительное выступает в роли подлежащего), а также выбор анафорического местоимения (*он, она, оно, который (-ая, -ое)* и др.). Очевидно, что определение этой грамматической характеристики важно для последующих этапов работы лингвистического процессора, связанных с анализом предложения и связного текста. Разумеется, можно приписывать сложносоставным существительным типа *диван-кровать, шкаф-купе, бизнес-сообщество* по два грамматических значения рода, но это повлечет за собой существенный рост синтаксической неоднозначности.

Проблема определения рода сложносоставных существительных не получила широкого освещения в литературе. Авторы теоретических статей обычно апеллируют к понятию приложения, что в контексте автоматического анализа неуместно. В статье [Большаков, Большакова 2012] предложен примитивный практический подход, согласно которому определять род и число сложносоставного существительного нужно исходя из соответствующих характеристик первого компонента. Его несостоятельность легко показать на многочисленных примерах.

В пособии Д. И. Розенталя сделан упор на понятие ведущего слова. Автор, в частности, пишет: «…ведущим (определяющим) является то слово, которое выражает более широкое понятие или конкретно обозначает предмет. <…> Как показывают примеры, обычно в этих случаях на первом месте стоит ведущее слово, с которым и согласуется сказуемое. Если же на первом месте стоит не ведущее слово, то при изменении падежной формы подобных сочетаний первая часть не изменяется» [Розенталь 1998: 239]. В качестве иллюстраций соответственно приводятся существительные *девушка-агроном* vs. *плащ-палатка*: ср. *девушки-агронома*, но *плащ-палатки*.

В целом соглашаясь с приведенными суждениями, заметим, однако, что даже человеку (а тем более компьютеру) бывает нелегко определить ведущее слово. Так, в существительных *хлеб-соль, братья-сестры, руки-ноги* компоненты семантически равноправны (объемы соответствующих понятий не пересекаются) и в одинаковой степени конкретны. В существительных типа *бизнес-сообщество, рок-группа, интернет-телевидение*, построенных по заимствованной из английского языка модели, тоже не наблюдается отмеченных смысловых отношений между компонентами. В самом деле, какое

понятие более широкое — *рок* (как музыкальное направление) или *группа* (как коллектив)? Какое слово обозначает конкретный предмет — *Интернет* или *телевидение*? К тому же сам автор цитированного пособия признает отдельные «колебания в согласовании глагола-сказуемого со сложносоставными существительными, в которых одна часть по функции напоминает приложение, ср. *кафе-столовая открыто/открыта*» [Розенталь 1998: 238–239].

Довольно замысловатые правила определения рода сложносоставного слова содержатся в работе [Волынец 2003: 11]:

1) если составное существительное обозначает лицо, то его род определяется по слову, указывающему на пол лица — *женщина-скульптор создала…, чудо-богатырь поскакал…*;
2) если составное существительное обозначает неодушевленные предметы, то род определяется по роду первого слова — *музей-квартира открыт…, школа-интернат построена…, кресло-кровать стояло в углу…*;
3) если в сложносоставном слове одно существительное несклоняемое, то род определяется по роду склоняемого слова — *пресс-конференция прошла успешно, кафе-столовая открыта…*.

Однако использование этих правил в автоматическом морфоанализе наталкивается на препятствия. Во-первых, информация о том, обозначает ли то или иное существительное лицо, указывает ли оно на пол лица, представляет собой семантические характеристики, которые, как правило, приписываются словоформам на более поздних этапах анализа. В нашем словаре словоформы характеризуются только признаком морфологической одушевленности/неодушевленности, которого в данном случае недостаточно. Таким образом, пункт 1) не несет полезной для нас информации.

Во-вторых, в процитированном выше методическом пособии ничего не говорится об обозначениях живых существ типа *рыба-меч, бабочка-адмирал* — похоже, что у них грамматический род определяется в соответствии с пунктом 2), т. е. так же, как у существительных, обозначающих неодушевленные предметы. Получается, что указанный выше признак морфологической одушевленности/неодушевленности, которым характеризуются словоформы в нашем словаре, опять нерелевантен.

Наконец, пункт 3) при ближайшем рассмотрении оказывается неточным (чтобы не сказать — неверным), так как есть много

слов, у которых первая часть сама по себе в отдельности склоняется, но в сложносоставном существительном становится неизменяемой, ср. *бизнес-школа, рок-группа, плащ-палатка, горе-чиновник, чудо-сад*. Формально они относятся к пункту 2), и их грамматический род должен определяться по роду первого компонента, но в действительности дело обстоит ровно наоборот: род сложносоставного слова совпадает с родом второго компонента. Значит, по сути, следовало бы причислять их к третьему типу. Но тогда возникает задача формального определения класса существительных, которые утрачивают способность склоняться, когда выступают в качестве первого компонента сложносоставных существительных. Непонятно, как она может быть решена[10].

Рассуждая о понятии морфосинтаксического локуса — элементе словосочетания, в котором морфологически выражается его связь с более широким контекстом, — Я. Г. Тестелец констатирует, что синтаксической вершиной в аппозитивных конструкциях может быть как первый элемент сочетания (*Мы увидели славный город-герой*), так и второй (*В путь-дорогу дальнюю я тебя отправлю*); более того, оба существительных могут оказаться морфосинтаксическим локусом для разных грамматических явлений, ср. *Черная птица-шофер на лету отвинтил правое переднее колесо* (М. А. Булгаков. Мастер и Маргарита): определение согласуется с одним существительным, а сказуемое — с другим [Тестелец 2001: 84–85].

Подводя итог затянувшимся размышлениям, приходится констатировать, что в настоящий момент мы не видим способа в полном объеме решить задачу вычисления грамматического рода (а следовательно, и словоизменительных характеристик) для всех сложносоставных существительных. В ряде случаев приходится допускать возможность альтернативного рода, что в итоге ведет к увеличению общего числа вариантов морфоанализа.

---

[10] Любопытно, что ситуация, в которой компоненты сложносоставного существительного имеют разное грамматическое значение числа, разрешается гораздо проще. Если эта разница происходит из-за того, что один из компонентов представляет собой существительное pluralia tantum, род составного слова (и его словоизменительные характеристики) определяются по другому компоненту, ср. *юбка-брюки, ясли-сад, часы-будильник*. Различие в числе также возможно тогда, когда первый компонент сложносоставного слова не склоняется, а второй стоит в форме множественного числа, ср. *чудо-йогурты, пресс-службы, генерал-майоры* — тогда, очевидно, род и другие морфологические характеристики сложносоставного существительного вычисляются по второму (изменяемому) компоненту.

Схожие проблемы возникают при попытке определения одушевленности/неодушевленности некоторых сложносоставных существительных, ср. *город-герой, страна-поставщик, бизнес-партнер* и т. д. Заметим, что одна и та же модель построения (неодушевленное существительное + одушевленное существительное) дает в приведенных примерах разный результат: первые два сложносоставных существительных являются неодушевленными, а последнее — одушевленным. Как и в случае с морфологическим родом, правильное определение значения данной грамматической категории важно для дальнейших этапов автоматического анализа; между тем литературы по данному вопросу практически нет (подробнее см. [Скребцова 2015б]).

Обратимся теперь к ситуации, когда в составе несловарной словоформы нет дефиса или, несмотря на описанные выше приемы, в словаре не удалось ничего найти. В этом случае к работе подключается алгоритм морфологического предсказания. Он опирается на списки псевдокорней (ср. *гран, границ, гранич, гранк, гранул* и т. п.), «концовок» (ср. *-зала, -жегши, -обраны, -етому, -нувшихся, -атизм, -итель, -чивый* и т. п.), а также приставок (в том числе двойных) и префиксоидов, созданные нами на базе «Толково-словообразовательного словаря» (Ефремова 2000) и проекта АОТ. Разработанный алгоритм пытается, двигаясь от конца словоформы, последовательно отделять возможные «концовки» и осуществлять поиск остатка в списке псевдокорней; аналогично от начала словоформы он пытается отсечь потенциальные префиксы и также искать подходящий корень. Если эта процедура дает результат, для словоформы восстанавливается предположительная лемма и определяются вероятные грамматические характеристики.

Разумеется, далеко не для всех неопознанных словоформ данный процесс заканчивается успешно. Во-первых, попытки отделения аффиксов могут оказаться неэффективными, в том смысле, что ни с начала, ни с конца словоформы не будет найдено подходящих фрагментов. Возможно, что некоторые грамматические сведения о такой словоформе впоследствии даст синтаксический анализ. Например, в следующей последовательности из пяти слов:

[прил], [прил] [*и*] [неопозн] [сущ] —

неопознанная словоформа не может быть ничем иным, кроме как однородным определением к существительному. (Кстати говоря, синтаксический анализ позволяет «задним числом» не только

приписать морфологические характеристики неопознанным словоформам, но и уточнить их для уже атрибутированных токенов, например определить падеж для аббревиатур, число для сокращений, тип числительного для записанных цифрами чисел.) Хотя такие информативные контексты в целом редки, синтаксический анализатор в любом случае стремится прикрепить неопознанный токен к какой-нибудь конструкции.

Во-вторых, не так уж редки случаи ложной омонимии флексий (подробно рассмотренные в [Мищенко, Калугина 2002]), приводящие к неверному морфемному членению. С этой проблемой можно было бы справиться, разработав полноценный морфологический анализатор, осуществляющий морфемное членение словоформ, однако на сегодняшний день он не готов (подробнее см. ниже параграф «Атрибуция других типов токенов»).

АТРИБУЦИЯ ДРУГИХ ТИПОВ ТОКЕНОВ

Обратимся теперь к процедуре атрибуции для токенов, которые, наряду с буквами, включают цифры и/или различные знаки. Начнем с буквенных сокращений (типа *кг, км, др., г., г-жа*). Как уже указывалось, они входят в словарь аббревиатур и сокращений. Если встречаются разные способы графического оформления сокращения (*р., руб.; млн., млн;, ж.д., ж.-д.*), в словаре приводятся все варианты. Для каждого сокращения приводится его «расшифровка» (не всегда совпадающая с леммой), порядковый номер ударного слога (для однословных «расшифровок»), а также (по возможности) частеречная характеристика и морфологические характеристики, ср.:

кг	килограмм	3	сущ; но; мр
др.	другие	2	прил; мн
пп.	пункты	1	сущ; но; мр; мн
т. е.	то есть		союз
т. п.	тому подобное		
ж.-д.	железнодорожный	5	прил
с.-х.	сельскохозяйственный	4	прил
ув.	уважаемый	3	прил
тов.	товарищ	2	сущ

Недостающая морфологическая информация может быть восполнена на этапе синтаксического анализа. Составные сокращения (типа *млн. долл., тыс. км*), подобно составным аббревиатурам, вхо-

дят в словарь устойчивых словосочетаний. Аналогично мы поступаем с аббревиатурами типа *км/ч, кг/кв.м.*

Кроме того, для сокращений введен специальный признак, призванный охарактеризовать то, к чему синтаксически примыкает то или иное сокращение. Возможные значения данного признака следующие: ‘к числу’, ‘к имени’ (с частным вариантом ‘к имени собственному’) и ‘самостоятельное сокращение’ (оно же — ‘концевое’, например *т. д., т. п., и др.*).

Это помогает снимать или по крайней мере сокращать неоднозначность. Так, *г.* может обозначать год, грамм, город, гору и господина. Благодаря указанному признаку, удается развести значения ‘год’ и ‘грамм’, с одной стороны, и ‘город’, ‘гора’, ‘господин’ — с другой. Далее, если перед сокращением стоит записанное арабскими цифрами четырехзначное число, скорее всего, оно обозначает год (для одно-, двух- и трехзначных арабских чисел, обозначающих год, после г. нередко идет еще одно сокращение — *н. э.* или *до н. э.*). Для цепочек типа *г. Москва*, если *Москва* найдена в словаре словоформ и опознана как локатив, остаются варианты интерпретации ‘город’ и ‘гора’. Что же касается, к примеру, цепочки *г. Пушкин*, то она может быть как географическим названием (ср. пригород Санкт-Петербурга с таким названием), так и обозначением лица.

Атрибуция подобных токенов (*2012 г., г. Москва, г. Эльбрус*) предполагает сравнение с так называемыми шаблонами. Шаблоны — это цепочки символов смешанной природы, которые могут включать буквы, цифры и другие знаки. Они были выделены для обработки часто встречающихся последовательностей символов, таких как даты, обозначения точного времени, буквенные сокращения с именами собственными в постпозиции или цифрами в препозиции, аббревиатуры или последовательности цифр с буквенными окончаниями, интернет-адреса, обозначения веков латинскими цифрами, арифметические выражения и др.

К примеру, учитываются разнообразные способы записи дат:

дд.мм.гг; дд.мм.гггг; дд/мм/гг; дд/мм/гггг; дд-мм-гг; дд-мм-гггг;
мм.дд.гг; мм.дд.гггг; мм/дд/гг; мм/дд/гггг; мм-дд-гг; мм-дд-гггг (американские способы записи дат).

При этом алгоритм проверяет соответствующие атомы с точки зрения ограничений на число месяцев в году и дней в месяце.

Существуют шаблоны, охватывающие различные способы записи имен людей, включающие фамилию с инициалами, а именно:

Фамилия И. О. (с пробелом между инициалами и без)
Фамилия И.
И. О. Фамилия (с пробелом между инициалами и между инициалом и фамилией или без)
И. Фамилия (с пробелом между инициалом и фамилией или без).

Процедура атрибуции производится применительно к слову, написанному полностью (фамилии), а начальные буквы имени и отчества отсекаются.

Необходимо заметить, что при полном написании компонентов Ф. И. О. любым из следующих способов:

Фамилия Имя Отчество
Имя Отчество Фамилия
Фамилия Имя
Имя Фамилия
Имя Отчество, —

сборка соответствующего комплекса осуществляется только на этапе синтаксического анализа. Напомним, что с графематической точки зрения каждая составляющая является отдельным токеном и предварительно характеризуется как имя собственное. На этапе атрибуции эти гипотезы проверяются по словарю словоформ. Аналогичным образом обстоит дело с составными наименованиями географических объектов (и не только), ср. *Нижний Новгород, Черное море, Петр Первый* (подробнее о различных сложностях, возникающих при морфоанализе имен собственных, в частности имен людей, см. [Кобзарева 2004]).

Обработка аббревиатуры с окончанием, будь то с дефисом или без (ср. *КВНы, КВН-ах*), предполагает вначале отсечение строчных букв и поиск той части токена, что набрана заглавными буквами (*КВН*), в словаре аббревиатур. Если она найдена, мы получаем лемму опорного слова аббревиатуры (в данном случае — *клуб*) и ее грамматический род. Далее имеющееся окончание сравнивается с окончаниями существительных данного рода на предмет совпадения. В случае успеха мы получаем (дополнительно к сведениям, пришедшим из словаря аббревиатур) морфологические характеристики токена по словоизменительным категориям (числа и падежа).

Схожей проверке подвергаются токены, состоящие из записанного цифрами числа и буквенного окончания типа *10-е, 3-мя, 126-ти* — для них производится поиск соответствующего окончания (точнее, псевдоокончания) по всем парадигмам количественных

и порядковых числительных. В итоге нередко получается несколько вариантов интерпретации. Например, токен *20-е* атрибутируется как порядковое числительное единственного числа, среднего рода, в именительном либо винительном падежа. А такие токены, как, скажем, *31-й* и *31-м*, неоднозначны даже с точки зрения того, количественные это числительные или порядковые, и индекс омонимии у них еще выше.

Если в тексте встречается цифровая запись числа без буквенного окончания, она может обозначать как количественное, так и порядковое числительное. До проведения синтаксического анализа невозможно квалифицировать сочетания типа *2 группы, 137 км* как выражения, обозначающие количество предметов или их порядок при счете. На этапе атрибуции токенов обработка цифровой записи предполагает отделение с правого конца цифр, выражающих числа от 0 до 20, и приписывание двух лемм, соответственно для количественного и порядкового числительных[11]. Например:

31 числ. колич. один
31 числ. порядк. первый
610 числ. колич. десять
610 числ. порядк. десятый
1113 числ. колич. тринадцать
1113 числ. порядк. тринадцатый

* * *

Возвращаясь к нашему опыту атрибуции буквенных токенов, следует заметить, что существуют разные методы проведения морфологического анализа. Предлагаемый подход, опирающийся на использование словаря словоформ, в настоящее время является наиболее распространенным, так как существенно экономит усилия лингвистов и разработчиков программного обеспечения. Такое «декларативное» обеспечение морфологического анализа, однако, имеет известный недостаток, а именно неспособность справляться со словами, которых нет в словаре (именами собственными, неологизмами, разнообразными дериватами, намеренно или случайно искаженными словами). Частично эта проблема может быть решена

[11] Это решение обусловлено дальнейшей задачей связывания токенов в ходе конструктивно-синтаксического анализа. Соответствие между членами количественных групп справедливо квалифицируется Л Л. Иомдиным [1990: 62] как одно из самых сложных и запутанных в русском синтаксисе.

при помощи механизма морфологического предсказания, однако он не всегда оказывается эффективным.

В качестве альтернативы мы рассматривали возможность разработки процедурного морфологического анализа, направленного на выявление морфемного состава словоформ. Разработка такого анализатора потребовала бы колоссальной подготовительной работы, но зато дала бы возможность обрабатывать слова, не идентифицированные по словарю. Эта идея казалась заманчивой нам еще и потому, что она была бы созвучна процедуре анализа текста на других этапах. Поясним сказанное.

Мы уже видели, что при графематическом анализе алгоритм работает по принципу постепенного, многоступенчатого наращивания единиц слева направо: от символов к атомам, от атомов к токенам, с последующей сборкой частей предложения, предложения и абзаца. Процедура опирается на словарь элементарных сегментов (символов) и правила интерпретации, разработанные для каждого шага процедуры связывания соседних сегментов между собой. Та же идеология лежит в основе конструктивно-синтаксического анализа, о котором речь пойдет ниже. С этой точки зрения было бы логично, если бы морфологический анализ строился на основе словаря морфем с приписанными им значениями признаков, обусловливающими возможность или невозможность соединения морфем друг с другом. Это способствовало бы более успешной обработке новых слов и дериватов вне зависимости от степени их закрепленности в языке, позволило бы делать предсказания относительно их семантики. Кроме того, мы могли бы существенно сжать словарь словоформ, в частности за счет многочисленных глагольных приставочных образований. Но, пожалуй, основная привлекательность этого подхода лежит в плоскости методологии: выбор такого способа морфологического анализа позволил бы говорить об универсальном алгоритме обработки языка на разных этапах — от графематики до синтаксиса.

В то же время очевидно, что создание морфемного анализатора с нуля — это весьма трудоемкое мероприятие, по сложности не уступающее синтаксическому анализу. Потребность в системном описании морфем русского языка, учитывающем их формальные варианты (алломорфы), значения и комбинаторные свойства была осознана более трех десятилетий тому назад [Милославский 1980], но само описание до сих пор не разработано. Имеющиеся морфемные и словообразовательные словари (ср. (Тихонов 1985; Кузнецова, Ефремова 1986; Ефремова 2000)) не решают эту комплексную задачу

в полном объеме. Но даже если существовал бы толковый словарь морфем, учитывающий не только присущие им значения, но и возникающие в различных окружениях семантические «приращения», автоматический разбор многоморфемных слов также не был бы простой задачей. Здесь, как и на других этапах анализа, осложняющим фактором является много-многозначное соответствие между формой и значением.

Что же касается идеи об универсальном алгоритме обработки языка на разных этапах, она, по-видимому, все равно не смогла бы быть воплощена. Ведь в языке, помимо словоформ, есть токены другой природы, требующие совсем иных подходов к интерпретации. Если говорить не о традиционном морфоанализе, а в целом об атрибуции всех токенов в тексте, данная процедура в принципе не может быть универсальной и последовательной — это обусловлено спецификой текстов на ЕЯ.