Метод структурных схем компьютерного морфологического анализа словоформ естественного языка

Е. Е. ЕГОРОВА
Национальный исследовательский университет «Высшая школа экономики»

А. М. ЛАВРЕНТЬЕВ
Национальный центр научных исследований Франции, Лионский университет, Франция

А. М. ЧЕПОВСКИЙ
Национальный исследовательский университет «Высшая школа экономики»
e-mail: achepovskiy@hse.ru

УДК 519.76+81'32

Ключевые слова: алгоритм выделения основ, алгоритм морфологического разбора, анализ текстов на естественном языке, структурная схема слова.

Аннотация

В работе предлагается метод структурных схем в качестве модели морфологического анализа словоформ естественного языка с развитым аффиксальным словообразованием и словоизменением. Дано описание алгоритма выделения псевдоосновы, его модификация, а также алгоритм восстановления грамматических характеристик словоформ. Описано применение предложенного метода для анализа словоформ французского языка. Представлены результаты работы предложенных алгоритмов.

Abstract

E. E. Egorova, A. M. Lavrentiev, A. M. Chepovskiy, A structural pattern based method for automated morphological analysis of word forms in a natural language, Fundamentalnaya i prikladnaya matematika, vol. 19 (2014), no. 3, pp. 91—109.

In this paper, a computerized model for morphological analysis of languages with word-formation based on affixation processes is proposed. The main idea consists in defining structural patterns of words and corresponding lists of suffixes. First, a detailed description of a stemming algorithm, its modification, and the technique of determining grammatical characteristics of word-forms are given. The next part of this work focuses on the application of the proposed algorithms for the French language. Finally, some results of execution of these algorithms are provided.

1. Введение

Одной из основных задач сферы информационно-поисковых систем является повышение качества и полноты поиска при одновременной оптимизации размера

Фундаментальная и прикладная математика, 2014, том 19, № 3, с. 91—109.

индекса таких систем. Основным алгоритмом, который позволяет осуществлять индексацию и последующий поиск информации в информационно-поисковых системах, является алгоритм выделения псевдоосновы словоформ. В связи с этим возникает задача разработки программного обеспечения, реализующего модели и алгоритмы анализа слов естественных языков и работающего при этом в условиях частичного или полного отсутствия морфологических словарей.

На сегодняшний день существует множество работ, посвящённых задаче автоматического морфологического анализа словоформ естественного языка, их подробный обзор можно найти в [1]. В данной работе мы остановимся на построении алгоритма морфологического анализа словоформ для языков с развитым аффиксальным словообразованием и словоизменением. Словоформы таких языков имеют достаточно сложную морфологическую структуру, так как изменение и формирование новых слов происходит при помощи присоединения к основе последовательности аффиксов. Однако такая сложная структура слов никоим образом не усложняет анализ. Напротив, она позволяет более точно выделять основу словоформы и определять её грамматические характеристики, так как каждый аффикс несёт в себе определённое словообразовательное или грамматическое значение.

На данный момент для языков рассматриваемого типа чаще всего используются алгоритмы, основанные на простом отсечении аффиксов. Однако такой подход обладает некоторыми недостатками, один из которых заключается в том, что не учитываются словоизменительные и словообразовательные механизмы языка. Первой попыткой устранить данный недостаток была морфологическая модель, описанная в [2]. В данной работе мы формулируем обобщённую модель структурных схем и рассматриваем её применение к французскому языку. Основная идея этой модели заключается в построении структурных схем окончаний, которые описывают всевозможные варианты последовательностей из аффиксов в словоформах исследуемого языка. Такой подход позволяет учитывать грамматические особенности языка и способен обрабатывать новые слова или специфичную лексику и, как следствие, даёт более точные результаты при выделении псевдооснов. Кроме того, в работе затрагивается вопрос определения грамматических характеристик. За счёт выделения значимых частей окончания техника структурных схем позволяет восстанавливать грамматические характеристики слов. Наконец, на примере французского языка как языка с развитым аффиксальным словообразованием и словоизменением мы проиллюстрируем процесс построения необходимых структур для предлагаемого подхода к морфологическому анализу.

2. Метод структурных схем

Будем рассматривать языки с развитым аффиксальным словообразованием и словоизменением. В таких языках словоформы состоят из корней (или основ),

несущих базовое лексическое значение, и аффиксов, каждый из которых выражает одно или сразу несколько грамматических или словообразовательных значений. Мы будем использовать термины «корень» и «основа» как синонимы, поскольку нашей целью является не безупречный с точки зрения лингвистической теории морфемный анализ, а объединение близких по смыслу словоформ. Аффиксы делятся на несколько типов в зависимости от их позиции относительно корня слова. Префиксы (или приставки) располагаются перед корнем, а суффиксы — после него. В лингвистике принято различать словообразовательные и словоизменительные аффиксы. Первые позволяют модифицировать лексическое значение или грамматический класс (часть речи) слова, вторые выражают исключительно грамматические значения (например, падеж существительного или время глагола). В рамках настоящей работы данное различие не является существенным, так как нашей основной задачей является выделение корня.

Рассмотрим несколько случаев представления слов. Под основой будем понимать всю ту часть слова, которая не является окончанием или приставкой. Возможны следующие три случая представления слов: приставка+основа+окончание, основа+окончание, приставка+основа. Таким образом, получаем, что словоформа состоит из основы и определённого множества аффиксов, которое в некоторых случаях может оказаться пустым. Опишем теперь более формально слова языков, обладающих вышеупомянутыми особенностями, и способ изменения этих слов. Пусть задан алфавит $\Sigma = \{\sigma_1, \sigma_2, \ldots\}$, где σ_i — буква.

Определение 2.1. *Словом* будем называть конечную последовательность $\sigma = \sigma_{i_1}\sigma_{i_2}\ldots\sigma_{i_n}$, где $\sigma_{i_j} \in \Sigma$. Пустое слово также допускается, будем обозначать его ε.

Определение 2.2. Множество Σ^* — множество всех допустимых слов над алфавитом Σ.

Определим структуру словоформы, для этого введём несколько обозначений. Основу как тип обозначим через r, а множество слов такого типа обозначим через R. Также рассмотрим множество типов аффиксов A, отвечающих за формирование окончания. Пусть число различных типов будет m, тогда $A = \{a_1, a_2, \ldots, a_m\}$. Для $i = 1, \ldots, m$ каждому типу a_i поставим в соответствие множество A_i его представителей (слов). Тип префиксов обозначим через p, а множество его представителей через P. Важно отметить, что существуют слова, тип которых не определяется однозначно, т. е. $P \cap A_1 \cap A_2 \cap \ldots \cap A_m \neq \varnothing$. Через a_0 будем обозначать тип слов, которые не принадлежат множеству $P \cup A_1 \cup \ldots \cup A_m \cup R$.

Рассмотрим структуру некорневой части слова более подробно. Как уже было сказано, она может состоять из приставки и окончания, или только из окончания, или только из приставки. В условиях данной модели приставка выражается одним аффиксом или пустым словом в случае её отсутствия. Структура окончания более сложная. Каждое окончание представляет из себя слово (возможно, пустое). Непустое окончание можно представить как последователь-

ность подслов, каждое из этих подслов является аффиксом. Таким образом, для каждого слова можно получить последовательность из типов аффиксов, которая будет описывать его окончание и представлять из себя упорядоченный вектор. В общем случае такая последовательность выглядит следующим образом: $(a_{i_1}, a_{i_2}, \ldots, a_{i_s})$, $a_{i_j} \in A$. Для описания всей некорневой части слова необходимо учитывать префиксную часть, поэтому в начало этой последовательности при наличии приставки нужно добавить p. Получим последовательность $(p, a_{i_1}, a_{i_2}, \ldots, a_{i_s})$, $a_{i_j} \in A$. Учитывая особенности рассматриваемых нами языков, можно утверждать, что количество таких последовательностей конечно. Обозначим через Sch множество всех допустимых для данного языка последовательностей типов аффиксов, которые могут описывать некорневую часть слова. Через l обозначим максимальную длину последовательности, описывающей окончание, тогда в общем виде Sch можно записать следующим образом:

$$\begin{aligned}\mathrm{Sch} := &\{(a_{i_1}, a_{i_2}, \ldots, a_{i_s}) \mid a_{i_j} \in A,\ s = 0, \ldots, l\} \cup \\ &\cup \{(p, a_{i_1}, a_{i_2}, \ldots, a_{i_s}) \mid a_{i_j} \in A,\ s = 1, \ldots, l\} \cup \{(p)\}. \quad (2.1)\end{aligned}$$

Видно, что такое множество содержит последовательности, описывающие все возможные представления некорневой части слова: окончание, приставка+окончание или приставка. Элементы множества Sch будем называть *описаниями* или *структурами* некорневой части слова.

Теперь опишем функцию, которая будет составлять основу будущего алгоритма: функцию типизации последовательности подслов.

Определение 2.3. *Функцией типизации* назовём такую функцию F, что

$$F\colon (\gamma_1, \gamma_2, \ldots, \gamma_k) \mapsto \{(c_1, c_2, \ldots, c_k) \mid c_i \in A\} \cup \{p, a_0\},\ i = 1, \ldots, k\},$$

где для $j = 1, \ldots, k$ $\gamma_j \in \Sigma^*$, а c_j — один из возможных типов γ_j.

Иначе говоря, функция типизации каждой последовательности подслов ставит в соответствие множество возможных последовательностей из типов этих подслов. Множественность таких последовательностей обусловлена тем, что один и тот же аффикс может характеризоваться сразу несколькими типами. Учитывая введённые понятия, перейдём теперь к более формальному описанию некорневой части слова.

Утверждение 2.1. Слово $\sigma \in \Sigma^*$ является окончанием тогда и только тогда, когда его можно разбить на подслова $\sigma = \sigma_1\sigma_2 \ldots \sigma_k$, $\sigma_i \in \Sigma^*$, $i = 1, \ldots, k$, таким образом, что $F[(\sigma_1, \sigma_2, \ldots, \sigma_k)] \cap \mathrm{Sch} \neq \varnothing$.

Иначе говоря, то, что слово является окончанием, равносильно тому, что его можно разбить на подслова таким образом, что соответствующая этому разбиению последовательность типов является одной из допустимых, т. е. принадлежит множеству Sch (2.1). Аналогичные утверждения можно сформулировать и для случаев приставки и приставки+окончания.

Утверждение 2.2. Слово $\sigma \in \Sigma^*$ является приставкой тогда и только тогда, когда $F[(\sigma)] = \{p\}$.

Утверждение 2.3. Слово $\sigma \in \Sigma^*$ — конструкция «приставка+окончание» тогда и только тогда, когда его можно разбить на подслова $\sigma = \sigma_1\sigma_2 \ldots \sigma_k$, $\sigma_i \in \in \Sigma^*$, $i = 1, \ldots, k$, таким образом, что $F[\sigma_1] = \{p\}$ и $F[(\sigma_1, \sigma_2, \ldots, \sigma_k)] \cap \mathrm{Sch} \neq \varnothing$.

Заметим, что из-за неоднозначности разбиения слова и того факта, что некоторые слова относятся сразу к нескольким типам, некорневая часть может соответствовать сразу нескольким описаниям. В рамках введённых определений опишем понятие «словоформа».

Определение 2.4. *Словоформа* — слово σ над алфавитом Σ, представимое в виде

$$\sigma = \sigma^p \sigma^r \sigma^d,$$

где $\sigma^p \in P$, $\sigma^r \in R$ и σ^d — окончание.

Заметим ещё раз, что приставка и окончание могут быть пустыми символами, поэтому данное определение описывает все возможные варианты конструкций слов, допустимые в данной модели. Словоформа будет являться главным объектом исследования.

Как было сказано раньше, каждая словоформа представляется как последовательность подслов, каждое из которых относится к определённому типу. Так как нас интересует бессловарный анализ словоформ, то множеством R мы не обладаем, поэтому основная задача заключается в корректном выделении приставки и окончания словоформы, что позволит грамматически правильно определить основу. Алгоритм, который будет описан далее, получает на вход словоформу, а возвращает множество её допустимых некорневых частей.

Алгоритм 2.1.

Вход: словоформа $\sigma = \sigma_{i_1} \ldots \sigma_{i_l}$, $\sigma_{i_j} \in \Sigma$, $|\sigma| = l$.

Выход: множество T допустимых некорневых частей словоформы σ.

Шаг 1. $T := \{\varepsilon\}$.

Шаг 2. Получить путём выделения приставки и разбиения окончания множество M всех последовательностей подслов, таких что каждое подслово содержится в множестве $A_1 \cup \ldots \cup A_m \cup P$.

Шаг 3. К каждому такому разбиению $(\sigma_1^d, \sigma_2^d, \ldots, \sigma_k^d) \in M$, $\sigma_i^d \in \Sigma^*$, применить функцию F. $\mathrm{Set} := F[(\sigma_1^d, \sigma_2^d, \ldots, \sigma_k^d)]$.

Шаг 4. Если $\mathrm{Set} \cap \mathrm{Sch} \neq \varnothing$, то $T := T \cup \{\sigma^d = \sigma_1^d \sigma_2^d \ldots \sigma_k^d\}$.

Проще говоря, на первом шаге мы должны получить все возможные разбиения словоформы, т. е. разбиения, для которых каждая из их составляющих является аффиксом, допустимым в данном языке. Далее для них применяем функцию типизации, чтобы понять, как выглядит последовательность типов подслов из разбиения. На финальном шаге определяем, является ли эта последовательность возможной, опять же в рамках исследуемого языка. В итоге описанный алгоритм для заданной словоформы позволяет определить возможные её разбиения на приставку, основу и окончания, которые согласуются с грамматикой данного языка. Может оказаться, что по окончании работы алгоритма множество T будет состоять только из пустого слова ε, это будет означать, что вся

словоформа сама является основой и не содержит приставки или окончания. Для каждого исследуемого языка здесь стоит учитывать множество его особенностей: возможные длины окончаний, длины аффиксов, буквы, с которых начинаются аффиксы, и т. д.

3. Грамматические характеристики в методе структурных схем

Описанный выше алгоритм позволяет для каждой конкретной словоформы определить множество возможных представлений её некорневой части. Каждый элемент этого множества представляет собой одну из допустимых для данного языка последовательность подслов, которые в дальнейшем мы будем называть аффиксами. Каждый аффикс несёт в себе определённую информацию о грамматических, словообразовательных или словоизменительных характеристиках словоформы. Таким образом, анализируя каждый аффикс по отдельности и всю некорневую часть слова в целом, можно определить грамматические характеристики данной словоформы. Рассмотрим алгоритм, основывающийся на результатах алгоритма 2.1, который позволяет восстанавливать грамматические характеристики словоформ. Для начала нужно более четко определить, какие грамматические характеристики бывают, как они соотносятся друг с другом и чем различаются. Будем различать два типа грамматических характеристик: грамматические характеристики первого уровня и грамматические характеристики второго уровня

Определение 3.1. Грамматическая характеристика первого уровня — это характеристика, которой обладают все аффиксы и которая однозначно определяет множество других возможных характеристик данной словоформы.

Пример 3.1. Для примера можно рассмотреть русский язык, в нём характеристика «часть речи» относится к первому уровню, так как по каждому из аффиксов можно понять, какую именно часть речи он характеризует. Также каждая часть речи (глагол, существительное, прилагательное и т. д.) однозначно определяет множество других свойств, которые имеют смысл и которые можно определить для какой-то конкретной части речи. Так, для глагола, в отличие от существительного, можно определить время, наклонение и т. д. Для существительного можно определить падеж, но для глагола этого сделать нельзя.

Далее будем предполагать, что грамматических характеристик первого уровня в языке может быть сразу несколько и каждый аффикс будет обладать ими всеми. Дадим теперь определение грамматических характеристик второго уровня.

Определение 3.2. Грамматическая характеристика второго уровня — это характеристика, присущая аффиксу, но не обязательно каждому. Каждый аффикс

может обладать сразу несколькими различными грамматическими характеристиками второго уровня.

Пример 3.2. Например, в русском языке для грамматической характеристики первого уровня «часть речи» при значении «существительное» можно выделить следующие характеристики второго уровня: род, число, падеж и т. д.

Перейдём теперь к более формальному описанию введённых понятий. В данной модели каждая характеристика первого и второго уровня имеет несколько «представителей», т. е., к примеру, в русском языке для характеристики первого уровня «часть речи» представителями являются «глагол», «существительное», «прилагательное» и т. д., а для характеристики второго уровня «род» представителями будут «мужской», «женский», «средний». Таким образом, мы располагаем конечным числом характеристик и для каждой из них можем определить конечное множество её представителей.

Пусть для данного языка имеются N различных характеристик первого уровня, обозначим их $C_1, C_2, \ldots, C_N$. Каждой характеристике поставим в соответствие множество её представителей, т. е., например, характеристике C_i будет соответствовать множество F_i её представителей. Более формально: $F_i := \{f_{i1}, f_{i2}, \ldots, f_{ik}\}$, где f_{ij} — это j-й представитель i-й характеристики, а через k обозначено количество представителей данной характеристики. Для каждой характеристики первого уровня число представителей варьируется, но мы не будем описывать это явно, так как в этом нет необходимости. Для каждой характеристики будем обозначать число её представителей как мощность соответствующего множества, т. е. для характеристики C_i число представителей равняется $|F_i|$.

Из определения характеристики первого уровня следует, что про каждый аффикс можно сказать, какой представитель в каждой из характеристик его описывает. Все возможные комбинации из представителей характеристик могут быть описаны как декартово произведение множеств $F_1, \ldots, F_N$, т. е.

$$F_1 \times F_2 \times \ldots \times F_N = \{(f_1, f_2, \ldots, f_N) \mid f_i \in F_i\}.$$

Таким образом, каждому аффиксу мы можем поставить в соответствие вектор из множества $F_1 \times F_2 \times \ldots \times F_N$. В связи с этим дадим ещё одно определение.

Определение 3.3. *Характеристическим профилем* (или просто *профилем*) аффикса будем называть такой вектор из множества $F_1 \times F_2 \times \ldots \times F_N$, что аффикс характеризуется всеми входящими в этот вектор элементами (представителями характеристик).

Так как один аффикс может быть описан сразу несколькими представителями характеристик первого уровня, то получаем, что каждому аффиксу может соответствовать сразу несколько характеристических профилей.

Число всевозможных характеристических профилей будет равняться

$$n = |F_1| \cdot |F_2| \cdot \ldots \cdot |F_N|.$$

Все эти профили можем упорядочить и составить вектор v длины n так, что каждая координата в нём будет соответствовать одному из профилей. Теперь

каждому аффиксу можно поставить в соответствие вектор v, причём единицы в нём будут стоять только на тех местах, которые соответствуют профилю, описывающему этот аффикс. Проще говоря, мы составляем вектор, который указывает на профили, соответствующие данному аффиксу. В общем случае данная модель достаточно сложна, но для рассматриваемых языков число характеристик первого уровня невелико, и ситуация значительно упрощается.

Для характеристик второго уровня ситуация немного проще. Пусть имеются M различных характеристик второго уровня, обозначим их через $D_1, \ldots, D_M$. Каждой из характеристик поставим в соответствие множество её представителей, т. е. для характеристики второго уровня D_i множество её представителей обозначим через $S_i = \{s_{i1}, \ldots, s_{ik}\}$, где s_{ij} — j-й представитель i-й характеристики. Количество представителей каждой характеристики второго уровня также будем обозначать через мощность множества, т. е. число представителей характеристики D_i равняется $|S_i|$.

Далее следует разграничивать два случая, от которых будет зависеть дальнейшее построение модели:

1) значение всех характеристик второго уровня однозначно определяется независимо от профиля;
2) значение некоторых характеристик второго уровня непосредственно зависит от профиля.

Рассмотрим первый случай. Здесь значение характеристик второго уровня не зависит от профиля, поэтому логично каждому аффиксу поставить в соответствие вектор из нулей и единиц аналогично тому, как это было сделано для характеристик первого уровня. Длина этого вектора будет равняться $m = |S_1| + |S_2| + \ldots + |S_M|$, тогда значение координаты вектора будет равняться единице, если соответствующая этой координате характеристика определяется данным аффиксом, и нулю в противном случае. Стоит заметить, что представители одной характеристики не являются взаимоисключающими.

Каждый профиль однозначно определяет множество характеристик второго уровня для данной словоформы. Для каждого профиля можно задать множество номеров координат вектора характеристик второго уровня, причём эти координаты будут соответствовать только таким характеристикам, которые имеют смысл для данного профиля (например, в русском языке определять спряжение для существительных не имеет смысла). Перейдём теперь непосредственно к алгоритму восстановления грамматических характеристик словоформы по её окончанию.

После применения алгоритма 2.1 для заданной словоформы мы можем получить множество её возможных некорневых частей и их разбиений на аффиксы. Опишем пошагово алгоритм для восстановления грамматических характеристик словоформы путём анализа таких частей.

Алгоритм 3.1.
Вход: словоформа $\sigma = \sigma_{i_1} \ldots \sigma_{i_l}$, $\sigma_{i_j} \in \Sigma$, $|\sigma| = l$, и множество T её допустимых некорневых частей, полученных алгоритмом 2.1.

Выход: множество характеристик первого и второго уровня для каждой из возможных некорневых частей.

Шаг 1. Вычисление характеристического профиля каждого аффикса. В действительности про каждый аффикс из некорневой части словоформы мы знаем, какие представители каждой из характеристик первого уровня ему соответствуют, а значит, знаем и то, как выглядит соответствующий ему вектор профилей.

Шаг 2. Вычисление характеристического профиля всей словоформы. Для этого необходимо выполнить покоординатное логическое умножение этих векторов. Проще говоря, словоформа будет характеризоваться профилем i, если во всех векторах профилей на i-м месте будет стоять единица. Вполне возможна ситуация, когда словоформу можно будет охарактеризовать сразу несколькими профилями. Также возможна ситуация, когда ни один из профилей не будет выбран, это означает, что соответствующее разбиение на аффиксы не является корректным и далее его не следует анализировать.

Шаг 3. Вычисление векторов характеристик второго уровня для каждого аффикса. На данном этапе нам известен набор возможных профилей словоформы. Для каждого профиля мы знаем номера координат вектора характеристик второго уровня, значение которых нам в дальнейшем и нужно узнать. Также для каждого аффикса мы имеем вектор из нулей и единиц, отражающий значения его характеристик второго уровня. Теперь нужно взять из этого вектора только те координаты, которые необходимы для данного профиля, и составить из них новые векторы.

Шаг 4. Вычисление характеристик второго уровня для всей словоформы. Выполним покоординатное логическое сложение этих векторов. Представители характеристик, на местах которых в получившемся векторе стоят единицы, и будут принадлежать рассматриваемой словоформе. Иными словами, если хотя бы в одном из векторов на i-м месте стоит единица, то говорим, что соответствующий этой координате представитель характеристики второго уровня описывает данную словоформу.

Шаг 5. Шаги 1—4 стоит применить для каждой возможной некорневой части словоформы.

Данный алгоритм был описан для случая, когда значение всех характеристик второго уровня однозначно определяется независимо от профиля. Рассмотрим теперь случай, когда существуют такие характеристики, которые зависят непосредственно от профиля. Получается, что теперь ставить каждому аффиксу в соответствие один вектор второстепенных характеристик (единый для всех профилей) некорректно. Каждому аффиксу для каждого профиля следует поставить в соответствие свой вектор характеристик второго уровня, причём только тех, которые имеют смысл для данного профиля. Это нас избавляет от того, чтобы хранить для каждого профиля множество номеров координат нужных характеристик. В целом алгоритм остаётся прежним. На первом шаге аналогичным образом определяем профиль, и уже в зависимости от него для каждого

аффикса выбираем соответствующий вектор характеристик второго рода и выполняем операцию логического сложения. В итоге словоформа будет обладать такими характеристиками второго уровня, на местах которых стояли единицы в суммарном векторе.

Модификация алгоритма выделения псевдоосновы

В этом разделе будет рассматриваться модификация базового алгоритма (алгоритма 2.1) выделения псевдооснов словоформ, использующего метод структурных схем. Будем рассматривать языки, в которых достаточно сильно развито формирование новых слов со схожим смыслом, но являющихся *другой частью речи*, путём присоединения к концу слова словообразовательных суффиксов. В связи с этим будет предложен алгоритм выделения псевдооснов словоформ с известной заранее частью речи. Заметим также, что префиксная часть в данном случае учитываться не будет, она будет (условно) входить в состав основы, задача же заключается в отсечении корректного окончания.

Определим введённые условия более формально. Будем рассматривать языки, у которых задана только одна характеристика первого уровня: $C_1 =$ $=$ «часть речи». Множество представителей данной характеристики конечно, обозначим его через $P = \{p_1, p_2, \ldots, p_n\}$, где p_i — это часть речи. Также нам дано множество типов суффиксов, обозначим его (как и раньше) через $A =$ $= \{a_1, a_2, \ldots, a_m\}$, где a_i — это тип суффикса. Через A_i будем обозначать множество представителей каждого из типов a_i, $i = 1, \ldots, m$. Кроме того, у нас есть множество всех возможных структурных схем данного языка Sch. Напомним, что множество Sch состоит из упорядоченных векторов, элементами которых являются типы аффиксов. Будем исходить из предположения, что все структурные схемы для таких языков могут быть поделены на группы в зависимости от части речи. Тогда множество Sch можно представить как $\mathrm{Sch} = \mathrm{Sch}_{p_1} \cup \ldots \cup \mathrm{Sch}_{p_n}$, где Sch_{p_i} — множество схем, отвечающих за часть речи p_i. Стоит заметить, что $\mathrm{Sch}_{p_1} \cap \ldots \cap \mathrm{Sch}_{p_n} = \{(\varepsilon)\}$, т. е. одинаковой для разных частей речи является только примитивная схема, т. е. такая схема, которая определяет слова без выраженных окончаний.

Используя введённые обозначения, опишем модификацию базового алгоритма. На вход, помимо слова σ, теперь также подаётся и его часть речи p_i. Тогда стоит провести разбор, как и раньше, но множество Sch заменить его подмножеством Sch_{p_i}. На выходе получим множество различных вариантов окончаний и их разбиений согласно схемам, но в данном случае они будут удовлетворять структуре слов заданной части речи. Такая модификация позволит сократить число вариантов возможных окончаний и, соответственно, повысить точность разбора.

Непосредственный алгоритм выделения псевдоосновы также подвергается некоторым модификациям. Учитывая введённые ограничения, этот алгоритм может быть реализован значительно проще. Если мы заранее знаем часть речи

словоформы, то количество соответствующих ей описаний значительно сокращается. Получаем, что после того, как мы узнали часть речи словоформы, нам остаётся проверить, есть ли в ней словообразовательные суффиксы, расположенные в одном из возможных порядков. Это возможно сделать путём анализа окончания словоформы, т. е. нам необходима функция, которая бы выделяла в конце слова один из возможных суффиксов, соответствующий определённому типу. Для этого введём функцию R, на вход такой функции будет подаваться слово и один из типов суффиксов, а на выходе будем иметь множество подслов, которые получились после отсечения от изначального слова всевозможных суффиксов заданного типа. Формально говоря,

$$R(\langle \sigma, a_i \rangle) = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}, \quad a_i \in A, \quad \sigma - \text{слово}, \quad \sigma_i - \text{подслово},$$

если невозможно ничего отсечь, то функция возвращает пустое множество.

Алгоритм 3.2.

Дано: слово σ, его часть речи p_l, все возможные схемы для такой части речи Sch_{p_l}. Все нижеизложенные шаги следует применять для каждой схемы s из множества Sch_{p_l}.

Шаг 1. Вычисляем множество $M = R(\langle \sigma, a_{|s|} \rangle)$, где $a_{|s|}$ — последний элемент в схеме s.

Шаг 2. Повторяем шаг 1 для всех подслов из множества M, но вместо $a_{|s|}$ берём $a_{|s|-1}$, т. е. следующий (с конца) тип аффикса в схеме s.

Шаг 3. Повторяем шаг 2, применяя функцию R для всех полученных на предыдущем шаге слов, но с изменением типа аффикса: каждый раз берём следующий с конца. Выполняем это до тех пор, пока не пройдём по всему вектору s или пока R не выдаст пустое множество.

Итог: все возможные основы, которые вернула функция R, остановившись при окончании вектора s. Если получили пустое множество, то возвращаем в качестве основы изначальное слово σ.

В целом подобный алгоритм подходит для всех случаев, когда схемы могут быть классифицированы по одному из признаков.

4. Компьютерный анализ словоформ французского языка

Опишем некоторые основные принципы построения структурных схем для словоформ французского языка. Основной акцент делается на морфологические и словообразовательные особенности французского языка, так как именно они позволяют производить корректный морфологический анализ словоформ.

Рассмотрим суффиксальный способ словообразования. Согласно [6], во французском языке выделяется девять частей речи: существительное (nom), прилагательное (adjectif), глагол (verbe), наречие (adverbe), артикль (article),

местоимение (pronom), предлог (préposition), союз (conjonction) и междометие (interjection). Суффиксальный способ словообразования чаще всего применяется для образования существительных, прилагательных, глаголов и наречий. Остановимся именно на этих четырёх частях речи.

Суффиксы делятся на словообразовательные и словоизменительные. Для прилагательных, глаголов и некоторых существительных флексивные суффиксы служат для изменения слова по роду (мужской/женский) и числу (единственное/множественное). Мы не будем выделять отдельный тип для таких суффиксов. При спряжении глагола изменяются лицо, число, время, род, наклонение и залог. Суффиксы, служащие для этого, также являются словоизменительными. Их тип обозначим через v. Словообразовательные суффиксы служат для формирования новых слов. С их помощью слова могут менять свою часть речи или обретать определённый оттенок значения (уменьшительность, многократность и т. п.).

Построение структур окончаний основывается на грамматике французского языка [3—6]. Для французского языка удобнее всего это делать, рассматривая каждую часть речи, поэтому при построении нашей модели анализировались и строились структурные схемы для четырёх основных частей речи: глаголов, существительных, прилагательных и наречий.

Глагол

Во французском языке при спряжении глаголы изменяют окончания в зависимости от лица, числа, рода, времени, наклонения и залога. По типам спряжения глаголы во французском языке делятся на три группы, различающихся по типам окончаний в инфинитиве. Глаголы первой и второй группы называются «правильными», так как имеют определённые правила спряжения, т. е. для них однозначно определён вид окончания в зависимости от различных грамматических характеристик. Глаголы третьей группы являются «неправильными», так как при спряжении они меняют свою основу и не имеют единых правил изменения окончания. Все возможные окончания инфинитивов и их формы, которые они приобретают при спряжении, будут учитываться в суффиксах, соответствующих типу v. Получаем первую возможную схему, описывающую окончания глаголов: основа-v. К примеру, глагол первой группы marcher (ходить) можно представить как «march-er». При спряжении этот же глагол может принять форму «march-ions» (ходили), «march-erai» (буду ходить) и т. п.

Кроме суффиксов «-er», «-ir» и т. п., во французском языке также можно выделить более сложные суффиксы, например «-ifier» и «-iser». Они служат для образования глаголов от существительных (нарицательных и имён собственных), прилагательных и аббревиатур. Сами эти суффиксы будем рассматривать как сложные, т. е. разбивать их на две части: «-ifi-er» и «-is-er». Это объясняется тем, что при спряжении будет меняться только финальная часть («-er») и тем, что от таких глаголов также можно образовывать новые слова, частички «-ifi» и «-is-» будут входить в их состав, а «-er» нет. Также во

французском языке есть суффиксы, которые не меняют часть речи, а только придают определённый оттенок. Таковы суффиксы «-aill-», «-ass-», «-ill-», «-och-», «-onn-», «-ot-», «-ouill-», «-ard-», «-âtr-», «-et-», «-in-», их также будем понимать как сложные. Примеры: «chant-er»—«chant-onn-er» (петь—напевать), «touss-er»—«touss-ot-er» (кашлять—покашливать) и т. д. Таким образом, можно ввести ещё один тип суффиксов, который будет отвечать за одну из частей в сложных глагольных окончаниях. Теперь можем предложить ещё одну схему, описывающую окончания глаголов: основа-c-v. Во французском языке возможно прибавление глагольного суффикса к слову, которое также было образовано суффиксальным способом. Иными словами, производное слово может стать производящим. Здесь стоит выделить два случая: образование глаголов на основе существительных и на основе прилагательных. В связи с этим можно предложить ещё две схемы, описывающие окончания глаголов: основа-n-v, через n будем обозначать суффиксы производных существительных, которые стали производящими; основа-av-c-v, через av будем обозначать суффиксы производных прилагательных, которые стали производящими.

Наречие

Бóльшая часть наречий во французском языке образуется от прилагательных с помощью присоединения к ним суффикса «-ment». Поэтому в окончании наречий сначала стоит суффикс, образующий прилагательное, а затем суффикс, образующий наречие. Различают несколько вариантов окончания «–ment» в зависимости от того, к какому прилагательному оно добавляется: «-ment», «-ement», «ément», «-amment», «-emment». Если окончание прилагательного не выражено явно, то наречие может быть описано следующей схемой: основа-d, где d — это тип суффиксов, которые образуют наречие.

Для случаев, где окончание прилагательных выражается явно, как, например, в слове «alphabét-ique-ment» (в алфавитном порядке), где суффикс «-ique» отвечает за формирование прилагательного, а «-ment» — за наречие, можно предложить такую схему: основа-a-d. Некоторые прилагательные во французском языке были образованы от существительных, имеющих собственные суффиксы. Соответствующая схема: основа-n-a-d. Также есть наречия, которые не были образованы суффиксальным образом, а просто исторически сформировались такими, какие они есть. Это такие наречия, как «droit» (прямо), «bien» (хорошо), «mal» (плохо) и т. п. Для них можно предложить следующее описание: основа-ε, где ε — пустой символ, или просто основа.

Существительное

Существительные во французском языке изменяются по числам и иногда по родам. Изменения в роде происходят, только если речь идёт о профессиях или о других понятиях, относящихся непосредственно к человеку или животному.

Такие изменения происходят путём замены финального суффикса или присоединения нового. В большинстве же своём существительные имеют фиксированный род и не могут изменяться. Множественное число образуется с помощью присоединения к слову определённого суффикса, который зависит от того, на какую букву заканчивается слово. Все эти особенности изменения существительных касаются грамматических характеристик и далее будут учтены в структурах с помощью суффиксов типа nt.

Во французском языке существует достаточно много существительных, которые не содержат какого-либо явного окончания: например, «lac», «mur», «sol», «tome», «foudre». Такие слова могут составлять основу производных слов, но сами не могут быть разбиты на значимые части. Отсюда получаем такое возможное описание: основа-ε. Ещё одну многочисленную группу составляют существительные с односложным окончанием: к примеру, «siffle-ment», «malad-ie»; описание: основа-nt. Во французском языке существуют отглагольные существительные, в частности такие, которые были образованы от основы глаголов, оканчивающихся на «-ifi-er» и «-is-er»: «humid-ifi-er», «humid-ifi-cation», «cristall-is-er», «cristall-is-ation». При образовании существительных от глаголов, оканчивающихся на «-ir», в слове возникает суффикс «-iss-»: например, «fin-ir»—«fin-iss-age/-eur/-euse»; описание: основа-c-nt. Также возможен вариант, когда от прилагательного с явно выраженным окончанием был получен глагол на «-is-er», а на основе этого глагола было получено существительное. Приведём пару примеров: «milit-aire»—«milit-ar-is-er»—«milit-ar-is-ation», «individu-el»—«individu-al-is-er»—«individu-al-is-ation»; описание: основа-av-c-nt. Во французском языке достаточно часто встречаются существительные, образованные от прилагательных: к примеру, «actu-al-ité», «techn-ic-ité», «act-iv-ité», «opt-ic-ien», «natur-al-isme/-iste» и т. д.; описание: основа-av-nt.

Прилагательное

Во французском языке прилагательные изменяются по родам и числам. Так же как и для существительных, это происходит путём присоединения в конец слова определённых суффиксов или изменения самого последнего суффикса. Такие грамматические особенности со всеми возможными исключениями будут учтены в суффиксах, соответствующих типу a. Однако стоит заметить, что большое число прилагательных не имеет формально выраженного рода или числа. Для таких слов эти грамматические характеристики должны определяться из контекста.

Опишем основные структуры окончаний. Пусть тип a отвечает за суффиксы, которые образуют прилагательные. Тогда возможными схемами будут: основа-ε, основа-a. Для прилагательных количество различных структур окончаний не очень велико. Первая описывает случаи, когда прилагательное было образовано от уже ранее образованного существительного: к примеру, «constitution»—«constitut-ionn-el/constitut-ionn-elle»; описание: основа-n-a. Вторая структура описывает окончания прилагательных, которые были получены от глаголов

на «-ifi-er» или «-is-er». К примеру, от глагола «mod-ifi-er» можно образовать прилагательные «modifi-ant(-e)», «mod-ifi-able», «mod-ifi-cateur», «mod-ifi-catrice», «mod-ifi-catif». Для окончания «-is-er» можно привести такой пример: «automat-is-er»—«automat-is-able». Получаем ещё одно описание: основа-c-a.

Подытожив все вышесказанное, для французского языка получаем следующие данные: множество типов аффиксов имеет вид $A = \{\mathrm{c}, \mathrm{v}, \mathrm{n}, \mathrm{av}, \mathrm{d}, \mathrm{a}, \mathrm{nt}\}$, множество Sch структурных схем состоит из 16 элементов:

$$\mathrm{Sch} = \{(\mathrm{v}), (\mathrm{c}, \mathrm{v}), (\mathrm{n}, \mathrm{v}), (\mathrm{av}, \mathrm{c}, \mathrm{v}), (\mathrm{d}), (\mathrm{a}, \mathrm{d}), (\mathrm{n}, \mathrm{a}, \mathrm{d}), (\mathrm{nt}), (\mathrm{c}, \mathrm{nt}), (\mathrm{av}, \mathrm{c}, \mathrm{nt}), (\mathrm{av}, \mathrm{nt}), (\mathrm{n}, \mathrm{av}, \mathrm{nt}), (\mathrm{a}), (\mathrm{n}, \mathrm{a}), (\mathrm{c}, \mathrm{a}), (\varepsilon)\}. \quad (4.1)$$

Для того чтобы восстанавливать грамматические характеристики словоформ, необходимо получить некоторую информацию из каждого аффикса, который входит в её состав. В случае французского языка имеем только одну характеристику первого уровня — «часть речи». Множество представителей этой характеристики выглядит так:

$$F = \{\text{существительное}, \text{прилагательное}, \text{наречие}, \text{глагол}\}.$$

К характеристикам второго уровня отнесём «род», его представителями будут «мужской»/«женский»; «число» — «единственное»/«множественное»; «лицо» — «первое»/«второе»/«третье»; «время» — «простое прошедшее»/«имперфект»/«презенс»/«футур», «наклонение» — «индикатив»/«кондиционалис»/«субъюнктив», «тип формы глагола» — «финитная»/«причастие»/«инфинитив». В случае французского языка «информативным» является самый последний аффикс окончания, именно он несёт информацию о части речи и о других грамматических характеристиках. Так же как и при построении схем, для каждой части речи опишем те характеристики, которые могут быть определены.

При анализе глаголов стоит разделять два случая: сложную и простую форму глаголов. Под сложной формой глагола подразумевается сочетание вспомогательного и смыслового глагола, выражающее единое в смысловом отношении значение определённого времени или залога. Общая схема такова: $\mathrm{A} + \mathrm{V_{pp}}$, где A — это вспомогательный глагол («avoir» или «être»), $\mathrm{V_{pp}}$ — смысловой глагол в форме причастия прошедшего времени. В данном случае, анализируя форму вспомогательного глагола, можно определить следующие грамматические характеристики: время, лицо, число. Под простой формой понимается такая форма глагола, которая не требует использования вспомогательного глагола. Здесь стоит анализировать последний аффикс глагола. Грамматические характеристики, которые возможно определить: наклонение, время, лицо, число.

В словоформах существительных и прилагательных также стоит анализировать самый последний аффикс в окончании. Грамматические характеристики для восстановления: род, число. Очевидно, здесь возникает проблема множественности. Некоторые аффиксы одновременно могут символизировать как существительное, так и прилагательное. Кроме того, род и число некоторых существительных и прилагательных не могут быть определены однозначно. Поэтому без анализа контекста избежать подобных неопределённостей невозможно.

Для наречий последний аффикс окончания может дать информацию только о части речи.

5. Результаты экспериментов

Для определения эффективности и корректности метода структурных схем окончаний на языке C++ были реализованы описанные выше алгоритмы в применении к французскому языку. В частности, был реализован алгоритм выделения псевдооснов с разбиением окончания на аффиксы, а также алгоритм восстановления грамматических характеристик. Была проведена серия вычислительных экспериментов. Во всех случаях использовались коллекции словоформ, взятые из обработанных (размеченных) художественных текстов на французском языке, общий объём которых составляет 42 000 слов. Это позволило группировать слова по части речи, а также производить анализ слов в различных формах (изменённых в роде, в числе; для глаголов — во времени и т. д.).

Во-первых, была проанализирована корректность разбиения на структурные схемы и определения грамматических характеристик случайных наборов слов. Считалось, что корректное разбиение существует, если во множестве возможных разбиений присутствует такое, которое оставляет правильную основу и имеет корректное разбиение с точки зрения словообразовательных процессов. Также считалось, что грамматические характеристики определяются правильно, если в списке возможных грамматических «профилей» есть верный. В итоге мы получили, что 95 из 100 слов имеют верное разбиение и 82 из 100 верные грамматические характеристики.

Во-вторых, был проведён ряд экспериментов для выявления зависимости между средним количеством подходящих структурных схем для словоформ разных частей речи от ограничения на количество букв в основе. В таблице 1 приведены результаты данного эксперимента: числа показывают среднее число выявленных структурных схем для соответствующей части речи и количества букв в основе.

Таблица 1. Зависимость среднего числа схем от ограничения на количество букв в корне

	3	4	5
глагол	3,8288	3,4701	2,9294
наречие	7,225	6,8605	6,2403
существительное	4,5446	3,9664	3,209
прилагательное	4,9874	4,5148	3,86

Видно, что чем большее число букв в основе задаётся, тем меньше возможных вариантов схем формируется. К примеру, при увеличении количества букв в основе с трёх до пяти множественность падает для прилагательных на 22 %, а для существительных на 29 %. Видно, что при помощи задания параметров имеется возможность уменьшать множественность результатов морфологического анализа.

Ещё один проведённый эксперимент заключался в подсчёте наиболее часто встречаемых схем для каждой из частей речи. Для реализации этого эксперимента анализировались списки слов одной части речи. В одном из случаев множество Sch состояло только из схем, предназначенных для исследуемой части речи. Таким образом было получено распределение по схемам для определённой части речи. На следующей гистограмме (рис. 1) отображены результаты данного типа эксперимента для случая глагола.

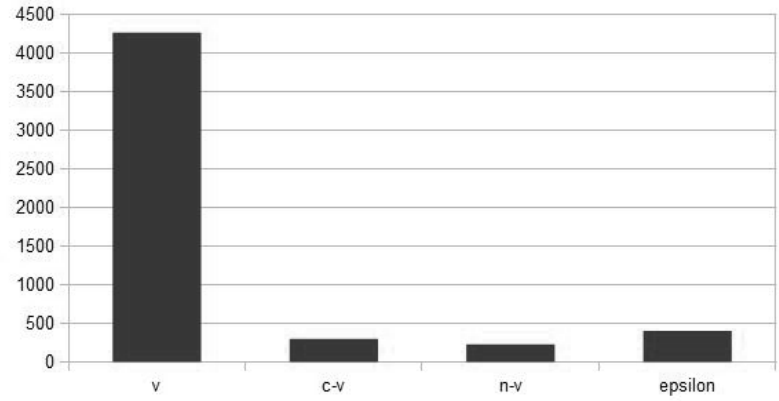

Рис. 1. Распределение по глагольным схемам

В другом случае на множество Sch не накладывалось никаких ограничений. В итоге мы получили распределение по всем возможным схемам. Для удобства отображения результатов схемы были объединены по части речи. Вторая гистограмма (рис. 2) показывает результаты данного типа эксперимента для случая глагола.

Приведённые гистограммы наглядно показывают, что в большинстве случаев в качестве результата мы имеем правильные схемы. Этот факт ещё раз доказывает корректность работы предложенных алгоритмов.

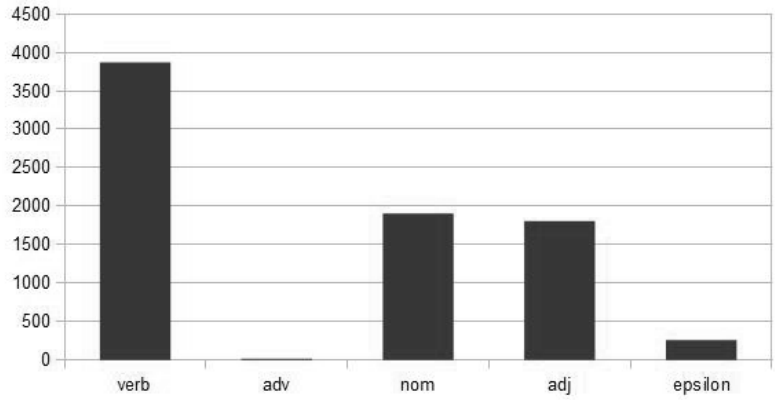

Рис. 2. Распределение по схемам всех частей речи

6. Заключение

В данной работе предложена модель морфологического анализа словоформ языков с развитым аффиксальным словообразованием и словоизменением. Главная задача, которую призвана решать предлагаемая модель, заключается в грамматически корректном выделении псевдооснов словоформ. Правильное и эффективное решение этой задачи играет ключевую роль в оптимизации процессов информационно-поисковых систем. В связи с этим был предложен алгоритм выделения псевдооснов словоформ и разбиения окончания на аффиксы, основанный на методе структурных схем. Основная идея этого метода заключается в представлении некорневой части слова в виде такой последовательности аффиксов, которая соотносится с одной из доступных (с грамматической точки зрения) схем исследуемого языка. Главным преимуществом данного подхода является то, что он не требует использования морфологического словаря и согласуется с грамматическими особенностями языка.

Также затронут вопрос восстановления грамматических характеристик словоформ. Предложен алгоритм, который базируется на результатах вышеописанного алгоритма: разбиение некорневой части слова позволяет получить информацию о грамматических характеристиках каждого аффикса, содержащегося в некорневой части, и, как следствие, всего слова в целом.

Для демонстрации возможности использования подобных алгоритмов на примере французского языка описан процесс построения всех необходимых структур. В частности, обозначены необходимые типы аффиксов, схемы, которые описывают окончания слов французского языка, и списки грамматических характеристик, которые возможно определить. На основе программы, реали-

зующей все предложенные алгоритмы, был проведён ряд экспериментов для получения численной оценки эффективности данного метода. Приведённые результаты иллюстрируют возможность применения предложенного подхода для решения задач информационного поиска.

Литература

[1] Болховитянов А. В., Чеповский А. М. Алгоритмы морфологического анализа компьютерной лингвистики. — МГУП им. Ивана Федорова, 2013.

[2] Егорова Е. Е., Чеповский А. М. Морфологическая модель для анализа и индексирования текстов на индоевропейских языках // Тр. Междунар. конф. по физико-технической информатике CPT2013, 12—19 мая 2013 г., Ларнака, Республика Кипр. — Протвино; Москва: Изд-во ИФТИ. — С. 154—159.

[3] Катагощина Н. А. Как образуются слова во французском языке. — М.: URSS; КомКнига, 2006.

[4] DuBois J., Lagane R. Livres de bord: Grammaire. — Larousse, 2010.

[5] Grevisse M., Goosse A. Le bon usage. — Paris: Duculot, 2011.

[6] Huot H. La morphologie: forme et sens des mots du français. — Armand Colin, 2006.