ОБРАБОТКА СИГНАЛОВ И ДАННЫХ, РАСПОЗНАВАНИЕ ОБРАЗОВ, ВЫЯВЛЕНИЕ ЗАКОНОМЕРНОСТЕЙ И ПРОГНОЗИРОВАНИЕ

УДК 81'322

МЕТОДЫ МОРФОЛОГИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

Пруцков Александр Викторович, кандидат технических наук, доцент, Рязанский государственный радиотехнический университет, 390005, Российская Федерация, г. Рязань, ул. Гагарина, 59/1, e-mail: mail@prutzkow.com

Розанов Алексей Константинович, аспирант, Рязанский государственный радиотехнический университет, 390005, Российская Федерация, г. Рязань, ул. Гагарина, 59/1

Выполнен обзор существующих отечественных и зарубежных методов морфологической обработки текстов. С целью разработки универсального метода (УМ) генерации и определения форм слов выявлены преимущества и недостатки существующих подходов. Морфологический анализ и синтез, ориентированный на один естественный язык, не подходит для построения УМ. Алгоритмы и структуры хранения данных (словари) таких методов специализированы под особенности одного языка и не могут быть изменены для других языков. Подходы к морфологическому анализу, предназначенные для нескольких языков, имеют «слабые места», которые могут быть скорректированы в УМ. На основе анализа рассмотренных в статье подходов к морфологическому анализу и синтезу сформулированы требования к УМ: обработка словоформ языков различных групп и семейств; универсальность структуры словарей, не требующей конвертации для решения задач определения или генерации словоформ; модель формообразования, на основе которой построен метод, должна описывать любые виды образования форм всей парадигмы слова. Разработанный авторами статьи метод генерации и определения форм слов соответствует этим требованиям.

Ключевые слова: автоматическая обработка текстов, морфологический анализ, морфологический синтез, генерация словоформ, определение словоформ, машинный перевод, диалог человека с ЭВМ, естественные языки

WAYS OF NATURAL LANGUAGE MORPHOLOGICAL PROCESSING

Prutskov Alexander V., Ph.D. (Engineering), Associate Professor, Ryazan State Radioengineering University, 59/1 Gagarin St., Ryazan, 390005, Russian Federation, e-mail: mail@prutzkow.com

Rozanov Aleksey K., postgraduate student, Ryazan State Radioengineering University, 59/1 Gagarin St., Ryazan, 390005, Russian Federation

We review existing home and foreign ways of morphological analysis and synthesis. For the purpose of universal word-form generation and recognition method development we expose existing way advantages and disadvantages. Morphological analysis and synthesis methods, that oriented for the one natural language, can not be the basis for the universal method development. This methods algorithmes and data storage structure (dictionaries) are specialized for the one language features and can not be changed for other languages. Morphological ways applicable for some languages have weak points, which can be fixed by the universal method. Analysis of morphological ways let us put in claim to developing method of word-form

generation and recognition. The universal method of word-form generation and recognition has to process word-forms of natural languages of different groups and families, have a universal dictionary structure, which do not require convertation for generation and recognithion task solvation, the model of word-form building, which the univesal method based of, has to describe of whole paradigm form building. We have developed the universal method satisfied the claim.

Keywords: natural language processing, morphological analysis, morphological synthesis, word-form generation, word-form recognition, machine translation, human-computer dialog, human languages

Задача морфологической обработки текстов на естественных языках является актуальной для разных областей человеческой деятельности. После появления ЭВМ (и особенно персональных ЭВМ) были разработаны и программно реализованы алгоритмы такой обработки. Однако существующие алгоритмы имеют ряд недостатков, ухудшающих качество получаемых результатов и не обеспечивающих универсальность для различных языков. Поэтому целями данной статьи являлись: 1) анализ существующих отечественных и зарубежных методов для сферы морфологической обработки текстов с целью их использования в качестве основы для создания универсального метода генерации и определения форм слов; 2) выявление преимуществ и недостатков этих методов; 3) формулирование на основе этого требований к разрабатываемому методу генерации и определения форм слов.

Морфологический уровень автоматической обработки текстов. Автоматическая (машинная) обработка текстов (АОТ) представляет собой действия, выполняемые с помощью вычислительной техники и программного обеспечения. АОТ применяется при машинном переводе, грамматической проверке текста, диалоге с компьютером на естественном языке, анализе текста (например, патентной информации по изобретениям), анализе и синтезе речи и в других областях. Эффективная реализация АОТ позволяет улучшить качество «человеко-машинного» управления системами и процессами, в том числе и в реальном масштабе времени. В общем случае АОТ осуществляется на трех уровнях: морфологическом, синтаксическом, семантическом.

Помимо деления текста на слова (при анализе) и сбора текста из слов (при синтезе) на морфологическом уровне решаются две основные задачи. Генерация формы слова (синтез, продукция), осуществляемая при морфологическом синтезе, – это процесс получения формы с использованием в качестве начальных параметров основы и грамматического значения. Определение формы слова (анализ, распознавание, интерпретация) при морфологическом анализе – это процесс обратный генерации, который заключается в нахождении по данной словоформе ее нормальной формы (основы) и грамматического значения.

Авторами был выполнен сравнительный анализ достоинств и недостатков методов/подходов, применяемых при морфологическом анализе и синтезе. На основе результатов анализа предложены модель формообразования [14], алгоритмы генерации и определения форм слов естественных языков различных семейств и групп [30], составляющие метод генерации и определения форм слов [15], а также программная система, реализующая этот метод [16].

Область исследования методов морфологической обработки текстов. В мире насчитывается по одним оценкам от 2 500 до 5 000 [22], а по другим – от 3 000 до 8 000 [8] естественных языков. Анализ морфологии всех естественных языков и разработок в области их морфологического анализа и синтеза потребовал бы значительного времени. Поэтому область исследования была разумно ограничена отечественными методами и алгоритмами морфологической обработки текстов на русском языке, а также зарубежными разработками, в которых заявлена возможность морфологического анализа и синтеза текстов на русском языке. Выбор русского языка в качестве основного обоснован следующими причинами: разработки в области АОТ и публикации на русском языке по данной тематике более доступны,

чем на других языках; методы морфологической обработки текстов на русском языке сложны и многообразны, при этом реализуются различные подходы к генерации и определению форм слов; русский язык – родной язык авторов данного анализа.

Причины многообразия методов морфологической обработки текстов на русском языке заключаются в сложности его морфологии. Этот язык обладает рядом особенностей, типичных для большинства естественных языков со сложной морфологией. 1. Супплетивизм – образование форм одного и того же слова от разных основ («*иду*» – «*шел*», «*хороший*» – «*лучше*»). 2. Различные способы слово- и формообразования: префиксальный («*читать*» – «*прочитать*»), суффиксальный («*бросать*» – «*бросить*»), чередование букв в основе («*собирать*» – «*собрать*»). 3. Синонимия – образование словоформы более чем одним способом; например, сравнительная степень прилагательных и наречий («*сильнее*» – «*сильней*») и творительный падеж единственного числа некоторых существительных («*водой*» – «*водою*»). 4. Омонимия – наличие у одной грамматической формы нескольких грамматических значений; например, словоформа «*стекло*» может быть существительным среднего рода, единственного числа, именительного или винительного падежа или формой глагола «*стекать*» изъявительного наклонения прошедшего времени совершенного вида. 5. Синтетические (простые: «*делать*», «*говорил*») и аналитические (сложные: «*буду делать*», «*говорил бы*») формы слов.

Перечисленные особенности (сложности) морфологии необходимо учитывать при разработке методов генерации и определения форм слов.

Любой метод морфологического анализа и/или синтеза включает две части: декларативную, которая состоит из данных, структурированных в словарях и таблицах, необходимых для морфологического анализа и синтеза; процедурную – включающую алгоритмы морфологического анализа и синтеза, вспомогательные процедуры. Далее мы рассмотрим наиболее известные методы морфологического анализа и синтеза, разработанные отечественными учеными [2, 4, 5, 10, 12, 21, 31] и зарубежными специалистами [24, 26, 29], выявим их преимущества и недостатки, которые возможно устранить.

Классификация, предложенная Чарльзом Ф. Хоккетом в середине 1950-х гг., делит модели морфологии, трактующие формообразование, на три группы [3, 5, 27, 33]: элементно-комбинаторную, предполагающую, что форму слова можно разбить на элементарные части и описать формообразование как последовательность данных частей; элементно-операционную, задающую правила образования формы слова из основы с влиянием одних частей слова на другие; словесно-парадигматическую, разделяющую все слова по определенным типам формообразования, имеющим специфичные признаки (формы слова образуются по правилам, свойственным типу формообразования).

Однако многие методы морфологического анализа и синтеза нельзя однозначно отнести к одной из групп данной классификации.

Преимущества и недостатки существующих методов морфологической обработки текстов. В нашей стране были разработаны методы морфологического анализа и синтеза форм слов русского языка (в основном они будут описаны далее), а также языков народов, проживающих на территории Российской Федерации (например, татарского [19]). Данные методы легли в основу систем АОТ и с некоторыми доработками используются до сих пор.

Г.Г. Белоногов совместно с Т.С. Белоноговой и А.К. Родионовой предложили в рамках автоматизированной информационно-поисковой системы точные и приближенные процедуры морфологического анализа и синтеза словоформ [4] (конец 1960-х гг.). Точные процедуры предназначены для анализа и синтеза форм слов, имеющихся в словарях системы, а приближенные – для слов, отсутствующих в словарях. В точных процедурах морфологического анализа и синтеза используются словарь основ слов (СОС) и следующие таблицы с указанием их функций: таблица окончаний (флексий) – каждому окончанию сопоставлен

номер; морфологическая таблица – номеру морфологического класса (типу основы) из СОС и номеру окончания (флексии) сопоставлен номер грамматической информации; таблица грамматической информации – номеру грамматической информации сопоставлено грамматическое значение; таблица для основ с чередованием гласных (тип II).

СОС содержит основы нормальных форм слов и вариантные основы (для двух типов изменения основы: чередования согласных (тип III) и нерегулярных основ (тип IV)). Вариантная основа – это основа, отличающаяся от основы нормальной формы слова чередованием букв.

Преимущества метода: простота структуры СОС и таблиц; простота реализации алгоритма для слов с неизменяемыми основами; возможность определение словоформ, отсутствующих в словарях системы. Недостатки метода: структура таблиц не универсальна – для синтеза форм слов необходимо преобразовывать морфологическую таблицу; СОС содержит несколько основ одного слова; каждый тип изменения основы русского языка обрабатывается отдельным алгоритмом; ориентация алгоритма и структуры словарей только на русский язык; ориентация на флективный анализ, т.е. определение основы и грамматического значения по окончанию словоформы; отсутствие возможности обработки аналитических форм слов, представляющих собой сочетание знаменательного и служебного слов (например, у глаголов русского языка «*буду говорить*», «*говорил бы*»).

М.Г. Мальковский разработал (начало 1980-х гг.) морфологический компонент [10, 11], являющийся составной частью системы общения с человеком на естественном языке TULIPS-2. Система состоит из СОС, словаря окончаний, таблицы чередований и таблицы исключений. В свою очередь СОС состоит из статей, содержащих морфосинтаксические показатели и лексико-семантические значения.

В словарных статьях СОС указываются такие данные: постоянные и переменные морфологические признаки, разделенные по классам: морфологическим (М-класс), парадигматическим (П-класс) и синтаксическим (С-класс); возможность образования тех или иных форм (например, невозможно образовать форму единственного числа слова русского языка «*ножницы*»); тип чередования; список окончаний, при которых данное чередование встречается.

Исключениями считаются слова со следующими характеристиками: нестандартным для данного П-класса окончанием (слова русского языка «*города*», но «*заводы*»); несовпадением М-класса и С-класса.

Морфологический компонент реализован в системе ПЛЭНЕР-БЭСМ, а статьи в СОС имеют ЛИСП-подобные структуры. Например, в статье СОС неизменяемого слова русского языка «*инкогнито*» описываются три варианта значения слова [10]: (((7 2 3 0 0) = пребывание под вымышленным именем =); ((7 1 1 0 0) = лицо, скрывающее свое имя =); ((1) = наречие =)).

Словарная статья для основы «*ед-*» слова «*ехать*» имеет следующий вид: ((1 1 5 0 0 0) * (1 1 1) * (1 1 3 0 0 0)

(ЕХА (6 9 1 2 3 0 0 0)) (1 2 6 0 0) ЕХАТЬ).

Числа в приведенном примере являются морфологическими характеристиками слова. Они показывают, что у слова отсутствуют формы повелительного наклонения, деепричастия настоящего времени; при образовании форм прошедшего времени используется словарная статья основы «*еха-*»; в неопределенной форме употребляется словоформа «*ехать*».

В СОС могут присутствовать несколько основ одного слова, связанные между собой ссылками. Например, для одной из основ слова «*цыпленок*» – «*цыплят-*» статья имеет следующий вид [10]:

((7 1 1 2 0) ((1 1 1 0) (ЦЫПЛЕНК((3 1 (())) 7 1 1 1 0))).

Здесь первый элемент списка указывает принадлежность слова к мужскому роду, а второй – обозначает ссылку на соответствующий вариант основы форм единственного числа.

В каждой статье словаря окончаний указывается собственно окончание и соответствующие ему номера М-классов, П-классов и тип чередования.

Алгоритм определения словоформы F заключается в разбиении формы слова на основу S и окончание, применении чередований и поиске полученной основы в СОС. Алгоритм генерации форм F по данной основе S и грамматическому значению G состоит в поиске основы в СОС, извлечении информации об основе из СОС и непосредственной генерации словоформы.

В настоящее время существуют реализации этого метода с помощью современных средств программирования и баз данных (БД) [20].

Преимущества метода: позволяет описывать любые изменения в синтетических формах слов; простота реализации алгоритма для слов с неизменяемыми основами. Недостатки метода: сложный формат статей в словарях; структура словарей не универсальна: для синтеза форм слов необходимо преобразовать словарь окончаний; СОС содержит несколько основ одного слова; реализация только для русского языка; сложность в обслуживании словарей; ориентация на флективный анализ по окончанию; отсутствие возможности обработки аналитических форм слов.

А.М. Андреев и коллектив авторов разработали блок морфологического анализа [2] (конец 1990-х гг.) лингвистического процессора информационно-поисковой системы. Морфологический анализ начинается со сравнения словоформы и основ в СОС по первым буквам. Для каждой найденной основы извлекается соответствующая ей строка аффиксов из словаря аффиксов. В стандартном понимании аффикс – это часть слова, видоизменяющая лексическое или грамматическое его значение.

Каждый аффикс из этой строки поочередно присоединяется к основе, и результат сравнивается с анализируемой словоформой. В случае их точного совпадения очередная запись добавляется в список результатов поиска. При этом по порядковому номеру аффикса в строке аффиксов определяется грамматическое значение, а по словарной информации данной основы – другие характеристики (например, для существительного – род и одушевленность).

Если в результате такого поиска не найдено ни одного совпавшего варианта, то проводится поиск среди исключений, хранящихся также в СОС и имеющих ссылку на словарь исключений.

При неудачном результате поиска среди исключений проверяется наличие у анализируемого слова постфикса (возвратного суффикса) *-ся*, *-сь* или приставок *не-*, *ни-*. Если данные суффиксы и приставки присутствуют у словоформы, то они отделяются от анализируемого слова, и затем процедура поиска повторяется сначала. При этом грамматические значения анализируемых основ изменяются специальной процедурой.

В случае, когда все этапы поиска дали отрицательный результат (не найдено ни одного варианта), пользователю предлагается ввести новую основу в словарь.

Преимущества метода: простота реализации алгоритма морфологического анализа; простота структуры словарей; возможность занести слова, имеющие особый тип формообразования, как слова-исключения в СОС; возможность определения словоформ, отсутствующих в словарях системы. Недостатки метода: возможен только анализ, но не синтез; СОС содержит несколько основ одного слова; реализация только для русского языка; ориентация на флективный анализ по окончанию; отсутствие возможности обработки аналитических форм слов.

В.З. Демьянков разработал экспертную систему, которая позволяет анализировать и синтезировать словоформы на основе введенных знаний эксперта [5, 6] (начало 1990-х гг.). Декларативная часть экспертной системы состоит из следующих словарей. 1. Лексикон

(словарь морфем), каждая статья которого содержит следующие данные: «заглавный морф» морфемы; набор номеров чередований, присущих данной морфеме; имя класса морфемы – квалификация морфемы как корня, суффикса, префикса и т.п.; указания на два набора чередований; соответственно, соседних слева и справа морфем; набор указаний на условия, которым должны отвечать морфемы слева и/или справа от данной морфемы в рамках интерпретируемого слова; номер первой основы по СОС, оканчивающейся на какой-либо альтернант данной морфемы. 2. СОС содержит только непроизводные основы, грамматические значения которых не «вычисляются» из составных частей. 3. Словарь флексий. Флексии представлены в русском языке окончаниями, но в семитских языках могут присоединяться в начало словоформы. Поэтому для русского языка эксперт может задать, что префикс «*наи-*» (как в форме «*наибольший*») является флексией. Экспертная система не исключает этой возможности. 4. Словарь постфиксов содержит описания единиц, следующих после всех флексий, но являющихся регулярными для всех словоформ. Так, в лексеме «*умываться*» выделяется возвратный постфикс «*-ся*», присутствующий в каждой словоформе парадигмы. Также постфиксами являются единицы «*-таки*», «*-нибудь*», «*-то*».

Приведем пример морфологического анализа из работы [5].

Введем следующие обозначения классов морфемы:

* – корень; # – флексия; _ – префикс; $ – постфикс; = – суффикс.

Словоформа «*вынашивать*» на первых этапах выдвижения гипотез будет интерпретироваться следующим образом: 1. *В_ЫНАШИВАТЬ*. 2. *ВЫ_НАШИВАТЬ*. В рамках гипотезы 1 имеем для *ЫНАШИВАТЬ*: 1.1. *ЫН*АШИВАТЬ* (где «*ын*» – альтернант морфемы «*ин*», как в слове «*иной*»), и т.д., в результате чего, скажем, получится такое разбиение: *В_ЫН*А=ШИ*В=АТЬ#*. В рамках же гипотезы 2 имеем: 2.1. *НА_ШИВАТЬ*. 2.2. *НАШ*ИВАТЬ*. Далее в рамках подгипотезы 2.1 имеем: 2.1.1. *Ш*ИВАТЬ*, а в рамках подгипотезы 2.2: 2.2.1. *ИВА=ТЬ*. Теперь сопоставим хотя бы две наиболее вероятные гипотезы: *ВЫ_НА_Ш*ИВА=ТЬ#* (корень *Ш* тот же, что и у глагола «*шла*»), *ВЫ_НАШ*ИВА=ТЬ#*. Алгоритм выбирает вторую гипотезу, так как строка содержит меньше морфем, чем первая и возвращает ее в качестве результата работы.

Преимущества метода: в качестве аффиксов могут использоваться как окончания, так и приставки; возможность реализации не только для русского языка. Недостатки метода: сложность алгоритма морфологического анализа; сложная структура словарей системы; сложность заполнения словарей системы; СОС содержит несколько основ одного слова; отсутствие возможности обработки аналитических форм слов.

Д.Е. Шуклин предложил для решения задач морфологического анализа текста и анализа словоизменения использовать семантическую нейронную сеть [21] (начало 2000-х гг.). В качестве структуры такой сети, выполняющей морфологический анализ, применяется синхронизированное линейное дерево. Автор так описывает свой метод.

«Линейное дерево состоит из подслоев нейронов. Каждому синхронизированному подслою соответствует фронт волны обработки. Нейроны первого подслоя соответствуют первой букве слова, второго – второй и так далее. Общее количество подслоев равно максимальному количеству букв в одном слове. Первый подслой состоит из нейронов, распознающих первую букву, второй слой состоит из нейронов, распознающих первые две буквы, третий – первые три буквы. Упрощенный фрагмент синхронизированного линейного дерева, распознающего слова «*мама*», «*маме*», «*машина*» и «*машине*», представлен на рис. 1.

Каждый нейрон в упрощенном синхронизированном линейном дереве является синхронизированным конъюнктором и имеет одну входную связь с нейроном из предыдущего подслоя (соответствующим предыдущей букве слова) и одну входную связь с нейроном из

слоя рецепторов, соответствующим текущей букве. Каждый нейрон может иметь выходную связь с неограниченным количеством нейронов из следующего подслоя обработки» [21].

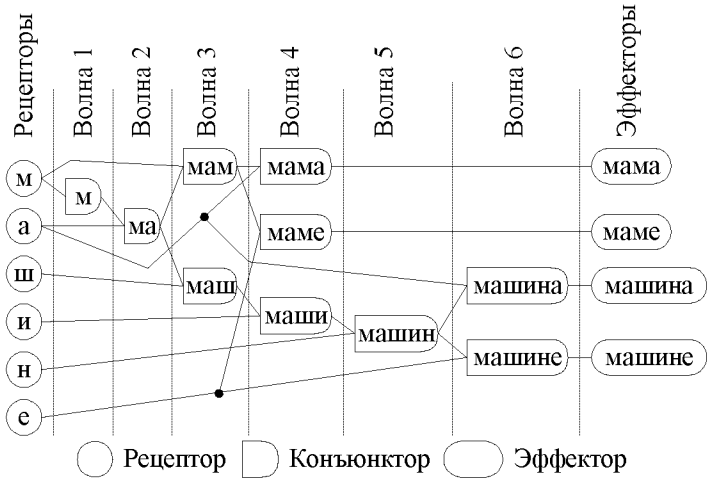

Рис. 1. Структура нейронной сети морфологического разбора слов «*мама*», «*маме*», «*машина*», «*машине*»

«Различным вариантам состояния входного потока символов в нейронной сети соответствуют различные варианты состояния этой сети. Результатом анализа, извлеченным из обработанной части текста в течение одного кванта времени, является мгновенное состояние синхронизированного линейного дерева. Мгновенное состояние включает в себя мгновенный снимок множества нейронов, множества связей между нейронами и множества внутренних состояний нейронов. Количество нейронов в сети ограничено, нейроны имеют конечное число состояний и связей, поэтому слой анализа текста в виде синхронизированного линейного дерева представляет конечный автомат. Переход из одного состояния в другое происходит при подаче на слой разбора текста очередного символа входной последовательности» [21].

Преимущества метода: возможность использования математического аппарата нейронных сетей для морфологического и синтаксического анализа; возможность описывать большое число типов изменений при формообразовании; в качестве аффиксов могут использоваться как окончания, так и приставки; возможность реализации не только для русского языка. Недостатки метода: необходимость построения нейронной сети сложной структуры для простых преобразований; при увеличении количества обрабатываемых слов значительно увеличивается количество нейронов и сложность структуры нейронной сети; отсутствие возможности синтеза форм слов; отсутствие возможности обработки аналитических форм слов.

И.М. Ножов в своей диссертационной работе предложил морфологический анализ без словаря [13] (конец 1990-х гг.) для системы автоматизированной индексации документов. Алгоритм анализа основан на самообучении программы на массивах текстов и совмещает два подхода: лингвистический (формализованная грамматика для построения морфологических гипотез); математический (метод корреляции, позволяющий унифицировать морфологическую гипотезу).

Морфологический анализ без словаря применяется для индексирования текстовой БД, в результате чего строится грамматический СОС и связанный с ним индекс документов, представляющий собой список слов и ссылки на тексты, в которых эти слова встречаются. Этот индекс предназначен для организации полнотекстового поиска по БД.

По результатам морфологического анализа словоформы соотносятся с той или иной лексемой; словоформы одной лексемы объединяются в отдельные классы; однозначно определяются грамматические значения словоформ и индексируются тексты по встретившимся в них основам.

Для морфологического анализа используется минимальный объем исходной информации: «таблица предлогов» и «таблица местоимений и числительных, имеющих нерегулярное склонение».

Результаты морфологического анализа объединяются в СОС данной БД, каждая статья которой задается тройкой значений [основа, часть речи, парадигматический класс]. Морфологический анализатор состоит из трех модулей: 1) статический массив флексий и правила формализованной грамматики русской морфологии, построенной на основе грамматического словаря А.А. Зализняка [7]; 2) модуль, использующий правила формализованной грамматики для построения морфологического дерева словоформы, в узлах которого хранятся все возможные гипотезы об основах и грамматических значениях словоформы; 3) метод, позволяющий отнести словоформу к лексеме и отсечь неверные варианты.

Преимущества метода: простота словарей системы, их минимальный объем; определение словоформ, отсутствующих в словарях системы; реализация методов во втором и третьем модулях, не зависящих от языка. Недостатки метода: возможен только анализ, но не синтез; правильный результата анализа может быть получен с вероятностью меньшей единицы; результат морфологического анализа не позволяет проводить в дальнейшем семантический анализ; трудоемкое построение в ходе анализа «леса» – нескольких деревьев и матричные вычисления для отсеивания неправильных вариантов; ориентация на флективный анализ по окончанию; отсутствие возможности обработки аналитических форм слов.

Рассмотрим методы морфологического анализа и синтеза, предложенные зарубежными специалистами и применимые, в том числе, и для словоформ русского языка.

Киммо Коскенниеми из Университета Хельсинки (Финляндия) предложил двухуровневую модель для определения и генерации форм слов [25, 26] (начало 1980-х гг.). Модель включает два уровня представления: лексический и поверхностный, между которыми вводятся правила (соответствия). Лексический уровень формируется путем применения правил естественного языка без учета контекста, а поверхностный уровень учитывает контекст, в котором эти правила используются.

Рассмотрим данную двухуровневую модель и задачу представления в ней каждого уровня на примере из работы [26]. Пусть необходимо получить форму множественного числа падежа партитив (*Partitiivi*, *part.*) существительного финского языка «*lasi*» – «*стекло*». Для этого на лексическом уровне к основе «*lasi*» добавляются показатели множественного числа «*I*» и партитива «*a*». На поверхностном уровне используются контекстные правила финского языка: 1) окончание основы «*i*» перед показателем множественного числа «*I*» заменяется на «*e*»; 2) показатель множественного числа «*I*», стоящий между гласными заменяется на «*j*».

Лексический уровень	l	a	s	I	I	a
Поверхностный уровень	l	a	s	E	j	a

Построим автомат, реализующий первое правило. На вход автомата (рис. 2) последовательно поступают пары с лексического и поверхностного уровней: l-l, a-a, s-s, i-e, I-j, a-a.

Чередование «*i-e*» будет разрешено, когда за «*i*» следует показатель множественного числа «*I*». Запись «=-=» означает любые другие пары.

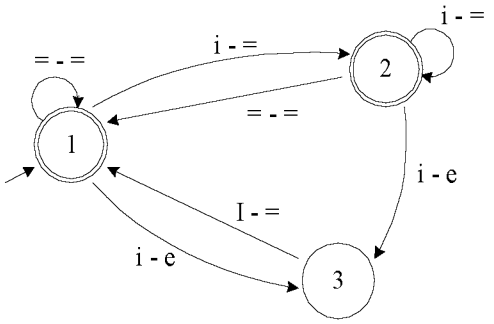

Рис. 2. Автомат для реализации чередования «*i-e*»

Данному автомату соответствует табличное представление (табл. 1).

Таблица 1

Табличное представление автомата для реализации чередования «i-e»

	i =	i e	I =	= =
1	2	3	0	1
2	2	3	0	1
3	0	0	1	0

На основе двухуровневой модели разработана система морфологического анализа и синтеза форм слов PC-KIMMO [23, 28]. Для каждого естественного языка создается файл правил, содержащий алфавит и фонологические правила, файл лексики со списком лексических объектов (слов или морфем) и еще ряд вспомогательных файлов.

Метод, предложенный К. Коскенниеми, – это пример элементно-операционной модели морфологии.

Преимущества метода: позволяет описывать любые изменения в форме слова; реализация для языков со сложным изменением основ: японского, русского, финского языков, а также для татарского языка [19]; универсальная структура словарей для анализа и синтеза. Недостатки метода: для каждого типа изменения форм слов необходимо разрабатывать и описывать автомат, проверять правильность его работы, что требует знания теории автоматов; отсутствие возможности обработки аналитических форм слов.

Мартин Портер разработал алгоритм стемминга [29] (от англ. *stem* – основа) (конец 1970-х гг.), который заключается в отделении от словоформы суффиксов и окончаний и получении основы для ее дальнейшей обработки. По Портеру, основу составляют корень и приставка. Метод стемминга позиционируется автором как чисто алгоритмический, в отличие от словарных методов, описанных выше. В то же время в программной реализации суффиксы и окончания присутствуют в программном коде, хотя рациональней было бы хранить их в отдельном словаре. Предложенный метод получения основы реализован в системе (стеммере) Snowball [32].

Преимущества метода: простота реализации алгоритма морфологического анализа; простая структура словарей; реализован для основных европейских языков. Недостатки метода: возможен только анализ, но не синтез; необходимо определять порядок отделения суффиксов и окончаний; при добавлении или модификации отделяемых суффиксов и окончаний необходимо перекомпилировать программный код; ориентация на флективный анализ по окончанию; отсутствие возможности обработки аналитических форм слов.

Джон Голдсмит из Чикагского университета (США) предложил алгоритм морфологического анализа без словаря (середина 1990-х гг.) [24]. Перед началом анализа необходимо выделить в языке сигнатуры – наиболее часто встречающаяся последовательность букв в словах. Сигнатурой может быть основа, префикс или суффикс. Выделение сигнатур происходит на основе текста, содержащего более 50 тысяч слов данного естественного языка.

Алгоритм выделения основ и суффиксов состоит из следующих шагов. 1. Произвести считывание всех слов текста и запомнить их всевозможные разбиения на основу и суффикс. 2. Все возможные основы записать в массив основ, а все возможные суффиксы – в массив суффиксов. 3. Для всех основ и суффиксов подсчитать количество их вхождений в разбиения слов. 4. Для всех вариантов разбиения одного слова на основу T и суффикс S подсчитать значение функции V: $V(T/S) = |T| \cdot \ln \langle T \rangle + |S| \cdot \ln \langle S \rangle$, где $|T|$, $|S|$ – длины основы и суффикса в буквах; $\langle T \rangle$, $\langle S \rangle$ – число вхождений основы и суффикса в разбиения. 5. Выбрать разбиения с максимальным значением функции V. 6. Отбросить основы, которые встречаются только с одним суффиксом, и суффиксы, которые встречаются только с одной основой. 7. Среди полученных основ выделить префиксы. 8. Сгруппировать основы по категориям в зависимости от сочетания с группой суффиксов. 9. Сформировать на основе полученных основ и суффиксов список сигнатур.

Сигнатуры позволяют быстро разбивать слова на суффиксы и основы, получать их грамматические значения.

Метод Дж. Голдсмита является типичным представителем элементно-комбинаторной модели морфологии.

Преимущества метода: простота реализации алгоритма морфологического анализа; отсутствие словарей; простой способ хранения данных; реализован для основных европейских языков. Недостатки метода: возможен только анализ, но не синтез; правильный результат анализа может быть получен с вероятностью меньшей единицы; результаты разбиения необходимо корректировать вручную; ориентация на флективный анализ по окончанию; отсутствие возможности обработки аналитических форм слов.

Информационно-поисковые системы в Интернете (например, Яндекс и Google) работают по принципу стеммера: от словоформы из поискового запроса отделяется окончание, и по полученной основе происходит поиск страниц, на которых эта основа встречается. Морфологические анализаторы информационно-поисковых систем не распознают аналитические формы слов, так как работают на втором уровне морфологического анализа – определении только основы слов.

Рассмотрим работу морфологического анализатора информационно-поисковой системы Яндекс [18, 31], разработанного И. Сегаловичем, М. Масловым (конец 1990-х гг.).

Алгоритм морфологического анализа основан на словарном анализе со словарями основ и окончаний, использующем электронный морфологический словарь лингвистического процессора ЭТАП [9], разработанного Лабораторией компьютерной лингвистики Института проблем передачи информации Российской академии наук. Объем словаря составляет около 90 тыс. лексем (120 тыс. словооснов) [18].

Алгоритм взаимодействует с декларативной частью, состоящей из следующих частей. 1. Список основ, описываемых следующими ключами: {основа, идентификатор лексемы, номер основы в лексеме}. Основы упорядочены в алфавитном порядке, начиная с конца слов, как в словаре А.А. Зализняка [7]. Основы слов с одинаковым или сходным типом формообразования в этом списке, как правило, находятся рядом. Лексемы могут иметь одну или несколько основ. 2. Список окончаний всех лексем словаря в алфавитном порядке, начиная с конца слов.

На вход алгоритма поступают лексема, основа и флексия (окончание) анализируемой формы. На выходе алгоритм возвращает количество вариантов разбора и наборы грамматических значений по каждому варианту разбора.

Алгоритм морфологического анализа состоит из следующих шагов. 1. Отделить от анализируемой словоформы все возможные окончания для получения вариантов основ. 2. Найти в СОС варианты основы, начиная с самого длинного. Если вариант основы в этом списке отсутствует, то найти «наиболее близкие» словарные основы, имеющие максимальный по длине общий «хвост». Запомнить позицию первой «наиболее близкой» основы и меру ее сходства. В качестве последней используется число совпавших символов в основе и длина окончания. 3. Сравнить вариант основы с одной из «ближайших» словарных основ. Если основы не равны, то анализируемое слово с данным вариантом основы в словаре отсутствует. В этом случае по варианту основы, окончанию и лексеме, соответствующей «ближайшей» словарной основе, генерируется гипотетическая лексема – модель формообразования для этого неизвестного слова. В случае успешной генерации эта гипотеза подается на вход морфологического анализатора. Успешные варианты разбора запоминаются в следующем виде: {лексема, варианты разбора}. Если среди лексем с одинаковой мерой сходства есть хотя бы один вариант разбора, то перейти к пункту 5 с положительным результатом. Если вариантов разбора нет, то уменьшить длину требуемого общего «хвоста» основы. Если после этого длина требуемого общего «хвоста» основы стала меньше двух, то перейти к пункту 5 с отрицательным результатом, иначе – перейти к пункту 3. 4. Унифицировать гипотезы по парадигмам и отфильтровать ложные гипотезы. 5. Закончить алгоритм.

Синтез форм слов не используется в информационно-поисковых системах.

Преимущества метода: простота реализации алгоритма морфологического анализа; простота словарей системы; возможность определения словоформ, отсутствующих в словаре. Недостатки метода: используется только для анализа форм слов; реализация только для русского языка; ориентация на флективный анализ по окончанию; отсутствие возможности обработки аналитических форм слов.

Заключение. Несмотря на большое число подходов к морфологической обработке текста, ни один из них не стал классическим, и задачи генерации и определения остаются интеллектуальными задачами.

Проанализировав преимущества и недостатки рассмотренных подходов к обработке текстов на морфологическом уровне, можно сформулировать следующее требование к разрабатываемому методу генерации и определения форм слов. Он должен быть универсальным по следующим критериям: К1) универсальность генерации и определения форм слов естественных языков различных групп и семейств; К2) универсальность структуры словарей, не требующей конвертации для решения задач определения или генерации; К3) универсальность метода для всех видов формообразования словоформ (любых видов аффиксов), обработки всей парадигмы слова, в том числе и аналитических словоформ.

Рассмотренные разработки в области АОТ на морфологическом уровне удовлетворяют лишь некоторым критериям универсальности (табл. 2).

Таблица 2

Соответствие существующих методов морфологической обработки текстов критериям универсальности

	Критерии		
Автор метода (алгоритма)	**К1**	**К2**	**К3**
Сегалович И.	–	±	–
Голдсмит Дж.	-	±	-
Портер М.	-	±	-
Коскенниеми К.	±	+	±
Ножов И.М.	-	±	-
Шуклин Д.Е.	-	±	-
Демьянков В.З.	-	+	±
Андреев А.М.	-	+	-
Мальковский М.Г	-	+	-
Белоногов Г.Г.	-	–	-

Примечание: «+» – полное соответствие; «–» – полное несоответствие; «±» – частичное соответствие.

Выводы. 1. На основе представленного в данной статье подробного анализа существующих подходов был разработан новый метод генерации и определения форм слов [15]. Он лишен недостатков, выявленных в существующих подходах, и соответствует всем трем перечисленным выше критериям универсальности. 2. Предложенный метод генерации и определения форм слов используется в электронном русско-английском математическом словаре [12], системе анализа коллективных договоров [1] и системе проверки знаний формообразования естественных языков [17].

Список литературы

1. Александров В. В. Автоматизированный анализ и оценка статей коллективных договоров / В. В. Александров , Н. П. Макаров , А. С. Шустов // Вестник Рязанского государственного радиотехнического университета. – 2013. – № 45. – С. 71–75.

2. Андреев А. М. Лингвистический процессор для информационно-поисковой системы / А. М. Андреев , Д. В. Березкин , А. В. Брик // Компьютерная хроника. – 1998. – № 11. – С. 79–100.

3. Бабина О. И. Автоматический морфологический анализ флективных языков / О. И. Бабина , Н. Ю. Дюмин // Наука ЮурГУ : мат-лы 62-й науч. конф. – Челябинск : Издательский центр ЮУрГУ, 2010. – Т. 2. – С. 35–38.

4. Белоногов Г. Г. Автоматизированные информационные системы / Г. Г. Белоногов, В. И. Богатырев ; под ред. К. В. Тараканова. – Москва : Сов. радио, 1973. – 328 с.

5. Демьянков В. З. Морфологическая интерпретация текста и ее моделирование / В. З. Демьянков. – Москва : Изд-во МГУ, 1994. – 206 с.

6. Демьянков В. З. Основы теории интерпретации и ее приложения в вычислительной лингвистике / В. З. Демьянков. – Москва : Изд-во Моск. ун-та, 1985. – 76 с.

7. Зализняк А. А. Грамматический словарь русского языка: ок. 100 000 слов / А. А. Зализняк. – Москва : Русский язык, 1977. – 880 с.

8. Ивин А. А. По законам логики / А. А. Ивин. – Москва : Мол. гвардия, 1983. – 208 с.

9. Лингвистическое обеспечение системы ЭТАП-2 / Ю. Д. Апресян, И. М. Богуславский, Л. Л. Иомдин и др. – Москва : Наука, 1989. – 296 с.

10. Мальковский М. Г. Диалог с системой искусственного интеллекта / М. Г. Мальковский. – Москва : Изд-во МГУ, 1985. – 214 с.

11. Мальковский М. Г. Анализатор системы TULIPS-2. Морфологический уровень / М. Г. Мальковский, И. А. Волкова // Вестник Моск. ун-та. Сер. Вычислительная математика и кибернетика. – 1981. – № 1. – С. 70–76.

12. Миронов В. В. Электронная информационно-поисковая система «Русско-английский математический словарь» / В. В. Миронов , А. И. Заволокин , А. К. Розанов // Информатизация образования и науки. – 2013. – № 3 (19). – С. 167–176.

13. Ножов И. М. Прикладной морфологический анализ без словаря / И. М. Ножов // Тр. конф. по искусственному интеллекту КИИ-2000. – Москва : Физматлит, 2000. – Т. 1. – С. 424–429.

14. Пруцков А. В. Алгебраическое представление модели формообразования естественных языков / А. В. Пруцков // Cloud Of Science. – 2014. – Т. 1, № 1. – С. 88–97.

15. Пруцков А. В. Генерация и определения форм слов естественных языков на основе их последовательных преобразований / А. В. Пруцков // Вестник Рязанского государственного радиотехнического университета. – 2009. – № 27. – С. 51–58.

16. Пруцков А. В. Определение и генерация сложных форм слов естественных языков при морфологическом анализе и синтезе / А. В. Пруцков // Известия Таганрогского государственного радиотехнического университета. – 2006. – Т. 70, № 15. – С. 10–14.

17. Пруцков А. В. Статический и динамический подходы к проектированию подсистем проверки знаний автоматизированных обучающих систем / А. В. Пруцков // Информационные ресурсы России. – 2006. – № 1. – С. 27–29.

18. Сегалович И. Русский морфологический анализ и синтез с генерацией моделей словоизменения для неописанных в словаре слов / И. Сегалович , М. Маслов // Диалог-98 : тр. Междунар. сем. по компьютерной лингвистике и ее приложениям. – Москва, 1998. – Т. 2. – С. 547–552.

19. Сулейманов Д. Ш. Двухуровневое описание морфологии татарского языка / Д. Ш. Сулейманов, А. А. Гильмуллин, Р. А. Гильмуллин // Языковая семантика и образ мира : тез. Междунар. науч. конф. – Казань : Изд-во КГУ, 1997. – Кн. 2. – С. 65–67.

20. Хахалин Г. К. Комплекс по разработке индивидуальных и/или корпоративных электронных толковых словарей / Г. К. Хахалин , Н. К. Богданов , С. В. Платонов // Обработка текста и когнитивные технологии. Когнитивное моделирование : тр. Междунар. конф. 1999 г. / МИСИС. – Москва, 2000. – Ч. 2. – С. 350–363.

21. Шуклин Д. Е. Морфологический и синтаксический разбор текстов как конечный автомат, реализованный семантической нейронной сетью, имеющей структуру синхронизированного линейного дерева // Новые информационные технологии : мат-лы 5-го науч.-практ. сем. / МГИЭМ – Москва, 2002. – С. 74–85.

22. Языкознание. Бол. энцикл. словарь / гл. ред. В. Н. Ярцева. – 2-е изд. – Москва : Бол. рос. энцикл., 1998. – 685 с.

23. Antworth E. L. PC-KIMMO: A Two-level Processor for Morphological Analysis. Number 16 in Occasional publications in academic computing / E. L. Antworth. – Dallas : Summer Institute of Linguistics, 1990.

24. Goldsmith J. Unsupervised Learning of the Morphology of a Natural Language / J. Goldsmith. – Chicago : University of Chicago Press, 1998. – P. 173–194.

25. Karttunen L. Short History of Two-Level Morphology / L. Karttunen , K. R. Beesley. – Xerox Palo Alto Research Center, Palo Alto, CA, 2001.

26. Koskenniemi K. Two-level Morphology: A General Computational Model for Word-form Recognition and Production / K. Koskenniemi. – University of Helsinki, Department of General Linguistics, 1983. – Publications No. 11.

27. Matthews P. H. Morphology / P. H. Matthews. – 2nd ed. – Cambridge : Cambridge University Press, 1991. – 251 p. – (Cambridge textbooks in linguistics).

28. PC-KIMMO Version 2. – Available at: http://www.sil.org/pckimmo/v2/index.html (accessed 10.04.2004).

29. Porter M. F. An algorithm for suffix stripping / M. F. Porter // Program. – 1980. – July. – Vol. 14, no. 3. – P. 130–137.

30. Prutskov A. V. Algorithmic Provision of a Universal Method for Word-Form Generation and Recognition / A. V. Prutskov // Automatic Documentation and Mathematical Linguistics. – 2011. – Vol. 45, no. 5. – P. 232–238.

31. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine / I. Segalovich // MLMTA-2003. – Las Vegas, 2003. – June.

32. Snowball. – Available at: http://snowball.sourceforge.net (accessed 15.05.2004).

33. Trask R. L. A Dictionary of Grammatical Terms in Linguistics / R. L. Trask. – London and New York : Routledge, 1993. – 352 p.

References

1. Aleksandrov V. V., Makarov N. P., Shustov A. S. Avtomatizirovannyy analiz i otsenka statey kollektivnykh dogovorov [Automated analysis and estimation of articles of collective treaty]. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta* [Bulletin of Ryazan State Radioengineering University], 2013, no. 45, pp. 71–75.

2. Andreev A. M., Berezkin D. V., Brik A. V. Lingvisticheskiy processor dlya informatsionno-poiskovoy sistemy [Linguistic processor for a information-searching system]. *Kompyuternaya khronika* [Computer Chronicle], 1998, no. 11, pp. 79–100.

3. Babina O. I., Dyumin N. Yu. Avtomaticheskiy morfologicheskiy analiz flektivnykh yazykov [Automatic morphological analysis of inflexional languages]. *Nauka Yuzhno-Uralskogo gumanitarnogo universiteta : materialy 62-y nauchnoy konferentsii* [Science of The South Ural State University: Proceedings of the 62nd Scienctific Confernce]. Chelyabinsk, Publishing Center of the South Ural State University, 2010, vol. 2, pp. 35–38.

4. Belonogov G. G., Bogatyrev V. I. *Avtomatizirovannye informatsionnye sistemy* [Automated information system]. Moscow, Sovetskoe radio, 1973. 328 p.

5. Demyankov V. Z. *Morfologicheskaya interpretatsiya teksta i ee modelirovanie* [Morphological text interpretation and its modelling]. Moscow, Moscow State University Publ, 1994. 206 p.

6. Demyankov V. Z. *Osnovy teorii interpretatsii i ee prilozheniya v vychislitelnoy lingvistike* [The interpretation theory fundamentals and its application in computational lingusitics]. Moscow, Moscow State University Publ., 1985. 76 p.

7. Zaliznyak A. A. *Grammaticheskiy slovar russkogo yazyka: okolo 100 000 slov* [The grammatical dictionary of the Russian language]. Moscow, Russkiy Yazyk, 1977. 880 p.

8. Ivin A. A. *Po zakonam logiki* [By The Laws of Logic]. Moscow, The Young Guard, 1983. 208 p.

9. Apresyan Yu. D., Boguslavskiyu I. M., Iomdin L. L. et al. *Lingvisticheskoe obespechenie sistemy ETAP-2* [Linguistic provision of ETAP-2 system]. Moscow, Nauka, 1989. 296 p.

10. Malkovskiy M. G. *Dialog s sistemoy iskusstvennogo intellekta* [Dialogue with artificial intellegence system]. Moscow, Moscow State University Publ., 1985. 214 p.

11. Malkovskiy M. G., Volkova I. A. Analizator sistemy TULIPS-2. Morfologicheskiy uroven [The analisator of TULIPS-2 system. Morphological level]. *Vestnik Moskovskogo universiteta. Seriya. Vychislitelnaya matematika I kibernetika* [Bulletin of Moscow University. Series. Computational Mathematics and Cybernetics], 1981, no. 1, pp. 70–76.

12. Mironov V. V., Zavolokin A. I., Rozanov A. K. Elektronnaya informatsionno-poiskovaya sistema «Russko-angliyskiy matematicheskiy slovar» [Electronic information searching system «Russian-English mathematical dictionary»]. *Informatizatsiya obrazovaniya i nauki* [Informatization of Education and Science], 2013, no. 3 (19), pp. 167–176.

13. Nozhov I. M. Prikladnoy morfologicheskiy analiz bez slovarya [Applied morphological analysis without dictionary]. *Trudy konferentsii po iskusstvennomu intellektu KII-2000* [Proceedings of Artificial Intellegence Conference KII-2000]. Moscow, Fizmatlit, 2000, vol. 1, pp. 424–429.

14. Prutzkov A. V. Algebraicheskoe predstavlenie modeli formoobrazovaniya yestestvennykh yazykov [An algebraic representation of the natural language word-form building model]. *Cloud of Science*, 2014, vol. 1, no. 1, pp. 88–97.

15. Prutzkov A. V. Generatsiya i opredeleniya form slov estestvennykh yazykov na osnove ikh posledovatelnykh preobrazovaniy [Generation and recognition of the natural languages word-forms based on consequent words transformation]. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta* [Bulletin of Ryazan State Radioengineering University], 2009, no. 27. pp. 51–58.

16. Prutzkov A. V. Opredelenie i generatsiya slozhnykh form slov estestvennykh yazykov pri morfologicheskom analize i sinteze [Recognition and Generation of natural languge compound word-forms for morphological analysis and synthesis]. *Izvestiya Taganrogskogo gosudarstvennogo radiotekhnicheskogo universiteta* [Proceedings of Taganrog State Radioengineering University], 2006, vol. 70, no. 15. pp. 10–14.

17. Prutzkov A. V. Staticheskiy i dinamicheskiy podkhody k proektirovaniyu podsistem proverki znaniy avtomatizirovannykh obuchayushchikh sistem [Static and dynamic ways to designing of testing subsystem of automated learning system]. *Informatsionnye resursy Rossii* [Information resources of Russia], 2006, no. 1, pp. 27–29.

18. Segalovich I., Maslov M. Russkiy morfologicheskiy analiz i sintez s generatsiey modeley slovoizmeneniya dlya neopisannykh v slovare slov [Russian morphological analysis and synthesis with generation of word-changing model for unknown words]. *Dialog-98 : trudy Mezhdunarodnogo seminara po kompyuternoy lingvistike i ee prilozheniyam* [Proceedings of International Conference of Computational Linguistika and its Application «Dialogue-98»], Moscow, 1998, vol. 2. pp. 547–552.

19. Suleymanov D. Sh., Gilmullin A. A., Gilmullin R. A. Dvukhurovnevoe opisanie morfologii tatarskogo yazyka [Two-level representation of the Tatar language morfology]. *Yazykovaya semantika i obraz mira: tezisy Mezhdunarodnogo nauchnogo konferentsii* [Language semantics and the world image: Proceedings of the International Conference]. Kazan, 1997, vol. 2, pp. 65–67.

20. Hahalin G. K., Bogdanov N. K., Platonov S. V. Kompleks po razrabotke individualnykh i/ili korporativnykh elektronnykh tolkovykh slovarey [Complex for development of individual and/or corporative explanatory dictionaries]. *Obrabotka teksta i kognitivnye tekhnologii. Kognitivnoe modelirovanie: trudy Mezhdunarodnoy konferentsii 1999 g.* [Text processing and cognitive technologies. Cognitive modeling: Proceedings of the International Conference]. Moscow, 1999, part 2, pp. 350–363.

21. Shuklin D. Ye. Morfologicheskiy i sintaksicheskiy razbor tekstov kak konechnyy avtomat, realizovannyy semanticheskoy neyronnoy setyu, imeyushhey strukturu sinkhronizirovannogo lineynogo dereva [Morphological and sintax text parsing as a finite automat realized by neuron net with synchronous linear tree structure]. *Novye informatsionnye tekhnologii: materialy 5-go nauchno-prakticheskogo seminara* [New information Technologies: Proceedings of the 5th Scientific-Practical Conference], Moscow, 2002. pp. 74–85.

22. *Yazykoznanie. Bolshoy entsiklopedicheskiy slovar* [Linguistics. The Big Encyclopaedia]. 2-nd ed. Moscow, Big Russian Encyclopaedia, 1998. 685 p.

23. Antworth E .L. *PC-KIMMO: A Two-level Processor for Morphological Analysis. Number 16 in Occasional publications in academic computing*. Dallas, Summer Institute of Linguistics, 1990.

24. Goldsmith J. *Unsupervised Learning of the Morphology of a Natural Language*. Chicago, University of Chicago Press, 1998, pp. 173–194.

25. Karttunen L., Beesley K. R. *A Short History of Two-Level Morphology*. Xerox Palo Alto Research Center, Palo Alto, CA. 2001.

26. Koskenniemi K. *Two-level Morphology: A General Computational Model for Word-form Recognition and Production*. University of Helsinki, Department of General Linguistics, 1983, Publications No. 11.

27. Matthews P. H. *Morphology*. 2nd ed. Cambridge, Cambridge University Press, 1991. 251 p. (Cambridge textbooks in linguistics).

28. PC-KIMMO Version 2. Available at: http://www.sil.org/pckimmo/v2/index.html (accessed 10.04.2004).

29. Porter M. F. An algorithm for suffix stripping. *Program*, 1980, July, vol. 14, no. 3, pp. 130–137.

30. Prutskov A. V. Algorithmic Provision of a Universal Method for Word-Form Generation and Recognition. *Automatic Documentation and Mathematical Linguistics*, 2011, vol. 45, no. 5, pp. 232–238.

31. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. *MLMTA-2003*. Las Vegas, 2003, June.

32. Snowball. Available at: http://snowball.sourceforge.net. (accessed 15.05.2004).

33. Trask R. L. *A Dictionary of Grammatical Terms in Linguistics*. London and New York, Routledge, 1993. 352 p.