УДК 004.912:811.512.111

НЕКОТОРЫЕ ВОПРОСЫ ПРОЕКТИРОВАНИЯ МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА ЧУВАШСКОГО ЯЗЫКА

Желтов В.П., Сергеев Е.С., Пушкин А.С., Скворцов А.В.

ФГБОУ ВО «Чувашский государственный университет имени И.Н. Ульянова», Чебоксары, e-mail: akaevgeniy@rambler.ru

Предметом исследования являются особенности разработки морфологических анализаторов естественных языков. Задачей морфологического анализатора (МА) чувашского языка является установление морфемного состава слов, а также морфологических признаков, используемых в задачах синтаксического и семантического анализаторов. База знаний морфологического анализа чувашского языка состоит из списка перечислимых типов (описывающих морфологические признаки), словарей (описывающих значение морфем), справочников (описывающих взаимосвязь морфем и признаков), структуры рабочей модели данных (для хранения текстов, слов и вариантов разбора слов) и метаправил (описывающих последовательность разбора слов разных частей речи). Цель работы: проектирование, разработка динамически подключаемой библиотеки (dll), предоставляющей набор методов и функций для морфологического анализа слов чувашского языка для внедрения в системы машинного перевода, лингвистических процессоров. Получена программная реализация правил для анализа слов-исключений. Разработанная модель является языконезависимой и может быть применена для русского и для чувашского языка. Приведен практический результат работы программы. Во всех случаях при наличии морфем в словаре программа производила точный анализ с правильным определением всех атрибутов.

Ключевые слова: морфологический анализ, компьютерная лингвистика, морфологический анализатор, чувашский язык, анализ текста, лингвистический корпус

SOME PROBLEMS OF DESIGNING THE MORPHOLOGICAL ANALYZER OF THE CHUVASH LANGUAGE

Zheltov V.P., Sergeev E.S., Pushkin A.S., Skvortsov A.V.

Federal State Budget Educational Institution of Higher Education «Chuvash State University named after I.N. Ulyanov», Cheboksary, e-mail: akaevgeniy@rambler.ru

The subject of the study are the features of the development of morphological analyzers of natural languages. The task of the morphological analyzer (MA) of the Chuvash language is the establishment of the morphemic composition of words, as well as the morphological features used in the tasks of the syntactic and semantic analyzers. The knowledge base of morphological analysis of the Chuvash language consists of a list of enumerated types (describing the morphological features), dictionaries (describing the meaning of morphemes), reference books (describing the relationship of morphemes and attributes), the structure of the working data model (for storing texts, words and word parsing options) and meta-rules (Describing the sequence of parsing the words of different parts of speech). The aim of the work: designing, developing a dynamically connected library (dll), which provides a set of methods and functions for the morphological analysis of the words of the Chuvash language for introduction into machine translation systems, linguistic processors. The program implementation of rules for the analysis of exception words is obtained. The developed model is language-independent and can be applied for Russian and for the Chuvash language. The practical result of the program is given. In all cases, in the presence of morphemes in the dictionary, the program produced an accurate analysis with the correct definition of all attributes.

Keywords: morphological analysis, computer linguistics, morphological analyzer, Chuvash language, text analysis, linguistic corpus

Задачей морфологического анализатора (МА) чувашского языка является установление морфемного состава слов, а также морфологических признаков, используемых в задачах синтаксического и семантического анализаторов.

База знаний морфологического анализа чувашского языка состоит из списка перечислимых типов (описывающих морфологические признаки), словарей (описывающих значение морфем), справочников (описывающих взаимосвязь морфем и признаков), структуры рабочей модели данных (для хранения текстов, слов и вариантов разбора слов) и метаправил (описывающих последовательность разбора слов разных частей речи).

Распределение морфологических характеристик по частям речи приведено в таблице.

Цель работы: проектирование, разработка динамически подключаемой библиотеки (dll), предоставляющей набор методов и функций для морфологического анализа слов чувашского языка для внедрения в системы машинного перевода, лингвистических процессоров.

Проект библиотеки морфологического анализатора состоит из семи классов (рис. 1).

Распределение морфологических характеристик

Существительное	Лицо, число, падеж, форма (1–17), одушевленность, время
Прилагательное	Лицо, число, падеж, форма (5, 12, 10, 3, 8, 17, 18), время
Числительное	Репрезентация, форма (4, 5, 10, 12, 10, 14, 2, 17), падеж
Местоимение	Репрезентация, форма (8, 14, 1, 2), падеж
Глагол	Репрезентация, вид, аспект, залог, время, падеж (В), форма (9)
Причастие	Вид, аспект, залог, время, форма (7, 1, 9, 11), падеж
Деепричастие	Вид, аспект, залог, время
Союз	Репрезентация, форма
Частица	Значение, форма
Послелог	Значение, форма

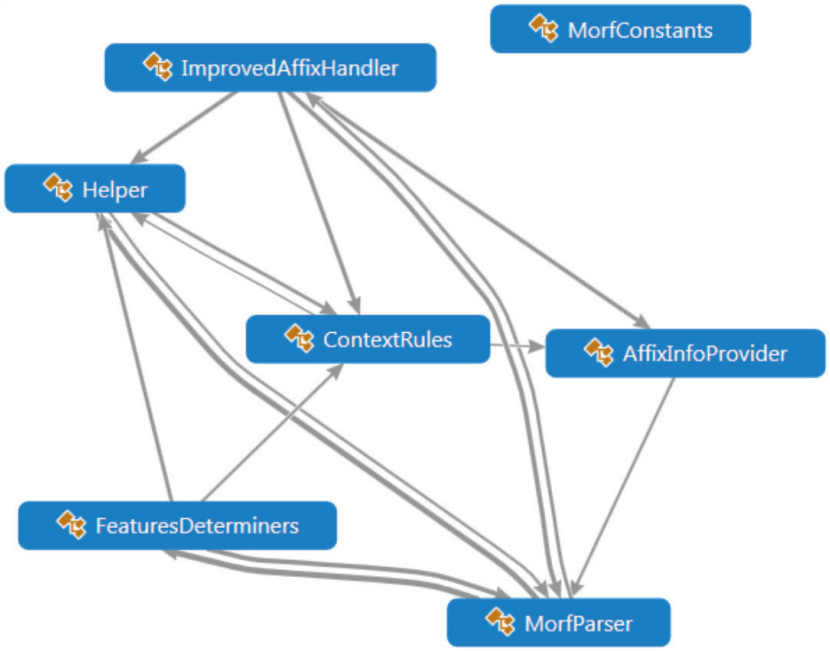

Рис. 1. Библиотеки морфологического анализатора чувашского языка

MorfParser.cs – Главный класс МА. Имеет два открытых метода для работы.

AffixInfoProvider.cs – Класс предоставляет набор полей и функций для работы с аффиксами.

ContextRules.cs – Класс отвечает за корректное определение правила восстановления согласно контексту.

FeaturesDeterminers.cs – Класс предоставляет набор функций для определения атрибутов (характеристик) слов.

Helper.cs – Класс содержит набор вспомогательных статических функций.

ImprovedAffixHandler – Класс принимает слово для разбиения его на составляющие и производит поиск в словаре.

MorfConstants – Класс представляет собой набор констант, введенных и использованных в проекте.

Требования к библиотеке [1]:

– Определение части речи слова и других морфологических характеристик

– Выделение корня и аффиксов.

– Распознавание контекстов.

– Анализ слов-исключений.

– Возможность вывода результатов в файл.

– Режим логгирования для удобной отладки.

Общий алгоритм:

Шаг 1. Начало.

Шаг 2. Разработка технического задания и составления критерия определения готовности работы.

Шаг 3. Выбор программы и языков создания библиотеки.

Шаг 4. Разработка алгоритмов решения задачи.

Шаг 5. Реализация функционала из технического задания.

Шаг 6. Тестирование всех функций библиотеки.

Шаг 7. Завершение работы.

Структура разработки представляет собой минимальный набор средств для работы приложения. Для использования библиотеки МА требуется платформа .NET Microsoft.

Структуру чувашских слов можно представить в виде суммы корней и аффиксов. Приставки и окончания, в отличие от русского языка, в чувашском языке отсутствуют, что упрощает разработку МА. Таким образом, для разработки требовался словарь основ (корней) и база аффиксов [2]. Исходный словарь представляет собой текстовый файл, в котором слова представлены следующим образом: слово, часть речи, информация об источнике. В нем собрано более тридцати одной тысячи слов чувашского языка.

В чувашском языке около ста семидесяти аффиксов. Исходная база аффиксов (БА) имела схожую структуру со словарем [3].

Рассмотрим пример. Слово «витресемпех», означающее «с ведрами же», раскладывается на следующие составляющие: «витре+сем+пе+х». Извлеченные аффиксы сохранят свой порядок и с другими словами («ачасемпех», «ĕçсемпех»). Из этого можно сделать вывод о возможности представления базы аффиксов в виде совокупности уровней. Где на каждом уровне будут храниться аффиксы, которые могут склеиваться с аффиксами, у которых уровни ниже. На основе ранее извлеченных аффиксов БА бы выглядела следующим образом: на первом уровне «сем», на втором «пе», на третьем «х». Вдобавок, каждый аффикс, помимо признака последовательности, хранит в себе характеристику типа. Согласно этому принципу, каждому аффиксу свойственна своя палитра частей речи, к которым он может присоединиться. Аффиксы из рассмотренного примера могут соединяться со следующими частями речи:

«Сем» – аффикс множественного числа. Существительные, прилагательные, числительные, местоимения.

«Пе» – аффикс дательного падежа. Существительные, прилагательные, числительные, местоимения.

«Х» – аффикс категории усиления. Существительные, прилагательные, числительные, местоимения, глаголы.

Список подходящих частей речи довольно обширный. Для компактности части речи разделены на условные группы. После анализа выявлены определены типы и которые применены в конечной версии базы аффиксов.

OnlyGlagol – тип, под которым объединены аффиксы глагольного типа, т.е. склеивающиеся только с глаголами. К таким относятся: «ма, ме, маç, меç».

NotGlagol – тип, объединяющий неглагольный класс аффиксов. По большей части это падежные аффиксы: «па, пе, ра, ре, та, те».

Any – общий тип аффиксов, способные соединяться и с глагольными и именными частями речи. Например, аффикс «а» относится и к дательному падежу (çурт -> çурта), так и к деепричастиям (чуп -> чупа).

Формирование базы аффиксов завершается определением одного из трех типов для каждого уровня аффиксов. Некоторые уровни, содержащие пересекающиеся типы, разбиваются до наличия лишь однородных аффиксов на один уровень. БА на данный момент разделен на тридцать семь (37) уровней.

В целях более удобного внесения изменений словарь и база аффиксов и другие вспомогательные документы хранятся в виде текстовых файлов. Для редактирования как специалисту, так и пользователю в качестве программного обеспечения достаточно стандартного блокнота. Тем не менее при необходимости БА легко может быть конвертирована в таблицы базы данных в силу своей структуры.

Разработку морфологического анализатора можно разделить на два этапа. Во-первых, слово в исходной форме ищется в словаре основ. Грамматические характеристики в данном случае определяются по умолчанию в зависимости от части речи. На втором этапе производится непосредственный анализ слова, разбиение его на пары «корень-аффиксы» и выявление характеристик. Оба этапа возвращают произвольное количество частей речи в зависимости от найденных совпадений. При отсутствии совпадений слово возвращается с «неопределенными» характеристиками. Рассмотрим поподробнее каждый из этих этапов.

Первая и наиболее простая часть – это прямой поиск входного слова в словаре основ. Эту задачу решает функция SearchInDictionaries. Алгоритм метода приводится ниже.

Алгоритм поиска совпадений в словаре.

Шаг 1. Отделить от входного слова аффиксы-частицы, такие как «-и», «-çке» и другие: «ăшă-çке» (тепло же), «пысăк-и?» (большой ли). Записать аффиксы в строковой переменной. Если слово «двойное», образованное с помощью повторения (кил-кил), убрать повторяющуюся часть слова.

Шаг 2. По началу слова определить диапазон поиска. Так как словарь отсортированный, непосредственно проход по всему словарю не требуется.

Шаг 3. В выявленном диапазоне производить поиск слова. Вне зависимости от найденных совпадений продолжать поиск до конца.

Шаг 4. Извлечь часть речи найденного слова. На основе него определить остальные грамматические характеристики.

Определение границ поиска непосредственно основано на рассмотрении всех возможных вариаций пар символов, с которых может начинаться слово (рис. 2).

```
"аб", "ав", "аг", "ад", "аз", "аи", "ай", "ак", "ал", "ам", "ан", "ао", "ап", "ар",
"ăв", "ăй", "ăл", "ăм", "ăн", "ăп", "ăр", "ăс", "ăç", "ăт", "ăф", "ăх", "ăш",
"ба", "бе", "би", "бл", "бо", "бр", "бу", "бы", "бю",
"ва", "вă", "ве", "вё", "вě", "вз", "ви", "вк", "вл", "во", "вр", "вс", "вт", "ву",
"га", "га", "ге", "гё", "ги", "гл", "гн", "го", "гр", "гу", "гя",
"да", "дв", "де", "дж", "дз", "ди", "дн", "до", "др", "ду", "ды", "дю",
"ев", "ег", "ед", "ей", "ек", "ел", "ем", "ен", "еп", "ер", "ес", "ет", "еф", "ех",
"ěк", "ěл", "ěм", "ěн", "ěп", "ěр", "ěс", "ěç", "ěт", "ěф", "ěх", "ěш",
"жа", "жг", "же", "жи", "жм", "жн", "жо", "жр", "жу", "жю",
"за", "зв", "зд", "зе", "зн", "зо", "зу", "зэ",
"иб", "ив", "иг", "ид", "ие", "иж", "из", "ий", "ик", "ил", "им", "ин", "ио", "ип",
"йа", "йă", "йё", "йо", "йу", "йы",
"ка", "кă", "кв", "ке", "кě", "ки", "кл", "кн", "ко", "кр", "кс", "ку", "кӳ", "кх",
"ла", "лă", "ле", "лё", "лě", "ли", "ло", "лу", "лӳ", "ль", "лю", "ля",
"ма", "мă", "ме", "мě", "ми", "мн", "мо", "мр", "му", "мӳ", "мы", "мэ", "мя",
"на", "нă", "не", "нě", "ни", "но", "нр", "ну", "нӳ", "нэ", "ня",
"оа", "об", "ов", "ог", "од", "оз", "ой", "ок", "ол", "ом", "он", "оп", "ор", "ос",
"па", "пă", "пе", "пě", "пи", "пл", "пн", "по", "пр", "пс", "пт", "пу", "пӳ", "пы",
"ра", "ре", "ри", "ро", "рт", "ру", "ры", "рэ", "рю", "ря",
"са", "сă", "сб", "св", "сд", "се", "сё", "сě", "си", "ск", "сл", "см", "сн", "со",
"ça", "çă", "çв", "çе", "çё", "çи", "çт", "çу", "çӳ", "çы",
"та", "тă", "тв", "те", "тё", "тě", "ти", "тк", "то", "тп", "тр", "тс", "ту", "тӳ",
"уа", "уб", "ув", "уг", "уд", "уе", "уз", "уй", "ук", "ул", "ум", "ун", "уп", "ур",
"ӳк", "ӳл", "ӳн", "ӳп", "ӳр", "ӳс", "ӳт", "ӳх",
```

Рис. 2. Структура вариаций пар символов

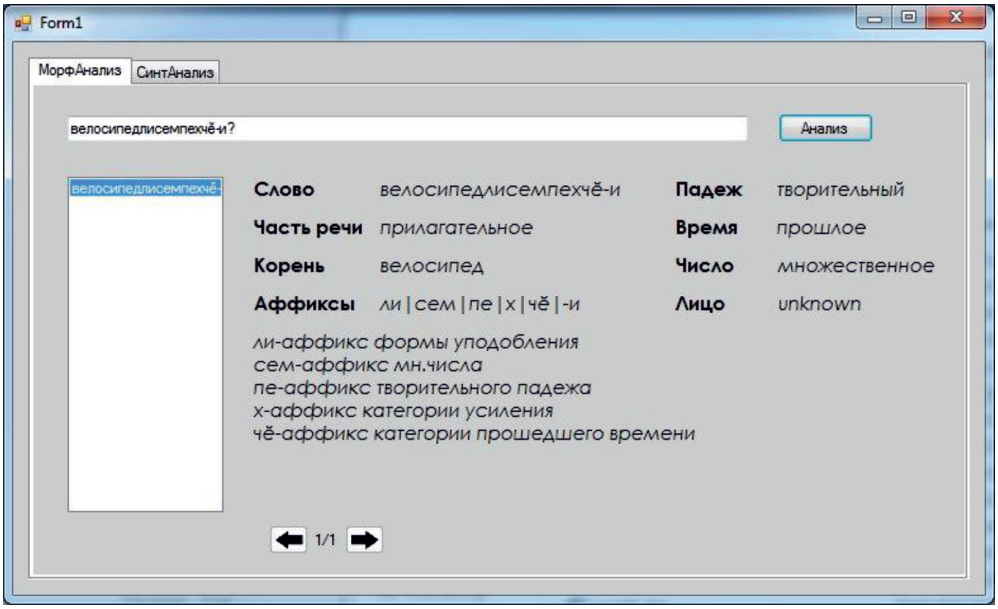

Рис. 3. Рабочее окно морфологического анализатора

Составив список пар, отсортировав их в алфавитном порядке и сопоставив их со словарем, получаем некую карту, где каждой паре слева справа соответствует номер строки, с которой начинаются подобные слова. Однако эти данные не определяются заранее. Так как в МА предусмотрена возможность динамического подключения разных словарей, выявление границ поиска определено в функцию. Таким образом, функция автоматически запускается при подключении нового словаря, обеспечивая анализатор актуальными данными [4, 5].

Как видно из рис. 3, программа корректно проводит анализ слов и правильно определяет корень и атрибуты. Как показали результаты тестирования, установление морфемного состава слов, а также морфологических признаков производится за 1–3 миллисекунды.

В предлагаемой работе разработана библиотека морфологического анализа чувашского языка. Библиотека создана на платформе NET.Framework в среде Visual Studio 2013 на языке C#. Приведен оригинальный алгоритм рекурсивного метода разбиения исходного слова на составляющие. Разработанная библиотека выполняет следующие задачи: определение части речи слова; извлечение корня и аффиксов; анализ контекстов, восстановление символов; определение морфологических характеристик; логгирование работы для удобной отладки.

Результаты работы библиотеки успешно могут быть применены на входе синтаксического анализатора и являются составной частью лингвистического процессора.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 15-04-00532.

Список литературы

1. Желтов П.В. Морфологический анализатор национального корпуса чувашского языка // Совершенствование методологии познания в целях развития науки: Сборник статей по итогам Международной научно-практической конференции (Самара, 30 июня 2017) / в 2 ч. Ч. 1. – Стерлитамак: АМИ, 2017. – С. 11–13.

2. Желтов П.В. Создание национального корпуса чувашского языка: проблемы и перспективы // Современные проблемы науки и образования. – 2015. – № 1–1.; URL: http://www.science-education.ru/ru/article/view?id=19046 (дата обращения: 18.07.2017).

3. Губанов А.Р. Морфологический стандарт для систем автоматической обработки текстов на чувашском языке и архитектура грамматического словаря // В сборнике: Актуальные вопросы истории и культуры Чувашского народа / Составитель и научный редактор Н.Г. Ильина; Чувашский государственный институт гуманитарных наук. – Чебоксары, 2015. – С. 146–161.

4. Сулейманов Д.Ш., Невзорова О.А., Гатиатуллин А.Р., Гильмуллин Р.А., Аюпов М.М., Пяткин Н.В. Основные компоненты прикладной грамматической модели татарского языка // В Трудах Междунар. научной конференции Диалог-2007. Компьютерная лингвистика и интеллектуальные технологии. – М.: Изд-во РГГУ, 2007. – С. 525–530.

5. Сулейманов Д.Ш., Невзорова О.А., Гатиатуллин А.Р., Гильмуллин Р.А. Лингвистические аспекты информационного поиска для тюркских языков // Искусственный интеллект. Интеллектуальные системы (ИИ-2009) / Материалы X Международной научно-технической конференции. – Таганрог: Изд-во ТТИ ЮФУ, 2009. – 294 с.: – С. 288–290.