# ПРОГРАММНЫЙ АЛГОРИТМ МОРФЕМНОГО АНАЛИЗА СЛОВ РУССКОГО ЯЗЫКА

Правосудов М. М.
Скирневский И. П.
Томский политехнический университет, Институт кибернетики
matvey@pravosudov.com

## Введение

Цель морфемного анализа – определение морфемного состава слова. Морфема – значащая часть слова: корень, приставка или суффикс [1]. То есть морфемный анализ (разбор) слова – определение состава его морфем, который является одним из самых общих в начальном образовании. На данный момент существует несколько словообразовательно-морфемных словарей, но почти все они являются бумажными, или оцифрованными PDF-файлами, что делает невозможным их использование «на ходу» в программах. Цель создания данного алгоритма – осуществление морфемного анализа слов в автоматических режиме.

Для создания, приведенного в этой статье алгоритма была рассмотрена работа Ронжина А. Л. О системе автоматического распознавания русской речи SIRIUS [2] и исследование Дикого П. В. «алгоритм и практическая реализация морфемного разбора» [3].

## Содержание алгоритма

Алгоритм совершает несколько шагов, которые описаны на Рис. 1.


Рис. 1. Схема работы алгоритма

В состав ПО алгоритма входит оцифрованный Морфемно-орфографический словарь русского языка Тихонова А. Н. (Рис. 2.) [4], а также словарь морфем в формате JSON из указателя словаря Ефремовой Т. Ф. [5] и данных сайта «Словород» [6].

```
абаз|аба'з/
абазин|абази'н/
абазинец|абази'н/ец/
абазинка|абази'н/к/а
```

Рис. 2. Оцифрованный словарь Тихонова А. Н.

Чтобы упростить процесс описания работы алгоритма, рассмотрим ключевые этапы алгоритма на примере разбора слова «абажуродержатель». Слово построчно ищется в словаре Тихонова. Далее, полученный результат ***(«абажур/о/держ/а'/тель/»)*** очищается от ударения и разбивается на массив морфем. На данном этапа алгоритм хранит информацию о морфемах слова, разделенных слешами, но неизвестной категории.

Происходит перебор всех приставок из словаря для n-первых морфем (n равно количеству всех морфем – 1, так как кроме приставки обязательно должен быть как минимум один корень). На каждой итерации проверяется, была ли предыдущая морфема приставкой. Если нет, то текущая морфема – корень. Пример представлен на Рис. 3.

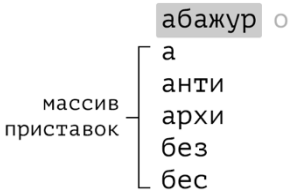


Рис. 3. Перебор приставок

В результат добавляется массив приставок. В слове абажуродержатель приставок нет.

Происходит перебор всех окончаний для последней морфемы ***(«тель»)*** и сохраняется маркер конца корневого массива (для последующего определения корней). Если окончание нулевое (в конце есть пустой слеш, наш случай), то оно помечается значением ***:empty:***. В дальнейшем, этот маркер используется для отображения графической схемы слова.

Следующим этапом работы алгоритма является выборка массива суффиксов с определенными условиями. Если в слове есть окончание – суффиксы нужно искать на одну морфему раньше. Если есть приставки, то нужно забронировать для них (и хотя бы одного корня) нетронутое суффиксальным циклом место. Суффиксы перебираются n раз (количество приставок плюс хотя бы один корень < n < индекс окончания) в обратном порядке. Последовательность суффиксов проверяется аналогично приставкам, чтобы не было «разрывов» в суффиксах. Определяются суффиксы «тель» и «а». В результат записывается реверсированный массив суффиксов.

Далее определяется маркер конца корневого массива (индекс первого по порядку следования морфем суффикса).

Проверяется постфикс и записывается маркер конца корневого массива.

Если есть приставки, маркер начала равен индексу последней приставки. Если есть суффиксы, то маркер конца равен первому индексу суффикса; если нет суффикса, то окончания; если нет окончания, то постфикса. Возможный массив корней ***(абажур/о/держ)*** представлен на Рис. 4.

```
абажур о держ а тель
```

Рис. 4. Массив корней

Всё в полученных ранее границах – корни или соединительные гласные «о» и/или «е». В слове абажуродержатель два корня, соединенных соединиительной гласной «о»: «абажур» и «держ».

Формируется результирующий массив данных (Рис. 5.), который впоследствии может быть использован при построении визуального представления слова.

```
array:3 [
  "roots" => array:3 [
    0 => "абажур"
    1 => "о"
    2 => "держ"
  ]
  "suffixes" => array:2 [
    0 => "а"
    1 => "тель"
  ]
  "ending" => ":empty:"
]
```

Рис. 5. Результат – массив морфем

Таким образом, алгоритм полностью выполняет задачи, которые ставятся ему морфемным анализом.

**Морфемный анализ исключений**

Для тестирования алгоритма предлагается использовать слова-исключения, которые имеют специфическую структуру. Например, слово вынуть – единственное в русском языке, не содержащее корня. Результат работы алгоритма приведен на Рис. 6.

```
array:2 [
  "prefixes" => array:1 [
    0 => "вы"
  ]
  "suffixes" => array:2 [
    0 => "ну"
    1 => "ть"
  ]
]
```

Рис. 6. Слово-исключение вынуть

Стоит заметить, что многие авторы разнятся в принадлежности «ть» окончанию или суффиксу, но алгоритм определяет его как суффикс.

Итогом представленной работы является: разработанный алгоритм морфемного анализа слов русского языка и реализован в веб-сервисе с использованием фреймворка Laravel [7].

**Заключение**

Представленный в работе алгоритм может быть использован для создания образовательных сервисов, позволяющих проводить морфемный анализ слов, или же для более глубокого исследования русского языка. Также с помощью алгоритма можно построить более полный морфемно-орфографический словарь, который будет использоваться без разбора в режиме реального времени.

**Список использованных источников**

1. Толковый словарь русского языка: В 4 т. / Под ред. Д. Н. Ушакова. Т. 1. М., 1935; Т. 2. М., 1938; Т. 3. М., 1939; Т. 4, М., 1940. (Переиздавался в 1947-1948 гг.)

2. Ронжин А. Л., Карпов А. А., Ли И. В. Система автоматического распознавания русской речи SIRIUS Automatic system for Russian speech recognition SIRIUS //Донецк, Украина. – 2005. – С. 590-601.

3. Petr V. Dikiy. Algorithm and practical implementation morphemic parsing // Theoretical and Applied Aspects of Cybernetics: proceedings of the International Scientific Conference of Students and Young Scientists — Kyiv: Bukrek, 2011.

4. Словари русского языка для скачивания [Электронный ресурс] / Архивы форума «Говорим по-русски» – URL: http://www.speakrus.ru/dict/index.htm (посл. обращение 09.08.2016).

5. Указатель морфов [Электронный ресурс] / Русская грамматика – URL: http://rusgram.narod.ru/morf1t.html (посл. обращение 09.08.2016).

6. Гаршин И. К. Словород: образование и история слов русского языка. Собирание и оживление славянских корней. [Электронный ресурс] / Словород – URL: http://www.slovorod.ru/ (посл. обращение 09.08.2016).

7. Словолит. [Электронный ресурс] / Словолит – URL: http://slovolit.herokuapp.com (посл. обращение 09.08.2016).