ВНЕСЕНИЕ АКЦЕНТУАЦИОННОЙ ИНФОРМАЦИИ В РУССКИЙ МОРФОЛОГИЧЕСКИЙ СЛОВАРЬ ЛИНГВИСТИЧЕСКОГО ПРОЦЕССОРА ЭТАП-3

В. Сизов

О. Подлесская

Краткое описание морфологического словаря ЭТАП-3:

Структура статьи:

- неизменяемая основа – морфема, общая для всех словоформ
- словоизменительные морфемы: приставка, тема(1-й суффикс), суффикс, окончание, частица (*-ся/сь*)
- Стандартные объекты (СТО), описывающие повторяющиеся последовательности словоизменительных морфем

По данным, приведенным в статье, транслятор генерирует все возможные словоформы с характеристиками.

В настоящее время русский морфологический словарь ЭТАП-3 содержит около 130 тыс. статей и около 1000 стандартных объектов.

Пример статьи КС и сгенерированной по ней парадигмы

- ENTRY:ТРАКТОР
- осн:тра`ктор ф:1,ок:'а'им,мн,'а'вин,мн т:6
- trs:tractor
- т:6 - ф1,ок:6
- ф:1 - хар:S,муж,неод
- ок:6 - '#'ед,им,'а'ед,род,'у'ед,дат,'#'ед,вин,'ом'ед,твор,'е'ед,пр, 'ы'мн,им,'ов'мн,род,'ам'мн,дат,'ы'мн,вин,'а`ми'мн,твор,'ах'мн,пр,'о'сл

- *трактор – S,ЕД,МУЖ,ИМ,НЕОД*
- *трактор|а – S,ЕД,МУЖ,РОД,НЕОД*
- *трактор|у – S,ЕД,МУЖ,ДАТ,НЕОД*
- *трактор – S,ЕД,МУЖ,ВИН,НЕОД*
- *трактор|ом – S,ЕД,МУЖ,ТВОР,НЕОД*
- *трактор|е – S,ЕД,МУЖ,ПР,НЕОД*
- *трактор|о – S,МУЖ,НЕОД,СЛ*
- *трактор|ы – S,МН,МУЖ,ИМ,НЕОД*
- *трактор|а – S,МН,МУЖ,ИМ,НЕОД*
- *трактор|ов – S,МН,МУЖ,РОД,НЕОД*
- *трактор|ам – S,МН,МУЖ,ДАТ,НЕОД*
- *трактор|ы – S,МН,МУЖ,ВИН,НЕОД*
- *трактор|а – S,МН,МУЖ,ВИН,НЕОД*
- *трактор|а`ми– S,МН,МУЖ,ТВОР,НЕОД*
- *трактор|ах – S,МН,МУЖ,ПР,НЕОД*

Чередование ударений в русском языке

- Чередование ударения в рамках сходных парадигм: *(мопе`д)-ами / (стол)-а`ми (т:6); (умн)- е`йш-(ий) /разу`мн)-ейш-(ий) (т:211)*
- Чередование ударений в рамках парадигмы одной лексемы: *кра`сн-ый / красн-е`йш-ий / красн-а`*.
- Чередование ударения в рамках ударной морфемы (редко): *дворяни`н / дворя`н-е, риск-ова`-ть/риск-о`ва-нн-ый*

Простейшее решение:

Добавить знаки ударения в существующий морфологический словарь.

Недостатки:

Чтобы отразить чередование ударений, потребуется дробление существующих СТО и введение новых чередований (разновидностей СТО) для основ слов. Это приведет к переписыванию почти всех статей словаря

Решение: использование специальных акцентуационных правил для реализации чередований ударений.

- Правила ставят ударение на нужную морфему в зависимости от сформулированных в них условий (которые могут быть взяты из описания схем ударений словаря А.А.Зализняка);
- Правила применяются после генерации парадигмы статьи к полученным в результате словоформам, не зависят от блочной структуры статьи и могут быть вставлены в любое место статьи

Формат правила

Правило содержит:

• логические выражения, проверяющие истинность сформулированных в них условий. Логические выражения состоят из **предикатов**, объединенных в конъюнкции, дизъюнкции и скобочные выражения.

•инструкции, которые выполняются, если проверка условий дала истину.

Предикаты

- Проверяют наличие/отсутствие морфологических характеристик. Совпадают с именем проверяемой характеристики (напр. *ЕД+(ИМ|ВИН+ОД)*).
- Проверяют буквенные цепочки в морфемах, входящих в словоформу. Имеют вид: *search(строка_поиска, имя_морфемы).*
  В строке поиска допустимы регулярные выражения.

Инструкция: имя, позиция ударения, приоритет.

- Имя инструкции (*преф:*, *осн:, тм:, сф:, ок:*,.и *чс:*) совпадает с именем морфемы, на которую на которую должно падать ударение. Если указанной морфемы в словоформе нет или в ней нет гласных, ударение ставится на последний слог предшествующей морфемы
- Если указана позиция ударения (порядковый номер слога в морфеме) – ударение ставится в указанной позиции. Если нет - остается проставленное ранее ударение по умолчанию.
- Если словоформа удовлетворяет условиям нескольких правил, то применяется инструкция с наивысшим приоритетом. Если таких инструкций несколько, то словоформа дублируется по числу инструкций, и каждая инструкция применяется к своей копии.

Типы правил

- Общие – применяются к словоформам всех статей. Записаны в специальном файле описания языково-специфичных свойств.
- Трафаретные – размещаются в файле стандартных объектов внутри специальных СТО acct. Применяются к статьям, в которых есть ссылки на эти СТО. СТО получают имена, соответствующие схемам ударения из словаря Зализняка.
- Словарные – размещаются в статье внутри специальных блоков acct. Применяются к словоформам статьи, внутри которой они размещены.

Алгоритм трансляции статьи

- Правила расстановки ударений из СТО и блоков в статье собираются в один набор
- Если в набор не попало ни одно правило, предварительно расставленные ударения стираются. В противном случае после генерации словоформ правила из набора и общие правила применяются к каждой полученной словоформе.
- Сперва проверяется истинность логических выражений правил. Из правил, получивших истинность, отбираются те, чьи инструкции имеют наивысший приоритет, эти инструкции применяются к словоформе.

Пример применения правила:

ENTRY:ТРАКТОР acct:c_a
осн:тра`ктор ф:1,ок:'а'им,мн,'а'вин,мн т:6
trs:tractor

Трафаретное правило: acct:c_a
end:=МН+^(ИМ|ВИН); end:{2}=МН+(ИМ|ВИН)+search("[а|я]",end:); end:=МН+ВИН+ОД;
bas:=S;
Общее правило:
осн:(0,"~")=СЛ+^V; [Ударение в форме СЛ всегда слабое и падает на основу]*

Парадигма:

- *тра`ктор – S,ЕД,МУЖ,ИМ,НЕОД*
- *тра`ктор|а – S,ЕД,МУЖ,РОД,НЕОД*
- *тра`ктор|у – S,ЕД,МУЖ,ДАТ,НЕОД*
- *тра`ктор – S,ЕД,МУЖ,ВИН,НЕОД*
- *тра`ктор|ом – S,ЕД,МУЖ,ТВОР,НЕОД*
- *тра’ктор|е – S,ЕД,МУЖ,ПР,НЕОД*
- *тра’ктор|о – S,МУЖ,НЕОД,СЛ*
- *тра`ктор|ы – S,МН,МУЖ,ИМ,НЕОД*
- *тра`тор|а – S,МН,МУЖ,ИМ,НЕОД*
- *тра`ктор|ов – S,МН,МУЖ,РОД,НЕОД*
- *тра`ктор|ам – S,МН,МУЖ,ДАТ,НЕОД*
- *тра`ктор|ы – S,МН,МУЖ,ВИН,НЕОД*
- *тра`ктор|а – S,МН,МУЖ,ВИН,НЕОД*
- *тра`ктор|а`ми– S,МН,МУЖ,ТВОР,НЕОД*
- *тра`ктор|ах – S,МН,МУЖ,ПР,НЕОД*

Морфологический анализ и синтез с учетом акцентуационной информации

Внесение в морфологический словарь ЭТАП-3 акцентуационной информации позволяет реализовать:

- распознавание морфологическим анализатором словоформ из входного текста как при наличии в них символов ударения и буквы *ё,* так и при их отсутствии;
- “строгий” режим разрешения омонимии *е/ё* для текстов с последовательно проставленной буквой *ё*. В этом режиме недопустимо распознавание буквы *е* как *ё*: (*он осел ≠ он осёл*);
- возможность генерации выходного текста как с акцентуацией и *ё*, так и без нее;
- при генерации словоформ с альтернативными ударениями (напр. *профе`ссорам / профессора`м*) - возможность вывода наиболее употребительной словоформы по умолчанию и явного выбора конкретной словоформы.