# Постморфологический анализ. Описание МаПоста

Программа постморфологического анализа состоит из
***нескольких довольно простых правил, разрешающих омонимию по контексту,***
***и правил, достраивающих морфологическую интерпретацию.***

Многие из этих правил можно было отнести либо к русской морфологии, либо к русскому синтаксису.

Этого не было сделано из-за того, что:

**существующая морфология имеет обобщенный характер, и нежелательно загромождать ее частными случаями из русского языка**

**в существующем синтаксическом анализе нет жесткого механизма удаления омонимов до построения клауз**

**Правило однобуквенных слов. Правило для цифровой записи числительных. Правило одинаковых лемм.**

Функция идет по омонимам слова. Пусть Z0 = <L,M> - первая интерпретация некоторого слова, где L - лемма, а M - тип морф. словаря. Если найдется среди омонимов слова другая интепретация Zi с такими же <L,M>, тогда Zi нужно уничтожить, а все ее грамматические коды вставить в Z0.

Это правило работает на словах, "обещать" (нс/св), "сервер" (од/но).

**Правило иностранных дефисных слов. Правило дефисных слов. Правило разделение причастий и личных форм**

Правило идет по всем словам.

Если найдена словоформа, которой одновременно приписаны личная форма глагола и причастие, тогда разделяем эту словоформу на два омонима (причастие и личная форма).

Например:

*"выдвинут + ВЫДВИНУТЬ фжцж"*

*фж Г св,пе,дст,буд,3л,мн*

*цж Г св,пе,прч,прш,стр,ед,мр,кр*

после работы алгоритма получаем

*выдвинут + ВЫДВИНУТЬ фж*

*выдвинут + ВЫДВИНУТЬ цж*

# Правило местоименного прилагательного

Правило пытается разрешить омонимию между МС (синтаксическим существительным) и МС-П (синтаксическим прилагательным) по ближайшему справа слову.

Правило идет по всем словам входного файла, пусть X - текущая русская словоформа =/= «это», «то», «их», «его», «ее» (эти исключения либо омонимичны части союзам, либо не склоняются). Пусть X одновременно приписано два омонима МС-П и МС (например, "свое", "всякий"). Проверяем, стоит ли после этой словоформы существительное, причастие, прилагательное или местоим. прил, согласованное с этим словом по роду, числу и падежу.

Если стоит, тогда удаляем омоним МС.

Например:

+ *свои люди*

+ *всякий  человек*

-*если я найду, то облако взлетит.*

-*Это облако.*

-*я считал его/ее великим учителем*

*- я считал их идиотами.*

# Правило сравнительного наречия

Правило идет по файлу, ищет слова, которые не были найдены в морфологии и которые начинаются с префикса «по-» и заканчиваются на «ее» или «ей».
Изымаем префикс «по-» и морфологизируем полученное слово. Если была найдена сравнительная степень прилагательного, тогда записываем в качестве леммы это прилагательное.
Примеры:
*побыстрей -> быстрее*
*побыстрее -> быстрее*

# Правило устойчивых словосочетаний

Правило ищет словосочетания по леммам.

Перечень словосочетаний загружается из файла «Registry\...\software\Dialing\mapost\collocs».

Описание одного словосочетания состоит из трех частей:

Набор лемм (для каждой леммы можно указать часть речи в квадратных скобках сразу же после самой леммы)

Номер главного слова словосочетания

Интерфейсная строка

Правило идет по словам входного файла и пытается найти хотя бы одно словосочетание из заданного перечня.

Проверяются лемма и часть речи (если она была задана).

Если **нашлось** **словосочетание**, тогда все строки, кроме строки, совпавшей с главным словом словосочетания, удаляются из файла.

В главной строке **лемма** меняется на **интерфейсную строку** найденного словосочетания.

Например, оборот
«кто бы то ни было» заменится на
«кто?»
Предполагается, что интерфейсная
строка содержится в семантическом
словаре.

# Правило прилагательных  с приставкой «полу-»

Правило проходит по всем словам, которые не были найдены в морфологии. Смотрит, не начинаются ли они с префикса **полу-**, если начинаются, и в морфологии есть такое же прилагательное без приставки, тогда ставим лемму без приставки, а в графематические пометы заносим графету **#ПОЛУ**

**Правило существительных с приставкой пол-**

Правило проходит по всем словам, которые не были найдены в морфологии.

Смотрит, не начинаются ли они с префикса **пол-** (**полу-**), если начинаются и в морфологии есть такое же существительное без приставки, тогда ставим лемму без приставки, а в графемематические пометы заносим графету **#ПОЛУ.**

Если слово без приставки стоит в родительном падеже, тогда нужно выставить граммемы им., вн. и рд. падежей, например,

*Полчаса оказались* (**им**)
*Я прождал полчаса* (**вн**)
*Мечта этого получаса* (**рд**)

Во всех остальных случаях нужно оставить граммемы падежа без изменений. Считаем, что все изменяемые этим правилом существительные имеют граммему множественного числа. Это нужно, чтобы на синтаксисе собралась группа "первые полчаса". Однако тогда получится, что примеры

*Прошло полчаса*
*Пол-Москвы знает об этом,*

где нужно согласование со сказуемым в единственном, не будут собираться на синтаксисе, но соберутся на семантике, где несогласование по числу не является фатальным.

# Правила для построения (ФИ1-ФИ2)

ФИО с инициалами строятся на графематике, а на MaPost строятся только полные имена.

Введем обозначения.

Пусть **ИМЯ** – это некоторое существительное из словаря имен с граммемой **имя** (*Олег, Маша и т.д.*).

Пусть **ОТЧЕСТВО** – слово с графематической пометой Бб, которое заканчивается на "-на"("Ивановна"), "-ич"("Иванович") "-ыч"("Иваныч").

Пусть **ФАМИЛИЯ** – это некоторое существительное из словаря имен с граммемой **фам** (*Иванов, Петров и т.д.*). Для нас неважно, найдены или предсказаны эти слова.

Это правило состоит из двух  подправил.

1.

**ИМЯ ОТЧЕСТВО ФАМИЛИЯ** *Петр Владимирович Иванов,*

**ФАМИЛИЯ ИМЯ ОТЧЕСТВО** *Иванов Петр Владимирович*

**ИМЯ ФАМИЛИЯ** *Петр Иванов,*

**ФАМИЛИЯ ИМЯ** *Иванов Петр*

**ИМЯ ОТЧЕСТВО** *Петр Владимирович*

Слова должны быть согласованы по падежу.

2.

**ИМЯ ИМЯ ФАМИЛИЯ** *Райнер Мария Рильке*

Так иногда записываются западные имена.

Максимальная из найденных цепочек объявляется "лучшей". Все омонимы, которые не удовлетворяют этой цепочке удаляются. В начале цепочки ставится графематическая помета ФИ1, в конце – ФИ2.

Если одна из частей цепочки попадает в уже построенную ФИ1-ФИ2, тогда эта цепочка не строится. Если одна из частей цепочки попадает а ОБ1-ОБ2, тогда ОБ1-ОБ2 стирается.

# Правила об "уже"

Правило идет по файлу и ищет словоформу УЖЕ. Если после нее нет запятой или слова с родительным падежом, тогда из морфологической интерпретации этой словоформы удаляется омоним прилагательного сравнительной степени Например:

+ *сказка о белом уже*

+ *уже ушел*

- *Эта кровать уже, чем диван.*

- *Эта кровать уже дивана.*

# Правила о несклоняемых существительных на "о"

Синтаксис иногда неправильно находит подлежащее. Например:

*С восточной стороны плато скалы были красного цвета,*

где *плато* становится подлежащим. Это происходит из-за того, что *плато* является неизменяемым. Поэтому было написано следующее правило.

Правило идет по неизменяемым существительным, которые заканчиваются на "-о". Если это существительное стоит во **множественном контексте** [3].

Иначе удаляем у него омоним множественного числа.

- *Большие пальто*
- *их пальто*
+ *большое пальто*
+ *первое пальто*
+ *мое пальто*

# Правила о словах с затрудненным множественным числом

Правило использует перечень лемм, имеющих затрудненную форму множественного числа. Перечень загружается из файла "Registry\...\software\Dialing\mapost\DeadPlurals".
Правило идет по всем леммам. Если лемма была найдена в указанном выше перечне и не находится во множественном контексте, тогда удаляем все грамкоды множественного числа у этой леммы.

Например:
+ *финансовой отчетности*
- *наши последние отчетности.*

# Правило о фамилиях

Правило использует перечень парадигм самых частых фамилий русского языка.. Перечень загружается из файла "Registry\...\software\Dialing\mapost\Surnames".

Одна парадигма состоит из пар вида (окончание, аношкинский код). Парадигмы пронумерованы начиная с 1. Одно окончание может входить много раз в одну парадигму или в несколько парадигм.

Аношкинским кодом называется уникальный двухбуквенный идентификатор, который соответствует некоторой комбинации значений селективных признаков и граммем.

Приведем фрагмент описания парадигмы для лексемы ‘рукоплескать’:

1740 %СКАТЬ*ка%СКАВШАЯ*мз%ЩУ*кб%ЩУТ*кж %ЩУЩЕГО*лблглп%….

РУКОПЛЕ 1740

‘Рукопле’ – основа слова в лексиконе; ‘1740’ – уникальный идентификатор парадигматического класса; ‘%’ маркирует начало псевдо-флексии; ‘*’ маркирует начало аношкинского кода; ‘ка’, ‘кб’, ‘лб’, ‘лг’, etc. – код.

| код | часть речи | словообразовательные характеристики | словоизменительные характеристики | Пример |
|---|---|---|---|---|
| ка | Г | нс,нп | дст,инф | рукоплескать, расти |
| мз | Г | нс,нп,прч | прш,дст,ед,жр,им | рукоплескавшая, росшая |
| кб | Г | нс,нп | дст,нст,1л,ед | рукоплещу, расту |
| кж | Г | нс,нп | дст,нст,3л,мн | рукоплещут, растут |
| лб | Г | нс,нп,прч | нст,дст,ед,мр,рд | рукоплещущего, растущего |
| лг | Г | нс,нп,прч | нст,дст,ед,мр,вн | рукоплещущего,растущий |
| лп | Г | нс,нп,прч | нст,дст,ед,ср,рд | рукоплещущего,растущего |

Правило идет по всем ненайденным в морфологии словам. Если слово начинается с большой буквы и заканчивается на одно из окончаний из вышеуказанного перечня, тогда

1.удаляем все омонимы, построенные морфологией;

2.проходим по всем окончаниям, с которыми совпало с конца данное слово;

3.и формируем множество интерпретаций М, которое и приписываем входному слову (число получаемых омонимов равно числу парадигм, в которых было найдено слово) .

# Правило о женских фамилиях

В морфологии часто нет женских фамилий, но есть соответствующие мужские. Это правило добавляет морфологические интрепретации (как женск. фамилия) к словоформе мужской фамилии.

Правило идет по найденным словам, если они являются мужской фамилией, а среди омонимов нет женской фамилии, и от этой словоформы можно образовать женскую фамилию (используется перечень Правила о фамилиях), тогда добавляем омоним женской фамилии.

Например, фамилия ФРОЛОВ есть в морфологии, а фамилии "ФРОЛОВА" нет. На словоформу "ФРОЛОВА", которая без этого правила интепретировалась бы только как род. падеж от ФРОЛОВ, навешивается омоним женской фамилии "ФРОЛОВА".

# Правила о словах типа "3-летний"

Правило ищет конструкции типа "2-летний". Эти слова занимают три графематические строки:

2 4 0 4 ЦК -1

- 1 4 1 ЗПР ДЕФ -1

летний 2 5 2 ЛЕ бб ?? -1

После работы этого правила получаем одну графематическую строку:

2-летний 2 5 2 ЛЕ бб ?? –1

Морф. интерпретацию новой строки получаем следующим образом:

1.если в морфологии есть слово, которое начинается с дефиса [4] , тогда выбираем эту интерпретацию;

2.если нет, тогда оставляем интерпретацию, которая уже была приписана этому прилагательному.

# Правило об обороте "как можно"

Правило ищет последовательность КАК МОЖНО **x**, где у **x** есть омоним [П сравн] или равен "БОЛЕЕ" или "МЕНЕЕ". Если нашло, удаляет леммы КАК и МОЖНО, оставляет у **x** омоним только на [П сравн] (у него могли быть омонимы на наречие). Добавляет в строку **x** графету #КАК_МОЖНО. Работает на фразах:

*я это сделаю как можно раньше*

*я это сделаю как можно более просто*

*я расскажу как можно более простым языком*