Проект по тестированию морфологических парсеров

*Захарова Елена, БКЛ-142 (**FreeLing**)*

1. Проанализируйте систему тегов:

1.1. Сколько частей речи учитывает система; какие части речи в системе отсутствуют, а Вы считаете, что эти части речи необходимо выделять (ответ мотивируйте)

Учитываемые части речи:

- прилагательное (`adjective` , A)
- существительное (`noun` , N)
- глагол (`verb` , V) (деепричастию приписывается тег глагола, то, что это деепричастие передается тегом G в поле залога)
- причастие (`participle` , Q)
- наречие (`adverb` , D)
- количественное числительное (`number, Z`)
- порядковое числительное (`ordinal` , Y)
- местоимение (`pronoun` , E)
- местоимение-наречие (`pronominal-adv` , P)
- местоимение-прилагательное (`pronominal-adj` , R)
- предлог (`preposition` , B)
- союз (`conjunction` , C)
- частица (`particle` , T)
- междометие (`interjection` , I/J)
- компаунд (`compound` , M)
- дата (`date` , W)

1.2. В какие pos-классы попадают местоимения

В зависимотси от типа местоимения попадут в классы pronoun (он) /pronominal-adj (весь) /pronominal-adv (как).

1.3. Как лемматизируются причастия?

Действительные причатсия лемматизируюся к глаголу, а страдательные к начальной форме самого причастия :

- *Кот кот NCNSMA0000 0.661396*
  , , Fc 1
  съевший съедать QNSMSFFA000 0.980769
  торт торт NCFSMI0000 0.641626

Но:

- *Торт торт NCNSMI0000 0.357934*
  , , Fc 1
  съеденный съеденный QNSMSFFSM00 0.343612
  котом кот NCCSMA0000 1

1.4. К одной или разным леммам бу дет отнесены словоформы нашедший и нах одившего, дал и давал

Нашедший и находить - разные леммы:

- нашедший находить
- находивший нахаживать

Дал и давал - одинковые:

- дал давать
- давал давать

1.5. Напишит е правило пересчета т егов систем ы на т еги из ЗС для анафорических мест оимений (он, она и т .п.) и наречий

pronoun(обозначается буквой E + присутствие в разборе указания на лицо) -> SPRO

adverb (D) -> ADV

2.Проведите функциональное т естирование выбранной Вами программы.

2.2. Ответьте на следующие вопросы:

2.2.1. Как решаются проблемы токенизации: что происходит с числами, десятичными числами, сокращениями типа г, словами с дефисами, апострофом, знаками препинания? спецзнаками типа $ или &, смешанными элементами (буквы+цифры, вкраплениями другого алфавита) etc.?

- Числа определяются как number (обозначается Z):

  9 9 Z00000 1
  цветков цветок NCGPMI0000 0.980769

- Десятичные числа также определяются как number:
  478,6 478.6 Z00000 1

- Сокращения типа г. определяются как noun:
  1999 1999 Z00000 1
  г. г. NCGSFI0000 0.0296805

- Знаки препинания считаются отдельными токенами, им присваиваются теги в зависимости от их типа:

Tag	Attributes
Fd	pos:punctuation; type:colon
Fc	pos:punctuation; type:comma
Flt	pos:punctuation; type:curlybracket; punctenclose:close
Fla	pos:punctuation; type:curlybracket; punctenclose:open
Fs	pos:punctuation; type:etc
Fat	pos:punctuation; type:exclamationmark; punctenclose:close
Faa	pos:punctuation; type:exclamationmark; punctenclose:open
Fg	pos:punctuation; type:hyphen

- Слова с дефисами обрабатываются хорошо:

  бело-кремовое бело-кремовое AFSA0F000 0.666667

  Но:

  по по B0 1
  7 7 Z00000 1
  - - Fg 1
  9 9 Z00000 1
  цветков цветок NCGPMI0000 0.980769

- Слова с апострофами разбиваются на несколько токенов по апострофу:

  Кот-д кот-д NP 1
  ’ ’ Frc 1
  Ивуар ивуар NP 1

- Имена собственные, состоящие из 2 слов считаются единым токеном (те, которые без апострофа):
  Западной_Европы западной_европы NP 1
  (в оригинальном тексте написано через пробел - Западной Европы)

- В целом тагер хорошо справляется со спецзнаками и вкраплениями другого алфавита
  - Строка 'кот$'
    Кот кот NCNSMA0000 0.661396
    $ $ Fz 1
  - Строка 'Кот&Кит'
    Кот кот NP 1
    & & Fz 1
    Кит кит NP 1
    . . Fp 1
  - Строка 'Kot съел tort.'
    Kot кот NP 1
    съел съедать VDSMS0F0000 1
    tort tort NCNSFI0000 0.0257503
    . . Fp 1

- Цифры + буквы (правильно лемматизировал, но почему-то неправильно приписал тег):
  31-я 31 JJ 1

2.2.2. Что происходит с незнакомыми словами? Насколько хорошо предсказываются их грамматические характеристики, их леммы?

В разборе текста про бутявок в основном все правильно, проблемы возникают с отделением частиц от незнакомых слов и междометиями:

- *Калушата калушата NP 1*

  присяпали присяпали VDP0S0N0M00 0.52381
  и и C0 1
  Бутявку бутявку NP 1
  стрямкали стрямкали VDP0S0FAA00 0.444444
  . . Fp 1
- *Бутявка-то бутявка-то NP 1*
  некузявая некузявая ANSF0F000 1
  ! ! Fat 1
- *A a C0 1*
  Калуша калуша NP 1
  волит волит VDS0FP3F0000 0.896057
  : : Fd 1
  - - Fg 1
  Оее оее NP 1
  ! ! Fat 1
  Оее оее NP 1
  ! ! Fat 1

2.2.3. Что происходит с омонимичными словоформами: предлагается только один максимально вероятный вариант; предлагаются все возможные варианты, предлагаются все варианты, за исключением очень маловероятных случаев или случаев, снимаемых "надежными" правилами и т.п.

Для омонимичных словоформ предлагается только один наиболее вероятный вариант.

2.2.4. Какие проблемные случаи омонимичных разборов разбираются хорошо, в каких часто возникают ошибки и т.п. (например, (а) частеречная омонимия: прилагательное vs. существительное, глагол vs. прилагательное, наречие vs. частица; (б) падежная омонимия; (в) омонимия различных местоименных форм и т.д.)

- Прилагательно/существительное - различает:
  Больной больной NCNSMA0000 0.387064
  поправился поправляться VDSMS0F0A00 1
  . . Fp 1

Больной больной ANSM0F000 0.274127
котик котик NCNSMA0000 0.493976
поправился поправляться VDSMS0F0A00 1
. . Fp 1

- Числительное/существительное - не различает:

Подали подавать VDP0S0F0000 1
на на B0 0.999321
первое первый YFSA0 0.454684 - тег числительного
. . Fp 1

Первое первый YNSA0 0.46852
собрание собрание NCNSAI0000 0.657867
будет быть VDS0F3F0A00 0.998115
завтра завтра NCDPAI0000 0.49854
. . Fp 1

- Причастие/отглагольное прилагательное - не различает:
У у B0 0.997789
меня я E0S0000 1
приподнятое приподымать QNSASFFS000 0.875 - прилагательно помечено как причастие
настроение настроение NCNSAI0000 0.626063
. . Fp 1

Приподнятый приподымать QNSMSFFS000 0.875
за за B0 1
лапы лапа NCFPFI0000 0.703019
кот кот NCNSMA0000 1
был быть VDSMS0N0A00 1
недоволен недоволен A0SM0S000 1
. . Fp 1

- Глагол/существительное - не различает:
Они они ENP0000 1
стали становиться VDP0S0F0A00 0.97383
*****петь петь VI0000N0M00 0.999582*
. . Fp 1

Кузов кузов NCNSMI0000 0.300142
сделан сделать Q0SMSSFSM00 1
из из B0 1
стали становиться VDP0S0F0A00 0.97383
. . Fp 1

- Падежная омонимия - различает:
Красивые красивый ANP00F000 0.999621
берёзы берёзы NCNPFI0000 0.197906 (NP - Nominative Plural)
, , Fc 1
там там P0 0.915529
нет нет NCFSMI0000 0.999903
берёзы берёзы NCGSFI0000 0.252299 (GS - Genetive Singular)
. . Fp 1

- Омонимия местоименных форм:
  Различает:
  ***Его его RNP000 0.116016** - местоимение -прилагательное**
  кот кот NCNSMA0000 1
  пошёл пошёл VDSMS0F0000 0.204378
  гулять гулять VI0000N0A00 1
  . . Fp 1

  Я я ENS0000 0.996729
  ***его он E0S0000 0.0938098** - местоимение**
  ударил ударять VDSMS0F0000 1
  . . Fp 1

  Не различает (обоим формам приписан тег местоимения-прилагательного):
  Его его RNP000 0.116016
  нет нет NCFSMI0000 0.999903
  . . Fp 1

  Его его RNP000 0.116016
  кота кот NCFSMA0000 0.623119
  тоже тоже T0 0.991101
  нет нет NCFSMI0000 0.999903
  . . Fp 1

3.Обработайте с помощью морфологическ ого анализатора файл.

- Итоговый файл лежит в приложенном архиве, называется test_result_freeling.txt
- Уровень оставшейся неоднозначности = 1
- Accuracy = 93%

Еще в приложенном архиве лежат текстовые файлы, на которых тестировались различные случаи и .mrf файлы - результат разметки этих текстовых файлов (на всякий случай).