Проект выполняли: **Сонина Полина, Стырина Соня**

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)

A. adjective
B. preposition
C. conjunction
D. adverb
E. pronoun
J. interjection (междометие)
N. noun
P. pronominal-adv
Q. participle
R. pronominal-adj
T. particle (частицы)
V. verb
Y. ordinal (порядковые числительные)
Z. number

Выделяется 14 категорий частей речи + compound, даты, пунктуация.
Нет предикативов (PRAEDIC), но они попадают в категорию adverb с пометой predicative, если анализатору удаётся их определить.
Причастия выделяются в отдельную категорию. Но странно то, что разбор причастий не содержит параметр падежа.

1.2. В какие pos-классы попадают местоимения

Кроме классов pronoun, pronominal-adj и pronominal-adv попадают также в adjective, adverb, conjunction, noun (точно ошибочно из-за дефиса) и number.

я	pronoun
ты	pronoun
он	pronoun
она	pronoun
оно	pronoun
мы	pronoun
вы	pronoun
они	pronoun
себя	pronoun
её	pronoun
кто	pronoun
что	pronoun
кто	pronoun
что	pronoun
никто	pronoun
ничто	pronoun
некого	pronoun

нечего	pronoun
некто	pronoun
нечто	pronoun
что-нибудь	pronoun
мой	pronominal-adj
твой	pronominal-adj
свой	pronominal-adj
ваш	pronominal-adj
наш	pronominal-adj
его	pronominal-adj
их	pronominal-adj
какой	pronominal-adj
чей	pronominal-adj
который	pronominal-adj
каковой	pronominal-adj
каков	pronominal-adj
какой	pronominal-adj
который	pronominal-adj
чей	pronominal-adj
каковой	pronominal-adj
каков	pronominal-adj
тот	pronominal-adj
этот	pronominal-adj
такой	pronominal-adj
таков	pronominal-adj
сей	pronominal-adj
всякий	pronominal-adj
каждый	pronominal-adj
сам	pronominal-adj
самый	pronominal-adj
любой	pronominal-adj
иной	pronominal-adj
другой	pronominal-adj
весь	pronominal-adj
никакой	pronominal-adj
ничей	pronominal-adj
весь	pronominal-adj
некоторый	pronominal-adj
где	pronominal-adv
кто-то	adjective
откуда	adverb
сколько	adverb
зачем	adverb
зачем	adverb
незачем	adverb
когда	conjunction
какой-либо	noun
сколько	number
столько	number
несколько	number

1.3. Как лемматизируются причастия?

Иногда правильно разбираются, как причастия, лемма глагол - при активном залоге, причастие - при пассивном; иногда разбираются ошибочно как глаголы или прилагательные, а лемма совпадает с формой

Список приведён в формат ЗС:

решающий	решающий	A	nom,sg,m
решающего	решающий	A	gen,sg
решающему	решающий	A	dat,sg,m
решающим	решающий	A	ins,sg,m
строящий	строящий	V	partcp,sg,m,fut,act
решаемый	решать	V	partcp,sg,m,fut,pass
решаемого	решаемого	V	imper,sg,m,fut,pass
решаемому	решаемому	V	imper,sg,m,fut,pass
решаемым	решать	V	partcp,sg,fut,pass
строимый	строить	V	partcp,sg,m,fut,pass
решавший	решать	V	partcp,sg,m,past,act
решавшего	решавшего	V	imper,sg,m,past,act
решавшему	решавшему	V	imper,sg,past,act
решавшим	решать	V	partcp,sg,m,past,act
нёсший	нёсший	A	nom,sg,m,supr
нёсшего	нёсшего	A	gen,sg,supr
нёсшему	нёсшему	A	dat,sg,supr
нёсшим	нёсшим	A	ins,sg,m,supr
принесший	приносить	V	partcp,sg,m,past,act
написанный	написанный	V	partcp,sg,m,past,pass
написанного	написанного	V	imper,sg,m,past,pass
написанному	написанному	V	imper,sg,m,past,pass
написанным	написанным	A	ins,sg
построенный	построенный	V	partcp,sg,m,past,pass
взятый	взять	V	partcp,sg,m,past,pass
взятого	взятого	A	gen,sg,m
взятому	взятому	A	dat,sg
взятым	взять	V	partcp,sg,past,pass
взят	взять	V	partcp,sg,m,past,pass
взята	взять	V	partcp,sg,f,past,pass
взято	взять	V	partcp,sg,past,pass
раскрытая	раскрытый	A	nom,sg,f
книга	книга	S	nom,sg,f
раскрытая	раскрывать	V	partcp,sg,f,past,pass
мальчиком	мальчиком	S	ins,sg,m
книга	книга	S	nom,sg,f

1.4. К одной или разным леммам будет отнесены словоформы *нашедший* и *находившего*, *дал* и *давал*

дал, давать -- к одной лемме
нашедший, находившего -- к разным леммам (см 1.3)

1.5. Напишите правило пересчета тегов системы на теги из ЗС для анафорических местоимений (*он, она* и т.п.) и наречий

Правила реализованы для всех частей речи в скрипте toGS.py

2. Проведите функциональное тестирование выбранной Вами программы.

2.1. Для этого подберите примеры, содержащие сложные и проблемные случаи для морфологического анализа: например, незнакомые слова, случаи различных типов омонимии, сложные случаи с точки зрения определения частей речи (например, отглагольное прилагательное vs. причастие, частица vs. союз, наречие vs. краткое прилагательное и т.д.)

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.?

Ответ:

Числа размечаются тэгом NUM:

словоформа:	лемма:	POS:	
7	7	NUM	
9	9	NUM	
170	170	NUM	
478,6	478.6	NUM	
Оба	оба	NUM	nom

Слова, написанные в другом регистре не разбираются грамматически:

Desmay	desmay NP	
Bullie_Brake	bullie_brake	NP

Знаки препинания - каждый имеет индивидуальный тэг:

&	&	Fz
?	?	Fit
:	:	Fd
,	,	Fc

—	—	Fz
.	.	Fp
«	«	Fra
"	"	Fe

Слова через дефис остаются одним токеном, плохо разбираются грамматически:

бело-кремовое	бело-кремовое	A	acc,sg
Тянь-Шаня	тянь-шаня	NP	

Топонимы написанные через underscore, вместо пробела, остаются одним токеном и грамматически не описываются:

Православного_Музея	православного_музея	NP
Епископом_Лукианом	епископом_лукианом	NP
Западной_Европы	западной_европы	NP

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

Ответ:

Топонимы распознаются как NP (Proper nouns), но лемма и грам. характеристики не предсказываются вообще:

словоформа:	лемма:	POS:	грам. разбор:
Тянь-Шаня	тянь-шаня	NP	
Западной_Европы	западной_европы	NP	
Азии	азии	NP	
США	сша	NP	
Резиденции	резиденции	NP	
Газпрома	газпрома	NP	
Москвы	москвы	NP	
России	россии	NP	
Венгрии	венгрии	NP	

В сравнительно новых/ заимствованных/не часто употребляемых словах хорошо предсказываются грам. характеристики, но плохо предсказывается лемма:

энергобезопасности	энергобезопасности	S	gen,sg,f
достижением	достижением	S	
станции	станции	S	gen,sg,f
миллионов	миллионов	S	gen,pl,m
атаки	атака	S	nom,pl,f

чувством	чувством	S	ins,sg
синдрому	синдрому	S	dat,sg,m
беспокойством	беспокойством	S	ins,sg

Нестандартные прилагательные и глаголы также не точно разбираются грамматически/ неправильно подбирается начальная форма:

внутрибрюшной	внутрибрюшной	A	prep,sg,f
нефтегазовом	нефтегазовый	A	prep,sg,m
аутистического	аутистического	A	gen,sg,m
характеризуются	характеризуются	V	pl,pres

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.

В стандартном формате выдачи предлагается самый вероятный вариант:

Прекрасен прекрасен NP 1
пушкинский пушкинский AFSMIF000 0.286932
стих стих NCFSMI0000 0.279258

Ветер ветер NCNSMI0000 0.908468
совсем совсем D000 1
стих стихать VDSMS0F0A00 0.557358

В в B0 1
форточке форточка NCOSFI0000 0.621951
нет нет NCFSMI0000 0.999903
стекла стекло NCGSAI0000 0.361131

Вода вода NCNSFI0000 0.999828
еще еще D000 0.940663
не не T0 1
стекла стекло NCNPAI0000 0.359353

В формате выдачи morfo все варианты разбора:

Прекрасен прекрасен NP 1
пушкинский пушкинский ANSM0F000 0.713068 пушкинский AFSMIF000 0.286932
стих стихать VDSMS0F0A00 0.557358 стих NCFSMI0000 0.279258 стих NCNSMI0000 0.163384

Ветер ветер NCNSMI0000 0.908468 ветер NCFSMI0000 0.0914325 ветер NP 9.92753e-005
совсем совсем D000 1
стих стихать VDSMS0F0A00 0.557358 стих NCFSMI0000 0.279258 стих NCNSMI0000 0.163384

нет нет NCFSMI0000 0.999903 нет NCNSMI0000 9.72316e-005
стекла стекло NCGSAI0000 0.361131 стекло NCNPAI0000 0.359353 стекло NCFPAI0000 0.279339 стекать VDSFS0F0A00 0.000177809

Вода вода NCNSFI0000 0.999828 вод NCGSMI0A00 8.61846e-005 вода NP 8.61846e-005
еще еще D000 0.940663 еще T0 0.0593368
не не T0 1
стекла стекло NCGSAI0000 0.361131 стекло NCNPAI0000 0.359353 стекло NCFPAI0000 0.279339 стекать VDSFS0F0A00 0.000177809

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.)

Хорошо разбираются случаи причастие vs. прилагательное при наличии контекста:

блестящий	**блестеть**	**V**	**partcp,sg,m,fut,act**	
на	на	PR		
солнце	солнце	S	prep,sg	
камень	камень	S	nom,sg,m	
блестящий	**блестящий**	**A**	**nom,sg,m**	
ответ	ответ	S	nom,sg,m	

Хорошо разбираются случаи существительное vs. прилагательное:

бедных	**бедный**	**A**	**gen,pl**
людей	человек	S	gen,pl,m
количество	количество	S	nom,sg
бедных	**бедный**	**S**	**gen,pl,m**
в	в	PR	
России	россии	S	

Хорошо разбираются случаи падежной омонимии:

росли	расти	V	pl,past
березы	**береза**	**S**	**nom,pl,f**
нет	нет	S	acc,sg,m
ни	ни	PART	
одной	один	ANUM	gen,sg,f
березы	**береза**	**S**	**gen,sg,f**

Плохо разбираются случаи наречие vs. прилагательное:

жарко	**жаркий**	**A**	**sg,brev**
молилась	молилась	V	sg,f,past

Жарко	**жаркий**		**A**	**sg,brev**
в	в	PR		
небе	небо	S	prep,sg	
солнце	солнце	S	nom,sg	
летнее	летнее	A	nom,sg	

Плохо разбираются случаи омонимии различных местоименных форм:

Что	**что**	**SPRO**	**nom,sg**	
же	же	PART		
ты	ты	SPRO	sg	
не	не	PART		
едешь	ехать	V	sg,fut,2p	
?	?			
Я	я	SPRO	nom,sg	
писал	писать	V	sg,m,past	
вам	вы	SPRO	pl	
,	,			
что	**что**	**SPRO**	**nom,sg**	
нас	мы	SPRO	pl	
захватили		захватывать	V	pl,past
штили	штиль	S	acc,pl,m	

Плохо разбираются случаи глагол vs. числительное:

три	**три**	**NUM**	**acc**
года	год	S	gen,sg,m
Не	не	PART	
три	**три**	**NUM**	**nom**
так	так	ADVPRO	

Плохо разбираются случаи глагол vs. прилагательное:

помидор	помидор	S	nom,sg,m
спел	**спеть**	**V**	**sg,m,past**
и	и	CONJ	
свеж	свежий	A	sg,m,brev
спел	**спеть**	**V**	**sg,m,past**
песню	песня	S	acc,sg,f

3. Обработайте с помощью морфологического анализатора файл.

Для 500 словоформ из файла ЗС определите:

- уровень оставшейся неоднозначности: число элементов в Output(W) для всех слов тестируемого текста, поделенное на число слов в тексте. Если алгоритм работает однозначно, то этот параметр равняется 1.

Ответ:

Число слов в тексте (ЗС) = 500

Число элементов в Output(W) программы FreeLing = 491

Уровень оставшейся неоднозначности = 491/500 = **0,982**

- лексическая точность алгоритма - число слов текста, для которых лемма приписана правильно, поделенное на общее число слов которым система приписала какие-то теги

Ответ:

Общее число слов которым система приписала какие-то теги = 500

Число слов текста, для которых лемма приписана правильно = 299

Лексическая точность алгоритма = 299/500 = **0,598**