Морфологический анализ

Павел Браславский

ВВЕДЕНИЕ

Морфология

- Система форм изменения слов в каком-л. языке, а также раздел грамматики, изучающий формы слов. (МАС)
- Раздел лингвистики, основным объектом которого являются *слова естественных языков, их значимые части и морфологические признаки*. (Википедия)

Морфологический анализ

- Морфологический анализ – используется как предварительный этап обработки текста практически во всех задачах ОЕЯ
- Хорошо разработанная область
- Детерминированные и статистические подходы
- Вручную подготовленные данные и автоматические методы

Разделы морфологии

- Словоизменение (inflection)
  - *бежать– бегу – бежишь – бежит – бегут – бежите – … – бегущий – …*
- Словообразование (derivation)
  - *рыба – рыбка – рыбный – рыбарь – рыбак – рыбалка – зарыбление – рыбзавод – рыболов – рыбачить – …*

Русская морфология

- Русский язык имеет развитое словоизменение (флективный язык)

  → более свободный порядок слов (синтетичность)

Морфология в разных языках

- DE:
  Sauerstoffverbrauchsrate
  Lebensversicherungsreformgesetz
- EN:
  OK: сущ, прил, глаг, нареч, межд
- TR:
  Bayramlasamadiklarimiz –
  «те из нас, кого мы не можем поздравить с байрамом»

OK

See also: **ok**, **ők**, **ök**, **-ök** *and*

Contents [hide]

Термины

- Словоформа (word form) – слово в тексте *бегущий*
- Лемма (lemma) – словарная форма слова *бежать*
- Морфема (morpheme) – минимальная морфологическая единица *бег, ущ*
- Граммема (grammeme) – грамматические значения *причастие, мужской род, настоящее время*
- Парадигма – список словоформ одной лексемы

Части речи

- Имя существительное (noun) *топор, философия*
- Имя прилагательное (adjective) *красивый*
- Местоимение (pronoun) *я, ты, этот, мой*
- Глагол (verb) *петь, рисовать*
- Причастие (participle) *бегущий, поющий*
- Деепричастие (gerund) пробегая, напевая
- Наречие (adverb) *красиво, быстро*
- Имя числительное (numeral) *два, третий*
- Предлог (preposition), союз (conjunction), частица (particle), междометье (interjection)

падеж	ед. ч.	мн. ч.
Им.	ча́й	чаи́
Р.	ча́я	чаёв
Д.	ча́ю	чая́м
В.	ча́й	чаи́
Тв.	ча́ем	чая́ми
Пр.	ча́е	чая́х
Разд.	ча́ю	—

падеж	ед. ч.	мн. ч.
Им.	И́ра	И́ры
Р.	И́ры	И́р
Д.	И́ре	И́рам
В.	И́ру	И́р
Тв.	И́рой И́рою	И́рами
Пр.	И́ре	И́рах
Зв.	Ир	—

падеж	ед. ч.	мн. ч.		
		обычное	причисл.	уст./шутл.
Им.	челове́к	лю́ди	—	челове́ки
Р.	челове́ка	люде́й	челове́к	челове́ков
Д.	челове́ку	лю́дям	челове́кам	челове́кам
В.	челове́ка	люде́й	—	челове́ков
Тв.	челове́ком	людьми́	челове́ками	челове́ками
Пр.	челове́ке	лю́дях	челове́ках	челове́ках
Зв.	челове́че	—	—	—

падеж	ед. ч.	мн. ч.
Им.	бе́рег	берега́
Р.	бе́рега	берего́в
Д.	бе́регу	берега́м
В.	бе́рег	берега́
Тв.	бе́регом	берега́ми
Пр.	бе́реге	берега́х
М.	берегу́	—

падеж		ед. ч.			мн. ч.
		муж. р.	ср. р.	жен. р.	
Им.		кра́сный	кра́сное	кра́сная	кра́сные
Рд.		кра́сного	кра́сного	кра́сной	кра́сных
Дт.		кра́сному	кра́сному	кра́сной	кра́сным
Вн.	одуш.	кра́сного	кра́сное	кра́сную	кра́сных
	неод.	кра́сный			кра́сные
Тв.		кра́сным	кра́сным	кра́сной кра́сною	кра́сными
Пр.		кра́сном	кра́сном	кра́сной	кра́сных
Кратк. форма		кра́сен	кра́сно красно́	красна́	кра́сны красны́

	наст.	прош.	повелит.
Я	пью́	пи́л пила́	—
Ты	пьёшь	пи́л пила́	пе́й
Он Она Оно	пьёт	пи́л пила́ пи́ло	—
Мы	пьём	пи́ли	—
Вы	пьёте	пи́ли	пе́йте
Они	пью́т	пи́ли	—
Пр. действ. наст.	пью́щий		
Пр. действ. прош.	пи́вший		
Деепр. наст.	—		
Деепр. прош.	пи́в, пи́вши		
Пр. страд. наст.	пи́тый		
Будущее	буду/будешь… пи́ть		

падеж		ед. ч.			мн. ч.
		муж. р.	ср. р.	жен. р.	
Им.		пью́щий	пью́щее	пью́щая	пью́щие
Рд.		пью́щего	пью́щего	пью́щей	пью́щих
Дт.		пью́щему	пью́щему	пью́щей	пью́щим
Вн.	одуш.	пью́щего	пью́щее	пью́щую	пью́щих
	неод.	пью́щий			пью́щие
Тв.		пью́щим	пью́щим	пью́щей пью́щею	пью́щими
Пр.		пью́щем	пью́щем	пью́щей	пью́щих

падеж		ед. ч.			мн. ч.
		муж. р.	ср. р.	жен. р.	
Им.		пивший	пившее	пившая	пившие
Рд.		пившего	пившего	пившей	пивших
Дт.		пившему	пившему	пившей	пившим
Вн.	одуш.	пившего	пившее	пившую	пивших
	неод.	пивший			пившие
Тв.		пившим	пившим	пившей пившею	пившими
Пр.		пившем	пившем	пившей	пивших

	будущ.	прош.	повелит.
Я	*побежу́	победи́л победи́ла	—
Ты	победи́шь	победи́л победи́ла	победи́
Он Она Оно	победи́т	победи́л победи́ла победи́ло	—
Мы	победи́м	победи́ли	—
Вы	победи́те	победи́ли	победи́те
Они	победя́т	победи́ли	—
Пр. действ. прош.	победи́вший		
Деепр. прош.	победи́в, победи́вши		
Пр. страд. прош.	побеждённый		

Задачи морфологического анализа

- Лемматизация (стемминг) – приведение словоформ к словарной форме (основе)
- Грамматические характеристики (POS-tagging, частеречная разметка)
- Генерация (постановка слова в заданную форму)

Лемматизация – пример

В больничном дворе стоит небольшой флигель, окруженный целым лесом репейника, крапивы и дикой конопли.

в больничный двор стоять небольшой флигель окружать целый лес репейник крапива и дикий конопля

Стемминг – пример

В больничном дворе стоит небольшой флигель, окруженный целым лесом репейника, крапивы и дикой конопли.

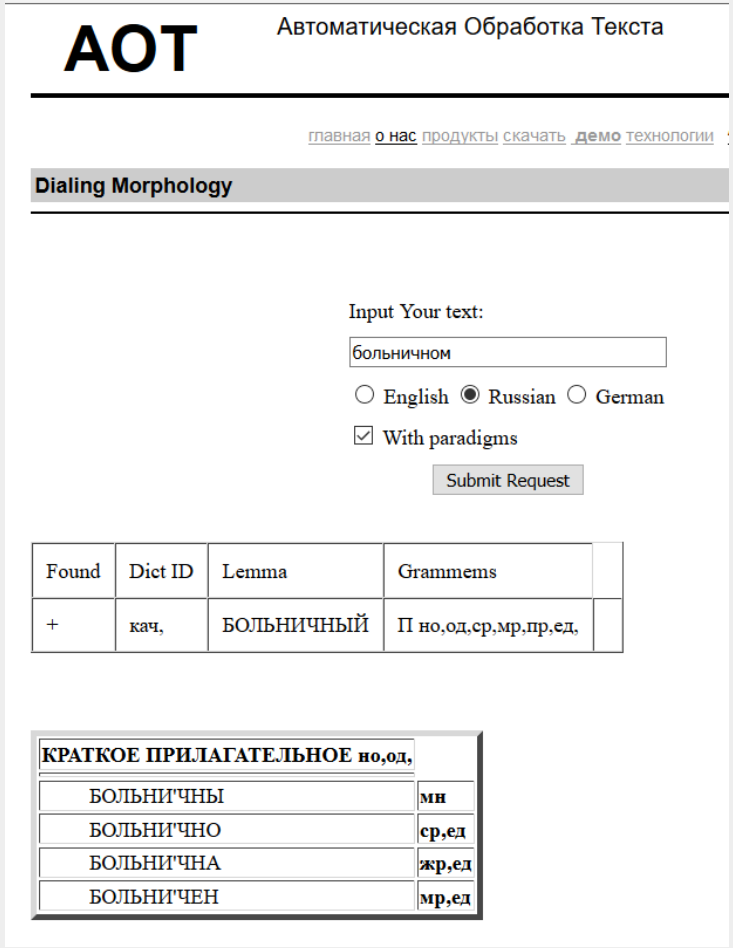
в больничн двор сто небольш флигел окружен цел лес репейник крапив и дик конопл

Грамматический разбор

В больничном дворе стоит небольшой флигель, окруженный целым лесом репейника, крапивы и дикой конопли.

{в=PR=}
{больничный=A=пр,ед,полн,муж}
{двор=S,муж,неод=пр,ед}
{стоять=V,несов,нп=непрош,ед,изъяв,3-л}
{небольшой=A=им,ед,полн,муж}
{флигель=S,муж,неод=им,ед}
{окружать=V,пе=прош,им,ед,прич,полн,муж,сов,страд}
{целый=A=твор,ед,полн,муж}
{лес=S,муж,неод=твор,ед}
{репейник=S,муж,неод=род,ед}
{крапива=S,жен,неод=род,ед}
{и=CONJ=}
{дикий=A=род,ед,полн,жен}
{конопля=S,жен,неод=род,ед}

Генерация



AOT Автоматическая Обработка Текста

главная о нас продукты скачать демо технологии

Dialing Morphology

Input Your text:

больничном

○ English ◉ Russian ○ German

☑ With paradigms

Submit Request

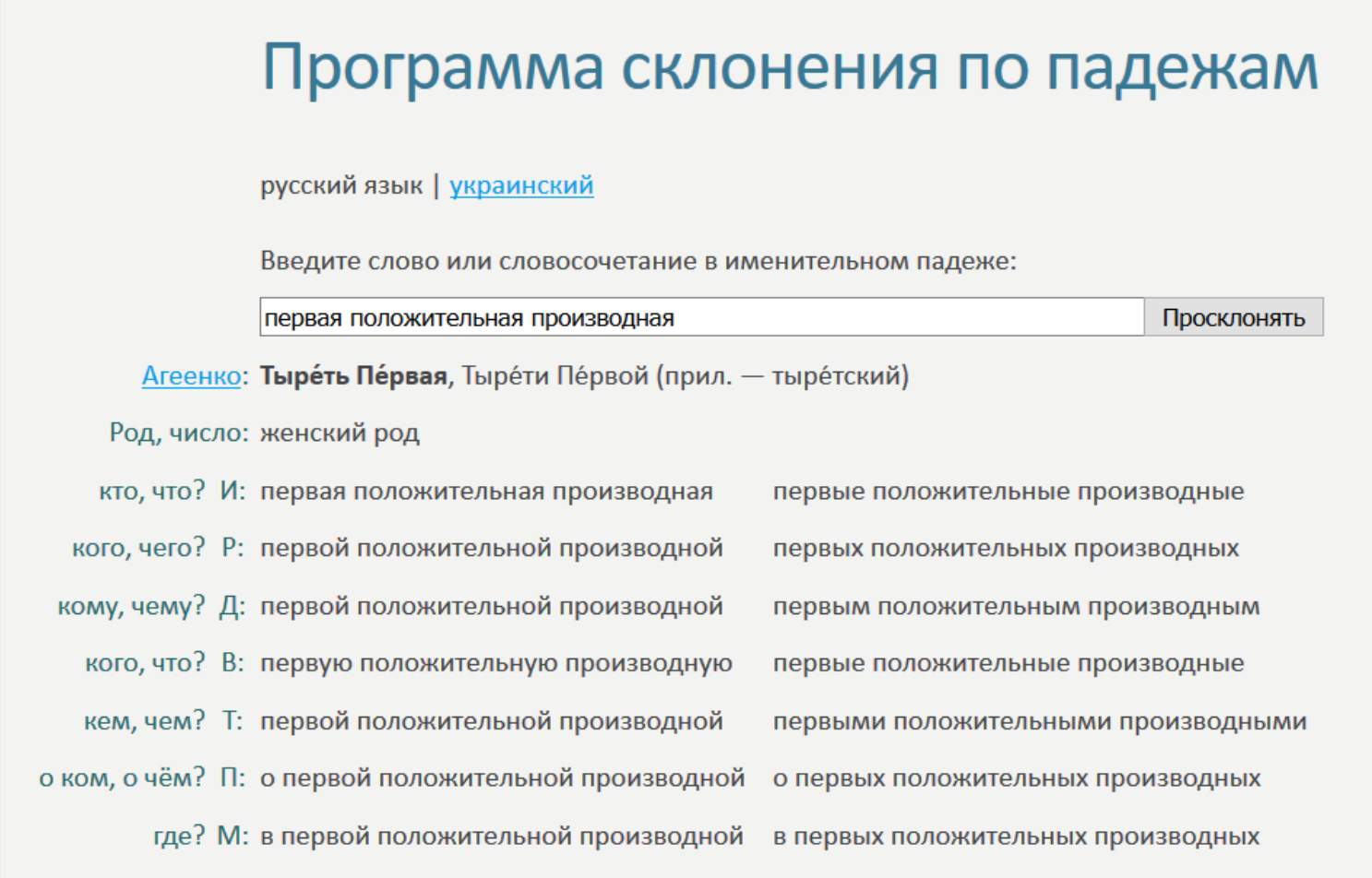
Found	Dict ID	Lemma	Grammems	
+	кач,	БОЛЬНИЧНЫЙ	П но,од,ср,мр,пр,ед,	

КРАТКОЕ ПРИЛАГАТЕЛЬНОЕ но,од,	
БОЛЬНИ'ЧНЫ	мн
БОЛЬНИ'ЧНО	ср,ед
БОЛЬНИ'ЧНА	жр,ед
БОЛЬНИ'ЧЕН	мр,ед

http://aot.ru/demo/morph.html

ПРИЛАГАТЕЛЬНОЕ	
мр,ед	
БОЛЬНИ'ЧНОГО	од,вн
БОЛЬНИ'ЧНЫЙ	но,вн
БОЛЬНИ'ЧНЫЙ	но,од,им
БОЛЬНИ'ЧНОГО	но,од,рд
БОЛЬНИ'ЧНОМУ	но,од,дт
БОЛЬНИ'ЧНЫМ	но,од,тв
БОЛЬНИ'ЧНОМ	но,од,пр
но,од,жр,ед	
БОЛЬНИ'ЧНАЯ	им
БОЛЬНИ'ЧНОЙ	рд
БОЛЬНИ'ЧНОЙ	дт
БОЛЬНИ'ЧНУЮ	вн
БОЛЬНИ'ЧНОЮ	тв
БОЛЬНИ'ЧНОЙ	тв
БОЛЬНИ'ЧНОЙ	пр
но,од,ср,ед	
БОЛЬНИ'ЧНОЕ	им
БОЛЬНИ'ЧНОГО	рд
БОЛЬНИ'ЧНОМУ	дт
БОЛЬНИ'ЧНОЕ	вн
БОЛЬНИ'ЧНЫМ	тв
БОЛЬНИ'ЧНОМ	пр
мн	
БОЛЬНИ'ЧНЫХ	од,вн
БОЛЬНИ'ЧНЫЕ	но,вн
БОЛЬНИ'ЧНЫЕ	но,од,им
БОЛЬНИ'ЧНЫХ	но,од,рд
БОЛЬНИ'ЧНЫМ	но,од,дт
БОЛЬНИ'ЧНЫМИ	но,од,тв
БОЛЬНИ'ЧНЫХ	но,од,пр
сравн,но,од	
БОЛЬНИ'ЧНЕЙ	
БОЛЬНИ'ЧНЕЕ	
ПОБОЛЬНИ'ЧНЕЙ	2
ПОБОЛЬНИ'ЧНЕЕ	2

Генерация – 2



Программа склонения по падежам

русский язык | украинский

Введите слово или словосочетание в именительном падеже:

первая положительная производная | Просклонять

Агеенко: **Тыре́ть Пе́рвая**, Тыре́ти Пе́рвой (прил. — тыре́тский)

Род, число: женский род

кто, что? И:	первая положительная производная	первые положительные производные
кого, чего? Р:	первой положительной производной	первых положительных производных
кому, чему? Д:	первой положительной производной	первым положительным производным
кого, что? В:	первую положительную производную	первые положительные производные
кем, чем? Т:	первой положительной производной	первыми положительными производными
о ком, о чём? П:	о первой положительной производной	о первых положительных производных
где? М:	в первой положительной производной	в первых положительных производных

http://morpher.ru/Demo.aspx

Генерация – pymorphy2

```
>>> pymorphy2.MorphAnalyzer().parse('кочерга')[0].inflect({'plur', 'gent'})
Parse(word='кочерёг', tag=OpencorporaTag('NOUN,inan,femn plur,gent'), norma
l_form='кочерга', score=1.0, methods_stack=((<DictionaryAnalyzer>, 'кочерёг
', 1794, 8),))
```

Сложности

- Неоднозначность
- Новые/неизвестные слова

МЕТОДЫ

Подходы

- Сгенерировать все данные и по ним искать
  - Структура данных?
- Обобщить типичные случаи, сформулировать правила

Как сгенерировать данные?

Зализняк А.А. Грамматический словарь русского языка (1977)

~100 тыс. входов

Модель русского словоизменения

Пример: лев мо 1*b (животное)

лев м 1a (денежная единица)

стричь нсв 8b (-г-)

прихожая ж (п 4a)

Основа большинства машинных морфологий РЯ

СКЛОНЕНИЕ

ЗНАЧЕНИЕ БУКВЕННЫХ СИМВОЛОВ И ЭЛЕМЕНТОВ ИНДЕКСА У ИМЕН[1]

БУКВЕННЫЕ СИМВОЛЫ

Основной буквенный символ — это символ, начинающий словарную статью (или подстатью). Символы «мн. одуш.», «мн. неод.», «мн. *от*» состоят из двух частей; прочие буквенные индексы при именах — единые (некоторые из них пишутся через дефис).

Если в словарной статье есть угловые скобки, то запись внутри этих скобок начинается с дополнительного буквенного символа. Дополнительные буквенные символы берутся из того же общего набора, что и основные.

Буквенные символы указывают: 1) основную синтаксическую характеристику имени (часть речи, ред. одушевленность—неодушевленность); 2) основную морфологическую характеристику имени (принадлежность к субстантивному, адъективному, местоименному склонению, к склонению числительных).

В нормальном случае между основной синтаксической и основной морфологической характеристиками имеется стандартное соотношение: существительные относятся к субстантивному склонению, прилагательные — к адъективному и т. д. В этом случае основной буквенный символ выполняет одновременно две роли: указывает основную синтаксическую и основную морфологическую характеристики слова. Например, **красúвый** п 1а: символ «п» указывает на то, что это слово — прилагательное, и на принадлежность его к адъективному склонению.

Гораздо реже слово склоняется по образцу слов другой («не своей») части речи (или другого рода), например, **запятáя, мужчúна**. В этом случае в словарной статье даются угловые скобки; основной буквенный символ указывает основную синтаксическую характеристику слова, дополнительный буквенный символ — основную морфологическую характеристику слова. Например, **запятáя** ж ⟨п 1b⟩: «ж» указывает на то, что это существительное женского рода, неодушевленное, «п» — на то, что это слово склоняется по адъективному склонению.

Основная синтаксическая характеристика имени

В роли указателей основной синтаксической характеристики имени основные буквенные символы имеют следующие значения:

м — существительное мужского рода неодушевленное

мо — существительное мужского рода одушевленное

ж — существительное женского рода неодушевленное

жо — существительное женского рода одушевленное

с — существительное среднего рода неодушевленное

со — существительное среднего рода одушевленное

мо-жо — существительное так наз. общего рода одушевленное (т. е. выступающее как существительное мужского рода при обозначении мужчин, женского рода — при обозначении женщин)

мн. (без слова *от*), мн. неод., мн. одуш.— существительные группы pluralia tantum (см. стр. 5); одушевленность — неодушевленность, если она не указана прямо, определяется дополнительным буквенным символом (который в этом случае обязательно имеется)

[1] Данный раздел «Грамматических сведений» (стр. 25—35) предназначен только для тех читателей, которые пожелают строить формы слов непосредственно по индексу, не обращаясь к образцам склонения. Для построения форм по образцам этот раздел не нужен.

мн. *от* — форма мн. числа от существительного, употребляющегося также и в единственном числе (см. стр. 5) и обозначенного после слова *от* его основным буквенным символом и индексом (или приведенного непосредственно)

п — прилагательное

мс — местоимение (точнее, местоимение-существительное; см. также стр. 6)

мс-п — местоименное прилагательное

числ.— числительное (количественное или собирательное; см. также стр. 6)

числ.-п — порядковое числительное (= счетное прилагательное)

Основная морфологическая характеристика имени

В роли указателей основной морфологической характеристики имени буквенные символы (основные или дополнительные) имеют следующие значения:

м, мо, ж, жо, с, со, мо-жо — субстантивное склонение

п — адъективное склонение

мс, мс-п — местоименное склонение

числ.— склонение числительных

Символы «мн.» (с уточнениями или без них) и «числ.-п.» в роли указателей морфологической характеристики не используются.

Субстантивное, адъективное и местоименное склонения характеризуются своими окончаниями. Ниже приводятся стандартные (т. е. наиболее распространенные и потому принимаемые за норму) окончания этих трех склонений. Что касается склонения числительных, то оно не представляет собой чего-либо единого и имеет много аномалий; поэтому в словаре при числительных непосредственно приводится их парадигма (или дается соответствующая отсылка); см. также стр. 6.

Окончания приводятся ниже в **графической** (не фонемной) форме. Они даны в двух вариантах: слева от косой черты окончания типа 1, справа — типа 2. (С содержательной точки зрения, окончания типа 1 — это стандартные окончания для основ на твердую согласную, типа 2 — для основ на мягкую согласную; подробнее о типах см. ниже, раздел «Цифра индекса».) Запись без косой черты означает, что окончания типов 1 и 2 здесь одинаковы. В некоторых случаях имеются, кроме того, варианты окончания, зависящие от места ударения (запись: *безуд.* ...; *уд.* ...).

Стандартные окончания субстантивного склонения

		Ед.			Мн.		
		м	с	ж	м	с	ж
И.		нуль/-ь	-о/*безуд.* -е / *уд.* -ё	-а/-я	-ы/-и	-а/-я	-ы/-и
Р.		-а/-я		-ы/-и	-ов/-ей	нуль/*безуд.* -ь / *уд.* -ей	
Д.		-у/-ю		-е	-ам/-ям		
В.	неод.	= И.	= И.	-у/-ю	= И.		
	одуш.	= Р.			= Р.		
Т.		-ом/*безуд.* -ем / *уд.* -ём		-ой [1]/*безуд.* -ей [1] / *уд.* -ёй [1]	-ами/-ями		
П.		-е			-ах/-ях		

[1] В этой форме имеются, кроме того, параллельные варианты окончаний: -ою, -ею, -ёю. В современном языке, в отличие от языка XIX века, эти варианты употребляются редко (преимущественно в поэзии).

Стандартные окончания адъективного склонения

Полные формы

		Ед. м	Ед. с	Ед. ж	Мн. (всех родов)
И.		*безуд.* -ый / *уд.* -ой / -ий	-ое/-ее	-ая/-яя	-ые/-ие
Р.		-ого/-его		-ой/-ей	-ых/-их
Д.		-ому/-ему		-ой/-ей	-ым/-им
В.	неод.	= И.	= И.	-ую/-юю	= И.
	одуш.	= Р.			= Р.
Т.		-ым/-им		-ой [1]/-ей [1]	-ыми/-ими
П.		-ом/-ем		-ой/-ей	-ых/-их

Краткие формы

Ед. м	Ед. с	Ед. ж	Мн. (всех родов)
нуль/-ь	-о / *безуд.* -е / *уд.* -ё	-а/-я	-ы/-и

Стандартные окончания местоименного склонения

		Ед. м	Ед. с	Ед. ж	Мн. (всех родов)
И.		нуль/-ь	-о / *безуд.* -е / *уд.* -ё	-а/-я	-ы/-и
Р.		-ого/-его		-ой/-ей	-ых/-их
Д.		-ому/-ему		-ой/-ей	-ым/-им
В.	неод.	= И.	= И.	-у/-ю	= И.
	одуш.	= Р.			= Р.
Т.		-ым/-им		-ой [1]/-ей [1]	-ыми/-ими
П.		-ом / *безуд.* -ем / *уд.* -ём		-ой/-ей	-ых/-их

[1] См. сноску к таблице окончаний субстантивного склонения.

Выбор нужной группы окончаний в приведенных выше таблицах

Прилагательные (обычные, местоименные, счетные). Прилагательные адъективного склонения, не местоименные, не счетные и не оканчивающиеся на **-ийся**, имеют весь набор форм адъективного склонения (т. е. все полные и краткие формы). Прочие прилагательные адъективного склонения (т. е. местоименные, счетные и оканчивающиеся на **-ийся**) имеют все полные формы, но не имеют кратких форм. Прилагательные местоименного склонения (любого типа) имеют весь набор форм местоименного склонения. Прилагательных субстантивного склонения не бывает.

Существительные имеют фиксированный род и являются либо одушевленными, либо неодушевленными. Соответственно в таблице (нужного склонения) берутся окончания только нужного рода и только нужной степени одушевленности. Например, для слова слон мо 1b в таблице субстантивного склонения берутся окончания мужского рода, а там, где различаются одушевленность и неодушевленность (в В. падеже),— одушевленные. Пример существительного местоименного склонения: **тре́тье** с ⟨мс 6*a⟩; для него в таблице местоименного склонения берутся окончания среднего рода, неодушевленные. Для существительных адъективного склонения из соответствующей таблицы берутся только окончания полных форм (краткие форм у существительных не бывает). Например, для слова **запята́я** ж ⟨п 1b⟩ в таблице адъективного склонения (раздел «полные формы») берутся окончания женского рода, неодушевленные.

Для существительных общего рода (т. е. с буквенным символом «мо-жо») берутся окончания женского рода, одушевленные (т. е. как при символе «жо»). Если в статье существительного субстантивного склонения имеются угловые скобки с дополнительным символом рода и одушевленности—неодушевленности, то окончания берутся в таблице в соответствии с этим дополнительным символом. Например, для слова **мужчи́на** мо ⟨жо 1a⟩ в таблице субстантивного склонения берутся окончания женского рода, одушевленные. Для существительных с буквенным символом «мн.» следует брать (с соблюдением всех предыдущих правил) только окончания мн. числа.

Для местоимений, а также для числительных (количественных и собирательных) в словарной статье непосредственно приводится их парадигма или дается соответствующая отсылка.

Примечания. 1) О сокращенном оформлении словарной статьи существительного в случаях типа **полице́йский** (п 3a✕—; мо) см. стр. 17.

2) Для существительных с пометой «мн. *от*» следует вначале построить форму И. ед. (см. правила на стр. 44, 50, 55, раздел Ж), после чего действовать по обычным правилам.

3) В нескольких особых случаях склонение слов, снабженных угловыми скобками, отклоняется в некоторых деталях от того, что вытекает из указаний, стоящих в угловых скобках. Эти случаи таковы: а) у одушевленных существительных мужского рода, склоняющихся по образцу среднего рода (например, **воронко́** мо ⟨со 3*b①②⟩), В. ед. = Р. ед.; ср. стр. 34; б) прилагательные и существительные местоименного склонения, оканчивающиеся на **-ин, -ов** (например, **дя́дин** п ⟨мс 1a⟩, **отцо́в** п ⟨мс 1a⟩, **ка́бельтов** м ⟨мс 1a⟩), а также **господень**, имеют некоторые особенности в окончаниях Р. ед., Д. ед. и П. ед.; см. стр. 63.

ЦИФРА ИНДЕКСА

Цифра индекса (от 0 до 8) обозначает тип склонения.

Цифра 0 означает, что слово неизменяемо (т. е. все формы внешне одинаковы).

Цифры 1—7 означают разные типы склонения, зависящие от того, на что оканчивается основа слова. Цифра 8 означает особый тип склонения (а именно, традиционное «3-е склонение существительных»), отличающийся сразу от всех типов 1—7[1].

Для различения типов 1—7 необходимо понятие «графической основы» слова. Под графической основой понимается буквенная последовательность, получаемая из исходной формы слова следующими операциями:

у слов субстантивного и местоименного склонений: если они оканчиваются на гласную, **-й** или **-ь** — отбрасывается одна конечная буква; если они оканчиваются на согласную (кроме й), графическая основа равна исходной форме;

у слов адъективного склонения (кроме оканчивающихся на **-ся**) отбрасываются две конечных буквы; у слов адъективного склонения на **-ся** отбрасываются четыре конечных буквы.

Примеры выделения графической основы: **аку́л-а**; **лыжн-я**, **ведр-о́**, **бель-ё**, **кра-й**, **то́пол-ь**, **ли́си-й**; **по́езд**, **стол**, **дя́дин**, **отцо́в**; **бе́л-ый**, **син-ий**, **часов-о́й**, **запят-а́я**, **лёгк-ое**, **пресмыка́ющ-ееся**.

Склонение слов разных типов различается окончаниями. Окончания типов 1 и 2 (основных) приведены в таблицах на стр. 26—27. Окончания типов 3—7 представляют собой разные ком-

[1] Таким образом, было бы более последовательно обозначить типы 1—7 вместе как единый класс 1 (разделяющийся на типы 1.1—1.7), которому противопоставлен класс 11 (т. е. тип 8). Используемая в настоящем словаре единая нумерация принята исключительно в целях упрощения символики.

Коды в Викисловаре

за-ря́

Существительное, неодушевлённое, женский род, 1-е склонение (тип склонения 2d по классификации А. А. Зализняка).

падеж	ед. ч.	мн. ч.
Им.	заря́	зо́ри
Р.	зари́	зо́рь
Д.	заре́	зо́рям
В.	зарю́	зо́ри
Тв.	зарёй зарёю	зо́рями
Пр.	заре́	зо́рях

бе-жа́ть

Глагол, двувидовой (может образовывать формы совершенного и несовершенного вида), переходный, тип спряжения по классификации А. Зализняка — 5bX^. В некоторых значениях «бежать» может использоваться как глагол совершенного вида, особенно в формах прош. времени: ◆ преступники **бежали** из тюрьмы ◆ враг **бежал** ◆ он **бежал** городской суеты ◆ суп **бежал** . В 3л. мн.ч. употребляется нестандартная форма «*бегу́т*». До XIX также равноправной формой была форма «*бежа́т*» (см. пример). Является переходным глаголом только в значении *пробега́ть какую-л. дистанцию* (**бежать** стометровку).

Корень: **-беж-**; суффикс: **-а**; глагольное окончание: **-ть** [Тихонов, 1996].

	наст./будущ.	прош.	повелит.
Я	бегу́	бежа́л бежа́ла	—
Ты	бежи́шь	бежа́л бежа́ла	беги́
Он Она Оно	бежи́т	бежа́л бежа́ла бежа́ло	—
Мы	бежи́м	бежа́ли	—
Вы	бежи́те	бежа́ли	беги́те
Они	бегу́т△	бежа́ли	—
Пр. действ. наст.	бегу́щий		
Пр. действ. прош.	бежа́вший		

Шаблоны словоизменения

- **Цифра индекса:**
  - 1 — слова с основой на твёрдый согласный (топор, комод, балда, кобра, олово, пекло; твёрдый, тусклый)
  - 2 — слова с основой на мягкий согласный (тюлень, искатель, цапля, Дуня, горе, поле; весенний)
  - 3 — слова с основой на г, к или х (петух, сапог, неряха, коряга, золотко, мягкий)
  - 4 — слова с основой на ж, ш, ч, щ (калач, лаваш, галоша, святоша, жилище, вече, вящий)
  - 5 — слова с основой на ц (немец, конец, девица, деревце, куцый)
  - 6 — слова с основой на гласный (кроме и) или й/j (бой, край; шея, здоровье)
  - 7 — слова с основой на и (полоний, сложение, мания, удостоверение)
  - 8 — слова с традиционным «3-м склонением» (боль, тетрадь, зыбь)
  - *Помощь в выборе цифры индекса*
- **Звездочка при цифре** * означает чередование в основе беглой гласной с нулём (1*a, 3*a и т. п.; платок, кошка, сердце).
- **Кружочек при цифре** ° обозначает признак особого систематического чередования (например, неравносложности основ или замены суффикса при словоизменении): время/времена́, мышонок/мыша́та, шу́рин - шурья́, господи́н - господа́.
- **Буква индекса**
  - a — ударение всегда на основу (парад, спонсор, мама, солнце, платёжный)
  - b — ударение всегда вне основы, если кроме основы вообще что-либо есть (топор, похвала, вещество, родной)
  - c — ударение на основу в ед. ч. и вне основы во мн. ч. (дар, место, поле)
  - d — ударение на окончание в ед. ч. и на основу во мн. ч. ({{сущ ru f ina 1d}}: *беда́ - бе́ды*)
  - e — ударение на основу в ед. ч. и им. п. мн. ч., вне основы в остальных падежах мн. ч. ({{сущ ru f ina 2e}}: *до́ля - до́ли - доля́ми*)
  - f — ударение на основу в им. п. мн. ч. и вне основы в остальных случаях ({{сущ ru f ina 1f}}: *слеза́ - слёзы - слеза́ми*)
  - d' или f' — ударение на основу в вин. п. ед. ч. ({{сущ ru f ina 1d'}}: *стена́ - сте́ну*; {{сущ ru f ina 1f'}}: *борода́ - бо́роду*)
  - f'' — ударение на основу в тв. п. ед. ч. ({{сущ ru f ina 8f''}}: *глушь - глуши́ - глу́шью*)

Пример: словарь oDict

```
1  бинокль,м,бинокля,биноклю,бинокль,биноклем,бинокле,бинокле,бинокли,биноклей,бин
2  бинокулярный,п,бинокулярного,бинокулярному,бинокулярный,бинокулярным,бинокулярн
3  бином,м,бинома,биному,бином,биномом,биноме,биноме,биномы,биномов,биномам,биномы
4  биномиальный,п,биномиального,биномиальному,биномиальный,биномиальным,биномиальн
5  биноминальный,п,биноминального,биноминальному,биноминальный,биноминальным,бином
6  бинормаль,ж,бинормали,бинормали,бинормаль,бинормалью,бинормали,бинормали,бинорм
7  бинт,м,бинта,бинту,бинт,бинтом,бинте,бинте,бинты,бинтов,бинтам,бинты,бинтами,би
8  бинтик,м,бинтика,бинтику,бинтик,бинтиком,бинтике,бинтике,бинтики,бинтиков,бинти
9  бинтовать,нсв,бинтовал,бинтовала,бинтовало,бинтовали,бинтую,бинтуешь,бинтует,би
10 бинтоваться,нсв,бинтовался,бинтовалась,бинтовалось,бинтовались,бинтуюсь,бинтуеш
11 бинтовка,ж,бинтовки,бинтовке,бинтовку,бинтовкой,бинтовке,бинтовке,бинтовки,бинт
12 бинтовой,п,бинтового,бинтовому,бинтовой,бинтовым,бинтовом,бинтовом,бинтовая,бин
13 биоаналог,м,биоаналога,биоаналогу,биоаналог,биоаналогом,биоаналоге,биоаналоге,б
14 биобиблиография,ж,биобиблиографии,биобиблиографии,биобиблиографию,биобиблиограф
15 биогенез,м,биогенеза,биогенезу,биогенез,биогенезом,биогенезе,биогенезе,биогенез
16 биогенетический,п,биогенетического,биогенетическому,биогенетический,биогенетиче
17 биогенный,п,биогенного,биогенному,биогенный,биогенным,биогенном,биогенном,биоге
18 биогеография,ж,биогеографии,биогеографии,биогеографию,биогеографией,биогеографи
19 биограф,мо,биографа,биографу,биографа,биографом,биографе,биографе,биографы,биог
20 биографический,п,биографического,биографическому,биографический,биографическим,
```

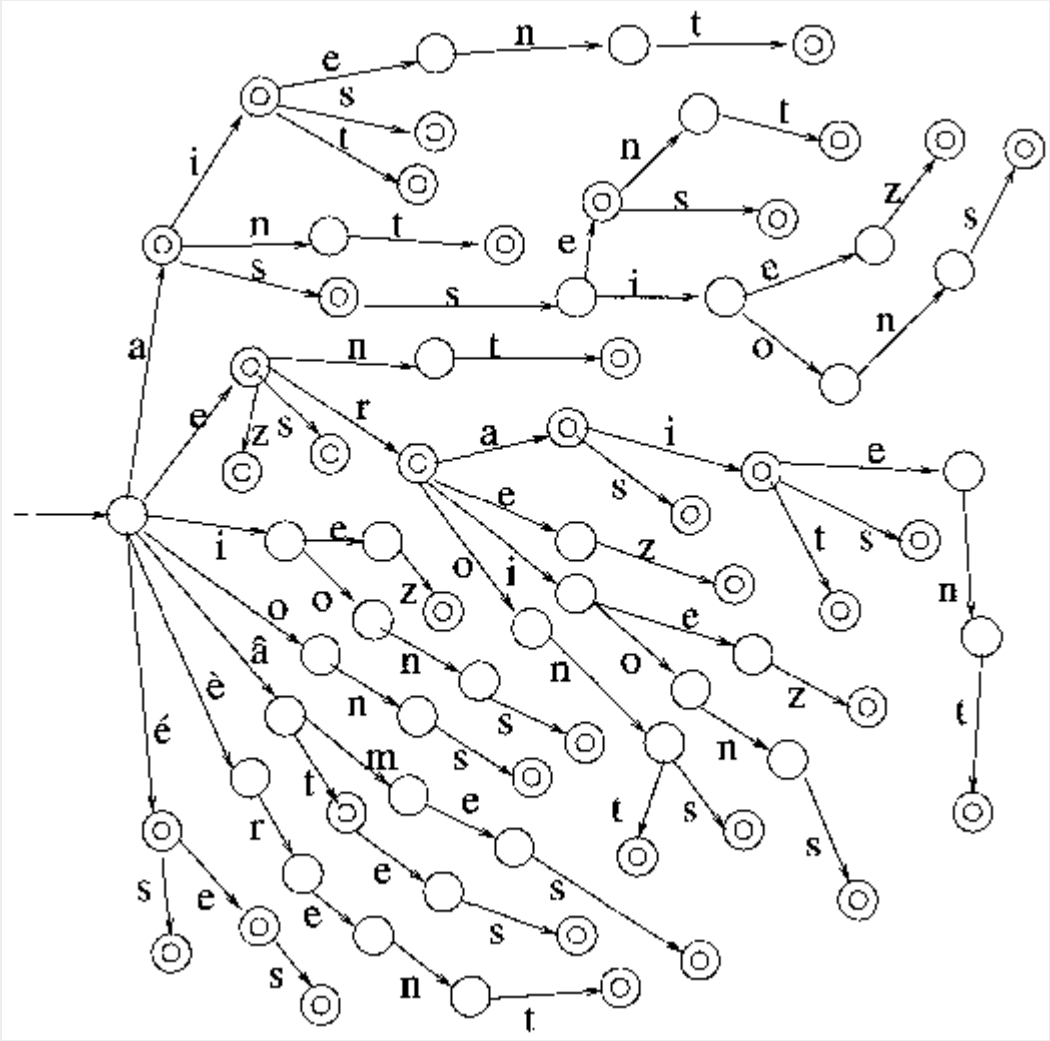
Структуры данных

- Tries – *mystem*
- Finite State Automata (FSA) – *aot, pymorphy2*
- Finite State Transducers – *английская морфология*

Критерии:

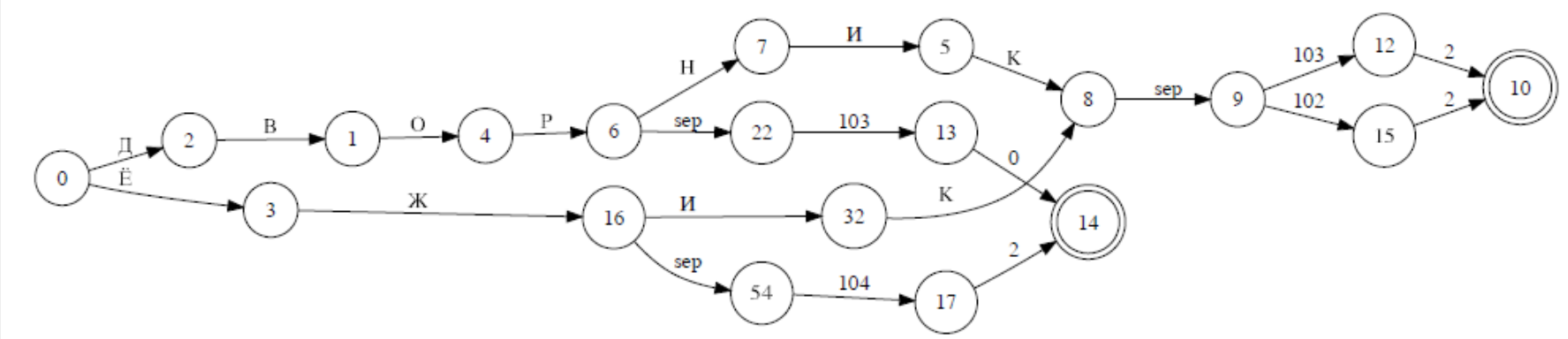
- Производительность
- Память
- Анализ/генерация
- Незнакомые слова

Префиксное дерево (trie)



[Daciuk et al., 2000]

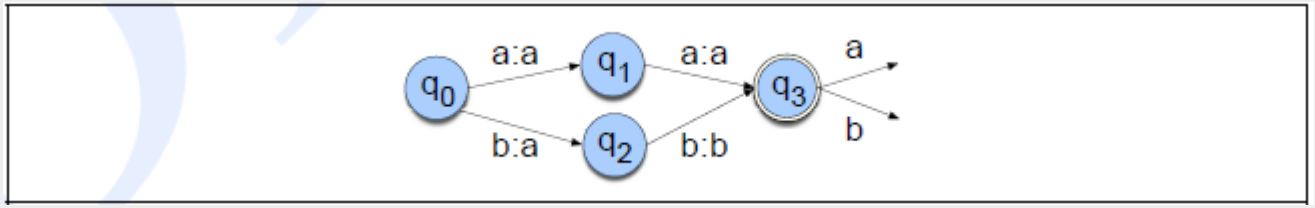
Конечный автомат



[Korobov, 2015]

Finite State Transducer, FST

- Обобщение конечных автоматов: переходы маркированы парами символов
- Переводят строки входного алфавита в строки выходного алфавита
  (cats, cat+N+PL), (walked, walk+V+PAST),…
- Хорошие алгебраические свойства (обратимость, композициональность)

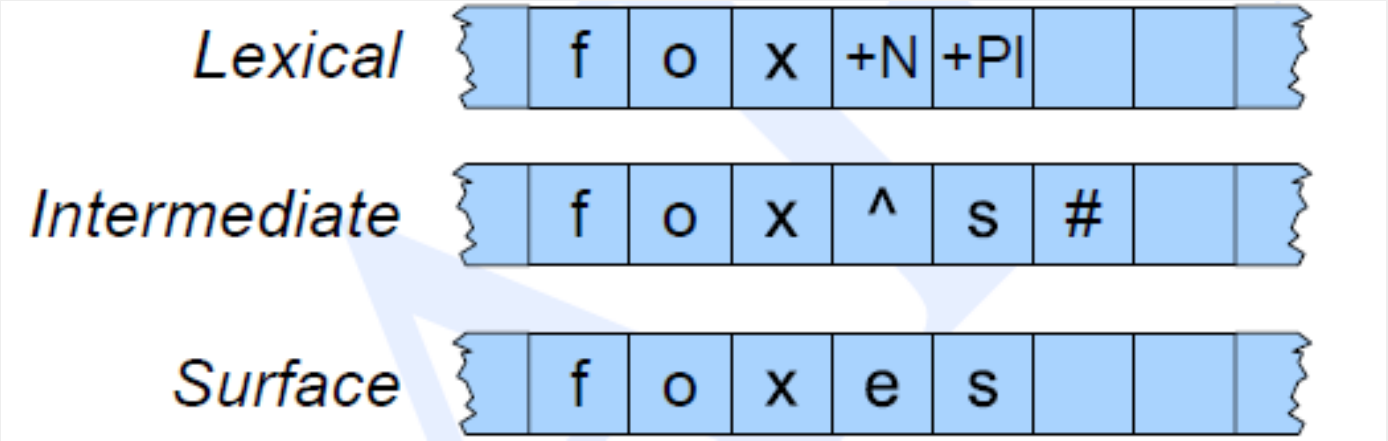


[SLP]

Три компонента морфологии

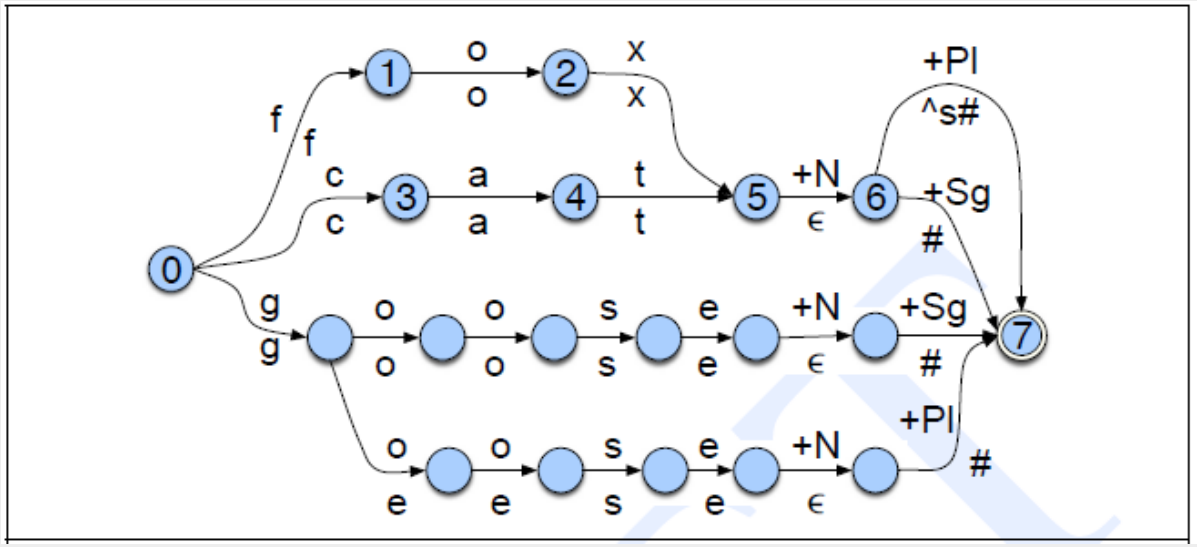
- Словарь (lexicon): основы (stems) + аффиксы (affixes) + свойства (части речи, число)
- Морфологические правила (morphotactics) – правила «нанизывания» морфем (например: *un*lock, cat*s*)
- Орфографические правила (party + -s →part*ie*s)

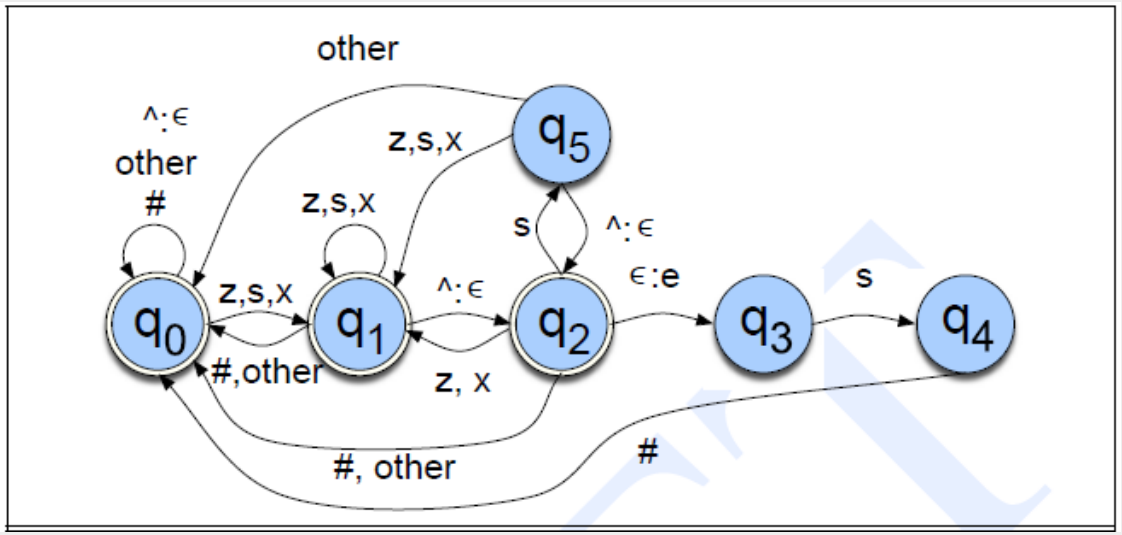
Пример



[SLP]

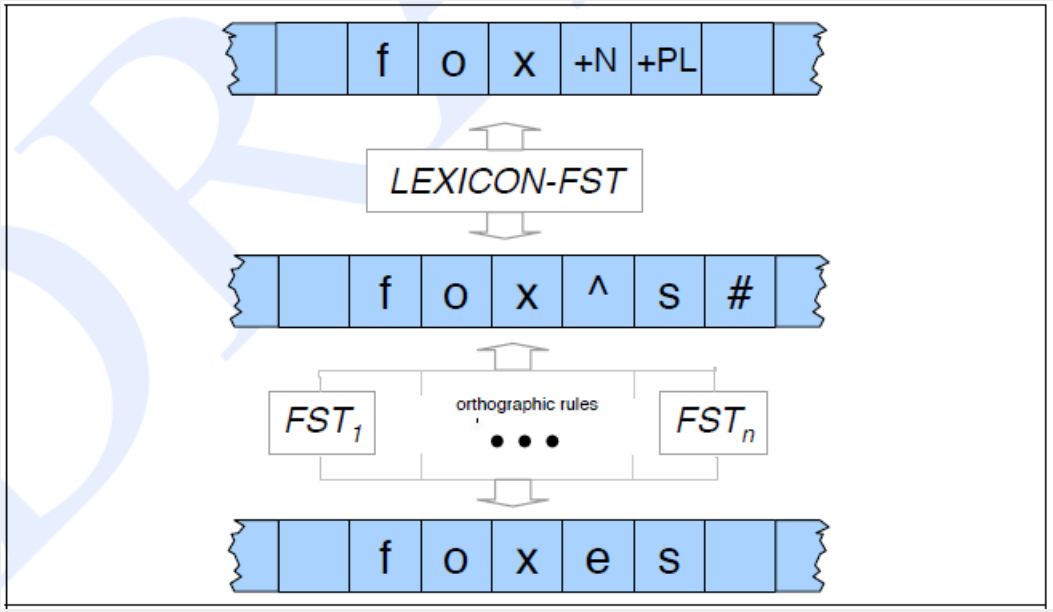
Преобразователи





[SLP]

Все вместе



[SLP]

НЕИЗВЕСТНЫЕ СЛОВА

Бармаглот

Варкалось. Хливкие шорьки
Пырялись по наве,
И хрюкотали зелюки,
Как мюмзики в мове.

(пер. Дины Орловской)

Подходы

- Сопоставление «с конца»
  - Отсечение префиксов (*квази-, псевдо-, …*)

mystem

варкалось{варкаться?=V,несов,нп=прош,ед,изъяв,сред}
хливкие{хливкий?=A=вин,мн,полн,неод|?=A=им,мн,полн}
шорьки{шорька?=S,жен,неод=вин,мн|?=S,жен,неод=род,ед|?=S,жен,неод=им,мн}
пырялись{пыряться=V,несов,нп=прош,мн,изъяв}
по{по=PR=}
наве{нав=S,имя,муж,од=пр,ед}
и{и=CONJ=} хрюкотали{хрюкотать?=V,несов,нп=прош,мн,изъяв}
зелюки{зелюк?=S,муж,од=им,мн}
как{как=CONJ=}
мюмзики{мюмзик?=S,имя,муж,од=им,мн}
в{в=PR=}
мове{мов?=S,гео,муж,неод=пр,ед}.

СТЕММИНГ

Стемминг – пример

В больничном дворе стоит небольшой флигель, окруженный целым лесом репейника, крапивы и дикой конопли.

в больничн двор сто небольш флигел окружен цел лес репейник крапив и дик конопл

Алгоритм Портера

- Самый распространенный стеммер для английского языка
- 5 циклов усечения
- Каждый цикл – набор команд
- В первую очередь выполняется операция над самым длинным суффиксом
- Широко используется (использовался) в информационном поиске

http://tartarus.org/~martin/PorterStemmer/

Алгоритм Портера – фрагмент

- *sses* $\rightarrow$ *ss*
- *ies* $\rightarrow$ *i*
- *ational* $\rightarrow$ *ate*
- *tional* $\rightarrow$ *tion*

- Правила с учетом длины
- *(m>1) EMENT* →
  - *replacement* → *replac*
  - *cement* → *cement*

[Manning&Raghavan,2005]

Стеммер Портера – пример

```
It is a truth universally acknowledged that a single man in possession of a
good fortune must be in want of a wife However little known the feelings or
views of such a man may be on his first entering a neighbourhood this truth
is so well fixed in the minds of the surrounding families that he is consid
ered the rightful property of some one or other of their daughters
```

Стемминг для русского

Step 1: Search for a PERFECTIVE GERUND ending. If one is found remove it, and that is then the end of step 1. Otherwise try and remove a REFLEXIVE ending, and then search in turn for (1) an ADJECTIVAL, (2) a VERB or (3) a NOUN ending. As soon as one of the endings (1) to (3) is found remove it, and terminate step 1.

Step 2: If the word ends with и (***i***), remove it.

Step 3: Search for a DERIVATIONAL ending in *R*2 (i.e. the entire ending must lie in *R*2), and if one is found, remove it.

Step 4: (1) Undouble н (***n***), or, (2) if the word ends with a SUPERLATIVE ending, remove it and undouble н (***n***), or (3) if the word ends ь (′) (soft sign) remove it.

http://snowball.tartarus.org/algorithms/russian/stemmer.html

Стемминг

кровать – крова

кроватью – кроват

лью – лью

лить – лит

людей – люд

человека – человек

лучше – лучш

хорошеет – хорошеет

бежит – беж

бегу – бег

Статистический стеммер

словарями → словарь → словар-ями → ар-ями

топорами → топор → топор-ами → ор-ами

летящего → лететь → лет-ящего → ет-ящего

летящего → летящий → летящ-его → ящ-его

+ правило: одна гласная в основе

имя	ра	546
има	ро	154
огещя	те	12
оге	щя	12

http://www.keva.ru/stemka/stemka.html

НЕОДНОЗНАЧНОСТЬ РАЗБОРА

Грамматическая омонимия

стекла{стекло=S,сред,неод=(вин,мн|род,ед|им,мн)|стекать=V,нп=прош,ед,изъяв,жен,сов}

падали{падать=V,несов,нп=прош,мн,изъяв|падаль=S,жен,неод=(пр,ед|вин,мн|дат,ед|род,ед|им,мн)}

печь{печь=S,жен,неод=(вин,ед|им,ед)|печь=V,несов,пе=инф}

черепах{череп=S,муж,неод=пр,мн|черепаха=S,жен,од=(вин,мн|род,мн)|черепаха=S,жен,неод=род,мн}

ученый{ученый=S,муж,од=им,ед|ученый=A=(вин,ед,полн,муж,неод|им,ед,полн,муж)}

работника{работник=S,муж,од=(вин,ед|род,ед)}

красиво{красиво=ADV=|красивый=A=ед,кр,сред}

пора{пора=ADV,прдк=|пора=S,жен,неод=им,ед}

бегу{бежать=V,нп=(непрош,ед,изъяв,1-л,несов|непрош,ед,изъяв,1-л,сов)|бег=S,ед,муж,неод=(дат|местн)}

гладь{гладь=S,жен,неод=(вин,ед|им,ед)|гладить=V,несов,пе=ед,пов,2-л}

Корпус

- Коллекция текстов с аннотациями
  - Морфологическая/синтаксическая/семантическая разметка
- Корпусная лингвистика (corpus linguistics)
- Русские корпусы:
  - НКРЯ http://ruscorpora.ru/
  - OpenCorpora http://opencorpora.org/
  - ГКРЯ http://www.webcorpora.ru/

НКРЯ

Подкорпус	Число текстов	Число предложений	Число словоупотреблений	% словоупотреблений
Основной корпус	76 882	17 574 752	209 198 275	57.3%
- в том числе со снятой омонимией	2 147	516 852	5 944 188	1.6%
Газетный корпус	181 175	8 553 495	113 292 003	31.0%
Диалектный корпус	197	20 273	194 283	0.1%
Обучающий корпус	229	65 666	664 751	0.2%
Параллельный корпус	370	1 609 609	24 022 437	6.6%
Поэтический корпус	41 448	638 861	6 738 474	1.8%
Устный корпус	3 034	1 604 626	10 122 579	2.8%
Мультимедийный корпус	31 741	148 619	648 576	0.2%
Всего:	**335 076**	**30 215 901**	**364 881 378**	**100%**

http://ruscorpora.ru/

Снятие на основе частоты

стекла{стекло:101.30|стекать:7.40}

падали{падать:83.20|падаль:1.90}

печь{печь:34.50=S|печь:7.40=V}

черепах{череп:26.10|черепаха:5.50}

ученый{ученый:126.00=S|ученый:19.39=A}

красиво{красиво:36.20|красивый:163.69}

пора{пора:72.30=ADV|пора:356.60=S}

бегу{бежать:142.30|бег:20.20}

гладь{гладь:6.10|гладить:18.39}

Снятие неоднозначности

в{в:30948.70}
больничном{больничный:15.10}
дворе{двор:165.50}
стоит{стоять:827.90|стоить:74.50}
небольшой{небольшой:194.30}
флигель{флигель:4.40}
окруженный{окружать:30.00}
целым{целый:224.20|целое:101.80}
лесом{лес:209.10|лесом:0.00}
репейника{репейник:1.20}
крапивы{крапива:5.90}
и{и:35302.69}
дикой{дикий:59.90}
конопли{конопля:1.80}

POS TAGGING

Идея

- Не отдельное слово, а последовательность
  - снятие неоднозначности лемматизатора
  - самостоятельный подход

ученый кот (A N) (N N?)
великий ученый (A N) (A A?)
прозрачного стекла (A N) (A V?)
вода стекла (N V) (N N?)

Теги Penn Treebank

DT	Determiner
IN	Preposition/subord conj
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present

(всего 36)

https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Теги НКРЯ / mystem

S — существительное (*яблоня, лошадь, корпус, вечн*
A — прилагательное (*коричневый, таинственный, м*
NUM — числительное (*четыре, десять, много*)
A-NUM — числительное-прилагательное (*один, седы*
V — глагол (*пользоваться, обрабатывать*)
ADV — наречие (*сгоряча, очень*)
PRAEDIC — предикатив (*жаль, хорошо, пора*)
PARENTH — вводное слово (*кстати, по-моему*)
S-PRO — местоимение-существительное (*она, что*)
A-PRO — местоимение-прилагательное (*который, т*
ADV-PRO — местоименное наречие (*где, вот*)
PRAEDIC-PRO — местоимение-предикатив (*некого,*
PR — предлог (*под, напротив*)
CONJ — союз (*и, чтобы*)
PART — частица (*бы, же, пусть*)
INTJ — междометие (*увы, батюшки*)

Род:

m — мужской род (*работник, стол*)
f — женский род (*работница, табуретка*)
m-f — «общий род» (*задира, пьяница*)
n — средний род (*животное, озеро*)

Одушевленность:

anim — одушевленность (*человек, ангел, утопленник*)
inan — неодушевленность (*рука, облако, культура*)

Число:

sg — единственное число (*яблоко, гордость*)
pl — множественное число (*яблоки, ножницы, детишки*)

Падеж:

nom — именительный падеж (*голова, сын, степь, сани, ко*
gen — родительный падеж (*головы, сына, степи, саней, ко*
dat — дательный падеж (*голове, сыну, степи, саням, котор*
dat2 — дистрибутивный дательный (*[по] многу, нескольку, с*
acc — винительный падеж (*голову, сына, степь, сани, кото*
ins — творительный падеж (*головой, сыном, степью, саням*
loc — предложный падеж (*[о] голове, сыне, степи, санях, к*
gen2 — второй родительный падеж (*чашка **чаю***)

http://ruscorpora.ru/corpora-morph.html

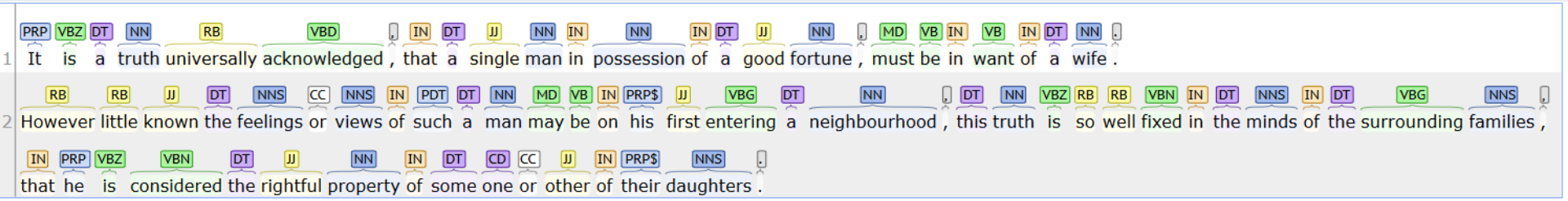
Набор тегов OpenCorpora

1	POST	ЧР	часть речи	—
2	NOUN	СУЩ	имя существительное	POST
3	ADJF	ПРИЛ	имя прилагательное (полное)	POST
4	ADJS	КР_ПРИЛ	имя прилагательное (краткое)	POST
5	COMP	КОМП	компаратив	POST
6	VERB	ГЛ	глагол (личная форма)	POST
7	INFN	ИНФ	глагол (инфинитив)	POST
8	PRTF	ПРИЧ	причастие (полное)	POST
10	PRTS	КР_ПРИЧ	причастие (краткое)	POST
11	GRND	ДЕЕПР	деепричастие	POST
12	NUMR	ЧИСЛ	числительное	POST
13	ADVB	Н	наречие	POST
14	NPRO	МС	местоимение-существительное	POST
15	PRED	ПРЕДК	предикатив	POST
16	PREP	ПР	предлог	POST
17	CONJ	СОЮЗ	союз	POST
18	PRCL	ЧАСТ	частица	POST
19	INTJ	МЕЖД	междометие	POST

http://opencorpora.org/dict.php?act=gram

Пример

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.
However little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth is so well fixed in the minds of the surrounding families, that he is considered the rightful property of some one or other of their daughters.



http://corenlp.run/

Пример

В больничном дворе стоит небольшой флигель, окруженный целым лесом репейника, крапивы и дикой конопли.

{в=PR=}
{больничный=A=пр,ед,полн,муж}
{двор=S,муж,неод=пр,ед}
{стоять=V,несов,нп=непрош,ед,изъяв,3-л}
{небольшой=A=им,ед,полн,муж}
{флигель=S,муж,неод=им,ед}
{окружать=V,пе=прош,им,ед,прич,полн,муж,сов,страд}
{целый=A=твор,ед,полн,муж}
{лес=S,муж,неод=твор,ед}
{репейник=S,муж,неод=род,ед}
{крапива=S,жен,неод=род,ед}
{и=CONJ=}
{дикий=A=род,ед,полн,жен}
{конопля=S,жен,неод=род,ед}

ДАННЫЕ

НКРЯ

Подкорпус	Число текстов	Число предложений	Число словоупотреблений	% словоупотреблений
Основной корпус	76 882	17 574 752	209 198 275	57.3%
- в том числе со снятой омонимией	2 147	516 852	5 944 188	1.6%
Газетный корпус	181 175	8 553 495	113 292 003	31.0%
Диалектный корпус	197	20 273	194 283	0.1%
Обучающий корпус	229	65 666	664 751	0.2%
Параллельный корпус	370	1 609 609	24 022 437	6.6%
Поэтический корпус	41 448	638 861	6 738 474	1.8%
Устный корпус	3 034	1 604 626	10 122 579	2.8%
Мультимедийный корпус	31 741	148 619	648 576	0.2%
Всего:	**335 076**	**30 215 901**	**364 881 378**	**100%**

http://ruscorpora.ru/

OpenCorpora

Размеченные тексты

Весь корпус, XML (XML Schema) обновлён 27.09.2016 05:08 MSK
предложений: 108958, токенов: 1966924, слов: 1522291

- целиком: архив .bz2 (31.91 Мб), архив .zip (54.79 Мб)
- один текст на файл: архив .bz2, архив .zip

Со снятой омонимией

Подкорпус со снятой омонимией*, XML | .bz2 (1.13 Мб) | .zip (1.78 Мб) обновлён 27.09.2016 05:08 MSK
предложений: 9798, токенов: 59307, слов: 36932

* В подкорпус включены целые предложения, не имеющие в своём составе ни одного неоднозначно разобранного слова — как изначально однозначные предложения, так и те, в которых неоднозначность была снята вручную.

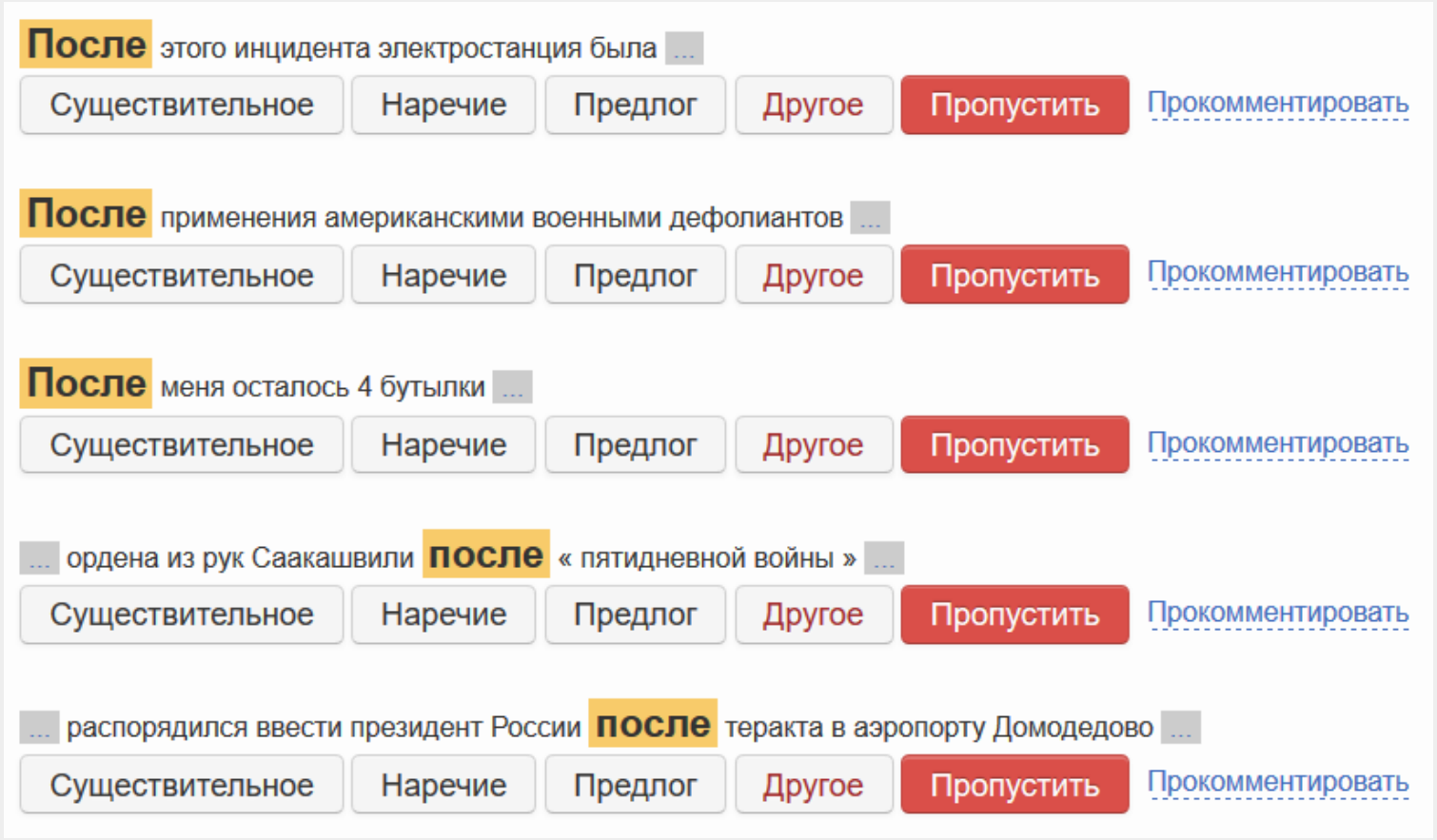
Подкорпус со снятой омонимией без UNKN, XML | .bz2 (0.86 Мб) | .zip (1.36 Мб) обновлён 27.09.2016 05:08 MSK
предложений: 7577, токенов: 41652, слов: 27276

Подкорпус со снятой омонимией (без модерации)*, XML | .bz2 (1.98 Мб) | .zip (3.06 Мб) обновлён 27.09.2016 05:08 MSK
предложений: 14781, токенов: 109010, слов: 74510

* В подкорпус включены предложения, где неоднозначность снята по ответам пользователей, в том числе не проверенным модераторами.

http://opencorpora.org/?page=downloads

Можно поучаствовать



Скрытые марковские модели

- Hidden Markov Model (HMM)
- Применяются в разных задачах ОЕЯ (языковые модели, распознавание речи, генерация текста, …)
- Концептуально простая модель, допускающая усложнение

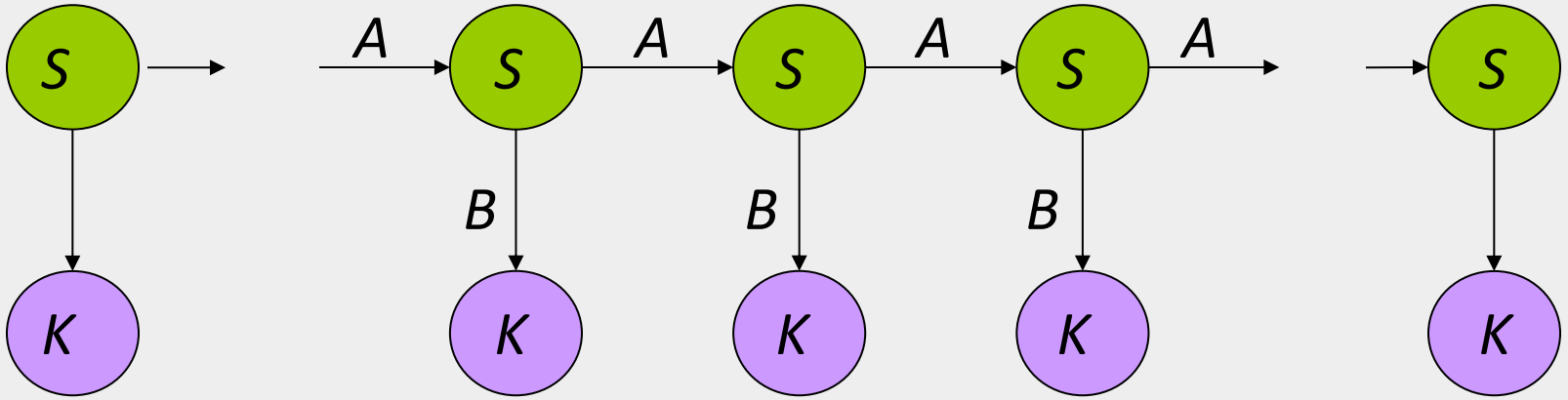
Задача разметки

- Скрытые состояния – граммемы (POS tags)
- Наблюдаемые состояния – слова
- Обучение вероятностей переходов и эмиссии – на корпусе со снятой омонимией
- Нахождение наиболее вероятной последовательности тегов – алгоритм Витерби (динамическое программирование)

Другими словами

- Раньше мы видели, что с вероятностью 0.4 за прилагательным следует существительное
- Еще мы знаем, что с вероятностью 0.7 [печь] – это существительное, а 0.3 – глагол (на самом деле вероятности «в другую сторону» – от тегов к словам)
- Мы видим слова, хотим восстановить последовательность тегов

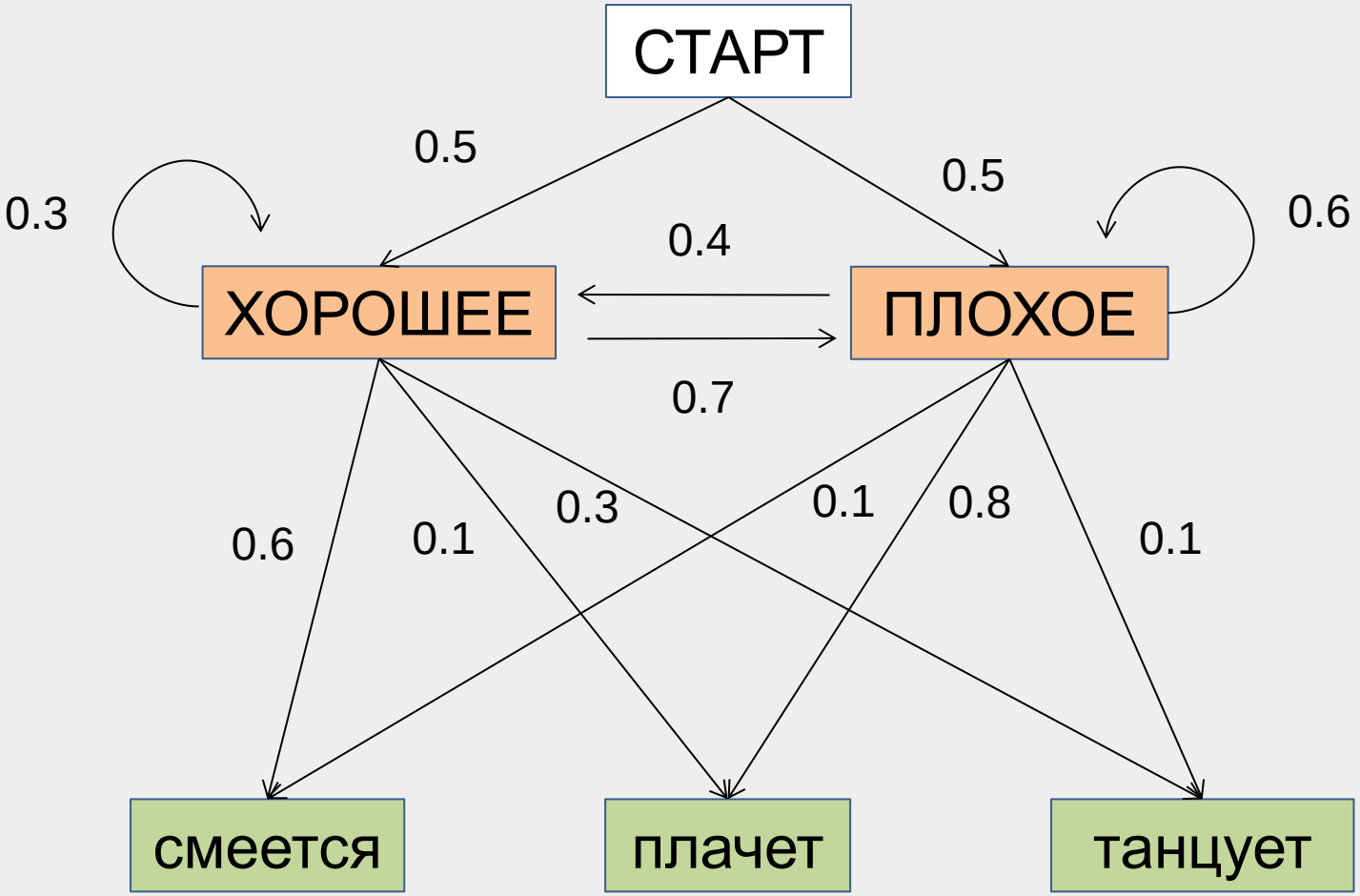
HMM

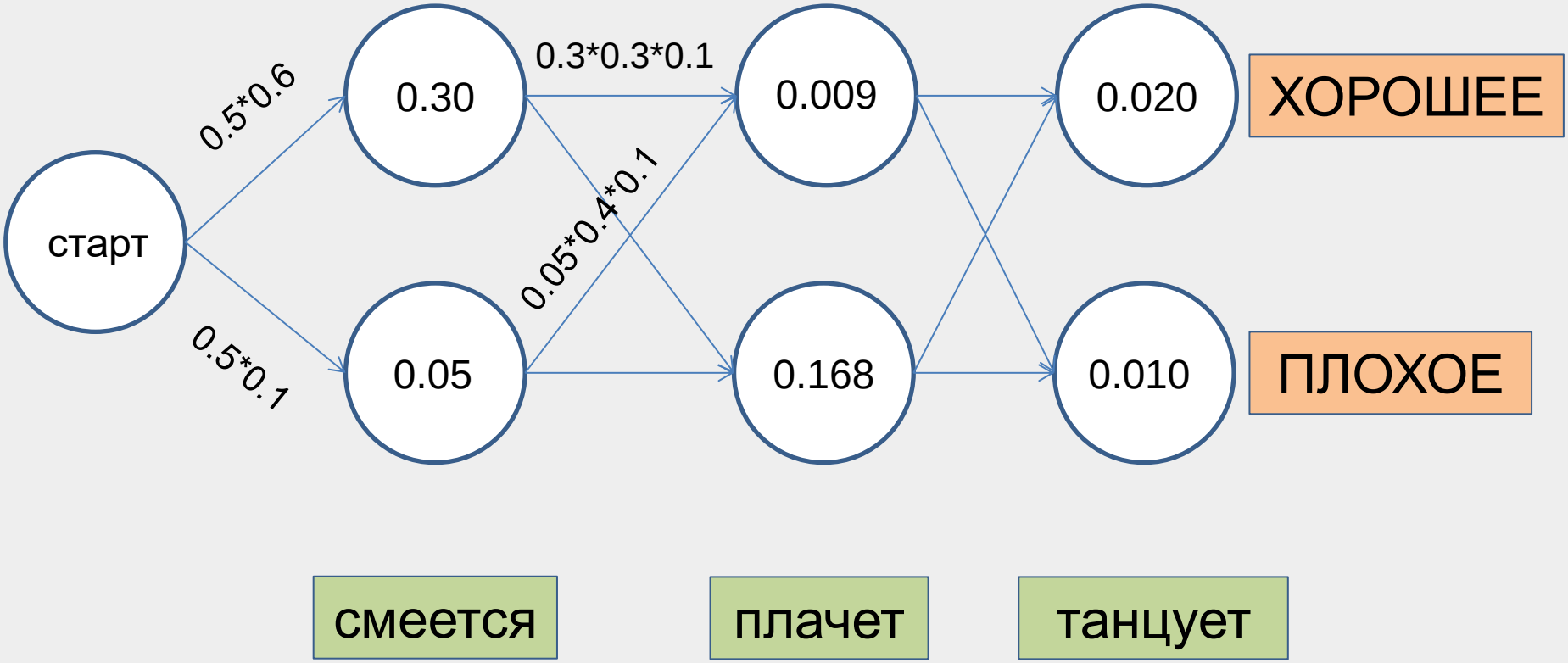


- $\{S, K, \Pi, A, B\}$
- $\Pi = \{\pi_i\}$ вероятности исходных состояний
- $A = \{a_{ij}\}$ вероятности перехода
- $B = \{b_{ik}\}$ вероятности состояний

[David Blei, 1999]

АЛГОРИТМ ВИТЕРБИ


СТАРТ
0.5
0.5
0.3
0.6
0.4
ХОРОШЕЕ
ПЛОХОЕ
0.7
0.6
0.1
0.3
0.1
0.8
0.1
смеется
плачет
танцует


старт
0.5*0.6
0.5*0.1
0.30
0.05
0.3*0.3*0.1
0.05*0.4*0.1
0.009
0.168
0.020
0.010
ХОРОШЕЕ
ПЛОХОЕ
смеется
плачет
танцует

Развитие

- Незнакомые слова – можно делать (вероятностные) предположения
- Можно учитывать более длинную «предысторию» : $t_{i-2}t_{i-1} \rightarrow t_i$
  - борьба с разреженностью данных: интерполяция:

$$P(t_i|t_{1,i-1}) = \lambda_1 P_1(t_i) + \lambda_2 P_2(t_i|t_{i-1}) + \lambda_3 P_3(t_i|t_{i-1,i-2})$$

  - сглаживание (smoothing) – подробнее в лекции про языковые модели (language models)

HMM [Сокирко, Толдова, 2005]

- Триграммная модель
- Сглаживание (разные наборы λ для разных групп по частоте триграмм)
- Данные – НКРЯ со снятой омонимией (5M словоформ)
- Полный набор тегов – 829

1) Точность по леммам – 99.71%

2) Точность по грамм. тегам – 98.34%

[Зеленков, Сегалович, Титов, 2005]

мал(1ть)	думал{думать=V,несов=прош,ед,изъяв,муж}
ере(1ь)	звере{зверь=S,муж,од=пр,ед}
огу(1а)	дорогу{дорога=S,жен,неод=вин,ед}
ной(4я)	мной{я=S,ед,од=(твор,жен|твор,муж)}
ись(9орачиваться)	повернувшись{поворачиваться=V=прош,деепр,сов}
щем(1)	туловищем{туловище=S,сред,неод=твор,ед}
ерх(0)	вверх{вверх=ADV=}
нно(1ый)	неуклонно{неуклонный=A=ед,кр,сред}

чал(0о|2инать)	начал{начинать=V=прош,ед,изъяв,муж,сов|начало=S,сред,неод=род,мн}
ерь(1ить|1ять)	поверь{поверить=V,сов=ед,пов,2-л|поверять=V=ед,пов,2-л,сов}
рка(0|1|1ий)	марка{марк=S,муж,од=(род,ед|вин,ед)|марка=S,жен,неод=им,ед|маркий=A=ед,кр,жен}
рую(1ить|1я)	струю{струить=V,несов=непрош,ед,изъяв,1-л|струя=S,жен,неод=вин,ед}
гко(0|1ий)	легко{легкий=A=ед,кр,сред|легко=ADV=}
мой(0|2ыть)	мой{мой=A=(им,ед,муж|вин,ед,муж,неод)|мыть=V,несов=ед,пов,2-л}
дит(0ь|4аживать)	выходит{выхаживать=V=непрош,ед,изъяв,3-л,сов|выходить=V=непрош,ед,изъяв,3-л,несов}

Словарь контекстов

ала(1о|Зинать)	[р]+2	1о	0.67
ала(1о|Зинать)	[р]+2	Зинать	0.33
ала(2новиться|2ть)	и(0)-2	2новиться	0.14
ала(2новиться|2ть)	и(0)-2	2ть	0.86
его(Зон|Зоно)	с(0)-1	Зон	0.98
его(Зон|Зоно)	с(0)-1	Зоно	0.02
его(Зон|Зоно)	с(0)-2	Зон	1
его(Зон|Зоно)	с(0)-3	Зон	0.94
его(Зон|Зоно)	с(0)-3	Зоно	0.06
его(Зон|Зоно)	с(0)+1	Зон	1
его(Зон|Зоно)	с(0)+2	Зон	0.95
его(Зон|Зоно)	с(0)+2	Зоно	0.05

Влияние контекстов

−1	1.00
+1	0.97
−2	0.93
−3	0.89
+2	0.88

ИНСТРУМЕНТЫ, ДАННЫЕ, ОЦЕНКА

Инструменты

- AOT http://aot.ru
- mystem https://tech.yandex.ru/mystem/
- pymorphy2 https://pymorphy2.readthedocs.io/
- TreeTagger http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Словари

- Словарь Зализняка
  см. ссылки на http://www.speakrus.ru/dict/
- oDICT http://odict.ru/
- AOT http://aot.ru/

Корпусы со снятой омонимией

- НКРЯ http://ruscorpora.ru/
- OpenCorpora http://opencorpora.org

RuMorphEval’2010

Задачи:

- Лемматизация (13/7) *полнота/точность*
- Часть речи (13/7) *полнота/точность*
- Морфология (12) *полнота*
- Редкие слова (8) : лемма/POS *полнота*

Данные:

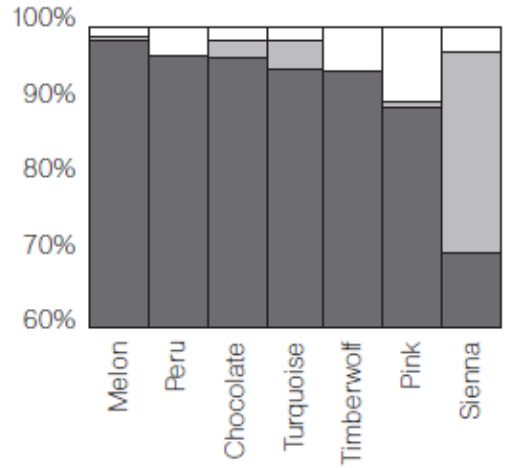
- Смешанная коллекция (1M / 2K)
- 75 редких слов с контекстами

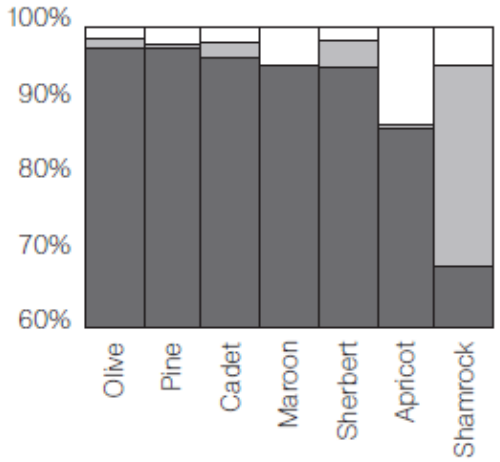
Оценка по подмножеству POS/граммем

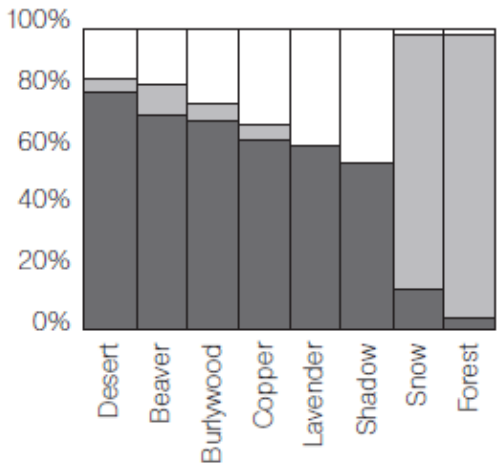
http://ru-eval.ru

Результаты

«Дизамбигуация: Леммы»					«Дизамбигуация: POS»					«Редкие слова»				
Участник	t	нет	f	Accur	Участник	t	нет	f	Accur	Участник	t	нет	f	Accur
Melon	2008	14	24	98,1 %	Olive	1991	22	33	97,3 %	Desert	59	3	13	78,7 %
Peru	1970	1	75	96,3 %	Pine	1991	5	50	97,3 %	Beaver	53	8	14	70,7 %
Chocolate	1964	43	39	96,0 %	Cadet	1958	43	45	95,7 %	Burlywood	52	4	19	69,3 %
Turquoise	1934	75	37	94,5 %	Maroon	1943	0	103	95,0 %	Copper	47	4	24	62,7 %
Timberwolf	1925	0	121	94,1 %	Sherbert	1934	75	37	94,5 %	Lavender	46	0	29	61,3 %
Pink	1831	11	204	89,5 %	Apricot	1769	11	266	86,5 %	Shadow	42	0	33	56,0 %
Sienna	1430	547	69	69,9 %	Shamrock	1394	547	105	68,1 %	Snow	10	63	2	13,3 %
										Forest	3	70	2	4,0 %
Всего ответов				2046					2046	Всего ответов				75
Медиана				94,5 %					95,0 %	Медиана				62,0 %







RuMorphEval'2017

Team name	team ID	Track	Number of the best try	Accuracy by tags	Accuracyby sentences	Lemmatization, accuracy by wordforms	Lemmatization, accuracy by sentences
MSU-1	C	Closed	2	**93.39**	**65.29**		
IQMEN	O	Closed	1	93.08	62.71	**92.22**	**58.21**
Sagteam	H	Closed	2	92.64	58.40	80.73	25.01
Aspect	A	Closed	2	92.57	61.01	91.81	56.49
Morphobabushka	M	Closed	2	90.07	48.10		
Pullenti Pos Tagger	G	Closed	4	89.96	47.23	89.32	45.18
	B	Closed	6	89.91	48.2		
	N	Closed	4	89.86	47.13	85.10	29.04
	K	Closed	4	89.46	48.54	88.47	44.78
	F	Closed	2	88.14	39.63	87.27	36.90
	I	Closed	2	86.05	34.62		
	L	Closed	2	71.48	6.48		
ABBYY	E	Open	3	**97.11**	**83.68**	**96.91**	**82.13**
Aspect	A	Open	4	92.38	60.90	87.66	41.12
	N	Open	5	90.88	51.77	85.91	32.57
	J	Open	1	83.51	29.69		
	D	Open	5	77.13	17.19		