



Этапы обработки текста. Часть 1

Ефремова Наталья Эрнестовна
Грацианова Татьяна Юрьевна

Содержание



- Этапы обработки текста
- Графематический анализ и сегментация текста
- Морфологический анализ и синтез текста
- Синтаксический анализ текста
- Частичный синтаксический анализ: особенности
- Лингвистические процессоры для РЯ

Этапы обработки текста



Структурная модель – набор правил, учитывает лингвистические особенности уровней текста

1. *Графематический анализ и сегментация* – предобработка текста
2. *Морфологический анализ*:
 - ✓ нормализация словоформ или их стемминг
 - ✓ определение морфологических характеристик
3. *Синтаксический анализ* – построение синтаксической структуры предложения
4. *Семантический и прагматический анализ*

Графематический анализ



- Цель – выделение и классификация основных единиц текста: слов, предложений, абзацев
- Анализ уровня графем
- *Графема* – минимальная единица письменного текста а, А, а, А, а
- Близкое понятие – *сегментация (segmentation)*, т.е. разбиение текста на части (от символов до разделов/глав)
- Сегментация – более широкое понятие, но часто используется как синоним
- ❖ В случае обработки речи части идентифицируются аудиально

Виды сегментации



Основное разделение:

- ❑ Сегментация нижнего уровня (*low-level*)
- ❑ Сегментация высокого уровня (*high-level*)

■ Сегментация нижнего уровня:

- ❖ *Токенизация (tokenization)* – выделение минимальных лингвистически значимых элементов текста (токенов); могут состоять из нескольких графем
- ❖ Разбиение текста на предложения

■ Сегментация высокого уровня:

- ❖ Синтаксическая сегментация (*inter-sentence*)
- ❖ Макросинтаксический анализ (*intra-sentence*)

Сегментация высокого уровня



- **Синтаксическая сегментация** (по сути – *частичный синтаксический анализ*):
 - ❖ выделение синтаксических групп – именных, глагольных и др.
 - ❖ выделение простых предложений в составе сложных
- **Макросинтаксический анализ:**
 - ❖ выделение композиционных элементов: абзацев, заголовков, сносок, примечаний и пр.
 - ❖ установление иерархии абзацев, заголовков, предложений
 - ✓ учитываются стандарты оформления текстов (*распоряжение, инструкция, договор* и т.д.)

Токенизация



- При выделении токенов опираемся на:
 - ❖ разделители: пробел, знак препинания, ...
 - ❖ вид знака: буква латиницы/кириллицы, цифра, ...
 - ❖ регистр букв
- **НО!** разделители неоднозначны → требуется анализ контекста знаков
- Виды токенов (примеры):
 - ✧ Слова ЕЯ *я иду гулять*
 - ✧ Знаки препинания *! – ;*
 - ✧ Числа *10 десять*
 - ✧ Буквенно-цифровые комплексы *Boeing-747*



Обработка сложных токенов (1)



- Различение тире, знака переноса и истинного дефиса, разбиение сложносоставных слов:
он – гений, сбор-ка, девица-красавица
- Сборка слов, написанных в разрядку
Заявление → заявление
- Нормализация (приведение к одному формату) дат и других особых токенов: *3 PM → 15.00, сто двадцать три → 123*
- Определение регистра букв и восстановление правильного: *МОСКВА, москва – Москва*

Обработка сложных токенов (2)



- Свертка устойчивых неизменяемых оборотов (не имеющих словоизменительных вариантов): *в связи с, так сказать, таким образом*
- Обработка сокращений слов и словосочетаний (неразрывных и разрывных): *they're, г. Москва, д/з*
- Выделение полного имени (фамилия, имя, отчество): *К.А. Петров, Анастасия Дорохова*

Для такой обработки требуется лингвистическая и лексическая информация: правила преобразования, словари (предлогов, словосочетаний, имен собственных) и пр.

Разделение на предложения



- ❑ Маркеры конца предложения – определенные знаки препинания (. ? ! ...)
- ❑ Маркер начала предложения – заглавная буква

НО! все эти критерии неоднозначны:

- ✓ встречаются сокращения, имена, числа, цитаты
- ✓ разное оформления, например, прямой речи
- ✓ возможно неверное использование знаков
- ❑ Требуется анализ контекста маркеров
- ❑ Токенизация и разделение на предложения могут проводиться одновременно

Разделение на предложения: примеры



15 марта в 9.00 около ДК им. Горбунова состоится митинг в поддержку И.И. Иванова.

Я смотрел на Кота Б. Б. Кот смотрел на меня.

Я живу на ул. Ленина и меня зарубает время от времени.

В данном случае использовать Yahoo! достаточно бессмысленно.

Я сказала: «Это Анин кот!». – I said: "It's Ann's cat!"

Парсер, Разбей эти... буквы, знаки и т.п. на предложения. И покажи пож. как со словом S.T.A.L.K.E.R. получится.

Подходы к реализации графематического анализа



- ❑ Инженерный подход. Опора на:
 - ✓ формального аппарата
 - ✓ эвристические правила определения токенов и предложений
 - ✓ словари сокращений, имен, знаков и пр.
- ❑ Машинное обучение по размеченным текстам
- ❑ Гибридный подход

Нередко:

- ❑ Более одного просмотра текста
- ❑ Для ЕЯ без разделения слов (*арабский*) и сложными словами (*немецкий*): объединяются графематический и морфологический анализы

Использование формального аппарата



- На базе теории формальных языков и грамматик
- Простейший графематический анализ – синтаксический анализ *регулярных языков*
- Более сложный графематический анализ – учет локального контекста, словари
- Средства описания регулярных языков:
 - ✓ регулярные (автоматные) грамматики
 - ✓ конечные автоматы
 - ✓ *регулярные выражения*: встроены в современные языки программирования (упрощает построение анализаторов)

Проект aot.ru



- Лингвистический анализ текстов на русском языке
- В том числе выполняет
 - ❖ графематический анализ
 - ❖ морфологический анализ
 - ❖ синтаксический анализ
 - ❖ семантический анализ
- Инженерный подход, структурная модель
- Программные модули с открытым кодом, C++
- На сайте есть тестовый он-лайн интерфейс



Графематика aot.ru

- Модуль выполняет
 - ❖ токенизацию и сегментацию на предложения
 - ❖ свертку словосочетаний
 - ❖ анализ макроструктуры текста
- При этом распознаются:
 - ❖ слова, разделители и предложения
 - ❖ аббревиатуры, ФИО, даты и числа, электронные адреса, имена файлов
 - ❖ тире и дефис
 - ❖ устойчивые словосочетания
 - ❖ абзацы, заголовки, перечисления (рубрики)
- **Вход** модуля – текстовый файл (плейн-текст)
Выход – таблица токенов с дескрипторами

Графематика aot.ru.

Дескрипторы



Последовательно вычисляются 3 вида дескрипторов

- Основные графематические дескрипторы (19):
 - RLE – русская лексема (последовательность букв кириллицы)
 - PUN – знак препинания (. , : ? ; -)
 - OPN и CLS – открывающая и закрывающая скобки
- Контекстные дескрипторы (19):
 - SENT_END – конец предложения
 - FAM1 и FAM2 – начало и конец ФИО
- Дескрипторы макросинтаксического анализа (5):
 - HDNG – конец заголовка

Графематика aot.ru: пример



Наконец я ей сказал: «Хочешь, пойдем прогуляться?»

Токен	Позиция начала	Длина	Дескрипторы
Наконец	777	7	RLE Aa NAM EXPR1 EXPR2 EXPR_NO 301
я	785	1	RLE aa
ей	787	2	RLE aa
сказал	790	6	RLE aa
:	796	1	PUN
«	798	1	PUN
Хочешь	799	6	RLE Aa NAM
,	805	1	PUN
пойдем	807	6	RLE aa
прогуляться	814	11	RLE aa
?	826	1	PUN SENT_END
»	827	1	PUN

Графематический анализ:

ВЫВОДЫ



- ❑ Нужно выделить последовательности графем, которые потом будут анализироваться как единое целое
- ❑ Важно ничего не потерять для дальнейшего анализа
- ❑ Задача кажется простой, на самом деле – сложный и громоздкий механизм
- ❑ Универсальных решений нет. Необходимо учитывать:
 - ❖ особенности ЕЯ
 - ❖ специфику текстов (тематику, жанр, количества сокращений, имен собственных, обозначений)
- ❑ Достигаемая R – 60-80 %, P – до 99 %

Основные понятия морфологии



- Слова ЕЯ состоят из *морфов* – минимальных значащих единиц ЕЯ
- Разбиение на морфы – морфемный разбор
- *Корень* (корневой морф) – носитель основного смысла слова
- *Аффикс* – служебный морф, придает дополнительный смысл:
 - ✧ *префикс (приставка), суффикс: сход, буровой,*
 - ✧ *флексия (окончание): начальн*ый*, shoes*
 - ✧ *постфикс: частица с*я*, с*ь**
- *Основа (stem)* – часть слова без окончания и постфикса, может изменяться: *бегать* – *бежал*
- *Псевдооснова* – неизменяемая начальная часть слова: *бегать* – *бежал*

Слово в тексте



- Слово меняется в зависимости от его синтаксической функции в предложении
- *Словоформа* – конкретная грамматическая форма слова: *носа, прибежал, синим*
- *Лексема* – множество всех словоформ слова, обозначают одно понятие
{*нос, носа, носу, носом, носе, носы, носов, носам, носами*}
- *Лемма, имя лексемы* – нормальная (базовая, каноническая, словарная) форма лексемы: *нос*
- *Словоизменительная/морфологическая парадигма* – система форм слова

Словоизменительная парадигма



- *Парадигма* (греч.) – образец, пример
- Характеризуется:
 - ✓ изменением *морфологических параметров* (МП)
 - ✓ инвариантной частью (корень, основа)
 - ✓ перечнем изменяемых компонентов (флексий)
- Бывают неизменяемые слова: *и, всего, пальто*
- В парадигме могут отсутствовать некоторые формы: *ножницы, победить*
- Много исключений: *идти – шел*

Словоизменительная парадигма: пример



Слов *нос*, *стол*, *завод* и пр.:

- изменяется *число* и *падеж*,
- перечень флексий (окончаний):

нос, *носа*, *носу*, *нос*, *носом*, *носе*,
носы, *носов*, *носам*, *носы*, *носами*, *носах*

Единственное число						Множественное число					
им.	род.	дат.	вин.	тв.	пр.	им.	род.	дат.	вин.	тв.	пр.
⊙	а	у	⊙	ом	е	ы	ов	ам	ы	ами	ах

Словоизменительный (флексивный) класс –
подкласс слов с одной морфопарадигмой

Морфологический анализ



Вход: словоформа текста ЕЯ

- ❑ *Теггинг* – определение части речи словоформы

Выход: словоформа и часть речи *красивее* → *ПРИЛ.*

- ❑ *Лемматизация (нормализация)* – преобразование окончания слова по правилам

Выход: лемма и часть речи *красивее* → *красивый*

- ❑ *Стемминг (stem – основа слова)* – отсечение окончания слова по правилам

Выход: основа/псевдооснова слова

водных → *вод/водн*, *бежит*, *бегут* → *бег/бе*

- ❑ *Полный морфологический анализ*

Выход: лемма, часть речи и значения МП

синих → *синий*, *ПРИЛ.*, *МНОЖ. ЧИСЛО*, *РОД. ПАДЕЖ*

Морфологический синтез



Цель – порождение нужной словоформы слова или его парадигмы исходя из леммы/основы

- Синтез всей парадигмы слова (лексемы):

Вход: лемма/основа слова

Выход: словоизменятельная парадигма

лук → лук, лука, луку, лук, луком, луке

- Синтез нужной словоформы

Вход: лемма/основа + набор конкретных значений морфологических параметров

Выход: словоформа

кипеть + ИЗЪЯВ. НАКЛ., НЕСОВ. ВИД, ПРОШ. ВР., ЕД. ЧИСЛО, МУЖ. РОД → кипел

- Возможен синтез разных вариантов словоформы:
(увидел) оператор/оператора

Модели морфологии и морфопроцессоры



- Структурные. Отличаются
 - ✓ способом представления лингвистической информации: *словарные* и *бессловарные*
 - ✓ учитываемыми флективными классами и значениями МП: *в РЯ 6 основных и от 3 до 7 дополнительных падежа*
- Статистические: на основе статистики слов и букв

Морфологические процессоры различаются:

- ✓ направлением обработки (анализ/синтез)
- ✓ используемой моделью морфологии

Грамматический словарь русского языка, Зализняк А.А., 1977



- Основа большинства моделей РЯ
- Системный подход к описанию морфологических парадигм РЯ, включает ударения
- 100 тыс. словарных входов (лексем)
- Много старых слов, нет новых

ТЕЧЬ

ж (жо): 8а, 8е, 8f'' — 47 | св (нсв): 8 — 118

утечь	св нп 8b/b (-к-), ё ● II	свóлочь	жо 8е	тòчь-в-тóчь	н
вытечь	св нп 8а (-к-) ● II	сволóчь	св 8b/b (-к-) [// простореч. сволочить] ● I (-а-)	запрячь	св 8b/b (-г-) ● II
дичь	ж 8а	отволóчь	св 8b/b (-к-) [// простореч. отволочить] ● I (-а-)	перезапрячь	св 8b/b (-г-) ● II
на́взничь	н	уволóчь	св 8b/b (-к-) [// простореч. уволочить] ● I (-а-)	напрячь	св 8b/b (-г-) ● II
опричь	предл.	вы́волочь	св 8а (-к-) [// простореч. вы́волочить] ● I (-а-)	поднапрячь	св 8b/b (-г-)
стричь	нсв 8b (-г-)	толóчь	нсв 8b/b (-к-) △ наст. тол- кú, толчёт, толкúт; прощ. толóк, толклá, толóкий;	перенапрячь	св 8b/b (-г-) ● II
застричь	св 8b (-г-) ● II			впрячь	св 8b/b (-г-) ● II
настричь	св 8b (-г-) ● II			подпрячь	св 8b/b (-г-) ● II
обстричь	св 8b (-г-) ● II			перепрячь	св 8b/b (-г-) ● II
подстричь	св 8b (-г-) ● II			припрячь	св 8b/b (-г-) ● II
перестричь	св 8b (-г-) ● II			сопрячь	св 8b/b (-г-) ● II
остричь	св 8b (-г-) ● II			спрячь	св 8b/b (-г-) ● II

Словарные морфологии



- Модели основаны на одном из видов словарей: **словаре словоформ** или **словаре основ**
- Обработка слова с конца путем отсечения возможных окончаний
- Используются вспомогательные словари флективных классов (ФК), исключений и т.д.
- Достоинства:
 - ❖ позволяют проводить синтез словоформ
 - ❖ реализуют полный морфоанализ словоформы, распознавание **морфоомонимии**: *после*, *косой*
- Недостатки:
 - ❖ проблема анализа новых слов (используются алгоритмы предсказания по аналогии)
 - ❖ не учитываются опечатки и другие ошибки

Словари словоформ и ОСНОВ



- Словарь словоформ:

- ✧ каждой возможной словоформе приписана лемма, часть речи, значения МП
- ✧ число словоформ зависит от степени флективности язык: для РЯ – большой объем

2609577	96056	одухотворяющие	прич	дст, но, од, нст, им, мн
2609578	96056	одухотворяющих	прич	дст, но, од, нст, рд, мн
2609579	96056	одухотворяющим	прич	дст, но, од, нст, дт, мн

- Словарь основ (или псевдооснов)

- ✧ перечень возможных в ЕЯ основ слов с морфологической информацией
- ✧ объем словаря ≈ числу лексем в ЕЯ
- ✧ связан со вспомогательными словарями (в т.ч. чередований, беглых гласных и пр.)

Пример: анализ слова *стола*



Основа	№ ФК
стол	001
столб	001

№ ФК	№ флексии	№ значений МП
001	22	40
001	66	06
002	67	26

Флексия	№	№ ФК
яя	64	
⊖	65	
а	66	001, 002
е	67	

№	Значения МП
04	им. ед.; вн. ед.; рд. мн.
05	им. ед.; рд. мн.; вн. мн.
06	рд. ед.
07	рд. ед.; дт. ед.; тв. ед.; пр. ед.

№ ФК	Флексии											
	Единственное число						Множественное число					
	им.	род.	дат.	вин.	тв.	пр.	им.	род.	дат.	вин.	тв.	пр.
001	⊖	а	у	⊖	ом	е	ы	ов	ам	ы	ами	ах
002	⊖	а	у	⊖	ом	е	ы	ов	ам	ы	ами	ах

Бессловарные морфологии



- Отсутствуют большие словари, **но** используются:
 - ❖ словарь псевдоокончаний/аффиксов
 - ❖ словарь неизменяемых слов и исключений
 - ❖ правила преобразования словоформ
*-ies → -y, -es → -ie, -s → ∅ – преобразование
мн. числа в един. в английском языке*
- Определение значений МП по последним буквам слова: -онок СУЩ, НЕОД., МР, ЕД.Ч., ИМ.П.
- Достоинства:
 - ❖ Не зависят от объемных словарей
 - ❖ Дают возможность обработки любых слов
- Недостатки:
 - ❖ Чувствительны к исключениям *шов, швы – ш?*
 - ❖ Применимы в ограниченном круге приложений, т.к. все слова – новые

Морфология aot.ru



- Базируется на словаре А.А. Зализняка (основ), включает 174 тыс. лемм, можно править
- При анализе выдается лемма, часть речи, значения МП
- Синтезируется полная парадигма
- Незнакомые слова обрабатываются по аналогии

	Found	Dict ID	Lemma	Grammems
<i>Делай</i>	+	пе, нс	ДЕЛАТЬ	Г дст,пвл,2л,ед
<i>Печь</i>	+	пе, нс	ПЕЧЬ	ИНФИНИТИВ, дст
	+	но	ПЕЧЬ	С жр,вн,им,ед,
<i>Покемон</i>	-	но	ПОКЕМОН	С мр,вн,им,ед
				АНЕМОН

Морфологический анализ и синтез: выводы



- ❑ Для построения морфопроцессоров в основном используется инженерный подход. Опора на:
 - ✓ словари основ, словоформ, флексий, исключений
 - ✓ правила преобразования словоформ
 - ✓ эвристические правила анализа новых слов и пр.
- ❑ Для построения статистических моделей требуется обработка больших объемов текстов
- ❑ Отдельная задача – снятие/разрешение омонимии (постморфологический анализ)
- ❑ Оценки качества работы – через аккуратность

$$A = (TP + TN) / (TP + FP + FN + TN)$$

Определение леммы – до 98%, новые слова – до 79%

Примеры морфоанализа для других ЕЯ



❑ Немецкий язык – необходимо делить на слова:
Donaudampfschiffahrtskapitän – капитан рейса,
выполняемого пароходом по Дунаю

❑ Турецкий язык – нужно выделять аффиксы:

Muvaffak – успешный

Muvaffakiyet – успех

Muvaffakiyetsiz – неуспешный

Muvaffakiyetsizleş – стать неуспешным



Muvaffakiyetsizleştiricileştiriveremeyebileceklerimizdenmişsinizcesine –
как будто бы ты оказался среди тех, кого нам непросто будет
сделать теми, кто делает кого-то неуспешными

Синтаксический анализ (СА)



- Единица обработки – предложение (выражает законченную мысль)
- **На входе:** результат морфологического анализа
На выходе: *синтаксическая структура* предложения
- СА предшествует семантическому анализу
- Иногда отдельно рассматриваются простые предложения и словосочетания: *частичный СА*
- Модель СА взаимосвязано включает:
 - ❖ способ представления синтаксической структуры предложения
 - ❖ способ описания грамматических правил
 - ❖ метод/алгоритм синтаксического анализа

Синтаксические модели



- Общее моделей: выход – *синтаксическое дерево*
- Отличие моделей:
 - ❖ выделяемые *синтаксические единицы*
 - ❖ учитываемые *синтаксические связи* между ними
- Основные подходы к моделированию синтаксической структуры:
 - ❖ *деревья составляющих* (возникли в американской лингвистике)
 - ❖ *деревья зависимостей/подчинения* (во французской и русской лингвистике)
 - ❖ гибридные модели, например, *теория синтаксических групп Гладкого*

Модель синтаксиса: деревья составляющих



- Синтаксические единицы – *непосредственные составляющие*: группы соседних слов текста
 - По факту, составляющие – словосочетания (фразы)
 - Два крайних случая составляющих: слово и предложение
 - Связь составляющих: *отношение вложения*
 - Вложенность изображается как *дерево составляющих* (ДС)
 - Любые две составляющие либо не пересекаются, либо одна целиком содержится в другой
- ((Серая птичка) (весело (поет (незатейливую песенку))))*

Деревья составляющих: грамматика



- Для предложения возможны разные деревья, некоторые из них грамматически правильны
- Грамматически правильная синтаксическая структура обычно фиксируется КС-грамматикой

- Пример *КС-грамматики*:

$S \rightarrow NP VP$

$NP \rightarrow A N \mid Det N \mid N \mid N PP$

$VP \rightarrow VP NP \mid AnV \mid Adv AnV$

$AnV \rightarrow Aux V$

$PP \rightarrow Prep NP$

- Нетерминалы соответствуют типам фраз и обозначениям частей речи слов
- А что есть терминалы?
- Если нетерминалы написаны в узлах, то дерево размеченное

Деревья составляющих: пример



S – предложение

NP – именная группа

VP – глагольная группа

PP – предложная группа

AnV – аналитическая форма
глагола

Det – местоименное прилагательное

N – имя существительное

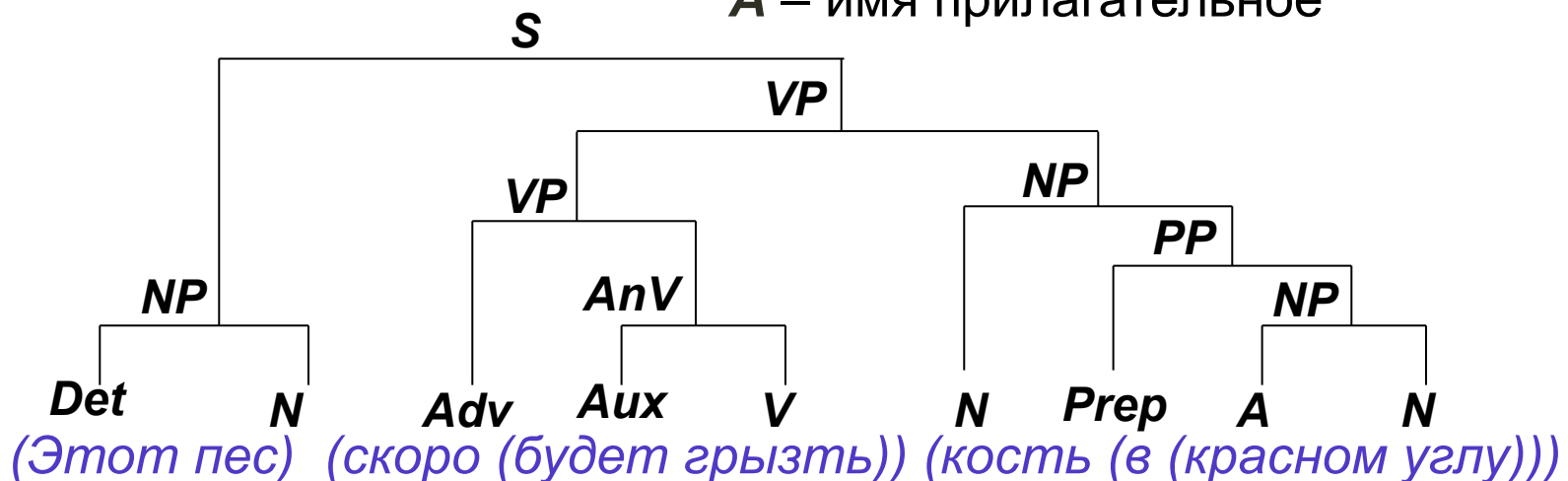
Adv – наречие

Aux – вспомогательный глагол

V – глагол

Prep – предлог

A – имя прилагательное



Деревья составляющих: дополнительные примеры



Пример 1:

– Дорогая, мне посуду мыть или ты сама потом помоешь?

– *Хорошо мой любимый.*

((Хорошо), (мой любимый)).

((Хорошо), (мой), (любимый)).

((Хорошо мой), (любимый)).

Пример 2:

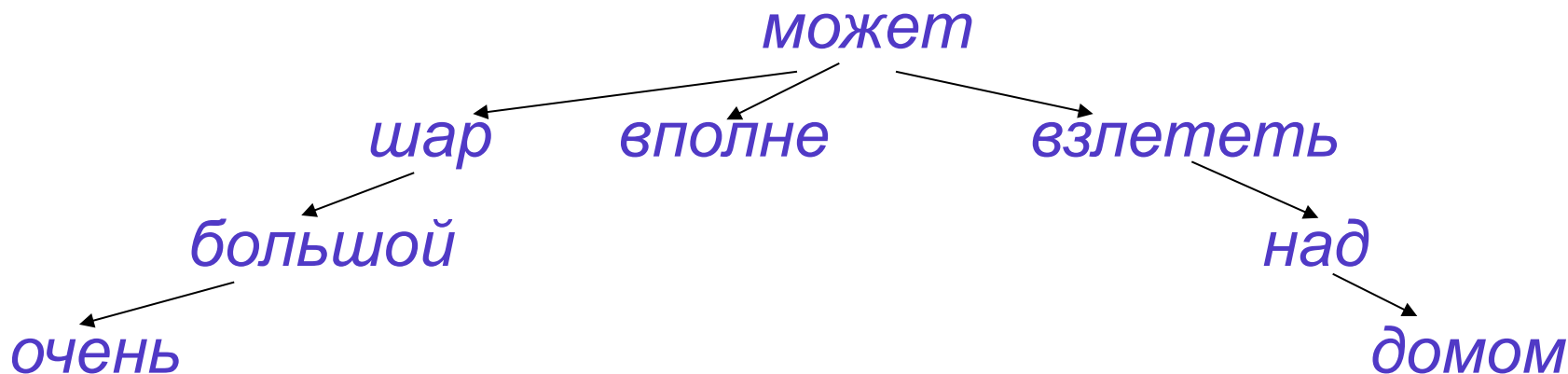
Декан назначен ректором.

(Декан) (назначен ректором).

Модель синтаксиса: деревья зависимостей



- ◆ Синтаксические единицы – слова, синтаксическая связь – подчинение (зависимость)
- ◆ Графическое представление – деревья зависимостей (ДЗ):
 - ✓ узлы – слова (корень – глагол, сказуемое)
 - ✓ дуги – *подчинительная связь* (зависимость)
- ◆ Особенность: нужно задавать порядок слов



Типы подчинительных связей



- ◆ В рамках модели ДЗ количество и набор типов связей слов может варьироваться
- ◆ Наиболее распространенные типы
 - прямообъектное: *уделить → внимание*
 - определительное: *очень ← хорошо*
 - отпредложное: *в → здание*
 - предикат – субъект: *спасатели ← обнаружили*
 - посессивное: *книга → врача*
 - аппозитивное: *мальчик ← Петя*
- ◆ При установлении связей нужно учитывать:
 - ❖ сочетаемость слов
 - ❖ валентность и модель управления
(определяются и синтаксисом, и семантикой)

Сочетаемость слов



Сочетаемость – способность образовывать единицы более высокого уровня

- Грамматическая (синтаксическая): требует определенной грамматической формы

яркий ← свет

песнь → птицы

- Лексическая: проявляется в избирательности лексем

оказать услугу/внимание, но не *заботу/интерес*

- Семантическая: предполагает согласованность ряда сем сочетающихся слов (отсутствие противоречий):

поет человек/птица, но не *апельсин/стол*

Пример установления отпредложной связи



Prep → *N* *на ковер*

- ◆ Если падеж *N* соответствует падежам, обслуживаемым предлогом *Prep*, то установить связь, сделав *Prep* главным словом
- ◆ Если *N* является неизменяемым, то установить связь, сделав *Prep* главным словом
- ◆ В иных случаях связь не устанавливать

N → *Prep* *освобождение от*

- Если *N* является отглагольным, то сделать его главным и установить связь
- В противном случае связь не устанавливать

Деревья зависимостей: проективность



- ◆ Дерево зависимостей часто изображается в виде точек на прямой, между которыми проведены направленные дуги (зависимости)
- ◆ Все дуги располагаются по одну сторону от прямой
- ◆ Проективность («правильность» фразы) – возможность изобразить зависимости слов так:
 - а) ни одна из дуг не пересекает другую дугу,
 - б) никакая дуга не накрывает вершину (корень дерева)
- ◆ Содержательный смысл проективности: в тексте синтаксически связанные слова близки к друг другу

Очень большой шар вполне может взлететь над домом

Слабая проективность и непроективность



Большинство правильных предложений русского языка проективны. Однако возможна:

- ◆ *Слабая проективность*: дуги не пересекаются, но допускается накрывание корня

Кубок все мечтают выиграть

- ◆ *Непроективные* предложения встречаются в художественной литературе и разговорной речи. Усложняют автоматический СА

Кубок все выиграть мечтают

Сравнение моделей СА



Представление	ДС	ДЗ
Линейной структуры	естественное	отсутствует
Подчинительных связей	через вложение	только между словами
Неподчинительных связей	естественное	неоднозначное
Словосочетаний (фраз)	явное	отсутствует
Отношений между фразами	частичное	отсутствует
Разорванных единиц и непроективности	отсутствует	естественное

ДС хороши для языков со строгим порядком слов,
ДЗ – для языков со свободным порядком

Частичный СА



- Для улучшения результатов СА применяется этап *предсинтаксического анализа (частичный СА)*:
 - ❖ выделение простых предложений
 - ❖ выделение словосочетаний
- Словосочетания различаются:
 - ❖ по составу: двучленные, трехчленные и т.п.:
синее небо, передать книгу преподавателю
 - ❖ по типу синтаксической связи
- Структура словосочетаний задается с помощью морфосинтаксических образцов
 - $A\ N\ <A=N>$ *голубой щенок*
 - $N\ N\ <c=gen>$ *сокровища нации*
 - $V\ N\ <c=acc>$ *поставить зачет*

Парсеры для СА: подходы к построению



- Инженерный подход: строится грамматика составляющих или зависимостей
 - ✧ экспертами-лингвистами
 - ✧ автоматизированно на основе корпусов текстов с синтаксической разметкой
 - ✓ грамматика может быть встроена в парсер или записана явно и отделена от него
- Статистический подход: сбор статистики на большом корпусе с синтаксической разметкой (для получения наиболее вероятного дерева)
- Гибридный подход

Стратегии и алгоритмы СА для грамматик

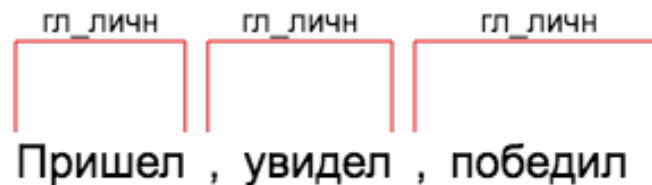
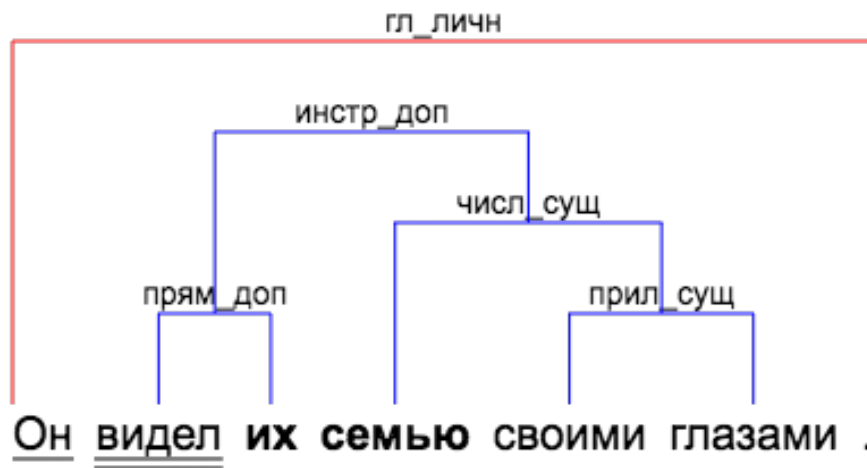


- Алгоритмы для грамматик составляющих – на основе КС-грамматик:
 - нисходящие, восходящие, комбинированные (алгоритм Кока-Янгера-Касами, Эрли)
 - в общем случае – недетерминированный разбор с возвратами
- Алгоритмы для грамматик зависимостей:
 - метод фильтров: применяются все правила установления связей, потом отбрасываются ошибочные структуры (путем фильтрации)

Синтаксис aot.ru



- Синтаксический анализ на основе синтаксических групп (гибридная модель синтаксиса)
- Особенности: нет цели получить полную синтаксическую структуру предложения



Синтаксический анализ:

ВЫВОДЫ



- ❑ Для отдельных предложений возможно более одной правильной синтаксической структуры
Я видел его молодым... Мать любит дочь.
- ❑ Иногда получить различные деревья можно только при привлечении семантики
Вася встретил Петю в коридоре.
Вася встретил Петю в костюме.
- ❑ Парсеры могут выдавать одно или все возможные деревья СА (но не все они верны)
- ❑ F-мера – до 95% (при учете семантики)

Морфопроцессоры для РЯ



- Mystem (Яндекс)
 - ✓ свободный доступ, закрытый код
 - ✓ словарь закрыт (НКРЯ), можно подключить свой
 - ✓ анализ и снятие омонимии
- Rymorphy2 (написан на Python)
 - ✓ свободный доступ
 - ✓ словарь на основе Opencorpora (можно править)
 - ✓ анализ, снятие омонимии, синтез
- Treetagger – языконезависимый инструмент морфологической разметки (нужен словарь)
- Snowball – язык программирования для создания алгоритмов стемминга

Парсеры для РЯ



- Лингвистический процессор ЭТАП
 - ✓ разрабатывается с 1972 г. в ИППИ РАН
 - ✓ анализ текста и извлечение из него смысла, машинный перевод
 - ✓ СА осуществляется в связке со словарем
 - ✓ грамматика зависимостей
- Link Grammar for Russian
 - ✓ построен на основе Link Grammar Parser
 - ✓ СА осуществляется в связке со словарем
 - ✓ грамматика связей
- MaltParser
 - ✓ грамматика зависимостей
 - ✓ SVM для обучения модели СА



Спасибо за внимание!