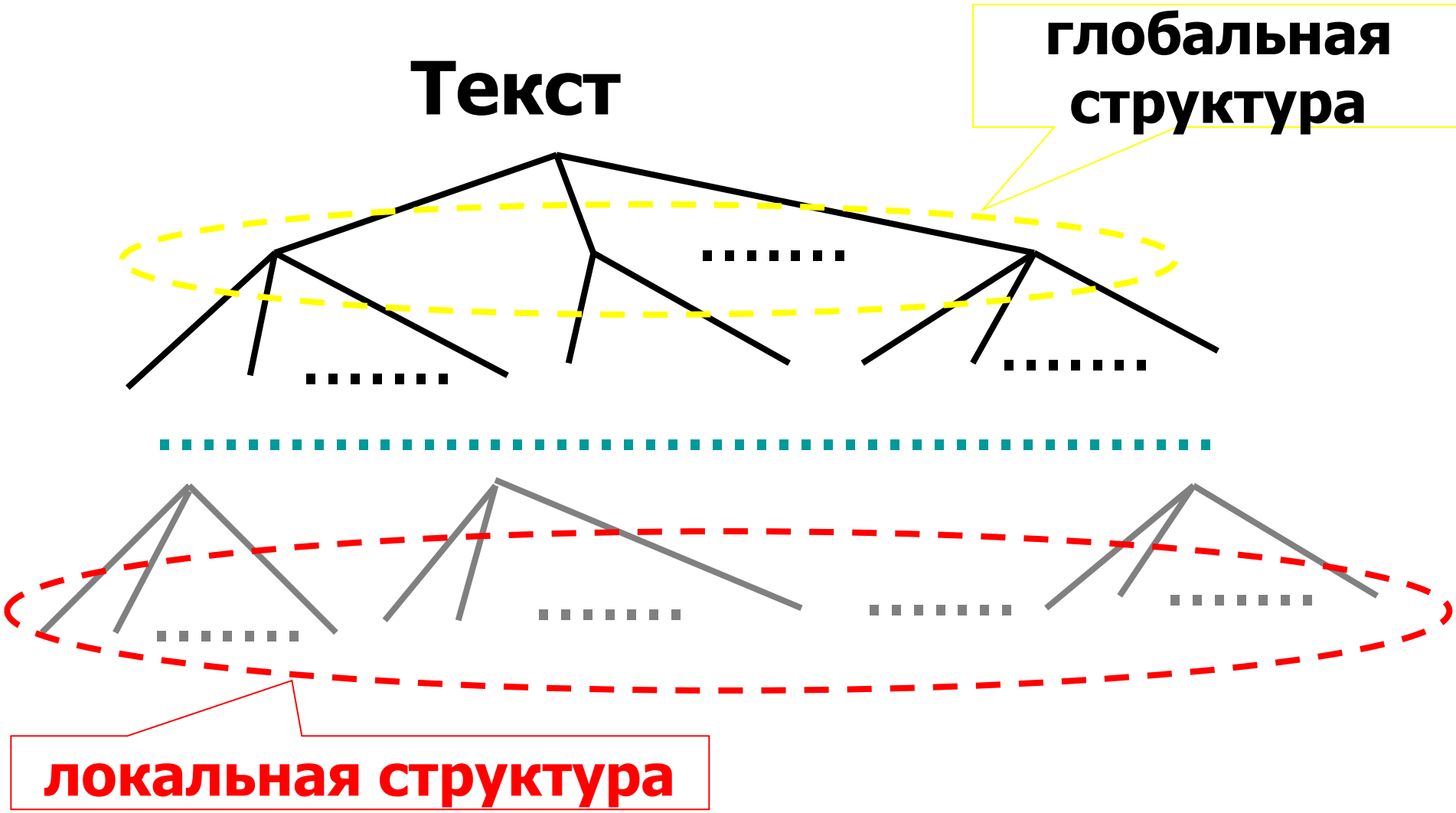
Структура связного текста

Связный текст (дискурс)

- Глобальная структура (макроструктура)
- Локальная структура (микроструктура)

- Глобальная связность
- Локальная связность

Структура как иерархия



Структурные единицы текста

- Элементарные дискурсивные единицы
  - Клаузы – простые предложения, обороты в составе сложного предложения
- Предложения
- Реплики в диалоге
- Абзацы в письменном тексте
- Разделы

Типы связности

- Типы связности по Т. Гивону
  - Референциальная
  - Локативная
  - Темпоральная
  - Событийная

- Типы связности по Halliday and Hasan
  - Связность на основе союзов и дискурсивных маркеров
  - Лексическая связность
  - Референция

Пример (А. Аверченко)

По пустому коридору **раздались** гулкие шаги, шелест женских юбок, и чья-то рука неожиданно громко **постучала** в мою дверь.

Машинально я **сказал**:

— Войдите!

Это была немолодая женщина, скромно одетая, с траурным крепом на шляпе.

Я **вскочил** с дивана, **сделал** по направлению к посетительнице три шага и удивленно **спросил**:

— Чем могу быть вам полезным?

Она внимательно **всмотрелась** в мое лицо.

Когезия: маркирование связности

По пустому коридору **раздались** гулкие шаги, шелест женских юбок, и чья-то рука неожиданно громко **постучала** в мою дверь.

Машинально я **сказал**:

— Войдите!

Это была немолодая женщина, скромно одетая, с траурным крепом на шляпе.

Я **вскочил** с дивана, **сделал** по направлению к посетительнице три шага и удивленно **спросил**:

— Чем могу быть вам полезным?

Она внимательно **всмотрелась**

референциальная

локативная

темпоральная

Дискурсивные маркеры

- Чаще всего – слова, иногда словосочетания
- Кодируют значения, отличные от пропозиционального содержания (или от истинностной оценки)
- Соотношение с традиционными частями речи:
  - частицы
  - союзы
  - наречия

НАТО за год заплатило талибам более $1,2 млн за отказ от насилия

- Более 2,7 тысячи талибов прошли организованную контингентом НАТО в Афганистане трехмесячную программу реабилитации…
- **Таким образом**, всего на эти цели пошло более 1,2 миллиона долларов..
- **Кроме того**, они должны также пройти особый трехмесячный курс**..**
- **В то же время** у программы находятся и критики**.**
- **Так**, экс-глава МВД Афганистана Ханиф Атмар (Hanif Atmar), считает ..

Классы дискурсивных слов

- неполнота: *едва, с трудом, почти*
- реальность: *действительно, в самом деле*
- обобщение: *вообще, в принципе*
- полнота: *вовсе, совсем*
- минимизация: *прямо, просто*
- элемент/множество: *только, всего лишь*
- реализация: *опять, наоборот*
- «установочная база»: *таки, все равно, как раз*
- гарант: *разве, небось, конечно*

Лексическая связность текста

- Типы лексической связности
  - Повторение
  - Синонимическое повторение
  - Связность через обобщение или специализацию (родовидовые отношения)
  - Связность через отношения часть-целое
  - Связность на основе других отношений

- Лексические цепочки

Лексическая связность: пример-1

О порядке оказания безвозмездной финансовой помощи на строительство (покупку) жилья и выплаты денежной компенсации за наем (поднаем) жилых помещений военнослужащим и гражданам, уволенным с военной службы

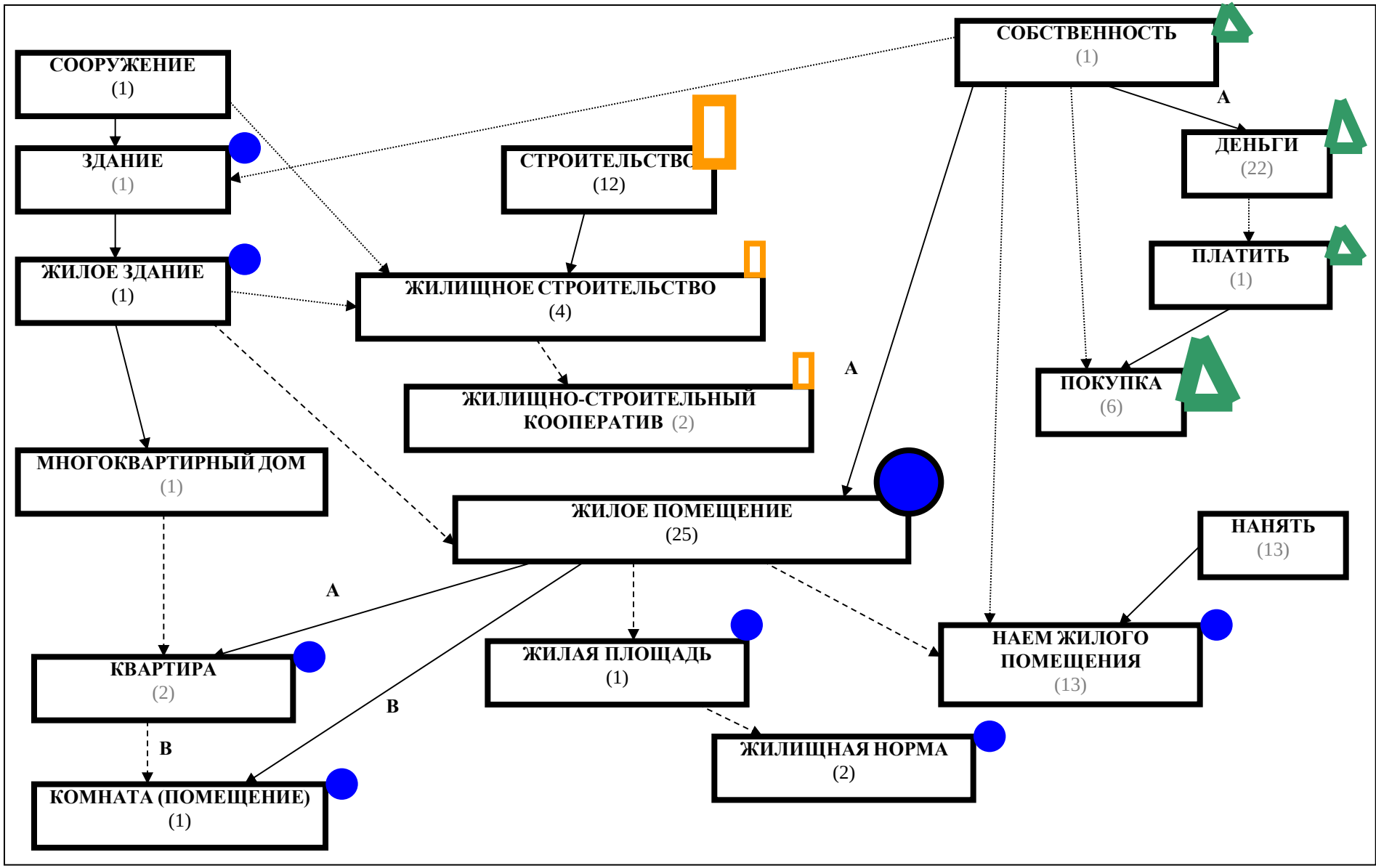
Во исполнение Закона Российской Федерации "О статусе **военнослужащих**" и в целях обеспечения прав на **жилище военнослужащих** и **граждан**, уволенных с **военной службы**, Правительство Российской Федерации п о с т а н о в л я е т :

1. Утвердить прилагаемое Положение о порядке оказания безвозмездной **финансовой помощи** на **строительство** (**покупку**) **жилья** и выплаты **денежной** компенсации за наем (**поднаем**) **жилых помещений** **военнослужащим** и **гражданам**, уволенным с **военной службы**.

2. Министерству обороны Российской Федерации и иным федеральным органам исполнительной власти, в которых предусмотрена **военная служба**:

в месячный срок разработать и утвердить формы и перечень документов, необходимых для принятия решения об оказании **военнослужащим** безвозмездной **финансовой помощи** на **строительство** (**покупку**) **жилья** и о выплате **денежной** компенсации за наем (**поднаем**) **жилых помещений**;

Тезаурусные отношения для документа

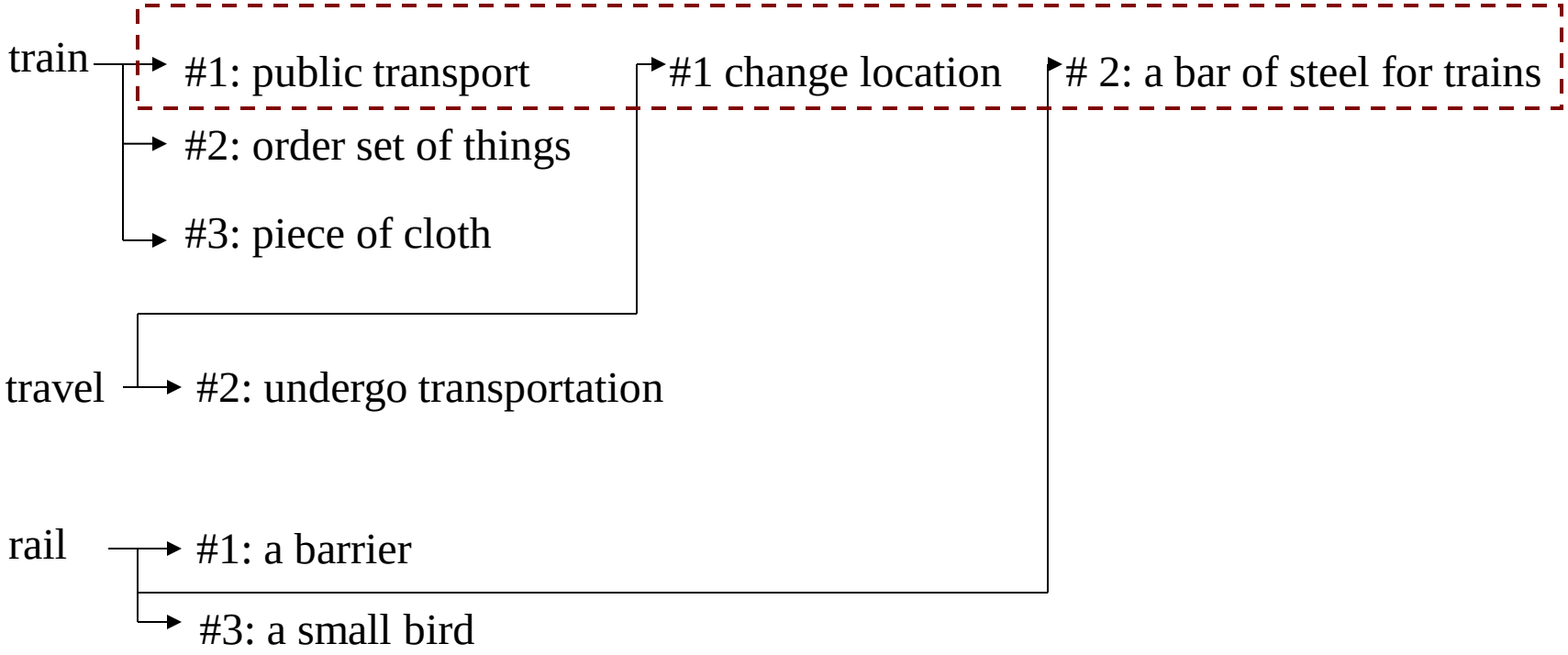


Лексические цепочки

- (Hirst and St-Onge 1988), (Haliday and Hassan 1976)
- Лексические цепочки – последовательность близких по смыслу слов, в которых проявляется лексическая связность связного текста
- Алгоритм создания лексических цепочек
  1. Извлечение слов из текста, между которыми может быть определена мера семантической близости
  2. Двигаясь сначала текста, для каждого очередного слова просматриваются имеющиеся цепочки
  3. Если есть цепочка, то слово присоединяется
  4. Если несколько, то цепочка, в которой близость больше
  5. Обычно есть ограничения на расстояние (предложение, абзац) до последнего слова цепочки

Пример: лексические цепочки и разрешение многозначности

A very long **train traveling** along the **rails** with a constant **velocity** v in a certain **direction** …



Использование лексических цепочек

- Уход от принципа мешка слов
- Уточнение весов слов
  - Слово имеет низкую частоту в тексте, но в тексте много близких по смыслу слов, -> вес слова должен быть больше
- Разрешение многозначности
  - Использование относительно дальнего (соседние предложения) контекста слова
- Автоматическое аннотирование
- Распознавание референции

  … вошел в здание. Дом…

Автоматическое разрешение референции

Понятия и определения

- **Референ́ция** — отнесенность актуализованных (включённых в речь) имён, именных групп или их эквивалентов к объектам внеязыковой действительности (референтам, денотатам).
- **Референ́т** — объект внеязыковой действительности, который имеет в виду говорящий в контексте конкретной языковой ситуации; предмет референции.
- **Корефере́нтность** — отношение между именами, имеющими один референт; то есть отношение между компонентами высказывания, которые обозначают один и тот же объект внеязыковой действительности.

Понятия и определения-2

- **Ана́фора** — ссылка на что-то, упомянутое ранее. Анафорическое слово или анафорическая ссылка соотносятся с антецедентом – другими словом или предложением.
- **Антецедент -** предыдущая единица высказывания (слово, словосочетание или предложение), замененная местоимением или какой-либо фигурой речи

- В предложении "Макс – миллионер, я тоже хочу им стать" – "миллионер" – антецедент, "им" – анафорическое слово.

Автоматическое разрешение референции

•Важная информация, относящаяся к одному событию, может содержаться не в одном предложении текста, а располагаться в соседних предложениях или в разных частях предложения

•Чтобы собрать все элементы сообщения о событии, нужно разобраться с референцией

*«**Иванов** разбил очки **Петрову**, за это **его** наказали.»*

Семантическая задача: определить кого наказали?

Два типа анафорических выражений

- Существительное или группа существительного
  - Нужно сначала определить наличие анафорического отношения
- Местоимения
  - Наличие анафорического отношения очевидно. Нужно найти это отношения

Референция существительны

1)Анафорическое отношение присутствует

*«**Президент Медведев** за дальнейшее сокращение часовых поясов. **Дмитрий Медведев** сегодня заявил, что считает возможным дальнейшее сокращение часовых поясов в России. **Президент** напомнил, что уже принят ряд решений по переводу пяти субъектов России в новые для них часовые пояса.»*

2) Анафорическое отношение отсутствует – абстрактное обозначение сущностей

*«**Президент** — выборная должность главы государства»*

*«Перед вступлением на должность **президент** обязан принять присягу государству»*

Референция местоимений

А) Личные местоимения (я, ты, Вы, он, она, …):
*«**Я** категорически против вступления России в ВТО» сказал **глава КПРФ***

Б) Возвратное местоимение (себя, себе, собой, собою): *«**Я** купил **себе** машину»*

В) Притяжательные местоимения (мой, твой, наш, Ваш, его, …)
*«**Ваш** автомобиль превысил скоростной режим» сказал инспектор **водителю**.*

Референция местоимений-2

Г) Вопросительные местоимения (какой, каков, чей, который): *«**Какая планета** третья от Солнца?»*

Д) Указательные местоимения (этот, это, тот, такой, таков): «**Этот пример** не самый подходящий»

Методы и подходы

Первым шагом, который присутствует во всех принципах и подходах разрешения референции, является определения кандидатов-референтов по номинационным свойствам:

- число и род
- одушевленность/неодушевленность
- и.т.д.;

т.е.эти свойства у антецедента(референта) и его анафоры должны совпадать или по крайней мере не различаться.

Кандидатами могут быть только слова и фразы из данного или предшествующих предложений текста.

Число кандидатов может быть очень велико.

Эвристические подходы

- Оценка по расстоянию и местоположению:
  - выбирается ближайший объект выше по тексту,
  - объекту приписывается некоторый вес, в зависимости от референциального выражения,
  - ограничение просмотра.

- Зависимость от типа референциального выражения:
  - для имен собственных ищется во всем тексте,
  - для личных местоимений - в текущем предложении и в нескольких предложениях

Упоминание референта в одном предложении

Двойное употребление референта в одном предложение
- только в составе двух разных пропозиций (базовой и осложняющей), т.е. разделяются запятой
- иначе имеется семантическое противоречие (референт участвует в одной ситуации в различных ролях)

Референт в единственном числе при последнем своем упоминании не должен входить в состав группы однородных членов предложения:
*«**Сидоров** столкнулся с Ивановым и Петровым в дверях, после чего **ему** не удалось избежать*

Проблемы и сложности

Неоднозначность текстов

*1) «Простой(**прил**.) солдат(**ед. ч., им. п.**)»*
и *«Простой(**сущ**.) солдат(**мн. ч., род. п.**)»*

*«Г-н **Песня [один из вариантов жр./од]** не уточнил, какую сумму он получил, продав компанию, сказав только, что **ее** оборот за 2008 год составил порядка…»*

Проблемы и сложности-2

2) Однородные члены предложения

Анафора множественного числа ссылается на группу однородных частей предложения

*«Выходец из питерских коридоров власти **Виталий Мутко и личный тренер премьера по горным лыжам Леонид Тягачев** считаются хорошими знакомыми Владимира Путина. … В ближайшее время **оба** будут вызваны на ковер.»*

3) Выделение именных групп

Необходимо выделять «***личный тренер премьера по горным лыжам Леонид Тягачев***», а не просто «***тренер***» или «***Леонид Тягачев***»

Проблемы и сложности-3

4) Референтное и нереферентное употребление

•Отношение присутствует:

*«**Президент Медведев** за дальнейшее сокращение часовых поясов. **Дмитрий Медведев** сегодня заявил, что считает возможным дальнейшее сокращение часовых поясов в России. **Президент** напомнил, что уже принят ряд решений по переводу пяти субъектов России в новые для них часовые пояса.»*

•Отношение отсутствует (абстрактное обозначение объектов или типов объектов):

*«**Президент** — выборная должность главы государства» «Перед вступлением на должность **президент** обязан принять присягу государству»*

Проблемы и сложности-4

5) Цитаты

Обычно часть текста находящаяся внутри цитаты оценивается отдельно от остального текста, однако иногда это может быть неверным:

*«М.Погосян сообщил, что уже "определены контуры **двигателя** второго этапа", но уточнил, что цикл **его** создания займет 10-12 лет.»*

6) Проблема вводных слов

При расположение референта довольно далеко от анафоры вводные слова могут получить наибольшую оценку (если их не исключить из рассмотрения):

в нужный момент, пользуясь случаем, по последним данным.

Проблемы и сложности

7) Достоверность определения рода

Не всегда можно по слову достоверно определить его род:

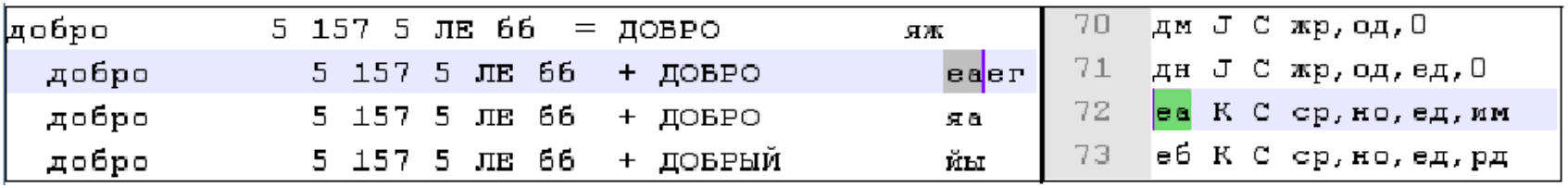
•*«В свою очередь сама **премьер-министр** заявила, что **ее** блок будет оставаться в коалиции, руководствуясь демократическими принципами.»*

Так здесь местоимение «**её**» ссылается на слово «***премьер-министр***», если не учитывать окружение этого слова, то оно не будет рассматриваться в качестве кандидата в референты из-за несовпадения рода.

Разрешение референции местоимений в новостных текстах

Предварительная обработка

- Разбивка текста на предложения
- Выделение слов, знаков препинания и прочих объектов в предложении
- Сопоставление слов из текста с результатом работы морфологического анализа



```
добро          5 157 5 ЛЕ бб  = ДОВРО          яж
  добро          5 157 5 ЛЕ бб  + ДОВРО          еаег
  добро          5 157 5 ЛЕ бб  + ДОВРО          яа
  добро          5 157 5 ЛЕ бб  + ДОБРЫЙ         йы
```

```
70  дм J С жр,од,0
71  дн J С жр,од,ед,0
72  еа К С ср,но,ед,им
73  еб К С ср,но,ед,рд
```

- Выделение цитат
- Определение однородных членов предложения

Базовый алгоритм

Факторы оценки потенциального референта:

- взаимное расположение местоимения и кандидата в референты – количество предложений между, количество грамматических основ между, положение внутри цитаты
- количество совпавших атрибутов – род, число
- одушевленность – наибольшая оценка одушевленным
- падеж кандидата

Устранение неоднозначности

- Фильтрация падежей слов на основе предшествующих им предлогов и предложных слов:

  *«**благодаря** фракции [~~рд,~~ дт, ~~пр, им, вн~~]»*

- Подключение синтаксического анализа (Диалинг АОТ) и корректировка на его основе:
  - Частей речи
    - ПОДЛ {дорога [СУЩ, ~~ПРИЛ~~] -> прокладывается}
  - Падежей
    - ПРЯМ_ДОП {дали -> показания [~~рд, им~~, вн]}
  - Множественности слов
    - ЧИСЛ_СУЩ {чиновника [~~ед~~, мн] -> оба}

Устранение неоднозначности - 2

Создание базы сущностей – наследование атрибутов

Для некоторых имен собственных морфологический анализатор не предоставляет никаких атрибутов, либо только неверные наборы

- *«Финская **компания Tieto []** намерена вложить более 130 млн долл. в создание центров разработки в российских технопарках.»*
- *«**Г-н Песня [жр]** не уточнил, какую сумму он получил, продав компанию, сказав только, что **ее** оборот за 2008 год…»*
- *«По словам аналитика iKS-Consulting **Константина Анкилова [жр]**,…»*

Устранение неоднозначности - 3

При проходе текста основные атрибуты наследуются от уточняющего слова, при этом создается следующая база сущностей

Основное слово	Дополнительные слова	Наборы атрибутов
Tieto	компания	жр, ед, неодуш
Песня	Юрий гы	мр, ед одуш

Основываясь на данной базе корректируются наборы атрибутов слов по всему тексту.

Вывод программы разрешения референции местоимений

1	**Глава** [0, 1] **Renault** [2] **верит** [3], **что** [4, 5, 6] **инвестиции** [7, 8, 9, 10, 11] **его** [**12** refTo {0 SC=**22,083**, }, **13** refTo {0 SC=**23,077**}, **14** refTo {0 SC=**23,071**}, 15, 16] **компании** [17, 18, 19, 20, 21] **в** [22] **автопром** [23] **РФ** [24] **окупятся** [25] .
2	**Глава** [26, 27] **французского** [28, 29] **автопроизводителя** [30, 31] **Renault** [32] **Карлос** [33] **Гон** [34, 35] **ожидает** [36], **что** [37, 38, 39] **инвестиции** [40, 41, 42, 43, 44] **его** [**45** refTo {26 SC=22,053, 30 SC=20,067, 31 SC=20,071, 33 SC=**22,083**, }, **46** refTo {0 SC=20,022, 26 SC=23,050, 30 SC=21,062, 31 SC=21,067, 33 SC=**23,077**, }, **47** refTo {0 SC=20,021, 26 SC=23,048, 30 SC=21,059, 31 SC=21,062, 33 SC=**23,071**, }, **48** refTo {24 SC=**11,542**}, **49** refTo {24 SC=**11,540**}] **компании** [50, 51, 52, 53, 54] **в** [55] **российский** [56] **автопромышленный** [57] **комплекс** [58] **со** [59] **временем** [60] **окупятся** [61].
3	**С** [62] **таким** [63] **заявлением** [64] **он** [**65** refTo {26 SC=20,026, 30 SC=18,029, 31 SC=18,029, 33 SC=**20,031**, }] **выступил** [66] **в** [67] **интервью** [68] **французской** [69, 70] **радиостанции** [71, 72] **RTL** [73].
4	**В** [74] **качестве** [75] **оснований** [76, 77] **для** [78] **подобного** [79] **оптимизма** [80] **он** [**81** refTo {0 SC=11,346, 26 SC=14,018, 30 SC=12,020, 31 SC=12,020, 33 SC=**14,021**, }] **указал** [82] **на** [83] **реструктуризацию** [84] **убыточного** [85] **АВТОВАЗа** [86], **в** [87] **котором** [88, 89] **Renault** [90] **принадлежит** [91] **25%** **акций** [92], **а** [93, 94, 95] **также** [96] **меры** [97, 98, 99] **правительства** [100, 101, 102], **направленные** [103, 104, 105, 106] **на** [107] **стимулирование** [108] **спроса** [109] **на** [110] **автомобили** [111].

Тестирование

- Тестирование относительно ручной разметки
- Тестируется точность разрешения

ЭТАП	**НАСТРОЕЧНЫЙ**	**ПРОВЕРОЧНЫЙ**
Базовый уровень	76,1%	75,32%
Устранение неоднозначности	81,2%	80,6%

Увеличение числа факторов

+машинное обучение

Признаки для разрешения анафоры-1

- 1. число имен собственных между анафором и антецедентом;
- 2. количество предложений, разделяющих анафор и антецедент;
- 3. стоит ли антецедент в именительном падеже;
- 4. является ли антецедент именем собственным;
- 5. количество существительных и местоимений, расположенных в предложениях между рассматриваемыми анафором и антецедентом;
- 6. совпадает ли падеж анафора и антецедента;
- 7. статистическая информация о том, в каком сегменте предложения располагается антецедент – насколько ближе к началу;

Признаки для разрешения анафоры-2

- 8. статистическая информация о том, в каком сегменте предложения располагается анафор – насколько ближе к началу;
- 9. количество анафоров, референтующих с текущим антецедентом по данным ручной разметки, расположенных между анафором и антецедентом;
- 10. число глаголов в сегменте, содержащем антецедент;
- 11. число причастий и деепричастий в сегменте, содержащем антецедент;
- 12. число местоименных прилагательных и союзов в сегменте, содержащем антецедент;
- 13. число существительных в именительном падеже в сегменте, содержащем антецедент;
- 14. род, падеж и число анафора и антецедента (в виде бинарных признаков);

Заключение

- Разрешение референции – важный этап анализа текста
- Разрешение референции – многофакторный процесс
- Привлечение большого числа факторов для разрешения референции требует разметки специализированного корпуса и применение методов машинного обучения