# Графематический и морфологический анализ текстов

# Графематический анализ

- Разделение текста на слова, разделители
- Выделение устойчивых оборотов, не имеющих словоизменительных вариантов
- Выделение предложений
- Выделение абзацев
- Выделение дат
- Определение языка слова (русский, нерусский)
- Определение формата написания слова (прописные, строчные буквы)

# Сегментация текста на слова

Англ. Tokenization

Принципиальные возможности

- в орфографии данного языка предусмотрены пробелы между словами;

- в орфографии данного языка нет пробелов или иных разделителей между словами.

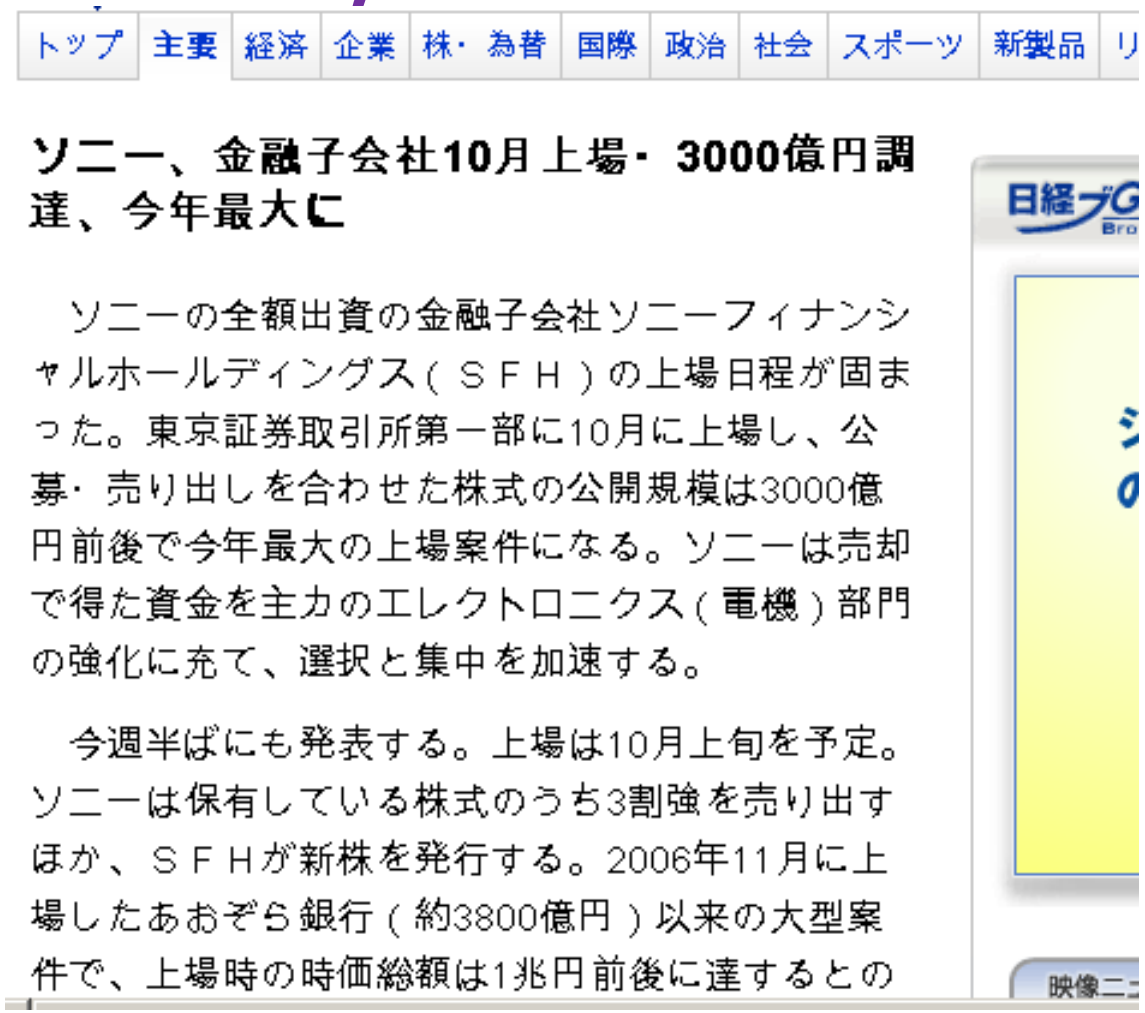
# Сегментация на слова текста с пробелами

Осложняющие факторы:

- сегменты текста между пробелами требуют переразложения
  - *du* = *de* + *le*; *au* = *à* + *le* (франц.), *gdybym* = *gdy* + *bym*, *bardzobym* = *bardzo* + *bym* (пол.); *neunzehnhundertzweiundfünfzig* (нем.)
  - *буду* (*часто*) *писать*; железная дорога; с разбегу; *Du holst mich ab* (нем.)

- словоформы могут разделяться не только пробелами
  - *наконец-то* (vs *кто-то, во-первых, по-моему*), *they're* и *isn't* (vs *friend's*), *and/or* (vs *accept/reject*)

# Сегментация текста без пробелов между словами



トップ 主要 経済 企業 株・為替 国際 政治 社会 スポーツ 新製品 リ

**ソニー、金融子会社10月上場・3000億円調達、今年最大に**

ソニーの全額出資の金融子会社ソニーフィナンシャルホールディングス（ＳＦＨ）の上場日程が固まった。東京証券取引所第一部に10月に上場し、公募・売り出しを合わせた株式の公開規模は3000億円前後で今年最大の上場案件になる。ソニーは売却で得た資金を主力のエレクトロニクス（電機）部門の強化に充て、選択と集中を加速する。

今週半ばにも発表する。上場は10月上旬を予定。ソニーは保有している株式のうち3割強を売り出すほか、ＳＦＨが新株を発行する。2006年11月に上場したあおぞら銀行（約3800億円）以来の大型案件で、上場時の時価総額は1兆円前後に達するとの

# Разбиение текста на предложения

- синтаксический анализ (парсеры)
- системы автоматического реферирования
- машинный перевод
- Извлечение терминов...

Текст, разбитый на предложения

# «Наивная» сегментация

По знакам препинания:

- точка,
- восклицательный и вопросительный знаки

# «Наивная» сегментация

По знакам препинания:

- точка,
- восклицательный и вопросительный знаки

В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).

# «Наивная» сегментация

По знакам препинания:

- точка,
- восклицательный и вопросительный знаки

Политический кризис в Сирии представляет опасность для ближневосточного региона и всего мира. Такое мнение высказал спецпосланник ООН и Лиги арабских государств (ЛАГ) **Л. Брахими**, передает Reuters.

# Практические решения

- Предложение должно содержать буквы
- Предложение должно начинаться с заглавной буквы
- Сокращения (из списка) требуют «особого внимания»
  - г., тыс., млн., ул …
- Отдельные большие буквы: А.Б. Иванов
- ...

# Операции со словами

- Словоизменение
  - изменение одного и того же слова
    - Лесной, лесная, лесного…
- Словообразование
  - образование новых слов
    - Лес, лесник, лесничество…

# Словообразование: есть ли достаточная предсказуемость?

- дневн(ой) — дневник
- вечерн(ий) — вечерник
- ночн(ой) — ночник
- утренн(ий) — утренник

# Есть ли достаточная предсказуемость?

АНАЛИЗ: нет регулярности

| | |
|---|---|
| дневник | {дневной} + тетрадь для записей, заполняемая с указанной периодичностью,<br>{дневной} + студент формы обучения, предусматривающей занятия в указанное время суток |
| вечерник | {вечерний} + студент формы обучения, предусматривающей занятия в указанное время суток |
| ночник | {ночной} + лампа, используемая в указанное время суток |
| утренник | {утренний} + представление, происходящее в указанное время суток |

# Морфологический анализ

- Морфологический анализ текста осуществляет приведение словоформ, встречающихся в тексте, к нормальному (словарному) виду и определяет морфологические характеристики словоформы
- Нормальная форма (=словарная форма=лемма):
  - Существительные, прилагательные
    - им. падеж
    - ед. число
    - мужской род
  - Глагол: инфинитив
- Упрощенная процедура: лемматизация (=восстановление нормальной формы)

- Обратная процедура: морфологический синтез

# Морфологические характеристики словоформ русского языка

- Имя существительное:

  6 падежей * 2 числа
- Имя прилагательное:

  6 падежей * 2 числа (в ед.ч. 3 рода)

  + 4 краткие формы

  + степени сравнения
- Глагол:

  (неопр.ф. + личные формы изъяв.накл. + повел.накл. + прич. + деепр.) * 2 вида
- Неизменяемые части речи…

# Обработка словоформы: морфологический анализ

| | |
|---|---|
| исследовать | {исследовать} +<br>+Неопр.ф. |
| исследую | {исследовать} +<br>+ Наст., Буд. вр. + Ед.ч. + 1 л. |
| исследуешь | {исследовать} +<br>+ Наст., Буд. вр. + Ед.ч. + 2 л. |
| исследует | {исследовать} +<br>+ Наст., Буд. вр. + Ед.ч. + 3 л. |
| … | |
| исследовал | {исследовать} +<br>+ Прош. вр. + Ед.ч. + М р. |
| исследовала | {исследовать} +<br>+ Прош. вр. + Ед.ч. + Ж р. |
| … | |

# Порождение словоформы: морфологический синтез

| | |
|---|---|
| {исследовать} + Неопр.ф. | исследовать |
| {исследовать} + Наст. вр. + Ед.ч. + 1 л. | исследую |
| {исследовать} + Наст. вр. + Ед.ч. + 2 л. | исследуешь |
| {исследовать} + Наст. вр. + Ед.ч. + 3 л. | исследует |
| … | |
| {исследовать} + Буд. вр. + Ед.ч. + 1 л. | исследую, буду исследовать |
| {исследовать} + Буд. вр. + Ед.ч. + 2 л. | исследуешь , будешь исследовать |
| {исследовать} + Буд. вр. + Ед.ч. + 3 л. | исследует, будет исследовать |
| … | |
| {исследовать} + Прош. вр. + Ед.ч. + М р. | исследовал |
| {исследовать} + Прош. вр. + Ед.ч. + Ж р. | исследовала |

…

# Морфологический анализ и лемматизация

| | | | |
|---|---|---|---|
| исследовать | {исследовать} + +Неопр.ф. | исследовать | {исследовать} |
| исследую | {исследовать} + + Наст., Буд. вр. + Ед.ч. + 1 л. | исследую | {исследовать} |
| исследуешь | {исследовать} + + Наст., Буд. вр. + Ед.ч. + 2 л. | исследуешь | {исследовать} |
| исследует | {исследовать} + + Наст., Буд. вр. + Ед.ч. + 3 л. | исследует | {исследовать} |
| … | | … | |
| исследовал | {исследовать} + + Прош. вр. + Ед.ч. + М р. | исследовал | {исследовать} |
| исследовала | {исследовать} + + Прош. вр. + Ед.ч. + Ж р. | исследовала | {исследовать} |
| … | | … | |

# Точнее: типы морфологического анализа

- Лемматизация – приведение к нормальной форме
  - Лесной, лесного, лесному->лесной
  - леса -> лес
  - Танцующая -> танцевать
- Стемминг –выделение псевдоосновы
  - Лесной, лесного, лес, лесистый -> лес
  - Система, системный, систематизировать->
  - систем

# Методы морфологического анализа

a) словарный
   - со словарем словоформ
     - Каждой словоформе поставлена в соответствие основа или лемма
   - со словарем основ

b) бессловарный (фактически – со словарем псевдоокончаний)

   + анализ по аналогии («предсказание»)

# Стемминг: Алгоритм Портера Snowball

- http://snowball.tartarus.org/algorithms/russian/stemmer.html
- Список служебных слов
  - Союзы, предлоги, наречия, частицы
- Для остальных слов отделяются псевдоокончания
- Стеммеры были созданы для распространенных индоевропейских языков

# Анализ алгоритмом Портера

- *Противоестественном*
- В основе слова должна остаться хотя бы одна гласная
- Фрагмент RV: *тивоестественном*
- К фрагменту RV применяются заданные списки окончаний
- При нескольких вариантах выбирается наиболее длинное окончание

# Окончание прилагательных в алгоритме Портера

- ее (***ee***) ие (***ie***) ые (***ye***) ое (***oe***) ими (***imi***) ыми (***ymi***) ей (***ei***) ий (***ii***) ый (***yi***) ой (***oi***) ем (***em***) им (***im***) ым (***ym***) ом (***om***) его (***ego***) ого (***ogo***) ему (***emu***) ому (***omu***) их (***ikh***) ых (***ykh***) ую (***uiu***) юю (***iuiu***) ая (***aia***) яя (***iaia***) ою (***oiu***) ею (***eiu***)

# Словарные морфологии

# Словарь Зализняка

1977

- «Грамматический словарь русского языка»

Автор – Андрей Анатольевич Зализняк (с 1997 г. академик РАН)

100тыс. входов - основа большинства компьютерных морфологий РЯ:

Проблемы:

- Автомобилестроения – мн.ч.
- Финансов – кр. Форма для финансовый
- При – пря
- Много старых слов
- Отсутствуют новые слова

# Фрагмент страницы словаря А.А.Зализняка

**ТЕЧЬ** ж (жо): 8a, 8e, 8f″ — *47* | св (нсв): 8 — *118*

**утéчь** св нп 8b/b (-к-), ё ◑II
**вы́течь** св нп 8a (-к-) ◑II
**дичь** ж 8a
**на́взничь** н
**опри́чь** предл.
**стричь** нсв 8b (-г-)
**застри́чь** св 8b (-г-) ◑II
**настри́чь** св 8b (-г-) ◑II
**обстри́чь** св 8b (-г-) ◑II
**подстри́чь** св 8b (-г-) ◑II
**перестри́чь** св 8b (-г-) ◑II
**остри́чь** св 8b (-г-) ◑II
**достри́чь** св 8b (-г-) ◑II
**постри́чь** св 8b (-г-) ◑II
**простри́чь** св 8b (-г-) ◑II
**состри́чь** св 8b (-г-) ◑II
**расстри́чь** св 8b (-г-) ◑II
**отстри́чь** св 8b (-г-) ◑II
**вы́стричь** св 8a (-г-) ◑II
**засти́чь** *см.* засти́гнуть
**насти́чь** *см.* насти́гнуть
**пристúчь** *см.* пристúгнуть
**достúчь** *см.* достúгнуть
**постúчь** *см.* постúгнуть
**жёлчь** ж 8a [// **желчь** =]

**сво́лочь** жо 8e
**своло́чь** св 8b/b (-к-) [// *простореч.* **сволочи́ть**] ◑I (-а-)
**отволо́чь** св 8b/b (-к-) [// *простореч.* **отволочи́ть**] ◑I (-а-)
**уволо́чь** св 8b/b (-к-) [// *простореч.* **уволочи́ть**] ◑I (-а-)
**вы́волочь** св 8a (-к-) [// *простореч.* **вы́волочить**] ◑I (-а-)
**толо́чь** нсв 8b/b (-к-) △ *наст.* толку́, толчёт, толку́т; *прош.* толо́к, толкла́, толо́кший; *прич. страд.* толчённый
**затоло́чь** св 8b/b (-к-) △ *буд.* затолю́, -чёт, -ку́т; *прош.* -о́к, -кла́, -о́кший; *прич. страд.* -чённый
**натоло́чь** св, *спряж. см.* затолóчь
**втолóчь** св, *спряж. см.* затолóчь
**подтолóчь** св, *спряж. см.* затолóчь
**перетолóчь** св, *спряж. см.* затолóчь
**потолóчь** св, *спряж. см.* затолóчь
**протолóчь** св, *спряж. см.* затолóчь
**столóчь** св, *спряж. см.* затолóчь
**растолóчь** св, *спряж. см.* затолóчь

**то́чь-в-то́чь** н
**запря́чь** св 8b/b (-г-) ◑II
**перезапря́чь** св 8b/b (-г-) ◑II
**напря́чь** св 8b/b (-г-) ◑II
**поднапря́чь** св 8b/b (-г-)
**перенапря́чь** св 8b/b (-г-) ◑II
**впря́чь** св 8b/b (-г-) ◑II
**подпря́чь** св 8b/b (-г-) ◑II
**перепря́чь** св 8b/b (-г-) ◑II
**припря́чь** св 8b/b (-г-) ◑II
**сопря́чь** св 8b/b (-г-) ◑II
**спря́чь** св 8b/b (-г-) ◑II
**распря́чь** св 8b/b (-г-) ◑II
**отпря́чь** св 8b/b (-г-) ◑II
**упря́чь** св 8b/b (-г-) ◑II
**вы́прячь** св 8a (-г-) ◑II
**нао́тмашь** н
**ро́спашь** ж 8a
**гуа́шь** ж 8a
**плешь** ж 8a
**флешь** ж 8a
**брешь** ж 8a
**ишь** част.; межд.
**бишь** част.
**вишь** част.
**дишь** (*часто безударн.*)

# Схема морфологического анализа со словарем

- Для неслужебных слов:
- Выделить возможные окончания слова длиной от 0 до 3 символов
- Для каждого полученного окончания определить код окончания по таблице окончаний и номер флективного класса по словарю основ (лемм)
- Если номер флективного класса и номер окончания найдены, то проверить их согласованность по морфологической таблице
- Если согласованность подтверждается, то сохранить данный вариант

# Процедура определения типовой парадигмы

- если слово оканчивается на *щийся*, то ТП 5;
- если слово оканчивается на *ин*, *ын*, то ТП 20;
- если слово оканчивается на *ов*, *ёв*, *ев*, то ТП 21;
- если слово оканчивается на *цый*, то ТП 6;
- если слово оканчивается на *ый*, то ТП 1;
- если слово оканчивается на *кий*, *гий*, *хий*, то ТП 3;
- если слово оканчивается на *щий*, то ТП 4;
- если слово оканчивается на *жий*, *ший*, *чий*, то ТП 4 или ТП 24;
- если слово оканчивается на *ий*, то ТП 2 или ТП 24;
- если слово оканчивается на *кой*, *гой*, *хой*, *жой*, *шой*, *чой*, *щой*, то ТП 8;
- если слово оканчивается на *ой*, то ТП 7.

# Морфологический анализ на базе словаря: проблемы

- Дают максимально полный анализ словоформы
- На реальных текстах дают сбои (опечатки, уникальные слова)
- Не существует абсолютно полных словарей – лексика языка непрерывно пополняется
- Для примера – невозможно включить в словарь всю существующую терминологию, имена, фамилии и т.д.

# Методы хранения словарей

- Хэширование – хэш-таблица, используется вспомогательная функция, которую называют хэшем.

- Дерево – использование древовидной структуры, поиск осуществляемся с помощью конечного автомата.

# Морфологический анализ слов, отсутствующих в словаре

# Предсказание в морфологическом анализе

- Функциональное назначение предсказания – морфологический анализ слов (словоформ), отсутствующих в словаре

- Метод предсказания – выявление аналогий со словоформами, распознаваемыми имеющимся словарем

# Алгоритм предсказания для новых слов

1) предсказание префиксального образования
2) предсказание по концовке, взятой из известных словоформ

# Предсказание по префиксу

- попытка найти существующую словоформу языка, которая максимально совпадала бы справа со входным словом.

- Если левая часть (потенциальный префикс) не длиннее M символов (пяти), а правая часть (совпавшая с известной словоформой) не короче N символов (четырех), то слово разбирается по образцу известной словоформы.

*[евро]технологию*, *[супер]коньками*

# Предсказание по концовке известной словоформы

Отделяются инвертированные концовки известных словоформа – длины K (пять букв),

Сопоставляются с морфологическими характеристиками:

- *Меина (записано справа налево)*
- как «ср. род, ед. ч., тв. пад.»

Такая строка заносится в исходный лексикон, если она встречается:

- не менее L раз (трех) и
- чаще конкурентов в пределах одной части речи

ВСЕГДА предусматривается разбор именем существительным, хотя бы неизменяемым.

# Проблема морфологической омонимии

*Пример:*

*На завод привезли **стекло**.*

*Масло **стекло** на пол.*

*Нес медведь, шагая к **рынку**,*

*На продажу меду **крынку**.*

*Вдруг на мишку - вот **напасть**!*

*Осы вздумали **напасть**.*

*Мишка с армией **осиной***

*Дрался вырванной **осиной**.*

*Мог ли в ярость он не **впасть**,*

*Если осы лезли в **пасть**,*

*Жалили куда **попало**,*

*Им за это и **попало**.*

## *Как решить?*

# Постморфологический анализ

- =предсинтаксический анализ
- Предназначен для устранения морфологической омонимии (многозначности) слов
  - –Выбор правильной леммы
  - –Уточнение морфологических характеристик

**Основные методы**

- Написание правил,
- Статистические методы, прежде всего, на основе морфологически размеченного корпуса

# Примеры правил постморфологического анализа

- Удаление признаков служебных частей речи для однобуквенных слов, за которыми следуют точки
- Удаление омонимов слова «уже», соответствующих прилагательным, если за ним не стоит запятая или слово в родительном падеже
- Удаление омонимов слова «сорока», если после слова следует числительное (сорок  пять)
- Обработка предлогов: удаление у слова, следующего за предлогом, всех омонимов, не соответствующих падежам, которыми обычно управляет данный предлог

# Статистические методы и морфологическая разметка корпуса

# Морфологическая разметка

- Частеречная разметка, морфологическая разметка (грамматическая разметка):

a) информация о морфологических (грамматических) характеристиках словоформ текста, включаемая в электронное представление этого текста (в виде тегов)

b) процедура добавления такой информации в электронное представление текста (как правило, частично или – редко – полностью автоматизированная)

# Теги морфологической разметки в Нац. корпусе русского языка

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

# Падежи в морфологической разметке

- nom — именительный падеж (*голова, сын, степь, сани, который*)
  gen — родительный падеж (*головы, сына, степи, саней*)
  dat — дательный падеж (*голове, сыну, степи, саням*)
  dat2 — дистрибутивный дательный (*[по] многу, нескольку)*
  acc — винительный падеж (*голову, сына, степь, сани*)
  ins — творительный падеж (*головой, сыном, степью, санями, которым*)
  loc — предложный падеж (*[о] голове, сыне, степи, санях)*
  gen2 — второй родительный падеж (*чашка **чаю***)
  acc2 — второй винительный падеж (*постричься в **монахи**; по **два** человека*)
  loc2 — второй предложный падеж (*в **лесу**, на **оси́***)
  voc — звательная форма (*Господи, Серёж, ребят*)
  adnum — счётная форма (*два **часа́**, три **шара́***)

# Фрагмент морфологической разметки в Национальном корпусе русского языка

- Я сидел на барском сиденье, дышал горячим ветром, бившим в лицо, ощущая в то же время не истребимую никакими сквозняками пыль и легкий запах духов -- катафалк с хорошей скоростью мчался по шоссе на юг. (Ю. Трифонов)

- <s>**Я**{я=S,ед,од=им} **сидел**{сидеть=V,несов=изъяв,прош,ед,муж} **на**{на=PR} **барском**{барский=A=ед,сред,пр} **сиденье**{сиденье=S,сред,неод=ед,пр}, **дышал**{дышать=V,несов=изъяв,прош,ед,муж} **горячим**{горячий=A=ед,муж,твор} **ветром**{ветер=S,муж,неод=ед,твор}**,** **бившим**{бить=V,несов=прич,прош,ед,муж,твор} **в**{в=PR} **лицо**{лицо=S,сред,неод=ед,вин}**,** **ощущая**{ощущать=V=несов,деепр,непрош} **в**{в=PR} **то**{тот=A=ед,сред,вин} **же**{же=PART} **время**{время=S,сред,неод=ед,вин} **не**{не=PART} **истребимую**{истребимый=A=ед,жен,вин} **никакими**{никакой=A=мн,твор} **сквозняками**{сквозняк=S,муж,неод=мн,твор} **пыль**{пыль=S,жен,неод,ед=вин} **и**{и=CONJ} **легкий**{легкий=A=ед,муж,вин,неод} **запах**{запах=S,муж,неод=ед,вин}…

# Процедура морфологической разметки

- Морфологический анализ всех словоформ текста
- Снятие неоднозначностей (или исправление ошибок)
- Добавление информации о результатах в электронное представление текста

# Процедура разметки в Нац. Корпусе русского языка

- Автоматический морфологический анализ (Mystem, Dialing)
- Промежуточная обработка – фильтрация маловероятных вариантов, принудительное введение синкретичных вариантов разбора (Grambat)
- Снятие омонимии – диалоговая утилита (макрос Gramedit)

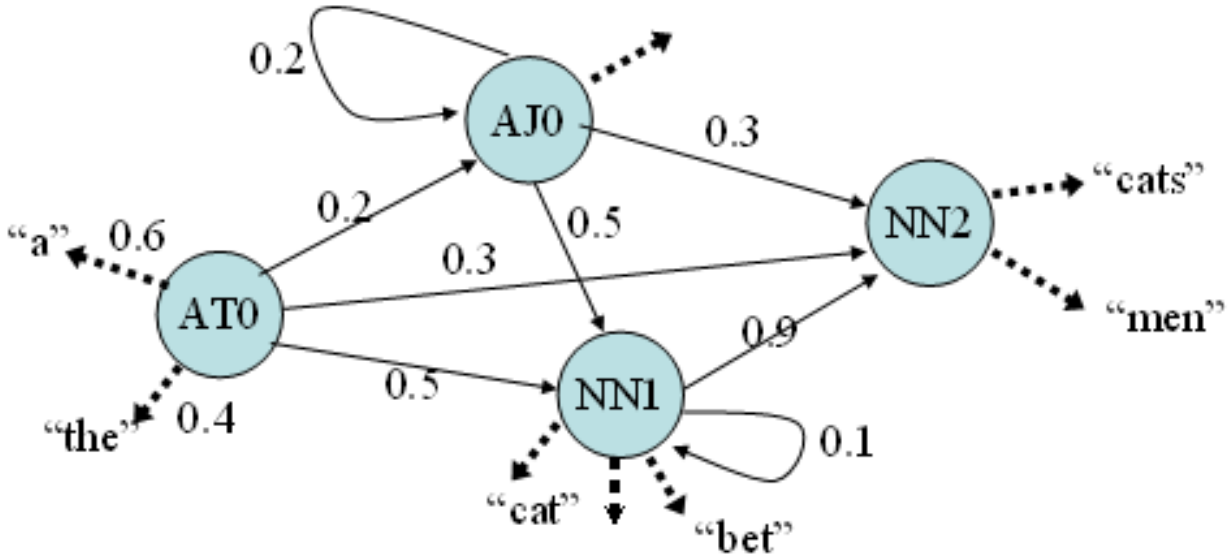
# Скрытые марковские модели как метод снятия омонимии (1)

Предположения о марковском характере зависимости (модель первого порядка):

- - встречаемость каждого тега в определенном месте цепочки зависит только от предыдущего тега;
- - то, какое слово находится в том или ином месте цепочки, полностью определяется тегом (а не, допустим, соседними словами).

Таким образом, порождение правильно построенной цепочки тегов уподобляется действию конечного автомата, где дуги помечены тегами с приписанными им вероятностями, а слова – это наблюдаемые реализации тегов. Состояния определяются парой «текущий тег + предыдущий тег»

# Скрытые марковские модели как метод снятия омонимии (2)



# Задание. Срок 9 октября

- Установить морфологический анализатор:
  - Mystem
  - http://company.yandex.ru/technology/mystem
  - или aot (aot.ru)
- Выбрать текст (литературное произведение)
- Сделать частотные списки лемм