Анализ текстов
на естественном языке

Николай Иванов
325 группа

План доклада

1. Задачи анализа текстов на ЕЯ *
2. Особенности текстов на ЕЯ
3. Этапы анализа текстов на ЕЯ
4. Лингвистические ресурсы

* ЕЯ — естественный язык

Задачи анализа текстов на ЕЯ

1. машинный перевод (Google Translate)
2. генерация текста (FOG)
3. реферирование (MS Word)
4. поисковые системы
   - выделение ключевых слов
   - кластерный анализ (Nigma)
5. человеко-машинные интерфейсы
   - голосовое управление
   - вопросно-ответные системы (WolframAlfa, Siri)

Поисковые системы. Кластерный поиск

курсы

В найденном | в Москве | Поисковики | Я

Фильтр

- курсы английского языка
- курсы иностранных языков
- в москве
- обучение
- компьютерные курсы
- english
- учебные курсы в москве
- курсы валют
- курсы повышения
- курсы дизайна

1. "Центральные **курсы** подготовки специалисто

   Проведение бухгалтерских, компьютерных **курсов**, а также

2. **Курсы**: **курсы** Москвы, популярные **курсы**,

   все танцевальные **курсы** Москвы. ТОП-15: рейтинг **курсов** центров, предлагающих большой выбор... ...

3. НОВОЕ ОБРАЗОВАНИЕ-2 УЧЕБНЫЙ ЦЕНТР

   Новое образование - О нас. О нас. О нас. Наши преимущест Москва, Каменщики Б. ул. 7, офис 302. (495) 9123645

WolframAlpha™ computational knowledge engine

what is the weather like today

Examples Random

Input interpretation:

weather	today

Recorded weather for Champaign, Illinois:

Show metric More

temperature	45 °F (wind chill: 40 °F)
conditions	clear
relative humidity	76% (dew point: 37 °F)
wind speed	9.2 mph

(1 hour 25 minutes ago)

Units »

Человеко-машинные интерфейсы. Siri

Особенности ЕЯ

1. универсальность
2. динамический характер

Особенности текстов на ЕЯ

1. многозначность
2. избыточность
3. эллипсис
4. лексическая сочетаемость

Особенности текстов на ЕЯ. Многозначность (омонимия)

Морфологическая:

«город**а**» — мн.ч., им.п.

«друг**а**» — ед.ч., род.п.

Лексическая:

«ключ»

- дверной ключ
- гаечный ключ
- скрипичный ключ
- родник
- пароль

Синтаксическая:

«Эти типы стали есть на складе»

Особенности текстов на ЕЯ. Многозначность (омонимия)

Морфологическая:

«город**а**» — мн.ч., им.п.

«друг**а**» — ед.ч., род.п.

Лексическая:

«ключ»

- дверной ключ
- гаечный ключ
- скрипичный ключ
- родник
- пароль

Синтаксическая:

«Эти типы стали есть на складе»

Особенности текстов на ЕЯ.

Избыточность — дублирование информации для облегчения ее восприятия.

Я вошел в комнату и увидел **там своего** друга.

Эллипсис — отсутствие слов, подразумеваемых контекстом.

- Ты куда **[идешь]** ?
- **[я иду]** В гости. Какие туфли надеть?
- **[надевай]** Белые **[туфли]** .

Лексическая сочетаемость

«стая волков» VS «стая коров»

Лексическая сочетаемость

«Бесцветные зеленые идеи яростно спят»

Основные этапы обработки текста

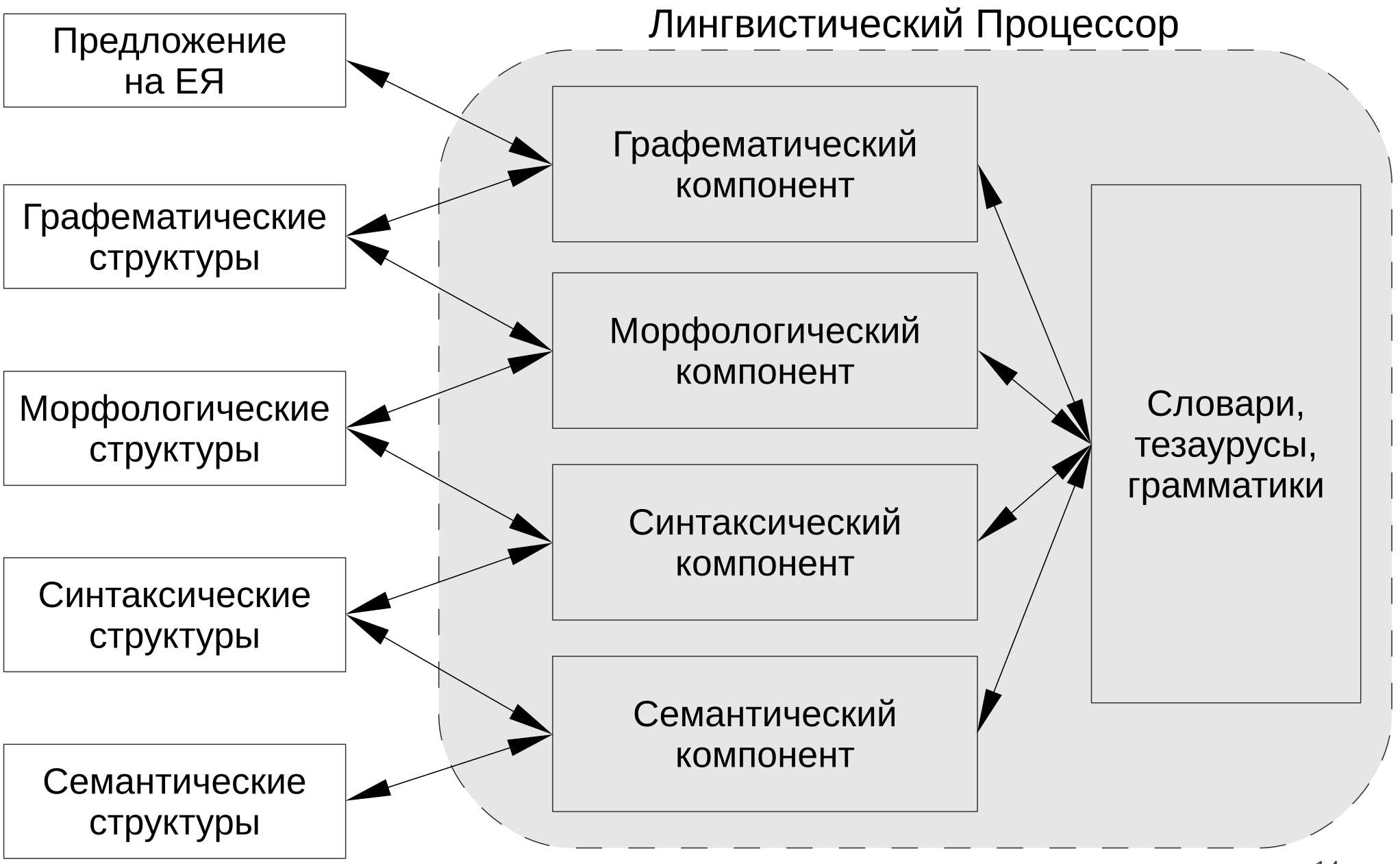
1. графематический анализ
   — разделение текста на графемы

2. морфологический анализ
   — определение морф. характеристик слов

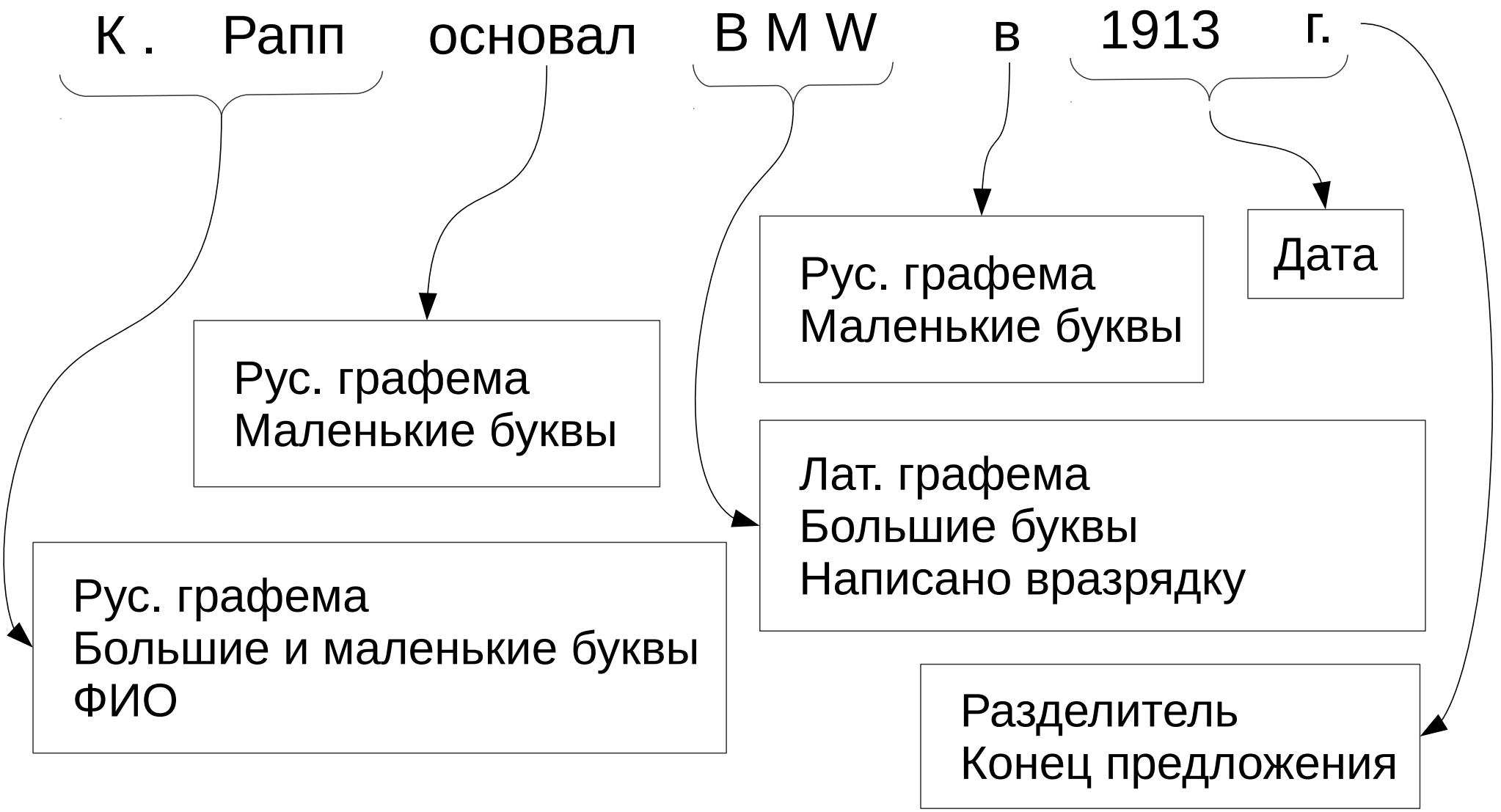
3. синтаксический анализ
   — построение синтаксической структуры предложения

4. семантический анализ
   — построение семантической структуры предложения

Общая схема автоматической обработки



Графематический анализ



Морфологический анализ.

Приближенный подход —

определение морфологических характеристик по последним буквам слова.

Пример для 2-ух последних букв:

	а	**б**	**в**	**г**	**д**	**...**
а		Сущ.	Сущ., Нар.	Сущ.	Сущ., Нар.	
б	Сущ.					
в	Сущ., Гл.		Сущ.		Сущ.	
г	Сущ.					
д	Сущ., Нар.					
...						

олимпиа**да**
ког**да**

ампли**туда**
туда

Морфологический анализ

Декларативный подход — поиск словоформы в словаре.

…	…	…	…	…
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
2609580	96056	одухотворяющих	ПРИЧ	дст,од,нст,вн,мн
2609581	96056	одухотворяющие	ПРИЧ	дст,но,нст,вн,мн
2609582	96056	одухотворяющими	ПРИЧ	дст,но,од,нст,тв,мн
2609583	96056	одухотворяющих	ПРИЧ	дст,но,од,нст,пр,мн
2609584	96056	одухотворявший	ПРИЧ	дст,но,од,прш,мр,им,ед
2609585	96056	одухотворявшего	ПРИЧ	дст,но,од,прш,мр,рд,ед
2609586	96056	одухотворявшему	ПРИЧ	дст,но,од,прш,мр,дт,ед
2609587	96056	одухотворявшего	ПРИЧ	дст,од,прш,мр,вн,ед
…	…	…	…	…

Морфологический анализ. Процедурный подход

Пример флективных классов:

№	Слово (пример)	Падежные окончания											
		Един. число						Множ. число					
		Им.	Род.	Дат.	Вин.	Твор.	Пр.	Им.	Род.	Дат.	Вин.	Твор.	Пр.
01	Телефон	_	а	у	_	ом	е	ы	ов	ам	ы	ами	ах
02	Тираж	_	а	у	_	ом	е	и	ей	ам	и	ами	ах
03	Огонь	ь	я	ю	ь	ем	е	и	ей	ям	и	ями	ях
04	Перебой	й	я	ю	й	ем	е	и	ев	ям	и	ями	ях
05	Санаторий	й	я	ю	й	ем	и	и	ев	ям	и	ями	ях
...	...	...						...					

Морфологический анализ. Процедурный подход

Пример анализа словоформы «стол**а**»

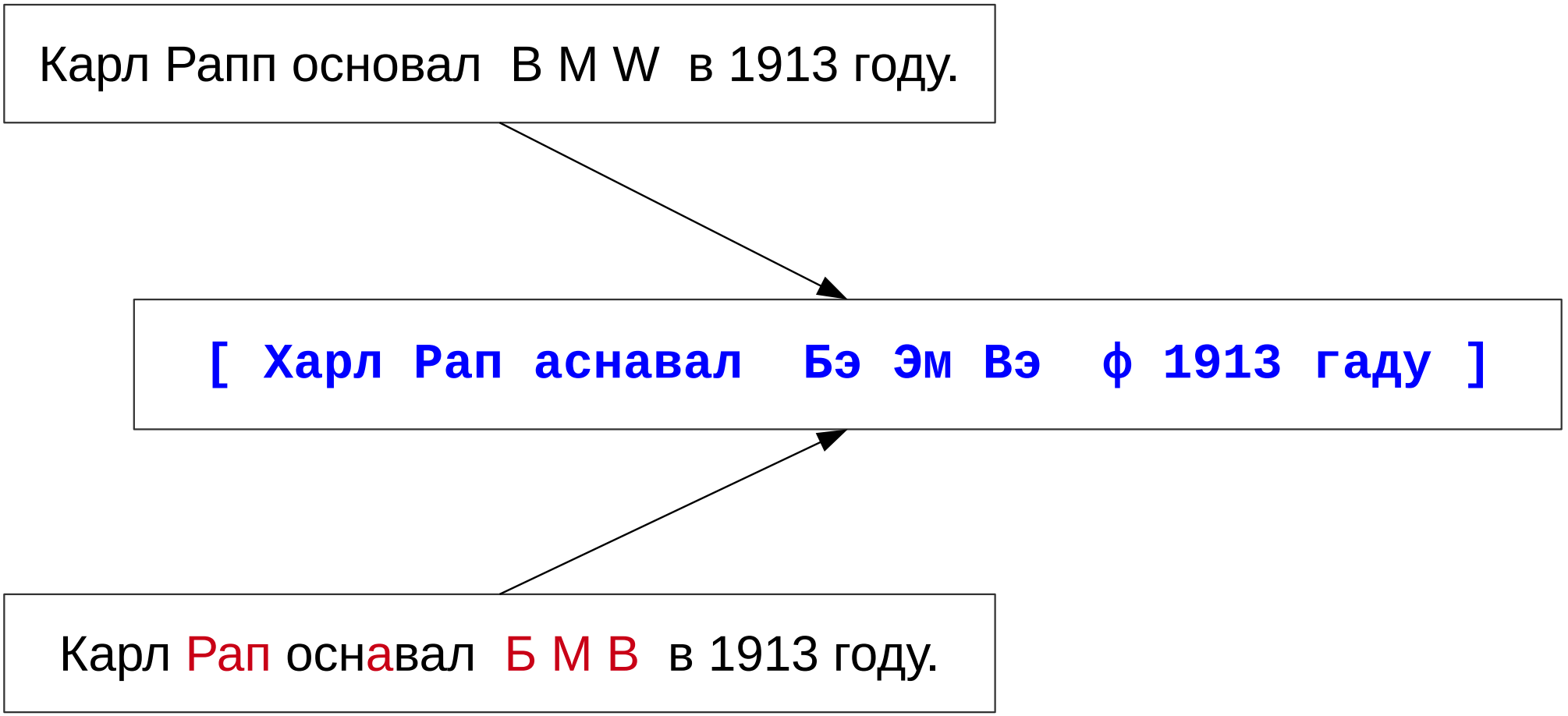
Основа	Номер флективного класса
…	…
СТОЛ	001
СТОЛБ	001
…	…

Номер флективного класса	Номер окончания	Номер грамматической информации
…	…	…
001	20	36
001	22	40
001	66	06
002	67	26
002	66	17
…	…	…

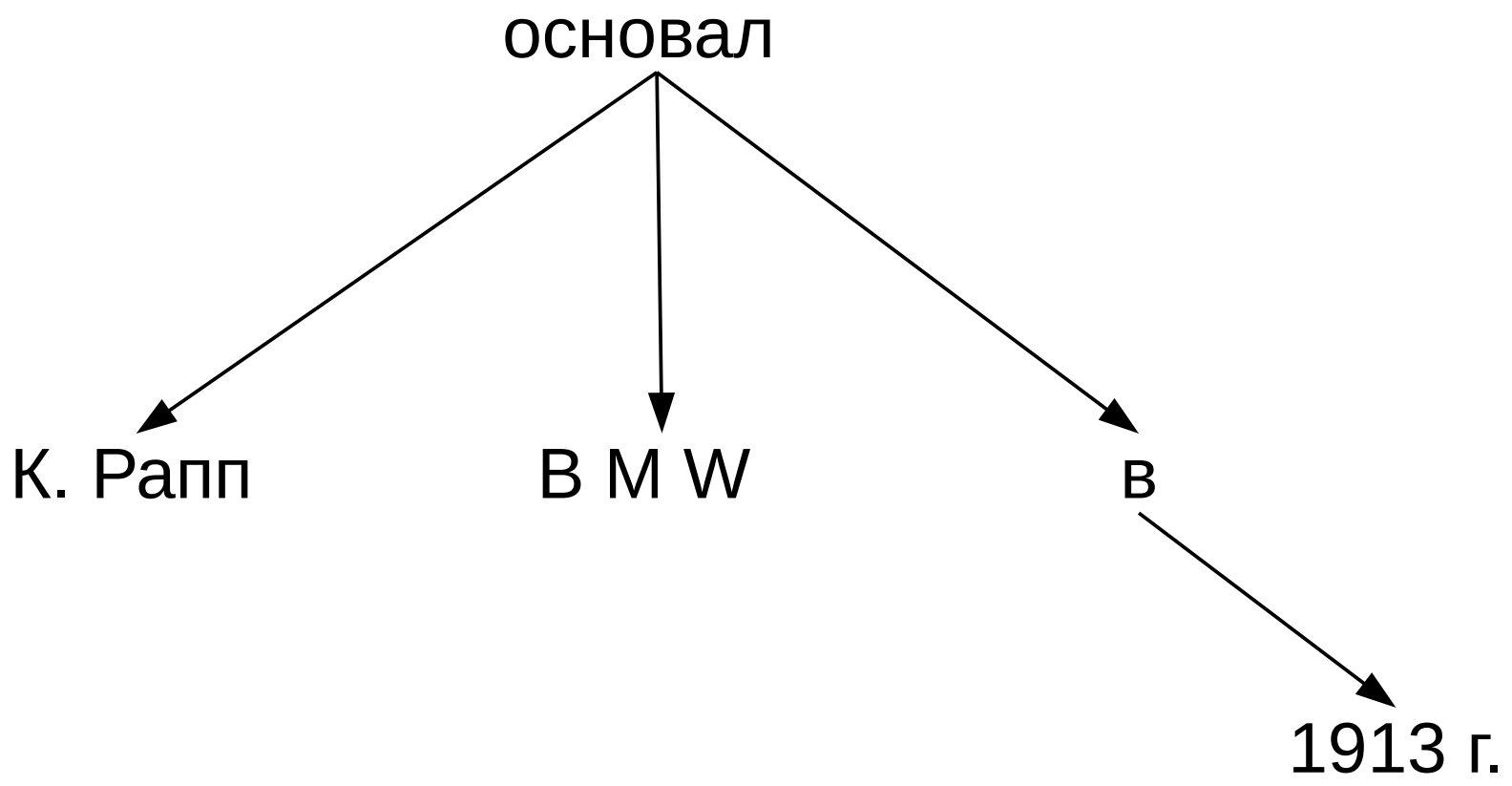
окончание	№
…	…
ят	62
ях	63
яя	64
_	65
а	66
е	67
и	70
…	…

№	Грамматическая информация
001	им. ед.
002	им. ед; вн. ед.
003	им. ед; вн. ед.; пр. ед.
004	им. ед; вн. ед.; рд. мн.
005	им. ед; рд. мн.; вн. мн.
006	рд. ед.
007	рд. ед., дт. ед.; тв. ед.; пр. ед.
…	…

Морфологический анализ. Фонетическое кодирование

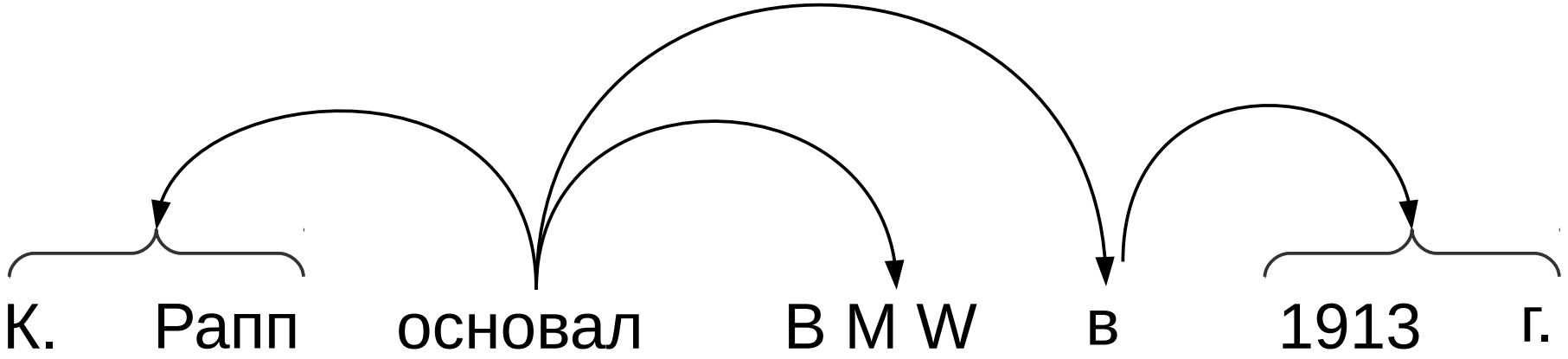


Синтаксический анализ. Деревья зависимостей

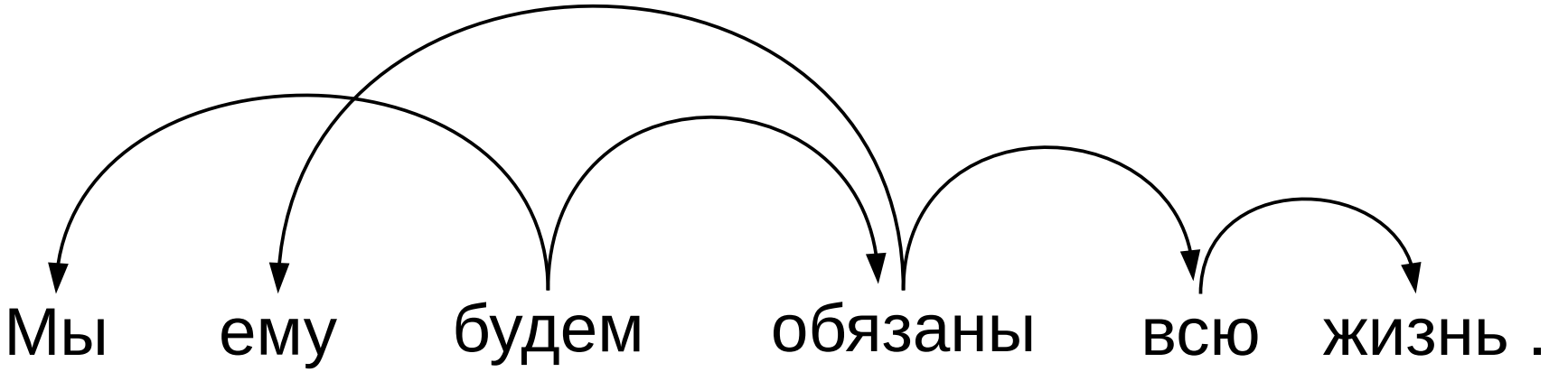


Синтаксический анализ. Деревья зависимостей

Проективное предложение:

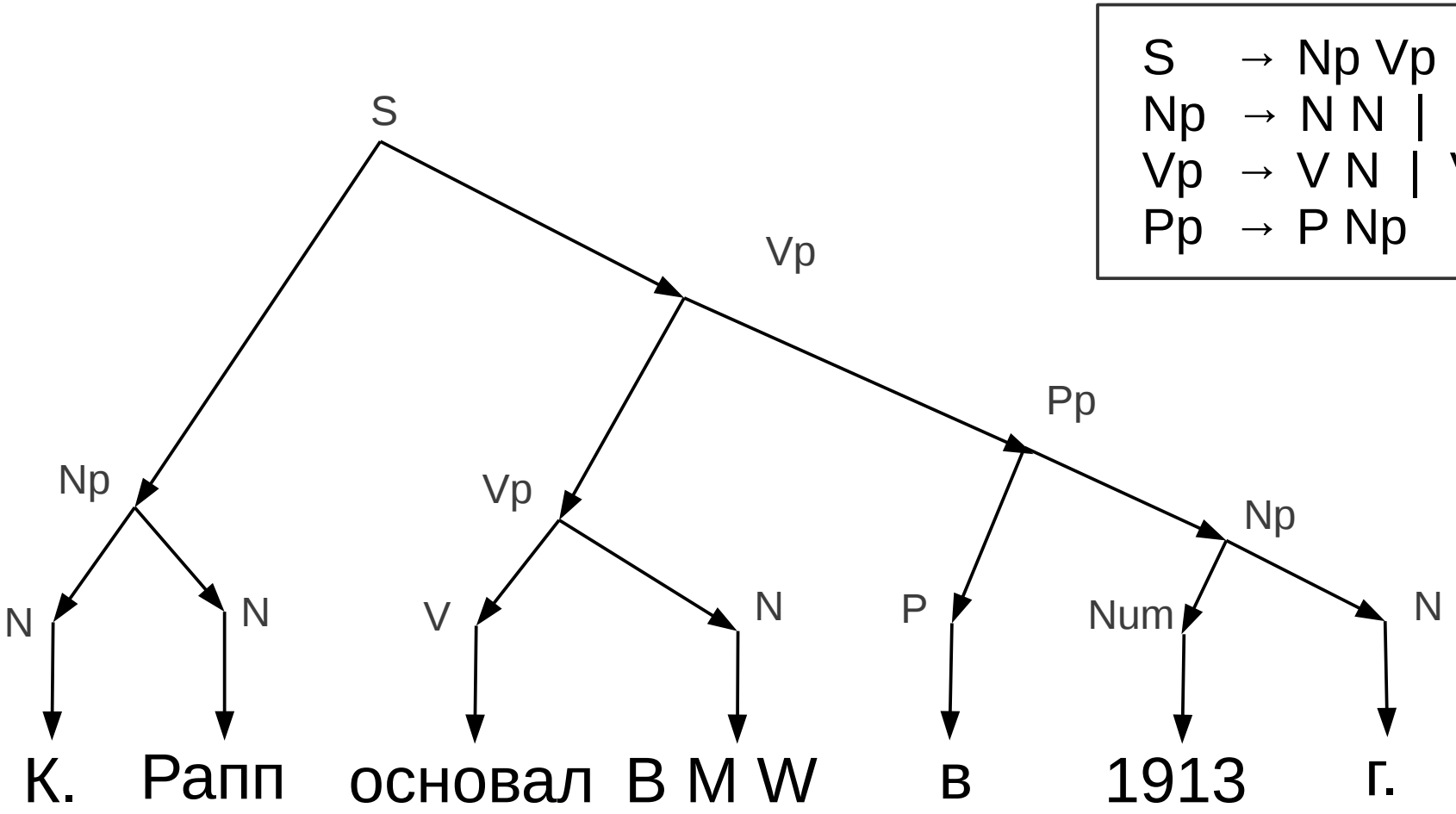


Непроективное предложение:



Синтаксический анализ. Системы составляющих

{ [К.Рапп] [(основал В М W) (в (1913 г.))] }



Семантический анализ

Для построения семант. графа используются:

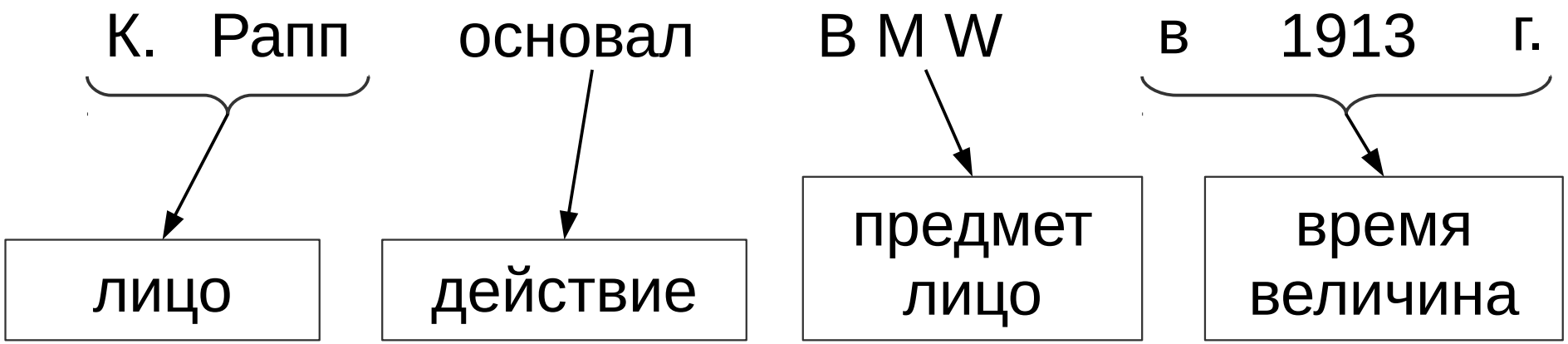
- семантическая категория (признак) — используется для толкования значения слова с помощью базовых понятий

- модель управления — описывает требования к другим словам, которые являются «параметрами» для данного слова

Семантический анализ

Семантические категории (признаки) слова:

- действие
- лицо
- предмет
- состояние
- свойство
- величина
- вещество
- информация
- время



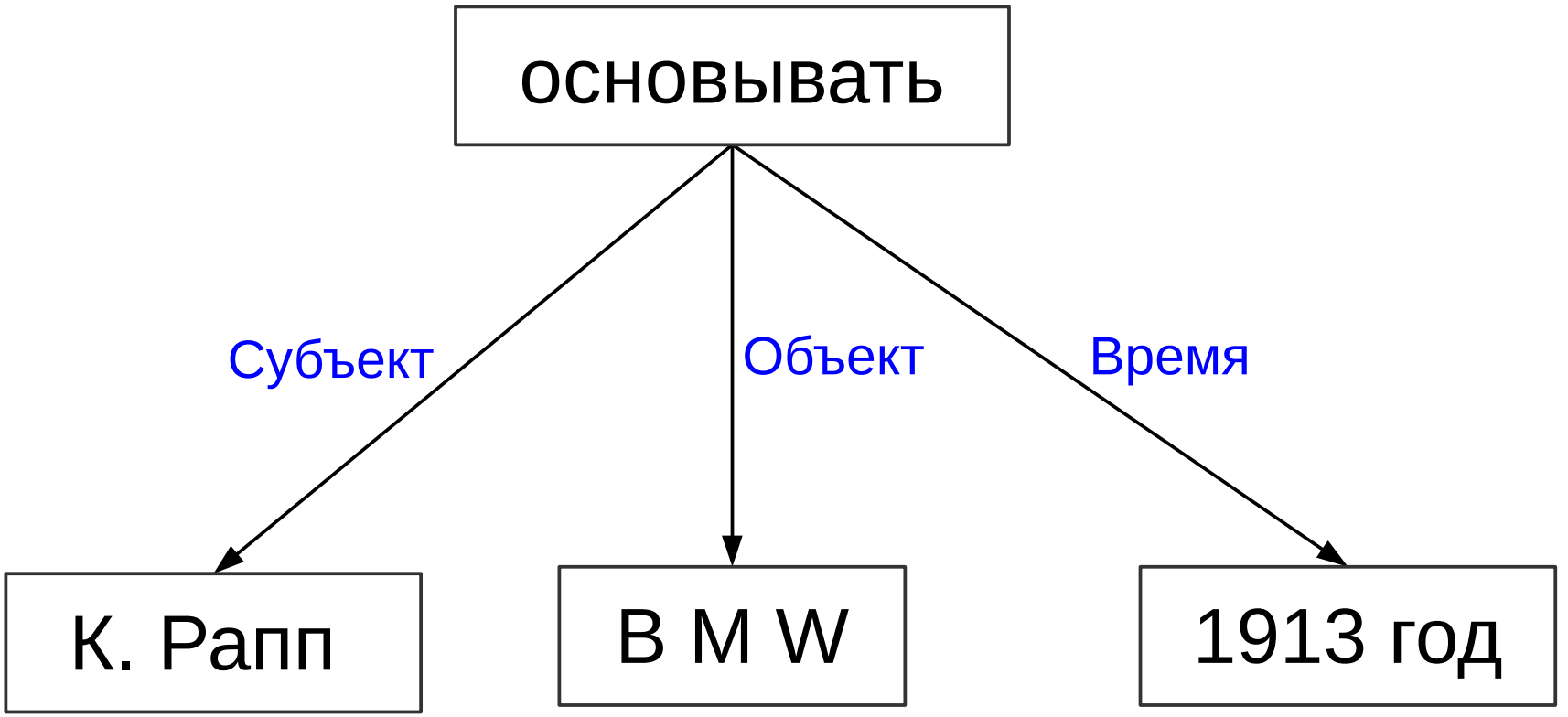
Семантический анализ. Модель управления

Глагол “приехать” :

Валентность	**Согласование**
Субъект	Им. п.
Место (откуда)	«из» + Род. п.
Место (куда)	«в» + Вин. п. «на» + Вин. п. «к» + Дат. п.
Время	Род. п. Вин. п.
Инструмент (на чем)	«на» + Предл. п. «в» + Предл. п.

Семантический анализ

Семантическая сеть:



Семантический анализ

Система фреймов:

ОСНОВЫВАТЬ	
Слот	Значение
Субъект	
Объект	
Время	
...	...

Человек	
Слот	Значение
Имя	К.Рапп
...	...

Организация	
Слот	Значение
Название	BMW
...	...

Дата	
Слот	Значение
Год	1913
...	...

Лингвистические ресурсы

1. грамматики
2. словари
    - электронные
    - компьютерные
3. тезаурусы
    - WordNet
4. корпуса текстов
    - с размеченной морфологией
    - с размеченным синтаксисом (tree-banks)

Грамматики.
Иерархия Хомского

0. Неограниченные

1. Контекстно-зависимые

2. Контекстно-свободные

3. Регулярные

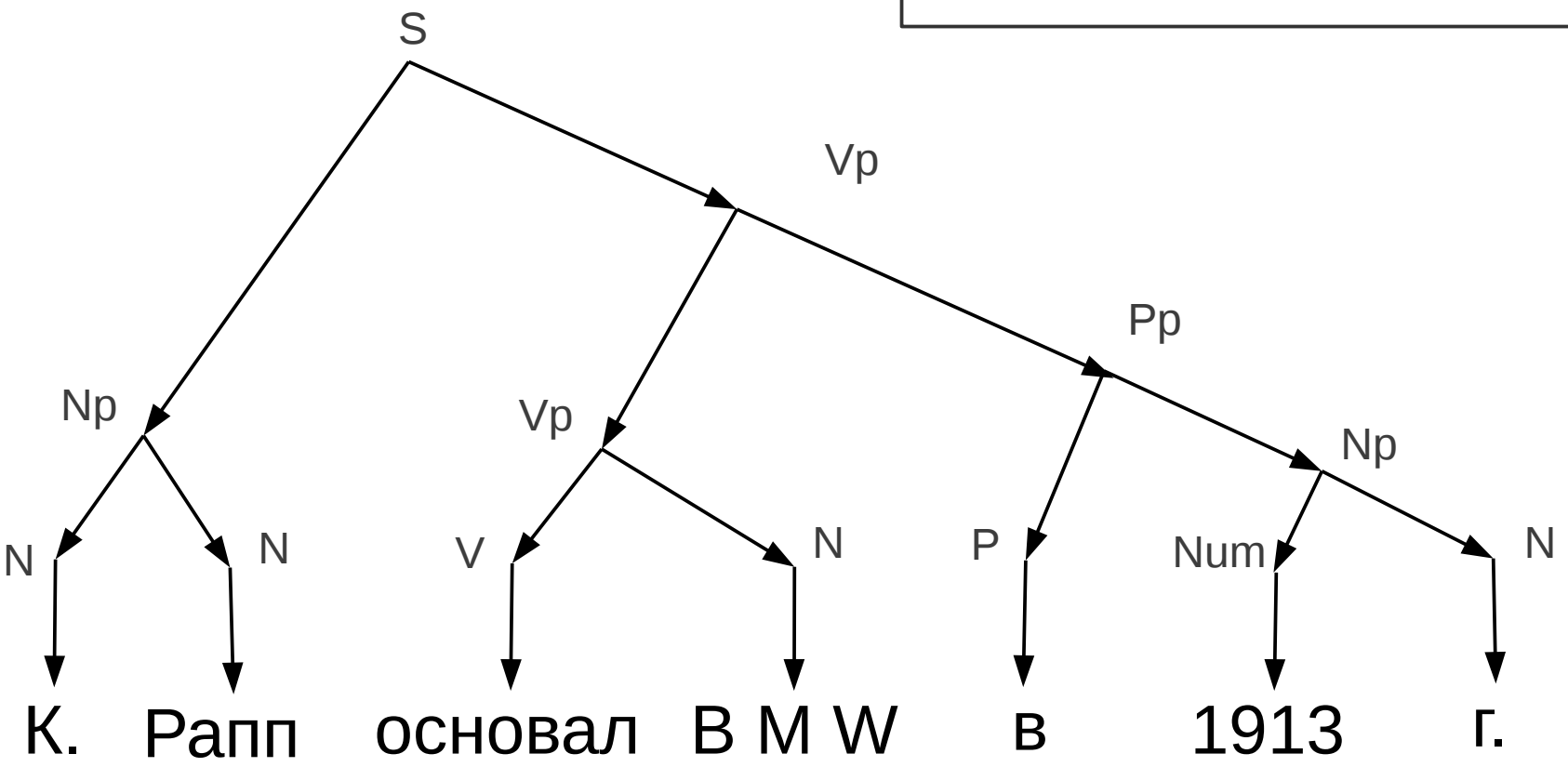
Грамматика отражает структуру языка

S → Np Vp | ...

Vp → Vp Np | Vp Pp | V | ...

Np → Np Np | Num Np | N | ...

Pp → P Np | ...



Компьютерные словари

Инверсионный порядок расположения слов:

-а по -мка

1.	а	1141.	надставка
2.	ба	1142.	представка
3.	аба	1143.	подставка
4.	кааба	1144.	приставка
5.	баба	1145.	доставка
6.	бой-баба	1146.	недоставка
7.	даба	1147.	сеноставка
8.	жаба	1148.	поставка
9.	раба	1149.	расставка
10.	полнеба	1150.	отставка

Тезаурусы. WordNet

intelligence ~ *noun* common

1. the ability to comprehend; to understand and profit from experience
2. a unit responsible for gathering and interpreting information about an enemy
3. secret information about an enemy (or potential enemy)

   we sent out planes to gather intelligence on their radar coverage
4. information about recent and important events

   they awaited news of the outcome
5. the operation of gathering information about an enemy

▼ Relatives

Synonyms	intelligence service
Antonyms	intelligence agency
Derivatives	intelligence information
Attributes	news
Similar	tidings
Kind of	word
	intelligence activity

Корпуса текстов.
Национальный корпус русского языка (НКРЯ)

НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА

что такое корпус?

статистика

частоты

морфология

обороты

синтаксис

семантика

параметры текстов

Частотное распределение популярных словоформ

Словоформы 2-граммы 3-граммы 4-граммы 5-граммы **6-граммы**

№	Словосочетание	Документы	Частота
1	во что бы то ни стало	1321	2435
2	как ни в чем не бывало	805	1360
3	ни с того ни с сего	837	1290
4	в одно и то же время	526	761
5	и т д и т п	326	488
6	...	330	398

Корпуса текстов. НКРЯ
Морфология

1. **С. Г. Керимов. Интеллектуальный поиск информации, основанный на онтологии // «Информационные технологии», 2004** [омонимия снята] **Все примеры (1)**

Язык для формулирования поискового запроса должен быть простым и удобным для пользователя, желательно близким к **естественному языку**. [С. Г. Керимов. Интеллектуальный поиск

естественному	
Лемма	естественный (см. в словарях)
Грамматика	прил, м, ед, дат, полн
Семантика основная	der:s, r:rel
Семантика дополнительная	der:s, r:rel
Доп. признаки	null

Корпуса текстов. НКРЯ

Тексты с размеченным синтаксисом:

Литература:

1. Васильев В.Г., Кривенко М.П.
   "Методы автоматизированной обработки текстов"

Ссылки:

1. http://aot.ru/ - Автоматическая обработка текста
2. http://ruscorpora.ru/ - Национальный корпус русского языка

Спасибо за внимание!