

NLP Linguistic Annotation



Сегментация

Деление на
предложение,
токенизация

Морфология

Лемматизация, POS,
грам. характеристики

Синтаксис

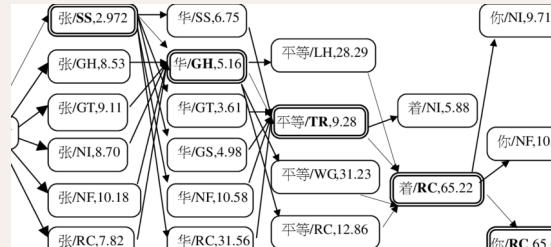
Зависимости,
составляющие

Семантика

Модели управления,
значения слов

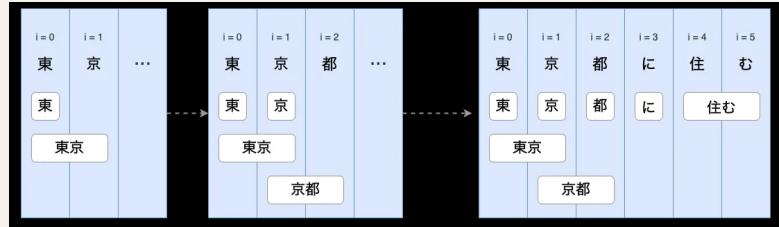
Токенизация

- Для языков, в которых есть четкое графическое деление на слова, токенизация правиловая (н-р, для русского есть библиотека razdel)
- Для языков без четкого деления на слова (с высоким индексом агглютинации, н-р, корейский, или с иероглифической письменностью) токенизаторы работают на словарях или нейронных сетях.
- Токенизаторы для языков типа японского (где нет явного деления на слова с помощью пробелов) используют lattice-based tokenization: lattice (решетка) – это графоподобная структура данных.
- В lattice содержатся все возможные токены в анализируемом предложении. Используется алгоритм Витерби для поиска наиболее вероятного варианта токенизации.



Токенизация

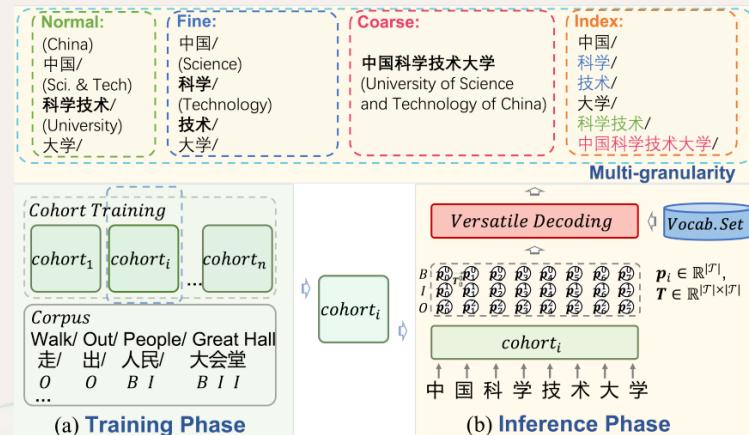
- Для японского языка существуют такие токенизаторы:
 - Kotori на Kotlin
 - Sudachi и Kuromoji на Java
 - Janome и SudachiPy на Python
 - Kagome на Go
- Для таких токенизаторов необходим словарь потенциальных токенов
- Такие специальные словари должны содержать следующие вещи:
 - Словоформа
 - ID левого/правого контекста (чтобы алгоритм Витерби мог вычислить наилучшую вероятность)
 - Собственно цена (алгоритм выбирает наименьшую цену). Чем выше это значение, тем реже используется этот вариант токена



Токенизация

Для китайского языка есть два варианта токенизации:

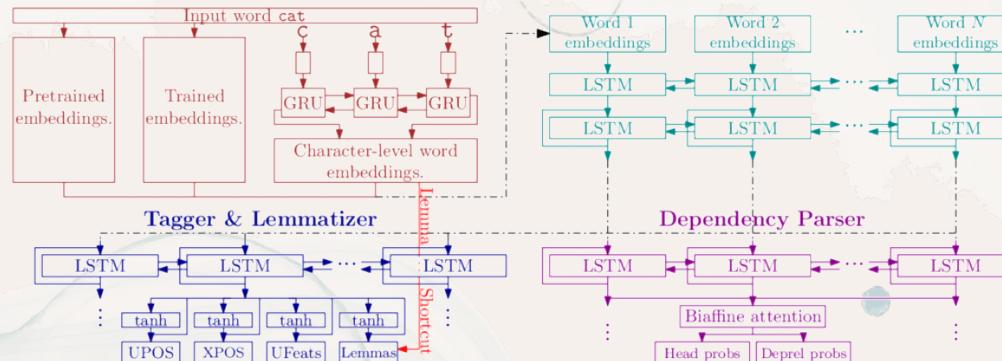
- N-граммы: оцениваются уни-, би- триграммы из иероглифов (работает плохо)
- На словарях (аналогично японскому)
- Библиотеки и методы:
 - ICU (на словарях)
 - Thulac (статистический алгоритм)
 - Jieba (словари с алгоритмом)
 - На нейронных сетях (CWSSeq)



Лемматизация

• Возможные алгоритмы:

- Искать по словарю парадигм (рус.яз.: Зализняк) – ресурсоемко
- Правиловые: пытаемся разделить слово на флексию и основу, ищем то и другое по словарям
- Классификация по типу парадигмы, ручные правила для преобразований – галлюцинации
- Топ-k

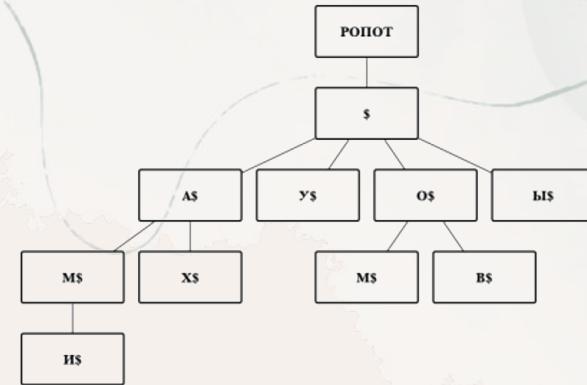


MyStem: идея алгоритма

- MyStem реализует алгоритм Сегаловича и позволяет эффективно находить для словоформы все возможные позиции в словоизменительных парадигмах лексем
- Идея алгоритма – поиск словоформы в двух префиксных деревьях (trie) – префиксном дереве инвертированных основ и префиксном дереве окончаний.
- Каждая лексема записана в виде $\text{inv\$A}$, где inv – инвертированная основа, а « A » – парадигма.
- Например, для слова «топор» будем иметь следующую форму записи – « $\text{ропот\$A}$ », где A – всевозможные окончания (« $ы$ », « $ом$ », « $ами$ » и т. п.)

Алгоритм Сегаловича

- 1 Начиная с правого конца слова найти все возможные разбиения на основу + окончание
- 2 Повторить следующие шаги, начиная с самого длинного возможного окончания (короткой основы)
- 3 Найти основу в префиксных деревьях для основ, проверить, есть ли форма с нужным окончанием. Если есть, разбор найден, если нет – перейти к следующему разбиению.



Морфопарсинг

- POS-tagging – определение части речи
- Построение морфологического анализа слова – POS + морфологические характеристики
- Необходим тагсет (tagset) – набор тегов

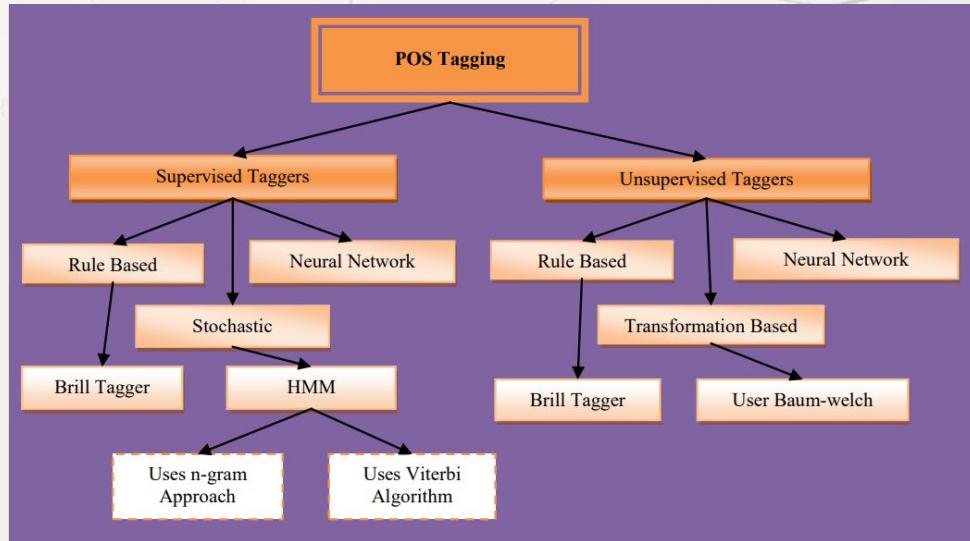
Проблемы:

- Что считать за части речи?..
- Омонимия, многозначность
- OOV-слова

Морфопарсинг

Подходы:

- Правиловый
- Статистический
- Нейросетевой
- Гибридные



Морфопарсинг

Статистические подходы:

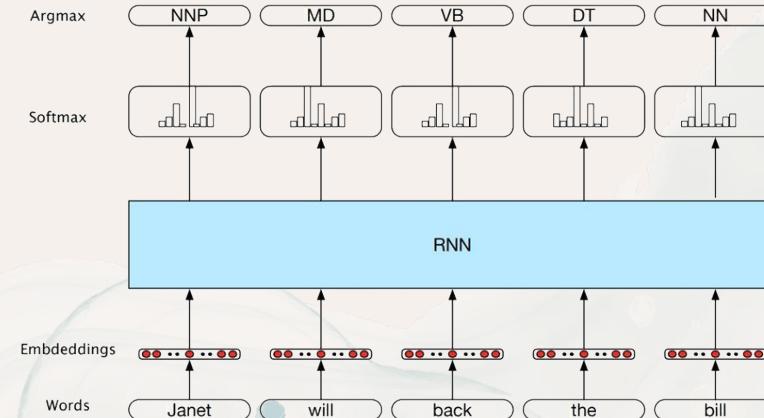
- Unigram tagging – выбор наилучшей гипотезы разбора без информации о контексте.
Выбираем самую частую/самую вероятную
- N-gram tagging – выбор наилучшей гипотезы разбора с использованием информации о n предыдущих токенах

[TreeTagger](#) (1995)
[TnT Tagger](#) (2000)

Морфопарсинг

Нейросетевой подход: задача Token Classification

- Лучше всего архитектуры, умеющие смотреть на контекст: RNN, Transformer
- SOTA: UDPipe (классический – на LSTM, есть варианты на Transformer)



Тагсеты

- Самый популярный – Universal Dependencies
- Для английского языка очень популярен Penn Treebank tagset
- Для русского языка есть тагсеты НКРЯ, OpenCorpora
- Для других языков есть свои тагсеты, которых может быть много
- Конвертация одного тагсета в другой – очень сложная задача

Sejong POS	UPOS
VA	ADJ (adjective)
MAG, MAJ	ADV (adverb)
IC	INTJ (interjection)
NNG, XR	NOUN (noun)
NNP, SL, SH	PROPN (proper noun)
VV	VERB (verb)
NNB, JKS, JKC, JKG, JKO, JKB, JKV, JKQ, JX	ADP (adposition)
VX, VCP, VCN, EP	AUX (auxiliary)
JC, EC	CCONJ (coordinate conjunction)
MM	DET (determiner)
NR, SN	NUM (numeral)
EF, ETN, ETM, XPN, XSN, XSA, XSV	PART (particle)
NP	PRON (pronoun)
EC	SCONJ (subordinate conjunction)
SF, SP, SE, SO, SS	PUNCT (punctuation)
SW	SYM (symbol)
NA, NF, NV	X (other)

Penn Treebank tagset

- Penn Treebank – один из самых известных морфосинтаксически размеченных корпусов английского языка (с 1989)
- 7 млн слов в морфологической разметке
- 3 млн слов в синтаксической разметке
- Свой формат разметки: 36 POS-тегов, 12 других для пунктуации и спецсимволов
- Список тегов с пояснениями [тут](#)

НКРЯ тагсет

- Тагсет Mystem совпадает с тагсетом НКРЯ
- Расшифровки можно посмотреть [тут](#)
- С 2023 года НКРЯ перешел на Universal Dependencies, но с доработками

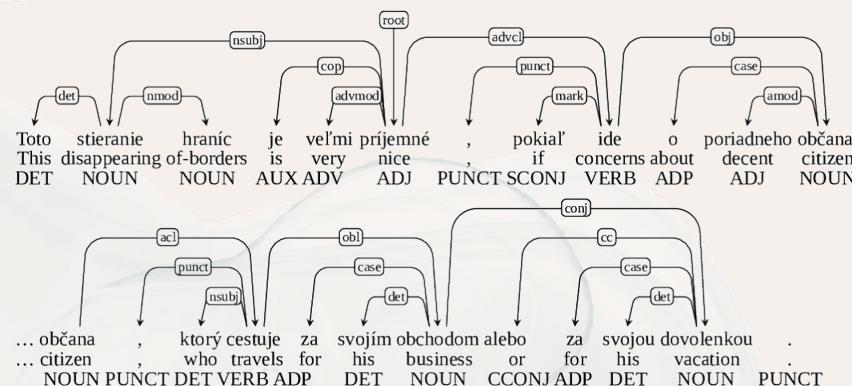
Грамматические признаки

Выбрать все Инвертировать выбор ×

<input type="checkbox"/> Часть речи <input type="checkbox"/> существительное <input type="checkbox"/> прилагательное <input type="checkbox"/> числительное <input type="checkbox"/> числ-прил <input type="checkbox"/> глагол <input type="checkbox"/> наречие <input type="checkbox"/> предикатив <input type="checkbox"/> вводное слово <input type="checkbox"/> мест-сущ <input type="checkbox"/> мест-прил <input type="checkbox"/> мест-предиктив <input type="checkbox"/> местоименное наречие <input type="checkbox"/> предлог <input type="checkbox"/> союз <input type="checkbox"/> частица <input type="checkbox"/> междометие	<input type="checkbox"/> Падеж <input type="checkbox"/> именительный <input type="checkbox"/> звательный <input type="checkbox"/> родительный <input type="checkbox"/> родительный 2 <input type="checkbox"/> дательный <input type="checkbox"/> винительный <input type="checkbox"/> винительный 2 <input type="checkbox"/> творительный <input type="checkbox"/> предложный <input type="checkbox"/> предложний 2 <input type="checkbox"/> счётная форма	<input type="checkbox"/> Наклонение / Форма <input type="checkbox"/> изъявительное <input type="checkbox"/> повелительное <input type="checkbox"/> повелительное 2 <input type="checkbox"/> инфинитив <input type="checkbox"/> причастие <input type="checkbox"/> деепричастие	<input type="checkbox"/> Степень / Краткость <input type="checkbox"/> сравнительная <input type="checkbox"/> сравнительная 2 <input type="checkbox"/> превосходная <input type="checkbox"/> полная форма <input type="checkbox"/> краткая форма
<input type="checkbox"/> Имена собственные <input type="checkbox"/> фамилия <input type="checkbox"/> имя <input type="checkbox"/> отчество	<input type="checkbox"/> Число <input type="checkbox"/> единственное <input type="checkbox"/> множественное	<input type="checkbox"/> Время <input type="checkbox"/> настояще <input type="checkbox"/> будущее <input type="checkbox"/> прошедшее	<input type="checkbox"/> Переходность <input type="checkbox"/> переходный <input type="checkbox"/> непереходный
	<input type="checkbox"/> Род <input type="checkbox"/> мужской <input type="checkbox"/> женский <input type="checkbox"/> средний <input type="checkbox"/> общий*	<input type="checkbox"/> Лицо <input type="checkbox"/> 1-е лицо <input type="checkbox"/> 2-е лицо <input type="checkbox"/> 3-е лицо	<input type="checkbox"/> Прочее <input type="checkbox"/> цифровая запись <input type="checkbox"/> аномальная форма <input type="checkbox"/> нестандартная запись <input type="checkbox"/> инициал <input type="checkbox"/> сокращение <input type="checkbox"/> несклоняемое <input type="checkbox"/> топоним**
	<input type="checkbox"/> Одушевленность <input type="checkbox"/> одушевленное <input type="checkbox"/> неодушевленное	<input type="checkbox"/> Вид <input type="checkbox"/> совершенный <input type="checkbox"/> несовершенный	

Деревья зависимости

- Один из самых старых и больших датасетов, размеченных синтаксически
 - Penn Treebank
- Парсить синтаксис пытались давно: старая диссертация аж 1993 года
- Сегодня это в основном Universal Dependencies
- Деревья зависимости удобнее использовать для практических задач, поэтому их парсят чаще, чем деревья составляющих



Universal Dependencies

- Основывается на Stanford Dependencies
- Первоначально SD были строго синтаксическими
- Universal Dependencies – попытка сделать их более универсальными, повлекшая к перекосу в сторону семантики
- Tagset тоже стремится к универсальности
- Иногда из-за этого плохо учитываются индивидуальные особенности языков, сильно отличающихся от английского
- Для русского UD разрабатывают О. Ляшевская, Т. Шаврина

Open class words	Closed class words	Other
<u>ADJ</u>	<u>ADP</u>	<u>PUNCT</u>
<u>ADV</u>	<u>AUX</u>	<u>SYM</u>
<u>INTJ</u>	<u>CCONJ</u>	<u>X</u>
<u>NOUN</u>	<u>DET</u>	
<u>PROPN</u>	<u>NUM</u>	
<u>VERB</u>	<u>PART</u>	
	<u>PRON</u>	
	<u>SCONJ</u>	

Universal Dependencies

Главные принципы:

- Вершинами могут быть только смысловые слова
- Различают номинал, клаузу и модификатор
- Опираются на слова (принцип lexical integrity – Chomsky 1970)

Размечаются:

- Леммы
- Универсальные теги (тагсет UD)
- Могут быть XPOS – другой тагсет
- Грамматические характеристики
- Вершины деревьев зависимостей
- Типы синтаксических связей

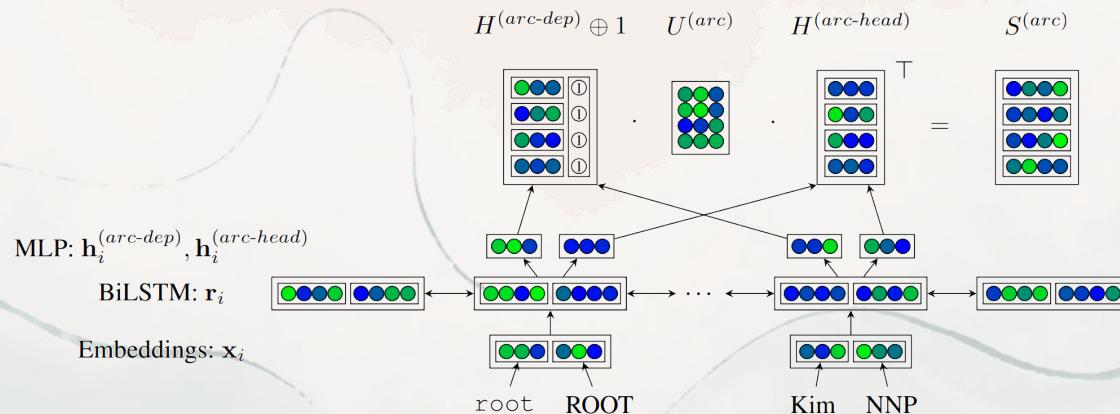
Universal Dependencies

- Использует формат CONLL-U (Plus)
- 10 колонок: index, form, lemma, UPOS, XPOS, feats, head, deprel, deps, misc
- В формате CONLL-U Plus можно добавлять свои колонки

```
# sent_id = 2
# text = एकुलता एक मुलगा महणून राजा-राणी तयाला जीव की पूरण करित.
# translit = ekulata eka mulagā mhaṇūna rājā-rāṇī tyālā jīva kī prāṇa karita.
1 एकुलता एकुलता ADJ _ Case=Nom|Gender=Masc|Number=Sing 3 amod _ Translit=ekulatā|LTranslit=ekulatā
2 एक एक DET _ Number=Sing|PronType=Ind 3 det _ Translit=eka|LTranslit=eka
3 मुलगा मुलगा NOUN _ Case=Nom|Gender=Masc|Number=Sing 0 root _ Translit=mulagā|LTranslit=mulagā
4 महणून महणून CCONJ _ 3 cc _ Translit=mhaṇūna|LTranslit=mhaṇūna
5 राजा राजा NOUN _ Case=Nom|Gender=Neut|Number=Sing 12 nsubj _ SpaceAfter=No|Translit=rājā|LTranslit=rājā
6 - - PUNCT _ 7 punct _ SpaceAfter=No|Translit=-|LTranslit=-
7 राणी राणी NOUN _ Case=Nom|Gender=Fem|Number=Sing 5 conj _ Translit=rāṇī|LTranslit=rāṇī
8 तयाला तो PRON _ Case=Dat|Deixis=Remt|Gender=Masc|Number=Sing|Person=3|PronType=Dem 12 obj _ Translit=tyālā|LTranslit=to
9 जीव जीव NOUN _ Case=Nom|Gender=Masc|Number=Sing 12 compound:lvc _ Translit=jīva|LTranslit=jīva
10 की की ADV _ 9 fixed _ Translit=ki|LTranslit=ki
11 पूरणपूरण NOUN _ Case=Nom|Gender=Masc|Number=Sing 9 fixed _ Translit=prāṇa|LTranslit=prāṇa
12 करित करित VERB _ Aspect=Hab|Number=Plur|Person=3|Tense=Past|VerbForm=Fin 3 conj _ SpaceAfter=No|Translit=karita|LTranslit=karita
13 . . PUNCT _ 12 punct _ Translit=-|LTranslit=-
```

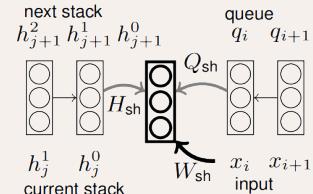
Universal Dependencies

- Парсер в формате: UDPipe (регулярно обновляется, много языков)
- Метрики: F1-score, UAS, LAS
- Для парсинга зависимостей использует биаффинную архитектуру Дозата-Маннинга

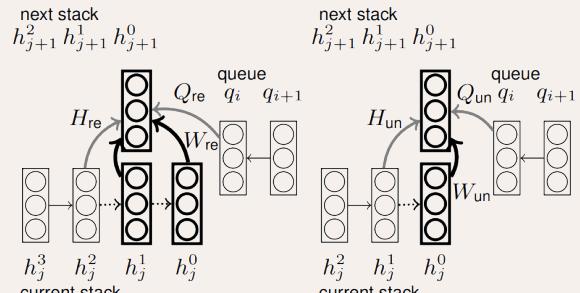


Деревья составляющих

- Модели: на спанах и на переходах (chart-based vs transition-based)
- Chart-based (span-based): мы максимизируем вероятность дерева для предложения, а дерево представляем как набор спанов (spans). Используется Conditional Random Fields (CRF)
- Transition-based: инкрементально строим дерево, представляя весь процесс как систему с состояниями и действиями. Действия: shift (строим терминал), unary (одиночный узел), binary (узел с двумя ветвями).



(a) shift- X action

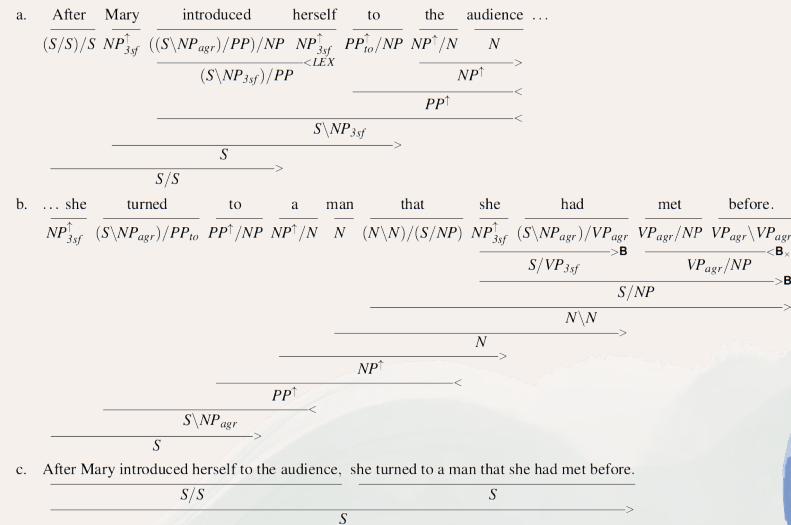


(b) reduce- X action

(c) unary- X action

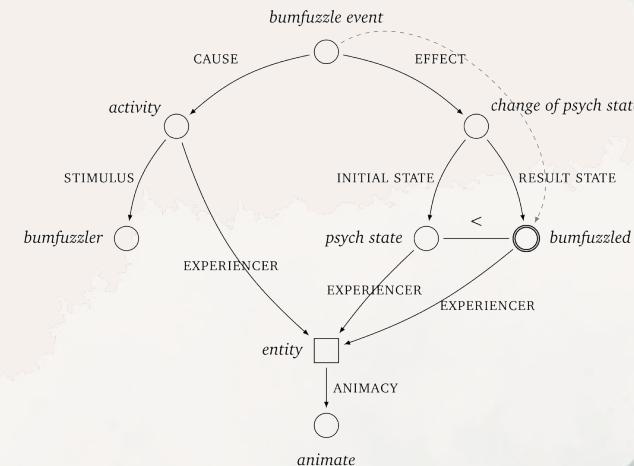
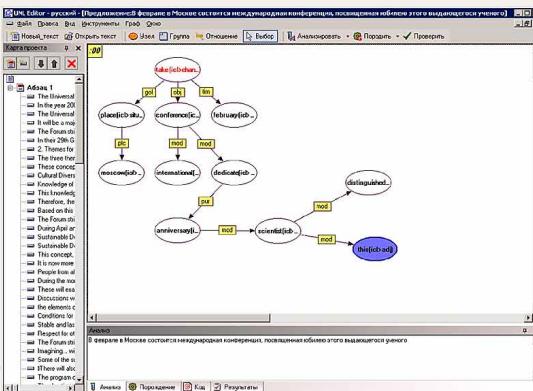
Combinatory Categorial Grammar

- CCG – лексикализованная теория грамматики, в которой вся языковая информация, в том числе по синтаксису, определяется в словаре (лексиконе). Steedman 2000
- Цель – создать такую объяснительную теорию, которую можно было бы использовать для практических задач в NLP
- Если вкратце – еще один формализм
- Для него тоже создаются парсеры



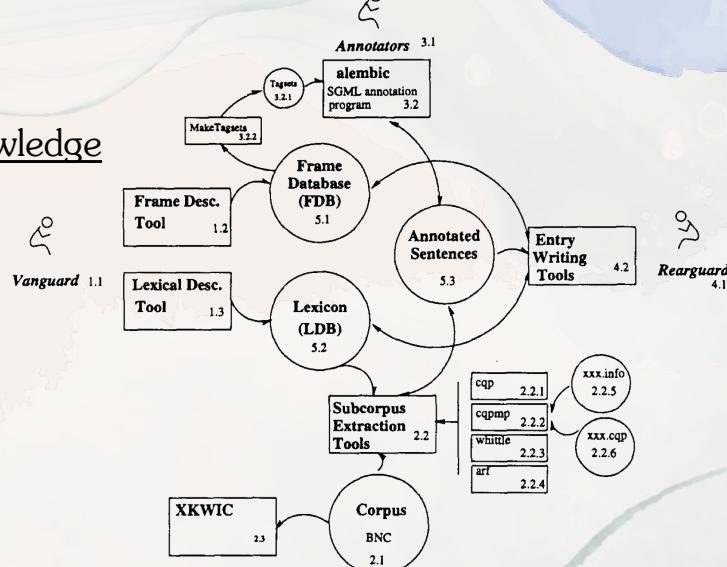
Семантическая разметка

- Теоретические попытки описывать семантику: Филлмор
 - Практические попытки создать interlingua для машинного перевода
 - Два подхода: с использованием онтологий и без



Фреймы

- Ч. Филлмор: Frame Semantics
- М. Минский: A Framework for Representing Knowledge



FrameNet

Linguistic_meaning

Definition:

A linguistic **Form** has a particular **Meaning**, possibly restricted to a particular **Textual_location**. Some linguistic **Forms**, nouns and nominal expressions, also can refer to an object in the real or an imagined world, a **Referent**: "Unicorn" DENOTES a mythological horse with a horn on its head.

In art, the word "oeuvre" MEANS "the collected works of an artist."

FEs:

Core:

Form [**Form**]

Form is the formal pole of a linguistic sign.

Meaning [**Meaning**]

Meaning is the sense of a linguistic **Form**.

Referent [**Referent**]

Referent identifies an object in the real (or a hypothetical) world that a linguistic expression is used to talk about.

Excludes: Meaning

Non-Core:

Degree []

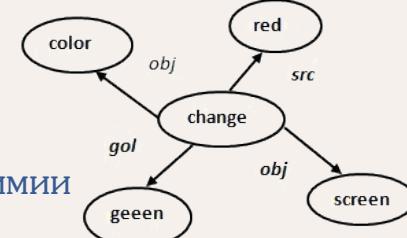
This frame element selects some gradable attribute of the **Form** and modifies the expected value for it.

Textual_location [**Tex**]

The FE **Textual_location** is used for expressions that indicate under which semantic, collocational, pragmatic or other circumstances a **Form** has a particular **Meaning**.

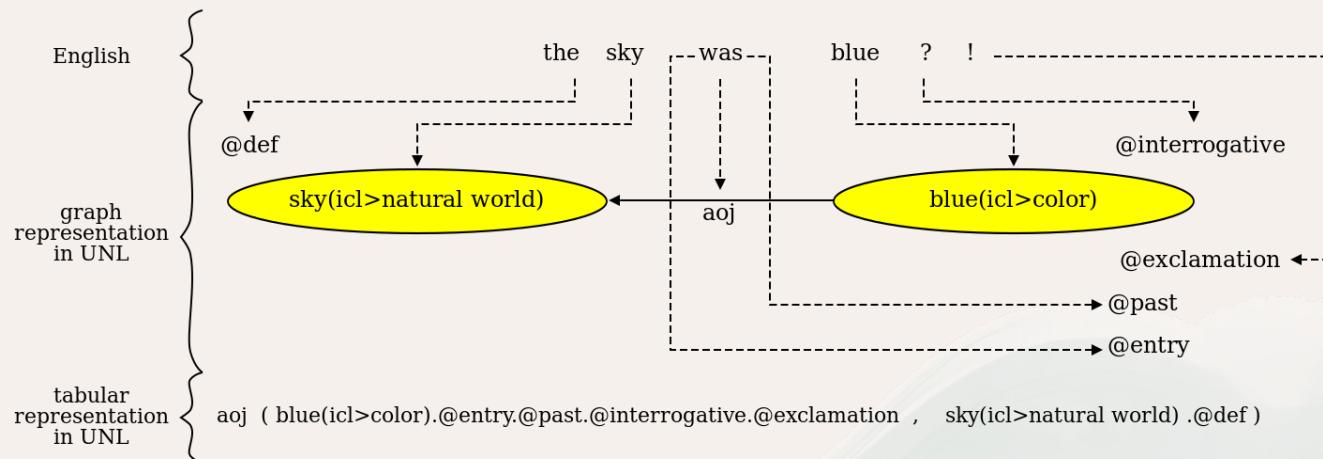
Universal Networking Language

- проект начался в 1994 году
- основная цель – преодолеть языковой барьер, создать универсальный семантический язык
- семантические графы, в узлах содержатся лексические единицы (универсальные слова) и характеристики, дуги несут семантические отношения
- универсальные слова организованы в базу знаний (knowledge base), которая собирается вручную; они объединены в основном через связи гипо- и гиперонимии, синонимии, меронимии
- абстрагируется от синтаксиса
- опирается на WordNet



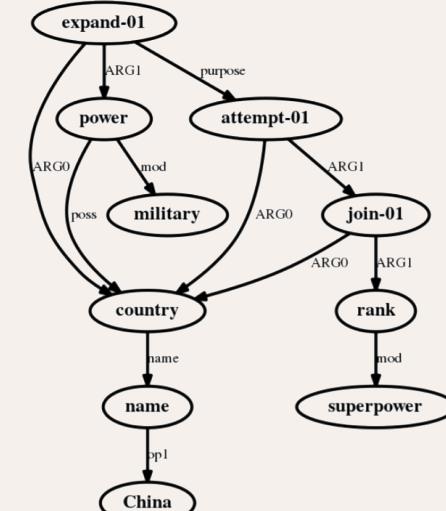
Universal Networking Language

- Парсер был имплементирован в Этапе-3
- Иногда используется в качестве вспомогательного инструмента
- Парсеров на нейронных сетях для него нет
- склонен в сторону английского языка – по сути, это дизамбигуированный английский



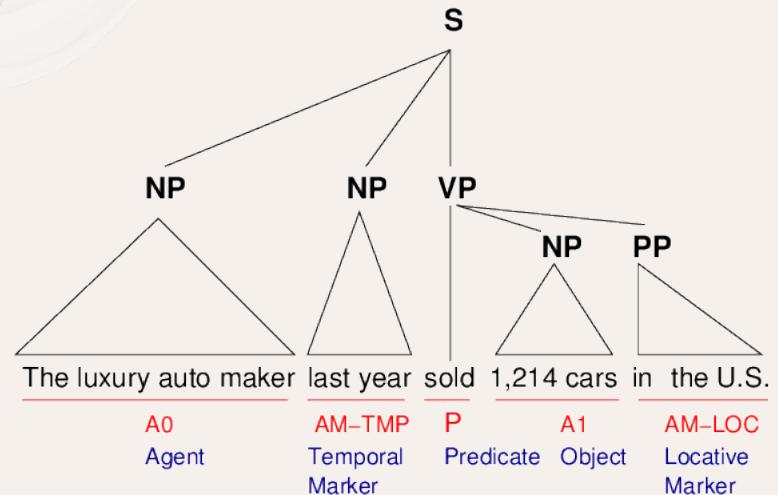
Abstract Meaning Representations

- AMR – 2013 (впервые придумали в 1998, но основная статья – 2013 года)
- язык семантического представления, предложения - графы со следующими свойствами:
 - У каждого графа есть вершина (root);
 - Все узлы графа обозначены ярлыками (labeled);
 - Графы – направленные (directed);
 - И незамкнутые (acyclic)
- В графах AMR узлы маркируются концептами, а ребра – отношениями
- Опирается на PropBank



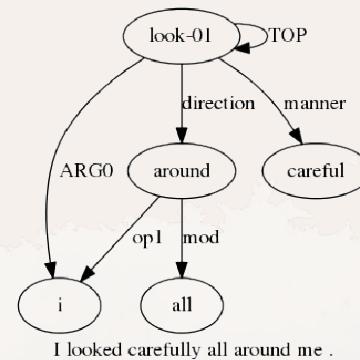
PropBank

- Это корпус с размеченными семантическими пропозициями:
- Синтаксические узлы размечены семантическими ярлыками, которые соответствуют ролям
- Тесно связано с понятием фреймов



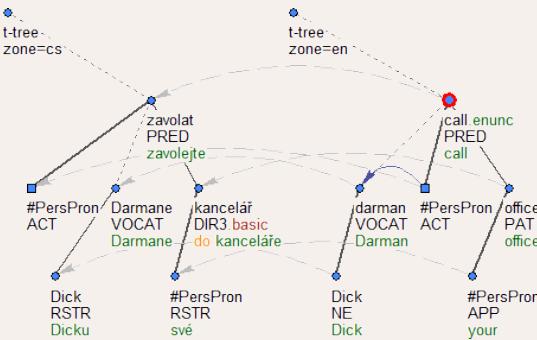
Abstract Meaning Representations

- абстрагируется от синтаксиса
- тоже есть смещение в сторону английского языка: они сами пишут, что «heavily biased»
- есть корпуса: английский (разметили «Маленького принца»), китайский
- размечаются сем. роли, кореференция, именованные сущности
- для него существует несколько парсеров
- парсеры работают примерно на тех же принципах, что и парсеры для деревьев составляющих



Prague Dependencies

- Prague Dependency Treebank (PDT) - первая версия в 2001 году, PDT-C (Consolidated) в 2006 году, теперь существует версия 2.0
- Чешский, английский, арабский
- В него конвертировали Синтагрус
- Размечает морфологию, синтаксис, семантику, кореферентные связи, актуальное членение предложения, восстанавливает эллипсис
- В основе теория Functional Generative Description
- База знаний у них собственная: Valency Lexicon (PDT-VALLEX)



Prague Dependencies

Верхний уровень представления языка в FGD называется тектограмматическим уровнем, и предполагается, что он представляет семантическую структуру предложения. Принципы, на которых строится тектограмматический уровень, таковы:

- базовый юнит - предложение
- для каждого хорошо сформированного (чешского) предложения возможно дать тектограмматическое представление
- в случае неоднозначности, в теории возможно приписывать более одного тектограмматического дерева одному предложению, однако в PDT приписывается только одно
- в случае синонимии разным предложениям может соответствовать одно тектограмматическое дерево

Prague Dependencies

Модель состоит из нескольких слоев:

- w-layer – токенизированный текст, в котором каждому слову присвоен ID.
- m-layer – морфологический слой, где токенам присваивается часть речи и проводится лемматизация
- a-layer – аналитический слой, который представляет синтаксическую структуру предложения.

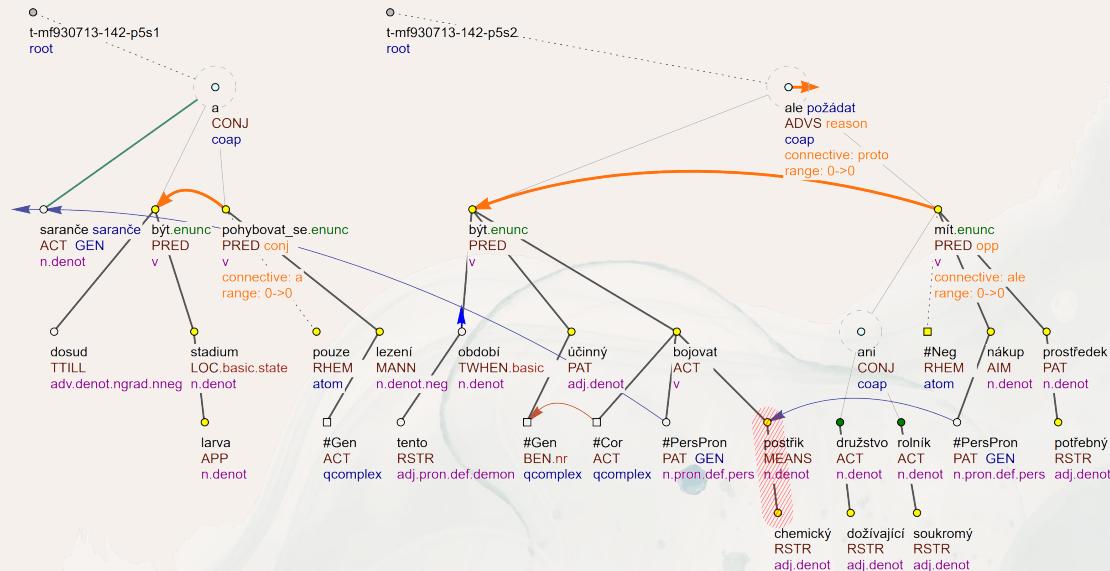
Токены получают тэги аналитических функций (субъект, объект, предикат и т.д)

- t-layer – тектограмматический слой, который представляет синтаксис и семантику в форме семантической разметки, разрешения анафоры и описания структуры аргументов, основанной на лексиконе валентности.

Prague Dependencies

Если такая крутая модель, почему она не так популярна, как могла бы быть?

- Есть проблемы с представлением непроективных предложений, хотя это зависимости
 - Очень. Сложная. Разметка. Оцените мануал

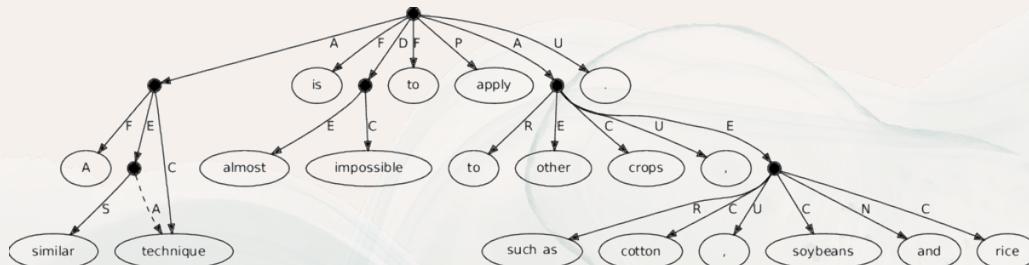


UCCA

- Universal Conceptual Cognitive Annotation – 2013
- Одна из важных целей создателей – простота формата. Могут размечать толокеры
- Содержит несколько слоев разметки
- Основной слой модели различает глагольные и номинативные предикаты и их аргументы, адъюнкты, прилагательные, копулы и отношения между клаузами
- Не использует онтологию: разметчики размечают тексты вручную, потом обучается нейронный парсер

UCCA

- Модель считывает слова и многосложные фразы как терминалы (terminals) - атомарные содержательные единицы
- Формальная базовая единица UCCA - это юнит (unit)
- На базовом слое размечаются элементы сцены (сцена – это некое событие), элементы, которые не создают сцену (~сирконстанты и прочие реляторы), отношения между сценами и некоторые другие

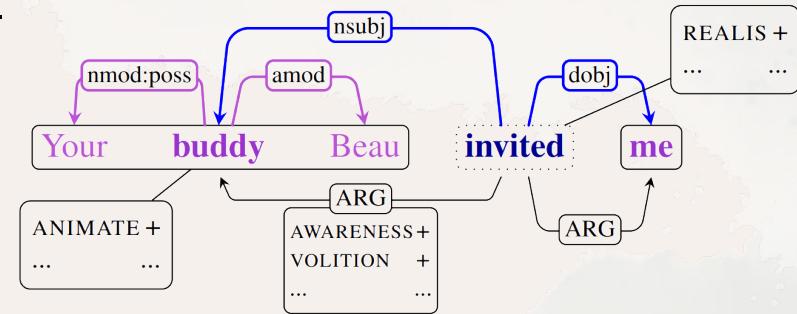


UCCA

- Сцена описывает некоторое действие, движение или состояние, и обычно имеет временной и пространственный локус
 - Каждая Сцена имеет одно главное отношение - Process или State, один или несколько Participants, которые могут быть конкретными и абстрактными
 - Вложенные сцены считаются Participants
 - Второстепенные отношения помечаются как Adverbials
-
- У них очень понятный мануал
 - Есть парсеры под это дело, конечно
 - Есть английский, французский, немецкий корпуса

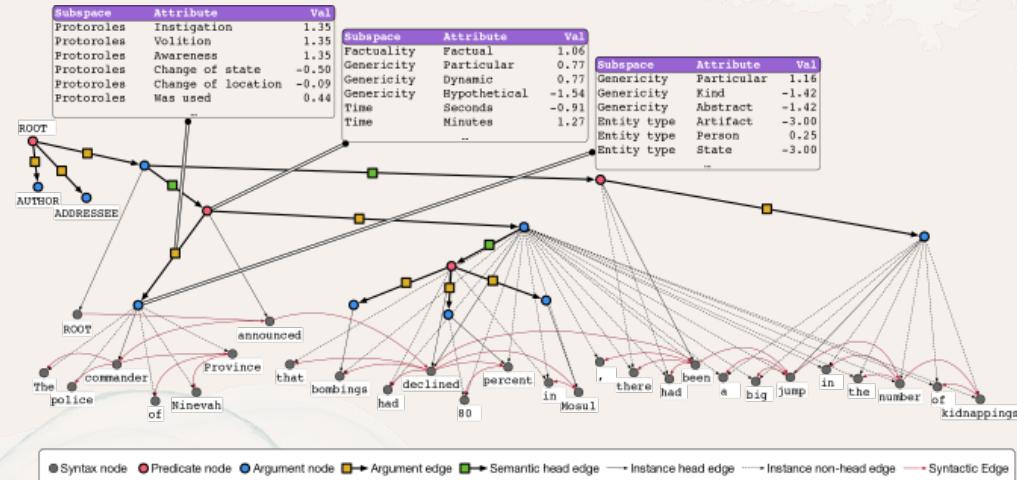
Decomp

- Universal Decompositional Semantics (Decomp)
- Сравнительно новые: статья вышла в 2016
- Их сайт
- Основная идея UDS - что значения слов декомпозициональны, т.е. в центре всего семантическое разложение
- У всех слов есть наборы характеристик
- Базируется на Semantic Proto-Role Labeling



Decomp

- Первоначальная разметка – ручная (краудфандинг)
- Потом обучаю парсер
- Онтологии тоже нет



CoBaLD

- Первая версия только в 2023 году
 - Трехуровневый формат: размечается морфология, синтаксис и семантика
 - Скрестили UD с Compreno
 - Активно разрабатывается: изменили внешний вид формата, хотим добавить синтаксические роли Compreno (т.к. это чистый синтаксис)
 - Морфология от UD, синтаксис (в версии 2.0) – Enhanced UD

```
# sent_id = 6
# text = На российско - грузинскую границу перебрасываются подразделения Северо - Кавказского военного округа.
1 НА НА ADP _ _ 3 case _ PREPOSITION
2 российско-грузинскую российско-грузинский ADJ _ Case=Acc|Degree=Pos|Gender=Fem|Number=Sing 3 amod Locative COUNTRY_AS_ADMINISTRATIVE_UNIT
3 границу граница NOUN _ Animacy=Inan|Case=Acc|Gender=Fem|Number=Sing 4 obl Locative_FinalPoint LINES
4 перебрасываются перебрасывать VERB _ Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin|Voice=Mid 0 root Predicate TO_SEND_TO_DELIVER
5 подразделения подразделение NOUN _ Animacy=Inan|Case=Nom|Gender=Neut|Number=Plur 4 nsubj Object MILITARY_FORCES_AS_ORGANIZATION
6 Северо-Кавказского Северо-кавказский ADJ _ Case=Gen|Degree=Pos|Gender=Masc|Number=Sing 8 amod Locative THE_EARTH_AND_ITS_SPATIAL_PARTS
7 военного военный ADJ _ Case=Gen|Degree=Pos|Gender=Masc|Number=Sing 8 amod Sphere CONFLICT_INTERACTION
8 округа округ NOUN _ Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing 5 nmod Locative ADMINISTRATIVE_REGION
9 . . PUNCT _ _ 4 punct _ _
```

CoBaLD

- Восстанавливает эллипсис
- Размечается кореференция (пока только для относительных местоимений)
- Корпуса: английский (2.0), русский (пока только 1.0)
- Активно ведутся исследования в области языкового переноса: турецкий, чешский, корейский языки

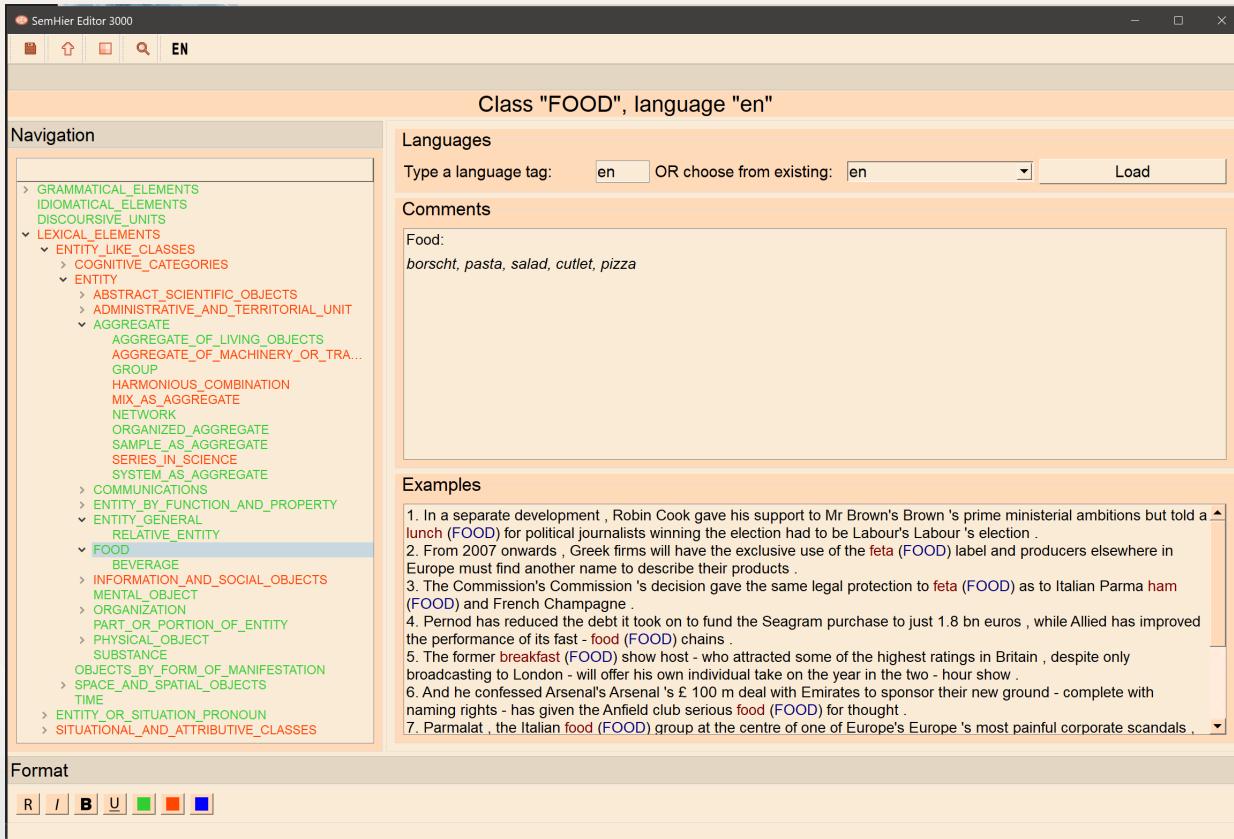
```
# sent_id = 13
# text = Well doesn't that include rock
1 Well well INTJ Interjection _ 5 discourse 5:discourse _ Parenthetical DISCOURSIVE_UNITS
2-3 doesn't _ _ AUX Verb _ Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin 5 aux 5:aux _ _ AUXILIARY_VERBS
3 n't not PART _ Polarity=Neg 2 advmod 2:advmod
4 that that PRON Pronoun Number=Sing|Person=1 5 nsubj 4.1:det _ Ch_Reference CH_REFERENCE_AND_QUANTIFICATION
4.1 #NULL #NULL NOUN Noun Number=Sing _ _ 5:nsubj ellipsis Possessor_Locative ENTITY_OR_SITUATION_PRONOUN
5 include include VERB Verb Mood=Ind|Number=Plur|VerbForm=Fin 0 root 0:root _ Predicate CONTAIN_INCLUDE_FORM
6 rock rock NOUN Noun Number=Sing 5 obj 5:obj _ Object DYNAMIC_ARTS
```

CoBaLD

- Семантическая часть: семантические роли (глубинные отношения) и семантические классы
- Семантические роли ~ модель управления (только активные)
- Семантические классы – значения слов
- Есть онтология (Compreno)
- Для CoBaLD формат Compreno сильно упростили, было очень много классов, стало 600+ (все равно много)
- Есть парсер, довольно высокого качества:

	Total	Lemma	POS	Features	UAS	LAS	SemSlot	SemClass
RuBERT-tiny	92.2%	96.1%	98.2%	95.3%	90.0%	85.6%	87.8%	92.2%
XLM-R	95.1%	97.3%	98.8%	96.8%	93.5%	89.8%	94.3%	94.8%

CoBaLD



Языковой перенос

- Языковые модели типа BERT бывают мультиязычными
- Учим модель на сырых текстах сразу нескольких языков
- Модель «знает» эти языки, у нее есть вероятностные распределения для них
- Когда используем модель для решения downstream-задачи, файнтюним ее на размеченных данных уже для конкретного языка
- Можно взять такую модель и применить ее на данных другого языка
- Это применяется в машинном обучении в принципе и называется transfer learning
- Когда речь идет о языках, говорят cross-lingual transfer
- Доказано, что переносе с английского качество теряется не больше чем на 25%

Языковой перенос

- Модель, обученную размечать в формате CoBaLD для русского языка, заставили размечать другие языки:

```
# sent_id = 2
# text = İsrail Gazze Şeridi'nin kuzeyindeki bir tarlaya hava saldırısı düzenledi.
# text = Israel Launched an airstrike on a field in the northern Gaza Strip.
1 İsrail israel PROPN _ Case=Nom|Number=Sing 9 nsubj Agent COUNTRY_AS_ADMINISTRATIVE_UNIT
2 Gazze gazze NOUN _ Case=Nom|Number=Sing|Person=3 3 nmod Name_Title INHABITED_LOCALITY
3 Şeridi'nin şeridt NOUN _ Case=Gen|Number=Sing|Number[psor]=Sing|Person=3|Person[psor]=3 4 nmod Whole THE_EARTH_AND_ITS_SPATIAL_PARTS
4 kuzeyindeki kuzey ADJ _ 6 amod Locative CH_DISPOSITION_AND_MOTION
5 bir bir DET _ Definite=Ind|PronType=Art 6 det Quantity CH_REFERENCE_AND_QUANTIFICATION
6 tarlaya tarla NOUN _ Case=Dat|Number=Sing|Person=3 9 obl Locative_FinalPoint PLACE
7 hava hava NOUN _ Case=Nom|Number=Sing|Person=3 8 nmod Agent_Metaphoric THE_EARTH_AND_ITS_SPATIAL_PARTS
8 saldırısı saldırı NOUN _ Case=Nom|Number=Sing|Number[psor]=Sing|Person=3|Person[psor]=3 9 obj Object_Situation AGGRESSIVE_ACTIONS
9 düzenledi düzenen VERB _ Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Past|VerbForm=Fin 0 root Predicate TO_COMMIT
10 . . PUNCT _ _ 9 punct _ _
```

```
# sent_id = 4
# text = مطلعٌ قَبْلَ قِرْطَلِ اهْنُ عَلَى حِجَّتِ لَتْسِ الْأَرْضِ (2023-2027).
1 بـ ADP P----- AdpType=Prep 2 case _ PREPOSITION
2 الْأَرْضِ NOUN N-----S2D Case=Gen|Definite=Def|Number=Sing 0 root Locative ADMINISTRATIVE_STRUCTURES
3 بـ ADV D----- _ 2 case Agent BEING
4 قَبْلَ X U----- _ 5 nmod Agent BEING
5 عَلَى X U----- _ 2 nsubj Predicate VERBAL_COMMUNICATION
6 حِجَّتِ NOUN N-----S2D Case=Gen|Definite=Def|Number=Sing 5 nmod Object METHOD
7 الْأَرْضِيِّيْنِ ADJ A-----FS2D Case=Gen|Definite=Def|Gender=Fem|Number=Sing 6 amod Characteristic COUNTRY_AS_ADMINISTRATIVE_UNIT
8 قَبْلَ قَبْلَ قَبْلَ قَبْلَ NOUN N-----S2D Case=Gen|Definite=Def|Number=Sing 6 nmod Object_Situation CULTURE
9 مَعْلُومَاتِ NOUN N-----P2D Case=Gen|Definite=Def|Number=Plur 6 nmod Time UNIT_OF_TIME
10 ( ( PUNCT G----- _ 11 punct
11 2023 2023 NUM Q----- NumForm=Digit 6 nummod OrderInTimeAndSpace CH_REFERENCE_AND_QUANTIFICATION
12 - - PUNCT G----- _ 13 punct
13 2027 2027 NUM Q----- NumForm=Digit 11 conj OrderInTimeAndSpace CH_REFERENCE_AND_QUANTIFICATION
14 ) ) PUNCT G----- _ 11 punct _ _
15 . . PUNCT G----- _ 2 punct _ _
```