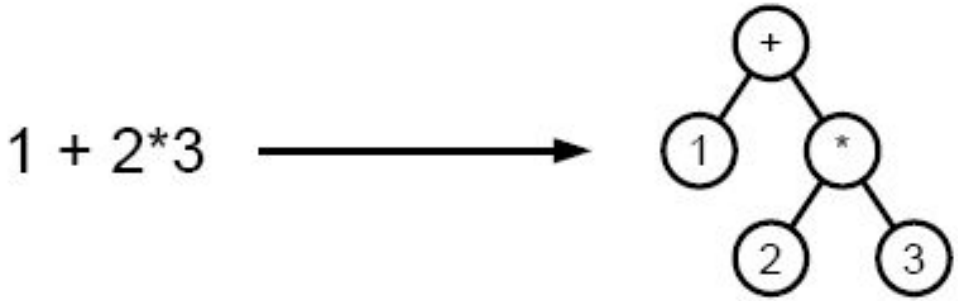
Natural Language Processing

Синтаксический анализ - часть 1

Синтаксический анализ - это процесс сопоставления линейной последовательности лексем (слов, токенов) естественного или формального языка с его формальной грамматикой.

Результатом обычно является **дерево разбора** (синтаксическое дерево)



Синтаксический анализ

Синтаксические деревья используются в таких задачах, как:

- исправление грамматических ошибок;
- семантический анализ;
- вопросно-ответные системы;
- извлечение информации;
- и др.

Пример: *What books were written by British women authors before 1800?*

Синтаксический анализ

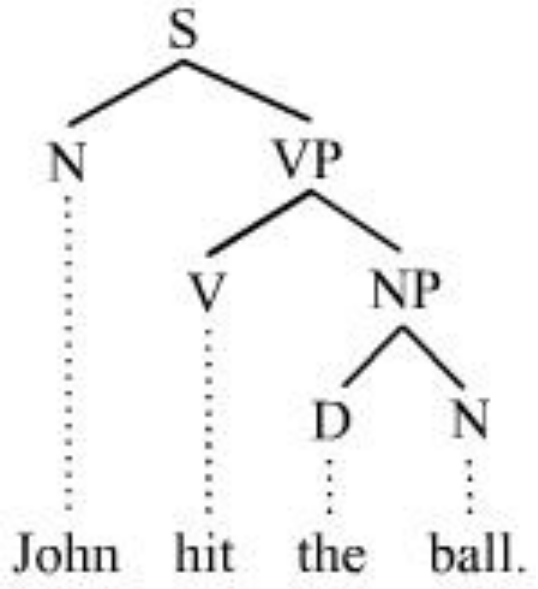
Синтаксические деревья:

- Деревья составляющих;
- Деревья зависимостей

Деревья составляющих

Составляющая - структурная единица (отрезок) предложения, целиком составленная из более тесно связанных друг с другом составляющих меньшего размера.

Грамматика составляющих (*метод составляющих*; англ. *constituency grammar*, *phrase structure grammar*) основана на постулате, согласно которому всякая сложная грамматическая единица складывается из двух более простых и не пересекающихся единиц, называемых её **непосредственными составляющими** (англ. *immediate constituent*).



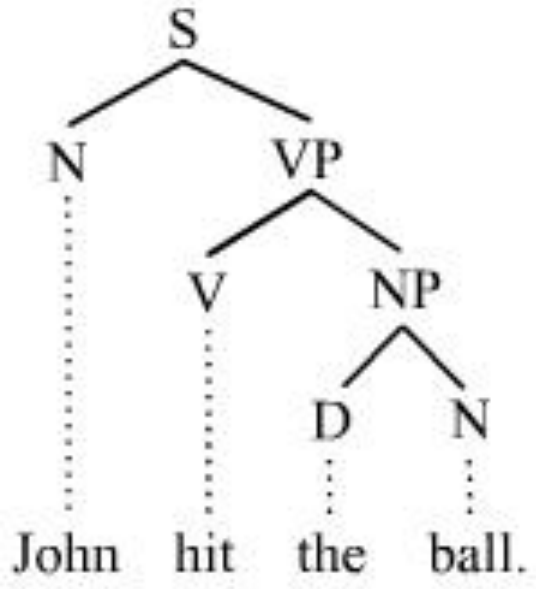
Деревья составляющих

Группа (англ. *phrase*) - это составляющая, включающая более одного слова.

Вершина (англ. *head*) группы - слово, соответствующее корневому узлу в дереве зависимостей, описывающем группу.

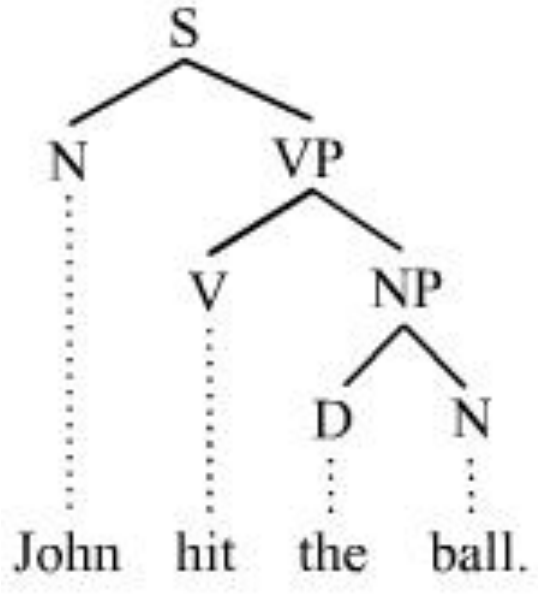
Единичная составляющая - это составляющая, состоящая из одного сегмента.

Полная составляющая - это сегмент, включающий в себя все составляющие данного предложения.



Деревья составляющих

(John (hit (the ball))).



Деревья составляющих

Классификация групп основывается на частеречной принадлежности их вершин:

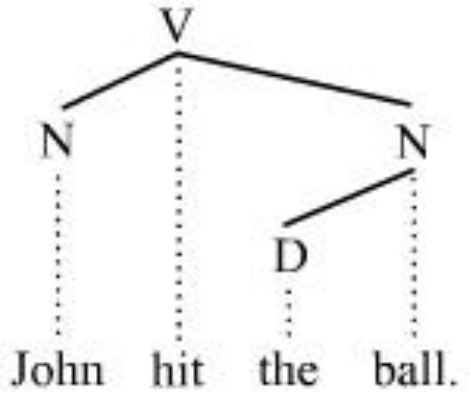
- **именная группа** (англ. *noun phrase*, *NP*) — возглавляется существительным;
- **группа прилагательного** (англ. *adjectival phrase*, *AP*) — возглавляется прилагательным;
- **наречная группа** (англ. *adverbial phrase*, *AdvP*) — возглавляется наречием;
- **предложная группа** (англ. *prepositional phrase*, *PP*) — возглавляется предлогом;
- **глагольная группа** (англ. *verb phrase*, *VP*) — возглавляется глаголом;
- **предложение** (англ. *sentence*, *S*).

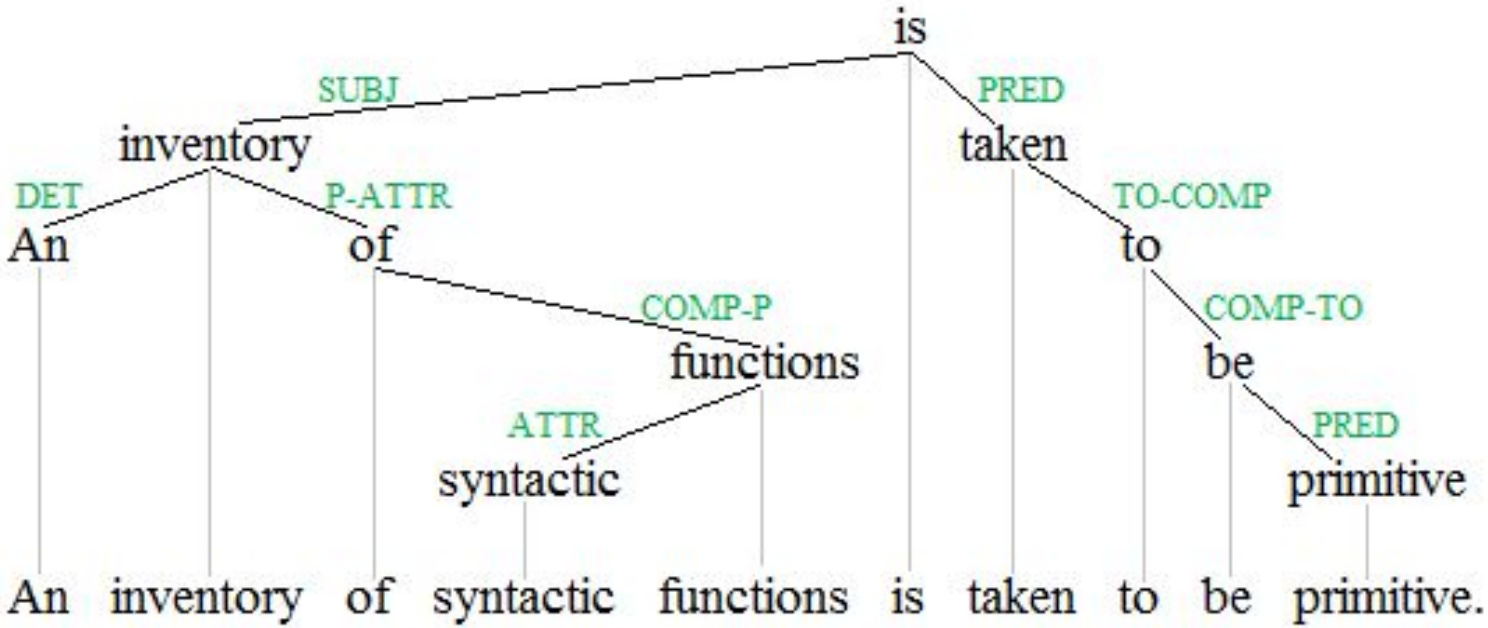
Некоторые фразовые категории, в частности именная группа и предложение, обладают **свойством рекурсивности** — способностью включать в себя составляющие той же фразовой категории.

Деревья зависимостей

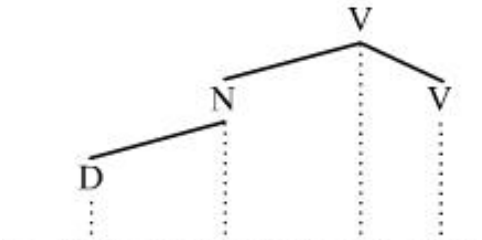
Представляет строй предложения в виде иерархии компонентов, между которыми установлено отношение зависимости. Таким образом, структура предложения рассматривается в терминах вершин и зависимых.

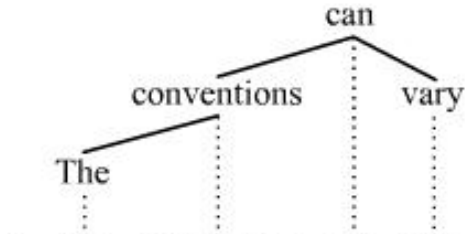
Грамматикой зависимостей в узком смысле называется теория синтаксической структуры предложения, в которой все связи в предложении рассматриваются как подчинительные, вершиной предложения признаётся сказуемое или его знаменательная часть, а предлоги описываются как управляющие связанными с ними формами существительных.

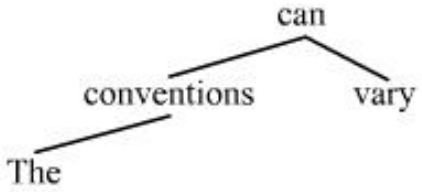


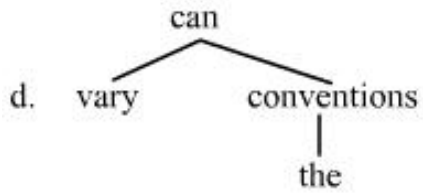


Annotated dependency tree

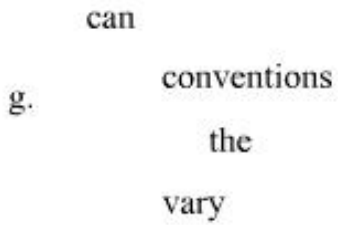








f. [[The] conventions] can [vary].



Representing dependencies

Universal Dependencies

http://universaldependencies.org/

Clausal Argument Relations	Description
NSUBJ	Nominal subject
DOBJ	Direct object
IOBJ	Indirect object
CCOMP	Clausal complement
XCOMP	Open clausal complement
Nominal Modifier Relations	**Description**
NMOD	Nominal modifier
AMOD	Adjectival modifier
NUMMOD	Numeric modifier
APPOS	Appositional modifier
DET	Determiner
CASE	Prepositions, postpositions and other case markers
Other Notable Relations	**Description**
CONJ	Conjunct
CC	Coordinating conjunction

Selected dependency relations from the Universal Dependency set

Universal Dependencies

Relation	Examples with *head* and **dependent**
NSUBJ	**United** *canceled* the flight.
DOBJ	United *diverted* the **flight** to Reno.
	We *booked* her the first **flight** to Miami.
IOBJ	We *booked* **her** the flight to Miami.
NMOD	We took the **morning** *flight*.
AMOD	Book the **cheapest** *flight*.
NUMMOD	Before the storm JetBlue canceled **1000** *flights*.
APPOS	*United*, a **unit** of UAL, matched the fares.
DET	**The** *flight* was canceled.
	Which *flight* was delayed?
CONJ	We *flew* to Denver and **drove** to Steamboat.
CC	We flew to Denver **and** *drove* to Steamboat.
CASE	Book the flight **through** *Houston*.

Examples of core Universal Dependency relations

Dependency Treebanks

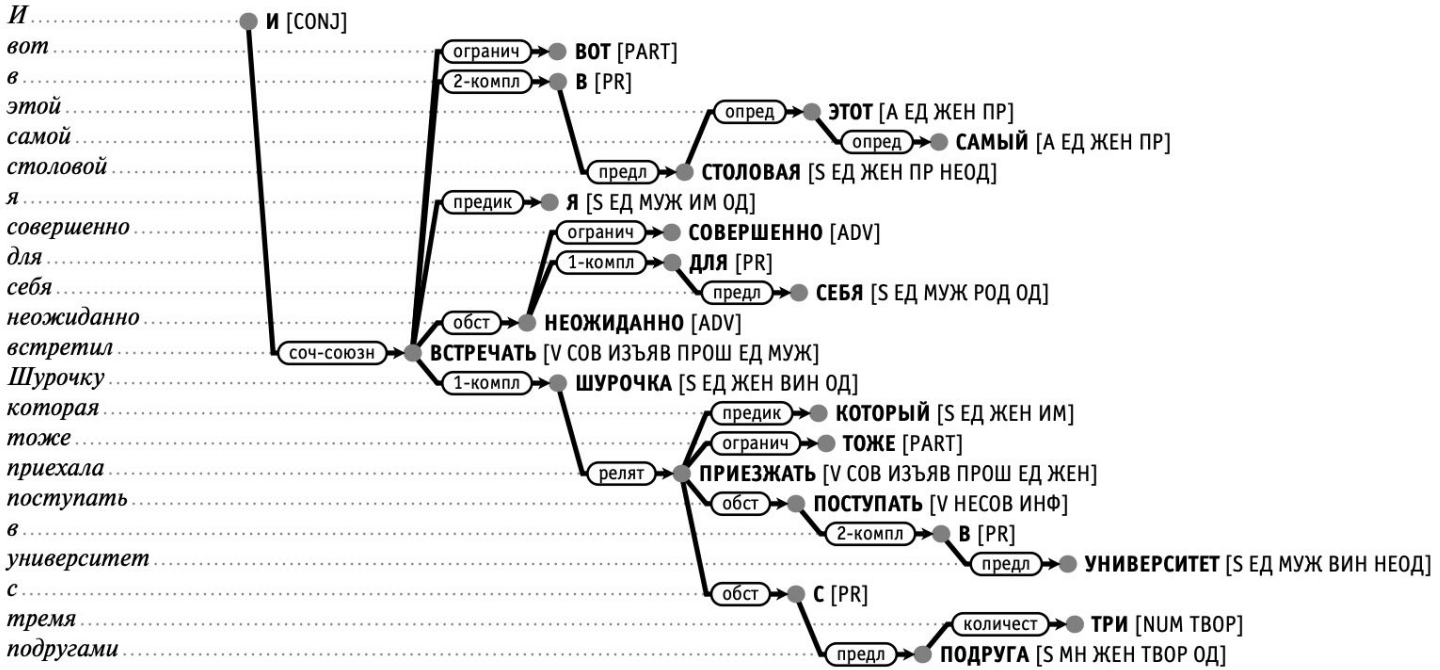
- http://universaldependencies.org
  - A Gold Standard Universal Dependencies Corpus for English (16,622 sentences);
  - Treebank of Learner English (TLE) (4,624 sentences)
  - some more…
- https://catalog.ldc.upenn.edu/ldc2013t19 - OntoNotes

Dependency Treebanks: SynTagRus

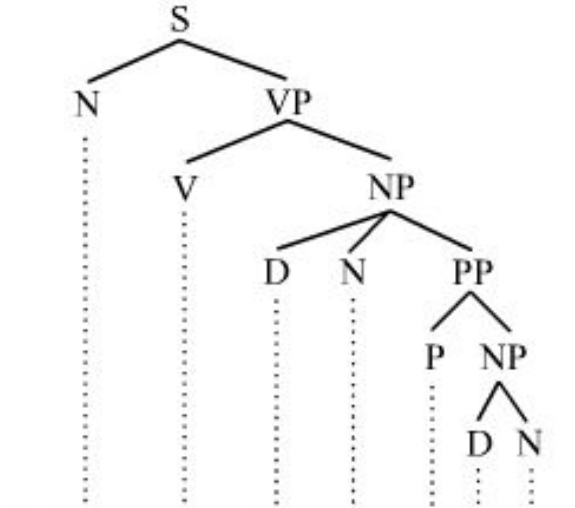
Содержит:

- 1,107,741 токен;
- 61,889 предложений

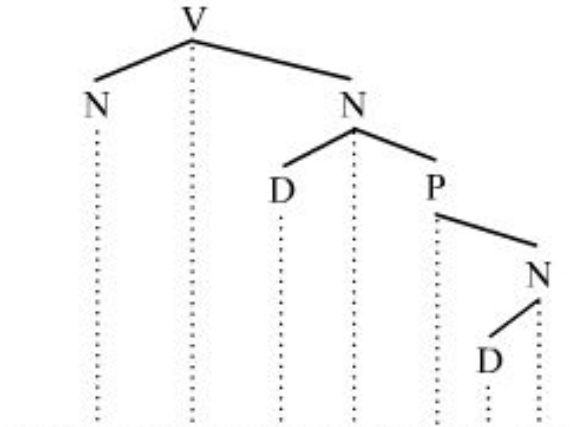
Dependency Treebanks: SynTagRus



Сравнение



Phrase structure grammar

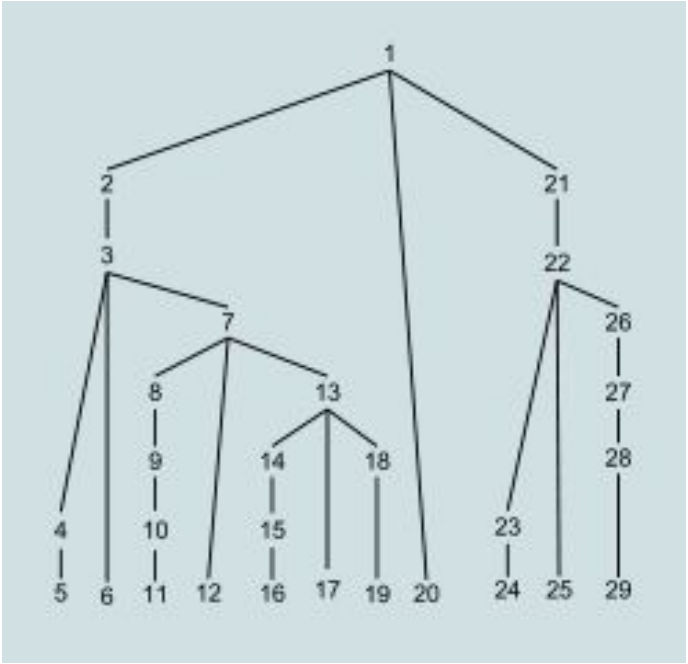


Dependency grammar

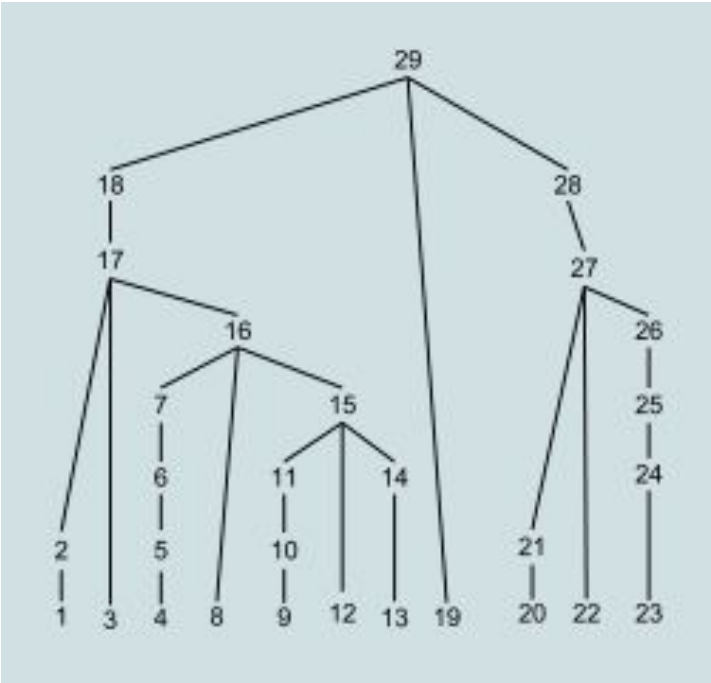
Сравнение

- Деревья зависимостей лучше подходят для языков с относительно свободным порядком слов;
- Деревья зависимостей содержат семантическую информацию об отношении между предикатом и аргументом, что позволяет использовать их в решении практических задач;
- Деревья составляющих основаны на идеи бинарного разделения, в грамматике зависимостей глагол является основой всего предложения.

Top-down parser



Bottom-up parser



Проблемы синтаксического анализа

- Частеречная разметка;
- Синтаксическая омонимия;
- Word-sense disambiguation;
- Синтаксическая синонимия;
- Разрешение кореференции

POS-tagging, word-category disambiguation

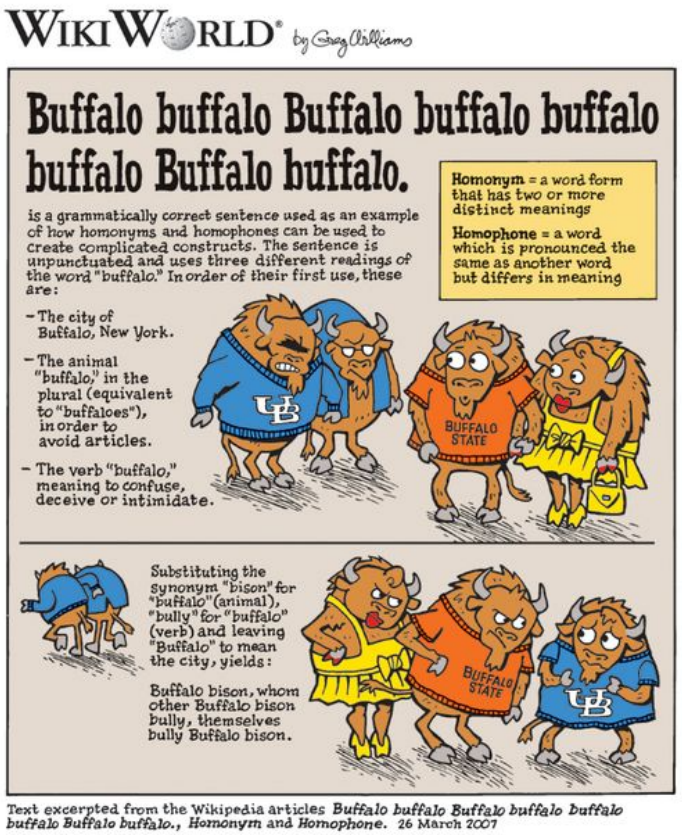
Morphological homonymy is phonetic and graphic equivalence of wordforms, which have different parts of speech. Example: *I read a book* vs. *Please, book a hotel*.

Methods of POS disambiguation:

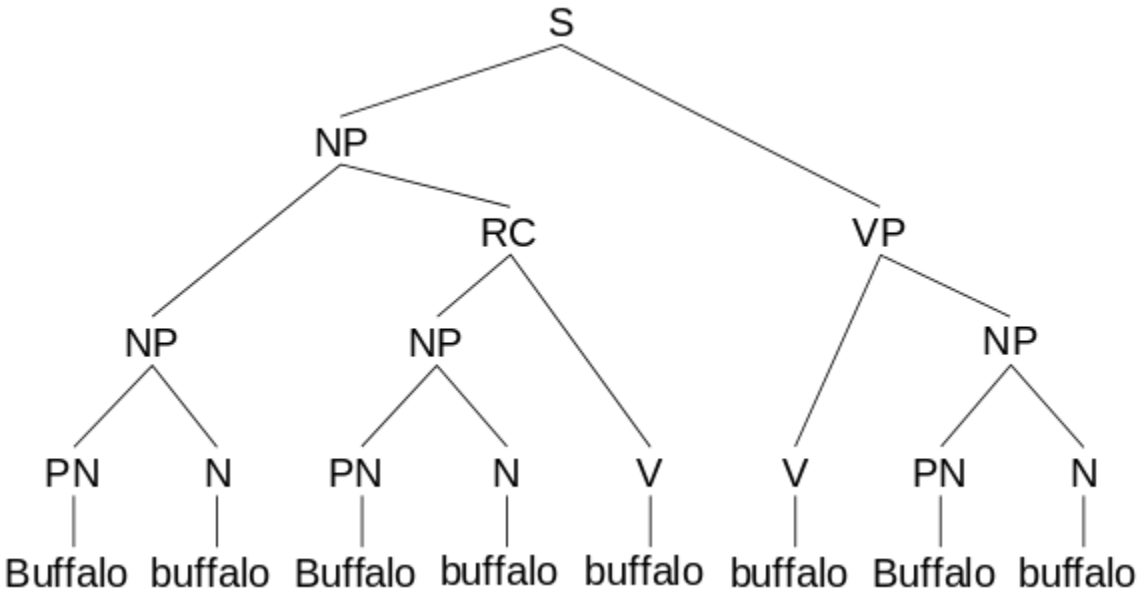
1. Deterministic;
2. Statistical.

POS-tagging, word-category disambiguation

«Буффальские бизоны, которых пугают (другие) буффальские бизоны, пугают буффальских бизонов»



POS-tagging, word-category disambiguation



POS-tagging, word-category disambiguation

За песчаной косой косой косой косой косой косой косой косил прокос.

За песчаной косой (за песчаным берегом) косой косой (кривой косою, то есть кривым орудием труда) косой косой (заяц с косоглазием) косой косил прокос (косил прокос, который у него выходил неровным, то есть прокошенная полоса получилась не похожей на ровный отрезок, поэтому "косой прокос").

Синтаксическая омонимия

Синтаксическая омонимия - это явление, возникающее, когда группы или части предложения имеют одинаковую форму, но разные синтаксические функции.

Примеры:

He hit a woman with a baby.

Он из Германии туманной привез учености плоды.

Лексическая омонимия

Лексическая омонимия - это фонетическая и графическая эквивалентность словоформ, имеющих одну и ту же часть речь, но разные семантические значения.

Примеры:

`Лук` (оружие) vs. `лук` (овощ)

`Bank` (пляж/берег) vs. `bank` (финансовая организация)

Разрешение кореференции (coreference resolution)

Кореференция – это попытка связать несколько разных отсылок в тексте к одному реальному объекту.

Примеры:

1. *Carol told Bob to attend the party. They arrived together.*
2. *If they are angry about the music, the neighbours will call the cops.*
3. *Я встретила друга с братом, он был очень вежливым.-?*

Разрешение кореференции (coreference resolution)

- Графическая: *photocontent* - *photo-content* - *photo content*;
- Транслитерация: Google - Гугл;
- Аббревиатуры: НГУ - Новосибирский государственный университет;
- Синонимы: aim - goal

На следующей лекции

- Обзор подходов к синтаксическому разбору предложений;
- Инструменты, современные парсеры;
- Оценка качества парсеров