

# Computational lexicology, morphology and syntax

Diana Trandabat

2022-2023

# Course overview

- Intro to Lexicology:
  - Vocabulary
  - Word, morpheme
  - Corpus
  - Dictionaries
  - WordNet
- Intro to Morphology
  - Finite state morphology

# Course overview

- Intro to Syntax
  - POS Tagging
  - Regular expressions
  - Chomsky hierarchy
  - CKY Parsing
  - Earley's Algorithm
  - Dependency parsing
  - Phrase Structure Grammar

# Prerequisites

**WILL TO LEARN!!!**

# Expectations

## LEARN

- Adequately use lexicology, morphology and syntax terminology
- Develop and/or implement NLP grammars
- Knowledge of text parsing techniques
- Better project presentations skills

## DO

- Assignments
- Project

# Evaluation

- Laboratory – 100 points
  - Attendance(10%)
  - Homework (90%)
- Project – 100 points
- Exam – 100 points
  - Midterm
  - Final exam

# Homework

- ~ Weekly
- In class delivery
  - 50% of the points for delivery in class; 50% for email submitted homework
- Late delivery for email submissions
  - 100% of the points for delivery on time
  - 80% of the points for 1 day late delivery
  - 60% of the points for 2 days late delivery
  - ...
- Name convention:  
LEX\_HomeworkNO\_StudentName\_ProgrammingLanguage
- Each implementation task is submitted with a short documentation (max. 1 page) with implementation details, challenges, methods/solutions, errors, problems etc.

# Projects

- We'll get to that latter...
- But in the meantime you can check [SemEval](#) and [Clef](#) competitions



# How do people communicate?

Different ways:

- speaking and listening
- making gestures
- specialized hand signals (such as when driving or directing traffic)
- sign languages for the deaf
- various forms of text

# Communication breakdown

- S (speaker) wants to convey P (proposition) to H (hearer) using W (words in a formal or natural language)

## 1. Speaker

- **Intention:** S wants H to believe P
- **Generation:** S chooses words W
- **Synthesis:** S utters words W

## 2. Hearer

- **Perception:** H perceives words  $W''$  (ideally  $W'' = W$ )
- **Analysis:** H infers possible meanings  $P_1, P_2, \dots, P_n$  for  $W''$
- **Disambiguation:** H infers that S intended to convey  $P_i$  (ideally  $P_i = P$ )
- **Incorporation:** H decides to believe or disbelieve  $P_i$

# What is NLP?

NLP - subdomain of Artificial Intelligence

≈ computational linguistics

≈ language technologies

Goal: communication man-machine in natural language

# Two major NLP directions

1. Natural Language Understanding and Analyzing
  - Input: spoken/written sentence
  - Output: some representation of the meaning of the sentence
2. Natural Language Generation
  - Input: some formal representation of what you intend to communicate
  - Output: expression of what we want to convey in a natural (human) language, i.e. a text or speech

# Examples of NLP applications

- Machine Translation
- Database Access
- Information Retrieval
- Text Categorization
- Extracting data from text
- Spoken language control systems
- Spelling and grammar checkers
- etc. etc. etc.

# Examples of NLP applications

- Microsoft: <https://www.microsoft.com/en-us/ai/ai-lab>
- Google: <https://experiments.withgoogle.com/collection/ai>
- Amazon: <https://ai-service-demos.go-aws.com>
- TalkToTransformer: <https://talktotransformer.com>

# Understanding a text

Goliat, the first Romanian nanosatellite, was successfully launched on the orbit Monday, 13th February 2012, from a base in the French Guyana, during the inaugural flight of the VEGA rocket. The satellite has been developed by a research team headed by the Romanian Space Agency between 2005-2009.

Source: <http://science.hotnews.ro/stiri-spatiul-11496155-live-video-ora-12-00-goliat-primul-satelit-romanesc-lansat-orbita.htm>

# Understanding a text - morphology

By morphological analysis we can identify part of speeches:

- Nouns;
- Verbs
- etc.

Goliat, the first Romanian nanosatellite, was successfully launched on the orbit Monday, 13th February 2012, from a base in the French Guyana, during the inaugural flight of the VEGA rocket. The satellite has been developed by a research team headed by the Romanian Space Agency between 2005-2009.

Source: <http://science.hotnews.ro/stiri-spatiul-11496155-live-video-ora-12-00-goliat-primul-satelit-romanesc-lansat-orbita.htm>



# Understanding a text - syntax

By syntactical analysis we can identify grammatical constituents:

- Subject;
- Predicate;
- etc.

**Goliat**, the first Romanian nanosatellite, **was** successfully **launched** on the orbit Monday, 13th February 2012, from a base in the French Guyana, during the inaugural flight of the VEGA rocket. The satellite has been developed by a research team headed by the Romanian Space Agency between 2005-2009.

Source : <http://science.hotnews.ro/stiri-spatiul-11496155-live-video-ora-12-00-goliat-primul-satelit-romanesc-lansat-orbita.htm>

# Understanding a text - semantics

By semantical analysis we can understand a text:

- Who, what, where, when, how, why etc. performs an action
- The meaning of words
- References/Anaphora

Goliat, the first Romanian nanosatellite, was successfully launched on the orbit Monday, 13th February 2012, from a base in the French Guyana, during the inaugural flight of the VEGA rocket. The satellite has been developed by a research team headed by the Romanian Space Agency between 2005-2009.

Source : <http://science.hotnews.ro/stiri-spatiul-11496155-live-video-ora-12-00-goliat-primul-satelit-romanesc-lansat-orbita.htm>

# Understanding a text

01000111011011110110110001101001011000010111010000101100001  
00000011100000111001001101001011011010111010101101100001000  
00011011100110000101101110011011110111001101100001011101000  
11001010110110001101001011101000010000001110010011011110110  
1101110000111010001001101110011001010111001101100011.....

Goliat, the first Romanian nanosatellite, was successfully launched on the orbit Monday, 13th February 2012, from a base in the French Guyana, during the inaugural flight of the VEGA rocket. The satellite has been developed by a research team headed by the Romanian Space Agency between 2005-2009.

Source : <http://science.hotnews.ro/stiri-spatiul-11496155-live-video-ora-12-00-goliat-primul-satelit-romanesc-lansat-orbita.htm>

# Tokenization

Breaking up a stream of characters into tokens: words, punctuation marks, numbers and other discrete items

Goliat, the first Romanian nanosatellite, was successfully launched on the orbit Monday, 13th February 2012, from a base in the French Guyana, during the inaugural flight of the VEGA rocket. The satellite has been developed by a research team headed by the Romanian Space Agency between 2005-2009.

Source : <http://science.hotnews.ro/stiri-spatiul-11496155-live-video-ora-12-00-goliat-primul-satelit-romanesc-lansat-orbita.htm>

# How many tokens?

今天天气晴朗

it's sunny today

c'est ensoleillé aujourd'hui

# What can we learn just through tokenization?

- Text statistics: no. of words, multi-word expressions, length of words/sentences, freq. of vowels/consonants > Language Identification
- Named Entity Recognition

Goliat, the first Romanian nanosatellite, was successfully launched on the orbit Monday, 13th February 2012, from a base in the French Guyana, during the inaugural flight of the VEGA rocket. The satellite has been developed by a research team headed by the Romanian Space Agency between 2005-2009.

Source : <http://science.hotnews.ro/stiri-spatiul-11496155-live-video-ora-12-00-goliat-primul-satelit-romanesc-lansat-orbita.htm>

# Named Entity Recognition

*“Dr. Watson, Mr. Sherlock Holmes”, said Stamford, introducing us.*

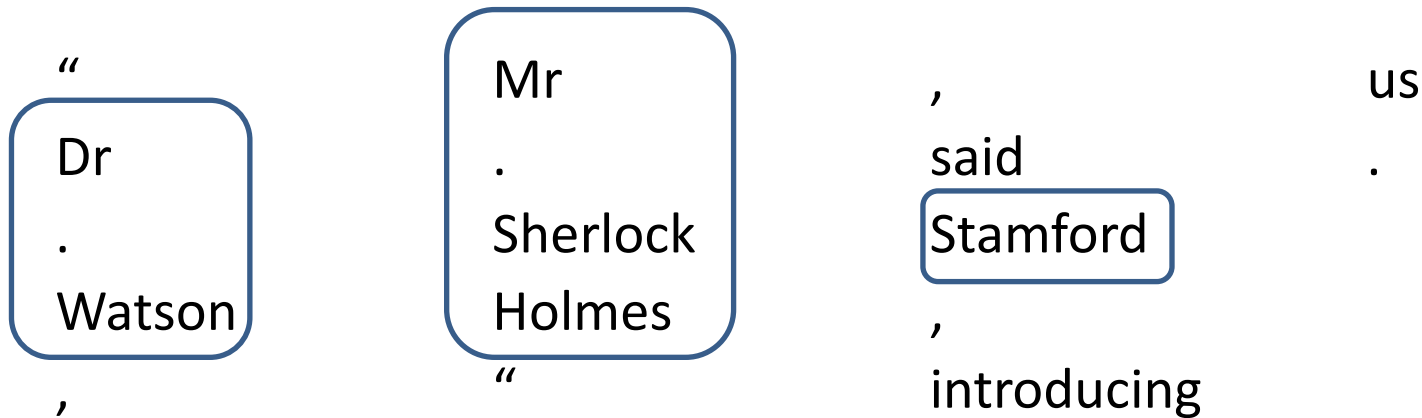
“  
Dr  
.  
Watson  
,

Mr  
.  
Sherlock  
Holmes  
”

,                    us  
said  
Stamford  
,  
introducing

# Named Entity Recognition

*“Dr. Watson, Mr. Sherlock Holmes”, said Stamford, introducing us.*



## *Examples of rules for NER*

- (Dr. | Professor | Mr. | Mrs. | Miss | Ms.) *Word*
- *Word Word, Integer*
- *Word* (said | thought | believed | claimed | argued | ...)



# What can we learn just through tokenization?

- Text statistics: no. of words, multi-word expressions, length of words/sentences, freq. of vowels/consonants > Language Identification
- Named Entity Recognition

Goliat, the first Romanian nanosatellite, was successfully launched on the orbit Monday, 13th February 2012, from a base in the French Guyana, during the inaugural flight of the VEGA rocket. The satellite has been developed by a research team headed by the Romanian Space Agency between 2005-2009.

# Natural Language Processing

... The **first** Romanian **nanosatellite** was ... **launched** ...

– Morphology:

- Use a dictionary to identify part of speeches

- First = **numeral**;

- nanosatellite = **noun**;

- launched = **verb**; ...

- Difficulty: ambiguity

- I like **research**

- I **research** natural language processing

# Computational morphology

- Computational morphology deals with
  - developing theories and techniques for
  - computational analysis and synthesis of word forms.
- Analysis: Separate and identify the constituent morphemes and mark the information they encode
- Synthesis (Generation): Given a set constituent morphemes or information be encoded, produce the corresponding word(s)

# Computational Morphology -Analysis

- Computational morphology deals with
  - developing theories and techniques for
  - computational analysis and synthesis of word forms.
- Extract any information encoded in a word and bring it out so that later layers of processing can make use of it

stopping    ⇒ stop+Verb+Cont

happiest    ⇒ happy+Adj+Superlative

went        ⇒ go+Verb+Past

books       ⇒ book+Noun+Plural

              ⇒ book+Verb+Pres+3SG.

# Computational Morphology -Generation

- In a machine translation applications, one may have to generate the word corresponding to a set of features
  - stop+Past  $\Rightarrow$  stopped
  - cânta+Past+1Pl  $\Rightarrow$  cântaserăm/cântasem
    - +2Pl  $\Rightarrow$  cântaserăți/cântasei

# WordNet

- What is missing in traditional dictionaries
  - It does not say, for example, that trees have roots, or that they consist of cells having cellulose walls, or even that they are living organisms
  - “Sense” of the super ordinate term aka **hypernym** (living plant or industrial plant)
  - **Coordinate terms** (bushes, shrubs, ...)
  - **Hyponyms** - types of trees (pine, tropical, deciduous..)
  - Information assumed to be known to everyone ( trees have barks and leaves, they grow from seeds, they make their own food by photosynthesis- probably information for encyclopedia!)

# What is WordNet?

- WordNet is a lexical database
- WordNet 3.0 has [1]:
  - 117,097 nouns (average noun has 1.23 senses)
  - 11,488 verbs (average verb has 2.16 sense)
  - 22,141 adjectives
  - 4,601 adverbs
- Created and maintained at Princeton University
- Accessible online @  
<http://wordnetweb.princeton.edu/perl/webwn>  
(Also Downloadable)
- Interfaces available in C, .Net , Java, Perl, Php, Python, Sql etc.

# What is a synset?

- Basic unit of WordNet
- A group of synonymous words which refer to a common semantic concept
- Words may belong to more than one synset – first sense is the most frequent sense
- Words also include collocations (“eye contact”, “mix up”)



# Synset examples

- “car” in
  - {car, auto, automobile, machine, motorcar}
  - {car, railcar, railway car, railroad car}.
- “Chocolate” in

## Noun

- **S: (n) cocoa, chocolate, hot chocolate, drinking chocolate** (a beverage made from cocoa powder and milk and sugar; usually drunk hot)
- **S: (n) chocolate** (a food made from roasted ground cacao beans)
- **S: (n) chocolate, coffee, deep brown, umber, burnt umber** (a medium brown to dark-brown color)

# Beyond WordNet

- eXtended WordNet
- SentiWordNet
  - Each term in WordNet database is assigned a score of 0 to 1 in SentiWordNet which indicates its polarity
- WordNet for languages other than English
- FrameNet
- SentiFrameNet

# English Parts of Speech

- Noun (person, place or thing)
  - Singular (NN): dog, fork
  - Plural (NNS): dogs, forks
  - Proper (NNP, NNPS): John, Springfields
- Pronouns
  - Personal pronoun (PRP): I, you, he, she, it
  - Wh-pronoun (WP): who, what
- Verb (actions and processes)
  - Base, infinitive (VB): eat
  - Past tense (VBD): ate
  - Gerund (VBG): eating
  - Past participle (VBN): eaten
  - Non 3<sup>rd</sup> person singular present tense (VBP): eat
  - 3<sup>rd</sup> person singular present tense: (VBZ): eats
  - Modal (MD): should, can
  - To (TO): to (to eat)

# English Parts of Speech (cont.)

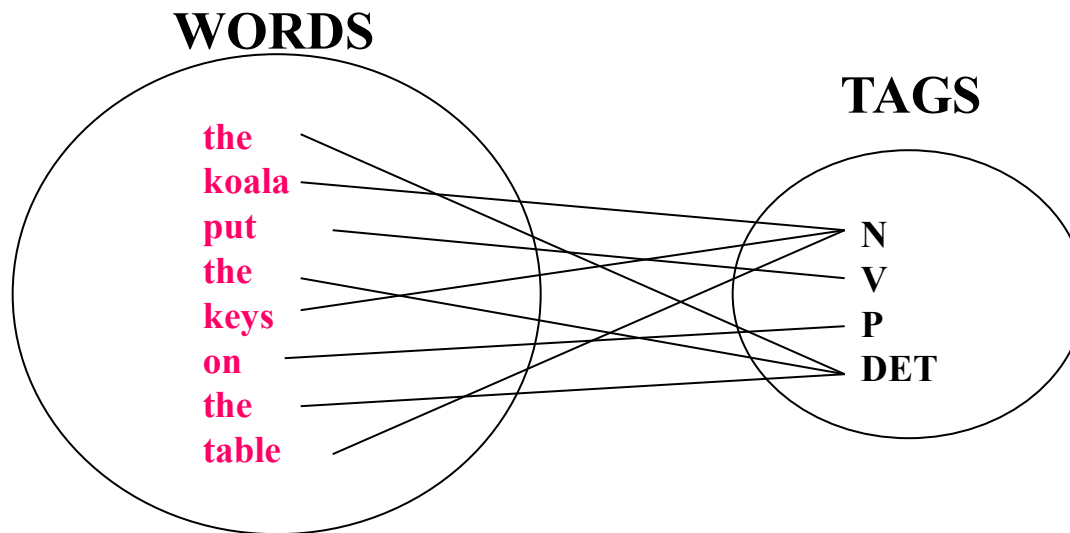
- Adjective (modify nouns)
  - Basic (JJ): red, tall
  - Comparative (JJR): redder, taller
  - Superlative (JJS): reddest, tallest
- Adverb (modify verbs)
  - Basic (RB): quickly
  - Comparative (RBR): quicker
  - Superlative (RBS): quickest
- Preposition (IN): on, in, by, to, with
- Determiner:
  - Basic (DT) a, an, the
  - WH-determiner (WDT): which, that
- Coordinating Conjunction (CC): and, but, or,
- Particle (RP): off (took off), up (put up)

# Penn Treebank Tagset

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	<i>and, but, or</i>	SYM	Symbol	<i>+, %, &amp;</i>
CD	Cardinal number	<i>one, two, three</i>	TO	“to”	<i>to</i>
DT	Determiner	<i>a, the</i>	UH	Interjection	<i>ah, oops</i>
EX	Existential ‘there’	<i>there</i>	VB	Verb, base form	<i>eat</i>
FW	Foreign word	<i>mea culpa</i>	VBD	Verb, past tense	<i>ate</i>
IN	Preposition/sub-conj	<i>of, in, by</i>	VBG	Verb, gerund	<i>eating</i>
JJ	Adjective	<i>yellow</i>	VBN	Verb, past participle	<i>eaten</i>
JJR	Adj., comparative	<i>bigger</i>	VBP	Verb, non-3sg pres	<i>eat</i>
JJS	Adj., superlative	<i>wildest</i>	VBZ	Verb, 3sg pres	<i>eats</i>
LS	List item marker	<i>1, 2, One</i>	WDT	Wh-determiner	<i>which, that</i>
MD	Modal	<i>can, should</i>	WP	Wh-pronoun	<i>what, who</i>
NN	Noun, sing. or mass	<i>llama</i>	WP\$	Possessive wh-	<i>whose</i>
NNS	Noun, plural	<i>llamas</i>	WRB	Wh-adverb	<i>how, where</i>
NNP	Proper noun, singular	<i>IBM</i>	\$	Dollar sign	<i>\$</i>
NNPS	Proper noun, plural	<i>Carolinas</i>	#	Pound sign	<i>#</i>
PDT	Predeterminer	<i>all, both</i>	“	Left quote	<i>( ‘ or “)</i>
POS	Possessive ending	<i>'s</i>	”	Right quote	<i>( ’ or ”)</i>
PRP	Personal pronoun	<i>I, you, he</i>	(	Left parenthesis	<i>( [ , ( , { , &lt;</i>
PRP\$	Possessive pronoun	<i>your, one's</i>	)	Right parenthesis	<i>( ] , ) , } , &gt;</i>
RB	Adverb	<i>quickly, never</i>	,	Comma	<i>,</i>
RBR	Adverb, comparative	<i>faster</i>	.	Sentence-final punc	<i>( . ! ?)</i>
RBS	Adverb, superlative	<i>fastest</i>	:	Mid-sentence punc	<i>( : ; ... - -)</i>
RP	Particle	<i>up, off</i>			

# Defining POS Tagging

- The process of assigning a part-of-speech or lexical class marker to each word in a corpus:



# Applications for POS Tagging

- Speech synthesis pronunciation
  - *Lead*                      *Lead*
  - *INsult*                     *inSULT*
  - *OBject*                    *obJECT*
  - *OVERflow*                *overFLOW*
  - *DIScount*                *disCOUNT*
  - *CONtent*                 *conTENT*
- Word Sense Disambiguation: e.g. *Time flies like an arrow*
  - Is *flies* an N or V?
- Word prediction in speech recognition
  - Possessive pronouns (*my, your, her*) are likely to be followed by nouns
  - Personal pronouns (*I, you, he*) are likely to be followed by verbs
- Machine Translation

# Natural Language Processing

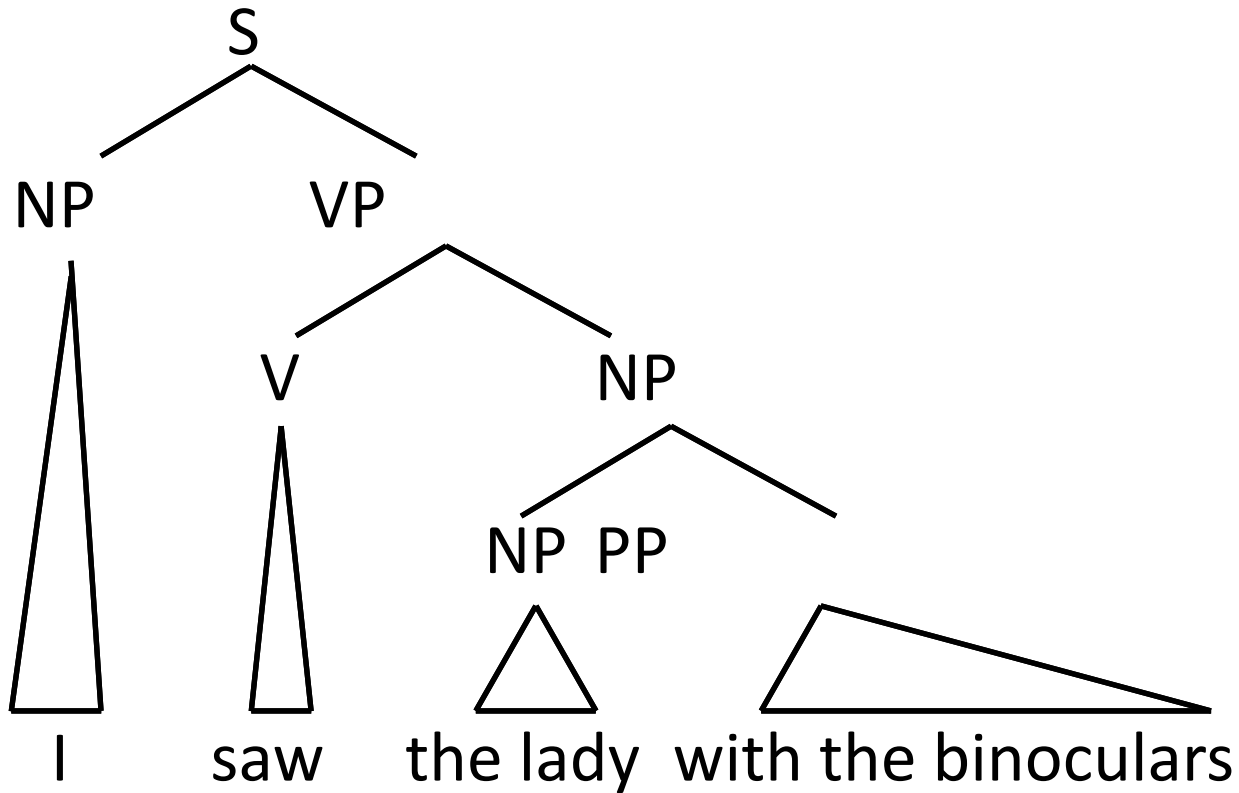
... The first Romanian nanosatellite was ... launched ...

## – Syntax:

- coded in formal grammars
  - determiner + numeral + noun+ adjective = **noun group**;
  - auxiliary + verb = **verb group**;
  - noun group + verb group = **sentence**; ...
- Difficulty: Ambiguity
  - I saw the lady with the binoculars.

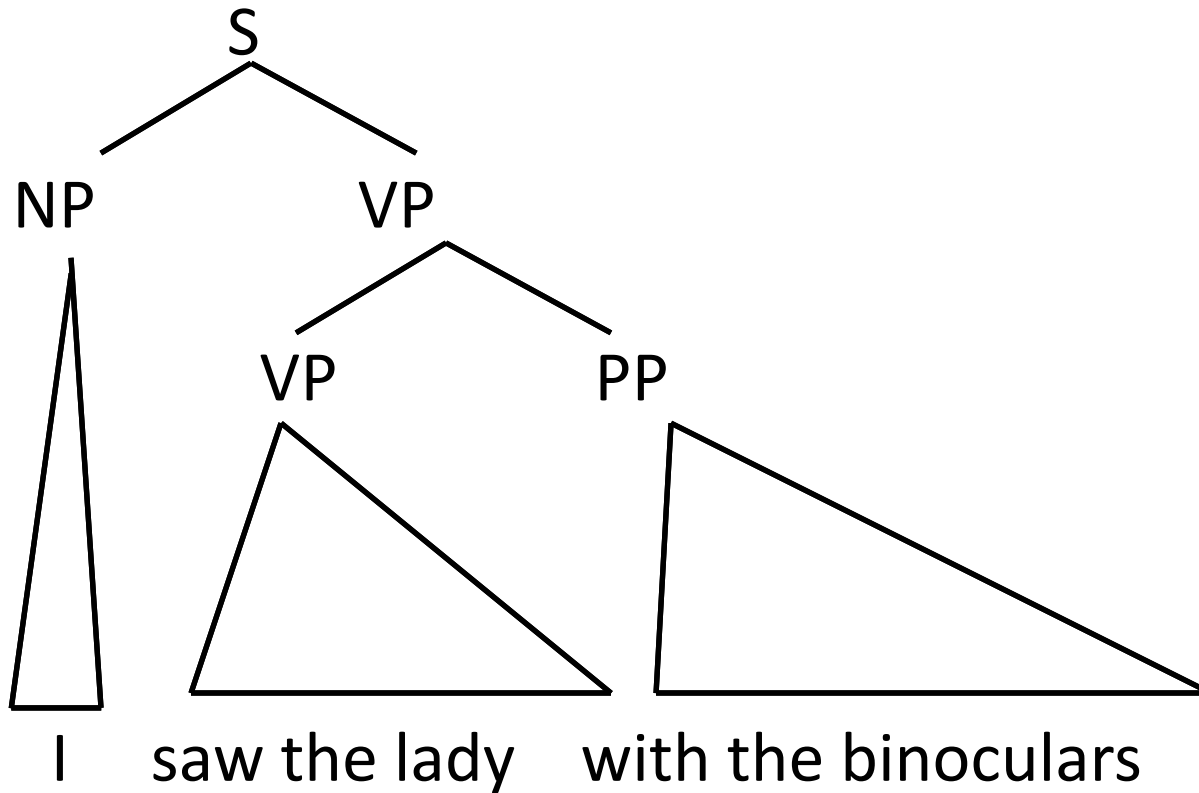


I saw the lady with the binoculars



I saw [the lady with the binoculars]

I saw the lady with the binoculars



I [ saw the lady ] [with the binoculars]

# What is a corpus?

- The word *corpus* comes from Latin (“body”) and the plural is *corpora*
- A corpus is a body of **naturally occurring language**
  - ...but rarely a random collection of text
  - Corpora “are generally assembled **with particular purposes in mind**, and are often assembled to be (informally speaking) **representative** of some language or text type.” (Leech 1992)
- “A corpus is a collection of (1) **machine-readable** (2) **authentic** texts (including transcripts of spoken data) which is (3) **sampled** to be (4) **representative** of a particular language or language variety.” (MXT 2006: 5)

# What is a corpus for?

- A corpus is made for the study of language in a broad sense
  - To test existing linguistic theory and hypotheses
  - To generate and verify new linguistic hypotheses
  - Beyond linguistics, to provide textual evidence in text-based humanities and social sciences subjects
- The purpose is reflected in a well-designed corpus

# What corpora cannot do

- Corpora do not provide negative evidence
  - Cannot tell us what is possible or not possible
  - Can show what is central and typical in language
- Corpora can yield findings but rarely provide explanations for what is observed
  - Interfacing other methodologies
- The findings based on a particular corpus only tell us what is true in that corpus
  - Generalisation vs. representativeness

# What corpora cannot do

<https://www.youtube.com/watch?v=aboZctrHfK8>

# Corpus classification

- Textual vs. Speech Corpus
- Public vs. Private Corpus
- Particular vs. Reference Corpus
  - Particular:
    - literature corpus classified by year/domain/author etc.
    - Corpus with the language of children, etc.
  - Reference:
    - Very large, covers all relevant language varieties and the common vocabulary of a language.
    - Is usually hierarchically structured in sub-corpora
    - Usually built by specialized linguistic institutions

# Corpus classification

- Diachronic corpus (language in its evolution)
- Monolingual vs. Multilingual Corpus
- Paralell vs. Comparable corpus



# Natural Language Processing

... The first Romanian nanosatellite was ... launched ...

RSA launched the first Romanian nanosatellite

– Semantics:

- Identify semantic roles around a predicate
  - X (subject) is launched by Y (object);
  - Y (subject) launches X (object).
- => concepts

# Extra reading

- [Romanian Language in the Digital Age](#)

Thank you!

[dtrandabat@info.uaic.ro](mailto:dtrandabat@info.uaic.ro)