Human Language Technologies

Diana Trandabăț

Academic year 2018-2019

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 P1,P2,…,Pn for W"
- **Disambiguation:** H infers that S intended to convey Pi (ideally Pi=P)
- **Incorporation:** H decides to believe or disbelieve Pi

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.

Understanding a text

Goliat, the first Romanian nanosatellite, was successfully launched on the orbit Monday, 13th February 2012, from a base in the Frech 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 Frech 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 Frech 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. peforms 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 Frech 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

0100011101101111011011000110100101100001011101000010110000100000011100000111001001101001011011010111010101101100001000000110011000010110111001101111011100110110000101110100011001010110110000110100101110100001000000111001001101111011010111001101100011.....

Goliat, the first Romanian nanosatellite, was successfully launched on the orbit Monday, 13th February 2012, from a base in the Frech 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 Frech 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?

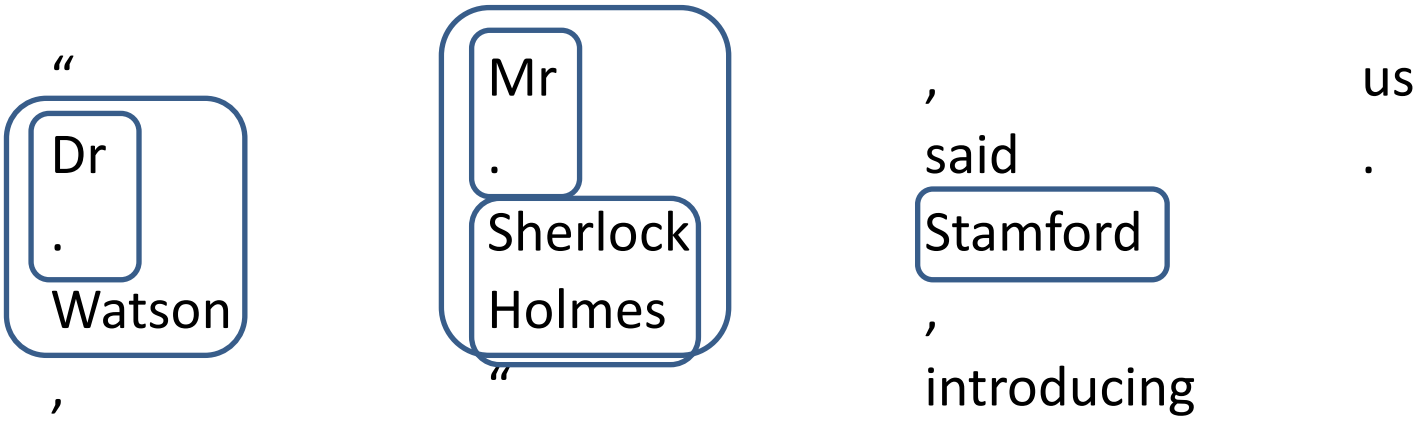
- Text statistics: no. of words, multi-word expressions, length of words/sentences, freq. of vowels/consonants > Language Identification
- Named Entity Recognition

Goliat, the first Romanian nanosatellite, was successfully launched on the orbit Monday, 13th February 2012, from a base in the Frech 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

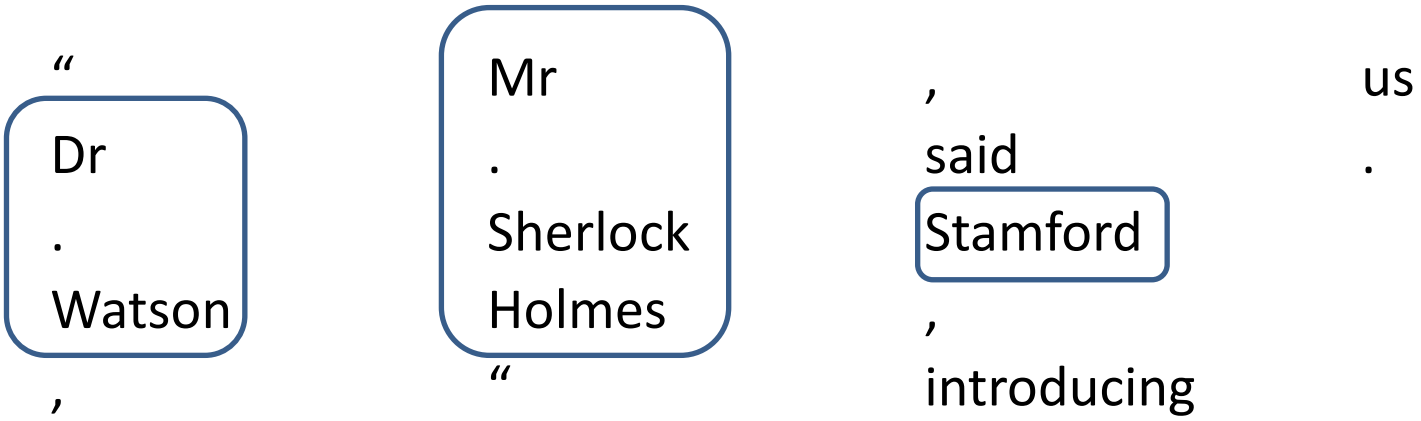
Named Entity Recognition

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



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 Frech 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

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 ⇒ stopped
  - cânta+Past+1Pl ⇒ cântaserăm/cântasem

    +2Pl ⇒ 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 3rd person singular present tense (VBP): eat
  - 3rd person singular present tense: (VBZ): eats
  - Modal (MD): should, can
  - To (TO): to (to eat)

English Parts of Speech (cont.)

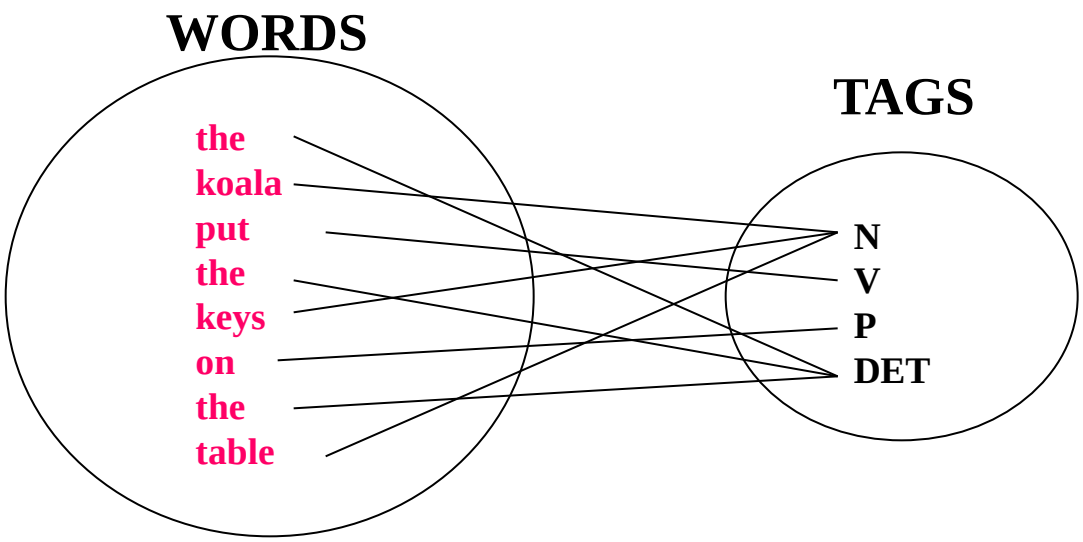
- Adjective (modify nouns)
  - Basic (JJ): red, tall
  - Comparative (JJR): redder, taller
  - Superlative (JJS): reddest, tallest
- Adverb (modify verbs)
  - Basic (RB): quickly
  - Comparative (RBR): quicker
  - Superlative (RBS): quickest
- Preposition (IN): on, in, by, to, with
- Determiner:
  -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

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

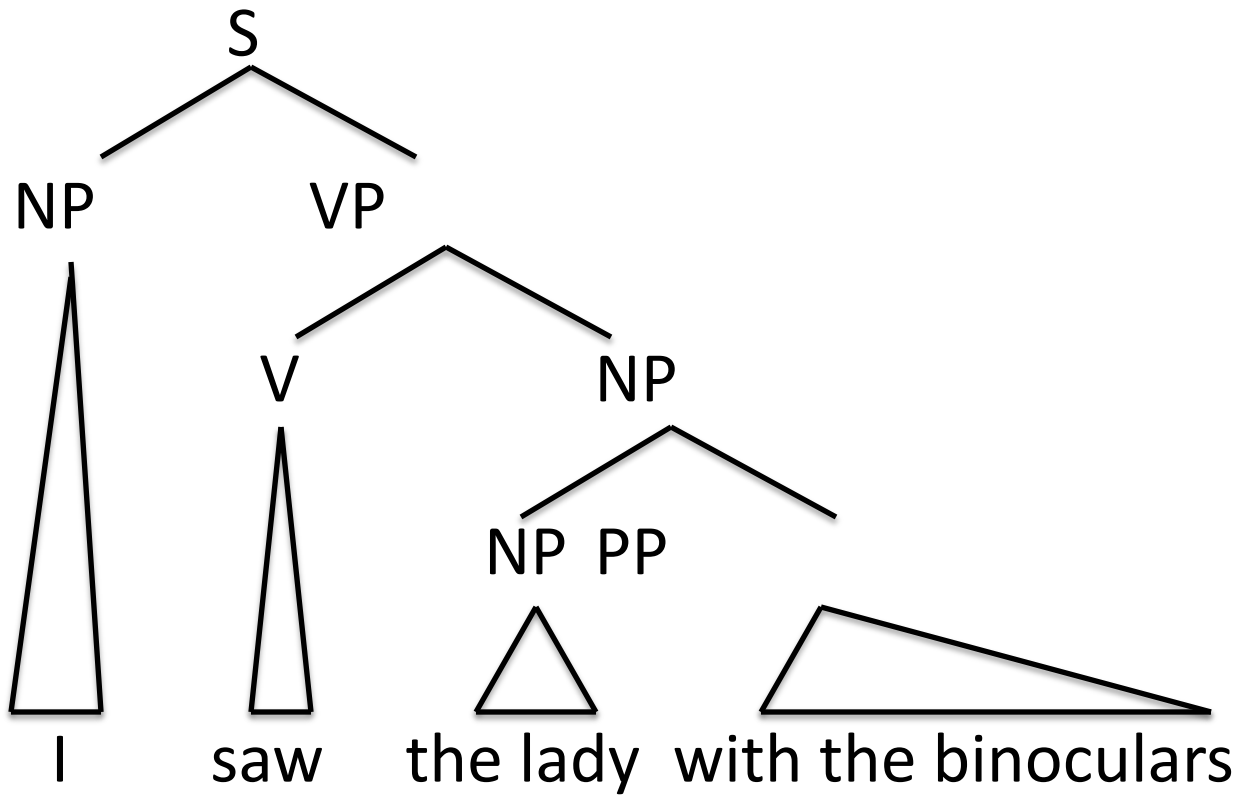
- Speech synthesis pronunciation
  - *Lead* *Lead*
  - *INsult* *inSULT*
  - *OBject* *obJECT*
  - *OVERflow* *overFLOW*
  - *DIScount* *disCOUNT*
  - *CONtent* *conTENT*
- Word Sense Disambiguation: e.g. *Time flies like an arrow*
  - Is *flies* an N or V?
- Word prediction in speech recognition
  - 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 ...

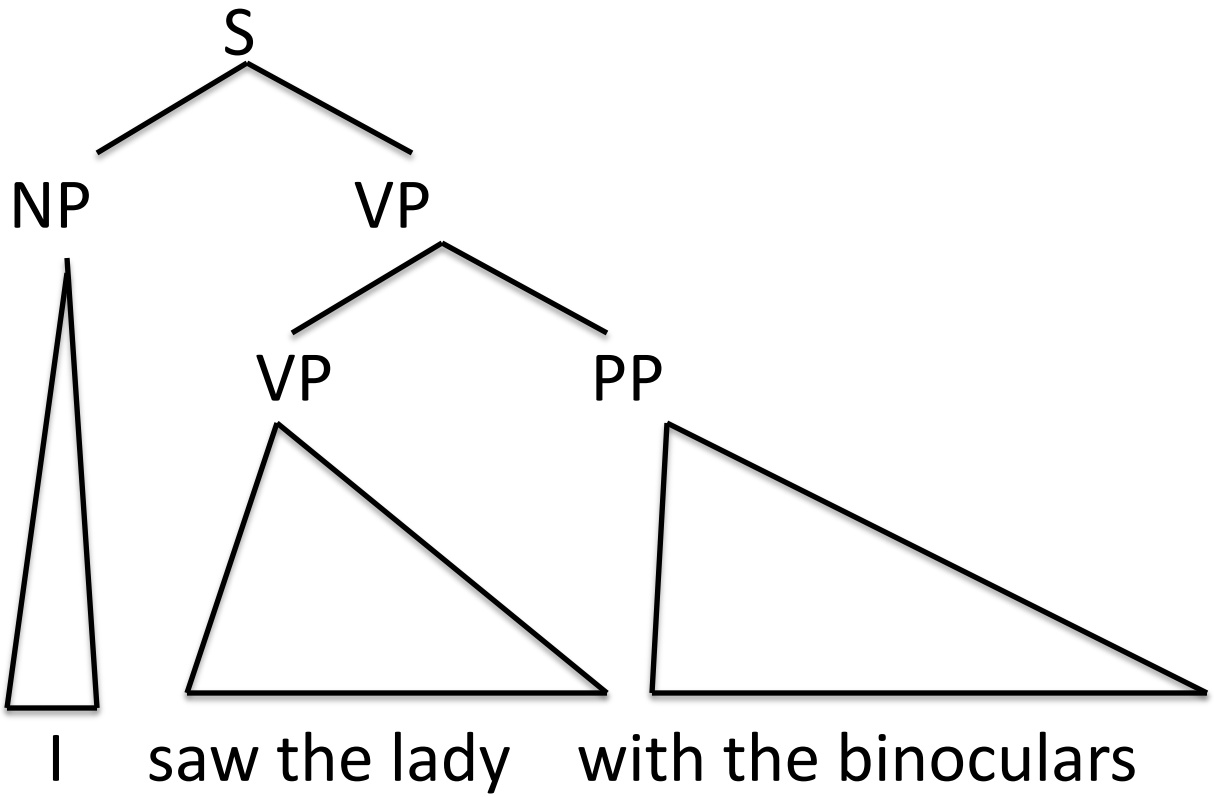
- Syntax:
  - coded in formal grammmars
    - determiner + numeral + noun+ adjectiv = noun group;
    - auxiliary + verb= verb group;
    - noun group + verb group = sentence; ...
  - Difficulty: Ambiguity
    - I saw the lady with the binoculars.

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

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 lauched the first Romanian nanosatellite

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

Which are the steps for a machine to translate?

Flowers are lovely!

Thank you!

dtrandabat@info.uaic.ro