ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

CENG 463: Introduction to Natural Language Processing
Rules, Lexicon, and Regular Expressions

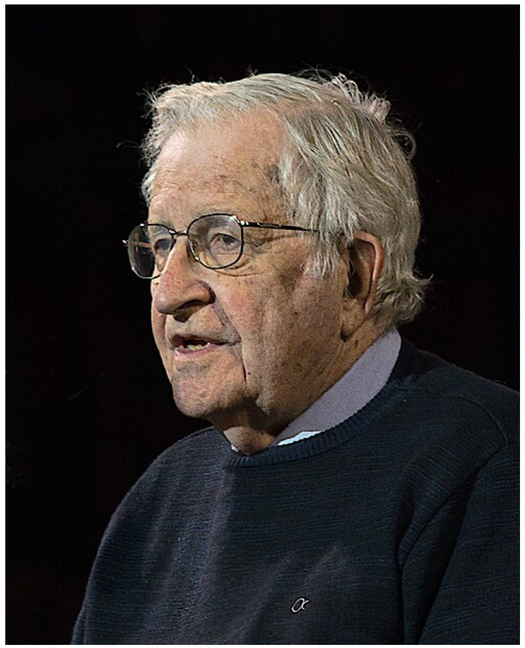
Asst. Prof. Cagri Toraman
Computer Engineering Department
ctoraman@ceng.metu.edu.tr

07.10.2025

Linguistics

Studies the relation between semantics (meaning) and syntax (form).

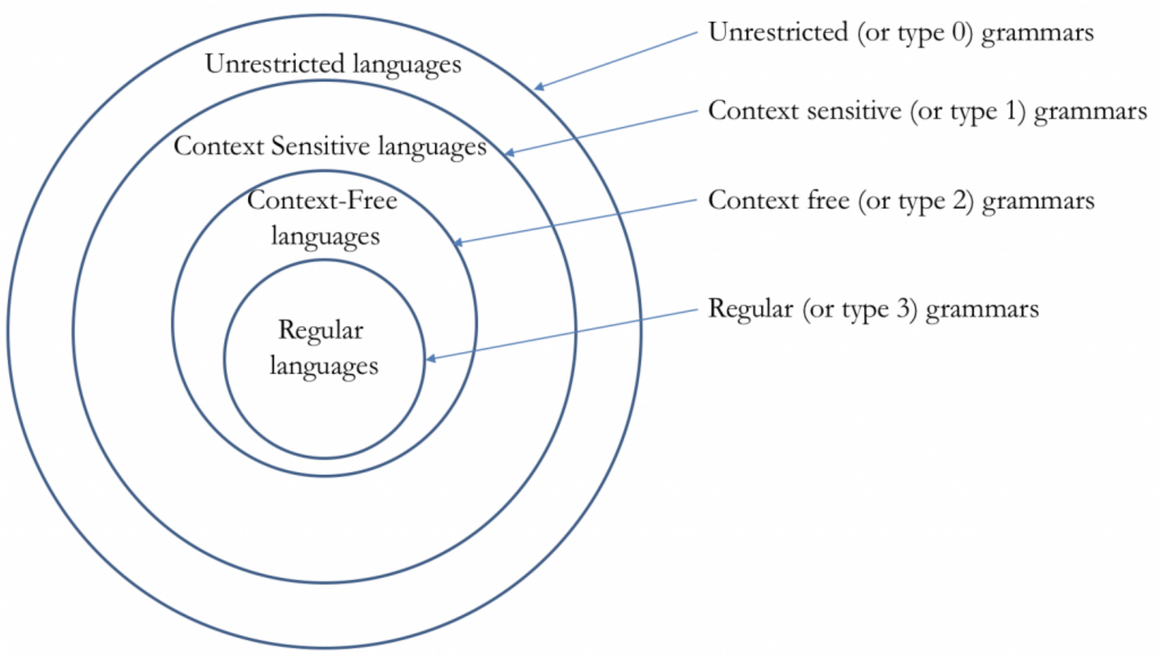
Modern Linguistics



“Noam Chomsky is widely recognized as having helped to spark the cognitive revolution in the human sciences, contributing to the development of a new cognitivistic framework for the study of language and the mind.” —Wikipedia

“Chomsky introduced the Chomsky hierarchy, generative grammar and the concept of a universal grammar, which underlies all human speech and is based in the innate structure of the mind/brain.” — From His Homepage

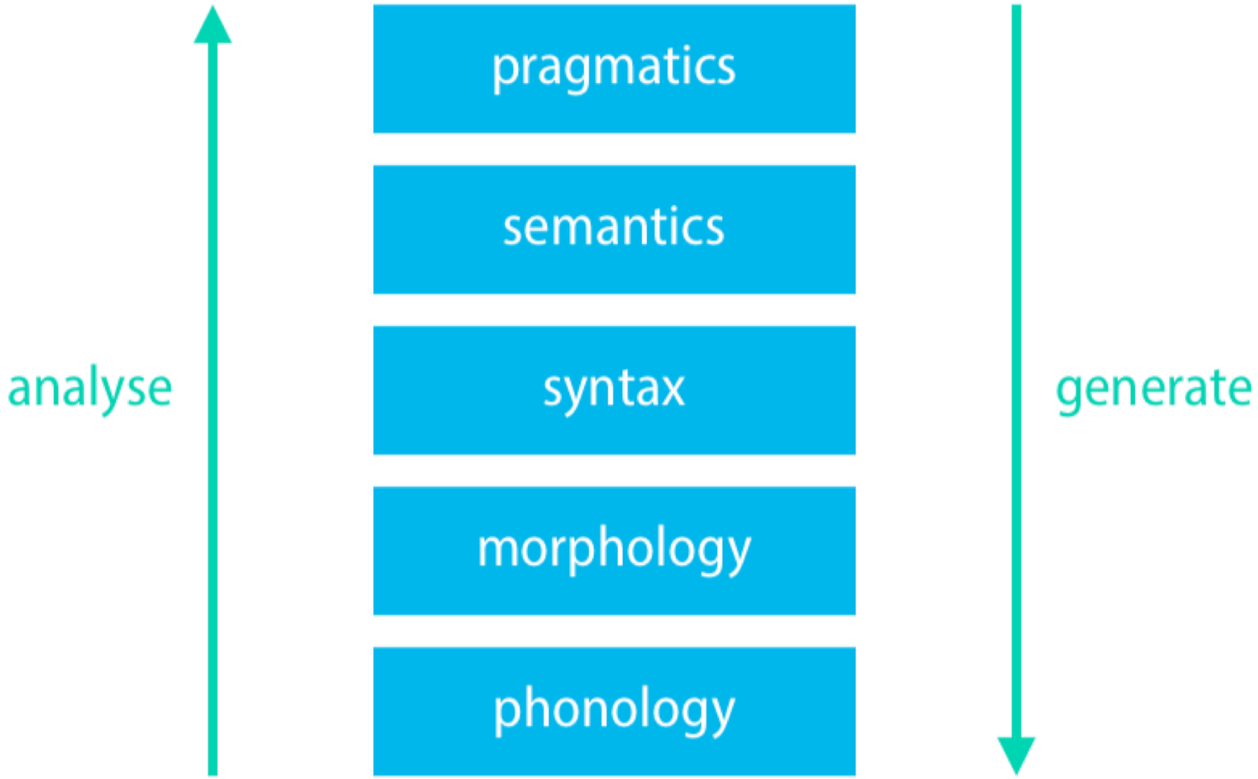
Modern Linguistics



Chomsky Grammar Classification

Grammar Type	Grammar Accepted	Language Accepted	Automaton
Type 0	Unrestricted grammar	Recursively enumerable language	Turing Machine
Type 1	Context-sensitive grammar	Context-sensitive language	Linear-bounded automaton
Type 2	Context-free grammar	Context-free language	Pushdown Automaton
Type 3	Regular grammar	Regular Language	Finite State Automaton

Research Topics in Linguistics



Phonology

Examines the rules that determine how sounds are patterned in human languages.

Example:

Pronunciation of the plural "-s" ending on nouns in English.
“Cats”, “Dogs”, “Buses”

Morphology

Examines the rules that determine how words are composed by morphemes.

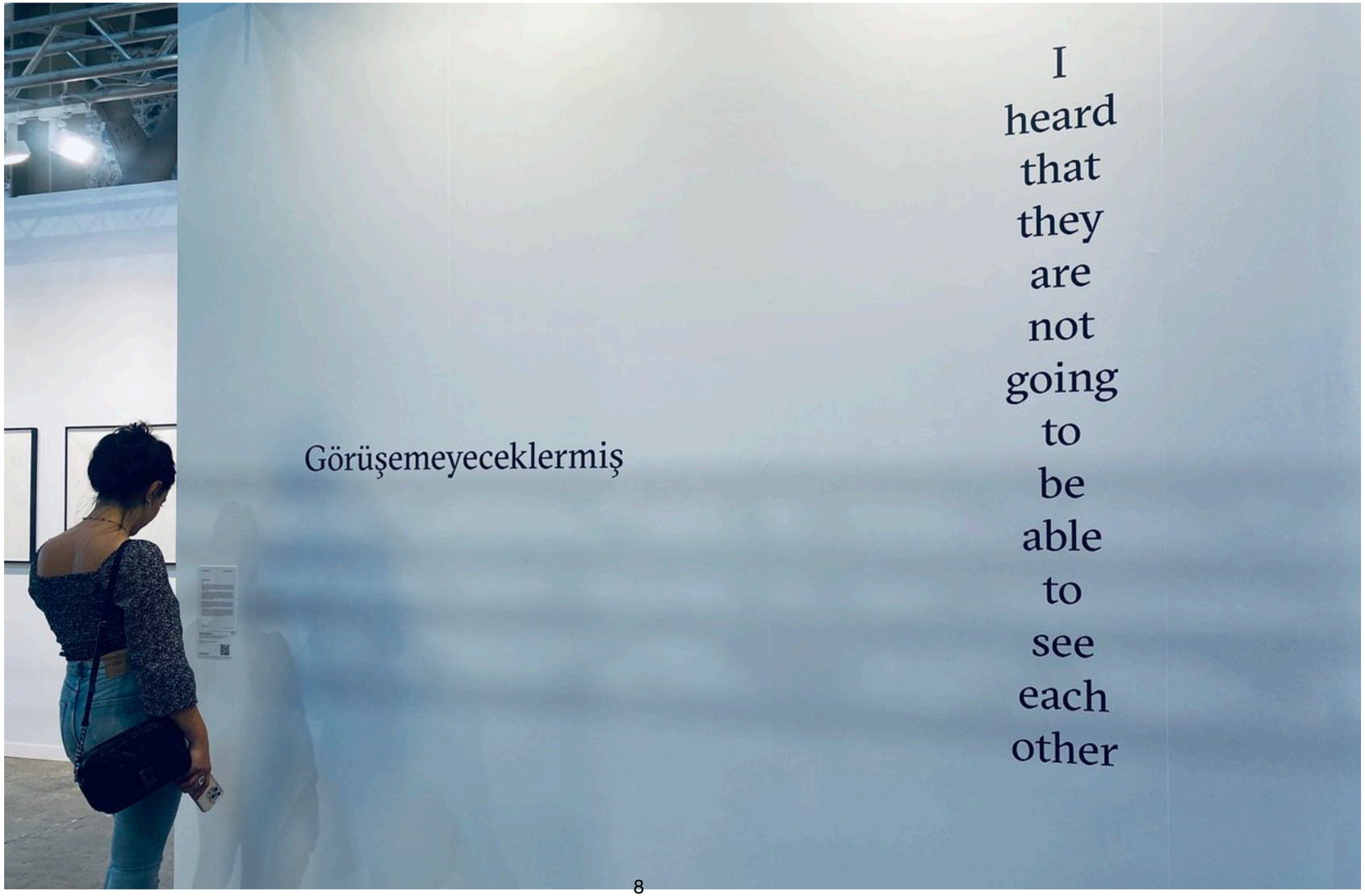
Morpheme: Smallest unit in language

Terminology: Root/stem, affix (prefix and suffix)

Example:

Agglutinative languages such as Turkish

Morphology


Görüşemeyeceklermiş
I
heard
that
they
are
not
going
to
be
able
to
see
each
other

Morphology

Lemmatization

Stemming (not linguistic stem)

Morphological analysis (suffixes, prefixes)

Example: Zemberek in Turkish

Syntax

Examines the rules that determine how words are organized into sentences.
Most sentences in English follow a Subject-Verb-Object word order.

Example:

“she only loves pizza”
“only she loves pizza”—same words, different meanings.

Syntax: Part-of-Speech

A category to which a word is assigned in accordance with its syntactic functions.

In English the main parts of speech are noun, pronoun, adjective, verb, adverb, preposition, conjunction, and interjection.

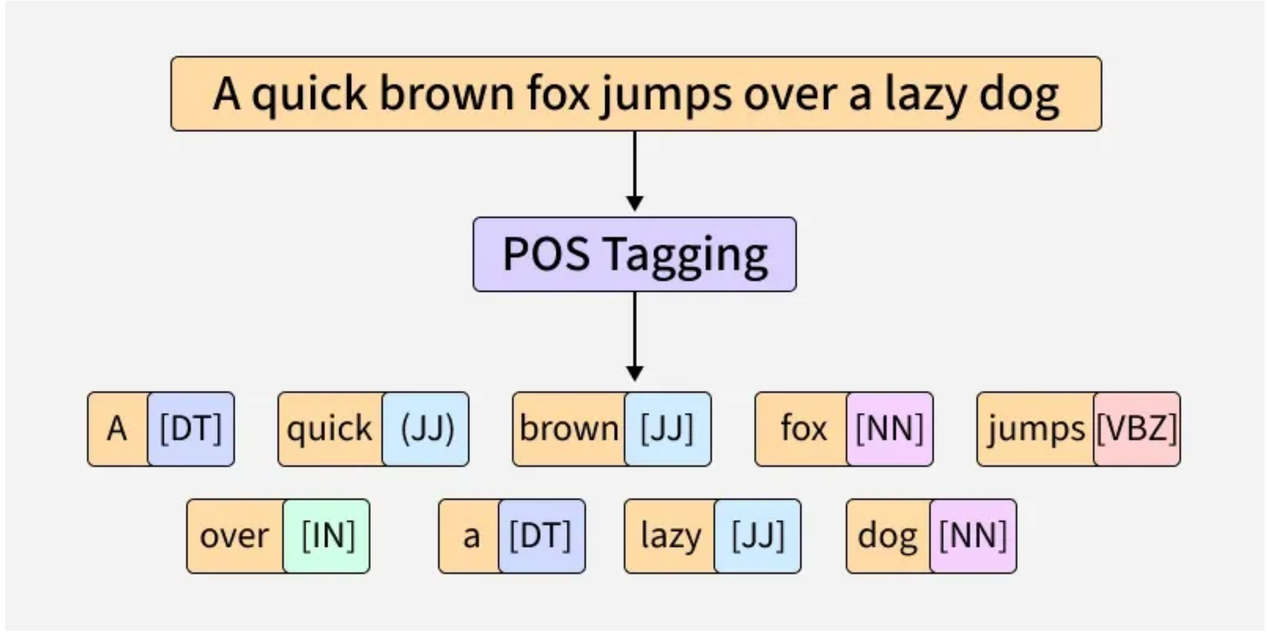
1. **Noun** Name of a person, place, thing, or idea. Dog, city, happiness	2. **Pronoun** It's a word that replaces noun He, she, they etc.
3. **Verb** It describes action that is taking place Run, write, play etc.	4. **Adverb** Adverb is a word that describes verb. quickly, now, there
5. **Adjective** It is used to describe noun. Happy, tall, blue	6. **Preposition** It is used to show relationship. In, on, under
7. **Interjection** It shows emotions. wow, ouch, oops etc.	8. **Conjunction** It joins words, phrases or ideas. And, but, or etc.

Syntax: Part-of-Speech

Content words: Noun, verb, adjective, and adverb. (New words are usually added)

Function words: Prepositions, conjunctions, etc.

Examples:



Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun

Tag	Description
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Whdeterminer
WP	Whpronoun
WP$	Possessive whpronoun
WRB	Whadverb

Syntax

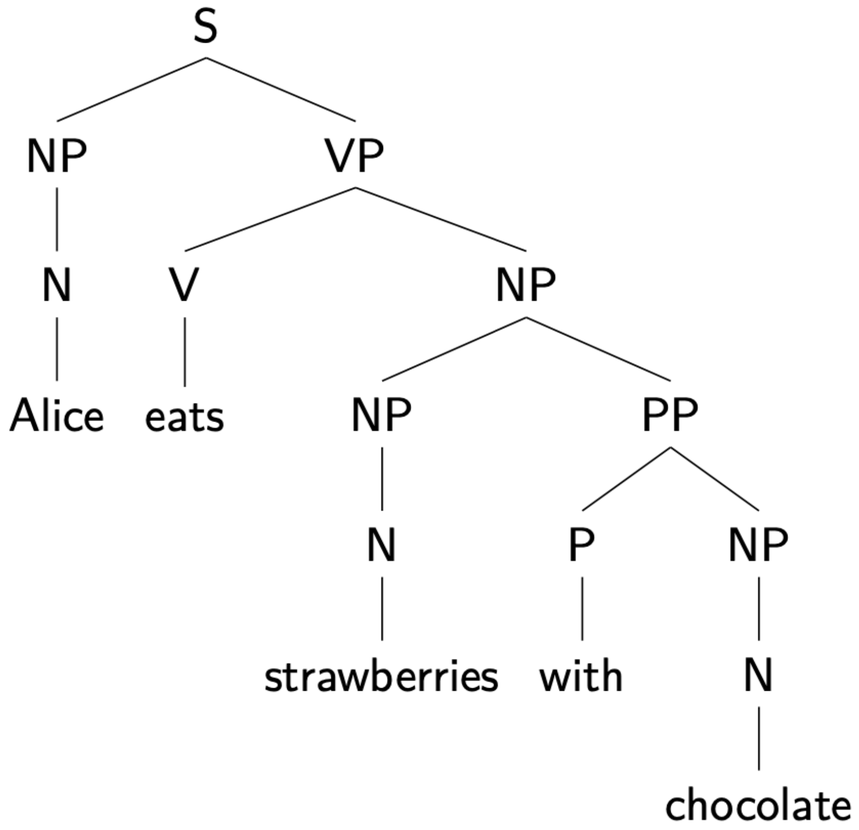
2 methods to represent syntactic structure:

Phrase structure

Dependency tree

Syntax: Phrase structure

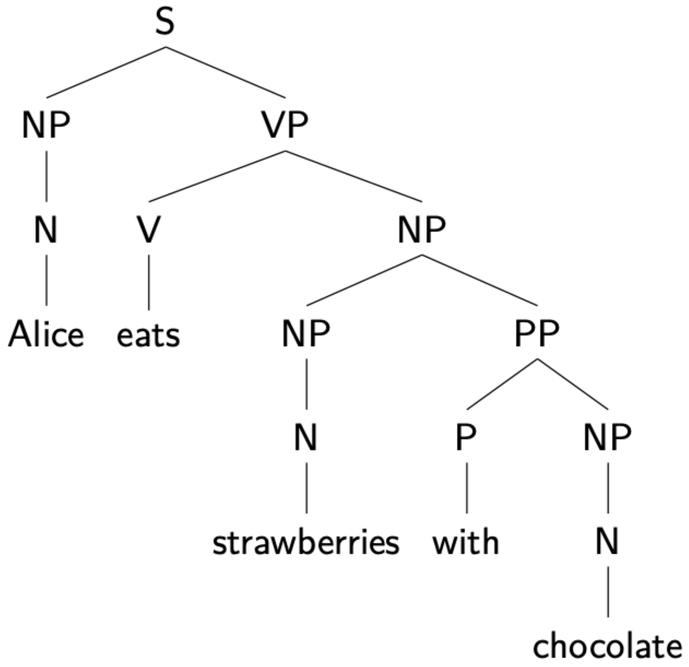
Alice/N eats/V strawberries/N with/P chocolate/N

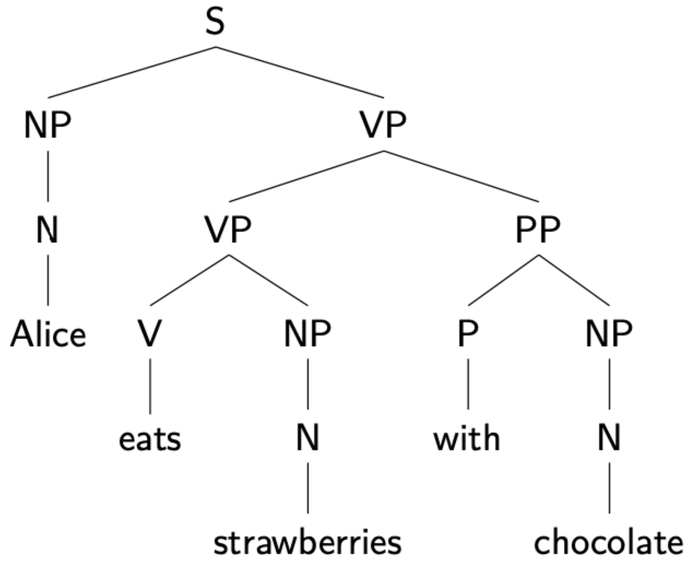


PoS tags	**Phrase tags**
—	S = Sentence
N = Noun	NP = Noun Phrase
V = Verb	VP = Verb Phrase
P = Preposition	PP = Prepositional Phrase
A = Adjective	AP = Adjectival Phrase
Det = Determiner	—
⋮	⋮

Syntax: Phrase structure

Alice eats strawberries with chocolate



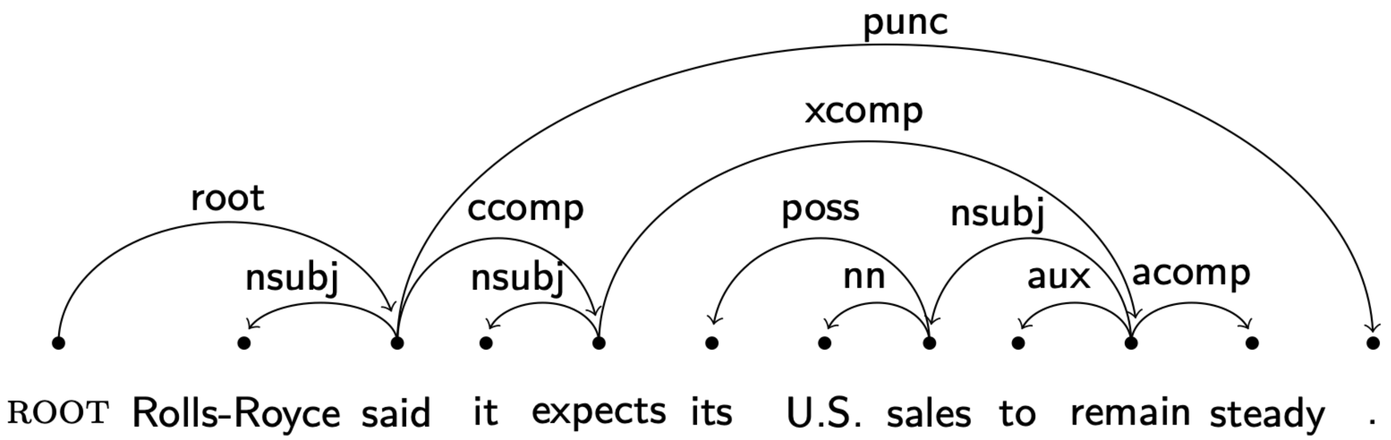


Are they both correct?

Syntax: Dependency tree

Arcs represent grammatical relations.

Rolls-Royce said it expects its U.S. sales to remain steady.



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

Semantics

Examines the meaning of linguistic components like words and sentences.

Semantics: Lexical Semantics

Examines the meaning of words.

Internal word semantics: Similarity with other words.
External word semantics: Combination with other words.

Examples:

“Oscar eats a bone”

“Oscar eats a car”

“Oscar eats a bone to Panda”

Semantics: Lexical Ambiguity

A word can have multiple meanings (word senses).

Pragmatics

Examines the meaning of linguistic expression in a given context.

Example:

“It is cold in here” (requesting to close the windows)

Lexicons

A lexicon is a structure that keeps meanings of each word.

A lexeme is the minimal unit represented in the lexicon.

A dictionary is a type of lexicon where meanings are expressed through definitions and examples.

Lexicons: WordNet

Lexical database for English, organized as a semantic network of senses.

Includes content words but not function words.

Lexemes are grouped into sets of cognitive synonyms (synset), each representing a distinct concept.

Lexicons: WordNet

A synset is a set of synonyms that define a concept or word meaning

- About half of the synsets (~54%) contains only one term, about one third (~29%) 2 terms, about 10% 3 terms
- An annotation (gloss) explaining the meaning is associated to each synset (especially to those containing a single term)
  - A synset contains ~1.75 terms in average)

2 senses of teacher

Sense 1

synset **teacher#1, instructor#1** -- (a person whose occupation is teaching)

=> educator#1, pedagogue#1, pedagog#1 -- (someone who educates young people)

Sense 2

synset **teacher#2** -- (a personified abstraction that teaches; "books were his teachers"; "experience is a demanding teacher")

=> abstraction#1, abstract#1 -- (a concept or idea not associated with any specific instance; "he loved her only in the abstract--not in person")

Lexicons: Word Sense Disambiguation

The task of selecting correct sense for a word in a given sentence.

Requires a dictionary that lists all senses of words.

I ate a cold **dish**
I washed a dirty **dish**
The *served* a cold **dish**

Lexicons: Dictionary-based Word Sense Disambiguation

Lesk Algorithm computes intersection among different meanings of words.

pine cone

PINE
1. kinds of evergreen tree with needle-shaped leaves
2. waste away through sorrow or illness

CONE
1. solid body which narrows to a point
2. something of this shape whether solid or hollow
3. fruit of certain evergreen trees

$pine_1 \cap cone_1 = 0$

$pine_1 \cap cone_2 = 0$

$pine_1 \cap cone_3 = 2$

$pine_2 \cap cone_1 = 0$

$pine_2 \cap cone_2 = 0$

$pine_2 \cap cone_3 = 0$

Lexicons: Word Similarity

Words are linked by different relations (as in WordNet).

Semantic distance between two words.

Semantic distance can be defined as the length of the minimum path leading from the first word to other.

Text Normalization: Regular Expressions

A method for cleaning text data with specialized expressions.

Most simple regular expression is a string: “Oscar”

How to extend regular expressions?

**Some slides in regular expressions are adapted from Cornell University CS 2112*

Text Normalization: Regular Expressions

The concatenation AB of two regular expressions A and B matches all strings with a first part matched by A followed by a second part matched by B.

- Regex `ab` is really just the concatenation of `a` and `b`.

The alternation $A|B$ of regexes A and B matches any string that is matched by *either* A or B.

- Regex `hello|goodbye` matches both `hello` and `goodbye`.
- Regex `d(aa|bb)c` matches both `daac` and `dbbc`.

Text Normalization: Regular Expressions

- `ab` matches only the string `ab`
- `(ab)` matches only the string `ab` (parentheses just do grouping)
- `(ab)*` matches any number of `ab`'s, including the empty string: "", "ab", "abab", etc.
  - Precedence: `ab*` matches an `a` followed by any number of `b`'s: "a", "ab", "abb", etc.
- `(ab)+` matches one or more `ab`'s. (Same as `ab(ab)*`)
- `(ab)?` matches "ab" or the empty string. (Same as `ab|`)
- `0{3,5}` matches 000, 0000, or 00000

Text Normalization: Regular Expressions

- Character classes specify a set of characters to match against: syntactic sugar for alternation.
- `[1]` is a trivial class that behaves just like "1".
- `[01]` matches 0 or 1 (but not both: same as `0|1`)
- `[01]{2}` matches 00, 11, 01, or 10
- Ranges let you match sets of consecutive characters without typing them all out:
  - `[a-z]` matches any lowercase letter, `[a-z]+` any lowercase word.
  - `[0-9]` matches any digit.

Text Normalization: Regular Expressions

- Character classes and Quantifiers mix to give useful expressions
- `[a-z]*` matches any number of consecutive lowercase characters
- `[0-9]+` matches all numbers
- `[0-9]{3}` matches all three digit numbers
- `[A-z]{4}` matches all four letter words

Text Normalization: Regular Expressions

- The ^ character beginning a character class is the logical negation operator
- `[^0]` matches any character but 0
- `[^abc]` matches any character but abc
- `[^a-z]` matches any character but lowercase letters

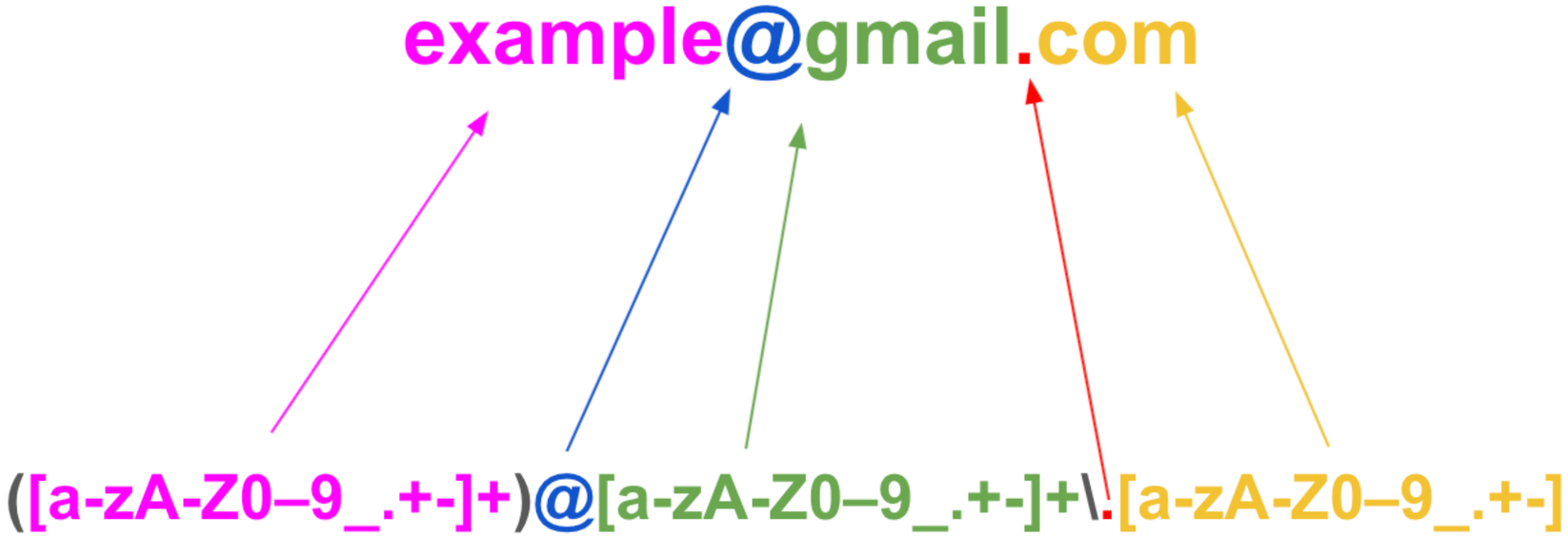
Text Normalization: Regular Expressions

- Predefined character classes are shorthand for commonly used character classes
- In most cases the capital letter is the negation of the lowercase
- `\d` = `[0123456789]`, `\D` = `[^0123456789]`
- `\s` matches white space (\t, \n, \r, etc.)
- `\w` matches "word" characters, basically not whitespace and punctuation.
- `.` matches anything but a newline. This is super useful.

Text Normalization: Regular Expressions

- Groups allow a section of the expression to be remembered for later
- \n matches the substring captured by the n^{th} capture group.
- (\d):\1 matches 1:1 or 7:7 but not 2:3
- (0|1) matches 0 or 1
- (0|1):\1 matches 1:1 or 0:0 but not 0:1
- (10) matches the string 10 but not 1 or 0 alone

Text Normalization: Regular Expressions



ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

Thanks for your participation!

Çağrı Toraman
07.10.2025