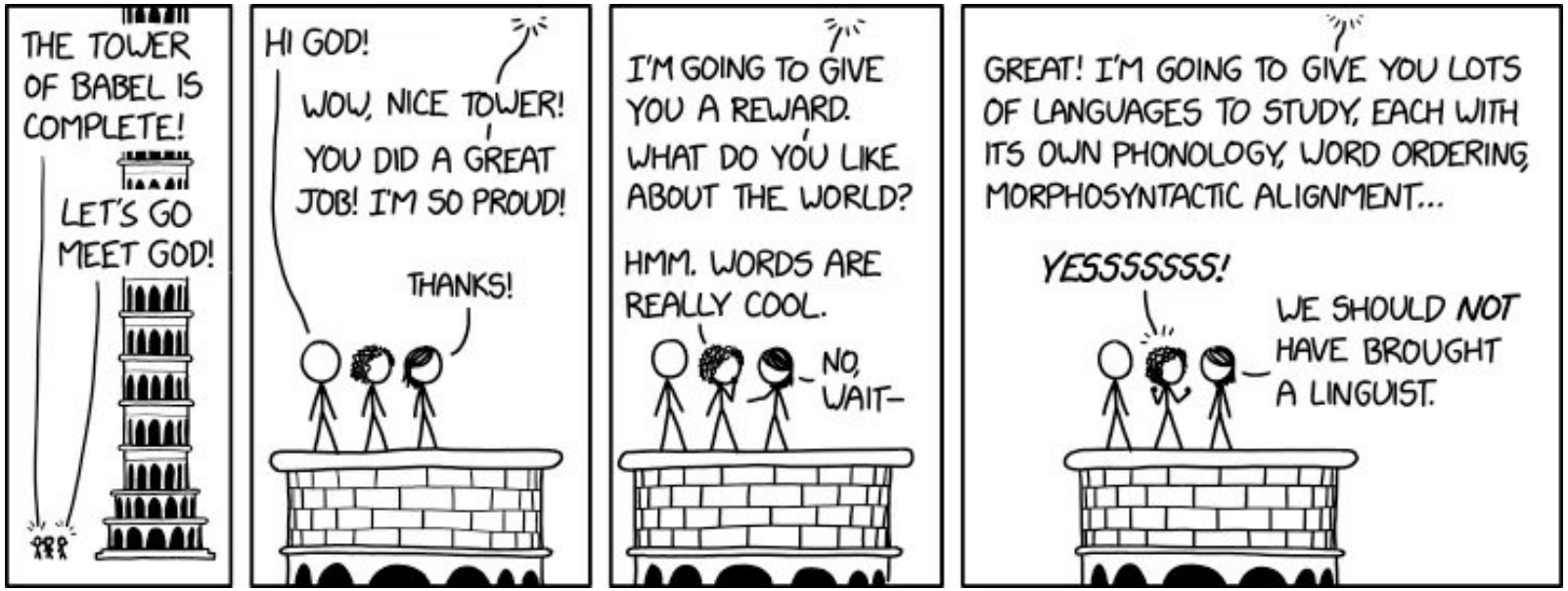
Natural Language Processing

Lecture 2 : Essentials of Linguistics

Master Degree in Computer Engineering
University of Padua
Lecturer : Giorgio Satta

Lecture based on material originally developed by :
Marco Kuhlman, Linköping University

Linguistics



What is natural language?

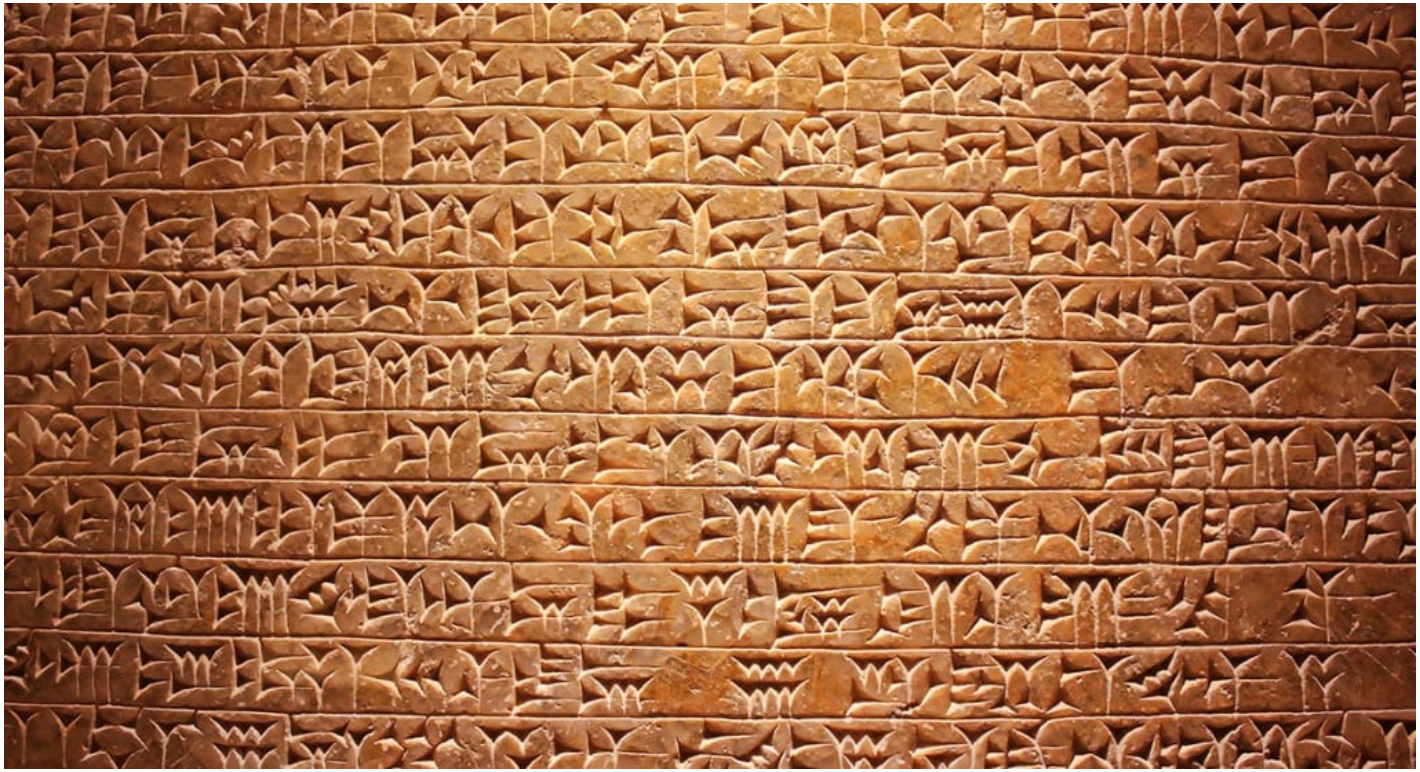


Image: Sirotinin Maksim/Shutterstock

What is natural language?

Natural language is a **structured** system for communication, consisting of a vocabulary and a grammar.

> *Language is, at its core, a system that is both digital (discrete) and infinite. To my knowledge there is no other biological system with these properties.*
>
> *Noam Chomsky*
> *Linguistics and Cognitive Science: Problems and Mysteries, 1991*

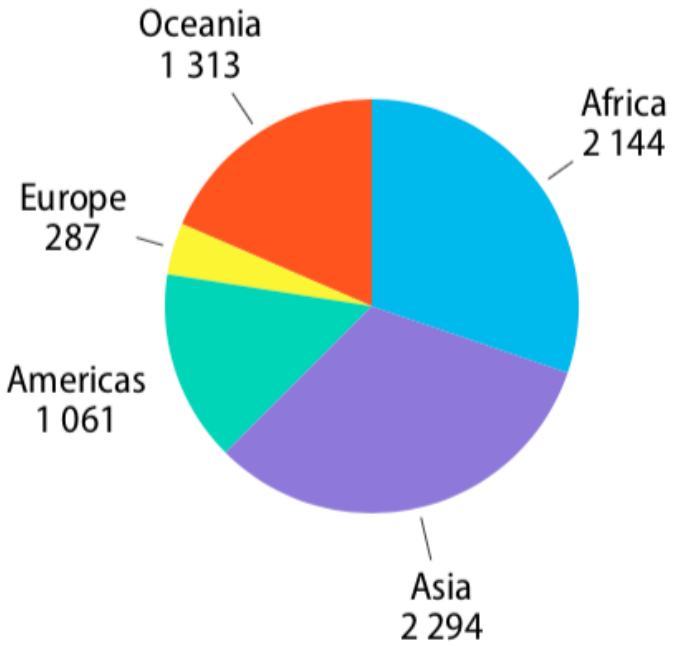
What is natural language?

Natural language is unique to *homo sapiens*, and originated between 50,000 and 100,000 years ago.

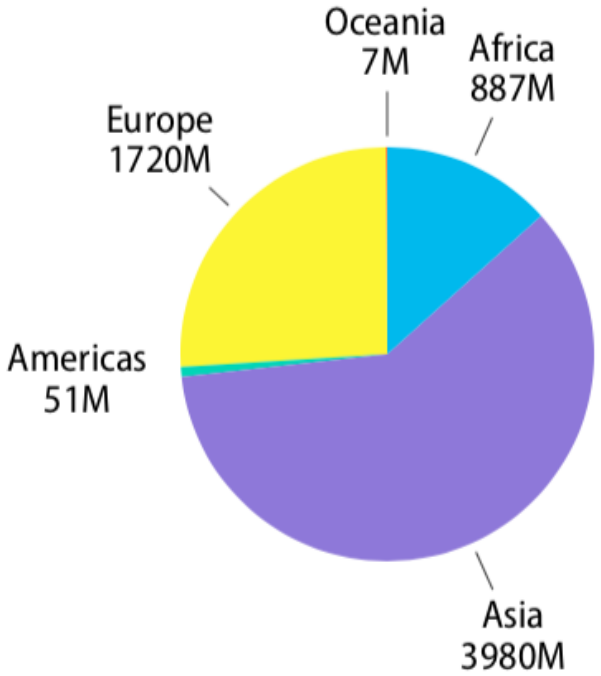
A variation of the FOXP2 gene seems related to natural language, but these studies are still very controversial.

The economist Jeffrey Sachs considers the appearance of natural language as the most important of the revolutions that have accompanied the development of humankind.

Languages of the world



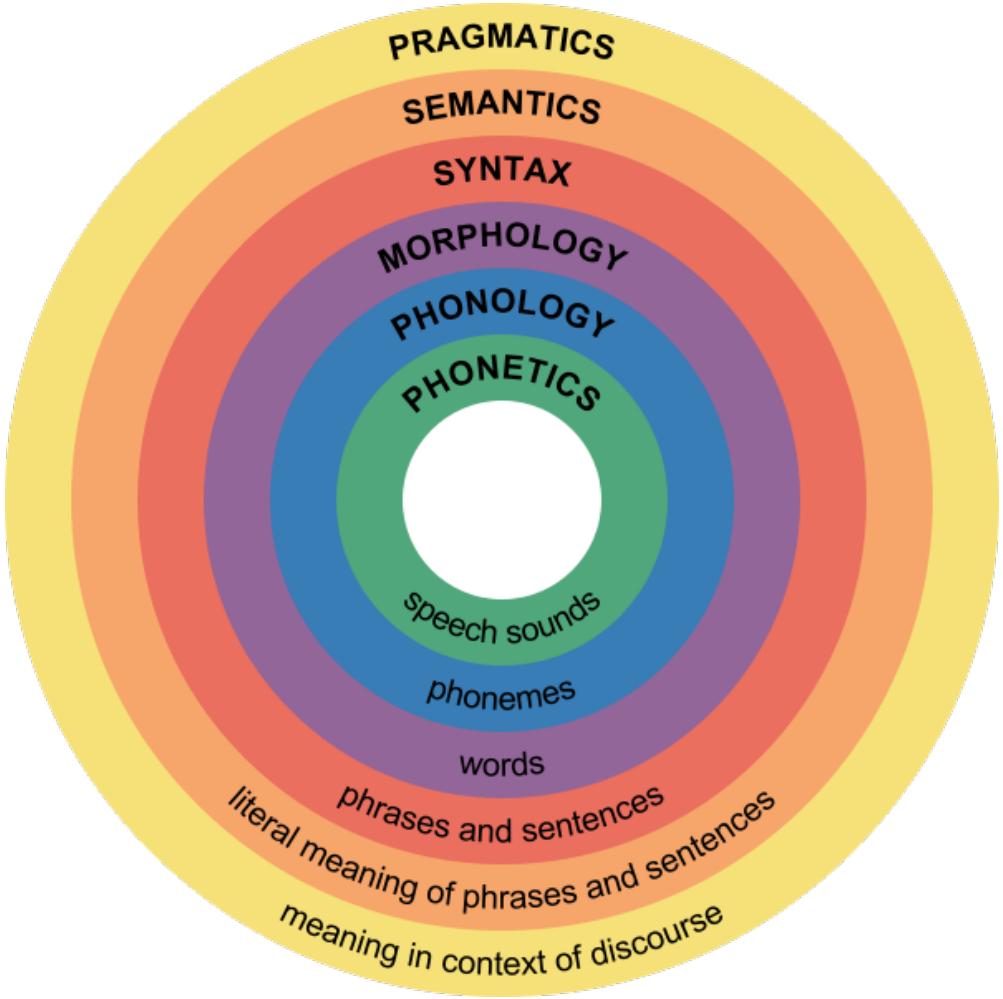
Languages by region of origin



Population by region of origin

Data elaborated from Ethnologue

What is linguistics?



https://commons.wikimedia.org/wiki/File:Major_levels_of_linguistic_structure.svg

What is linguistics?

Linguistics is the **scientific study** of language, and in particular the relationship between language form and language meaning.

Besides form and meaning, another important subject of study for linguistics is how language is **used in context**.

The earliest activities in the documentation of language have been attributed to the 6th-century-BC Indian grammarian Pāṇini, who wrote a formal description of the Sanskrit language.

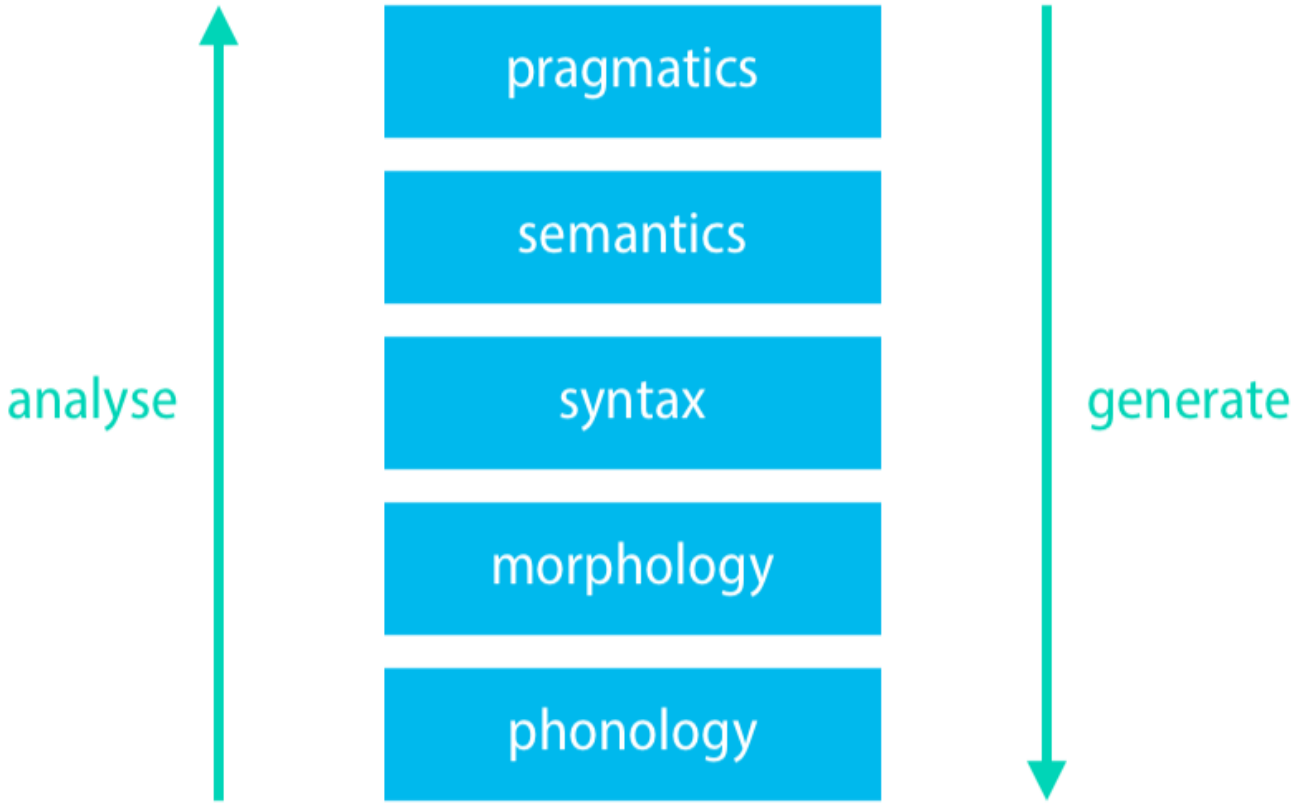
What is linguistics?

Noam Chomsky, sometimes called "the father of modern linguistics", is an American scientist who has started the development of a new framework for the study of language, called **generative linguistics**.

The kind of structures that are studied in generative linguistics seem to be **universal** across languages.

Mention notions of universal grammar and parameters.

Levels of linguistic description



We broadly overview the individual areas above. More detailed introduction will be presented in due course.

Phonology

Phonology studies the rules that organize patterns of sounds in human languages.

Example : Japanese speakers who learn English as a second language have difficulty in hearing and producing the sounds /r/ and /l/ correctly, because in Japanese these sounds are assigned the same category.

Example: right/light, arrive/alive

Phonology is different from **phonetics**, which is concerned with the production, transmission and perception of sounds, without prior knowledge of the language being spoken.

Morphology

Morphology is the study of how words are composed by **morphemes**, which are the smallest meaningful units of language.

The structure of a word consistis of several morphemes:

- one root or stem
- zero or more affixes, such as prefixes and suffixes

Example : draw, draw+s, draw+ing+s, un+draw+able

Morphology

Inflectional morphology: no change in the grammatical category

Example : give, given, gave, gives

Derivational morphology: change in the grammatical category

Example : process, processing, processable, processor, processabilty

Morphology

Isolating language: a language in which each word form consists typically of a single morpheme. Example: classical Chinese.

Analytic language: no inflection to indicate grammatical relationships, may still contain derivational morphemes. Example: English.

Synthetic language: uses inflection or agglutination to express syntactic relationships within a sentence. Example: Turkish.

The word may describe the whole sentence, incorporating subject, object, and tense relation, etc.

The term **morphologically rich language** (MRL) refers to a language in which substantial grammatical information is expressed at word level.

Example: Arabic, Hebrew, Latin, Russian, Turkish, etc.

Morphology

What we have seen in our examples so far is **concatenative morphology**: morphemes are placed one after the other.

Some languages, as for instance semitic languages, are based on **template morphology**.

Also known as root-and-pattern morphology.

Example :
Hebrew root **l-m-d**, represented by three consonants C_1-C_2-C_3, means 'study'

pattern	form	meaning
C_1aC_2aC_3	lamad	he studied
C_1iC_2eC_3	limed	he taught
C_1uC_2aC_3	lumad	he was taught
C_1aC_2 C_3an	lamdan	scholar

Syntax

Syntax studies the rules and constraints that govern how words can be organized into sentences.

Connection with formal language theory and rewriting grammars.

Differently from formal language theory, NLP systems need to be **robust** to input that does not follow the rules of grammar.

Part of speech

A **part of speech** (PoS) is a category for words that play similar roles within the syntactic structure of a sentence.

PoS can be defined

- **distributionally**: Kim saw the {elephant, movie, mountain, error} before we did.
- **functionally**: verbs = predicates; nouns = arguments; adverbs = modify verbs, etc.

Distributional and functional methods are alternative approaches that provide the same categories.

Part of speech

Open class tags: noun, verb, adjective, and adverb. New words in the language are usually added to these classes.

Sometimes referred to as content words.

Closed class tags: determiners, prepositions, conjunctions, etc. Closed word classes rarely receive new members.

Sometimes referred to as function words, as they help to organize the components of the sentence.

PoS tags can only be assigned in syntactic context:
Example : Contrast the two POS tag assignments for **fish**
They/N can/Aux fish/V
They/N eat/V fish/N

Syntactic structure

There are several representations for syntactic structure. Most common are

- phrase structure
- dependency tree

Phrase structure

A **phrase structure** is a tree-like representation with

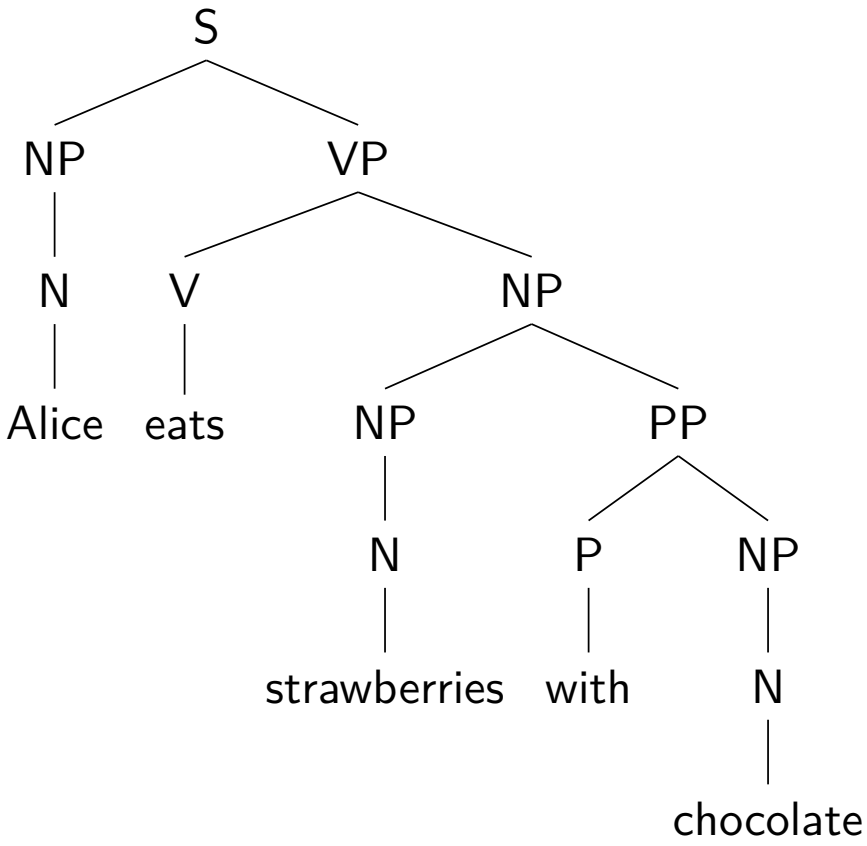
- leaf nodes representing sentence words
- internal nodes representing word groupings called **phrases**

Phrase structures use specialized labels

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	—
⋮	⋮

Phrase structure

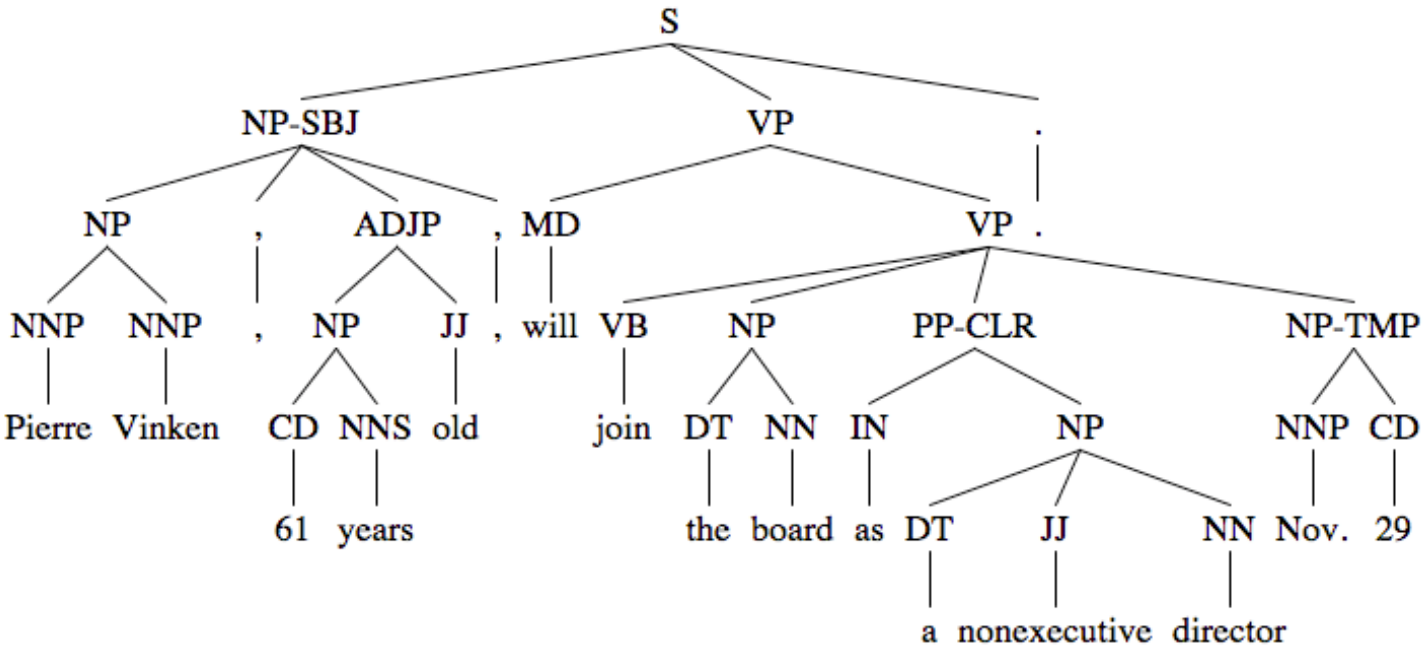
Example : Alice/N eats/V strawberries/N with/P chocolate/N



Phrase structure

Example : Pierre Vinken, 61/CD years/NNS old/JJ, will join the board as a nonexecutive director Nov. 29.

Penn Treebank, the first large-scale treebank.



Dependency tree

A **dependency tree** is a tree-like representation where

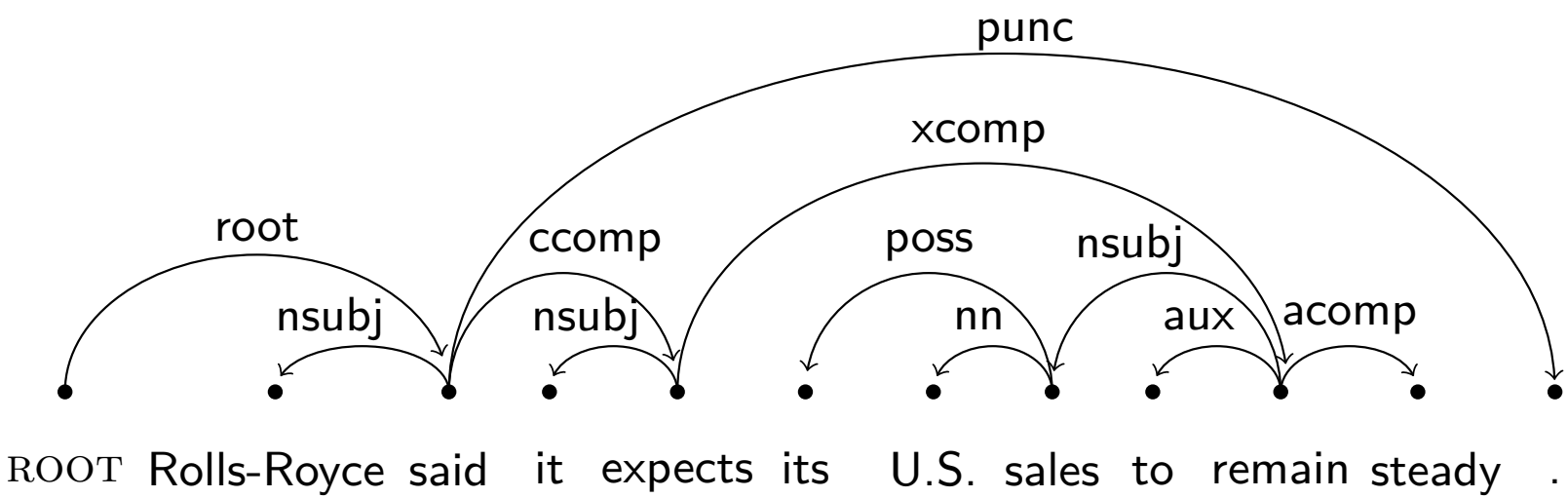
- nodes represent words (and punctuation) in the sentence
- arcs represent grammatical relations between a **head** and a **dependent**

Dependency trees use labels at arcs, representing **grammatical relations**: SBJ, OBJ, COMP, etc.

Arcs in a dependency tree are often called **dependencies**.

Dependency tree

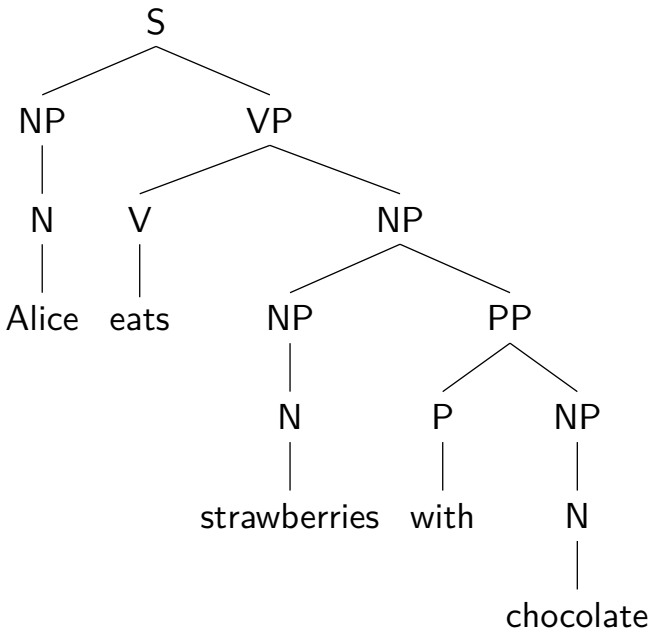
Example : Rolls-Royce said it expects its U.S. sales to remain steady.

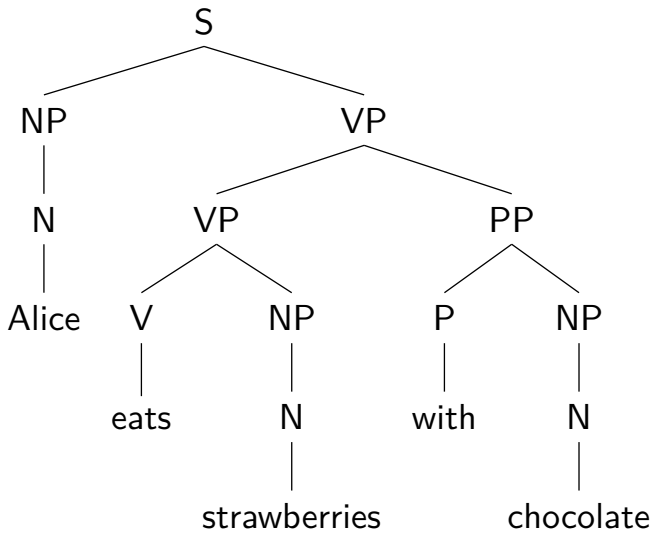


Syntactic ambiguity

A sentence can be assigned more than one syntactic structure

Example : Alice eats strawberries with chocolate





Right structure represents a wrong interpretation.

Semantics

Semantics is the study of the **meaning** of linguistic expressions such as words, phrases, and sentences.

The focus is on what expressions conventionally/abstractly mean, rather than on what they might mean in a particular context.

The latter is the focus of pragmatics.

Lexical semantics

The linguistic study of word meaning is called **lexical semantics**.

The **internal** semantic structure of a word refers to the similarity with other words.

Question: how meaningful is it to interchange this word with other words?

The **external** semantic structure of a word refers to the allowability to combine with other words.

Question: how meaningful is it to combine this word with other words?

Example : Contrast the following sentences
Alice eats an apple
?Alice eats a thunderstorm (internal violation for thunderstorm)
*Alice eats an apple to John (external violation for eat)

? marks weak ungrammaticality; * marks strong ungrammaticality.

Lexical ambiguity

Lexical ambiguity arises because a word can have different meanings, called **word senses**.

Example :
bank^1: 'financial institution'
bank^2: 'sloping mound'

Example :
plant^1: 'living organism'
plant^2: 'place where industrial or manufacturing processes are carried over'

Principle of compositionality

The meaning of a whole expression is a **function** of the meanings of its parts and of the way they are syntactically combined.

Firstly proposed in writings by the German philosopher, logician, and mathematician Friedrich Ludwig Gottlob Frege (1848–1925).

Syntax provides the scaffolding for semantic composition: the meaning of a sentence isn't just the amalgamation of the meaning of its component words.

Example : Contrast the following sentences:
the brown dog saw the striped cat
cat dog the the saw striped brown

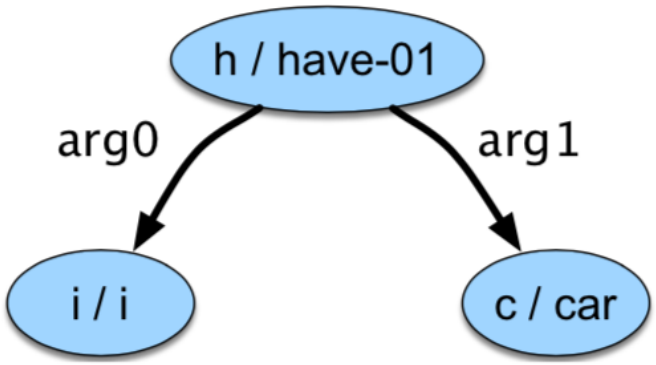
There is an analogy here with compilers for programming languages, where the translation process is driven by the syntactic structure.

Meaning representation

Many representations for semantic structure. Most common are

- logical representation
- predicate-argument structure
- graph representation

Graph representation for the sentence 'I have a car'



Pragmatics

Pragmatics studies the way linguistic expressions with their semantic meanings are used for specific **communicative** goals.

In contrast to semantics, pragmatics explicitly asks the question what an expression means in a given **context**.

An important concept in pragmatics is the **speech act**, which describes an action performed through language.

Example : **Can you pass the salt?** is a requesting act.

Discourse analysis

Discourse analysis studies written and spoken language in relation to its social context.

Discourse refers to a piece of text with multiple sub-topics and coherence relations between them, such as

- explanation
- elaboration
- contrast

Discourse analysis

Example : Contrast the two following texts

- Kim switched off the lights (and) Kim drew the blinds. The room became dark.
- Kim switched off the light. The room became dark. Kim drew the blinds.

There are many formalisms for representing discourse structure, as for example discourse representation theory and rhetorical structure theory.

Dialogue is a cooperative kind of discourse, where two or more participants are involved.

Research papers

Iñaki del Olmo from Unsplash

Research papers

Title: Linguistic Fundamentals for Natural Language Processing: 100 Esentials from Morphology and Syntax
Author: Emily M. Bender
Publisher: Morgan & Claypool Publishers, 2013
Content: The purpose of this book is to present in a succinct and accessible fashion information about the morphological and syntactic structure of human languages that can be useful in creating more linguistically sophisticated, more language-independent, and thus more successful NLP systems.

https://www.morganclaypool.com/doi/abs/10.2200/S00493ED1V01Y201303HLT020

Research papers

Title: Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics
Authors: Emily M. Bender and Alex Lascarides
Publisher: Morgan & Claypool Publishers, 2013
Content: The purpose of this book is to present a selection of useful information about semantics and pragmatics, as understood in linguistics, in a way that's accessible to and useful for NLP practitioners with minimal (or even no) prior training in linguistics.

https://www.morganclaypool.com/doi/abs/10.2200/S00935ED1V02Y201907HLT043

Research papers

Title: Q&A: What is human language, when did it evolve and why should we care?
Author: Mark Pagel
Journal: BMC Biology volume 15, Article number: 64 (2017)
Content: Language evolution shares many features with biological evolution, and this has made it useful for tracing recent human history and for studying how culture evolves among groups of people with related languages.

`https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-017-0405-3`