TRADITIONAL NATURAL LANGUAGE PROCESSING

Mehmet Can Yavuz, PhD.

Adapted from Julia Hockenmaier, NLP S2023 - course material
https://courses.grainger.illinois.edu/cs447/sp2023/

WHAT IS NATURAL LANGUAGE PROCESSING (NLP)?

Natural language processing is the set of methods for making human language accessible to computers

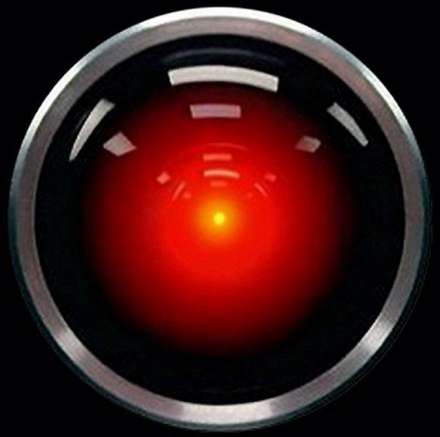
Jacob Eisenstein

Natural language processing is the field at the intersection of Computer science (Artificial intelligence) and linguistics.

Christopher Manning

EXAMPLE: CONVERSATIONAL AGENT

HAL is an artificial agent capable of such advanced language-processing behavior as speaking and understanding English, and at a crucial moment in the plot, even reading lips



2001: A Space Odyssey – HAL 9000

- Conversational agents contain:
  - Speech recognition
  - Language analysis
  - Dialogue processing
  - Information retrieval
  - Text to speech

- **David Bowman**:
- Open the pod bay doors, Hal.
- **HAL:**
- I'm sorry, Dave, I'm afraid I can't do that.
- **David Bowman:**
- What are you talking about, Hal?
- **…HAL:**
- I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.

NATURAL LANGUAGE PROCESSING VS COMPUTATIONAL LINGUISTICS

In linguistics, *language is the object of study*:

- Computational methods may be brought to bear, just as in scientific disciplines like computational biology and computational astronomy, but they play only a supporting role

In contrast, natural language processing is focused on the design and analysis of *computational algorithms and representations for processing natural human language*:

- The goal of natural language processing is to provide new computational capabilities around human language: for example, extracting information from texts, translating between languages, answering questions, holding a conversation, taking instructions

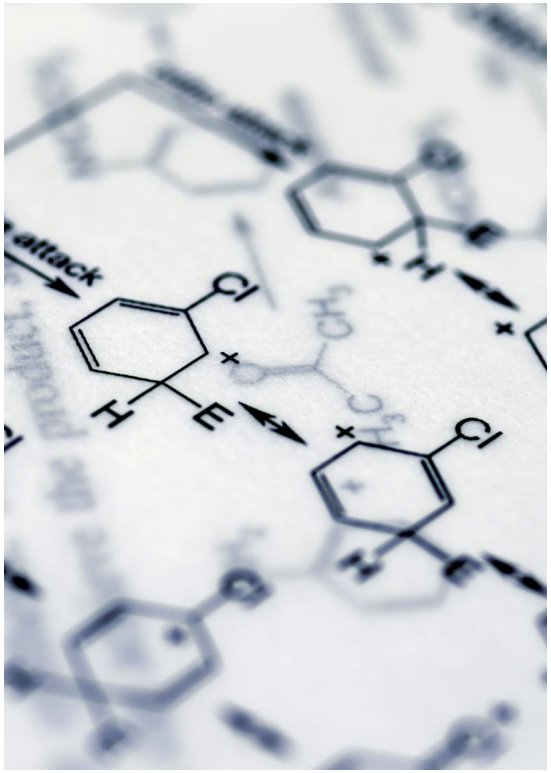
PHONETICS AND PHONOLOGY

Phonetics and Phonology: knowledge about linguistic sounds

The study of:

language sounds how they are physically formed;	systems of discrete sounds, e.g. languages' syllable structure

dis-k&-'nekt disconnect



MORPHOLOGY

Morphology: knowledge of the meaningful components of words

The study of the sub-word units of meaning

Even more necessary in some other languages,

e.g. Turkish:

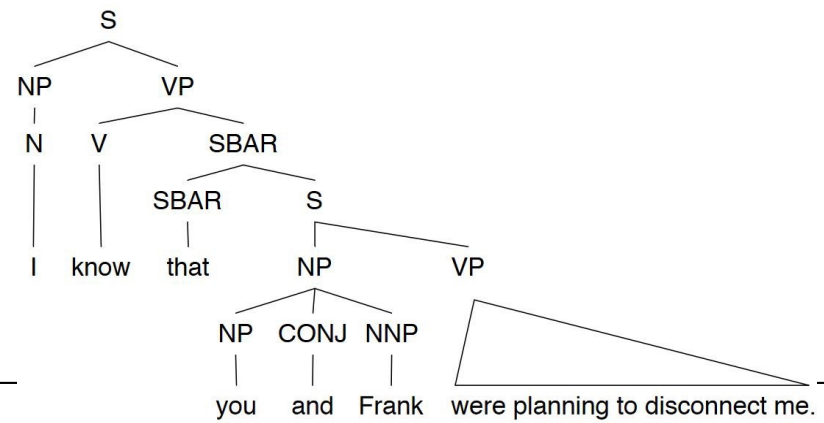
- Uygarlastiramadiklarimizdanmissinizcasina
- uygar las tir ama dik lar imiz dan mis siniz casina

SYNTAX

Syntax: knowledge of the structural relationships between words

The study of the structural relationships between words

- I know that you and Frank were planning to disconnect me.



SEMANTICS

Semantics: knowledge of meaning

The study of the literal meaning

- I know that you and Frank were planning to disconnect me.
- ACTION = disconnect
- ACTOR = you and Frank
- OBJECT = me

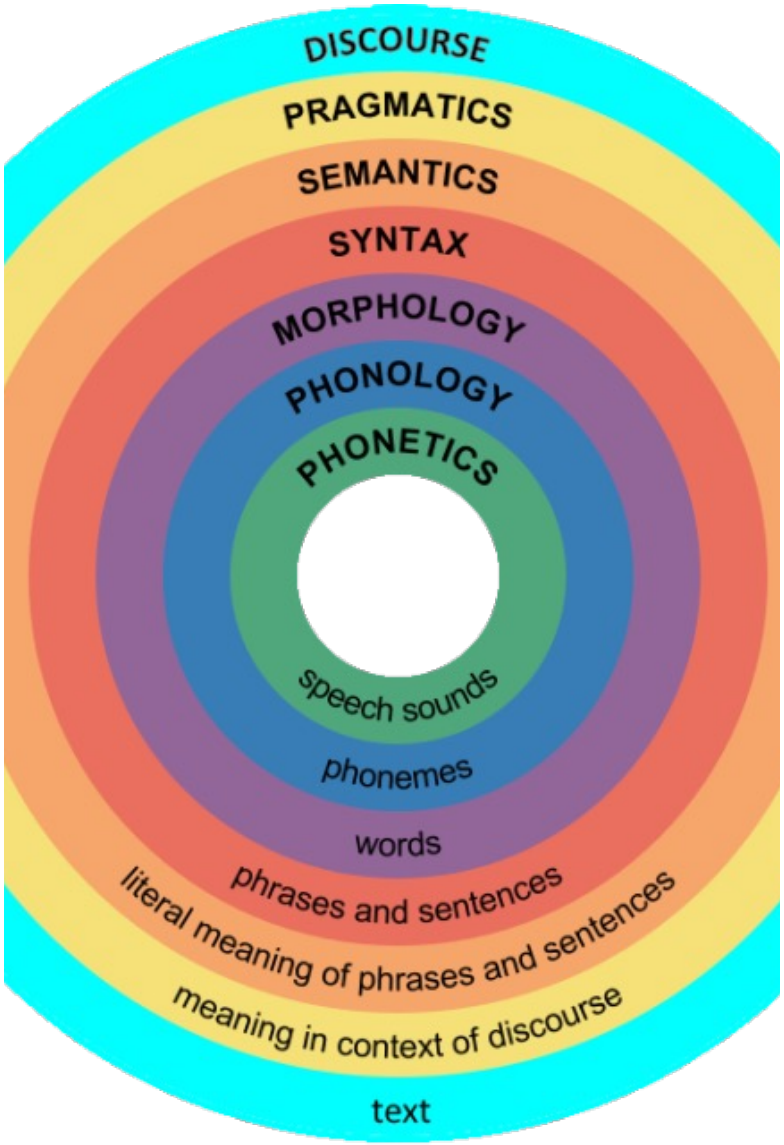
PRAGMATICS

When a diplomat says *yes*, he means ‘perhaps’;
When he says *perhaps*, he means ‘no’;
When he says *no*, he is not a diplomat.—*Voltaire* (Quoted, in Spanish, in Escandell 1993.)

Pragmatics: knowledge of the relationship of meaning to the goals and intentions of the speaker

The study of how language is used to accomplish goals

- What should you conclude from the fact I said something?
- How should you react?
  - I’m sorry Dave, I’m afraid I can’t do that.
  - Includes notions of polite and indirect styles



DISCOURSE

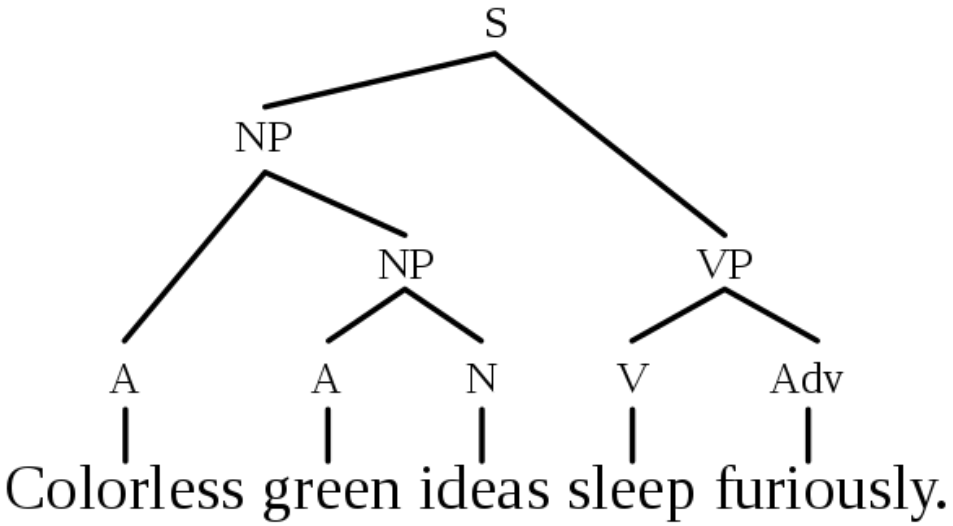
Discourse: knowledge about linguistic units larger than a single utterance

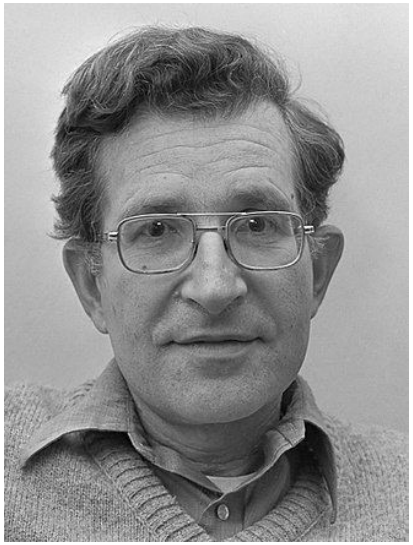
The study of linguistic units larger than a single utterance. Discourse is a generalization of the notion of a conversation to any form.

Syntax vs. Semantics

Colorless green ideas sleep furiously.

(example by Noam Chomsky 1957)





Noam Chomsky
The most cited person alive

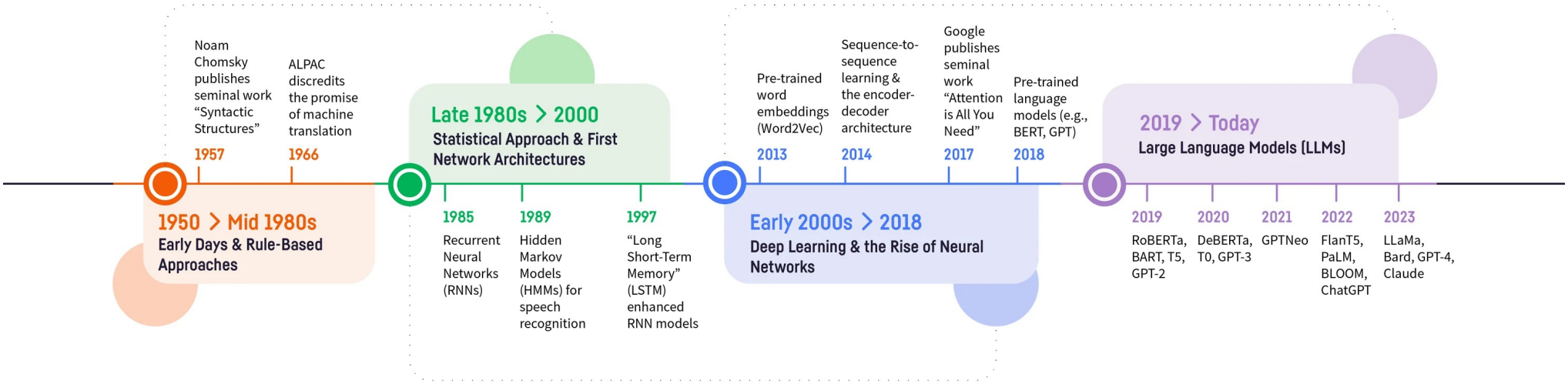
Semantics vs. Pragmatics

What does "You have a green light" mean?

- □ You are holding a green light bulb?
- □ You have a green light to cross the street?
- □ You can go ahead with your plan?

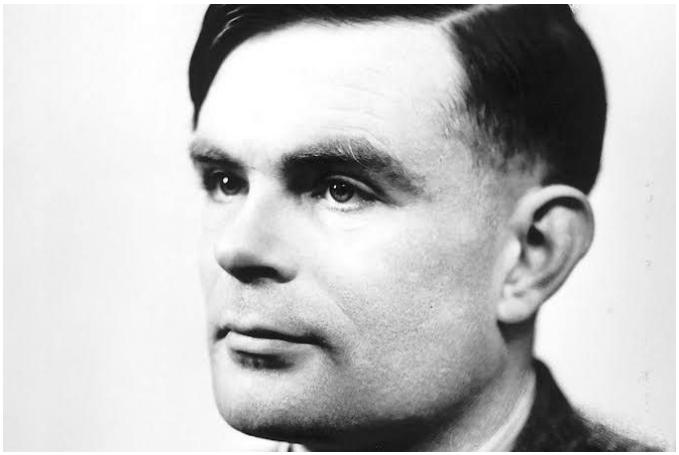
History of NLP

The History of NLP



Turing Test

"Computing Machinery and Intelligence"
Mind, Vol. 59, No. 236, pp. 433-460, 1950

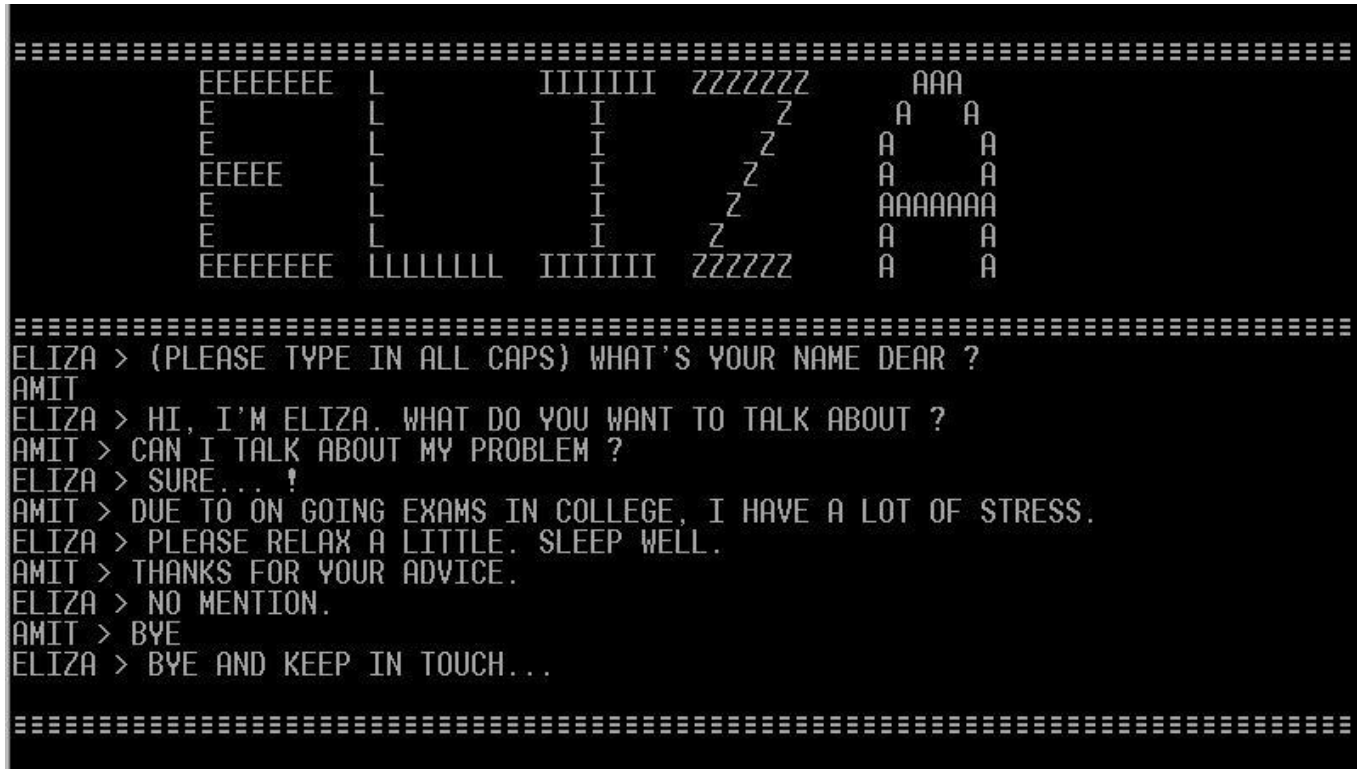


I propose to consider the question
"Can machines think?"...
We can only see a short distance ahead, but
we can see plenty there that needs to be done

In Turing's game, there are three participants: two people and a computer.
One of the people is a contestant who plays the role of an interrogator. To win, the interrogator must determine which of the other two participants is the machine by asking a series of questions via a teletype. The task of the machine is to fool the interrogator into believing it is a person by responding as a person would to the interrogator's questions. The task of the second human participant is to convince the interrogator that the other participant is the machine and that she is human.

Q: Please write me a sonnet on the topic of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: (Pause about 30 seconds and then give answer as) 105621.

ELIZA



ELIZA was an early natural language processing system capable of carrying on a limited form of conversation with a user

1950 – 1970

Mid 1950's – Mid 1960's: Birth of NLP and Linguistics

- At first, people thought NLP is easy! Researchers predicted that "machine translation" can be solved in 3 years or so
- Mostly hand-coded rules / linguistic-oriented approaches
- The 3-year project continued for 10 years, but still no good result, despite the significant amount of expenditure

Mid 1960's – Mid 1970's: A Dark Era

- After the initial hype, a dark era follows
- People started believing that machine translation is impossible, and most abandoned research for NLP

1970 – 2000

1970's and early 1980's – Slow Revival of NLP

- Some research activities revived, but the emphasis is still on linguistically oriented, working on small toy problems with weak empirical evaluation

Late 1980's and 1990's – Statistical Revolution!

- By this time, the computing power increased substantially
- Data--driven, statistical approaches with simple representation win over complex hand-coded linguistic rules
- "Whenever I fire a linguist, our machine translation performance improves." (Jelinek, 1988)

2000's – Statistics Powered by Linguistic Insights

- With more sophistication with the statistical models, richer linguistic representation starts finding a new value

Recent Years

2010's – Emergence of embedding model and deep neural networks

- Several embedding models for text using neural networks and deep neural networks were proposed including Word2Vec, Glove, fastText, Elmo, BERT, COLBERT, GTP[1-4]
- New techniques brought attention to more complex tasks