NLP Lecture Notes #1.1: An Introduction

Mehmet Can Yavuz, PhD

September 22, 2025

Introduction Natural Language Processing (NLP) is the field of study focused on making human language accessible to computers. In other words, NLP sits at the intersection of computer science (especially artificial intelligence) and linguistics, designing computational algorithms and representations to process natural human language.

Example

For example, a sophisticated conversational agent like HAL 9000 from *2001: A Space Odyssey* (Stanley Kubrick) can speak, understand English, and even read lips – demonstrating advanced NLP capabilities in understanding and generating language.

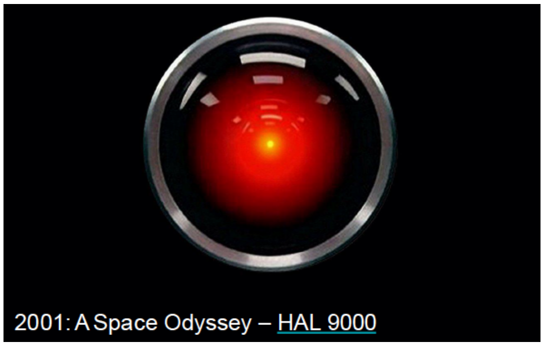


Figure 1: The ominous, glowing red "eye" of the HAL 9000, the sentient artificial intelligence that controls the spacecraft.

Figure 2: The graceful approach of the spaceplane to the rotating space station, a scene famous for its groundbreaking visual effects and depiction of future space travel.

NLP is often contrasted with computational linguistics: while linguistics treats language as an object of study (often using computational methods as support), NLP is engineering-oriented, aiming to create systems that perform useful language tasks.

A core challenge in NLP is *ambiguity in natural language*. Language is inherently ambiguous at multiple levels:

- *Structural ambiguity*: A sentence can have multiple parse trees.

  Example

  "I eat sushi **with tuna**" vs. "I eat sushi **with chopsticks**"

Example

"I saw **the man with the telescope** on the hill" vs "I saw the man with **the telescope on the hill**" (e.g., "I saw the man with the telescope on the hill" has different meanings depending on whether the phrase "with the telescope" attaches to the seeing action or to the man)

- *Lexical ambiguity*: A single word can have multiple meanings

Example

(e.g., "I went to **the bank**": "bank" can mean a financial institution or a riverbank).

- *Referential ambiguity*: Pronouns or references can be unclear

Example

(e.g., "**John** saw **Jim**. **He** was drinking coffee." – it's unclear who he refers to).

- *Others*: pragmatic ambiguity (intentions) and so on.

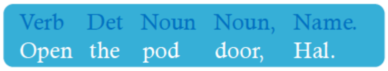
These ambiguities mean that understanding language is *non-deterministic* and *context-dependent*. Dealing with ambiguity requires algorithms to search over multiple interpretations, structured representations (like parse trees or graphs), and scoring functions (such as statistical models) to choose the most likely interpretation. In fact, statistical or machine learning models are a key tool in resolving ambiguity in modern NLP. By training on large amounts of language data, these models can learn to disambiguate words and structures based on context. However, they need to be trained (learned from data) before use, and even then they are not perfect. Another major issue is coverage: real-world text will always contain unfamiliar words or constructions. NLP systems must generalize from what they've seen in training to new, unseen inputs. This is why robust models and generalization strategies are crucial.

Key Concepts Early approaches to NLP often followed a *pipeline of sub-tasks (often called traditional NLP)* each enriching the input text with linguistic information:

Definition

- *Tokenization/Segmentation*: Splitting raw text into words (tokens) and sentences.
- *Morphological Analysis & Part-of-Speech Tagging*: Identifying the internal structure of words and assigning grammatical categories (noun, verb, etc.) to each word.
- *Syntactic Parsing*: Analyzing the grammatical structure of sentences (e.g., identifying phrases, subjects, objects).
- *Semantic Analysis*: Deriving the meaning of sentences, often by building a formal representation of meaning or identifying word senses.
- *Coreference Resolution & Discourse*: Linking pronouns and expressions to the entities they refer to, and understanding relationships across sentences.
- *Other tasks*: Word Sense Disambiguation (determining which sense of a word is intended), named entity recognition, sentiment analysis, etc., which play roles in understanding language.

Each step in this pipeline produces a symbolic representation of some aspect of language – for example, a POS tag for each word, a parse tree for each sentence, or a set of coreference links between mentions. These representations abstract away from surface text and highlight generalizations (e.g., grouping words into noun phrases, or recognizing all mentions of the same entity). Designing good representations often



Verb	Det	Noun	Noun,	Name.
Open	the	pod	door,	Hal.

Figure 3: Morphological Analysis

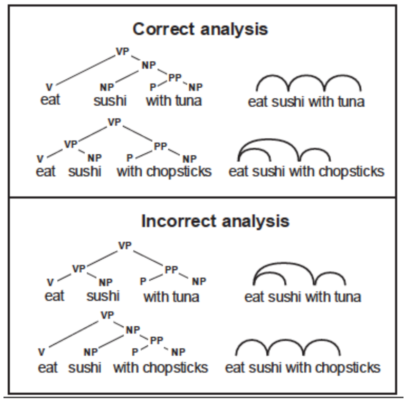


Figure 4: Syntactic Analysis

```
∃x∃y(pod_door(x) & Hal(y)
  & request(open(x, y)))
```

Figure 5: Semantic Analysis

requires linguistic insight, and each step typically relies on a learned model to predict the appropriate labels or structure. Notably, errors can cascade: if tokenization or tagging is wrong, it can negatively impact downstream parsing or semantic interpretation. Moreover, creating annotated training data for each layer (tags, parse trees, etc.) is expensive and time-consuming.

In recent years, end-to-end neural approaches have become popular, sidestepping the traditional pipeline. Instead of producing explicit intermediate linguistic representations, these models (often deep neural networks) learn to map raw text inputs directly to the desired output (for example, translating a sentence or answering a question). With sufficient training data, end-to-end models often outperform pipeline approaches, because they can jointly learn the features that are most useful for the final task. However, they raise new questions: how to integrate knowledge and reasoning, how to deal with limited data (low-resource languages), and how to interpret what the model is doing internally. The trade-off between structured, knowledge-rich approaches and purely data-driven approaches is a central theme in modern NLP development.

Another key concept is that NLP spans multiple *linguistic levels*:

Definition

- *Phonetics/Phonology*: The sounds of language (mostly relevant for speech processing).
- *Morphology*: The structure of words and meaningful subword units (morphemes).
- *Syntax*: The structure of sentences and the relationships between words (grammar).
- *Semantics*: The literal meaning of words and sentences.
- *Pragmatics*: How language is used in context to achieve goals, including implied meanings and speaker intentions.
- *Discourse*: Language units larger than a sentence – how sentences connect to form coherent discourse and conversations.

Each level introduces its own challenges, and NLP systems may handle some or all of these layers depending on the task. For instance, a machine translation system needs to handle syntax and semantics to produce correct translations, while a dialogue system must also handle pragmatics and discourse to maintain context over a conversation.

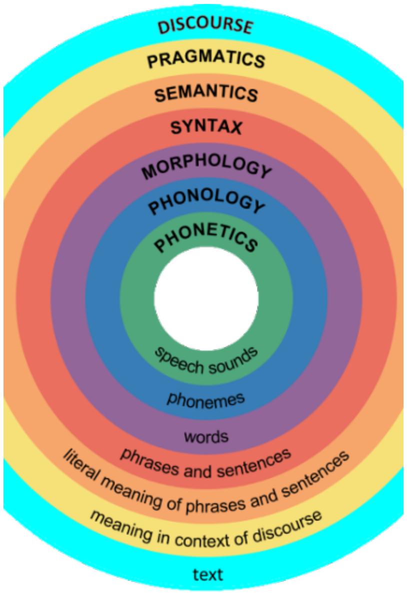


Figure 6: Linguistics Circle

Example

NLP or Not? – Which of the following scenarios involve NLP?

1. A program that converts speech audio into text transcripts.
2. A system that predicts stock prices from historical data.
3. An app that answers questions by understanding and looking up information in Wikipedia.
4. A robot that uses sensors to avoid obstacles.

Solution: (a) Yes – converting speech to text is NLP (speech recognition). (b) No – stock price prediction doesn't primarily involve language. (c) Yes – a question-answering app needs NLP to understand the question and retrieve answers. (d) No – obstacle avoidance is not about processing language.

Example

Ambiguity Identification: Consider the sentence: "The chicken is ready to eat." Identify two different interpretations of this sentence. Why does the ambiguity arise?
Solution: One interpretation is that the chicken (the animal) is ready to eat something (it's hungry). Another is that the chicken (as food) is ready to be eaten (cooked and prepared). The ambiguity arises from the structure and the word "eat": the phrase can be interpreted as the chicken being the agent about to eat, or the object that is to be eaten (this is a form of lexical/structural ambiguity, sometimes called a garden-path or amphiboly). Human listeners use context to decide which meaning is intended.

Example

Coreference Resolution: "Alice gave Joan her book after she had finished reading." In this sentence, who does "she" refer to? How could you rephrase the sentence to make it clearer?
Solution: The pronoun "she" is ambiguous – it could refer to Alice or Joan. Without additional context, it's impossible to be certain, which illustrates referential ambiguity. To make it clearer, we could rephrase: "After Alice had finished reading, she gave Joan Alice's book." or "After Joan had finished reading, Alice gave her Joan's book." depending on the intended meaning. This replaces the ambiguous pronoun with a specific name or possessive, resolving the ambiguity.