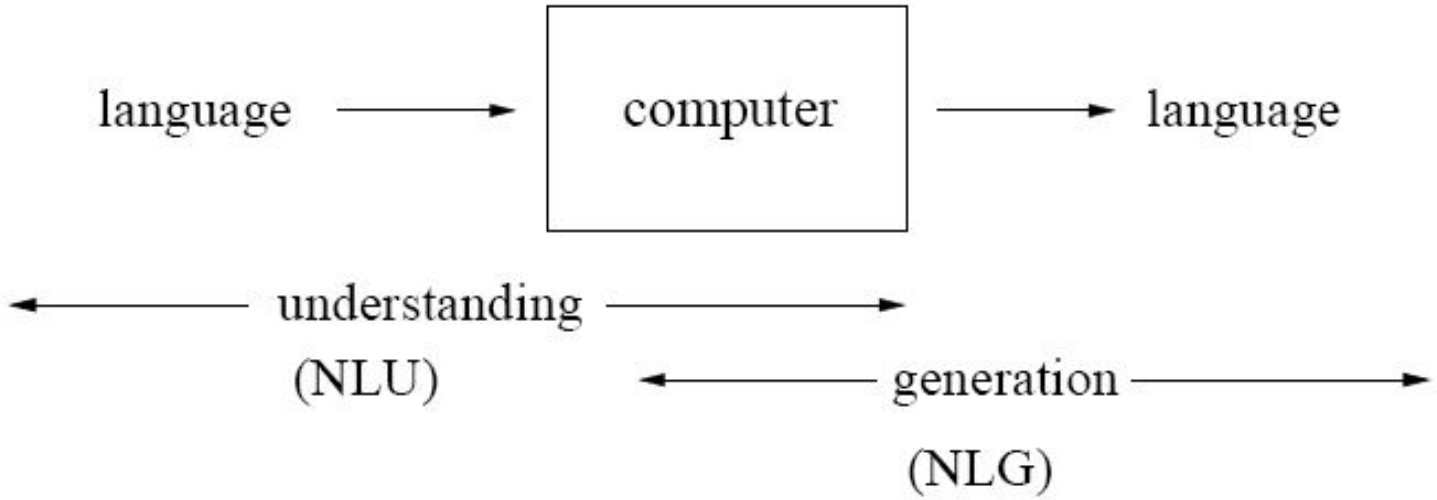
# Natural Language Processing

# What is Natural Language Processing?

computers using natural language as input and/or output




*

# NLP: applications

- Speech recognition and synthesis
- Machine translation
- Document processing
  - information extraction
  - summarization
- Text generation
- Dialog systems (typed and spoken)

# Levels of language analysis

- Phonology: What words (or sub words) are we dealing with?
- Morphology: How words are constructed from more basic meaning units?
- Syntax: What phrases are we dealing with?
- Semantics: What's the context-free meaning?
- Pragmatics: What is the more exact (context-dependent) meaning?
- Discourse Knowledge: how the immediately preceding sentences affect the interpretation of the next sentence?
- World knowledge: Using general knowledge about the world

# Levels of language analysis

- Phonetics: sounds -> words
  - /b/ + /o/ + /t/ = boat
- Morphology: morphemes -> words
  - friend + ly = friendly
- Syntax: word sequence -> sentence structure
- Semantics: sentence structure + word meaning -> sentence meaning
- Pragmatics: sentence meaning + context -> more precise meaning
- Discourse and world knowledge

# Levels of language analysis (cont.)

1. Language is one of fundamental aspects of human behavior and is crucial component of our lives.
2. Green frogs have large noses.
3. Green ideas have large noses.
4. Large have green ideas nose.
5. I go store.

# Why is NLP Hard?

“At last, a computer that understands you like your mother”

# Ambiguity

- "At last, a computer that understands you like your mother"

1. (*) It understands you as well as your mother understands you
2. It understands (that) you like your mother
3. It understands you as well as it understands your mother

- 1 and 3: Does this mean well, or poorly?

# Ambiguity at Many Levels

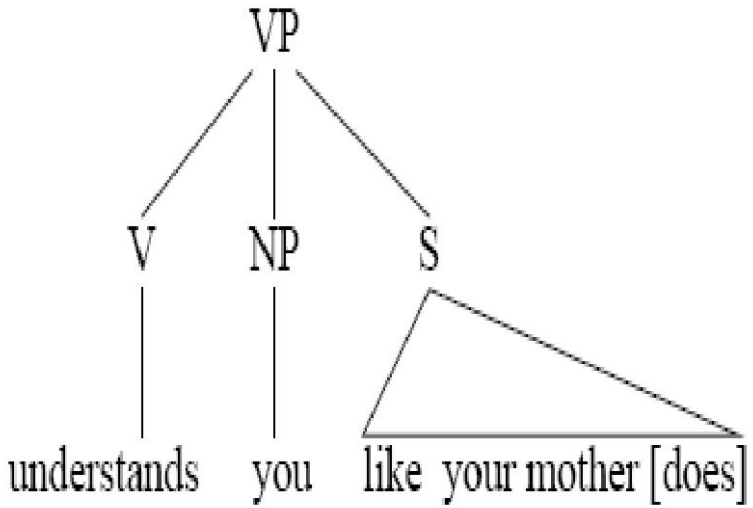
At the acoustic level (speech recognition):

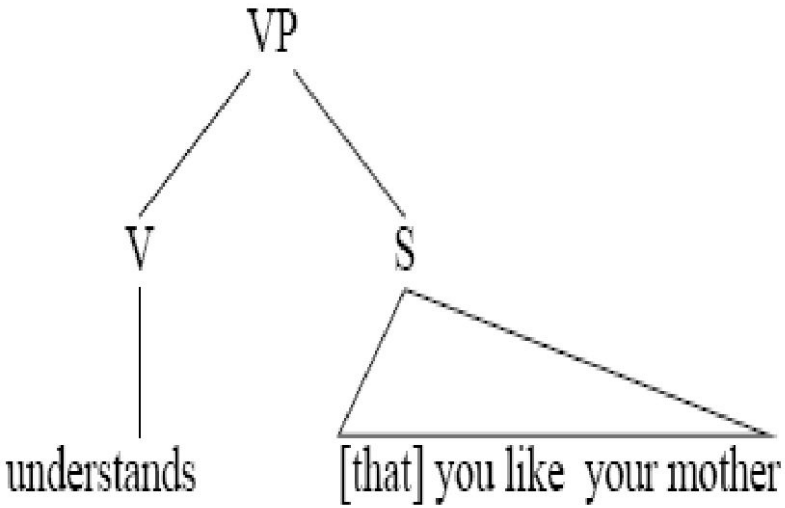
1. “ ... a computer that understands you like your mother”
2. “ ... a computer that understands you lie cured mother”

# Ambiguity at Many Levels

At the syntactic level:

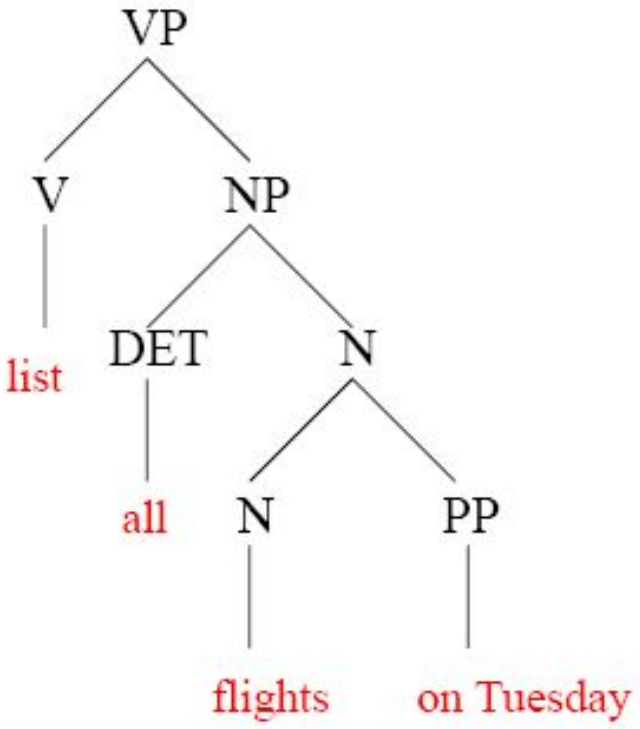
D

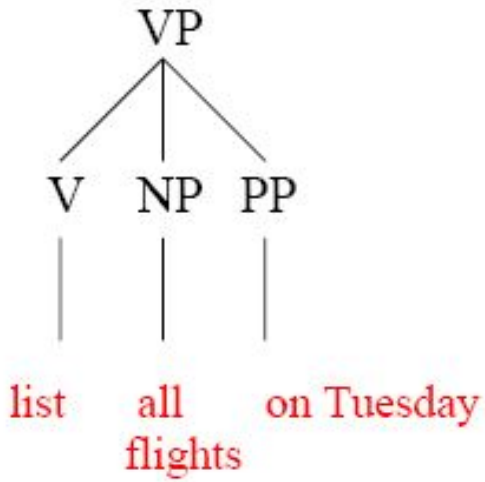



*

# More Syntactic Ambiguity




*

# Ambiguity at Many Levels

At the semantic (meaning) level:

Two definitions of “mother”

- a woman who has given birth to a child
- a substance consisting of bacteria, used to produce vinegar (i.e., mother of vinegar)

This is an instance of word sense ambiguity

# Ambiguity at Many Levels

At the discourse level:

- Alice says they've built a computer that understands you like your mother
- But she ...
- ... doesn't know any details
- ... doesn't understand me at all

# Syntactic analysis

- Syntax can make explicit when there are several possible interpretations
  - *(Rice flies) like sand.*
  - *Rice (flies like sand).*
- Knowledge of ‘correct’ grammar can help finding the right interpretation
  - *Flying planes are dangerous.*
  - *Flying planes is dangerous.*

# Syntax shows how words are related in a sentence.

Visiting aunts ARE boring.

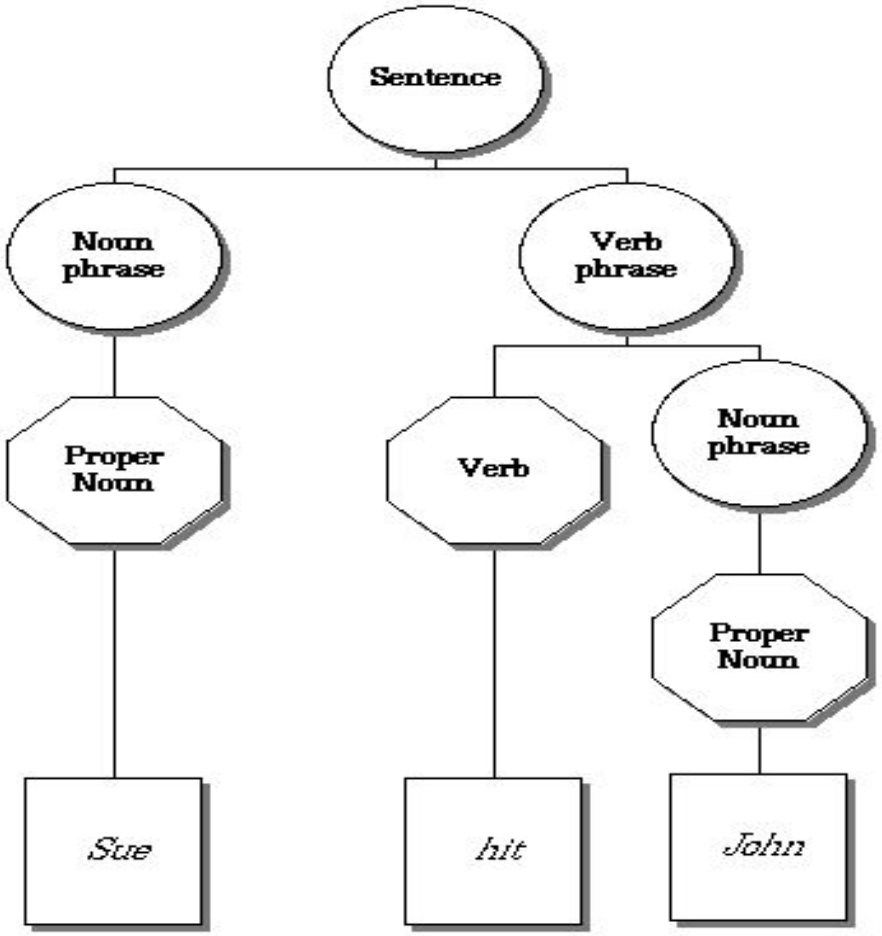
vs

Visiting aunts IS boring.

Subject verb agreement allows us to disambiguate here.

# How do we represent syntax?

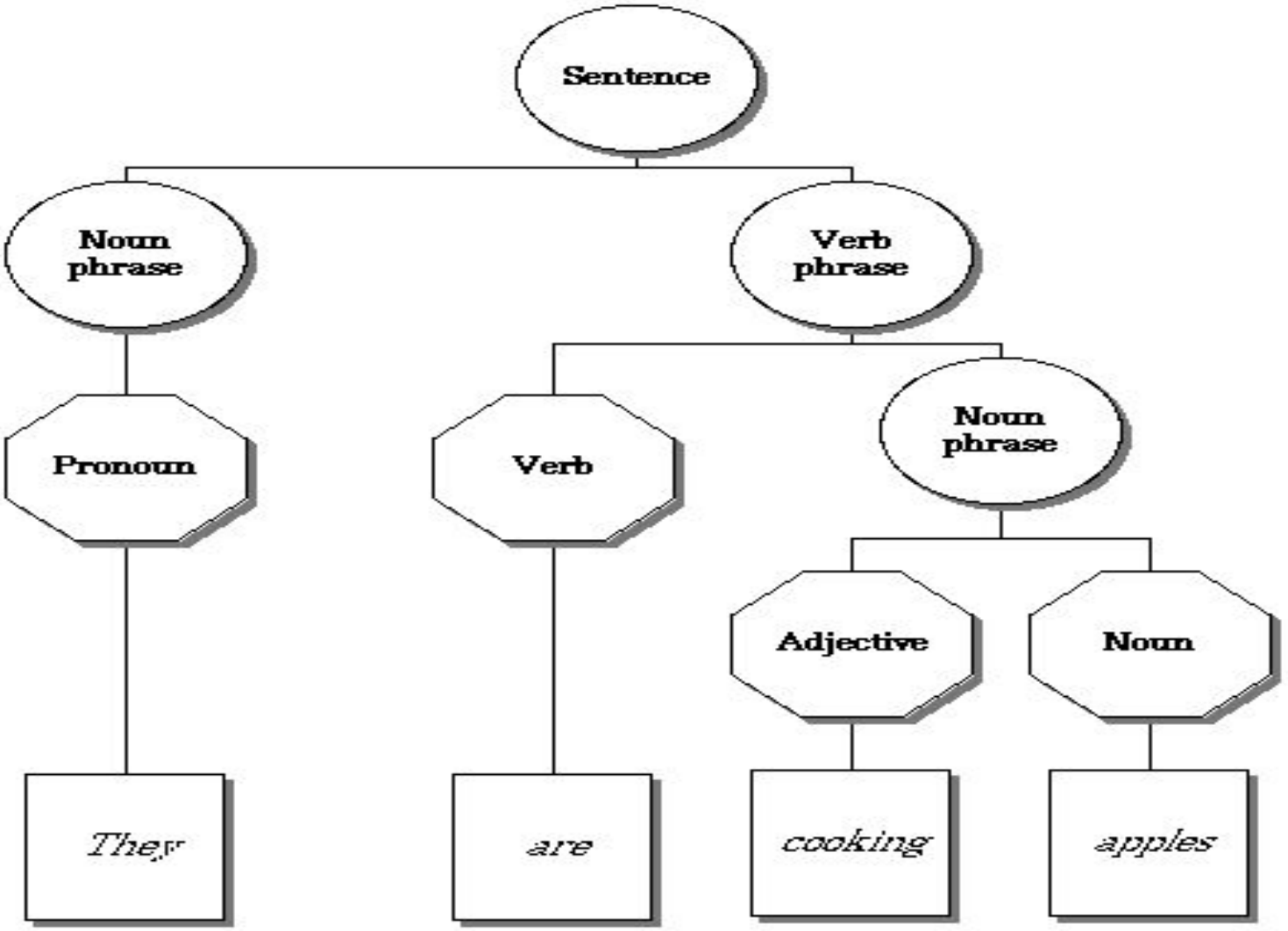
Parse Tree



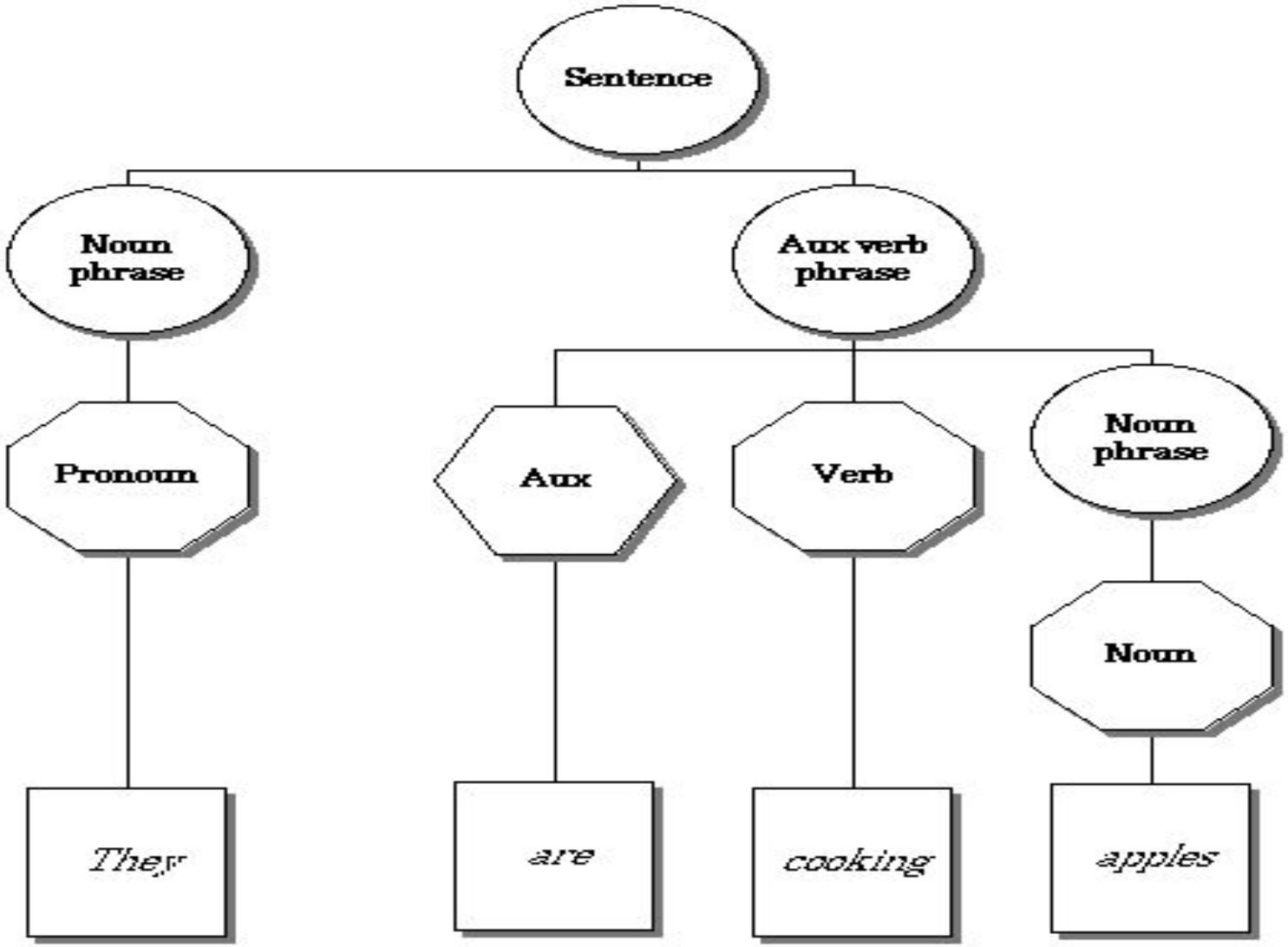
# An example:

- Parsing sentence:
- "They are cooking apples."

# Parse 1


Sentence
Noun phrase
Verb phrase
Pronoun
Verb
Noun phrase
Adjective
Noun
They
are
cooking
apples

# Parse 2


Sentence
Noun phrase
Aux verb phrase
Pronoun
Aux
Verb
Noun phrase
Noun
They
are
cooking
apples

*

# How do we represent syntax?

List

Sue hit John

[ s, [np, [proper_noun, Sue] ] ,

[vp, [v, hit],

[np, [proper_noun, John] ] ] ]

# What is Natural Language Processing (NLP)

- The process of computer analysis of input provided in a human language (natural language), and conversion of this input into a useful form of representation.
- The field of NLP is primarily concerned with getting computers to perform useful and interesting tasks with human languages.
- The field of NLP is secondarily concerned with helping us come to a better understanding of human language.

# Forms of Natural Language

- The input/output of a NLP system can be:
  - **written text**
  - **speech**
- We will mostly concerned with written text (not speech).
- To process written text, we need:
  - **lexical, syntactic, semantic knowledge about the language**
  - **discourse information, real world knowledge**
- To process spoken language, we need everything required to process written text, plus the challenges of speech recognition and speech synthesis.

# Components of NLP

- **Natural Language Understanding**
  - Mapping the given input in the natural language into a useful representation.
  - Different level of analysis required:

    ***morphological analysis,***

    ***syntactic analysis,***

    ***semantic analysis,***

    ***discourse analysis****, …*
- **Natural Language Generation**
  - Producing output in the natural language from some internal representation.
  - Different level of synthesis required:

    ***deep planning*** (what to say),

    ***syntactic generation***
- NL Understanding is much harder than NL Generation. But, still both of them are hard.

# Why NL Understanding is hard?

- Natural language is extremely rich in form and structure, and **very ambiguous**.
  - How to represent meaning,
  - Which structures map to which meaning structures.
- One input can mean many different things. Ambiguity can be at different levels.
  - Lexical (word level) ambiguity -- different meanings of words
  - Syntactic ambiguity -- different ways to parse the sentence
  - Interpreting partial information -- how to interpret pronouns
  - Contextual information -- context of the sentence may affect the meaning of that sentence.
- Many input can mean the same thing.
- Interaction among components of the input is not clear.

# Knowledge of Language

- **Phonology** – concerns how words are related to the sounds that realize them.
- **Morphology** – concerns how words are constructed from more basic meaning units called morphemes. A morpheme is the primitive unit of meaning in a language.
- **Syntax** – concerns how can be put together to form correct sentences and determines what structural role each word plays in the sentence and what phrases are subparts of other phrases.
- **Semantics** – concerns what words mean and how these meaning combine in sentences to form sentence meaning. The study of context-independent meaning.

# Knowledge of Language (cont.)

- **Pragmatics** – concerns how sentences are used in different situations and how use affects the interpretation of the sentence.

- **Discourse** – concerns how the immediately preceding sentences affect the interpretation of the next sentence. For example, interpreting pronouns and interpreting the temporal aspects of the information.

- **World Knowledge** – includes general knowledge about the world. What each language user must know about the other's beliefs and goals.

# Ambiguity

**I made her duck.**

- How many different interpretations does this sentence have?
- What are the reasons for the ambiguity?
- The categories of knowledge of language can be thought of as ambiguity resolving components.
- How can each ambiguous piece be resolved?
- Does speech input make the sentence even more ambiguous?
  - Yes – deciding word boundaries

# Ambiguity (cont.)

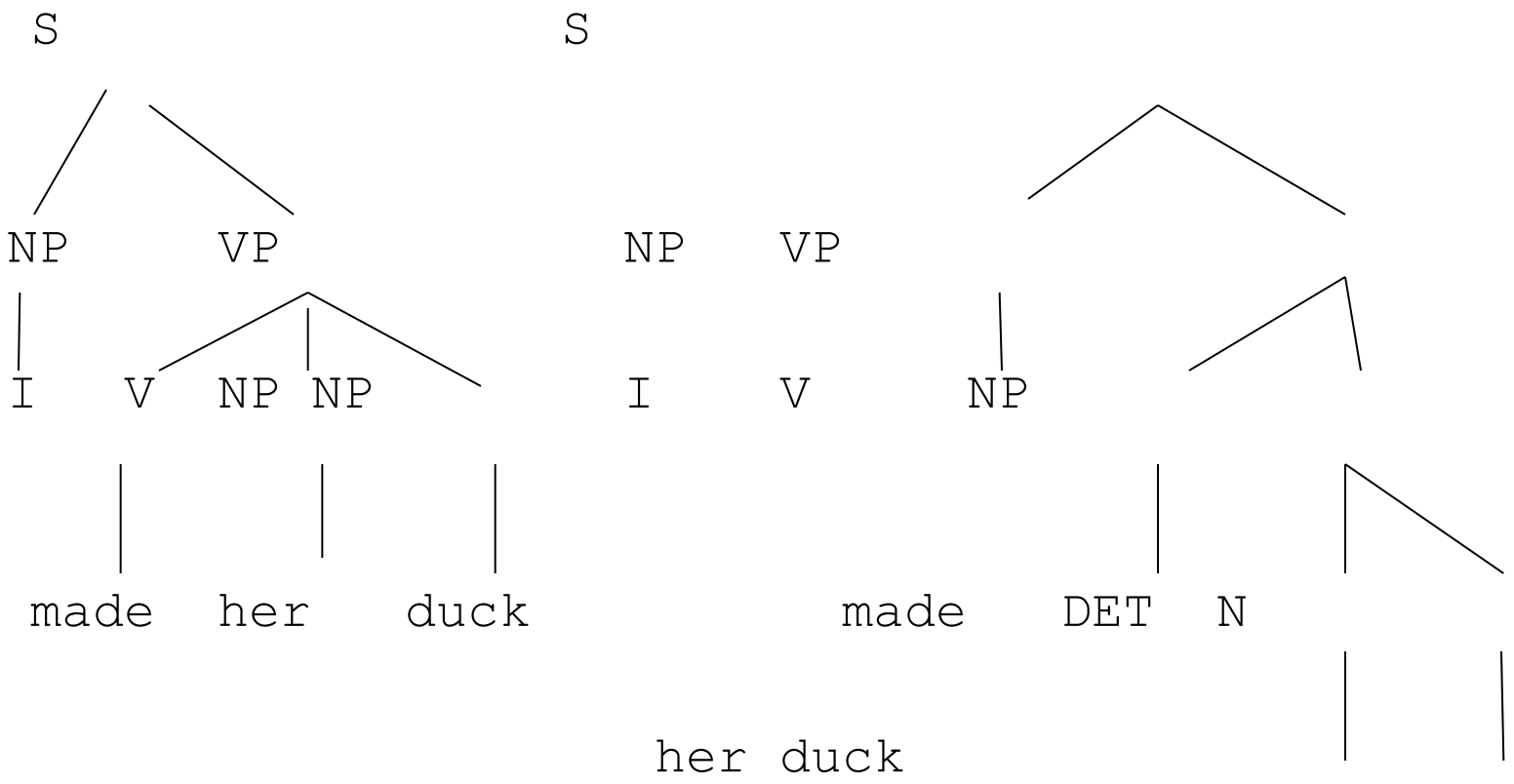
- Some interpretations of : **`I made her duck.`**
  1. I cooked *duck* for her.
  2. I cooked *duck* belonging to her.
  3. I created a toy duck which she owns.
  4. I caused her to quickly lower her head or body.
  5. I used magic and turned her into a *duck*.
- `duck` – morphologically and syntactically ambiguous: noun or verb.
- `her` – syntactically ambiguous: dative or possessive.
- `make` – semantically ambiguous: cook or create.
- `make` – syntactically ambiguous:
  - Transitive – takes a direct object. => 2
  - Di-transitive – takes two objects. => 5
  - Takes a direct object and a verb. => 4

# Resolve Ambiguities

- We will introduce *models* and *algorithms* to resolve ambiguities at different levels.
- **part-of-speech tagging** -- Deciding whether `duck` is verb or noun.
- **word-sense disambiguation** -- Deciding whether `make` is `create` or `cook`.
- **lexical disambiguation** -- Resolution of part-of-speech and word-sense ambiguities are two important kinds of lexical disambiguation.
- **syntactic ambiguity** -- `her duck` is an example of syntactic ambiguity, and can be addressed by probabilistic parsing.

# Resolve Ambiguities (cont.)

**I made her duck**




*

# Models to Represent Linguistic Knowledge

- We will use certain formalisms (*models*) to represent the required linguistic knowledge.
- **State Machines** -- FSAs, FSTs, HMMs, ATNs, RTNs
- **Formal Rule Systems** -- Context Free Grammars, Unification Grammars, Probabilistic CFGs.
- **Logic-based Formalisms** -- first order predicate logic, some higher order logic.
- **Models of Uncertainty** -- Bayesian probability theory.

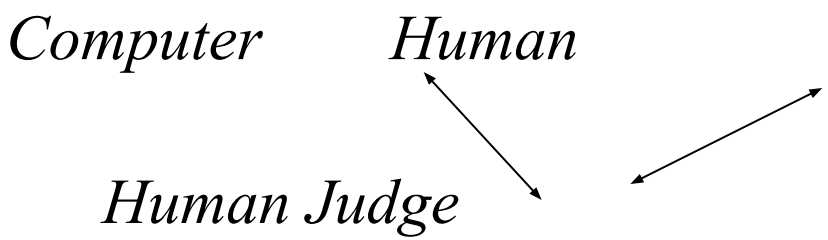
# Algorithms to Manipulate Linguistic Knowledge

- We will use *algorithms* to manipulate the models of linguistic knowledge to produce the desired behavior.
- Most of the algorithms we will study are **transducers** and **parsers**.
  - These algorithms construct some structure based on their input.
- Since the language is ambiguous at all levels,

  these algorithms are never simple processes.
- Categories of most algorithms that will be used can fall into following categories.
  - state space search
  - dynamic programming

# Language and Intelligence

## Turing Test




- *Human Judge* asks tele-typed questions to *Computer* and *Human.*
- *Computer's* job is to act like a human.
- *Human's* job is to convince Judge that he is not machine.
- *Computer* is judged "intelligent" if it can fool the judge
- Judgment of intelligence is linked to appropriate answers to questions from the system.

# NLP - an inter-disciplinary Field

- NLP borrows techniques and insights from several disciplines.
- **Linguistics**: How do words form phrases and sentences? What constraints the possible meaning for a sentence?
- **Computational Linguistics**: How is the structure of sentences are identified? How can knowledge and reasoning be modeled?
- **Computer Science**: Algorithms for automatons, parsers.
- **Engineering**: Stochastic techniques for ambiguity resolution.
- **Psychology**: What linguistic constructions are easy or difficult for people to learn to use?
- **Philosophy**: What is the meaning, and how do words and sentences acquire it?

# Some Buzz-Words

- NLP – Natural Language Processing
- CL – Computational Linguistics
- SP – Speech Processing
- HLT – Human Language Technology
- NLE – Natural Language Engineering
- SNLP – Statistical Natural Language Processing
- Other Areas:
  - Speech Generation, Text Generation, Speech Understanding, Information Retrieval,
  - Dialogue Processing, Inference, Spelling Correction, Grammar Correction,
  - Text Summarization, Text Categorization,

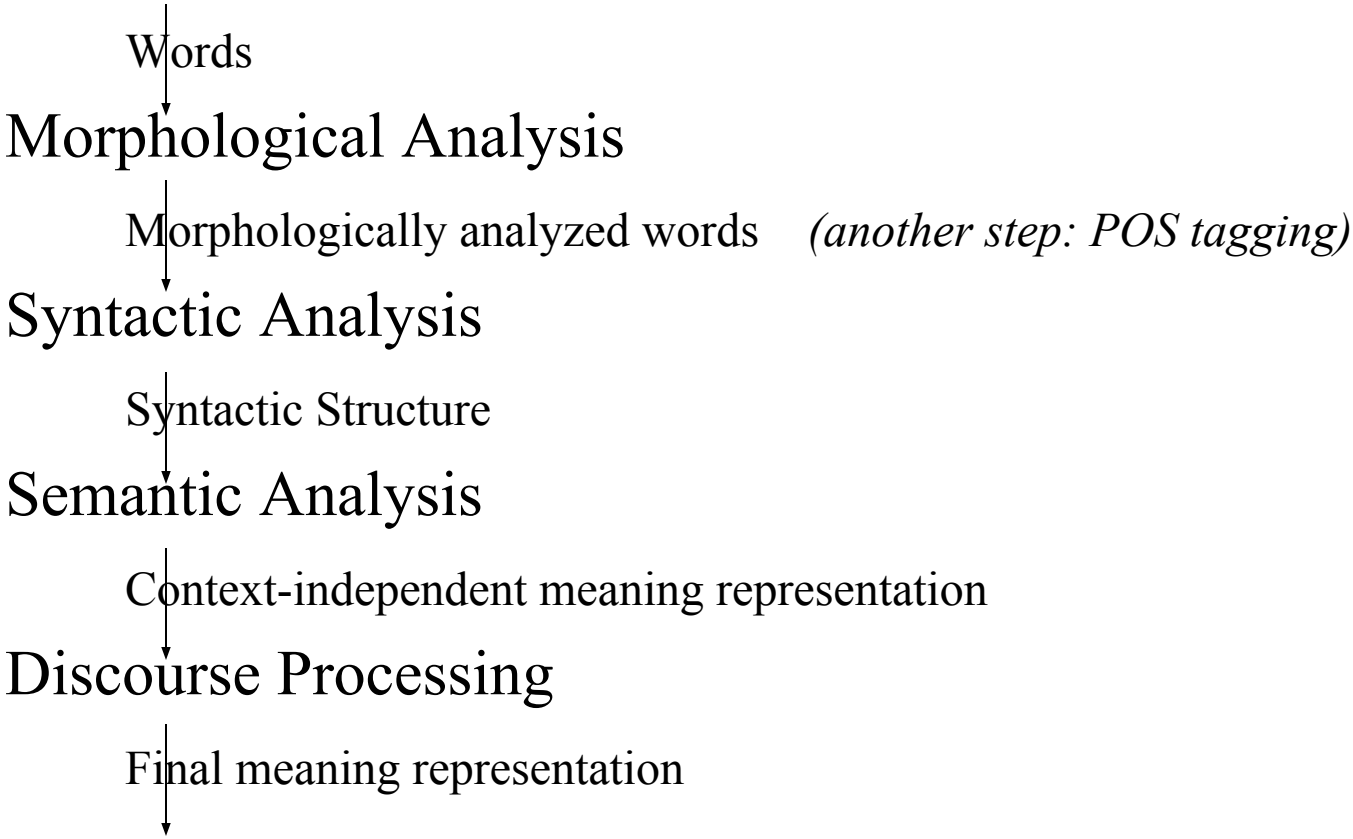
# Some NLP Applications

- Machine Translation – Translation between two natural languages.
  - See the Babel Fish translations system on Alta Vista.
- Information Retrieval – Web search (uni-lingual or multi-lingual).
- Query Answering/Dialogue – Natural language interface with a database system, or a dialogue system.
- Report Generation – Generation of reports such as weather reports.
- Some Small Applications –
  - Grammar Checking, Spell Checking, Spell Corrector

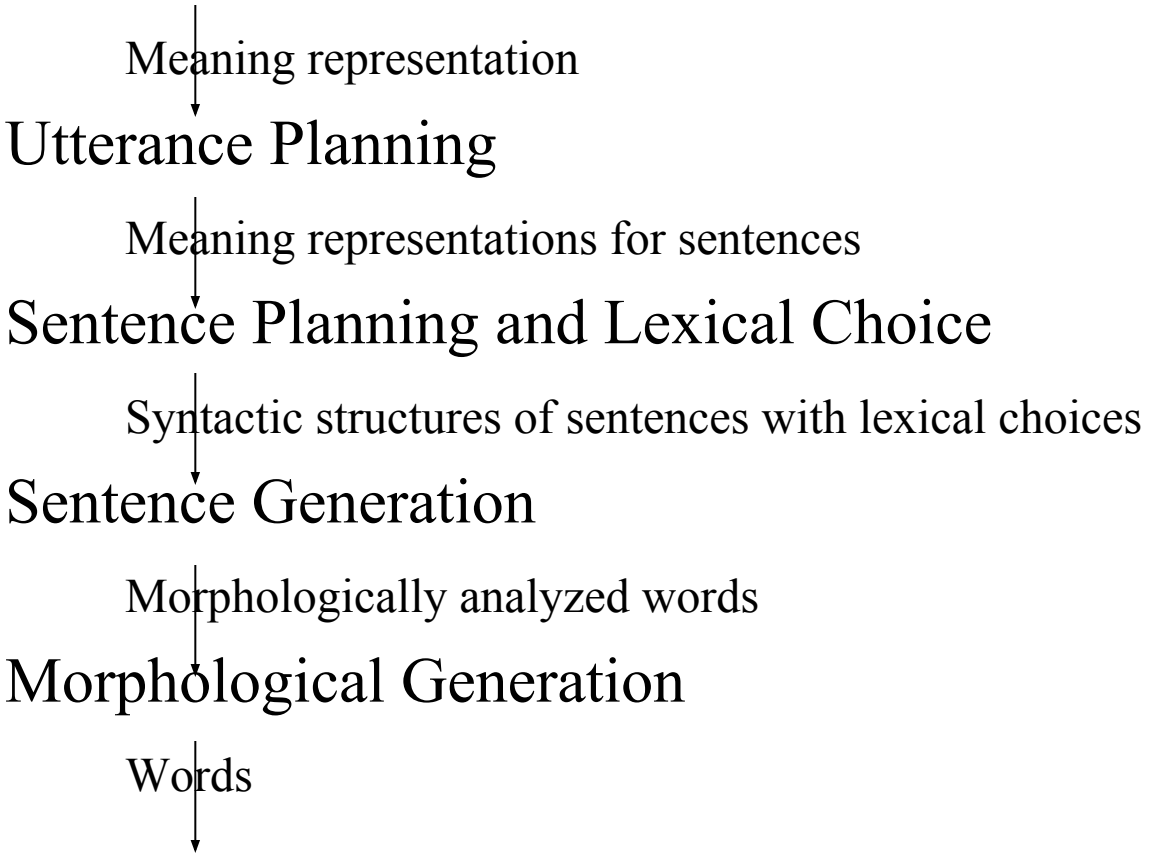
# Brief History of NLP

- 1940s –1950s: Foundations
  - Development of formal language theory (Chomsky, Backus, Naur, Kleene)
  - Probabilities and information theory (Shannon)
- 1957 – 1970s:
  - Use of formal grammars as basis for natural language processing (Chomsky, Kaplan)
  - Use of logic and logic based programming (Minsky, Winograd, Colmerauer, Kay)
- 1970s – 1983:
  - Probabilistic methods for early speech recognition (Jelinek, Mercer)
  - Discourse modeling (Grosz, Sidner, Hobbs)
- 1983 – 1993:
  - Finite state models (morphology) (Kaplan, Kay)
- 1993 – present:
  - Strong integration of different techniques, different areas.

# Natural Language Understanding



# Natural Language Generation



# Morphological Analysis

- Analyzing words into their linguistic components (morphemes).
- Morphemes are the smallest meaningful units of language.

  cars car+PLU

  giving give+PROG

  geliyordum gel+PROG+PAST+1SG - I was coming
- Ambiguity: More than one alternatives

  flies $fly_{VERB}$+PROG

  $fly_{NOUN}$+PLU

  adamı adam+ACC - the man (accusative)

  adam+P1SG - my man

  ada+P1SG+ACC - my island (accusative)

*

# Morphological Analysis (cont.)

- Relatively simple for English. But for some languages such as Turkish, it is more difficult.

  uygarlaştıramadıklarımızdanmışsınızcasına

  uygar-laş-tır-ama-dık-lar-ımız-dan-mış-sınız-casına

  uygar +BEC +CAUS +NEGABLE +PPART +PL +P1PL +ABL +PAST +2PL +AsIf

  "(behaving) as if you are among those whom we could not civilize/cause to become civilized"

  +BEC is "become" in English
  +CAUS is the causative voice marker on a verb
  +PPART marks a past participle form
  +P1PL is 1st person plural possessive marker
  +2PL is 2nd person plural
  +ABL is the ablative (from/among) case marker
  +AsIf is a derivational marker that forms an adverb from a finite verb form
  +NEGABLE is "not able" in English

- Inflectional and Derivational Morphology.
- Common tools: Finite-state transducers

# Part-of-Speech (POS) Tagging

- Each word has a part-of-speech tag to describe its category.
- Part-of-speech tag of a word is one of major word groups (or its subgroups).
  - **open classes** -- noun, verb, adjective, adverb
  - **closed classes** -- prepositions, determiners, conjuctions, pronouns, particples
- POS Taggers try to find POS tags for the words.
- `duck` is a verb or noun? (morphological analyzer cannot make decision).
- A POS tagger may make that decision by looking the surrounding words.
  - `Duck!` (verb)
  - `Duck is delicious for dinner.` (noun)

# Lexical Processing

- The purpose of lexical processing is to determine meanings of individual words.
- Basic methods is to lookup in a database of meanings -- **lexicon**
- We should also identify non-words such as punctuation marks.
- Word-level ambiguity -- words may have several meanings, and the correct one cannot be chosen based solely on the word itself.
  - bank in English
  - yüz in Turkish
- Solution -- resolve the ambiguity on the spot by POS tagging (if possible) or pass-on the ambiguity to the other levels.

# Syntactic Processing

- **Parsing** -- converting a flat input sentence into a hierarchical structure that corresponds to the units of meaning in the sentence.
- There are different parsing formalisms and algorithms.
- Most formalisms have two main components:
  - **grammar** -- a declarative representation describing the syntactic structure of sentences in the language.
  - **parser** -- an algorithm that analyzes the input and outputs its structural representation (its parse) consistent with the grammar specification.
- CFGs are in the center of many of the parsing mechanisms. But they are complemented by some additional features that make the formalism more suitable to handle natural languages.

# Semantic Analysis

- Assigning meanings to the structures created by syntactic analysis.
- Mapping words and structures to particular domain objects in way consistent with our knowledge of the world.
- Semantic can play an import role in selecting among competing syntactic analyses and discarding illogical analyses.
  - `I robbed the bank` -- bank is a river bank or a financial institution
- We have to decide the formalisms which will be used in the meaning representation.

# Knowledge Representation for NLP

- Which knowledge representation will be used depends on the application -- Machine Translation, Database Query System.
- Requires the choice of representational framework, as well as the specific meaning vocabulary (what are concepts and relationship between these concepts -- ontology)
- Must be computationally effective.
- Common representational formalisms:
  - first order predicate logic
  - conceptual dependency graphs
  - semantic networks
  - Frame-based representations

# Discourse

- Discourses are collection of coherent sentences (not arbitrary set of sentences)
- Discourses have also hierarchical structures (similar to sentences)
- **anaphora resolution** -- to resolve referring expression
  - `Mary bought a book for Kelly. **She** didn't like **it**.`
    - **She** refers to Mary or Kelly. -- possibly Kelly
    - **It** refers to what -- book.
  - `Mary had to lie for Kelly. **She** didn't like **it**.`
- Discourse structure may depend on application.
  - Monologue
  - Dialogue
  - Human-Computer Interaction

# Natural Language Generation

- NLG is the process of constructing natural language outputs from non-linguistic inputs.
- NLG can be viewed as the reverse process of NL understanding.
- A NLG system may have two main parts:
  - **Discourse Planner** -- what will be generated. which sentences.
  - **Surface Realizer** -- realizes a sentence from its internal representation.
- **Lexical Selection** -- selecting the correct words describing the concepts.

# Machine Translation

- Machine Translation -- converting a text in language A into the corresponding text in language B (or speech).
- Different Machine Translation architectures:
  - interlingua based systems
  - transfer based systems
- How to acquire the required knowledge resources such as mapping rules and bi-lingual dictionary? By hand or acquire them automatically from corpora.
- Example Based Machine Translation acquires the required knowledge (some of it or all of it) from corpora.