Introduction to Natural Language Processing

Lecture #1

SNU 4th Industrial Revolution Academy:
Artificial Intelligence Agent

What is Natural Language Processing?

- The study of human languages and how they can be represented computationally and analyzed and generated algorithmically
  - *The cat is on the mat.* --> on (mat, cat)
  - on (mat, cat) --> *The cat is on the mat*

- Studying NLP involves studying natural language, formal representations, and algorithms for their manipulation

What *is* Natural Language Processing?

Building computational models of natural language comprehension and production

Other Names:

- Computational Linguistics (CL)
- Human Language Technology (HLT)
- Natural Language Engineering (NLE)
- Speech and Text Processing

Engineering Perspective

Use CL as part of a larger application:

- Spoken dialogue systems for telephone based information systems
- Components of web search engines or document retrieval services
  - Machine translation
  - Question/answering systems
  - Text Summarization
- Interface for intelligent tutoring/training systems

Emphasis on

- Robustness (doesn't collapse on unexpected input)
- Coverage (does something useful with most inputs)
- Efficiency (speech; large document collections)

Cognitive Science Perspective

Goal: gain an understanding of how people comprehend and produce language.

Goal: a model that explains actual human behaviour

Solution must:

explain psycholinguistic data

be verified by experimentation

Knowledge needed to understand and produce language

- *Phonetics and phonology:* how words are related to sounds that realize them
- *Morphology:* how words are constructed from more basic meaning units
- *Syntax:* how words can be put together to form correct utterances
- *Lexical semantics:* what words mean
- *Compositional semantics*: how word meanings combine to form larger meanings
- *Pragmatics:* how situation affects interpretation of utterance
- *Discourse structure:* how preceding utterances affects processing of next utterance

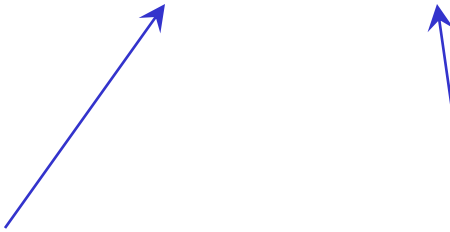
What can we learn about language?

- Phonetics and Phonology: speech sounds, their production, and the rule systems that govern their use
  - tap, butter
  - nice white rice; height/hot; kite/cot; night/not...
  - city **hall**, **parking** lot, **city** hall **parking** lot
  - The cat is on the mat. The cat is on the mat?

Morphology

- How words are constructed from more basic units, called *morphemes*

friend + *ly* = *friendly*



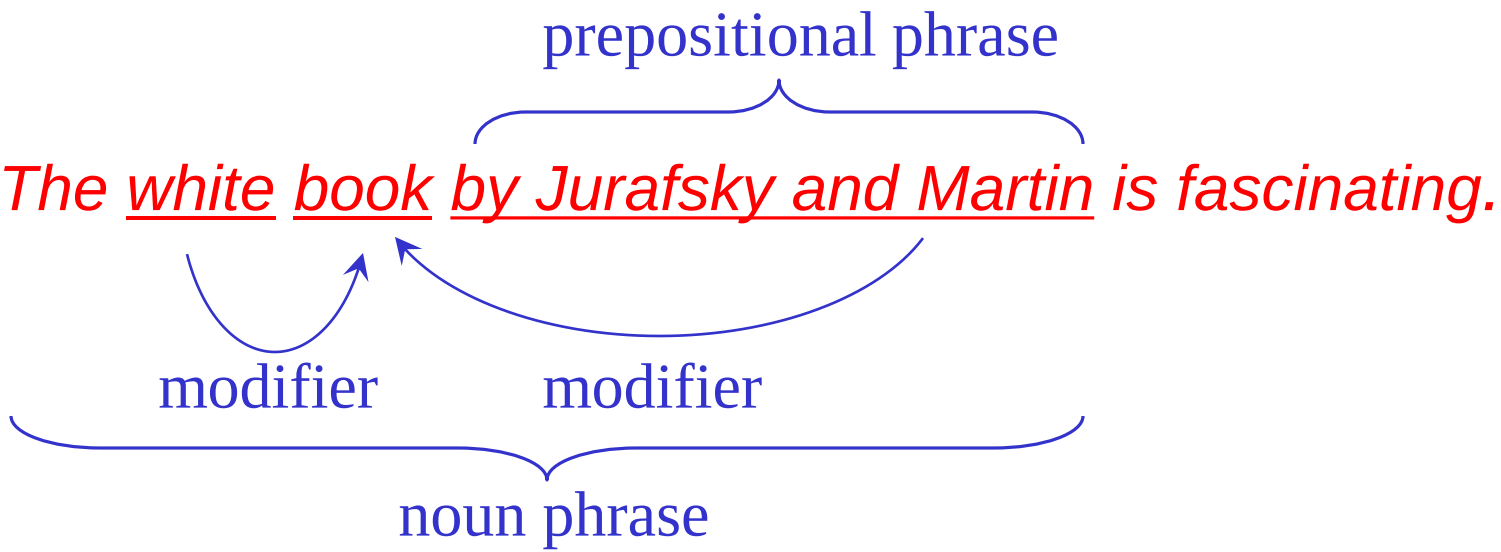
noun

Suffix -*ly* turns noun into an adjective (and verb into an adverb)

- Morphology: words and their composition
  - cat, cats, dogs
  - child, children
  - undo, union

Syntactic Knowledge

- how words can be put together to form legal sentences in the language
- what structural role each word plays in the sentence
- what phrases are subparts of other phrases

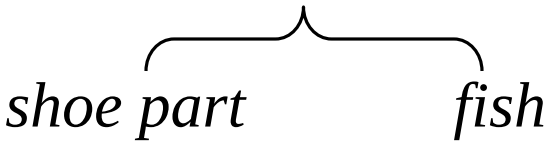


- Syntax: the structuring of words into larger phrases
  - John hit Bill
  - Bill was hit by John (passive)
  - Bill, John hit (preposing)
  - Who John hit was Bill (wh-cleft)

Semantic Knowledge

- What words mean
- How word meanings combine in sentences to form sentence meanings

The sole died. *(selectional restrictions)*



Syntax and semantics work together!

(1) What does it taste like?
(2) What taste does it like?

N.B. Context-independent meaning

- Semantics: the (truth-functional) meaning of words and phrases
  - gun(x) & holster(y) & in(x,y)
  - fake (gun (x)) (compositional semantics)
  - The king of France is bald (presupposition violation)
  - bass fishing, bass playing (word sense disambiguation)

- Pragmatics and Discourse: the meaning of words and phrases in context
  - George got married and had a baby.
  - George had a baby and got married.
  - Some people left early.
  - Prosodic Variation
    - German teachers
    - Bill doesn’t drink because he’s unhappy.
    - John only introduced Mary to Sue.
    - John called Bill a Republican and then he insulted him.
    - John likes his mother, and so does Bill.

Pragmatic Knowledge

- What utterances mean in different *contexts*

Jon was hot and desperate for a dunk in the river.

Jon suddenly realised he didn't have any cash.

He rushed to the bank.

financial institution river bank

Discourse Structure

Much meaning comes from simple conventions that we generally follow in discourse

- How we ***refer*** to entities
  - *Indefinite NPs* used to introduce new items into the discourse

    A woman walked into the cafe.
  - *Definite NPs* can be used to refer to subsequent references

    The woman sat by the window.
  - *Pronouns* used to refer to items already known in discourse

    She ordered a cappuccino.

Discourse Relations

- Relationships we infer between discourse entities
- Not expressed in either of the propositions, but from their juxtaposition

1. (a) I'm hungry.

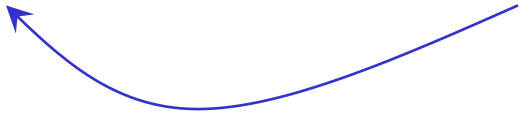
(b) Let's go to the Fuji Gardens.

2. (a) Bush supports big business.

(b) He'll vote no on House Bill 1711.

Discourse and Temporal Interpretation

Max fell. John pushed him.



explanation

Syntax and semantics: "him" refers to Max

Lexical semantics and discourse: the pushing occurred before the falling.

Discourse and Temporal Interpretation

John and Max were struggling at the edge of the cliff.

Max fell. John pushed him.

Here discourse knowledge tells us the pushing event occurred **after** the falling event

World knowledge

- What we know about the world and what we can assume our hearer knows about the world is intimately tied to our ability to use language

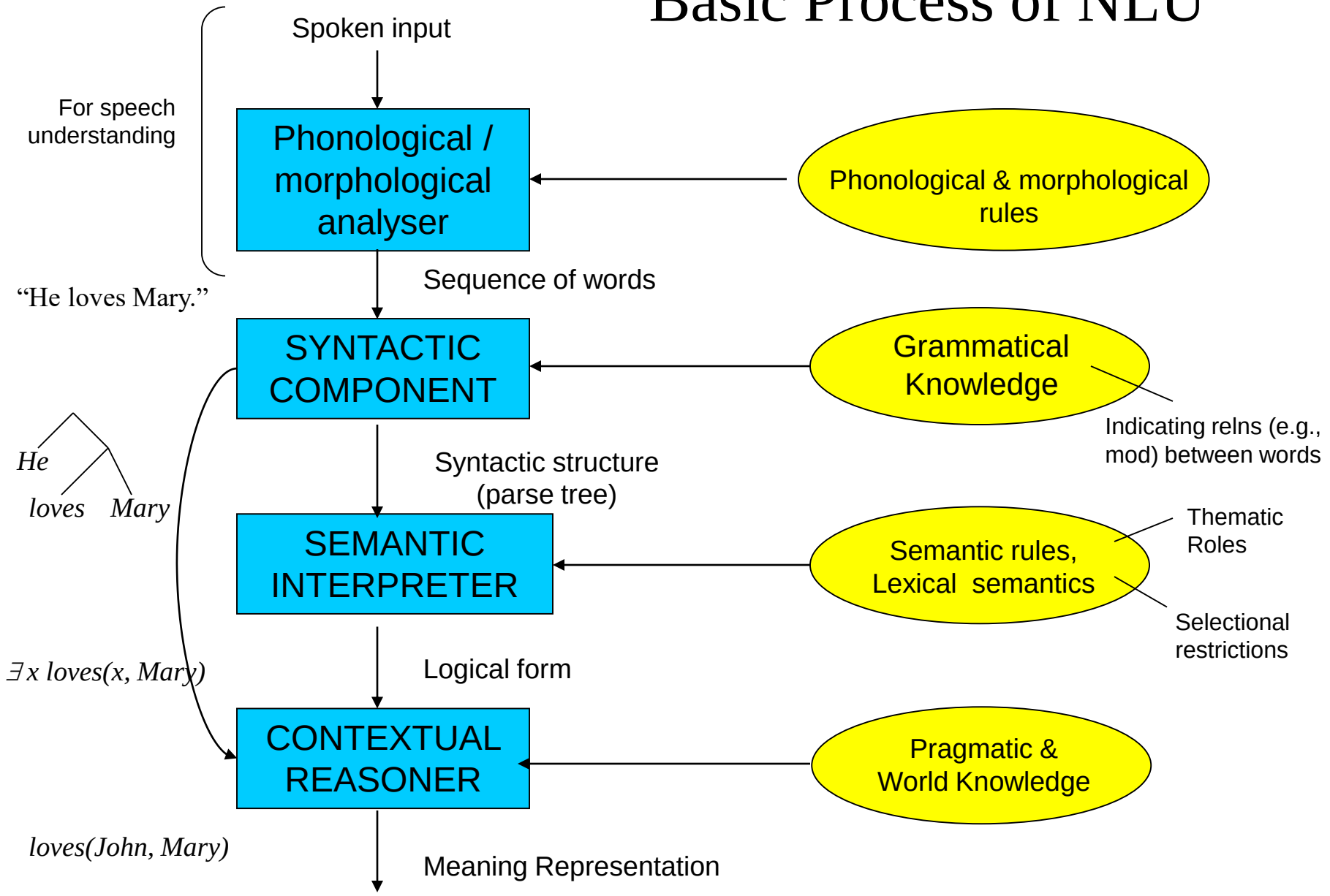
I took the cake from the plate and ate it.

Ambiguity

I made her duck.

- The categories of knowledge of language can be thought of as ambiguity-resolving components
- How many different interpretations does the above sentence have?
- How can each ambiguous piece be resolved?
- Does speech input make the sentence even more ambiguous?

Basic Process of NLU

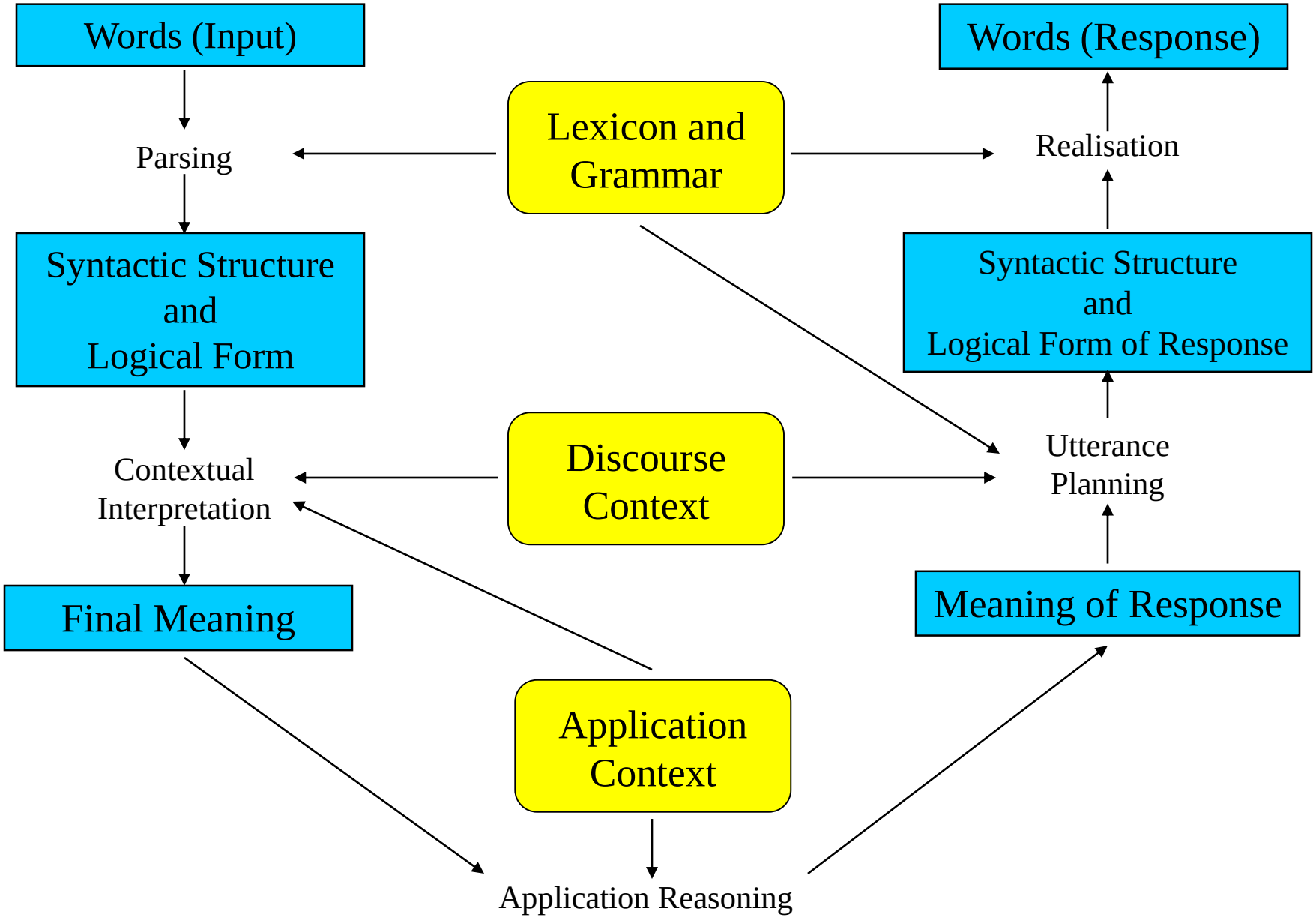


It’s not that simple

- Syntax affects meaning
  1. (a) *Flying planes is dangerous.*
     (b) *Flying planes are dangerous*.

- Meaning and world knowledge affects syntax
  2.* (a) *Flying insects is dangerous.*
     (b) *Flying insects are dangerous.*

  3. (a) *I saw the Grand Canyon flying to LA.*
     (b) *I saw a condor flying to LA.*


Words (Input)
Parsing
Syntactic Structure and Logical Form
Contextual Interpretation
Final Meaning
Lexicon and Grammar
Discourse Context
Application Context
Application Reasoning
Words (Response)
Realisation
Syntactic Structure and Logical Form of Response
Utterance Planning
Meaning of Response

Can machines think?

- Alan Turing: the *Turing test* (language as test for intelligence)
- Three participants: a computer and two humans (one is an interrogator)
- Interrogator's goal: to tell the machine and human apart
- Machine's goal: to fool the interrogator into believing that a person is responding
- Other human's goal: to help the interrogator reach his goal

Examples

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: 105621 (after a pause)

Example (from a famous movie)

Dave Bowman: Open the pod bay doors, HAL.
HAL: I'm sorry Dave, I'm afraid I can't do that.

Deconstructing HAL

- Recognizes speech and understands language
- Decides how to respond and speaks reply
- With personality
- Recognizes the user's goals, adopts them, and helps to achieve them
- Remembers the conversational history
- Customizes interaction to different individuals
- Learns from experience
- Possesses vast knowledge, and is autonomous

The state of the art and the near-term future

- World-Wide Web (WWW)
- Sample scenarios:
  - generate weather reports in two languages
  - provide tools to help people with SSI to communicate
  - translate Web pages into different languages
  - speak to your appliances
  - find restaurants
  - answer questions
  - grade essays (?)
  - closed-captioning in many languages
  - automatic description of a soccer gams

NLP Applications

- Speech Synthesis, Speech Recognition, IVR Systems (TOOT: more or less succeeds)
- Information Retrieval (SCANMail demo)
- Information Extraction
  - Question Answering (AQUA)
- Machine Translation (SYSTRAN)
- Summarization (NewsBlaster)
- Automated Psychotherapy (Eliza)