Indian Institute of Technology Kharagpur
Department of Computer Science & Engineering

CS60075
Natural Language Processing
Autumn 2020

Lecture 1B : Introduction

Sep 3 2020

What is natural language

- Created by humans for communication
- Learned from experience
- A symbolic/discrete system
- Encoded by continuous signals
    - Sounds (phoneme)
    - Image (grapheme)

Natural Language is Hard

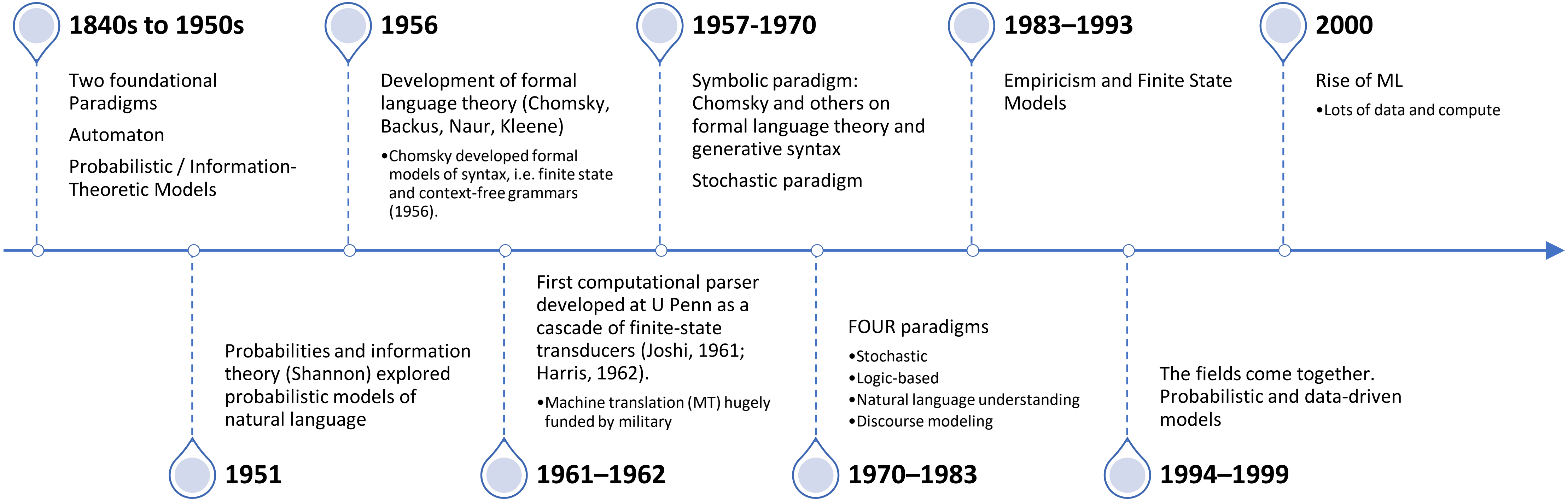
- Ambiguity
- Variability
- Sparsity
- Grounding

Examples of Applications

- Build systems that help people deal with text

- Text classification
- Machine translation
- Information extraction
- Dialog interfaces
- Question answering
- Virtual Assistants
- Human-level comprehension

NLP History



1840s to 1950s

Two foundational Paradigms

Automaton

Probabilistic / Information-Theoretic Models

1951

Probabilities and information theory (Shannon) explored probabilistic models of natural language

1956

Development of formal language theory (Chomsky, Backus, Naur, Kleene)

- Chomsky developed formal models of syntax, i.e. finite state and context-free grammars (1956).

1961–1962

First computational parser developed at U Penn as a cascade of finite-state transducers (Joshi, 1961; Harris, 1962).

- Machine translation (MT) hugely funded by military

1957-1970

Symbolic paradigm: Chomsky and others on formal language theory and generative syntax

Stochastic paradigm

1970–1983

FOUR paradigms

- Stochastic
- Logic-based
- Natural language understanding
- Discourse modeling

1983–1993

Empiricism and Finite State Models

1994–1999

The fields come together. Probabilistic and data-driven models

2000

Rise of ML

- Lots of data and compute

MT

Machine translation was one of the first non-numerical applications of computers. Until the mid 1960’s it was an area of intensive research activity and the focus of much public attention

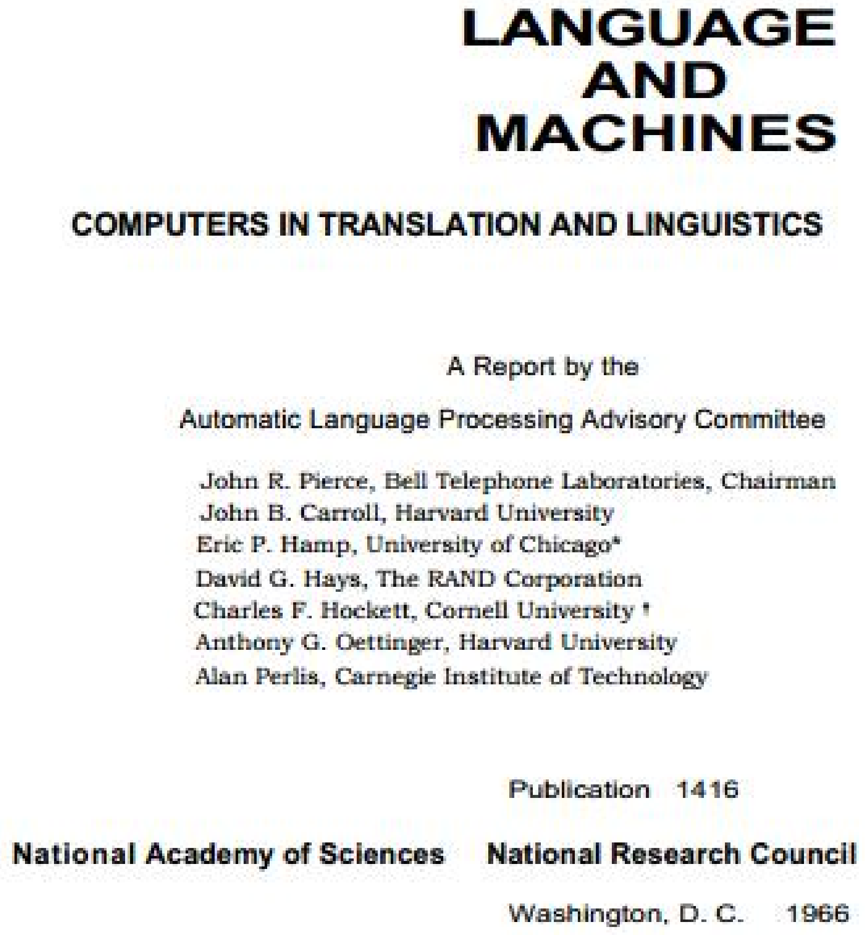
Early expectations were not fulfilled, linguistic complexities became increasingly apparent and seemed to be ever more intractable

ALPAC report: MT was generally considered to have been a ‘failure’

Machine Translation: the ALPAC report

ALPAC report (Automatic Language Processing Advisory Committee, 1966)

- Its effect was to bring to an end the funding of MT research for some twenty years.
- A message to the general public and the rest of the scientific community that MT was hopeless.
- To this day, the 'failure' of MT is still repeated by many as an indisputable fact.



LANGUAGE
AND
MACHINES

COMPUTERS IN TRANSLATION AND LINGUISTICS

A Report by the

Automatic Language Processing Advisory Committee

John R. Pierce, Bell Telephone Laboratories, Chairman
John B. Carroll, Harvard University
Eric P. Hamp, University of Chicago*
David G. Hays, The RAND Corporation
Charles F. Hockett, Cornell University †
Anthony G. Oettinger, Harvard University
Alan Perlis, Carnegie Institute of Technology

Publication 1416

National Academy of Sciences National Research Council

Washington, D. C. 1966

MT ridiculed

- Russian into English

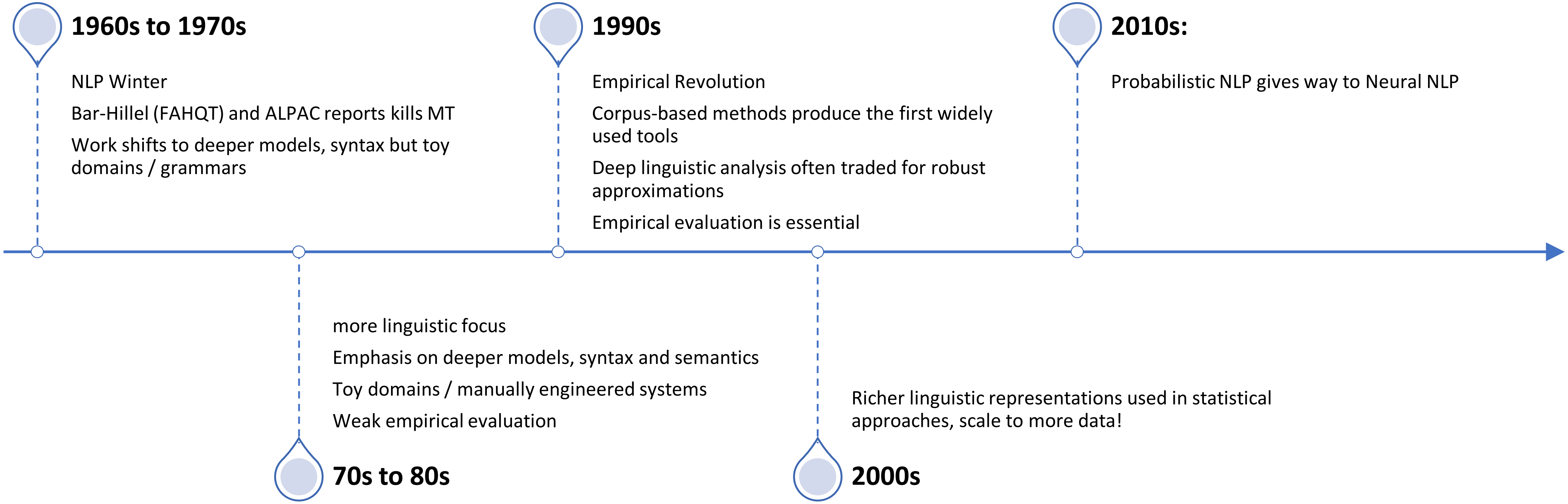
1. Out of sight, out of mind

   “invisible insanity”

2. The spirit is willing but the flesh is weak

   “The vodka is good but the meat is rotten”

NLP History



1960s to 1970s

NLP Winter

Bar-Hillel (FAHQT) and ALPAC reports kills MT

Work shifts to deeper models, syntax but toy domains / grammars

70s to 80s

more linguistic focus

Emphasis on deeper models, syntax and semantics

Toy domains / manually engineered systems

Weak empirical evaluation

1990s

Empirical Revolution

Corpus-based methods produce the first widely used tools

Deep linguistic analysis often traded for robust approximations

Empirical evaluation is essential

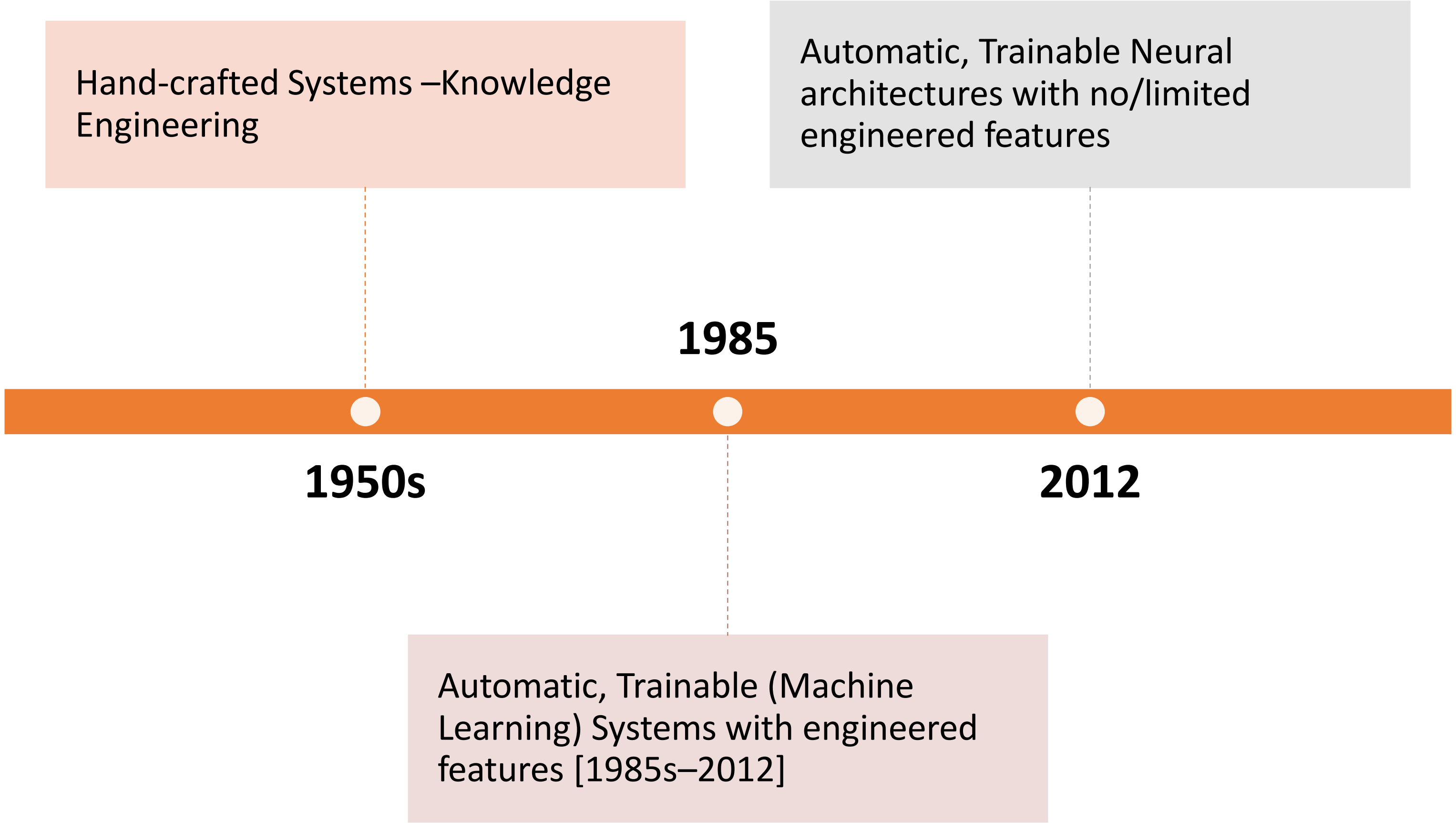
2000s

Richer linguistic representations used in statistical approaches, scale to more data!

2010s:

Probabilistic NLP gives way to Neural NLP

Three Generations of NLP



Hand-crafted Systems –Knowledge Engineering

1950s

1985

Automatic, Trainable (Machine Learning) Systems with engineered features [1985s–2012]

2012

Automatic, Trainable Neural architectures with no/limited engineered features

Components of NLP

Natural Language Understanding

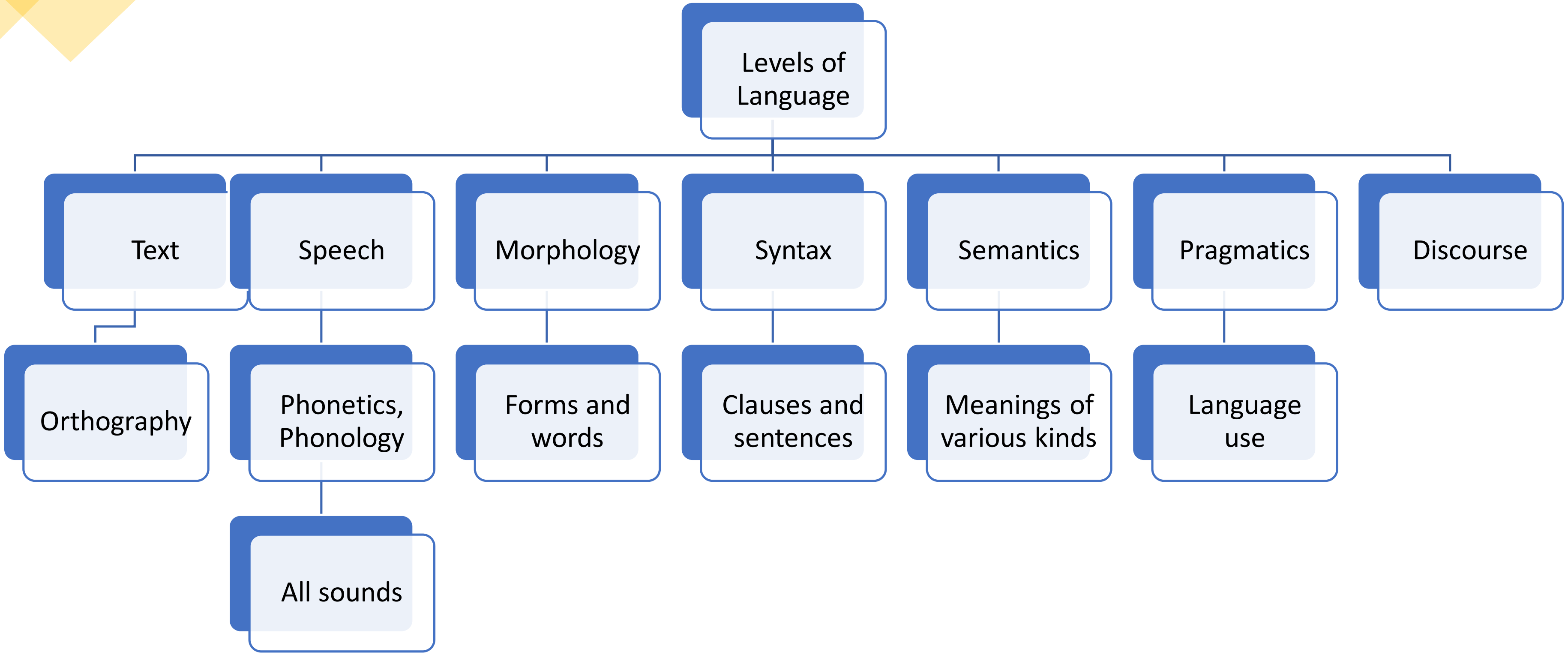
- Mapping the given input in the natural language into a useful representation.
- Different level of analysis required: morphological, syntactic, semantic, discourse

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

Natural Language Understanding

- Uncovering the mappings between the linear sequence of words and the meaning that it encodes.
- Representing this meaning in a useful (usually symbolic) representation.
- By definition - heavily dependent on the target task
  - Words and structures mean different things in different contexts
  - The required target representation is different for different tasks.
- Appropriateness of a representation depends on the application.


Levels of Language
Text
Speech
Morphology
Syntax
Semantics
Pragmatics
Discourse
Orthography
Phonetics, Phonology
Forms and words
Clauses and sentences
Meanings of various kinds
Language use
All sounds

Orthography

- Linking the symbols of an alphabet to the sounds of a language.
- Enable written communication
- How the symbols (graphemes) represent the sounds (phonemes) used in spoken language.

Phonology: concerns how words are related to the sounds that realize them.

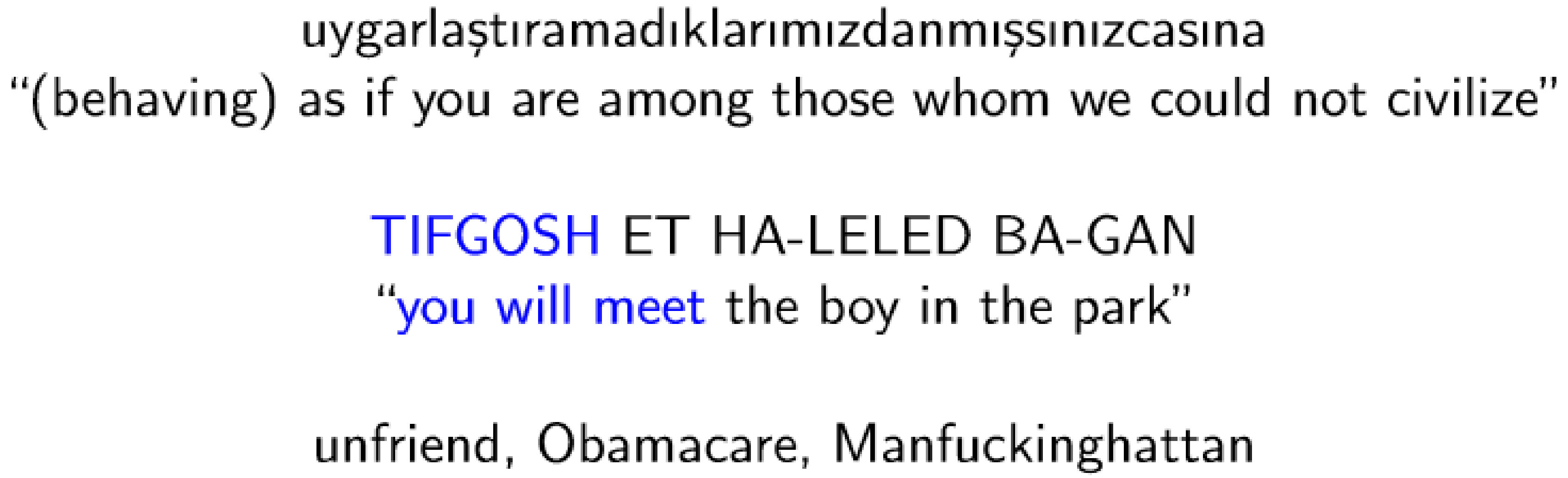
Morphology

- The identification, analysis and description of the structure of words

- Morpheme: the smallest linguistic unit with semantic meaning
- Lexeme: corresponds to a set of forms taken by a single word.

Morphology

The identification, analysis and description of the structure of words



uygarlaştıramadıklarımızdanmışsınızcasına

“(behaving) as if you are among those whom we could not civilize”

TIFGOSH ET HA-LELED BA-GAN

“you will meet the boy in the park”

unfriend, Obamacare, Manfuckinghattan

The Challenge of Words

- Segmenting text into words (Thai)
- Sandhi splitting (Sanskrit)
- Morphological variation
- Words with multiple meanings (based on context, domain)
- Multiword expression

Lexicon

- The lexicon contains information about particular idiosyncratic properties of words; eg. what sound or orthography goes with what meaning

Syntax

- Syntax concerns the way in which words can be combined together to form (grammatical) sentences
- Determines what structural role each word plays in the sentence and what phrases are subparts of other phrases.

1. revolutionary new ideas appear infrequently
2. colourless green ideas sleep furiously
3. *ideas green furiously colourless sleep

Syntax

- Words combine syntactically in certain orders in a way which mirrors the meaning conveyed
  - John loves Mary
  - Mary loves John
- John gave her dog biscuits
  - (john (gave (her) (dog biscuits)))
  - (john (gave (her dog) (biscuits)))

Semantics

- The manner in which lexical meaning is combined morphologically and syntactically to form the meaning of a sentence
  - Concerns the meaning of words, phrases and sentences
  - The meaning of a sentence is usually a productive combination of the meaning of its words

- **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.