Module 1
Introduction to NLP

Prepared By

Prof. Suja Jayachandran

Reference :

1.Speech and Language Processing by Daniel Jurafsky,James and Martin

2. https://datascience.foundation.com

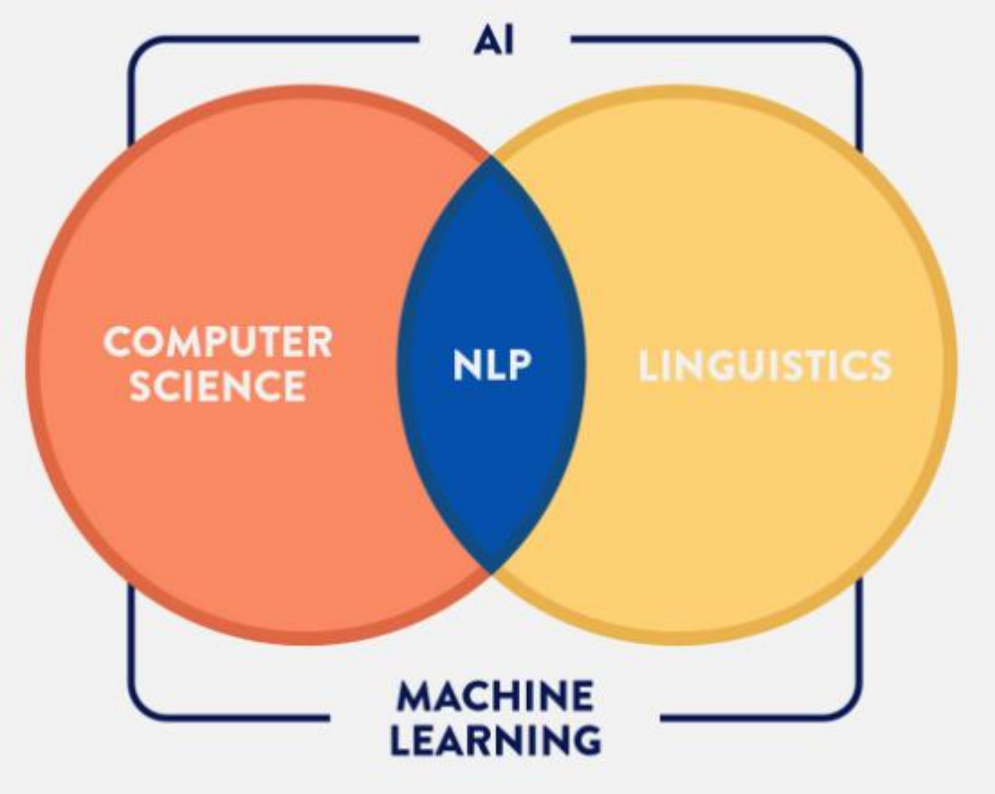
3. https://www.ibm.com/demos/live/natural-language-understanding/self-service/home

What is NLP??

- **NLP** is a subfield of computer science and artificial intelligence concerned with interactions between computers and human (natural) languages.
- It is used to apply **machine learning** algorithms to **text** and **speech**.
- It is the technique that enables the machine to interpret and understand the way humans communicate.

Why NLP??

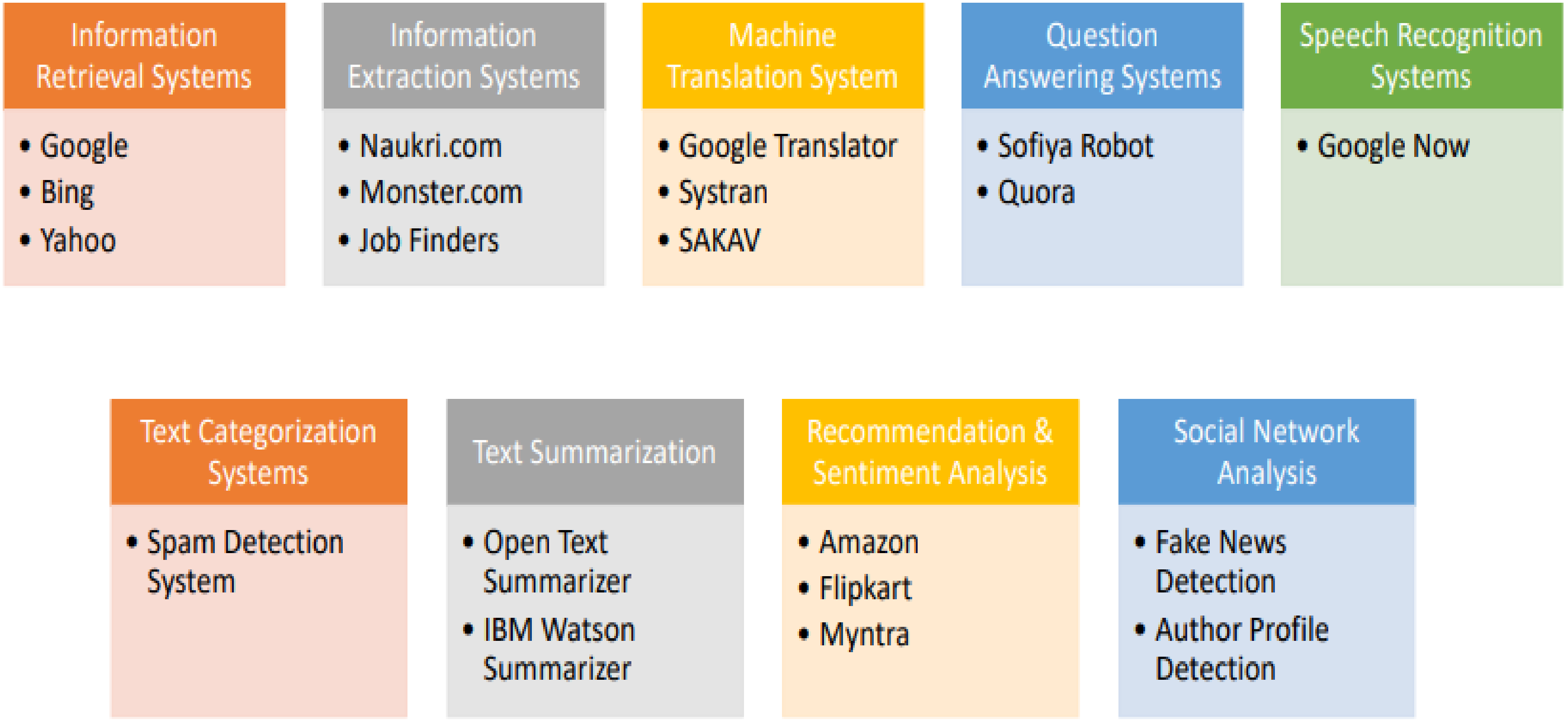
- 80% of total generated data is Unstructured
- To process Unstructured data we require NLP Expertise



Goals of NLP

- **Fundamental Goal**: Deep Understanding of broad language.
- **Engineering Goal**: Design, implement and test systems that process natural language for practical application.

Applications of NLP



Information Retrieval Systems

- Google
- Bing
- Yahoo

Information Extraction Systems

- Naukri.com
- Monster.com
- Job Finders

Machine Translation System

- Google Translator
- Systran
- SAKAV

Question Answering Systems

- Sofiya Robot
- Quora

Speech Recognition Systems

- Google Now

Text Categorization Systems

- Spam Detection System

Text Summarization

- Open Text Summarizer
- IBM Watson Summarizer

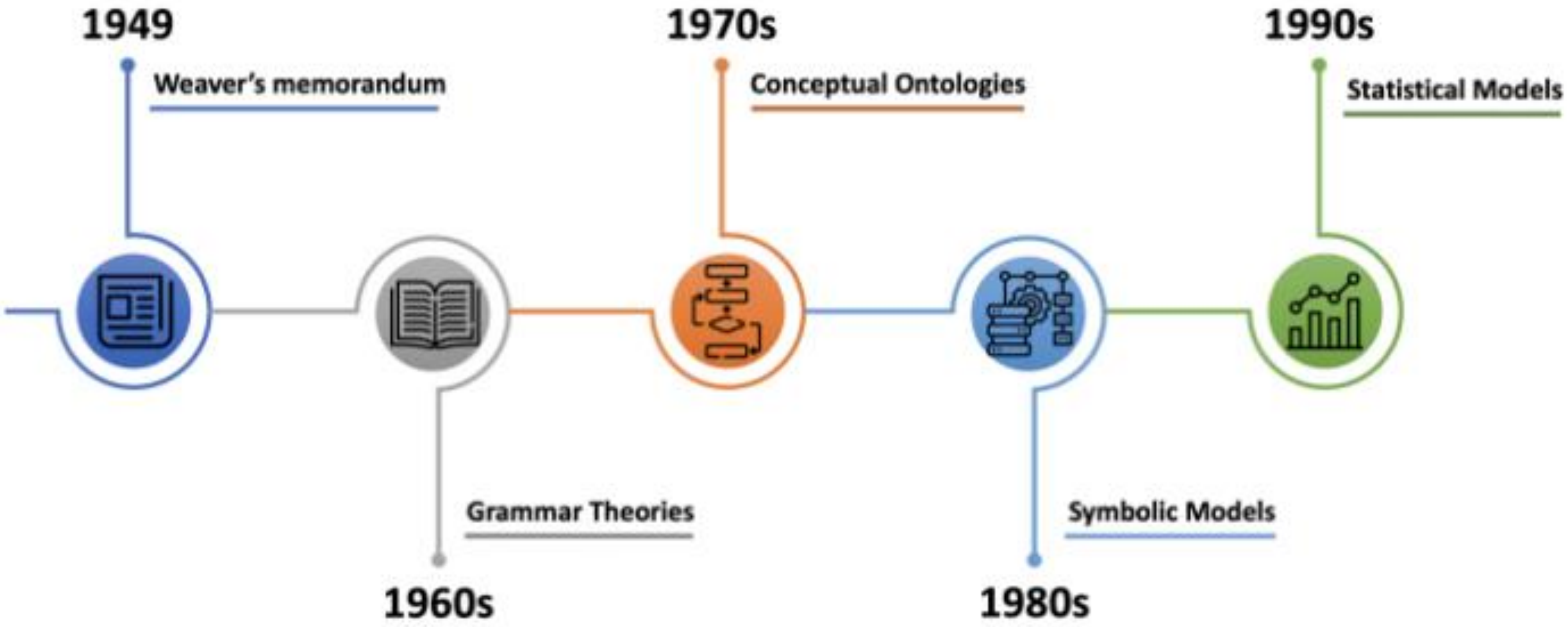
Recommendation & Sentiment Analysis

- Amazon
- Flipkart
- Myntra

Social Network Analysis

- Fake News Detection
- Author Profile Detection

History of NLP



The big stages of NLP before the deep learning era.

1949

- **Weaver’s memorandum** *(Shannon and Weaver, 1949)* brought the idea of the first computer-based application related to natural language: machine translation (MT).
- It subsequently inspired many projects, notably the Georgetown experiment *(Dostert, 1955)*, a joint project between IBM and Georgetown University that successfully demonstrated the machine translation of more than 60 Russian sentences into English.
- The researchers accomplished this feat by using hand-coded language rules, but the system failed to scale up to general translation.
- most systems used dictionary-lookup of appropriate words for translation and reordered the words after translation to fit the word-order rules of the target language.

1957-60

- introduce the idea of generative grammar *(Chomsky, 1957)*, a rule-based system of syntactic structures that brought insight into how mainstream linguistics could help machine translation.
- Theoretical work in the late 1960s and early 1970s mainly focused on how to represent meaning.
- Researchers developed new **theories of grammar** that were computationally tractable for the first time, particularly after the introduction of transformational generative grammars *(Chomsky, 1965)*, which were criticized for being too syntactically oriented and not lending themselves easily to computational implementation.

1970

- The 1970s brought new ideas into NLP, such as building **conceptual ontologies** which structured real-world information into computer-understandable data.
- Examples are MARGIE *(Schank and Abelson, 1975)*, TaleSpin *(Meehan, 1976)*, QUALM *(Lehnert, 1977)*, SAM *(Cullingford, 1978)*, PAM *(Schank and Wilensky, 1978)* and Politics *(Carbonell, 1979)*.

1980

- In the 1980s, many significant problems in NLP were addressed using **symbolic approaches** *(Charniak, 1983; Dyer, 1983; Riesbeck and Martin, 1986; Grosz et al., 1987; Hirst, 1987)*, i.e., complex hard-coded rules and grammars to parse language.
- Practically, text was segmented into meaningless tokens (words and punctuation).
- Representations were then manually created by assigning meanings to these tokens and their mutual relationships through well-understood knowledge representation schemes and associated algorithms.
- Those representations were eventually used to perform deep analysis of linguistic phenomena.

1990

- **Statistical models** came as a revolution in NLP *(Bahl et al., 1989; Brill et al., 1990; Chitrao and Grishman, 1990; Brown et al., 1991)*, replacing most natural language processing systems based on complex sets of hand-written rules.
- This progress was the result of both the steady increase of computational power, and the shift to machine learning algorithms.
- While some of the earliest-used machine learning algorithms, such as decision trees *(Tanaka, 1994; Allmuallim et al., 1994)*, produced systems similar in performance to the old school hand-written rules, statistical models broke through the complexity barrier of hand-coded rules by creating them through automatic learning, which led research to increasingly focus on these models.
- At the time, these statistical models were capable of making soft, probabilistic decisions.

Language, Knowledge & Grammar

- Language is the expression of ideas by means of speech sounds combined into words. A language is a system of communication that consists of a set of sounds and written symbols that are used by the people of a particular country or region for talking or writing.
- Knowledge deals with how words can be put together to form correct sentences. It also determines what structural role each word plays in the sentence and what phrases are subparts of what other phrases. It concerns the meanings of the words and sentences.
- Grammar is defined as the rules for forming well-structured sentences. Grammar denotes syntactical rules that are used for conversation in natural languages.

Knowledge of Language

- Phonetics and Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Disclosure

Knowledge of Language

Phonetics and Phonology: The study of linguistic sounds.

Morphology: The study of the meaningful components of words.

Syntax: The structural relationships between words.

Semantics: Study of meaning.

Pragmatics: Study of how language is used to accomplish goals.

Disclosure: Study of linguistic units larger than a single utterance.

STAGES OF NLP

Phonetics & Phonology

Morphology

Lexical Analysis

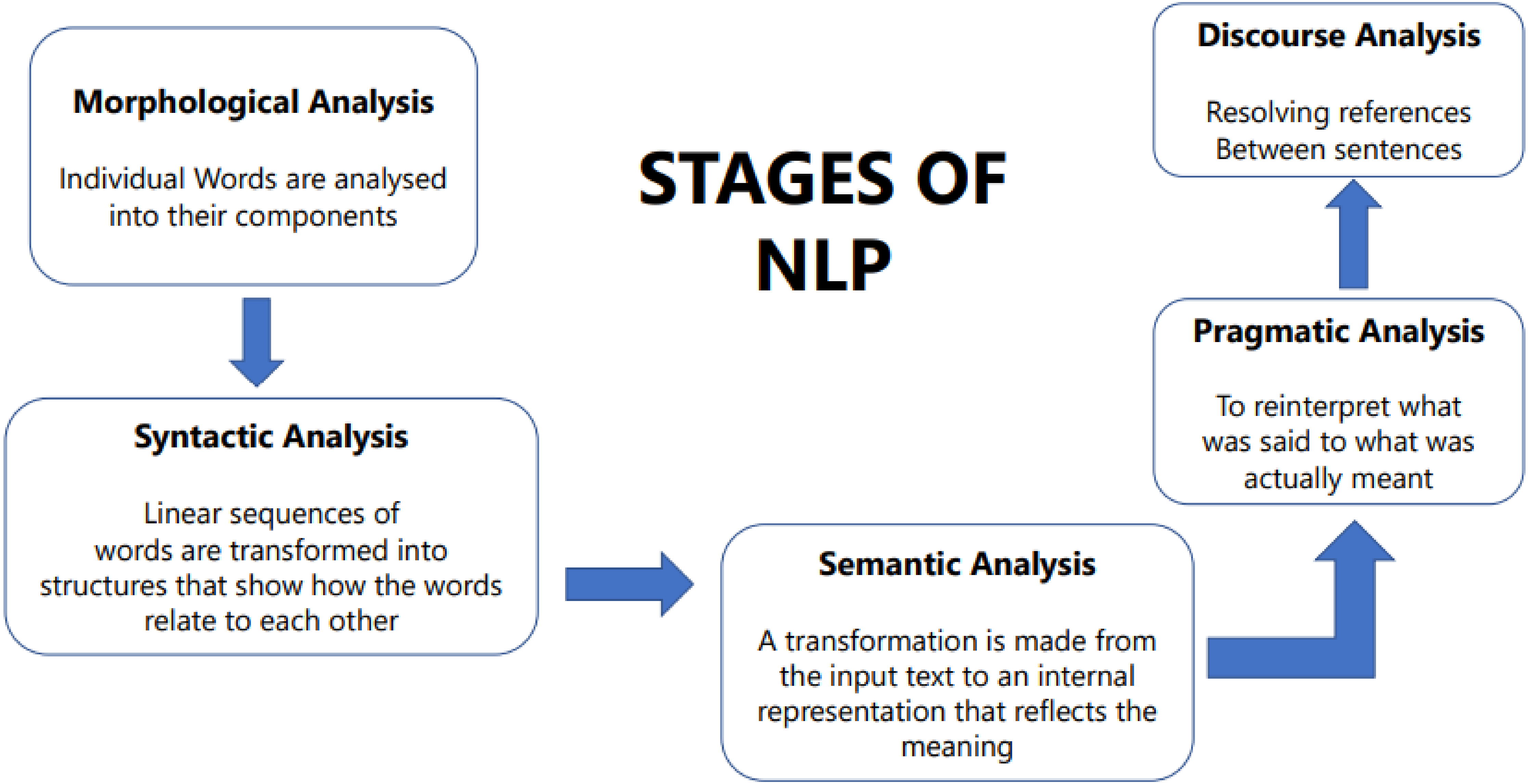
Syntactic Analysis

Semantic Analysis

Pragmatics

Discourse

STAGES OF NLP



Morphological Analysis

Individual Words are analysed into their components

Syntactic Analysis

Linear sequences of words are transformed into structures that show how the words relate to each other

Semantic Analysis

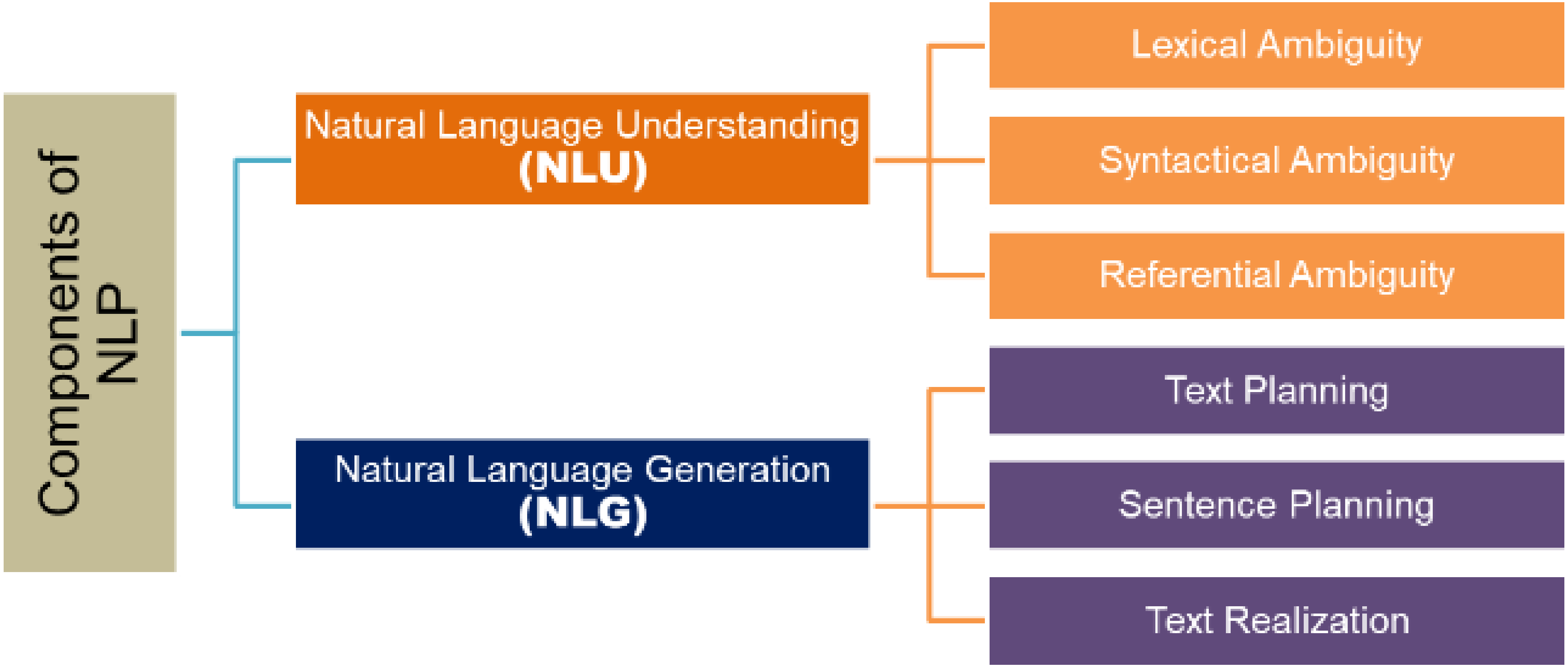
A transformation is made from the input text to an internal representation that reflects the meaning

Pragmatic Analysis

To reinterpret what was said to what was actually meant

Discourse Analysis

Resolving references Between sentences



Components of NLP

Natural Language Understanding **(NLU)**

Lexical Ambiguity

Syntactical Ambiguity

Referential Ambiguity

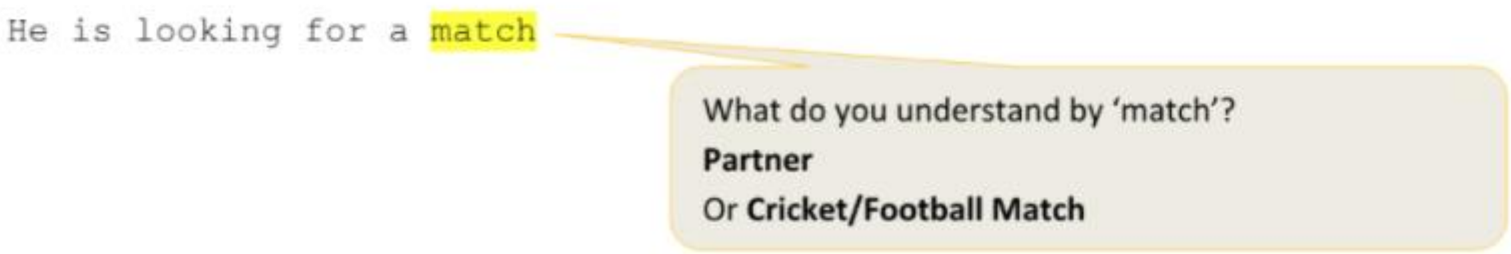
Natural Language Generation **(NLG)**

Text Planning

Sentence Planning

Text Realization

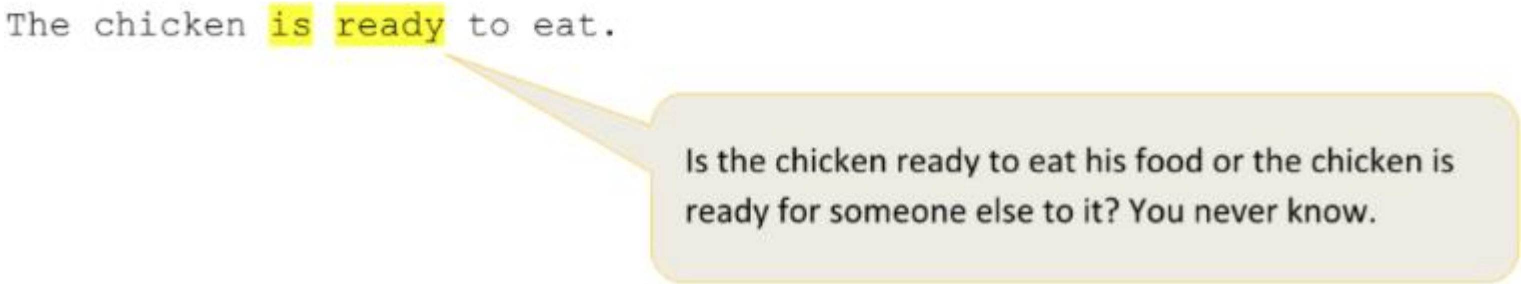
- Ambiguity:



He is looking for a match

What do you understand by 'match'?
Partner
Or **Cricket/Football Match**

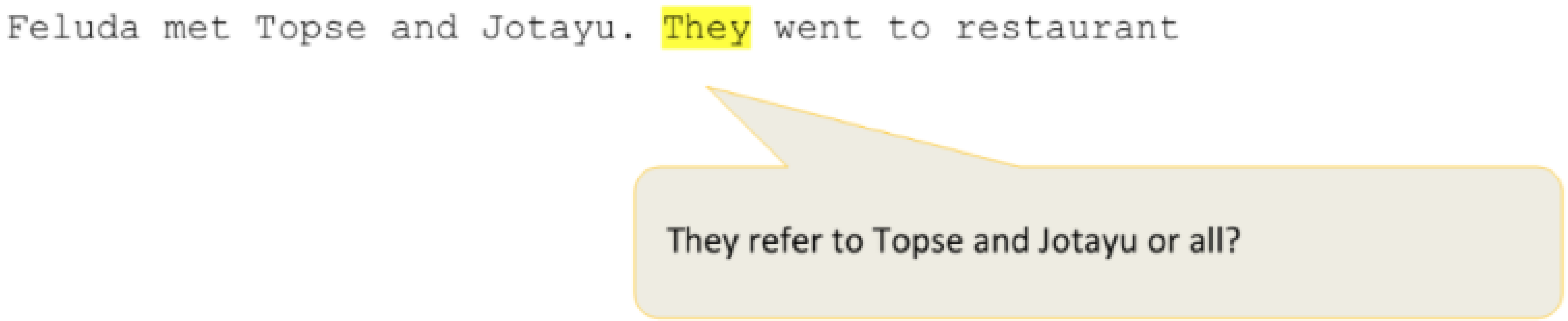
- Lexical Ambiguity: It occurs when a word carries a different sense, i.e. having more than one meaning and the sentence in which it is contained can be interpreted differently depending on its correct sense. Lexical ambiguity can be resolved to some extent using parts-of-speech tagging techniques.



The chicken is ready to eat.

Is the chicken ready to eat his food or the chicken is ready for someone else to it? You never know.

Syntactical Ambiguity means when we see more than one meaning in a sequence of words. It is also termed as grammatical ambiguity.



Feluda met Topse and Jotayu. They went to restaurant

They refer to Topse and Jotayu or all?

Referential Ambiguity: Very often a text mentions as an entity (something/someone), and then refers to it again, possibly in a different sentence, using another word. Pronoun causes ambiguity when it is not clear which noun it is referring to.

Example: The boy told his father about the theft. He was very upset.

He is referentially ambiguous because it can refer to both the boy and the father.

EXAMPLES

- I **bank** on the **bank** for my transactions .[English]

  noun **verb**

- Mujhe **Khaana** **khaana** hai. [Hindi]

  noun **verb**

- **पुजा** देवीची **पुजा** कर. [Marathi]

  noun **verb**

AMBIGUITY OF LANGUAGE

- **I made her duck.**

- Some interpretations of : **I made her duck.**
  - I cooked *duck* for her.
  - I cooked *duck* belonging to her.
  - I created a toy duck which she owns.
  - I caused her to quickly lower her head or body.
  - I used magic and turned her into a *duck*.

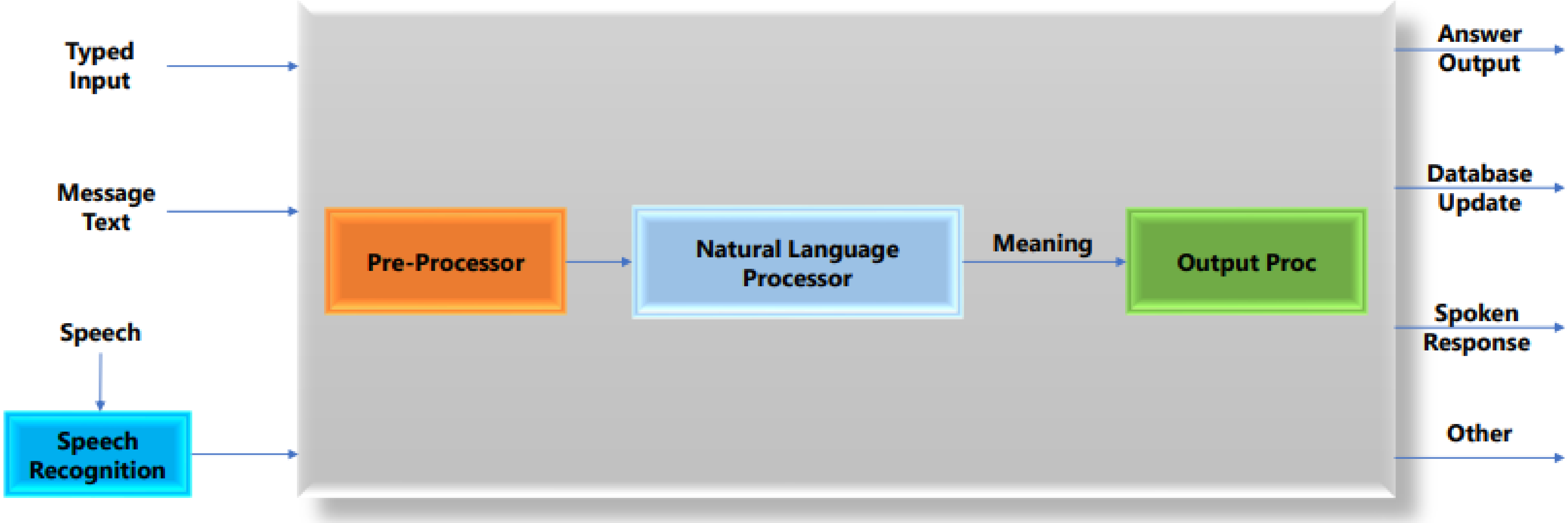
- duck – morphologically and syntactically ambiguous: noun or verb.
- her – syntactically ambiguous: dative or possessive.
- make – semantically ambiguous: cook or create.
- make – syntactically ambiguous:

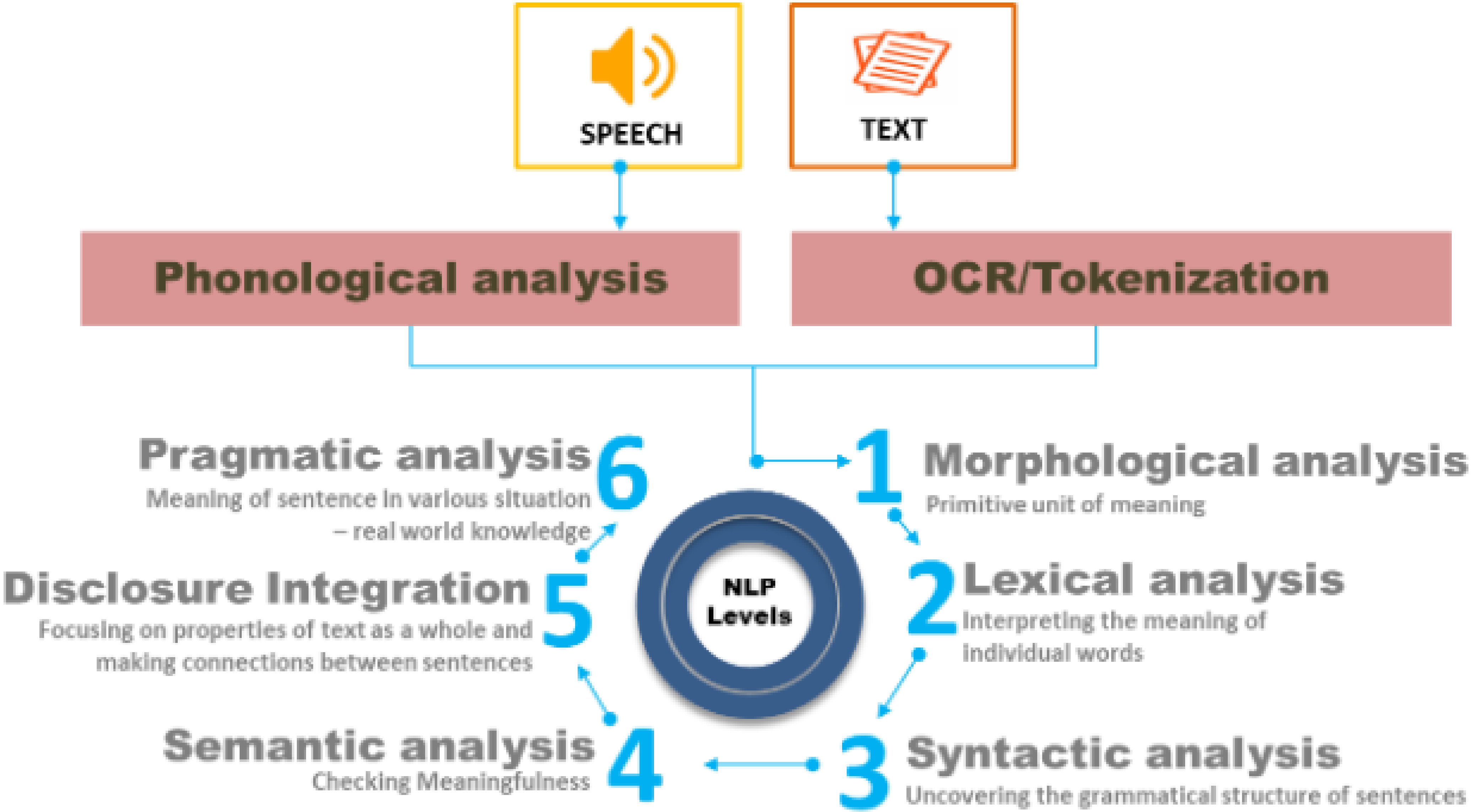
Natural Language Generation (NLG)

- It is the process of producing meaningful phrases and sentences in the form of natural language from some internal representation.

- It involves –
- Text planning – It includes retrieving the relevant content from the knowledge base.
- Sentence planning – It includes choosing required words, forming meaningful phrases, and setting the tone of the sentence.
- Text Realization – It is mapping sentence plan into sentence structure.

Generic NLP System





SPEECH

TEXT

Phonological analysis

OCR/Tokenization

NLP Levels

1 Morphological analysis
Primitive unit of meaning

2 Lexical analysis
Interpreting the meaning of individual words

3 Syntactic analysis
Uncovering the grammatical structure of sentences

4 Semantic analysis
Checking Meaningfulness

5 Disclosure Integration
Focusing on properties of text as a whole and making connections between sentences

6 Pragmatic analysis
Meaning of sentence in various situation – real world knowledge

- Phonological Analysis: This level is applied only if the text origin is a speech. It deals with the interpretation of speech sounds within and across words. Speech sound might give a big hint about the meaning of a word or a sentence.
- Morphological Analysis: Deals with understanding distinct words according to their morphemes (the smallest units of meanings) .
- Example: “unhappiness ”. It can be broken down into three morphemes (prefix, stem, and suffix), with each conveying some form of meaning:
- the prefix un- refers to “not being”
- the suffix -ness refers to “a state of being”.
- The stem- happy is considered a free morpheme since it is a “word” in its own right.
- Bound morphemes (prefixes and suffixes) require a free morpheme to which they can be attached, and can therefore not appear as a “word” on their own.

Lexical Analysis: It involves identifying and analyzing the structure of words. Lexicon of a language means the collection of words and phrases in a language.

- Lexical analysis is dividing the whole chunk of text into paragraphs, sentences, and words.
- The most common lexicon normalization practices are Stemming:

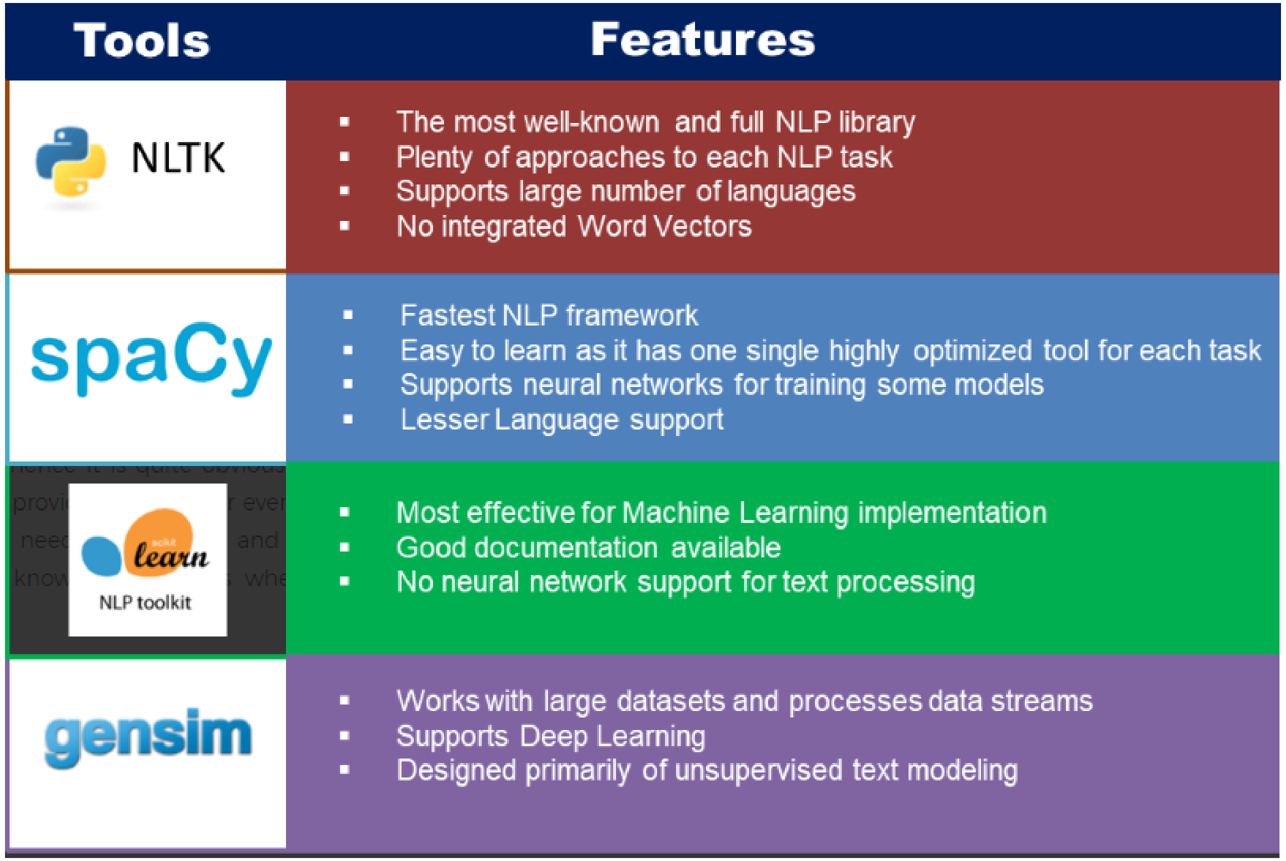
1. Stemming: Stemming is a rudimentary rule-based process of stripping the suffixes (“ing”, “ly”, “es”, “s” etc) from a word.

2. Lemmatization: Lemmatization, is an organized & step by step procedure of obtaining the root form of the word, it makes use of vocabulary (dictionary importance of words) and morphological analysis (word structure and grammar relations)

- Syntactic Analysis: Deals with analyzing the words of a sentence so as to uncover the grammatical structure of the sentence. E.g.. "Colourless green idea." This would be rejected by the Symantec analysis as colorless here; green doesn't make any sense.
- Semantic Analysis: Determines the possible meanings of a sentence by focusing on the interactions among word-level meanings in the sentence. Some people may think it's the level that determines the meaning, but actually all the level do. The semantic analyzer disregards sentence . Example:"hot ice-cream".
- Discourse Integration: Focuses on the properties of the text as a whole that convey meaning by making connections between component sentences. It means a sense of the context. The meaning of any single sentence which depends upon that sentences. It also considers the meaning of the following sentence. For example, the word "that" in the sentence "He wanted that" depends upon the prior discourse context. Example: Ram and Shyam were playing. After few hours He left the game after sometime.

- Pragmatic Analysis: Explains how extra meaning is read into texts without actually being encoded in them. This requires much world knowledge, including the understanding of intentions, plans, and goals. Consider the following two sentences:
- Example:
- The city police refused the demonstrators a permit because they feared violence.
- The city police refused the demonstrators a permit because they advocated revolution.
- The meaning of “they” in the 2 sentences is different. In order to figure out the difference, world knowledge in knowledge bases and inference modules should be utilized.
- Pragmatic analysis helps users to discover this intended effect by applying a set of rules that characterize cooperative dialogues. E.g., "close the window?" should be interpreted as a request instead of an order.

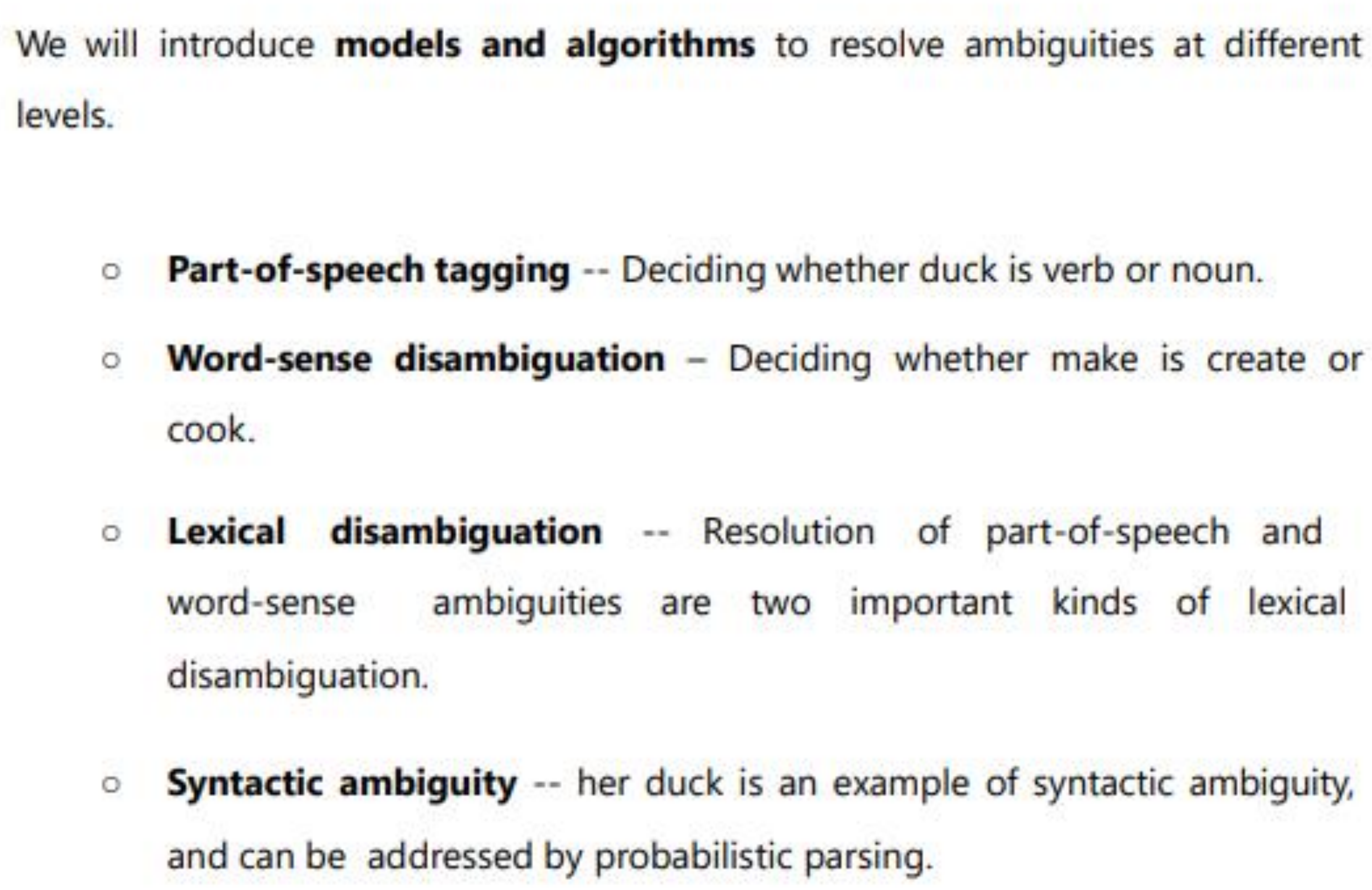
Widely used NLP Libraries



Tools	Features
NLTK	▪ The most well-known and full NLP library ▪ Plenty of approaches to each NLP task ▪ Supports large number of languages ▪ No integrated Word Vectors
spaCy	▪ Fastest NLP framework ▪ Easy to learn as it has one single highly optimized tool for each task ▪ Supports neural networks for training some models ▪ Lesser Language support
scikit learn NLP toolkit	▪ Most effective for Machine Learning implementation ▪ Good documentation available ▪ No neural network support for text processing
gensim	▪ Works with large datasets and processes data streams ▪ Supports Deep Learning ▪ Designed primarily of unsupervised text modeling

https://www.ibm.com/demos/live/natural-language-understanding/self-service/home

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.

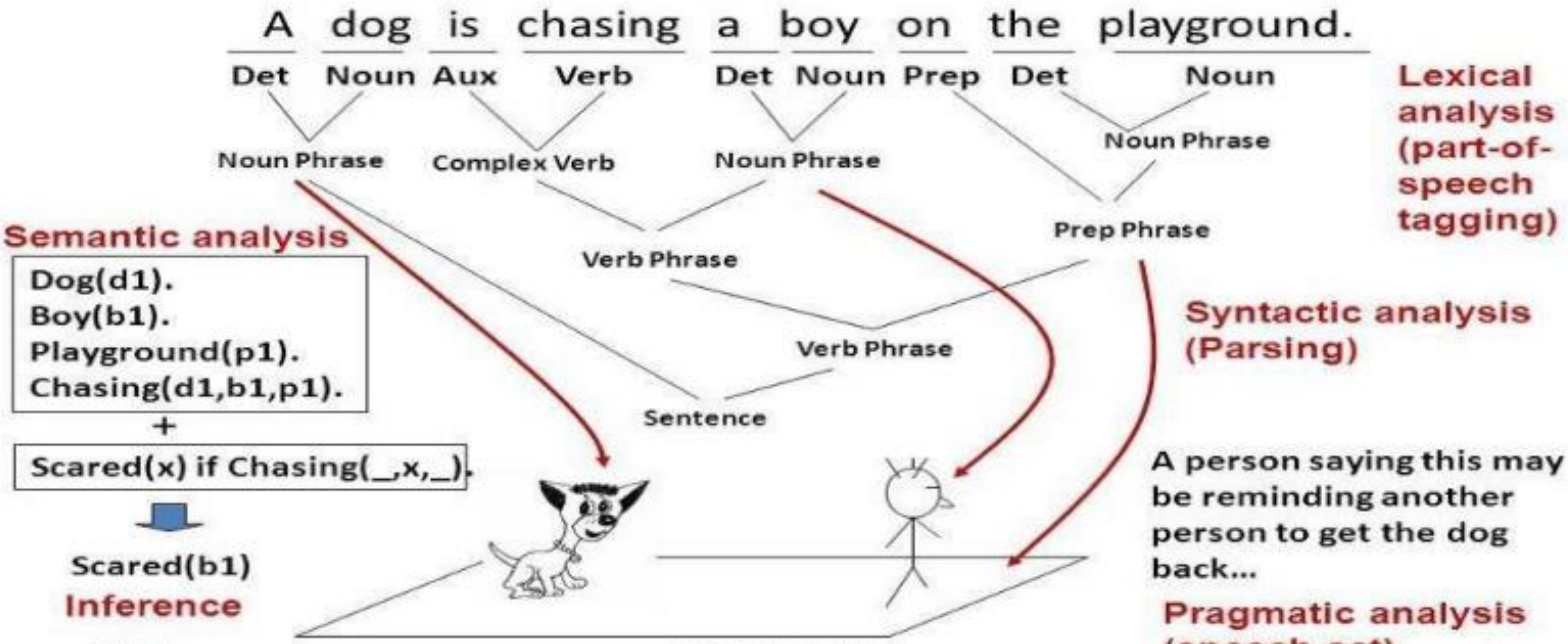
- Statistical approach to handle ambiguity:

1.Probabilistic model

2.Part of speech

- Rule-Based
- Markov Model
- Maximum entropy
- HMM-based tagger
- 3.Machine Learning Approach

An example of NLP



Lexical analysis (part-of-speech tagging)

Syntactic analysis (Parsing)

Semantic analysis

Dog(d1).
Boy(b1).
Playground(p1).
Chasing(d1,b1,p1).

+

Scared(x) if Chasing(_,x,_).

Scared(b1)

Inference

A person saying this may be reminding another person to get the dog back…

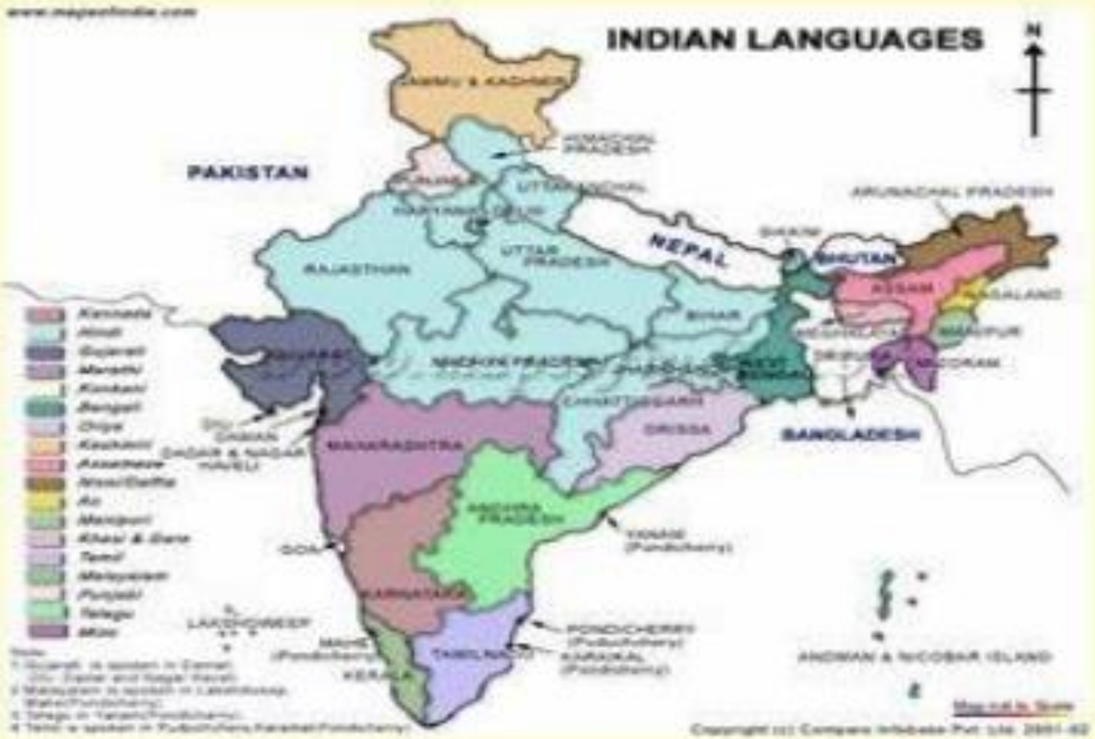
Pragmatic analysis (speech act)

Challenges of NLP

- Breaking the sentence
- Tagging the parts of speech (POS) and generating dependency graphs
- Building the appropriate vocabulary
- Linking different components of vocabulary
- Word Sense Disambiguation
- Pronoun Resolution
- Setting the context
- Extracting semantic meanings
- Extracting named entities
- Transforming unstructured data into structured format

Great Linguistic Diversity

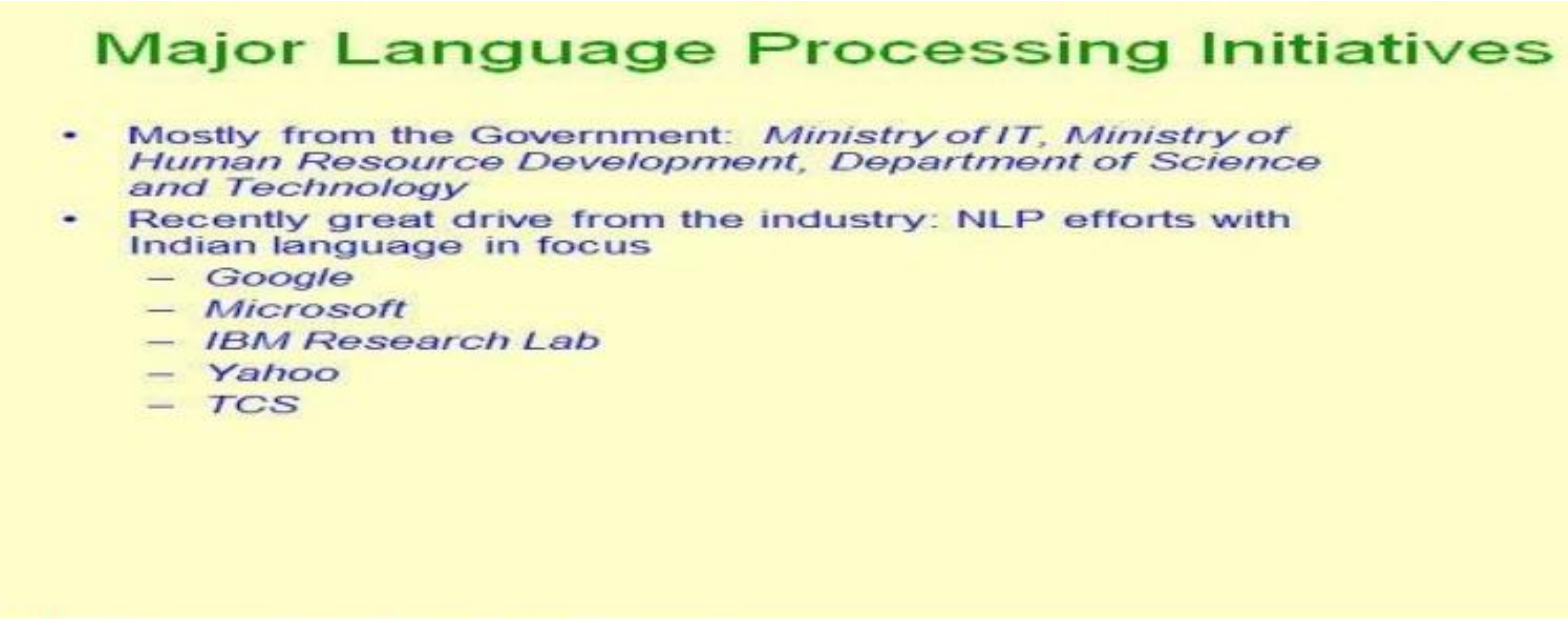
- Major streams
  - Indo European
  - Dravidian
  - Sino Tibetan
  - Austro-Asiatic
- Some languages are ranked within 20 in the world in terms of the populations speaking them
  - Hindi and Urdu: 5th (~500 milion)
  - Bangla: 7th (~300 million)
  - Marathi 14th (~70 million)



Technology Development in Indian Languages (TDIL)

- Started by the Ministry of IT in 2000
- 13 resource center across the country
- Responsibility for two languages: one major and one minor
- For example,
  - IIT Bombay: *Marathi* and *Konkani*
  - IIT Kanpur: *Hindi* and *Nepali*
  - ISI Kolkata: *Bangla* and *Santhaali*
  - Anna University: *Tamil*

INDIAN LANGUAGES
N
PAKISTAN
NEPAL
BHUTAN
BANGLADESH
JAMMU & KASHMIR
RAJASTHAN
UTTAR PRADESH
MADHYA PRADESH
BIHAR
ASSAM
NAGALAND
MANIPUR
MIZORAM
ORISSA
CHHATTISGARH
MAHARASHTRA
ANDHRA PRADESH
KARNATAKA
TAMILNADU
KERALA
GOA
LAKSHDWEEP
DAMAN
DADAR & NAGAR HAVELI
PONDICHERRY
ANDMAN & NICOBAR ISLAND
Kannada
Hindi
Gujarati
Marathi
Konkani
Bengali
Oriya
Kashmiri
Assamese
Ao
Manipuri
Tamil
Malayalam
Punjabi
Telegu
Mizo



Major Language Processing Initiatives

- Mostly from the Government: *Ministry of IT, Ministry of Human Resource Development, Department of Science and Technology*
- Recently great drive from the industry: NLP efforts with Indian language in focus
  - *Google*
  - *Microsoft*
  - *IBM Research Lab*
  - *Yahoo*
  - *TCS*

NLP Association of India: 4 years old:

Recently efforts are on making tools and resources freely available on the websit of **NLPAI**

Industry Scenario: NLP

•How to use NLP to increase the **search engine performance** (*precision, recall, speed)*

•*Google, Rediff, Yahoo, IRL, Microsoft*:

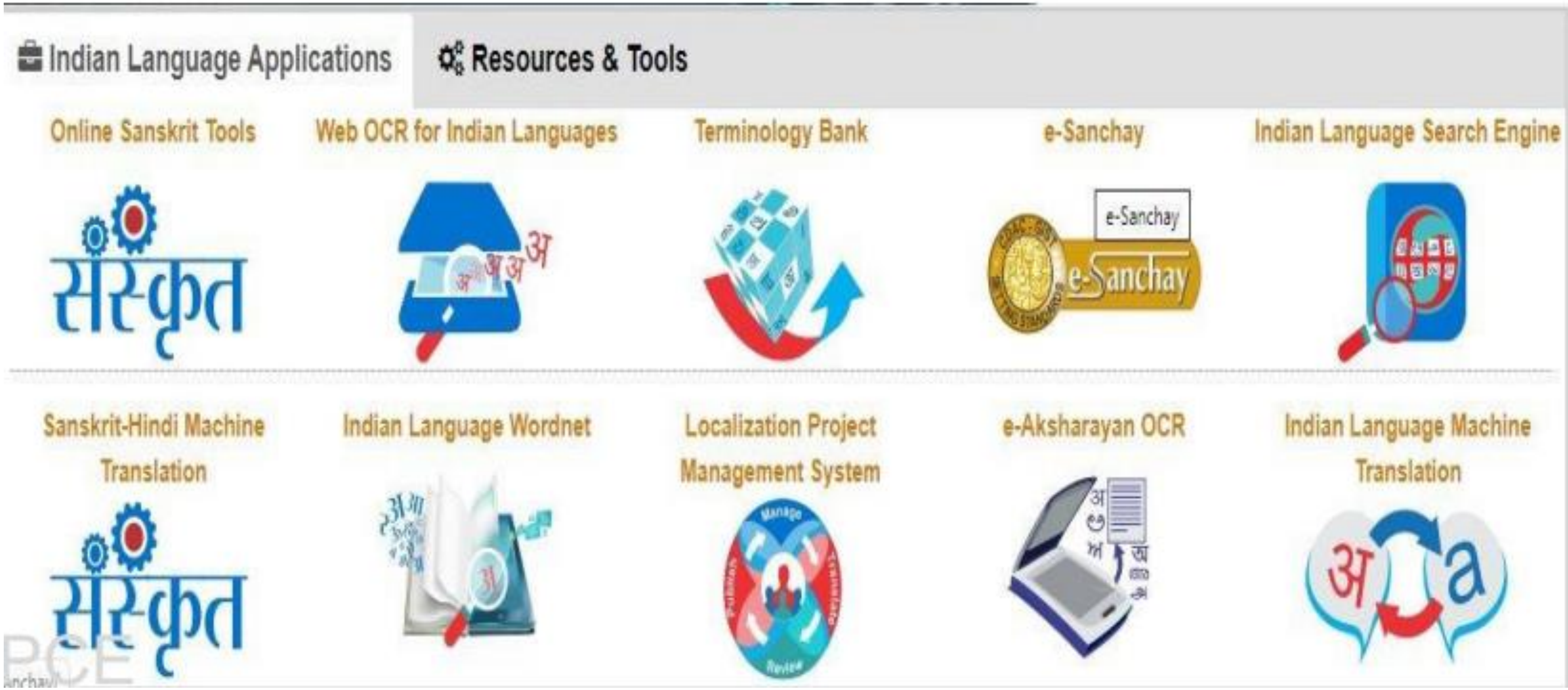
all have **search engine, IR, IE R& D projects** outsourced from USA and being carried out in India.

•IBM Research lab has massive **English Hindi Parallel Corpora** (news domain) :Statistical Machine Translation

•Microsoft India at Bangalore has opened a **Multilingual Computing Division**

•*Google* and *Yahoo India* is actively pursuing **IL search engine:**

http://tdil-dc.in/ : Indian Language Technology development center


Indian Language Applications
Resources & Tools
Online Sanskrit Tools
Web OCR for Indian Languages
Terminology Bank
e-Sanchay
Indian Language Search Engine
संस्कृत
e-Sanchay
e-Sanchay
Sanskrit-Hindi Machine Translation
Indian Language Wordnet
Localization Project Management System
e-Aksharayan OCR
Indian Language Machine Translation
संस्कृत
Manage
Translate
Review
Publish
अ
a

Text Pre Processing

- It is a process of cleaning and preparing the text.
- NLTK and RE python libraries are used.
- Techniques are:

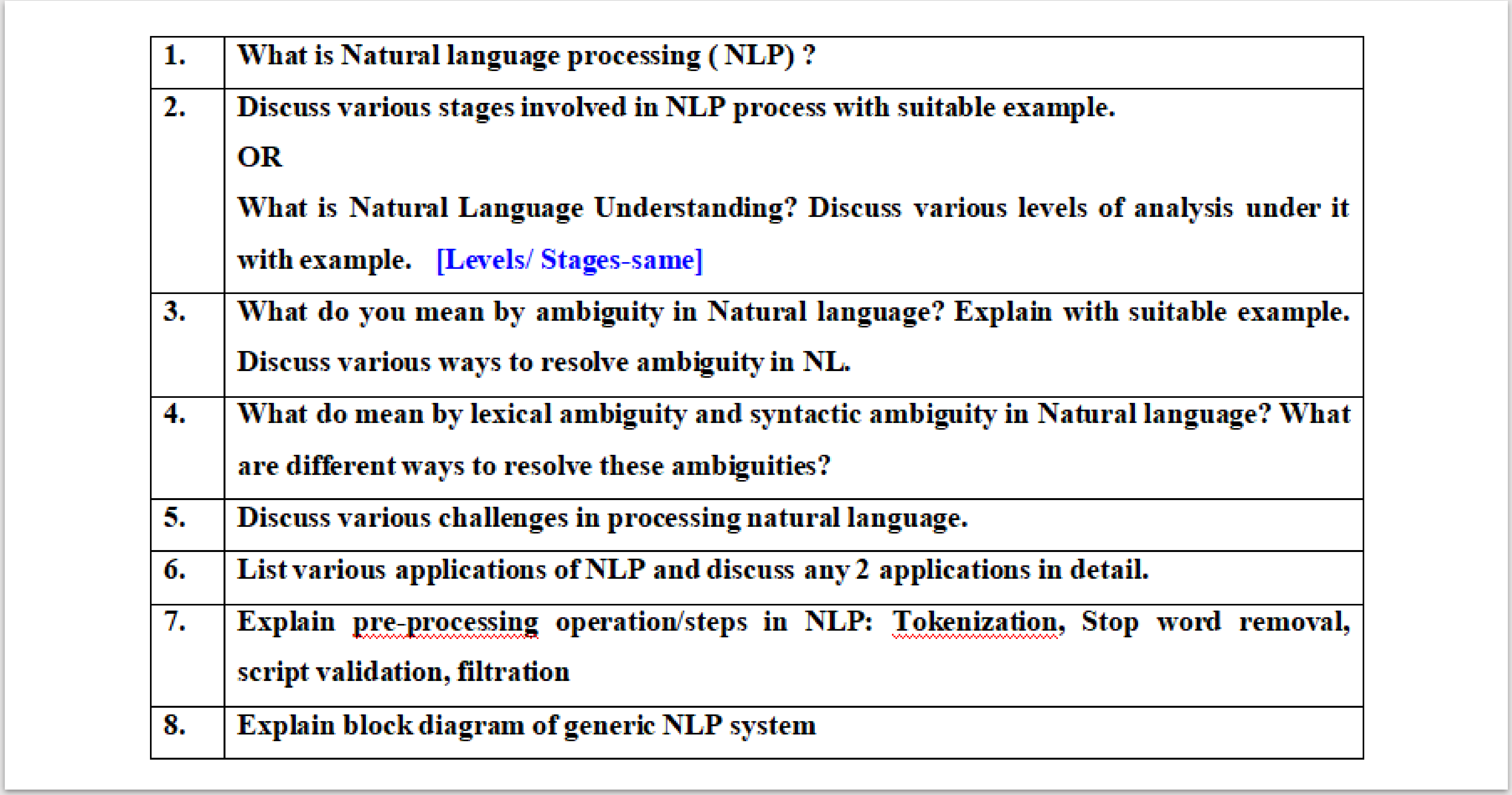
1. Removing HTML tags
2. Remove White spaces
3. Accented character
4. Expand Contractions
5. Remove special character
6. Convert all words into Lower Case
7. Remove Punctuation mark
8. Remove words and digits containing digits
9. Remove Stop words
10. Stemming and Lemmatization
11. Script Validation

- **Tokenization**: Tokenization is the process of tokenizing or splitting a string, or text into a list of tokens. One can think of tokens as parts like a word is a token in a sentence, and a sentence is a token in a paragraph. It is important because the meaning of the text can be interpreted through analysis of the words present in the text.

➢ Token: Non-unique words of a sentence

➢ Type: Unique words of a sentence.

- **Stopword Removal**: Stop word removal is one of the most commonly used preprocessing steps across different NLP applications. The idea is simply to remove the words that occur commonly across all the documents in the corpus. Typically, articles and pronouns are generally classified as stop words. These words have no significance in some of the NLP tasks like information retrieval and classification.



1.	What is Natural language processing (NLP) ?
2.	Discuss various stages involved in NLP process with suitable example. OR What is Natural Language Understanding? Discuss various levels of analysis under it with example. [Levels/ Stages-same]
3.	What do you mean by ambiguity in Natural language? Explain with suitable example. Discuss various ways to resolve ambiguity in NL.
4.	What do mean by lexical ambiguity and syntactic ambiguity in Natural language? What are different ways to resolve these ambiguities?
5.	Discuss various challenges in processing natural language.
6.	List various applications of NLP and discuss any 2 applications in detail.
7.	Explain pre-processing operation/steps in NLP: Tokenization, Stop word removal, script validation, filtration
8.	Explain block diagram of generic NLP system