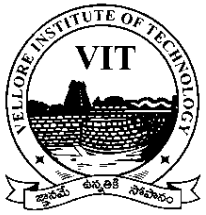


VIT-AP UNIVERSITY

Natural Language Processing

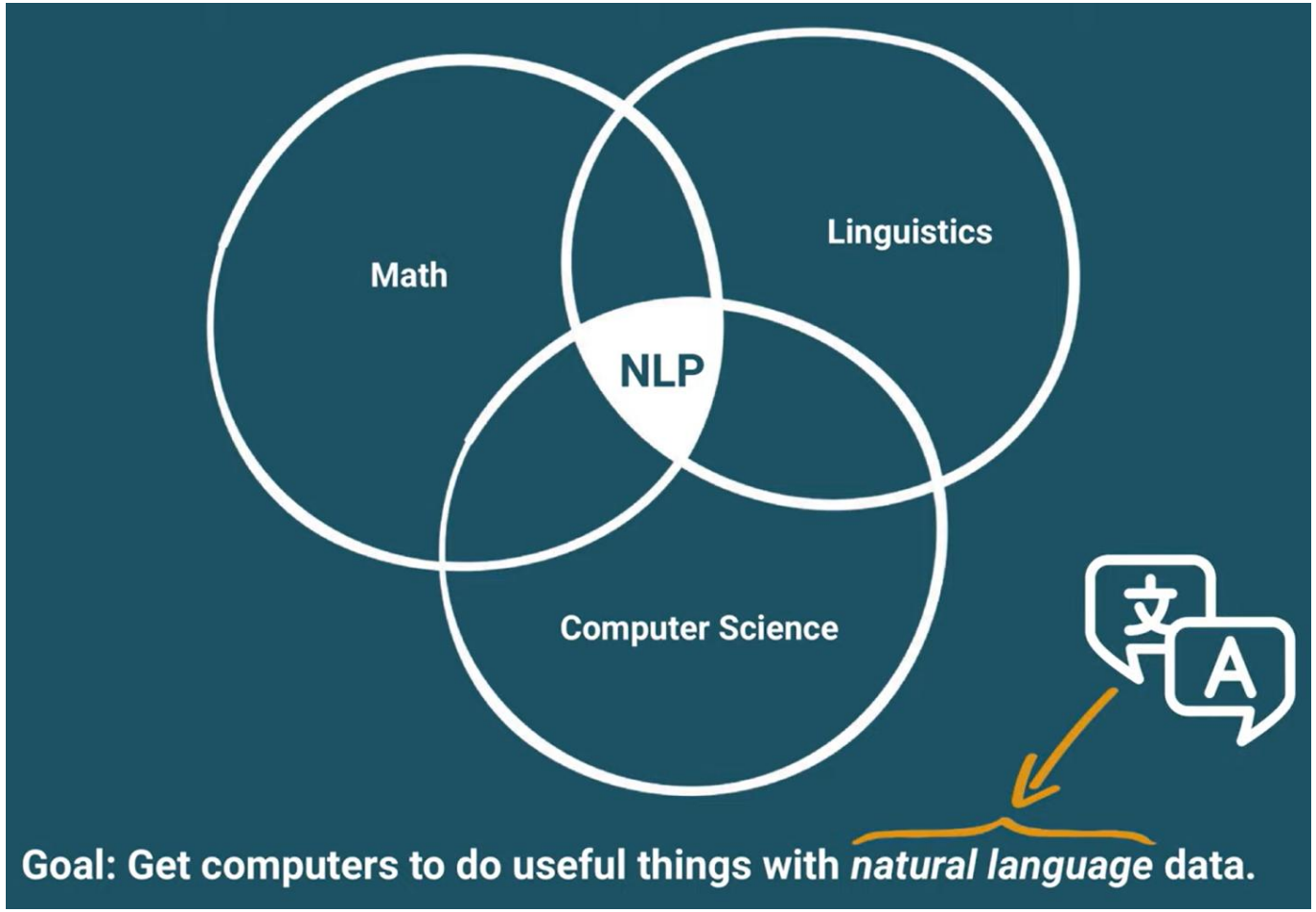
(Course Code: CSE 3015)

Module-1:Lecture-1: INTRODUCTION TO NLP

Gundimeda Venugopal, Professor of Practice, SCOPE

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.

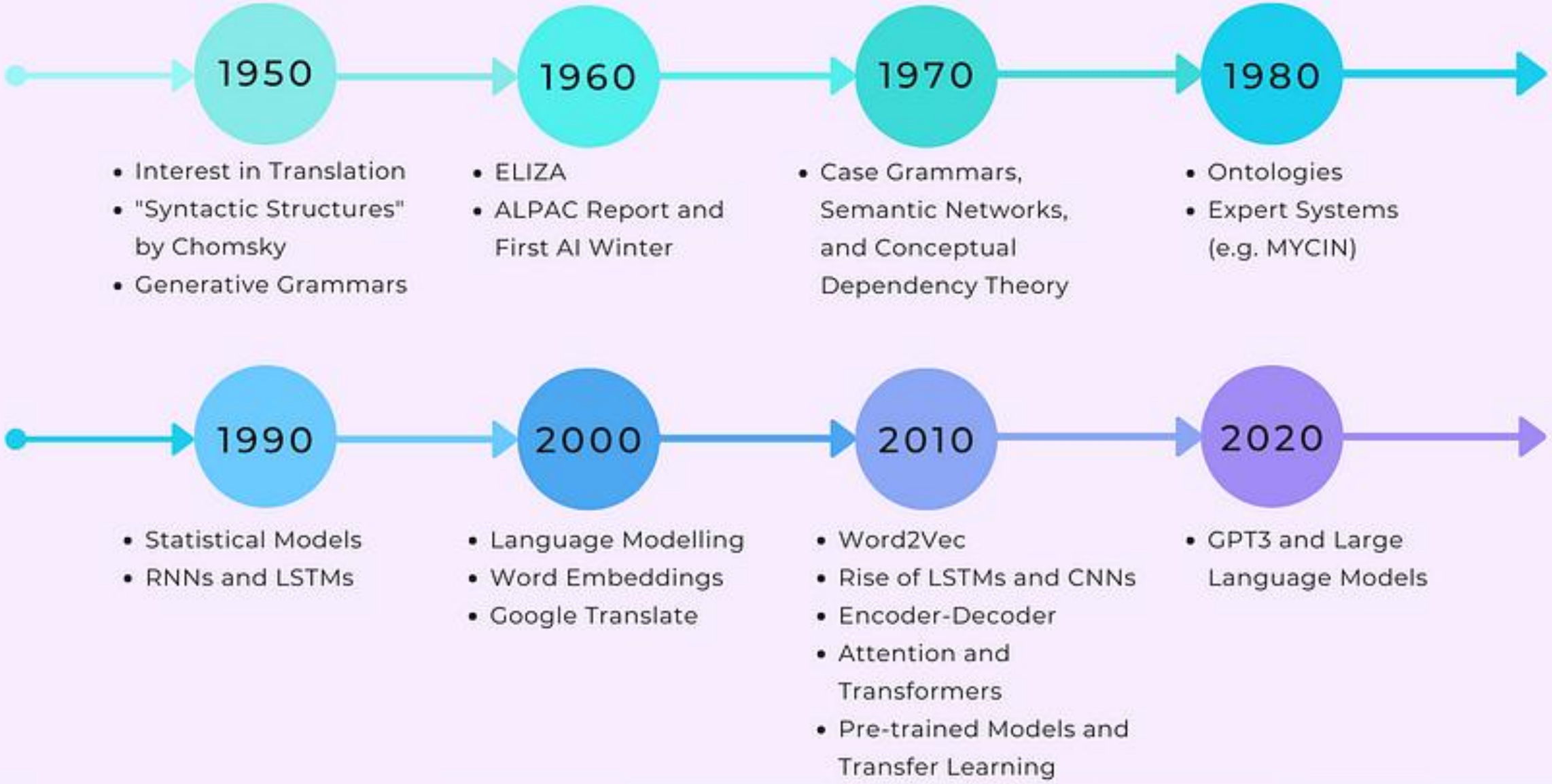


Linguistics is the scientific study of language.

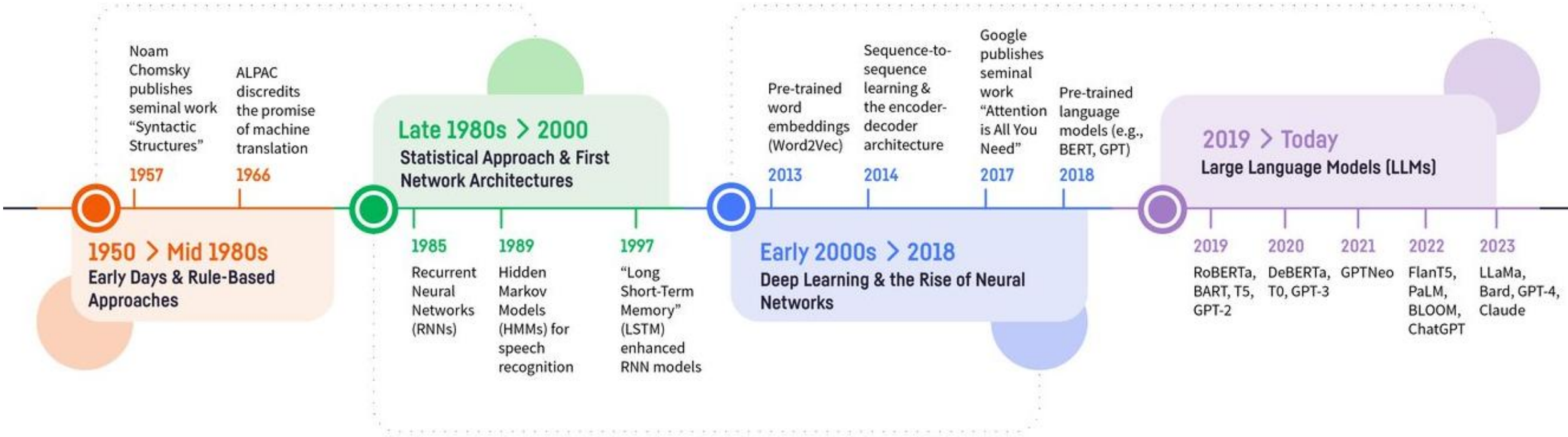
NLP - an inter-disciplinary Field

- 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 identified? How can knowledge and reasoning be modeled?
- **Computer Science**: Algorithms for automation, 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?

A Brief Timeline of NLP



The History of NLP



Linguistics

The Scientific Study of Language

Linguistics

Sounds and Sound systems

Words and sentences

Study of meaning in a language

Phonetics

How speech sounds are produced, perceived and analysed acoustically?

Phonology

Principles of sounds In a language

Morphology

How words are built?

Syntax

Sentence structure

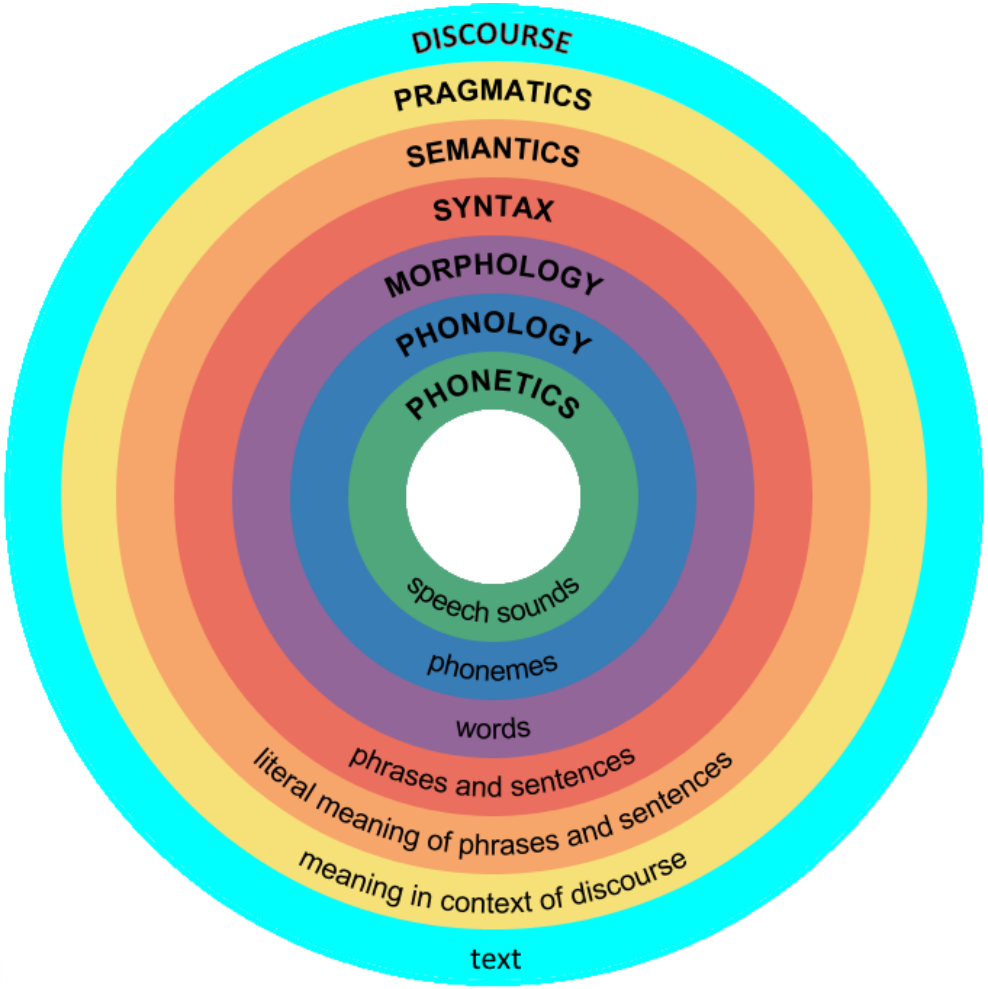
Semantics

Meaning of words and sentences

Pragmatics

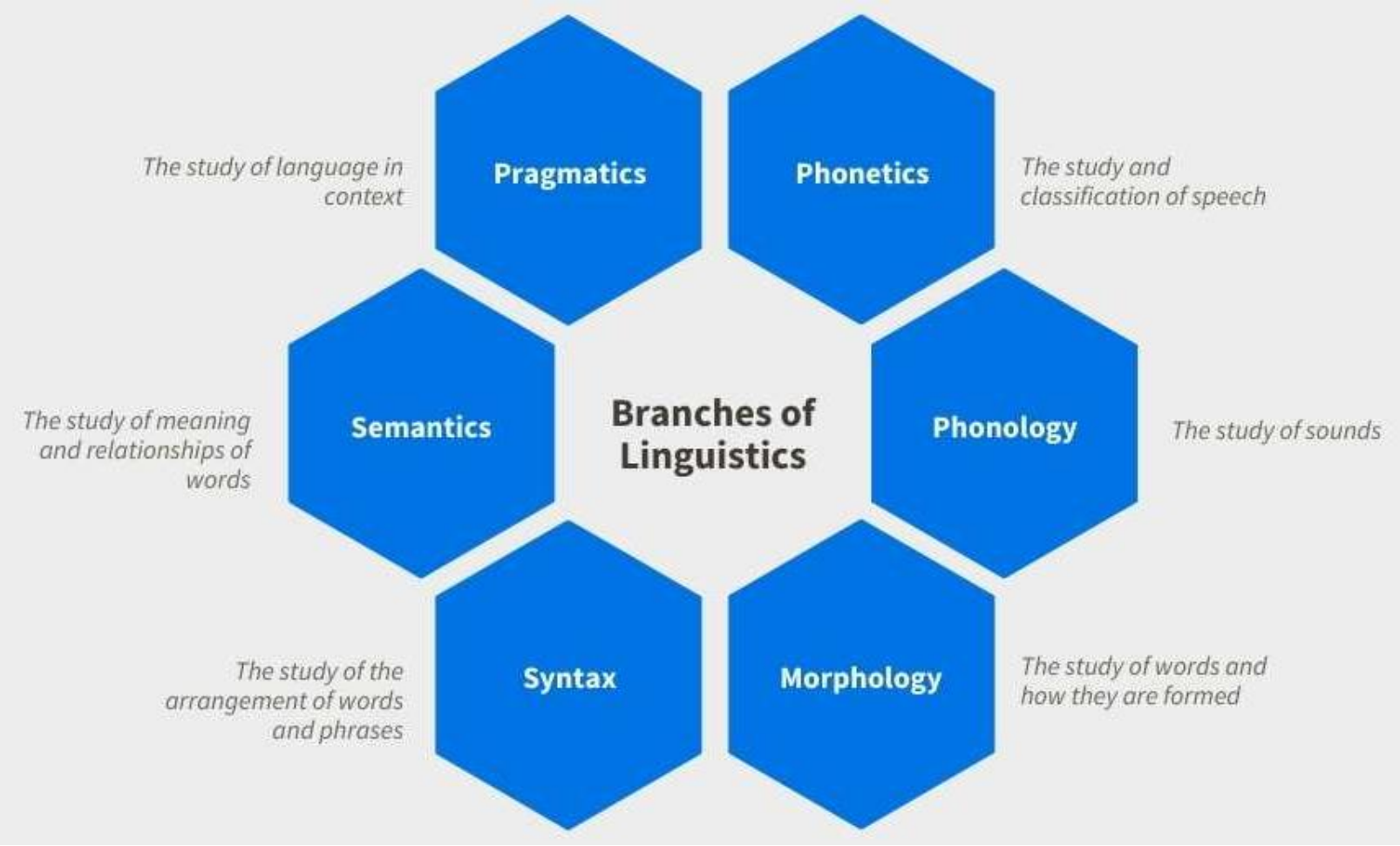
Use of language and its effects

Major Levels of Linguistic Structure



Linguistics

Linguistics is the scientific study of language.



- Phonetics – how do humans produce and perceive acoustic or visual signals?
- Phonology – how are acoustic signals organized in spoken languages or dialects?
- Morphology – how are words formed?
- Syntax – how are sentences formed?
- Semantics – what do linguistic expressions or signals mean?
- Pragmatics – how does meaning depend on context?

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.

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.

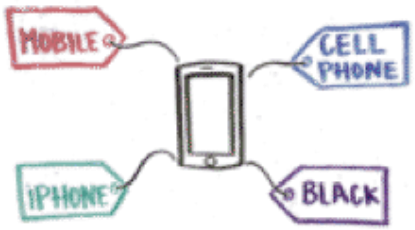
  e.g., Batsman wears gloves while playing cricket. South Indians eat with their hands. Neem Plant leaves are green. Cricket pitch length is 22 yards. 1 feet is 12 inches. 1 m is 100 cm.

Knowledge Representations

Knowledge Representations

Free-text tags

FOLKSONOMY



CONTROLLED LIST

List of pre-defined terms.
Improves consistency.

Pre-defined terms & synonyms.
Hierarchical relationships.
Improves consistency.
Allows for parent/child content relationships.

TAXONOMY

THESAURUS

Pre-defined terms & synonyms.
Hierarchical relationships.
Associative ("related to") relationships.
Scope notes.
Increased expressiveness.

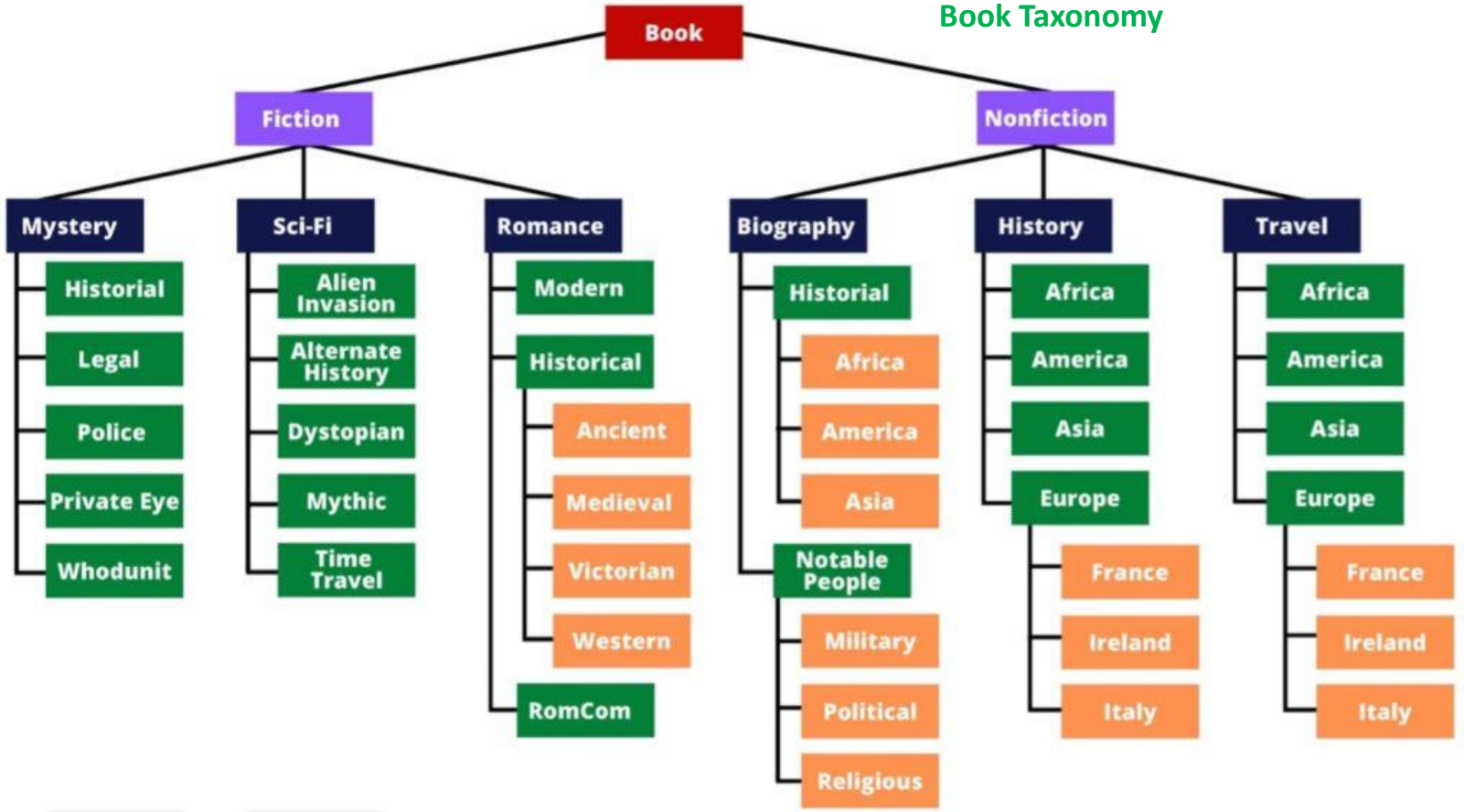
Scope Notes.
Pre-defined classes & properties.
Expanded relationship types.
Increased expressiveness.
Inference.

ONTOLOGY

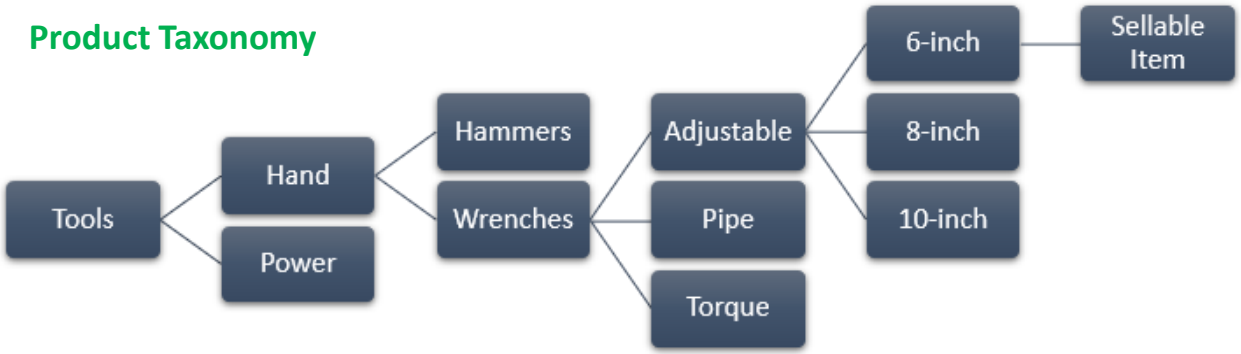
Taxonomy

- A taxonomy is a hierarchical framework, or schema, for the organization of organisms, inanimate objects, events, and/or concepts.
- We see taxonomies daily as humans, and we don't give them much thought.
- Taxonomies are the facets, filters, and search suggestions commonly seen on modern websites.

Book Taxonomy

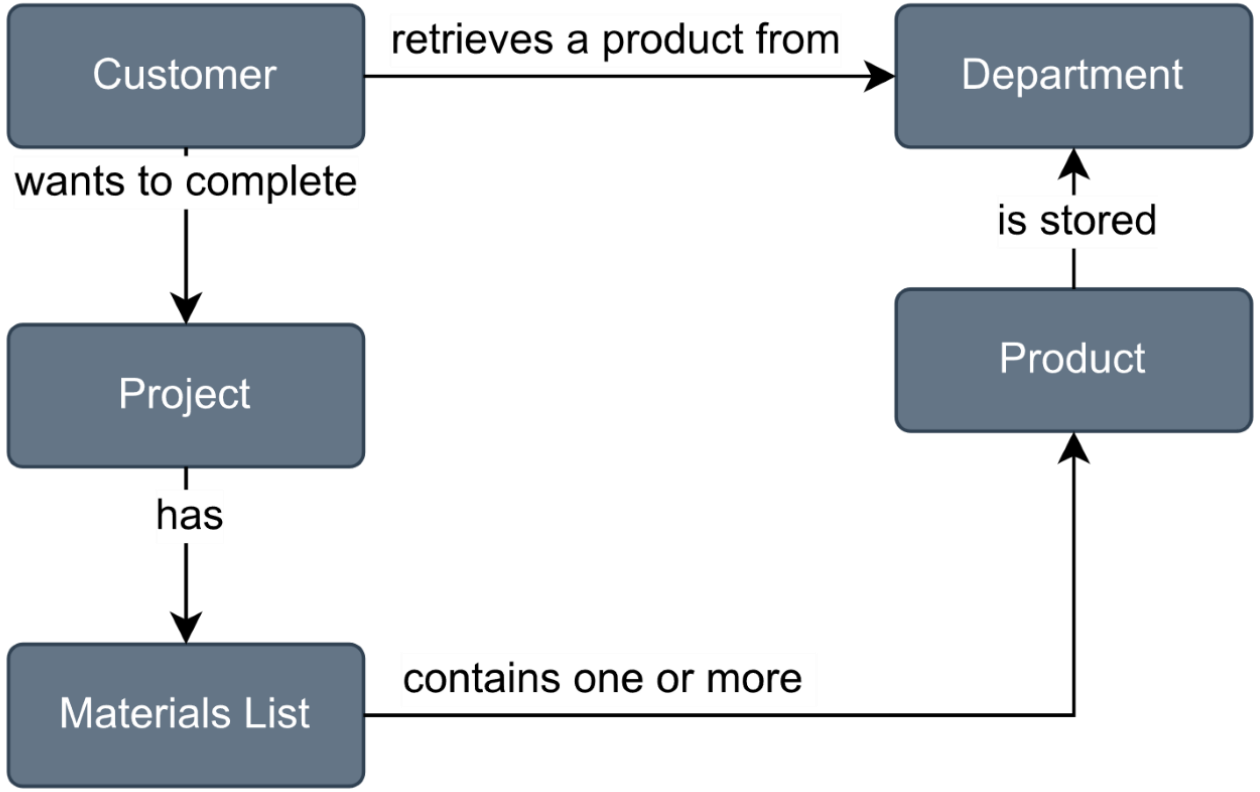


Product Taxonomy



Ontology

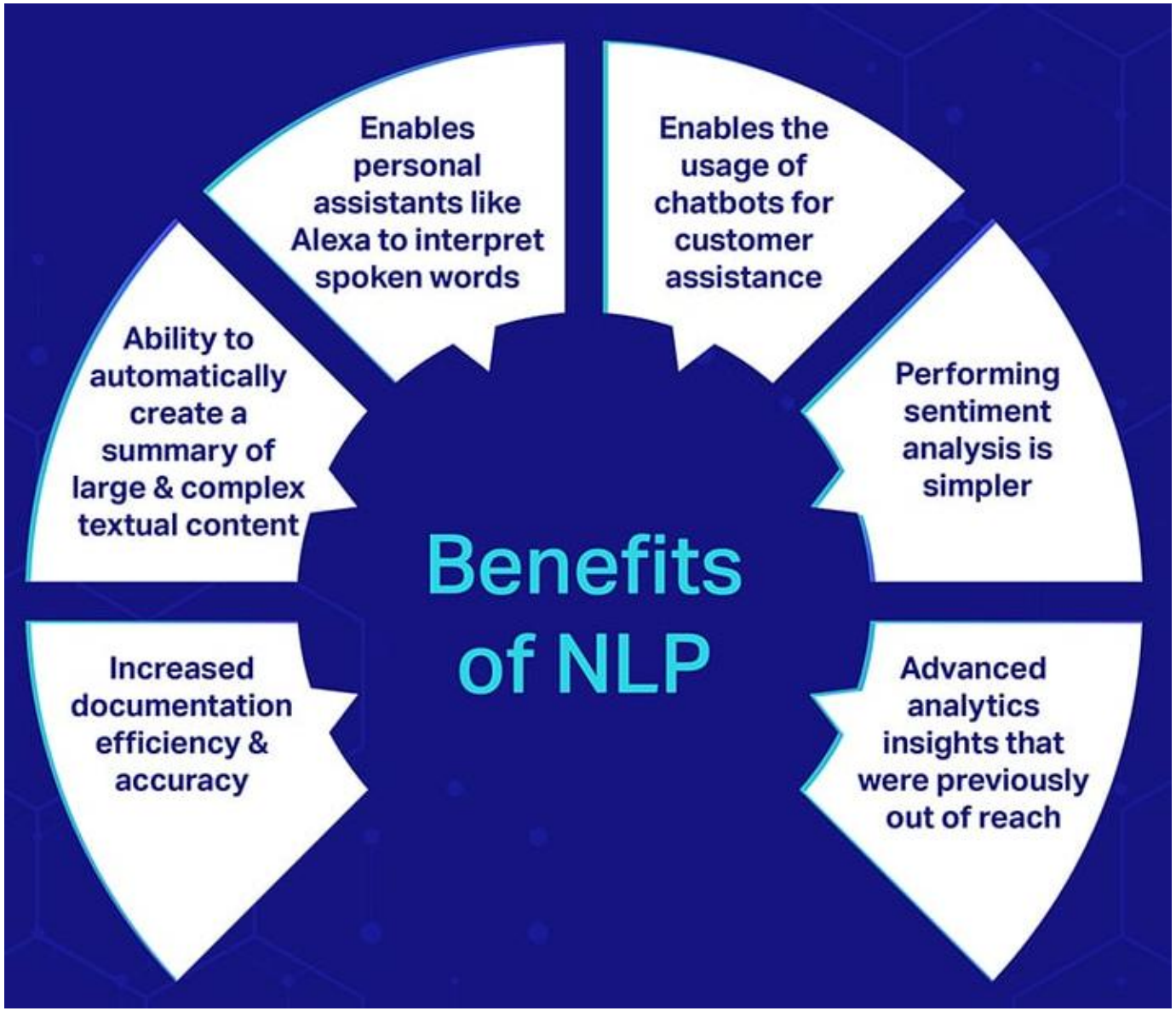
- an ontology "encompasses a representation, formal naming, and definition of the categories, properties, and relations between the concepts, data, and entities that substantiate one, many, or all domains of discourse."
- In other words, ontologies allow us to organize the jargon of a subject area into a controlled vocabulary, thereby decreasing complexity and confusion. Without ontologies, you have no frame of reference.
- Ontologies are "essential in modern architectural patterns to ensure data quality, governance, findability, interoperability, accessibility, and reusability."
- Represented using Resource Description Format (RDF)



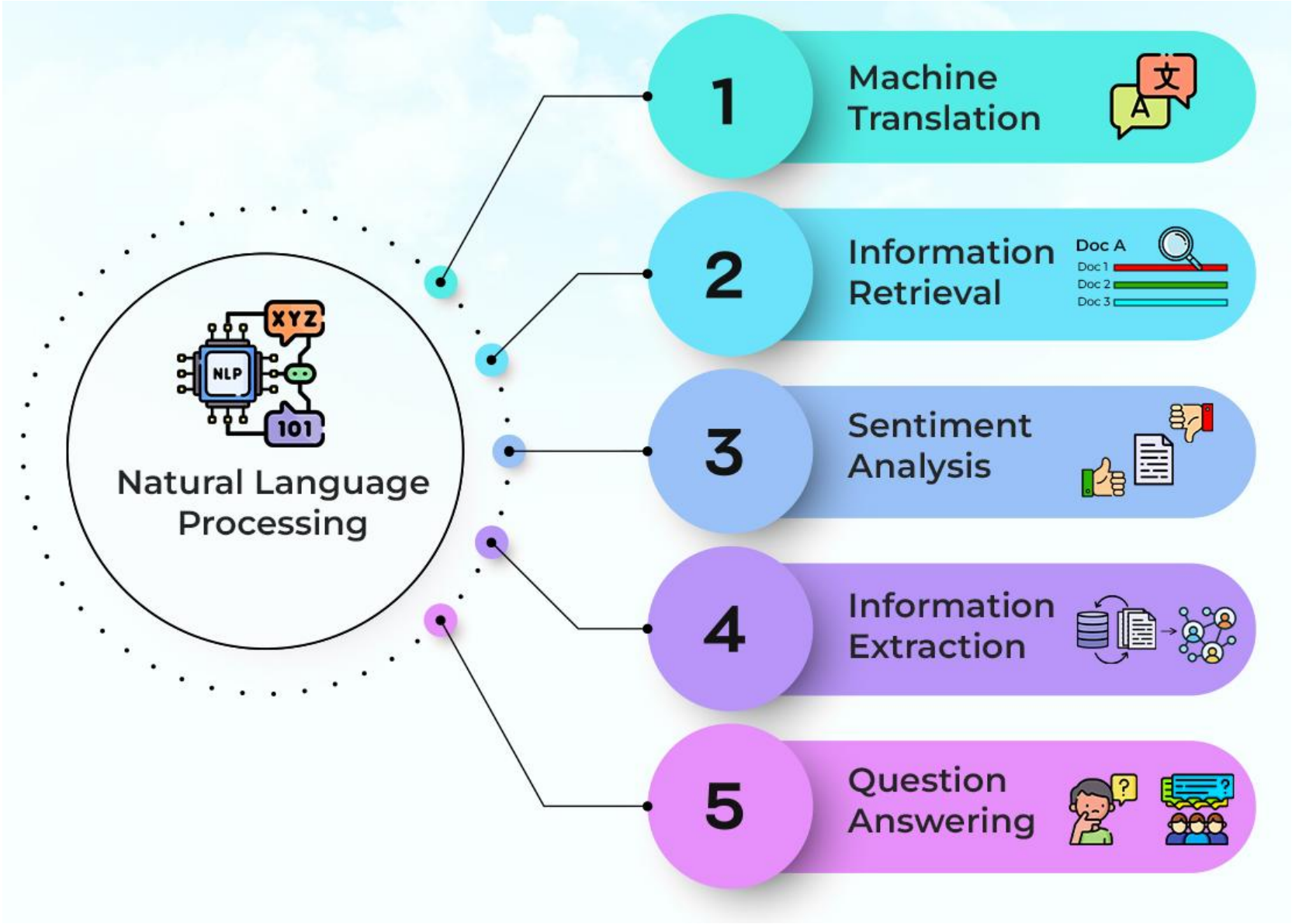
Each box is an *entity* in the ontology.
The arrows represent the relationships among entities, where each label describes the specific nature of that relationship.

NLP Applications

Benefits of NLP



NLP Applications (1/3)

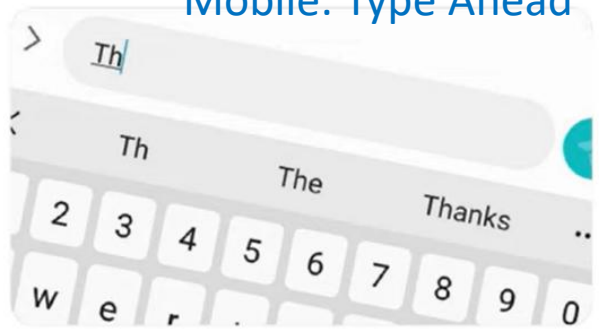


NLP Applications (2/3)

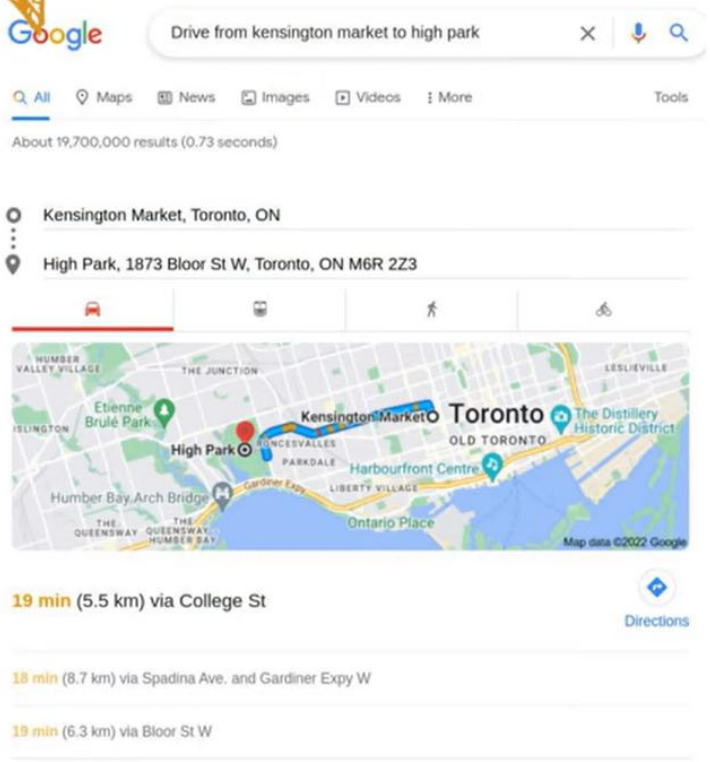
NLP Applications 3/3

Email: Spam Filter

Mobile: Type Ahead

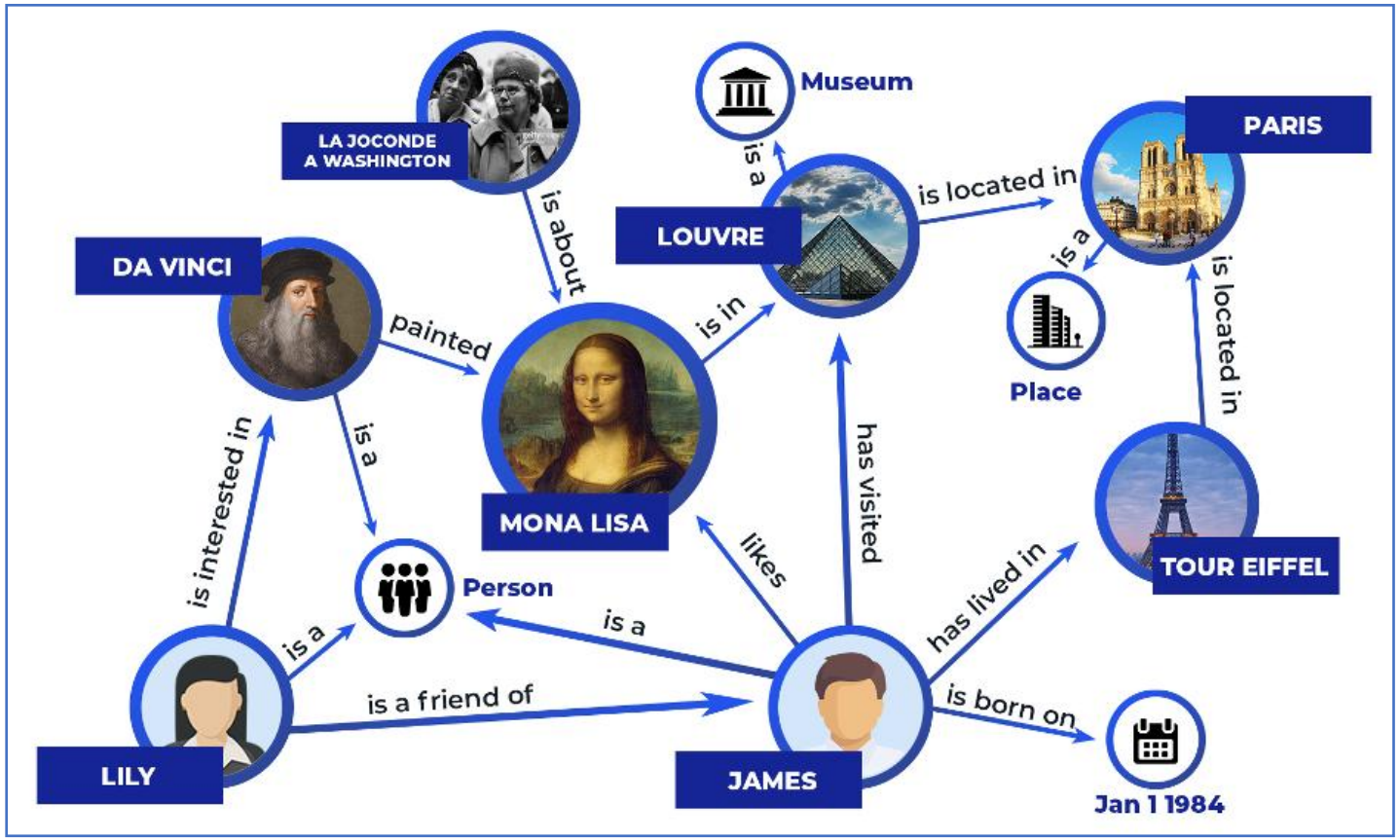


Alexa: Voice chatbot



Google Maps: Driving Directions

Knowledge Graph



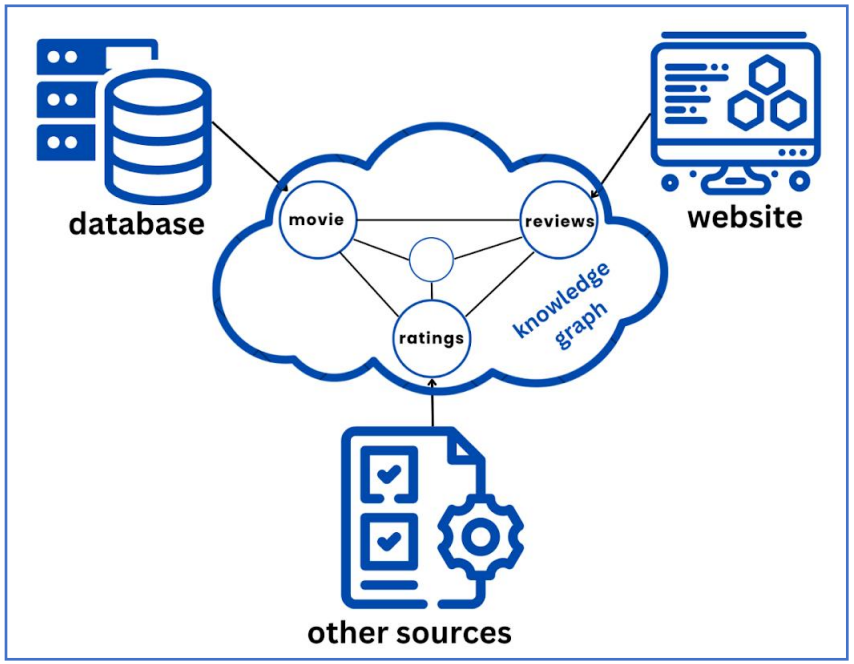
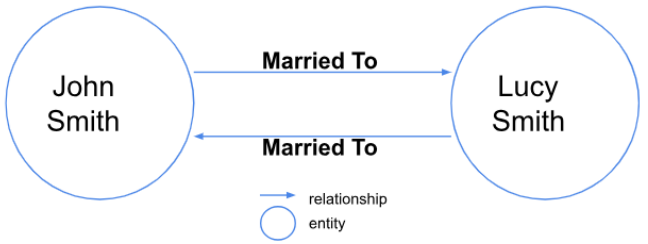


Table Representation

Name	Married to
John Smith	Lucy Smith
Lucy Smith	John Smith

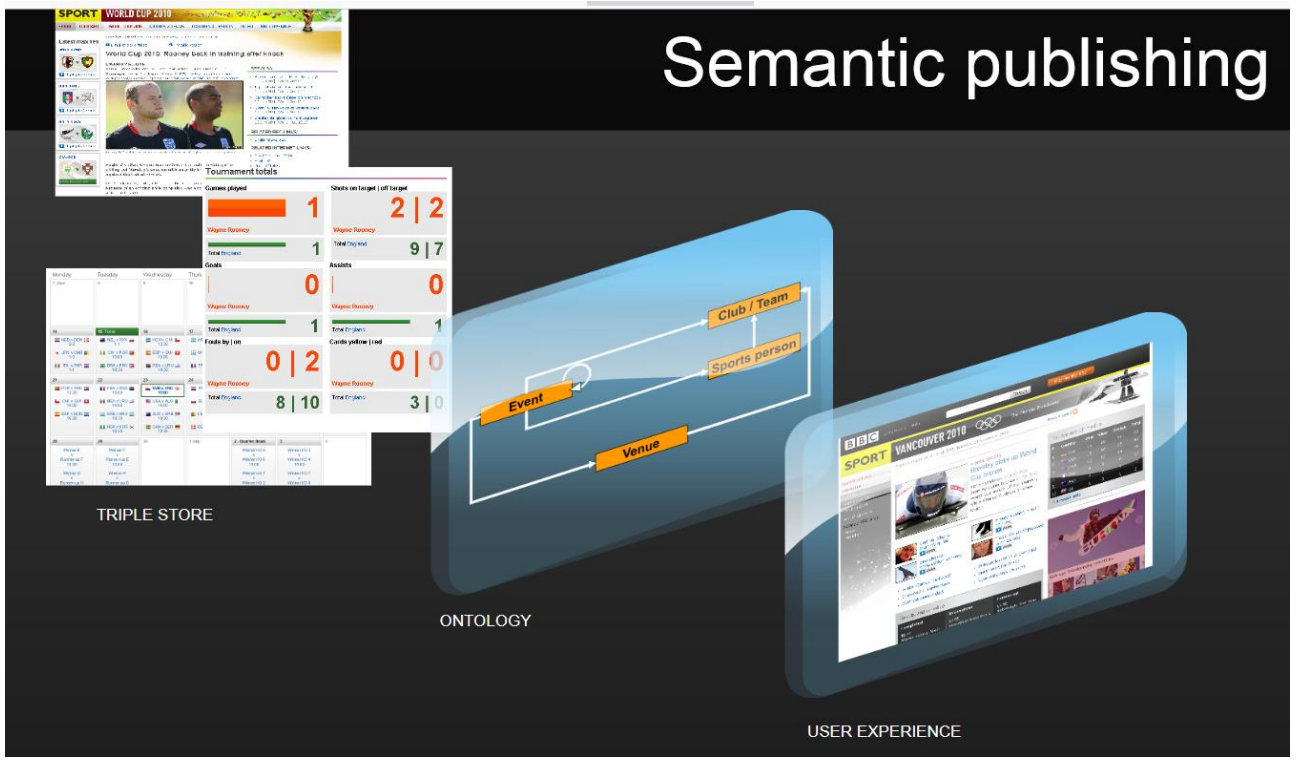
Graph Representation

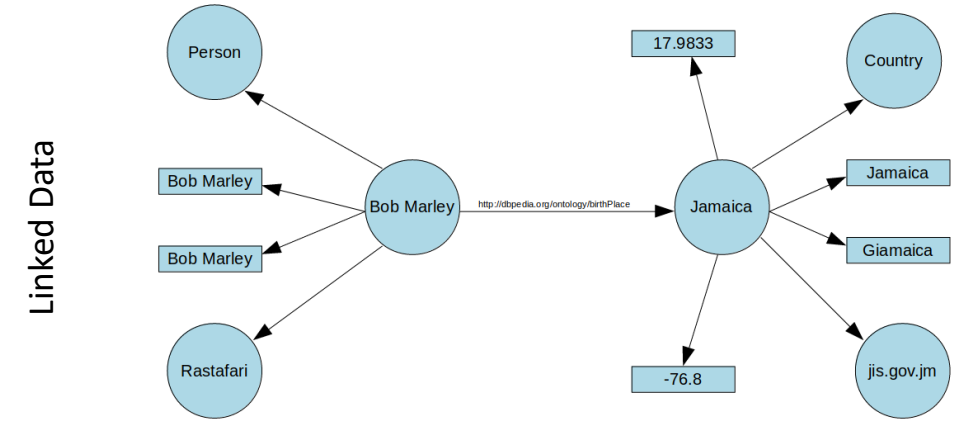


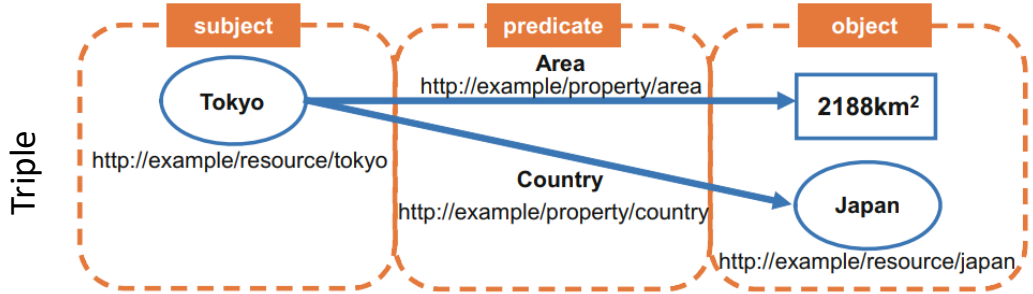
[Google: Introducing the Knowledge Graph]

Data represented in the table and graph

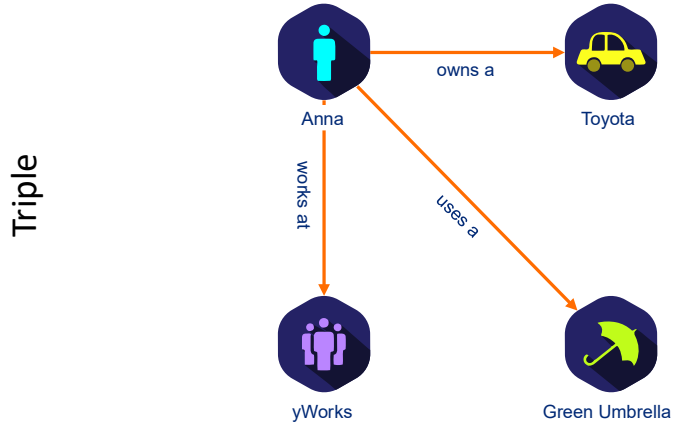
Dynamic Publishing: Scaling to Olympic Proportions





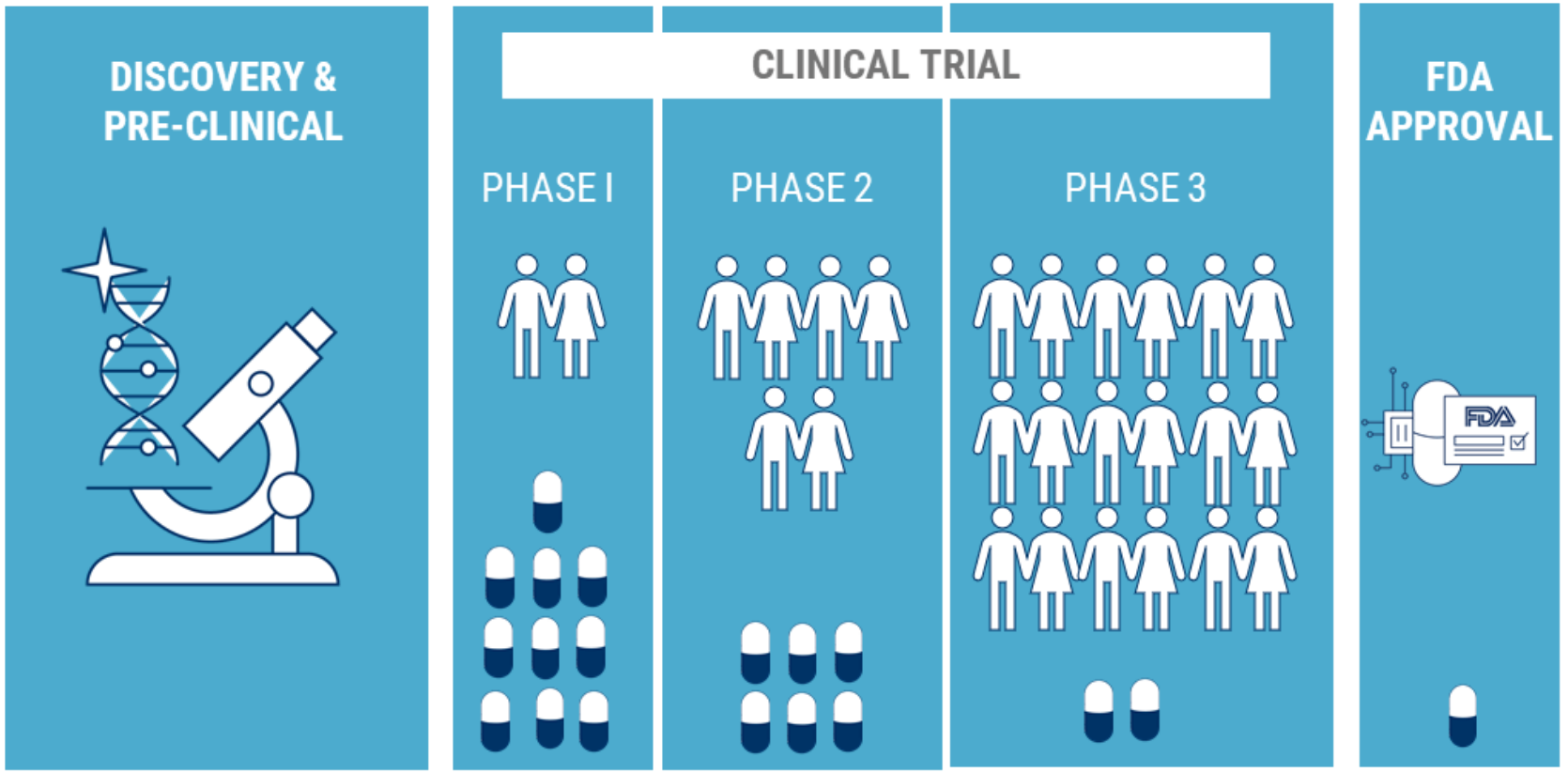


- 106 million requests for BBC Olympic video content
- 55 million global browsers across the games
- 2.8 Petabytes of Data on the busiest day
- A daily record of 7.1 million UK browsers



Clinical Trials

Clinical Trials



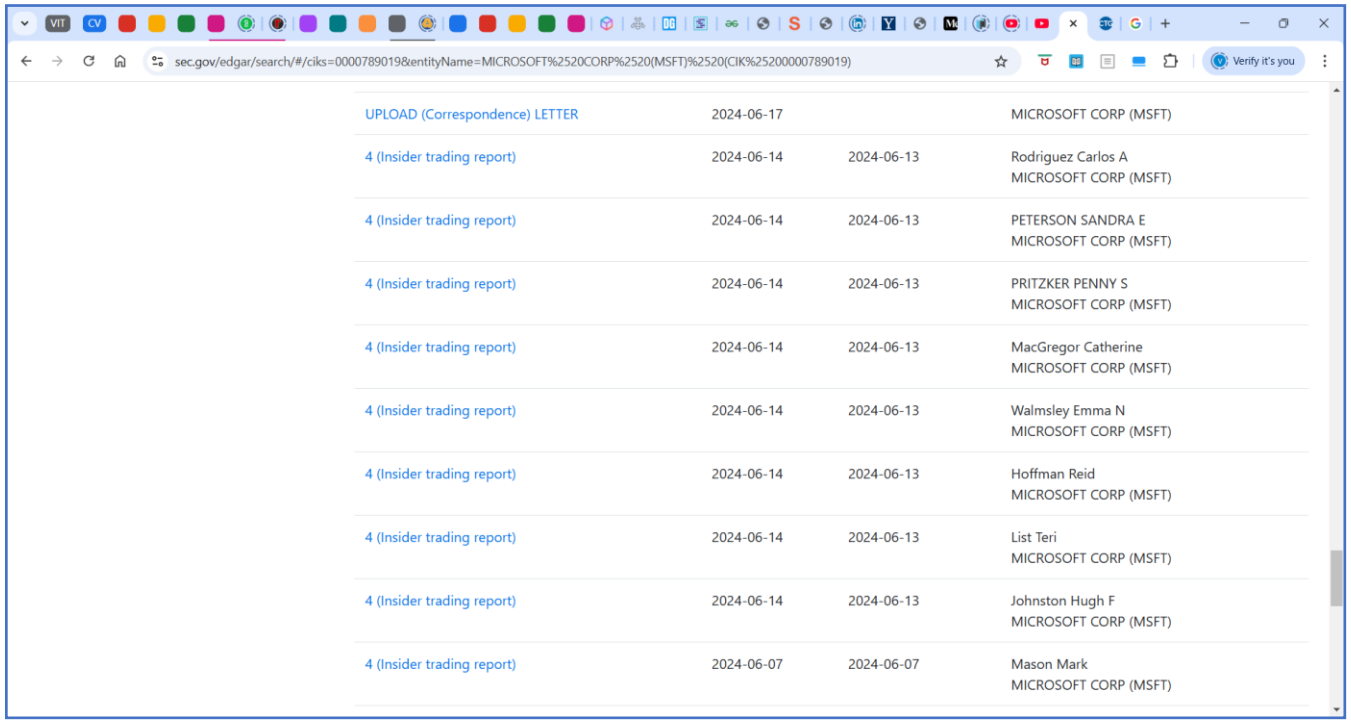
Source: cbinsights.com

https://Clinicaltrials.gov/

Financial Information Extraction: SEC

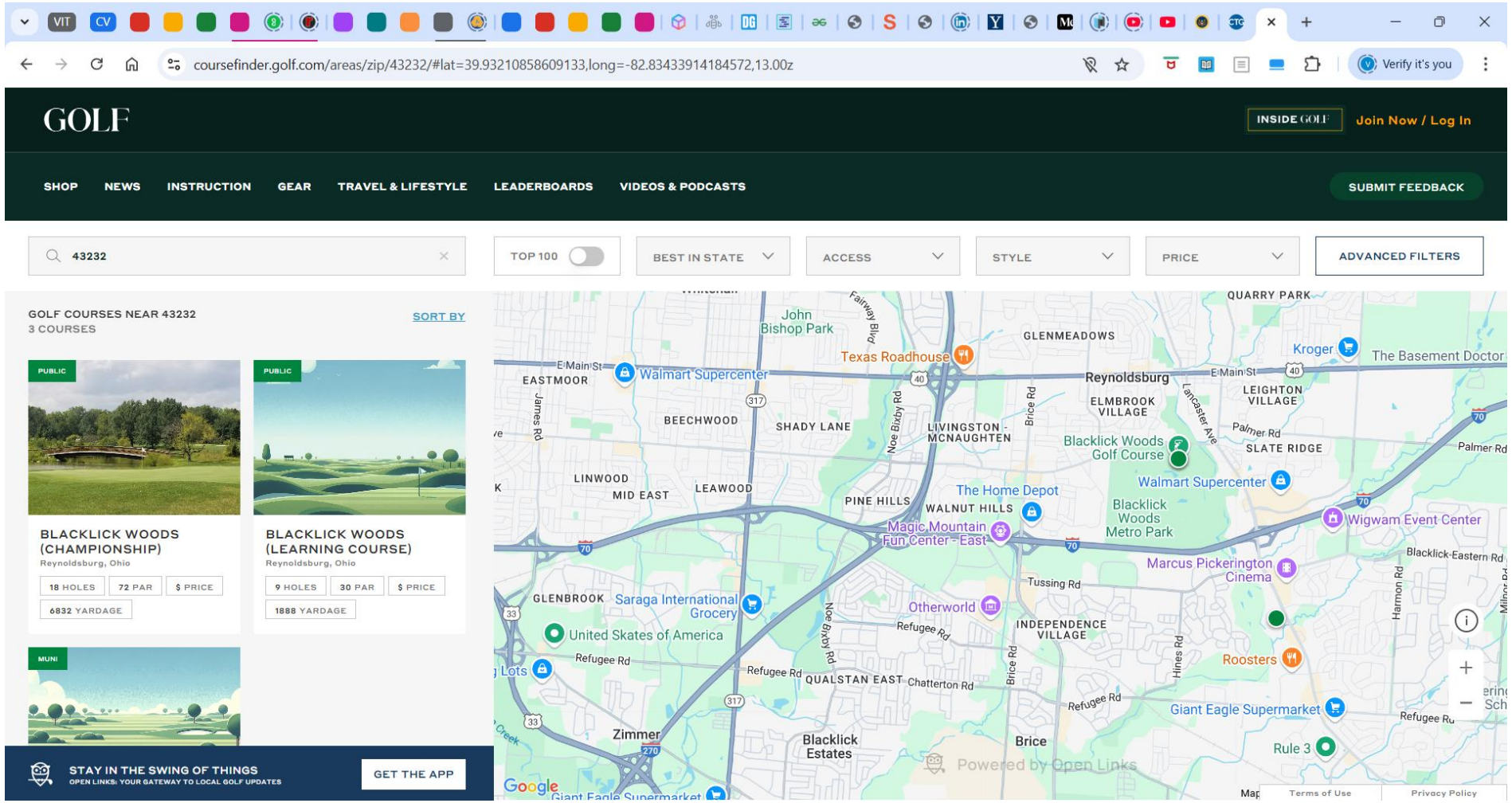
U.S. Securities and Exchange Commission

http://sec.gov

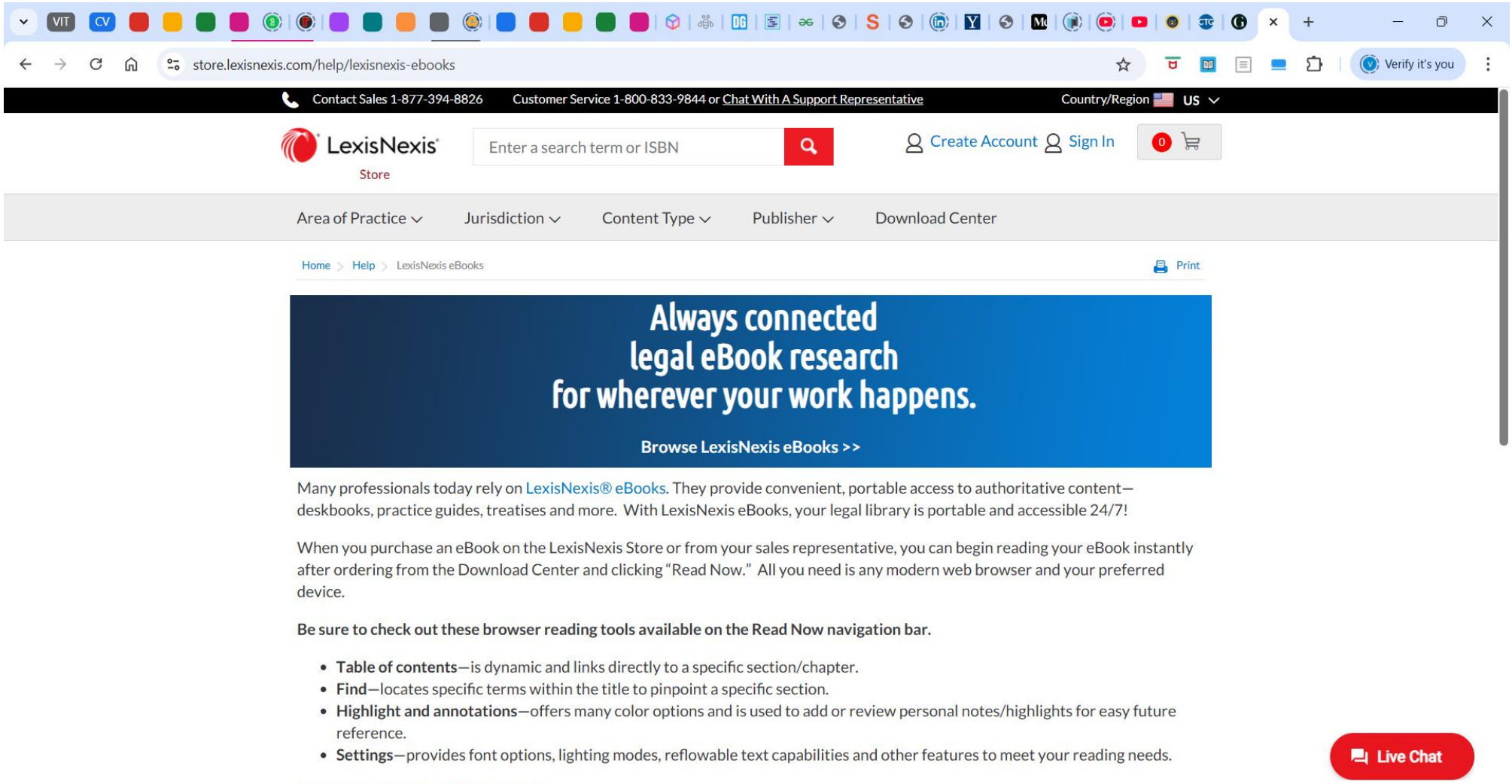


Golf.com

❖Golf.com



Lexis Nexis eBooks



Personal Assistant

- "Stop"
- "Turn it up"
- "Volume level 6"
- "Repeat that"
- "What can you do?"

- "Play some music"
- "Play music by [artist]"
- "Play dance music on YouTube"
- "Play KEXP radio on TuneIn"
- "Play the latest episode of Radiolab"
- "Pause"
- "Next song"

- "When's my first appointment tomorrow?"
- "Wake me up at 6am tomorrow"
- "Tell me about my day"
- "How long will it take to get to work?"
- "What's the weather today?"

Examples: Google Home. Amazon Alexa, Apple Siri

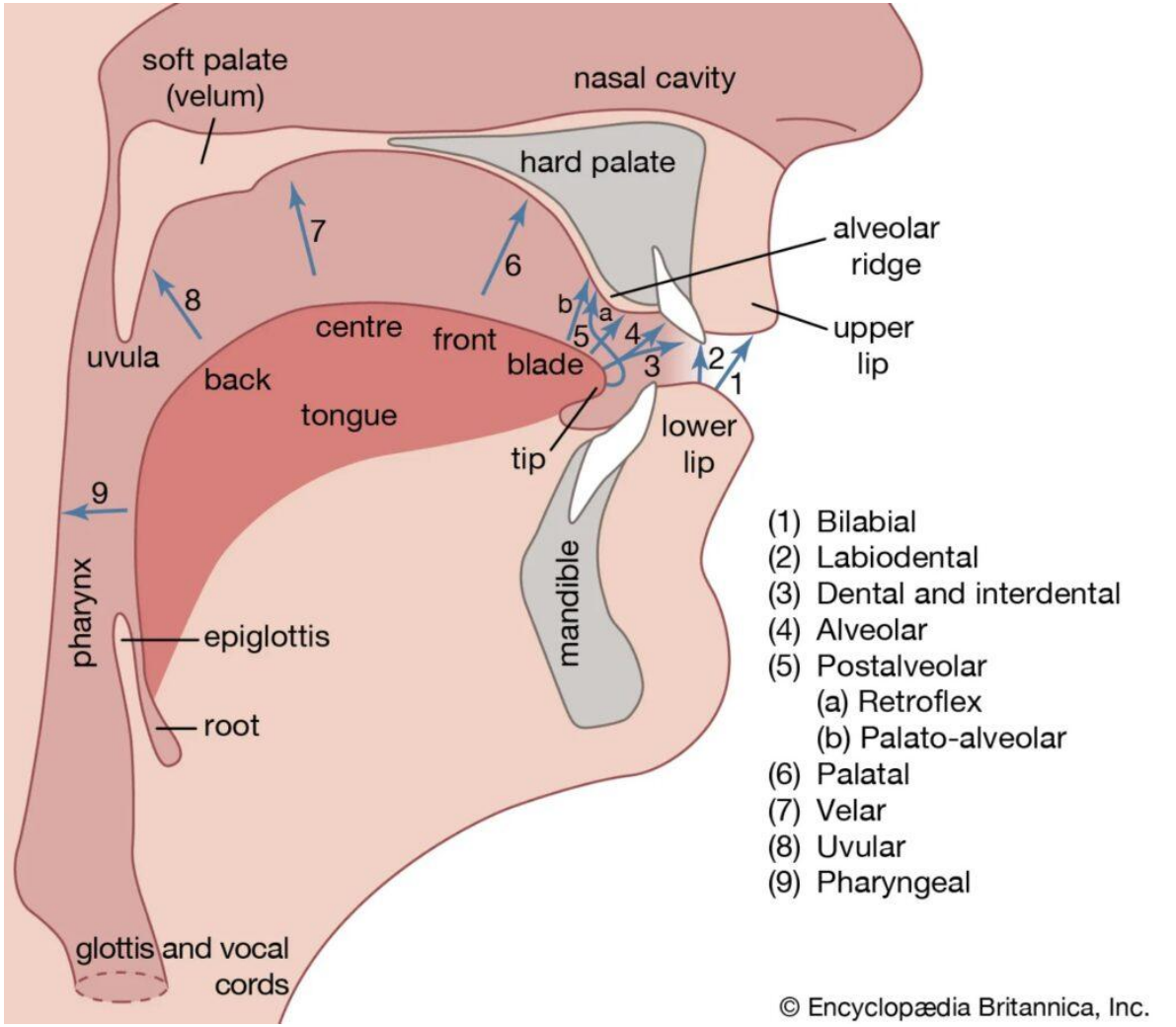
IBM's Watson: Jeopardy Challenge

- In 2011, IBM's question-answering computer Watson roundly defeated the television game show Jeopardy! champions Ken Jennings and Brad Rutter with a three-day total of $77,000, as compared to Jennings' $24,000 and Rutter's $22,000, winning the first-place prize of $1 million.

Phonetics

Phonetics

- **phonetics**, the study of speech sounds and their physiological production and acoustic qualities.
- It deals with the configurations of the vocal tract used to produce speech sounds (articulatory phonetics), the acoustic properties of speech sounds (acoustic phonetics), and the manner of combining sounds so as to make syllables, words, and sentences (linguistic phonetics).
- Knowing a language includes knowing the sounds of that language
- Phonetics is the study of speech sounds
- We are able to segment a continuous stream of speech into distinct parts and recognize the parts in other words
- Everyone who knows a language knows how to segment sentences into words and words into sounds
- Our linguistic knowledge allows us to ignore nonlinguistic differences in speech (such as individual pitch levels, rates of speed, coughs)
- We are capable of making sounds that are not speech sounds in English but are in other languages



human vocal organs and points of articulation Diagram depicting the location of human vocal organs and possible places of articulation used for speech.

https://scholar.harvard.edu/files/adam/files/phonetics.ppt.pdf

The Phonetic Alphabet

- In 1888 the International Phonetic Alphabet (IPA) was invented in order to have a system in which there was a one-to-one correspondence between each sound in language and each phonetic symbol
- Someone who knows the IPA knows how to pronounce any word in any language
- Dialectal and individual differences affect pronunciation, but the sounds of English

A Phonetic Alphabet for English Pronunciation

Consonants						Vowels			
p	**p**ill	t	**t**ill	k	**k**ill	i	b**ee**t	ɪ	b**i**t
b	**b**ill	d	**d**ill	g	**g**ill	e	b**ai**t	ɛ	b**e**t
m	**m**ill	n	**n**il	ŋ	ri**ng**	u	b**oo**t	ʊ	f**oo**t
f	**f**eel	s	**s**eal	h	**h**eal	o	b**oa**t	ɔ	b**o**re
v	**v**eal	z	**z**eal	l	**l**eaf	æ	b**a**t	a	p**o**t/b**a**r
θ	**th**igh	tʃ	**ch**ill	r	**r**eef	ʌ	b**u**tt	ə	sof**a**
ð	**th**y	dʒ	**g**in	j	**y**ou	aɪ	b**i**te	aʊ	b**ou**t
ʃ	**sh**ill	ʍ	**wh**ich	w	**w**itch	ɔɪ	b**oy**		
ʒ	mea**s**ure								

Consonants: Place of Articulation

- **Bilabials**: [p] [b] [m]
  - Produced by bringing both lips together
- **Labiodentals**: [f] [v]
  - Produced by touching the bottom lip to the upper teeth
- **Interdentals** [θ] [ð]
  - Produced by putting the tip of the tongue between the teeth

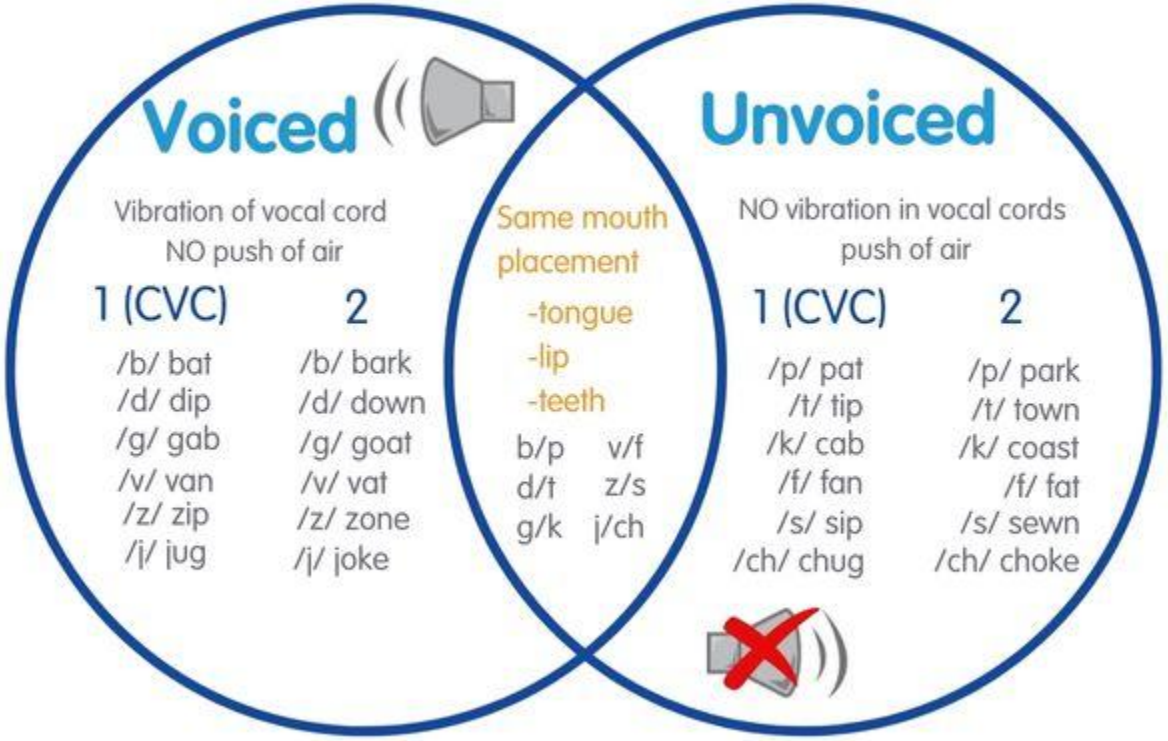
- **Palatals**: [ʃ] [ʒ] [ʧ] [ʤ][j̠]
  - Produced by raising the front part of the tongue to the palate
- **Velars**: [k] [g] [ŋ]
  - Produced by raising the back of the tongue to the soft palate or velum
- **Uvulars**: [ʀ] [q] [ɢ]
  - Produced by raising the back of the tongue to the uvula
- **Glottals**: [h] [ʔ]
  - Produced by restricting the airflow through the open glottis ([h]) or by stopping the air completely at the glottis (a **glottal stop**: [ʔ])

Consonants: Manner of Articulation

- The manner of articulation is the way the airstream is affected as it flows from the lungs and out of the mouth and nose
- Voiceless sounds are those produced with the vocal cords apart so the air flows freely through the glottis
- Voiced sounds are those produced when the vocal cords are together and vibrate as air passes through

- **Alveolars**: [t] [d] [n] [s] [z] [l] [r]
  - All of these are produced by raising the tongue to the **alveolar ridge** in some way
    - [t, d, n]: produced by the tip of the tongue touching the alveolar ridge (or just in front of it)
    - [s, z]: produced with the sides of the front of the tongue raised but the tip lowered to allow air to escape
    - [l]: the tongue tip is raised while the rest of the tongue remains down so air can escape over the sides of the tongue (thus [l] is a **lateral** sound)
    - [r]: air escapes through the **central** part of the mouth; either the tip of the tongue is curled back behind the alveolar ridge or the top of the tongue is bunched up behind the alveolar ridge

Voiced and Unvoiced Sounds



Vowels:

All vowels are voiced sounds in the English language. So is the semi-vowel ‘y’ when pronounced like a vowel rather than a consonant. The production of vowel sounds requires little to no restriction of airflow.

Diphthongs:

All diphthongs (where the sound begins as one vowel sound and glides to another vowel sound) are voiced e.g. ‘ow’, ‘ou’, ‘aw’, ‘au’, ‘oi’ and ‘oy’.

Consonants:

Consonants can be either voiced or unvoiced.

•The voiced consonants are: b, d, g, j, l, m, n, r, v, w, y and z. The digraph ‘ng’ and the /sz/ heard in ‘treasure’ are also voiced.

•The unvoiced consonants are f, h, c/k/q (all of which share the same unvoiced sound, /k/), p, s, t, x (pronounced as a combination of two unvoiced consonants).

•The digraphs /ch/ as in ‘church’, /sh/ and /wh/ are also unvoiced.

•The digraph ‘th’ can be voiced, as in “that” or unvoiced as in “thing.”

Vowel

- Vowels are classified by how high or low the tongue is, if the tongue is in the front or back of the mouth, and whether or not the lips are rounded

- **High** vowels: [i] [ɪ] [u] [ʊ]
- **Mid** vowels: [e] [ɛ] [o] [ə] [ʌ] [ɔ]
- **Low** vowels: [æ] [a]

- **Front** vowels: [i] [ɪ] [e] [ɛ] [æ]
- **Central** vowels: [ə] [ʌ]
- **Back** vowels: [u] [ɔ] [o] [æ] [a]

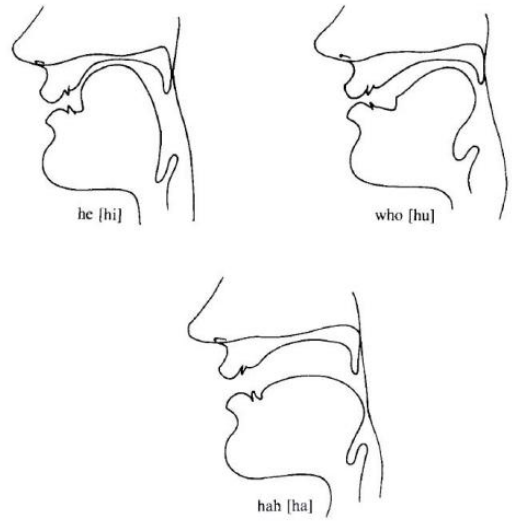
Part of the Tongue Involved

Tongue Height	FRONT	CENTRAL	BACK	ROUNDED
HIGH	i beet		boot	u
↑	ɪ bit		put	ʊ
MID	e bait		boat	o
	ɛ bet	ə Rosa		
↓		ʌ butt	bore	ɔ
LOW	æ bat		bomb	a

- **Round** vowels: [u] [ʊ] [o] [ɔ]
  - Produced by rounding the lips
  - English has only back round vowels, but other languages such as French and Swedish have front round vowels

- **Diphthongs**: [aɪ] [aʊ] [ɔɪ]
  - A sequence of two vowel sounds (as opposed to the **monophthongs** we have looked at so far)

- **Nasalization**:
  - Vowels can also be pronounced with a lowered velum, allowing air to pass through the nose
  - In English, speakers nasalize vowels before a nasal sound, such as in the words *beam*, *bean*, and *bingo*
  - The nasalization is represented by a diacritic, an extra mark placed with the symbol: *bean* [bĩn]



Language Nuances

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 refers to the presence of multiple possible meanings or interpretations within a statement, word, or any other form of communication. 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.

Morphology

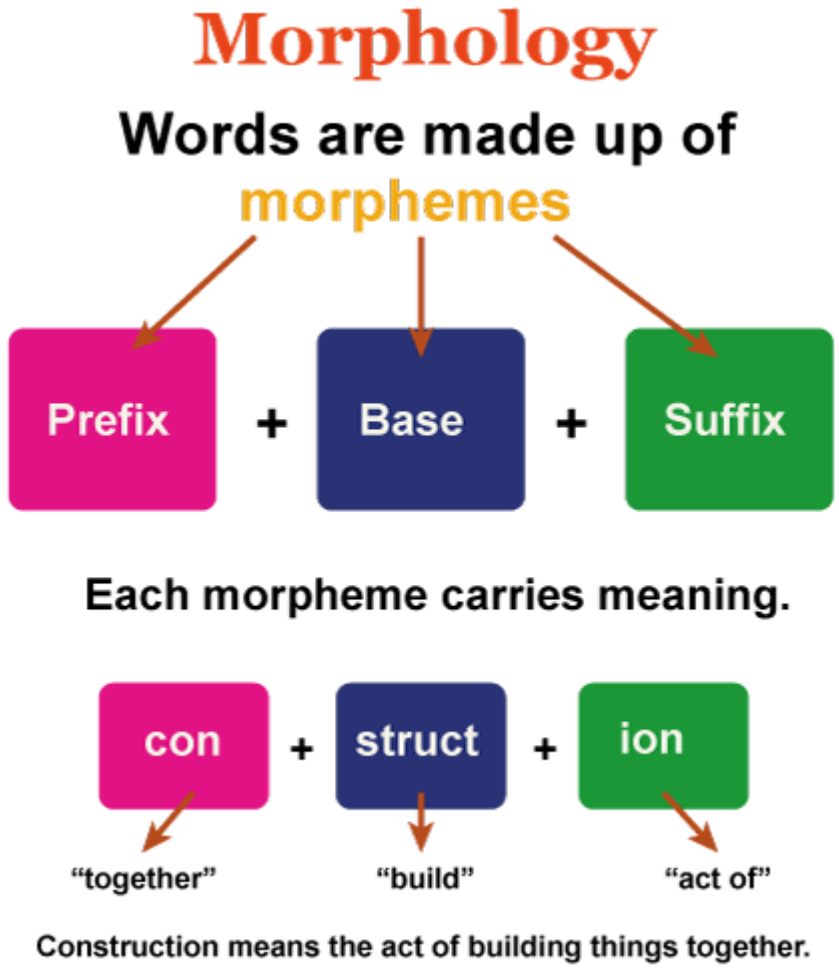
The study of how morphemes are combined to form words.

Words are made up of morphemes:

Prefixes | Roots/Bases | Suffixes

Morphemes are the smallest unit of *meaning*.

un	reach	able	dis	tract	ion	s
"not"		"able to"	"away"	"pull, drag"	"state of"	plural



Examples of Morphemes

Word	Morphemes	# of Morphemes	Type of Morpheme
house	house	1	base
jumped	jump + ed	2	base + suffix
football	foot + ball	2	base + base
prediction	pre + dict +ion	3	prefix + root + suffix
deconstructed	de + con + struct + ed	4	prefix + prefix + root + suffix
informational	in + form + ate + ion+ al	5	prefix + root + suffix + suffix + suffix

Idioms and Phrases in Languages

❖सोने पे सुहागा

❖Slam dunk

తెలుగు: నక్కకి నాగలోకానికి ఉన్న తేడా

English: Like night and day" or "Worlds apart

తెలుగు: అందితే జుట్టు లేకపోతే కాళ్ళు

English: Give them an inch, and they'll take a mile

తెలుగు: **ఆలస్యం అమృతం విషం**

English: Be/Do on time (Else bad thing can happen)

తెలుగు: దూరపు కొండలు నునుపు

Hindi: दूर के ढ़ोल सुहावने होते है।

English: Distant drums sound well.

Hindi: करनी कथनी से ताकतवर होती है।

English:Actions speak Louder than words.

Hindi: बूँद बूँद से सागर भरता है।

English: many a pickle makes mickle.

Hindi: बहती गंगा में हाथ धोना।

English: Make hay while the Sun shines.

Why is NLP difficult?

Ambiguity

"He's at the bank."

vs.

lexical ambiguity

"She flew in last night."

Ambiguity

"Stolen painting found by tree"

Prepositional Phrase

By a tree, a stolen painting was found.

vs.

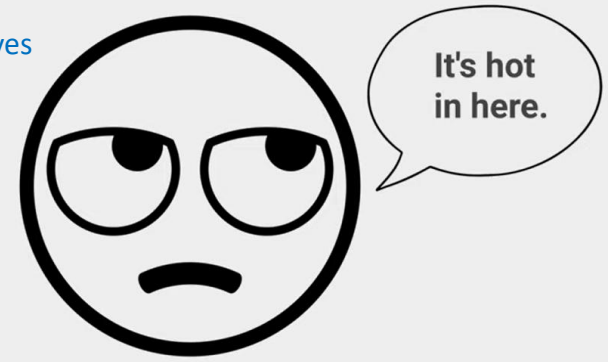
A tree found a stolen painting.

PP-attachment problem

syntactic ambiguity

Beyond text (tones and gestures)

While Rolling your Eyes
+ Fanning your face



Really: "I'm bored and tired and I hate you for bringing me here."

Pragmatics (how people use language)

How you say it is as important as what you say it.

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

Ambiguity in a Telugu and Hindi Sentences

A word, phrase, or sentence is ambiguous if it has more than one meaning. Ambiguity resolution has always been the most important testing ground in linguistics for parsing models.

We can group ambiguity into two, lexical and structural ambiguity.

Lexical Level Ambiguity

- Definition: a word belongs to two or more word ("part of speech") classes
- Example: the *round* table (adjective), to *round* the corner (verb), dance in a *round* (noun), come *round* and see us (adverb), he walked *round* the room (preposition)
- Finite state grammars can be used for resolving lexical ambiguity

పాత్ర / vessel/ role	Role	Item
కొట్టు / *beat*	Shop	Beat
రసం / *Juice*	Fluid	Expression
గాలి / air	Air	Name

Table.2 Example words in different senses

Word	$Sense_1$	$Sense_2$
శూన్యం / Emptiness	Atmosphere	Value
ప్రగతి / Progress	Improvement	Name
రాజు / king	King	Name

Sense	Examples (keyword in context)	Tag
1	కష్టపడి పని చేస్తే ప్రగతి సాధించవచ్చు / *Progress* can be achieved with hard work	ADJECTIVE
1	యువత పైనే దేశ ప్రగతి ఆధారపడినది. *Progress* is based on the youth of the country.	ADJECTIVE
2	ప్రగతి బాగా పాడుతుంది. / PRAGATI [progress] sings well	NOUN
2	ఈ పుస్తకం ప్రగతి సంస్థ వారిచే ముద్రించబడినది. / The book, published by the PRAGATI [Progress] company.	NOUN

Structural Level Ambiguity

Structural ambiguity occurs when the meaning of the component words can be combined in more than one way

Maine ne gali se aate hue ladke ko dekhaa.

I saw a boy coming from the street.

This is an ambiguous sentence because it has more than one meaning.

a) I am coming from the street.
b) Boy is coming from the street

Understanding Ambiguities in Natural Language Processing

Lexical Ambiguity

- Lexical means relating to words of a language. During Lexical analysis given paragraphs are broken down into words or tokens. Each token has got specific meaning. There can be instances where a single word can be interpreted in multiple ways. The ambiguity that is caused by the word alone rather than the context is known as Lexical Ambiguity.
- Example: “Give me the bat”, “The fisherman wen to the bank”, “The pilot was banking on a safe landing.”, “We fly over a million people” (above/ more than) .

Syntactic Ambiguity/ Structural ambiguity

- Syntactic meaning refers to the grammatical structure and rules that define how words should be combined to form sentences and phrases. A sentence can be interpreted in more than one way due to its structure or syntax such ambiguity is referred to as Syntactic Ambiguity.
- Example: “John saw the boy with telescope. “, “Visiting Relatives can be uplifting”, “I saw a man with binoculars”

Semantic Ambiguity

- Semantics is nothing but “Meaning”. The semantics of a word or phrase refers to the way it is typically understood or interpreted by people. Syntax describes the rules by which words can be combined into sentences, while semantics describes what they mean.
- Semantic Ambiguity occurs when a sentence has more than one interpretation or meaning.
- Example: “He ate the burnt lasagna and pie.”, “Seema loves her mother and Sriya does too.”

Anaphoric Ambiguity

- A word that gets its meaning from a preceding word or phrase is called an anaphor.
- Example: “Susan plays the piano. She likes music.”, “The horse ran up the hill. It was very steep. It soon got tired.”, “I went to the hospital, and they told me to go home and rest.”

Pragmatic ambiguity

- Pragmatics focuses on the real-time usage of language like what the speaker wants to convey and how the listener infers it. Situational context, the individuals’ mental states, the preceding dialogue, and other elements play a major role in understanding what the speaker is trying to say and how the listeners perceive it.

Ambiguities in NLP: Anaphoric ambiguity

- A word that gets its meaning from a preceding word or phrase is called an anaphor.
- Example: "**Susan** plays the piano. *She* likes music."

In this example, the word *she* is an anaphor and refers back to a preceding expression i.e., *Susan*. The linguistic element or elements to which an anaphor refers is called an antecedent. The relationship between anaphor and antecedent is termed '**anaphora**'. 'Anaphora resolution' or 'anaphor resolution' is the process of finding the correct antecedent of an anaphor.

Ambiguity that arises when there is more than one reference to the antecedent is known as Anaphoric Ambiguity.

- Example 1: "The horse ran up the hill. It was very steep. It soon got tired."

In this example, there are two 'it', and it is unclear to which each 'it' refers, this leads to Anaphoric Ambiguity. The sentence will be meaningful if first 'it' refers to the hill and 2nd 'it' refers to the horse. Anaphors may not be in the immediately previous sentence. They may present in the sentences before the previous one or may present in the same sentence.

Anaphoric references may not be explicitly present in the previous sentence rather they might refer to the part of the antecedent.

- Example 2: "I went to the hospital, and they told me to go home and rest."

In this sentence, 'they' does not explicitly refer to the hospital instead it refers to the Dr or staff who attended the patient in the hospital. Anaphors are mostly pronouns, or they can even be noun phrases in some instances.

- Example 3: "Darshan plays keyboard. He loves music. "

In this case 'He' is a pronoun.

- Example 4: "A puppy drank the milk. The cute little dog was satisfied."

Here Anaphor is 'cute little dog' which is a noun phrase.

Ambiguities in NLP: Pragmatic ambiguity

Sentence	*Direct meaning (semantic meaning)*	*Other meanings (pragmatic meanings)*
Do you know what time is it?	Asking for the current time	Expressing anger to someone who missed the due time or something `
Will you crack open the door? I am getting hot	To break	Open the door just a little
The chicken is ready to eat	The chicken is ready to eat its breakfast, for example.	The cooked chicken is ready to be served

Language nuances

A finite state language is a finite or infinite set of strings (sentences) of symbols (words) generated by a finite set of rules (the grammar), where each rule specifies the state of the system in which it can be applied, the symbol which is generated, and the state of the system after the rule is applied.

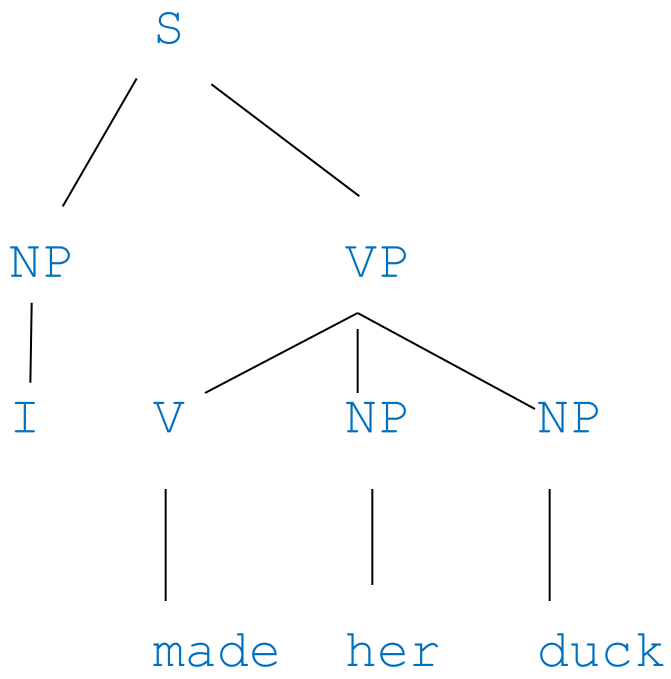
Telugu Conversation Sentences	English Meaning	Annotated Pattern
నీతో మాట్లాడాలి, కొంచెం టైమ్ ఇస్తావా? సరే చెప్పు శేఖర్, ఏం మాట్లాడాలి?	I want to talk to you, can you spare few minutes for me? Ok, tell me sekhar, what do you want to talk?	Normal Question and Normal Reply
నీతో మాట్లాడాలి, కొంచెం టైమ్ ఇస్తావా? అయ్యో నా వాచ్లో బ్యాటరీ అయిపోయిందే!	I want to talk to you, can you spare few minutes for me? Oh, the battery is dead on my watch!	Normal Question and Sarcastic Reply
ఏంటి బంగారం ఈరోజు ఆఫీస్ నుండి ఇంత తొందరగా వచ్చేశావు? అదా, ఈరోజు కొంచెం పని ఎక్కువగా ఉంది,అందుకే లేట్ అయ్యింది.	Hai dear, it seems today you came early from the office? Actually, today I had bit more work than usual, so I was late.	Sarcastic Question and Normal Reply
ఏంటి బంగారం ఈరోజు ఆఫీస్ నుండి ఇంత తొందరగా వచ్చేశావు? అదా, నీ మీద ప్రేమ ఎక్కువైపోయి మా బాస్ ని పర్మిషన్ అడిగి వచ్చేశాను.	Hai dear, it seems today you came early from the office? Since, I had more love on you, I have requested my boss for a permission to leave a bit early from the office.	Sarcastic Question and Sarcastic Reply

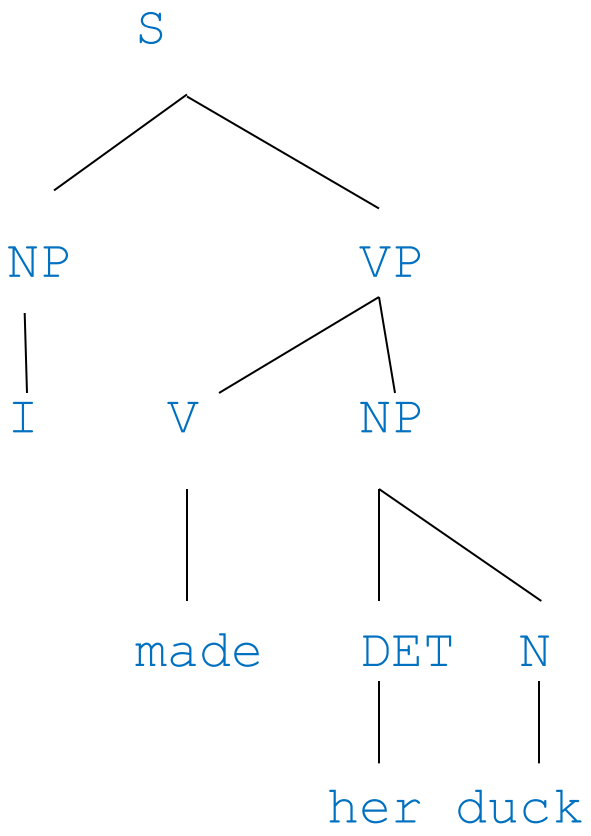
Resolve Ambiguities

- 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.

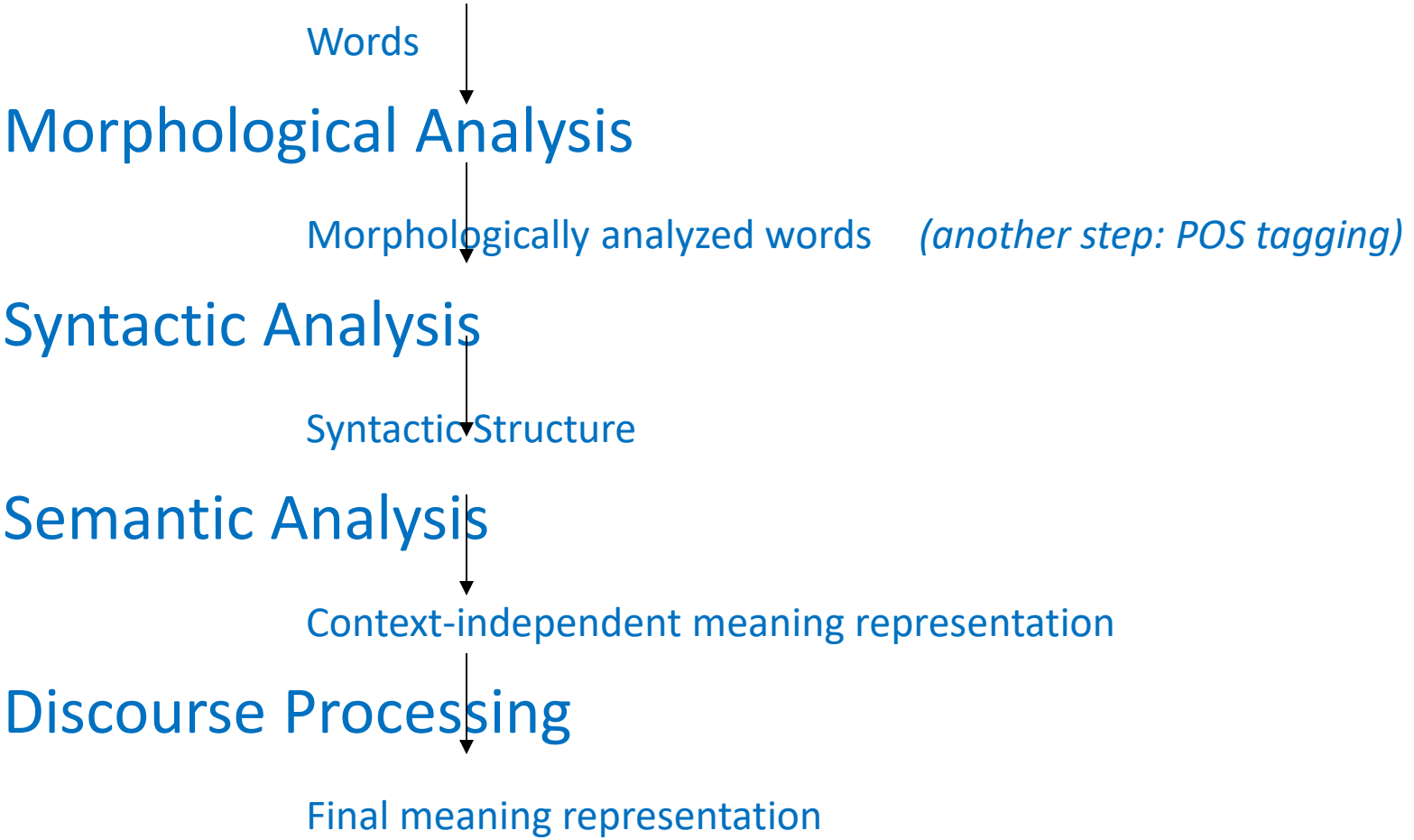
Resolve Ambiguities (cont.)

I made her duck

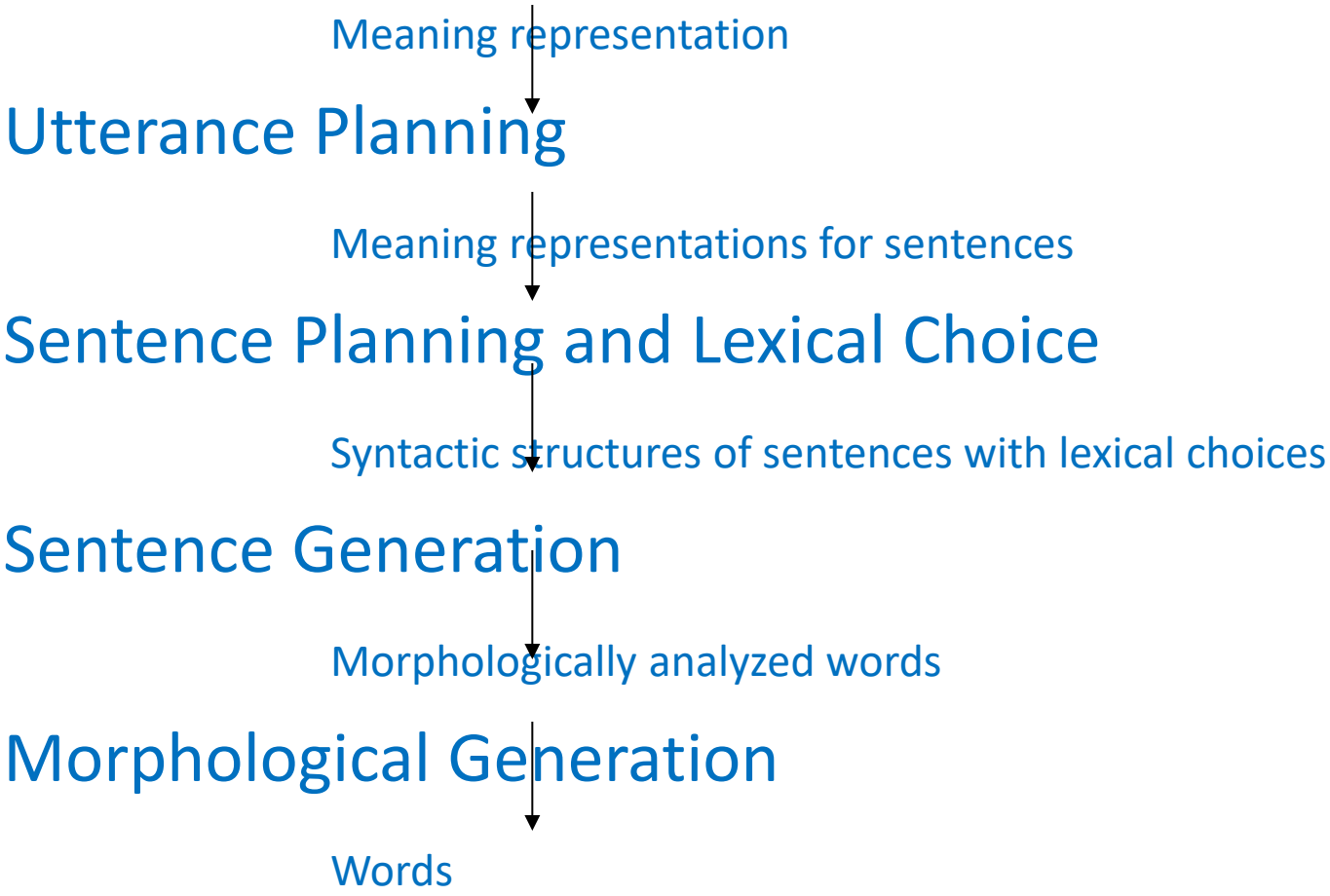




Natural Language Understanding



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

❖Ambiguity: More than one alternatives

flies	fly_{VERB}+PROG
	fly_{NOUN}+PLU

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
- 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

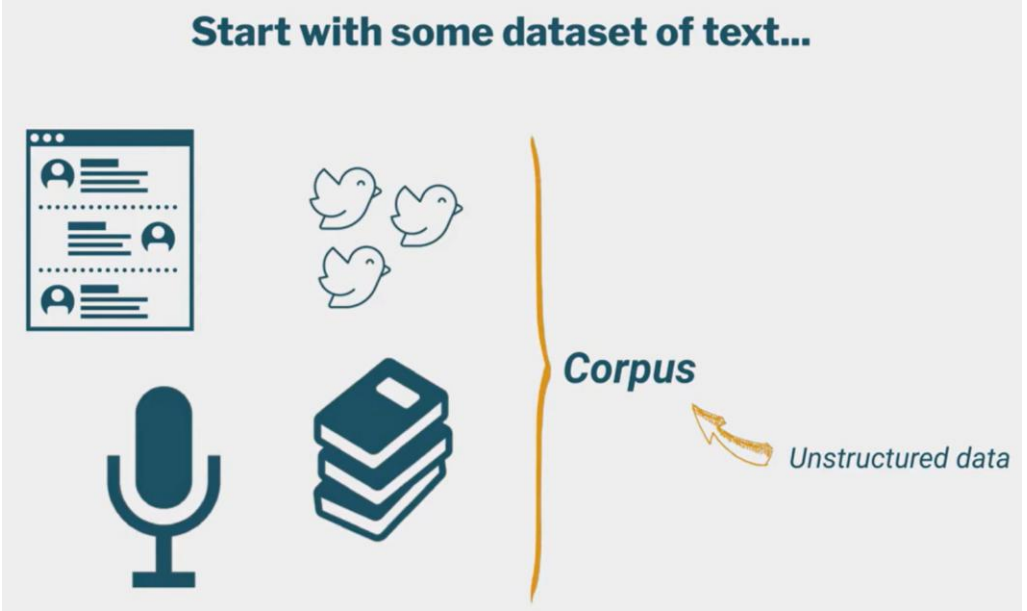
- Discourses are collection of coherent sentences (not arbitrary set of sentences)
- Discourses have also hierarchical structures (similar to sentences)
- **anaphora resolution** -- to resolve referring expression
  - Mary bought a book for Kelly. **She** didn't like **it**.
    - **She** refers to Mary 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

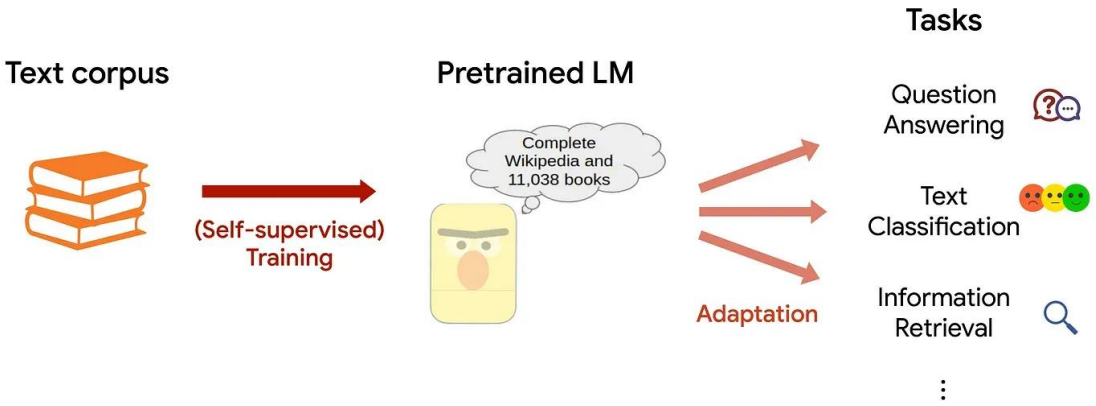
Natural Language Generation

- 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.

Corpus

- A corpus (plural corpora), also known as a text corpus in linguistics, is usually a large collection of texts, and it could be compared to a database in which each text is a record.
- A corpus can be assembled from a variety of sources and genres. Such a corpus can be used for general NLP tasks. On the other hand, a corpus might be from a single source, domain or genre. Such a corpus can be used only for a specific purpose.
- In NLP, corpora are used to train AI algorithms and develop statistical models. These corpora can be used by linguists, lexicographers, data scientists, and experts in NLP for various tasks, including word frequency analysis, part-of-speech tagging, and text classification.
- Ex1: Brown Corpus: Compiled at Brown University in the 1960s with 500 texts, each containing a little over 2,000 words for a total of 1 million words of American English sampled from 15 different categories
- Example 2: British National Corpus (BNC), created by Oxford University Press, which contains 100 million words
- Example 3: Corpus of Contemporary American English (COCA), 1 billion words
- Example 4: POS Tagging: Penn Treebank's WSJ section is tagged with a 45-tag tagset. Use Ritter dataset for social media content
- Example 5: Named Entity Recognition: CoNLL 2003 NER task is newswire content from Reuters RCV1 corpus. It considers four entity types. WNUT 2017 Emerging Entities task and OntoNotes 5.0 are other datasets.

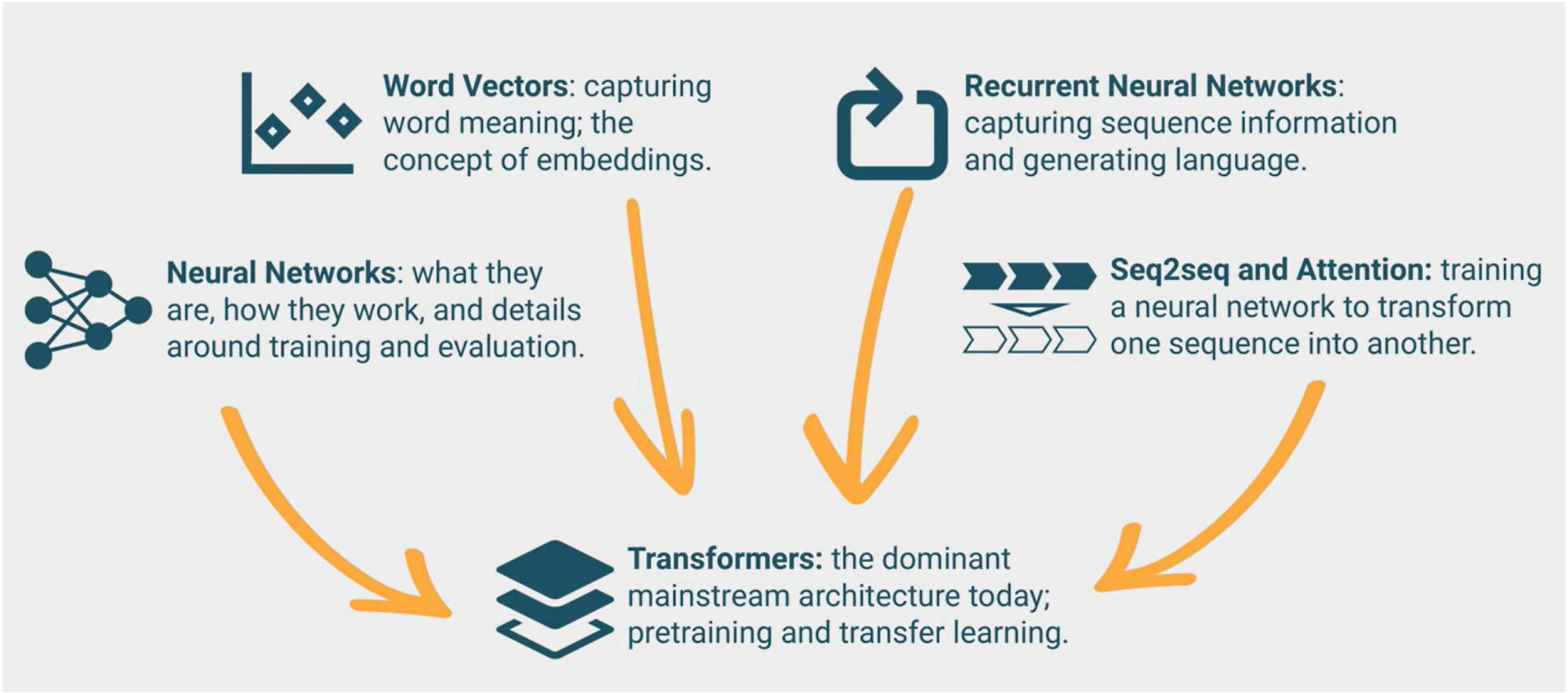




Characteristics of a Good Corpus

- large
- systematically assembled
- natural texts
- often available to other researchers
- spoken and/or written language
- usually in electronic form
- can be tagged for use with text manipulation programs

Deep Learning for NLP



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,

What is Covered in the Course

Module No. 1	Introduction to NLP	8 Hours
Overview: Origins and challenges of NLP-Need of NLP, Preprocessing techniques- Text Wrangling, Text cleansing, sentence splitter, tokenization, stemming, lemmatization, stop word removal, rare word removal, spell correction. Word Embeddings, Types : One Hot Encoding, Bag of Words (BoW), TF-IDF Static word embeddings: Word2vec, GloVe, FastText		
Module No. 2	Parts of Speech Tagging	6 Hours
Parts of Speech Tagging and Named Entities –Tagging in NLP, Sequential tagger, N-gram tagger, Regex tagger,Brill tagger,NER tagger; Machine learning taggers-MEC,HMM,CRF		
Module No. 3	Parsing Structure in Text	10 Hours
Shallow vs Deep parsing, Approaches in parsing, Types of parsing- Regex parser, Dependency parser, Constituency Parsing Meaning Representation: Logical Semantics, Semantic Role Labelling, Distributional Semantics Discourse Processing: Anaphora and Coreference Resolution		
Module No. 4	NLP Using Deep Learning	6 Hours
Types of learning techniques,–Chunking, Information extraction & Relation Extraction, Recurrent neural networks, LSTMs/GRUs, Transformers, Self-attention Mechanism, Sub-word tokenization, Positional encoding,		
Module No. 5	Web Crawling and Social Media Mining	6 Hours
Web crawler – Writing first crawler–Data flow in Scrapy–Scrapy shell. Social Media Mining-Data Collections, Data Extraction, Geo visualization.		
Module No. 6	NLP latest Techniques and applications	9 Hours
Contextualized word embeddings: ELMo, BERT, GPT Pre-trained Language Models (PLMs): BERT, GPT, ELMo, Large Language Models (LLMs). Applications of NLP: Transforming text, Sentiment Analysis, Information retrieval, text summarization, Question and Answering, Automatic Summarization		

Text Books and References

❖ Text Books

- Daniel Jurafsky, James H.Martin. Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Pearson, 2nd Edition, January 2013.
- Daniel Jurafsky and James H. Martin. 2024. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition.

  https://web.stanford.edu/~jurafsky/slp3/
- Nitin Hardeniya, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, Iti Mathur, "Natural Language Processing: Python and NLTK", Packt publisher, 2016.
- Steven Bird, Ewan Klein, Edward Loper, "Natural Language Processing with Python", O'Reilly, 1stEdition 2009.
- Jacob Perkins, "Python 3 Text Processing with NLTK 3 Cookbook", Pearson Education, Second Edition, 2014.
- Deepti Chopra, Nisheeth Joshi, Iti Mathur, " Mastering Natural Language Processing with Python", Packt ,2016.

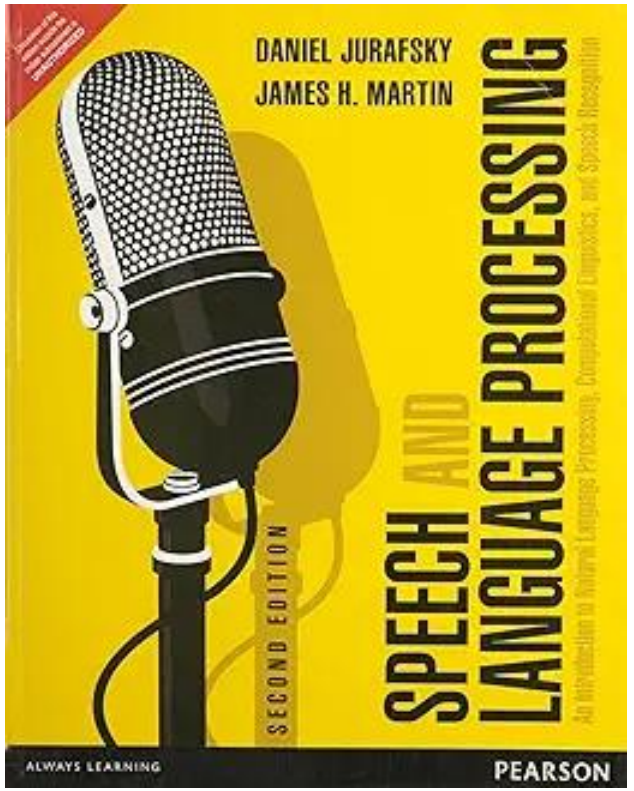
❖ References

- Natural Language Processing by Jacob Eisenstein, 2018. Available online: https://cseweb.ucsd.edu/~nnakashole/teaching/eisenstein-nov18.pdf
- Natural Language Processing with Python by Steven Bird, Ewan Klein, and Edward Loper

  https://www.nltk.org/book/

➢ Credits: Slides and Video content borrowed from:

- Stanford CS224N: Natural Language Processing with Deep Learning Course | Winter 2019
- Lecture Collection | Natural Language Processing with Deep Learning (Winter 2017)
- Stanford CS224N: Natural Language Processing with Deep Learning | 2023
- Stanford CS224N: Natural Language Processing with Deep Learning | Winter 2021
- **GEN102 - What is Linguistics (not)?**



3rd Edition online

NLP Progress and Tools

1950s: Rules-based systems.

Late 1980s: "Statistical revolution"; incorporating Machine Learning.

Early 2010s: Neural Networks; Deep Learning.

Lots of tools available today

PyTorch

NLTK

spaCy

TensorFlow

AllenNLP

Mode of Evaluation + Assignments

- Theory (75%)
  - Continuous Assessment Test-1 15
  - Continuous Assessment Test-2 15
  - Digital Assignments/Quizzes (Min) 30
  - Final Assessment Test 40
- Laboratory (25%)

Assignments:

- a lot of programming in NLP.

Participation:

- Be around for lectures.
- Ask and answer questions.

Contact information

- Feel free to email me on venugopal.g@vitap.ac.in

  or reach me at AB-1, Room#: 161, Cabin#: 165

- Please use [NLP] in email title!

- LinkedIn: https://in.linkedin.com/in/gundimedavenugopal

Models to Represent Linguistic Knowledge

- We will use certain formalisms (*models*) to represent the required linguistic knowledge.
- **State Machines** -- FSA, Hidden Markov Models (HMMs)
- **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.

CONTEXT FREE GRAMMARS

- The grammar is called **Context Free** because,
  - using these rules, the decision to **replace a non-terminal** by some other sequence is made **without looking at the context** in which the non-terminal occurs.

$S \rightarrow aSb$
$S \rightarrow aS$
$S \rightarrow bS$
$S \rightarrow \varepsilon$

"aaabbb"

① $S \Rightarrow aSb$
$\Rightarrow aaSbb$
$\Rightarrow aaaSbbb$
$\Rightarrow aaabbb$

② $S \Rightarrow aS$
$\Rightarrow aaS$
$\Rightarrow aaaS$
$\Rightarrow aaabS$
$\Rightarrow aaabbS$
$\Rightarrow aaabbbS$
$\Rightarrow aaabbb$

③ $S \Rightarrow aSb$
$\Rightarrow aaSbb$
$\Rightarrow aaaSbbb$