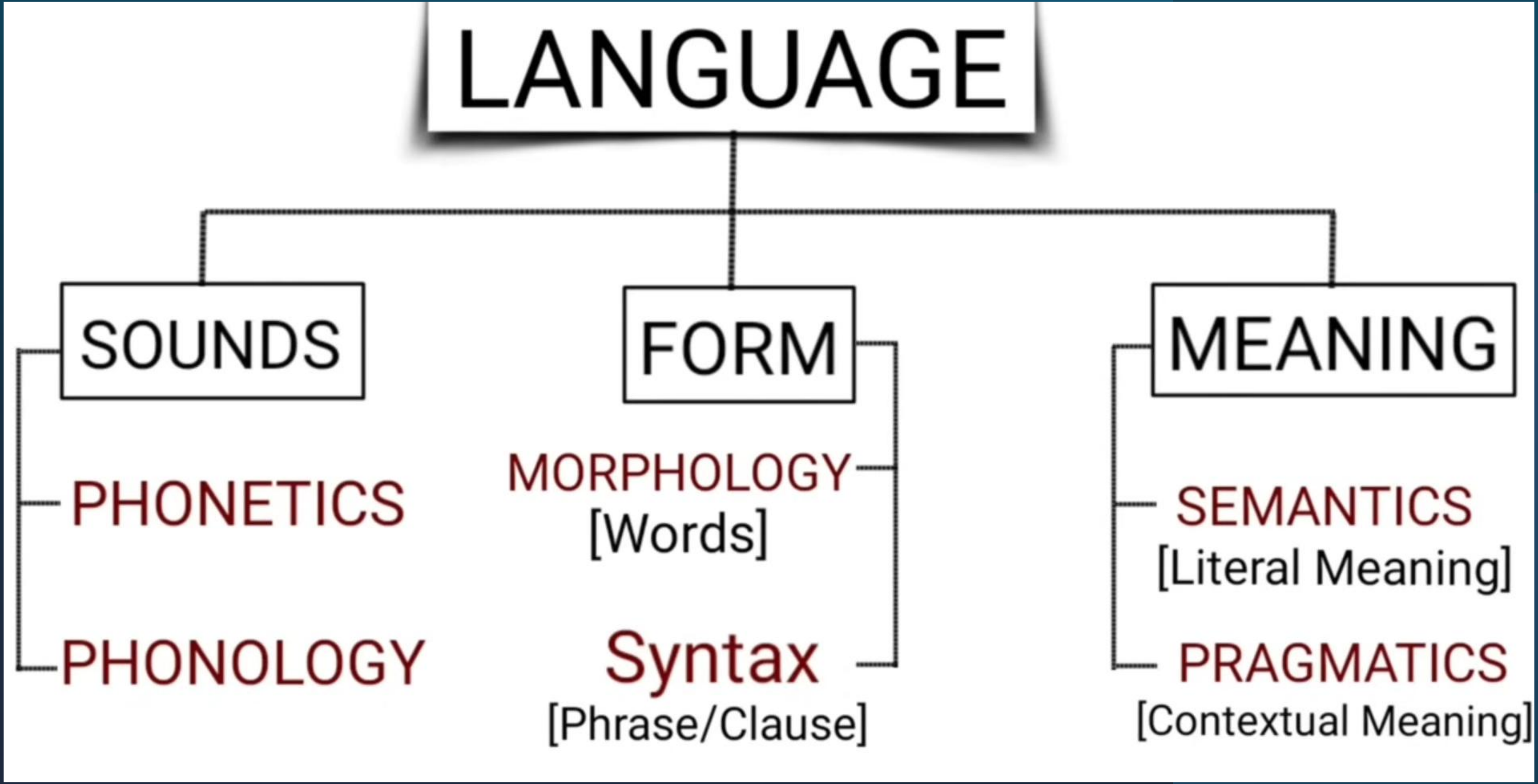
Natural Language Processing

Linguistic Terminology

What is Linguistic Terminology in NLP?

- In **NLP (Natural Language Processing)**, the term **linguistic terminology** refers to the set of concepts, categories, and labels borrowed from **linguistics**—the scientific study of language—that are used to describe, analyze, and process human language in computational systems.
- NLP is about teaching computers to understand, interpret, and generate human language.
  Since linguistics already studies:
  - How words are **formed** (morphology)
  - How they **sound** (phonology)
  - How they are **written** (orthography)
  - How they are **structured** (syntax)
  - How they **mean things** (semantics)
  - How they are **used in context** (pragmatics)

...NLP borrows this terminology to design algorithms, models, and datasets.



LANGUAGE

SOUNDS

- PHONETICS
- PHONOLOGY

FORM

- MORPHOLOGY [Words]
- Syntax [Phrase/Clause]

MEANING

- SEMANTICS [Literal Meaning]
- PRAGMATICS [Contextual Meaning]

Morpheme

- A morpheme is the smallest unit of meaning in a word.
- The smallest unit of meaning (e.g., “un-” in “undo”)
- Example:

  "played" = *play* (base meaning) + *-ed* (shows past tense)

  "cats" = *cat* + *-s* (shows plural)

  "unfriendly" = *un-* + *friend* + *-ly*
- Morphs are the actual parts of a word that carry meaning.

Types of Morphemes

1. Stems (Root Words)

- These are the **main part** of the word with core meaning.
- Examples:
  - *play*, *cat*, *friend*

Examples of Morphological Structure

- **replayed** → *re-* + *play* + *-ed*
- **computerized** → *comput* + *-er* + *-ize* + *-ed*

2. Affixes

- These are **added to stems** to change their meaning or form.
- Two types:
  - **Prefixes** – added **before** the stem (e.g., *un-* in *unhappy*)
  - **Suffixes** – added **after** the stem (e.g., *-ed* in *played*)

Graphemes

- A grapheme is the smallest unit of writing in a language.
- It represents how a sound (phoneme) is written down.
- Graphemes are visual symbols – the way we write language.
- They are not always the same as sounds (phonemes).
- One sound can be written in different ways:

  /f/ sound → f (fan), ph (phone), gh (enough)

Grapheme	Represents this sound	Example Word
a	/æ/ (as in cat)	cat
b	/b/	bat
sh	/ʃ/	shop
ch	/tʃ/	chair

Types of Graphemes:

- **Single-letter graphemes**: One letter = one sound

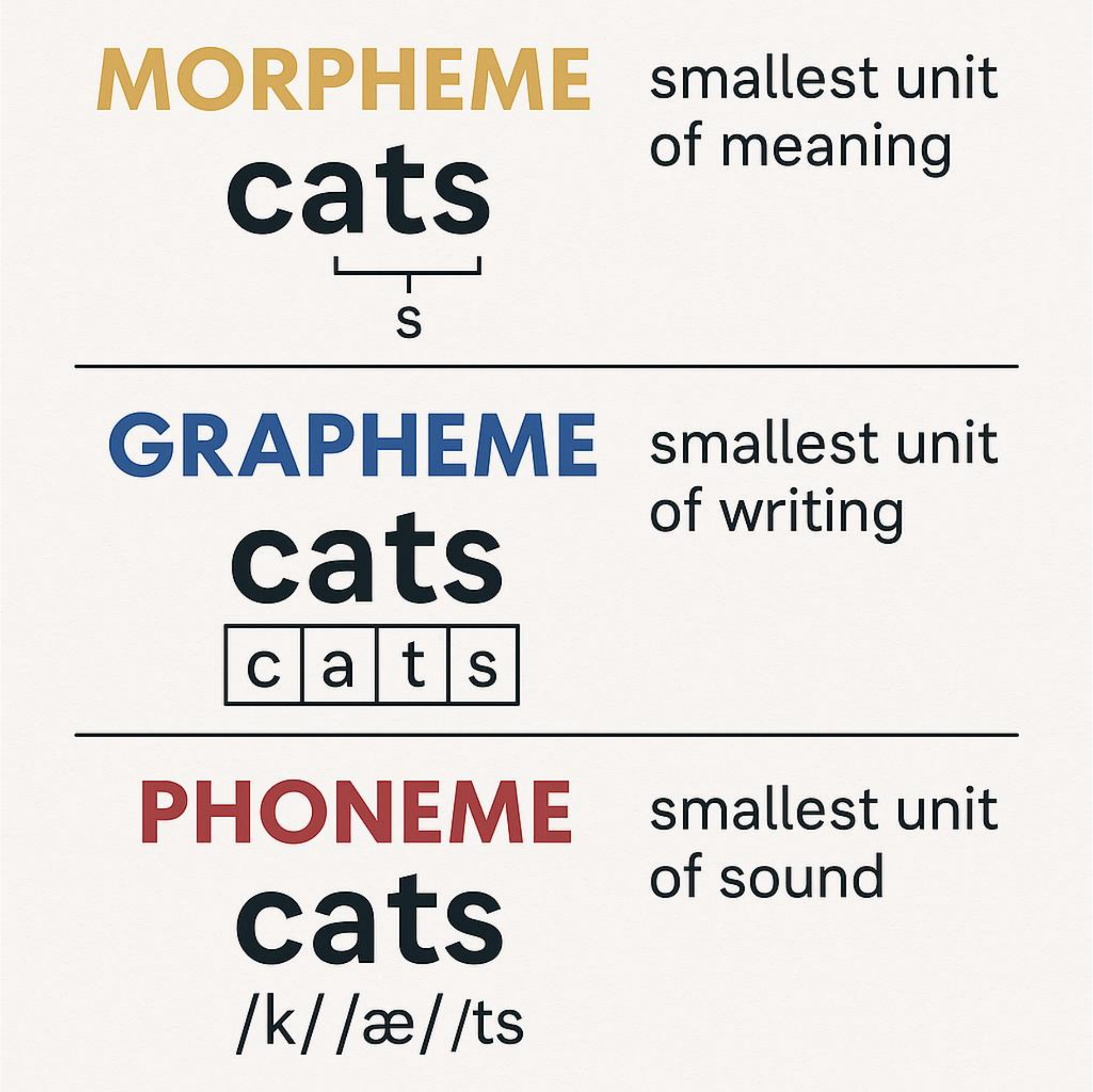
  e.g., **b**, **t**, **m**

- **Multi-letter graphemes**: Two or more letters = one sound

  e.g., **sh**, **th**, **ph**, **ee**

Phoneme

- A phoneme is the smallest unit of sound in a word that helps distinguish meaning.
- Example: cat vs rat – only the first phoneme is different.
- A letter can represent one phoneme, but often one phoneme has multiple spellings in English.
- Some letters (like x) represent two phonemes (e.g., /k/ + /s/).
- English has around 44 phonemes, including:
  1. Short & long vowels
  2. Consonants
  3. Digraphs like /sh/, /th/, /ch/
  4. Diphthongs like /oy/, /ou/
  5. Special sounds like /ng/, /ar/, schwa, etc.
- Phonemes are about sound, not spelling. They define how we hear and understand words.



MORPHEME

cats

s

smallest unit of meaning

GRAPHEME

cats

c a t s

smallest unit of writing

PHONEME

cats

/k/ /æ/ /ts

smallest unit of sound

Classical Approaches of NLP

1. Rule-Based Approach:

- Uses **manually written grammar rules** to understand and generate language.
- Example:
  - Rule: "If a word ends in -ed, it may be past tense."
- Requires **lots of human effort** and **language expertise**.

Good for: Simple, predictable language tasks
Bad at: Handling exceptions, slang, or ambiguity

2. Lexicon-Based Approach

- Uses a **dictionary (lexicon)** of words with their meanings, grammar, and roles.
- Each word is tagged with info like part of speech, synonyms, etc.
- Used in tasks like **sentiment analysis** and **part-of-speech tagging**.

Good for: Understanding word roles and meaning
Bad at: Handling new or unknown words

3. Grammar-Based (Parsing) Approach:

- Focuses on **sentence structure** (syntax).
- Tries to **build a tree** that shows how words relate grammatically.
- Based on **Context-Free Grammar (CFG)** rules.

Good for: Analyzing sentence structure
Bad at: Free-flowing or ungrammatical language

4. Statistical Approach (Early ML):

- Uses **mathematics and probability** based on real language data.
- Analyses large text datasets to find patterns (corpus-based).
- Common models: **N-grams, Hidden Markov Models (HMMs)**

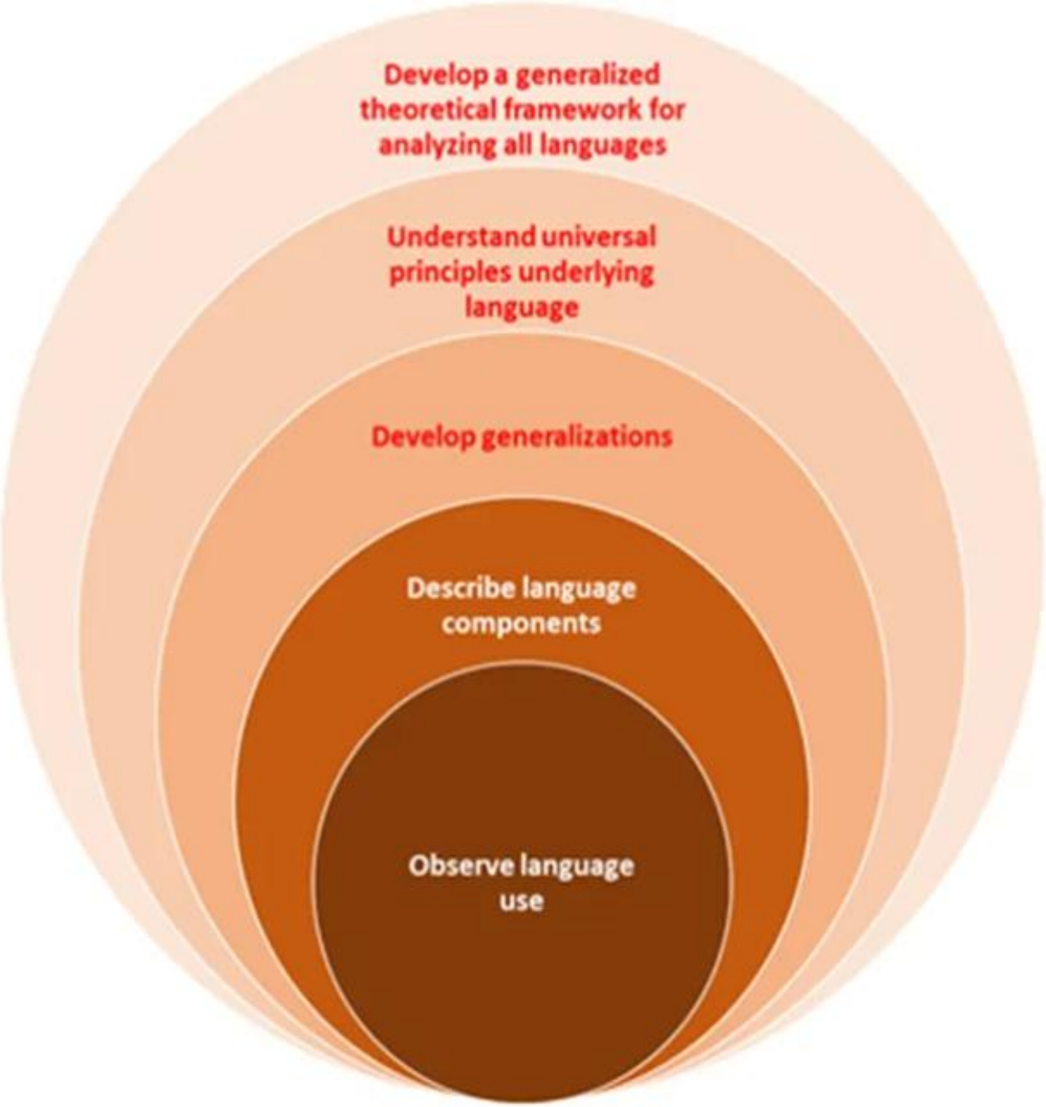
Good for: Predicting likely word sequences
Bad at: Understanding deep meaning or logic

Approach	Based On	Strengths	Weaknesses
Rule-Based	Handwritten rules	Precise control	Hard to scale
Lexicon-Based	Word dictionaries	Good word meaning	Struggles with new words
Grammar-Based	Syntax rules	Sentence structure understanding	Complex grammar required
Statistical	Data + probability	Handles large data, flexible	Lacks deep understanding

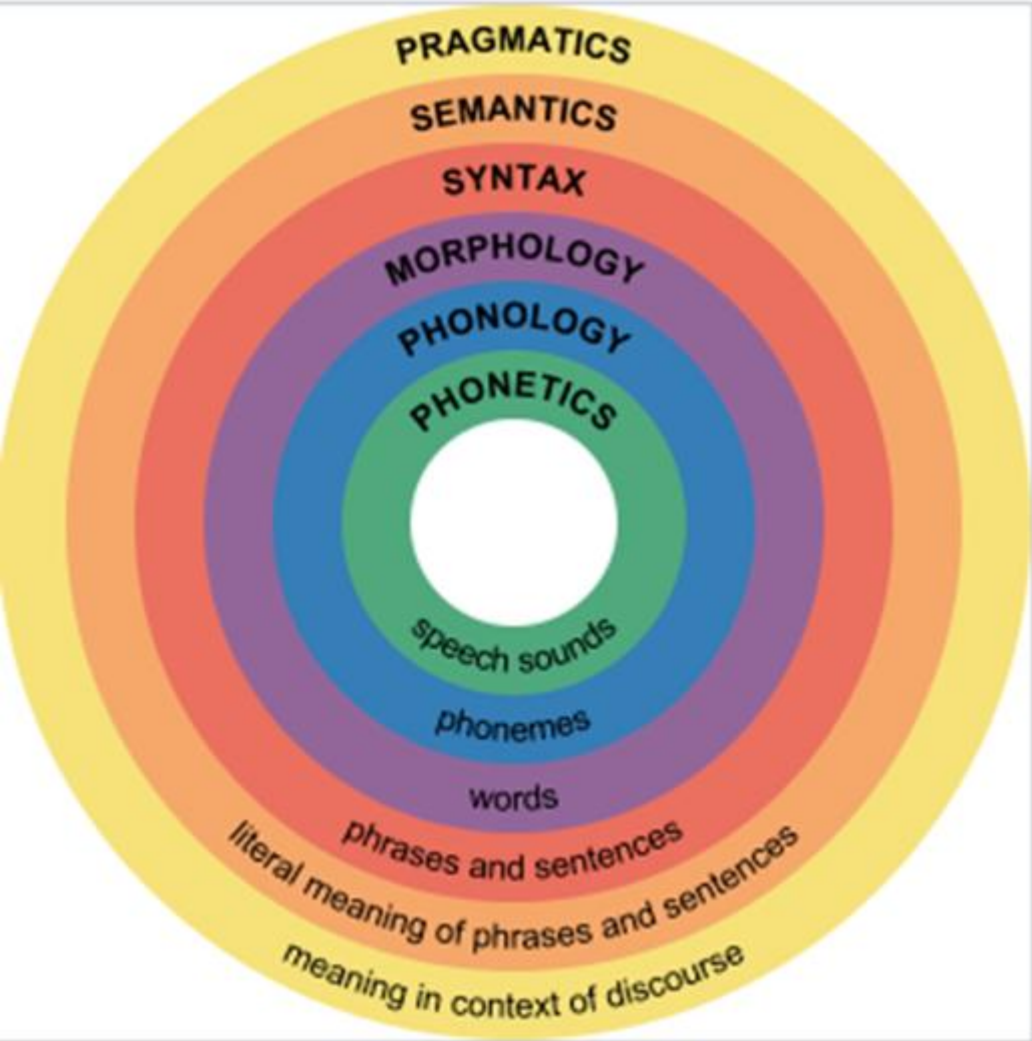
What is Linguistic

- Linguistics is the scientific study of language, and its focus is the systematic investigation of the properties of particular languages as well as the characteristics of language in general.

- It encompasses not only the study of sound, grammar and meaning, but also the history of language families, how children and adults acquire languages, and how language use is processed in the mind and how it is connected to race and gender.


Develop a generalized theoretical framework for analyzing all languages
Understand universal principles underlying language
Develop generalizations
Describe language components
Observe language use

What linguistics does (Image credit: Author created)

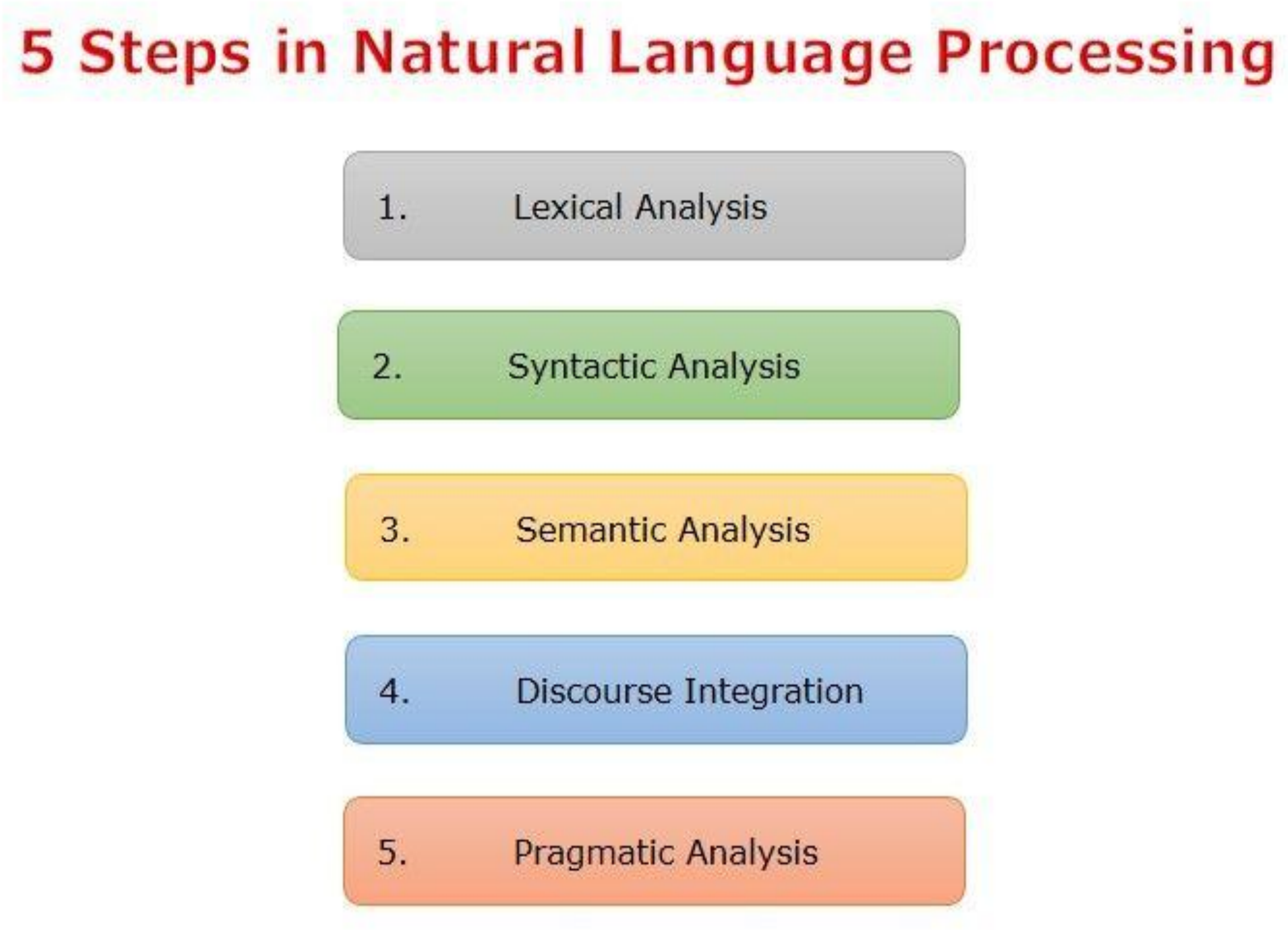


Major levels of linguistic structure

Important subfields of linguistics include

- **Phonetics** - the study of how speech sounds are produced and perceived
- **Phonology** - the study of sound patterns and changes
- **Morphology** - the study of word structure
- **Syntax** - the study of sentence structure
- **Semantics** - the study of linguistic meaning
- **Pragmatics** - the study of how language is used in context
- **Historical Linguistics** - the study of language change
- **Sociolinguistics** - the study of the relation between language and society
- **Computational Linguistics** - the study of how computers can process human language
- **Psycholinguistics** - the study of how humans acquire and use language

Understanding Linguistic



5 Steps in Natural Language Processing

1. Lexical Analysis
2. Syntactic Analysis
3. Semantic Analysis
4. Discourse Integration
5. Pragmatic Analysis

1. Phonetics: the study of speech sounds

- Phonetics is the study of speech sounds in any language.
- It focuses on how sounds are produced, transmitted, and received, not their meanings.
- It is the most basic level of language analysis.
- **Phone**: The smallest unit of sound in phonetics (e.g., a single speech sound).
- Phonetics deals with physical sounds, not the meaning of those sounds.
- Involves **vocal organs** like the **lips, tongue, and teeth**.
- Sounds are organized and studied using tools like the **International Phonetic Alphabet (IPA)**.

- 🔹 **Subfields of Phonetics**
  - **Articulatory Phonetics** – How speech sounds are physically produced.
  - **Acoustic Phonetics** – The physical properties of speech sounds (like frequency, amplitude).
  - **Auditory Phonetics** – How the ear and brain perceive speech sounds.
- 🔹 **Use in NLP**
- Early **Automatic Speech Recognition (ASR)** systems used phonetic elements.
- These were **not very effective** due to the huge variability in how sounds are spoken.
- **Modern ASR** uses **transformer-based models**, which perform much better.

2. Phonology

- **Phonology** is the study of **how speech sounds (phonemes)** are **organized and used** in a language.
- It focuses on how sounds are **represented in the mind** and how they help **convey meaning**.
- It looks at **patterns and rules** for **pronouncing words** correctly in a given language.
- Phonology studies **rules** like:
    - Which sounds can start a word
    - What sound combinations are allowed
    - Stress and intonation patterns

➢**Uses of Phonology in NLP**

- Mostly used in **academic and theoretical** linguistics.
- **Limited use** in practical NLP systems.
- Some use in **speech recognition**, but early systems using phonology weren't very successful.
- **Deep learning models** (like transformers) now handle speech recognition more effectively.

Aspect	Phonetics	Phonology
Focus	Physical sound production	Mental representation of sounds
Concerned With	How sounds are made	How sounds are **used and understood**
Example	Each 'p' sound may be slightly different	All 'p' sounds are treated as the same phoneme

3. Morphology: The study of word structure

- Morphology is the **study of word structure** — how words are **built** and how they **relate** to other words.
- It breaks words into **morphemes**, which are the **smallest units of meaning**.

◆ **Examples of Morphemes**

- **"cats"** → *cat* (stem) + *-s* (plural suffix)
- **"played"** → *play* (stem) + *-ed* (past tense suffix)

- Languages differ in how they use morphology:
  1. **Isolating Languages** – Use few or no affixes (e.g., **Chinese**).
  2. **Agglutinative Languages** – Words are built from **clear morpheme chains** (e.g., **Turkish**).
  3. **Inflectional Languages** – Morphemes are **fused into word endings** with complex meaning (e.g., **Latin)**
- **Use of Morphology**

1. **Stemming** – Cutting off word endings (e.g., *played* → *play*)

   Tool: **Porter Stemmer**

2. **Lemmatization** – Converts word to its **dictionary (base) form**, considering grammar.

   Tool: **WordNet Lemmatizer**

Type	Description	Examples
Root/Stem	Main part of a word	play, cat
Affix	Added part to modify meaning	-ed, -s, un-
Prefix	Before the stem	*un-* in *undo*
Suffix	After the stem	*-ing* in *playing*

4. Syntax: study of grammar

- Syntax is the study of grammar and sentence structure.
- It focuses on how words and morphemes combine to form phrases and sentences that make sense
- **Syntactic Structure** = The rules that decide which word comes where in a sentence.
- Every language has a set word order (like Subject → Verb → Object in English).
- **Constituents** = Words or phrases that form a single unit in a sentence.
- Words are grouped into **categories**: noun, verb, adjective, etc.

🔶 **Common NLP Tools:**

- **Treebanks**: Annotated corpora with parse trees (e.g., Penn Treebank, Prague Treebank).
- **POS Taggers**: Mark each word with its part of speech (noun, verb, etc.).
- **Parsers**: Break sentences into grammatical structures (syntax trees).

5. Semantics: study of word meanings

- **Semantics** is the **study of meaning** in language.
- It focuses on **what words, phrases, and sentences mean**, and how meaning changes with context.
- Example: The word **"bank"** can mean:

  A financial institution

  The side of a river

- **Word Vectors** (embeddings) represent word meanings as numbers.
- **Earlier tools**:

  Pre-2010: Used similarity scores, corpus statistics

  Word2Vec, **GloVe**: Created **dense word embeddings** using context

- **Current trend**:
  - Transformers (like BERT, GPT) learn **dynamic** context-based word meanings

1. Lexical Semantics

- Deals with **word meanings** and their **relationships**.
- Studies how words behave in grammar and in combination (compositionality).
- Key Concepts:
  - ✓ **Synonymy** – same meaning (e.g., big/large)
  - ✓ **Antonymy** – opposite meaning (e.g., hot/cold)
  - ✓ **Polysemy** – one word, many meanings (e.g., bank)
  - ✓ **Hypernymy, Troponymy** – advanced meaning relations (tree > plant, walk > move)

2. Distributional Semantics

- Based on the idea that **"a word is known by the company it keeps"** – J.R. Firth
- Word meaning is learned from **context**.
- If two words appear in similar contexts, they are considered **semantically similar**.

6. Pragmatics – Study of Language in Social Contexts

- **Pragmatics** is the study of how **language is used in real-life situations**.
- It focuses on the **implied meaning** behind what people say, depending on **context**, **tone**, **social norms**, etc
- It goes **beyond the literal meaning** of words.

🔷 Speech Acts (Direct vs Indirect)

- **Direct Speech Acts**:

  "Give me water." → Clearly a command.

- **Indirect Speech Acts**:

  "I'm really thirsty." → Implies a request for water.

🔷 Types of Pragmatic Analysis

- **Speech Acts** – Language used to perform actions (e.g., request, promise, order).
- **Conversational Implicature** – Meaning inferred from indirect speech.
- **Rhetorical Structure** – How ideas are organized in conversation/text.
- **Reference Management** – How speakers refer to people/things (e.g., using "he," "they").

Syntax	**Semantics**
Structure of sentences	Meaning of words and sentences
Rules of word combinations	Word meanings and relationships
Example: “Colourless green ideas sleep furiously.” — Grammatically correct (syntax), but meaningless (semantics).	