Definitions of scientific terms within **Linguistics:**

Phonetics: The study of speech sounds, including their production, articulation, and acoustic properties. Phonetics helps analyze the physical aspects of language sounds.

Phonology: The study of how speech sounds function within a particular language. It involves examining patterns of sound distribution and the rules governing their combinations.

Morphology: The study of the internal structure of words and the rules that govern the formation of words. Morphology investigates morphemes, which are the smallest meaningful units of language.

Morpheme: The smallest meaningful unit in a language. Morphemes can be words or parts of words (prefixes, suffixes, roots) that contribute to the meaning of a word.

Free Morpheme: A morpheme that can stand alone as a complete word with its own meaning, such as "book" or "run."

Bound Morpheme: A morpheme that cannot stand alone as a word and must be attached to a free morpheme, like the "-ed" in "walked" or "un-" in "unhappy".

Affix: A bound morpheme attached to a root word to modify its meaning. Affixes can be prefixes (added before the root), suffixes (added after the root), or infixes (added within the root).

Prefix: An affix added to the beginning of a root word to modify its meaning, such as "un-" in "undo" or "unhappy".

Suffix: An affix added to the end of a root word to modify its meaning, like "-able" in "comfortable."

Root Morpheme: The core lexical unit to which affixes can be added. It carries the central meaning of a word.

Stem: A morpheme that results from attaching affixes to a root. It includes the root and any additional affixes.

Derivation: The process of forming new words by adding affixes to a root or stem, resulting in changes to the word's meaning or grammatical category.

Inflection: The process of adding affixes to a word to indicate grammatical relationships such as tense, number, gender, case, etc., without changing the word's fundamental meaning.

Inflectional Morpheme: A bound morpheme that adds grammatical information to a word without changing its basic meaning, such as plural "-s" in "dogs."

Derivational Morpheme: A bound morpheme that creates a new word with a different meaning or part of speech, such as adding "-ness" to "happy" to form "happiness."

Agglutination: A morphological process in which affixes are added to a root or stem without changing their form, resulting in long, visibly separate strings of morphemes.

Compounding: The process of forming new words by combining two or more independent words to create a single word with a unified meaning, like "toothbrush" or "blackboard."

Syntax: The study of sentence structure and the rules that govern how words are combined to form grammatical sentences.

Syntactic Tree: A graphical representation of the hierarchical structure of a sentence, with nodes representing constituents and branches representing syntactic relationships.

Part of Speech (POS): The concept of "Part of Speech" (POS), also known as "word class" or "lexical category," is a fundamental linguistic classification that categorizes words based on their grammatical and syntactic roles within sentences. Part of speech categories help us understand how words function in context and contribute to the overall structure and meaning of sentences.

Open Word Classes: Open word classes, also known as "content words" or "lexical categories," are groups of words that typically carry the main semantic content of a sentence. These classes have a more flexible and dynamic nature, allowing for the addition of new words over time. Open word classes include: Nouns, Verbs, Adjectives…

Closed Word Classes: Closed word classes, also known as "function words" are groups of words that serve primarily functional or grammatical roles in sentences. These classes are relatively stable and resistant to the addition of new words. Closed word classes include: Prepositions, Determiners, Conjunctions…

Semantics: The study of meaning in language. Semantics examines how words, phrases, and sentences convey meaning, including denotation (literal meaning) and connotation (associative meaning).

Pragmatics: The study of how context influences the interpretation of language. Pragmatics deals with the use of language in real-world situations, including implicature, speech acts, and conversational implicatures.

Definitions of words relevant to the first lecture:

Natural Language Processing (NLP): A subfield of artificial intelligence and linguistics that focuses on enabling computers to understand, interpret, and generate human language.

Language Modeling: Creating statistical models that estimate the likelihood of sequences of words, often used in tasks like speech recognition and machine translation.

Rule-Based Systems: Systems that use explicit linguistic rules to process and analyze text, often in tasks like parsing and information extraction.

Machine Learning: A subset of AI that focuses on training algorithms to improve their performance on a task using data, without being explicitly programmed.

Deep Learning: A subset of machine learning that utilizes artificial neural networks with multiple layers to model complex patterns in data.

Tokenization: The process of breaking text into individual units, such as words or subwords, known as tokens, for further analysis.

Types: Number of unique tokens in a text.

Type-token Ratio: The number of types divided by the number of tokens for a given text or corpus.

Hapax (Hapax legomenon): A token or word that only occurs once in a given text or corpora.

Lemmatization: Reducing words to their base or dictionary form (lemma) to group together inflected forms and improve analysis accuracy.

Collocations: (disputed term within linguistics) a set of words that commonly occur together

Sentence segmentation (sentence splitting): the process of splitting a text into sentences.

Preprocessing: Cleaning and preparing text data for analysis, which may involve tasks like lowercasing, removing punctuation, and stemming.

Feature extraction: The process of defining and extracting relevant features from data.

Case folding: converting all tokens to the same case (usually lowercase)

> **True casing:** replace all sentence initial tokens with the most common case-form for that token (Capitalized, or non-capitalized)

Zipf's law: The product of the frequency of a word f and its position in the frequency list (rank) r is constant: $f \cdot r = k$

Definitions of words relevant to the second lecture:

Training data (training set): The dataset used to train the model.

Development set (Validation set): The dataset used to tune hyperparameters and test different models.

> **Hyperparameters:** Parameters that the model is unable to learn on its own, and therefore has to be specified by the human(s) creating the model.

Test data (test set): The dataset used to test the final model.

K-fold cross validation: A technique that trains K models with different parts held out for validation.

Evaluation measures:

> **Accuracy:** What proportion of our predictions where correct..?
>
> $$\frac{\text{Num\_Correct\_Predictions}}{\text{Num\_Predictions}}$$

	Gold label YES	Gold label NO
System prediction YES	true positive	false positive
System prediction NO	false negative	true negative

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

> **F-measure:** an evaluation measure that weighs precision and recall.

F1-measure (f1-score): $$F_1 = \frac{2 * P * R}{P + R}.$$

To be continued…