# NLP Assignment 2

**Corpora:**

Corpora are essential resources in linguistics and natural language processing (NLP) research. They comprise large collections of text or speech data, carefully curated for linguistic analysis. These collections can range from written texts like books, articles, and websites to transcriptions of spoken language, such as recorded conversations or broadcasts. Researchers use corpora to study language patterns, semantics, syntax, and other linguistic phenomena. By analyzing corpora, linguists can gain insights into language structure, usage, and variation across different contexts and domains.

**Tokens:**

Tokens are fundamental units of text obtained through tokenization, the process of breaking down a text into smaller components. Tokens can include words, punctuation marks, numbers, symbols, or other linguistic elements. Tokenization involves splitting a text into individual tokens based on predefined criteria, such as whitespace, punctuation, or specific patterns. Once tokenized, the text can be further analyzed and processed for tasks like part-of-speech tagging, named entity recognition, and sentiment analysis. Tokens serve as building blocks for various natural language processing tasks and algorithms.

**Unigrams, Bigrams, Trigrams:**

* + Unigrams: Unigrams are single words or tokens in a text. They represent the most basic units of analysis in natural language processing. Unigrams are often used to calculate word frequencies, analyze vocabulary usage, and build basic language models.
  + Bigrams: Bigrams are pairs of adjacent words or tokens in a text. They capture the co-occurrence patterns and relationships between two consecutive words. Bigrams are useful for tasks like language modeling, text classification, and sentiment analysis.
  + Trigrams: Trigrams are sequences of three adjacent words or tokens in a text. Like bigrams, trigrams capture patterns and relationships, but they provide additional context by considering sequences of three words. Trigrams are commonly used in language modeling, machine translation, and text generation tasks.

**Generating n-grams from text:**

To generate n-grams from text, you start by tokenizing the text into individual words or tokens. Then, you create sequences of n consecutive tokens, where n represents the desired length of the n-grams (e.g., 2 for bigrams, 3 for trigrams). By sliding a window of size n across the tokenized text, you extract all possible n-grams. These n-grams can be used for various NLP tasks, such as building language models, analyzing text structure, and extracting meaningful patterns from text data.

**Lemmatization:**

Lemmatization is the process of reducing words to their base or canonical form, known as the lemma. Lemmatization helps normalize words by converting them to their dictionary or root form, which facilitates comparison and analysis. For example, the lemma of "running" is "run," and the lemma of "better" is "good." Lemmatization considers factors like morphology, part of speech, and context to determine the correct lemma for each word. Lemmatization is commonly used in information retrieval, text mining, and natural language processing tasks to improve accuracy and reduce vocabulary complexity.

**Stemming:**

Stemming is the process of reducing words to their root or stem form by removing affixes like prefixes and suffixes. Unlike lemmatization, stemming does not consider the context or meaning of words. Instead, it applies heuristic rules to chop off affixes and produce the stem. Stemming algorithms aim to produce valid stems, but the resulting stems may not always be actual words. Stemming is a more straightforward and computationally efficient approach compared to lemmatization, making it suitable for tasks like text indexing, information retrieval, and document clustering.

**Part-of-speech (POS) tagging:**

Part-of-speech tagging is the process of assigning grammatical categories or tags to words in a text based on their syntactic role and behavior within a sentence. POS tags represent the grammatical properties of words, such as their part of speech (noun, verb, adjective, etc.), tense, number, gender, and case. POS tagging helps disambiguate word meanings and structure sentences correctly. POS tagging is used in various natural language processing tasks, including text analysis, parsing, machine translation, and information extraction.

**Chunking or shallow parsing:**

Chunking, also known as shallow parsing, involves grouping words or tokens into syntactically meaningful chunks or phrases based on their grammatical structure. Chunking does not analyze the full grammatical structure of sentences but focuses on identifying and extracting chunks such as noun phrases (NP), verb phrases (VP), and prepositional phrases (PP). Chunking helps identify higher-level linguistic units and extract relevant information from text data. It is commonly used in information extraction, text summarization, and named entity recognition tasks.

**Noun Phrase (NP) chunking:**

Noun phrase (NP) chunking is a specific type of chunking that focuses on identifying and extracting noun phrases from a text. Noun phrases consist of a noun and any associated modifiers, such as adjectives, determiners, and prepositional phrases. NP chunking helps identify key entities and concepts in text data, making it useful for tasks like information extraction, question answering, and sentiment analysis. NP chunking is often performed as part of larger NLP pipelines to extract structured information from unstructured text.

**Named Entity Recognition (NER):**

Named Entity Recognition is a natural language processing task that involves identifying and categorizing named entities or proper nouns in a text. Named entities can include people's names, locations, organizations, dates, times, monetary values, and other entities of interest. NER systems typically use machine learning algorithms and linguistic patterns to recognize and classify named entities in text data. NER is an essential component of information extraction systems, chatbots, question answering systems, and entity linking applications.