**1. Question: What are Corpora?**

Answer: Corpora (singular: corpus) are large collections of written or spoken texts used for linguistic analysis and research. They serve as a source of data for studying language patterns, syntax, semantics, and other linguistic properties. Corpora can consist of various types of texts, including literature, news articles, transcripts, social media posts, and more.

**2. Question: What are Tokens?**

Answer: Tokens are individual units of text that result from splitting a sentence or document into smaller segments, such as words or punctuation marks. Tokens are the building blocks used for text analysis and natural language processing tasks. For instance, the sentence "The quick brown fox jumps over the lazy dog." can be tokenized into the words: ["The", "quick", "brown", "fox", "jumps", "over", "the", "lazy", "dog", "."].

**3. Question: What are Unigrams, Bigrams, Trigrams?**

Answer: Unigrams, bigrams, and trigrams are different types of n-grams, which are sequences of consecutive words within a text.

- Unigrams: N-grams of size 1, representing individual words in the text.

- Bigrams: N-grams of size 2, representing pairs of adjacent words.

- Trigrams: N-grams of size 3, representing sequences of three consecutive words.

**4. Question: How to generate n-grams from text?**

Answer: To generate n-grams from text, you slide a window of n words across the text and create a sequence of words within that window. For example, given the sentence "Natural language processing is fascinating," and n=2 (bigrams), you'd generate: ["Natural language", "language processing", "processing is", "is fascinating"].

**5. Question: Explain Lemmatization.**

Answer: Lemmatization is a text normalization technique that reduces words to their base or root form, known as a lemma. It involves removing inflections and variations to unify words with the same semantic meaning. For instance, lemmatizing "running" and "ran" both results in "run."

**6. Question: Explain Stemming.**

Answer: Stemming is a text normalization technique that reduces words to their root or stem by removing suffixes or prefixes. Unlike lemmatization, stemming might not always result in valid words. For instance, stemming "running" and "ran" might both result in "run," but "runned" would also become "run."

**7. Question: Explain Part-of-speech (POS) tagging.**

Answer: Part-of-speech tagging is the process of assigning grammatical categories (such as noun, verb, adjective) to each word in a sentence. This helps in understanding the syntactic structure of sentences and is used in various language processing tasks like parsing, information retrieval, and machine translation.

**8. Question: Explain Chunking or shallow parsing.**

Answer: Chunking, also known as shallow parsing, involves grouping words in a sentence into meaningful chunks based on their part of speech. It's a step towards identifying more complex syntactic structures in text. Chunking doesn't generate full parse trees but focuses on recognizing phrases like noun phrases (NP) or verb phrases (VP).

**9. Question: Explain Noun Phrase (NP) chunking.**

Answer: Noun Phrase (NP) chunking is a type of chunking that identifies and groups words that together form a noun phrase. Noun phrases typically consist of a noun and its associated modifiers, like adjectives and determiners. For instance, in the sentence "The big red apple," the NP chunk would be "The big red apple."

**10. Question: Explain Named Entity Recognition.**

Answer: Named Entity Recognition (NER) is a process of identifying and categorizing specific entities within text, such as names of people, organizations, locations, dates, and more. NER helps extract structured information from unstructured text and is crucial for tasks like information retrieval, text summarization, and question answering.