Answers to Learning Objectives

1. Tokenization, Stemming, and Lemmatization:
Tokenization is the process of breaking text into smaller units such as words or sentences. Stemming reduces words to their root form by removing suffixes, often producing non-dictionary words. Lemmatization reduces words to their base or dictionary form using linguistic rules.

2. English Morphology:
Inflectional morphology modifies words to express grammatical features like tense or number without changing meaning. Derivational morphology creates new words by changing meaning or word class using prefixes or suffixes.

3. Regular Expressions:
Regular expressions are patterns used to match character combinations in text. Common types include literals, character classes, quantifiers, anchors, and groups.

4. Dictionary Lookup and Finite State Morphology:
Dictionary lookup checks words against a lexical database to validate or retrieve information. Finite state morphology uses finite state machines to model morphological rules efficiently.

5. N-gram Models:
N-gram models predict the next word based on the previous N-1 words. Bigram models use two-word sequences, while trigram models use three-word sequences.