Computational Linguistics (CL3.101)

Summer 2021, IIIT Hyderabad

07 June, Monday (Lecture 7)

Taught by Prof. Radhika Mamidi

An Exercise in Morph Analysis

"Word" is not a well-defined term; therefore, in NLP, the term "token" is used.

A "type" is a *distinct* occurrence of a word in a corpus. Inflected and derived versions are counted separately.

Tokens that occur exactly once in the corpus are called *hapax*.

The "root" of a word is either its original form (in etymology), or its uninflected form (in CL and morphology); it is sometimes called the "base form" for disambiguation. A "lemma" is a headword in a dictionary.

Articles, pronouns, conjunctions, etc. come under "function words"; they carry very little meaning. "Lexical" or "content" words – nouns, verbs, etc. – contain most of the meaning of a sentence.
"Stop words" are those which occur highly frequently and which add little to no meaning. For example, in a program to generate a title for a given paragraph, stop words can be entirely ignored. However, they are important in NLP.