Summary

This unit introduced n-grams as a fundamental concepts of computational and theoretical linguistics alike.

- A negative n-gram grammar is a set of n-grams.
  - A string is well-formed iff it does not contain any forbidden n-grams.
  - An n-gram grammar is fixed if all n-grams have the same length, and mixed otherwise.
- Every negative n-gram grammar has an equivalent positive n-gram grammar, and the other way round.
- Positive grammars must be fixed, the length of n-grams cannot vary.
- The maximum size of an n-gram grammar grows polynomially with the size of the alphabet, and exponentially with the value of n.
- Multiple grammars can be combined into a single grammar.
  - Negative grammars: union of sets
  - Positive grammars: intersection of sets
- Whereas an n-gram grammar is a set, a bag-of-words model is a multiset.
  - The multiset counts for each word type its number of word tokens.
  - Multiset sum and scalar multiplication can be used to combine and modify counts.
- Due to Zipf's law, a small number of words make up the majority of each text. Very frequent words that contribute little information are called stop words.
- The function del_S removes all stop words.
- Mathematically, this is the same as constructing phonological tiers.
- Tiers make it possible to handle long-distance dependencies in an elegant fashion with much smaller grammars.

1 Some additional terminology

We now have two types of n-gram grammars: those that regulate strings, and those that regulate tiers. The former are commonly referred to as **strictly local** (SL) grammars, whereas the latter are **tier-based strictly local** (TSL). A TSL grammar consists of both an SL grammar G and a set T of tier symbols. A string s is well-formed with respect to the TSL grammar iff $\text{del}_{+T}(s)$ is well-formed with respect to G.

By default SL and TSL grammars are negative, but positive counterparts can be defined as usual.