**n-Gram Models of Morphological Derivations**

Aniello De Santo and Alëna Aksenova
Department of Linguistics @ SBU

**Introduction**

Computational linguistics is concerned with modeling natural language, and modeling the structure of words – morphology – is an essential part of it. Morphology studies what different morphemes mean, how they apply to the stem, and so on. For example, *of* is a morpheme that can apply to a root node in order to derive its past tense *walked*.

Here we present a modification of the traditional n-gram approach to modeling language that allows the following:

- capture ambiguity, in cases such as undeckable;
- more efficiently limit behavior of morphemes.

This will result in better semantic extraction while parsing, and will improve our understanding of possible morpheme combinations.

**1 n-Grams Models in Natural Language**

Widely known in natural language processing, n-gram models can be used to decide whether a word belongs to a language or not, by banning specific substrings of length n listed in a grammar.

**Example 1**: Intercovalic voicing in German

- in **German**, *Af* is realized as [ʔ] between two vowels:
  1. Faser → [fazər] fibre
  2. reisen → [rɛi.zen] to travel
- other consonants are unaffected:
  3. Wasser → [wa.zer] water
  4. reiste → [ri.ztɛ] traveled
- banned 3-grams:

  - [asə, iəsə, eːsə, isə, iəs, ...]
  - [ɛ i ɛ n]

**Example 2**: Word-final devoicing in German

- in **German**, *Af* is realized as [ʔ] at the end of the word:
  1. Kind → [kind] child
  2. Kinder → [ki.nder] children
- banned 2-grams:

  - [dvə]
  - [v n, v n, v n]

**2 Limits**

A surprising number of natural language patterns can be captured by these very simple models. But not all patterns are well-behaved.