ANTONIS ANASTASOPOULOS
CS499 INTRODUCTION TO NLP
WORDS AND MORPHOLOGY

https://cs.gmu.edu/~antonis/course/cs499-spring21/
DEFINITIONS
A word is an ill-defined concept: 

- do — do not — don’t
- Lebensversicherungsgesellschaftsangestellter (life insurance company employee)
- 莎拉波娃现在居住在美国东南部的佛罗里达。 (Sharapova now lives in Us southeastern Florida)

**Type:** a class of tokens that use the same character sequence

**Token:** an individual occurrence of a type in speech or writing

**Vocabulary:** the set of types

https://en.wikipedia.org/wiki/Type%E2%80%93token_distinction
A rose is a rose is a rose.

#Types: 4

Vocabulary: \{a, rose, is, .\}

#Tokens: 9
Corpus: a computer-readable collection of text or speech
“Don’t think of an elephant!,” says George.

Elephants are not something you should be thinking, according to Lakoff.

Dr. Lakoff asks that you do not think of an elephant.
"Don’t think of an elephant!," says George.

Elephants are not something you should be thinking, according to Lakoff.

Dr. Lakoff asks that you do not think of an elephant.
“don’t think of an elephant!”, says George.
elephants are not something you should be thinking, according to Lakoff.
dr. Lakoff asks that you do not think of an elephant.

Tools:
- NLTK (https://www.nltk.org/)
- spacy (https://spacy.io/)
MORPHOLOGY
Words are not atoms:

• they have internal structure
• they are composed of morphemes
• most languages make extensive use of morphology, but English and Chinese do not

Morphology is the study of the structure of words

mis - understand - ing - s
un - dead
re - implement - ation

同志们 (tong-zhi-men) comrades

The minimal meaningful units are called morphemes.
HIERARCHICAL STRUCTURE

Words are not necessarily sequences of morphemes

kingfishers
  kingfisher
    king
    fish
  -s
  fisher
    fish
    er
WHAT IS A WORD?

Defining a word is not straightforward:

• Whitespace? — some languages don’t use it.
• Listed in dictionary? — but dictionaries can list multi-word expressions (listemes) which are idiosyncratic
• A single phonological domain?
• Speakers don’t always intuitively agree.
TWO TYPES OF MORPHOLOGY

Inflectional Morphology

- Adds grammatical information to a word
- The word doesn’t change part-of-speech
  
  argument — arguments
  walk — walks
  she — hers — her

Derivational Morphology

- Creates new words with new meanings (and often with new part-of-speech)
  
  argument — argumentation
  parse — parser
  repulse — repulsive

mis - understand - ing - s
TYPES OF MORPHEMEs

Root — the central morpheme that carries the main meaning

Affixes:

Prefix  
pre-nuptual, ir-regular

Suffix  
conceptual-ize, regulat-or

Infix  
Pennsyl-fu%&!n-vania

Circumfix  
ge-sammel-t (German)

Non-concatenative morphology

Umlaut  
Tooth-teeth —— foot-feet

Ablaut  
sing, sang, sung

Reduplication  
anak (child) —> anak-anak (children)

Root-and-pattern (templatic)

Common in Arabic, Hebrew, and other Afroasiatic languages

Roots made of consonants, vowels are shoved into the root

<table>
<thead>
<tr>
<th>Active</th>
<th>Passive</th>
<th>Active</th>
<th>Passive</th>
<th>Perfect</th>
<th>Passive</th>
</tr>
</thead>
<tbody>
<tr>
<td>I katab</td>
<td>kuttib</td>
<td>ktab</td>
<td>ktab</td>
<td>kaatib</td>
<td>ktab</td>
</tr>
<tr>
<td>II kattab</td>
<td>kuttib</td>
<td>kattab</td>
<td>kattab</td>
<td>kattab</td>
<td>kattab</td>
</tr>
<tr>
<td>III kaatib</td>
<td>kuttib</td>
<td>kaatib</td>
<td>kaatib</td>
<td>kaatib</td>
<td>kaatib</td>
</tr>
<tr>
<td>IV khatib</td>
<td>kuttib</td>
<td>ktab</td>
<td>ktab</td>
<td>ktab</td>
<td>ktab</td>
</tr>
</tbody>
</table>
Tagalog, the basis of Filipino, makes extensive use of both infixation and reduplication in its grammar:

<table>
<thead>
<tr>
<th>Stem</th>
<th>Perfective</th>
<th>Contemplative</th>
<th>Imperfective</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>kain</td>
<td>kumain</td>
<td>kakain</td>
<td>kumakain</td>
<td>‘eat’</td>
</tr>
<tr>
<td>sulat</td>
<td>sumulat</td>
<td>susulat</td>
<td>sumusulat</td>
<td>‘write’</td>
</tr>
<tr>
<td>hanap</td>
<td></td>
<td></td>
<td></td>
<td>‘seek’</td>
</tr>
</tbody>
</table>
NOT EVERYTHING IS REGULAR

Formal Irregularities

Inflectional marking depend on the root

walk — walked — walked
sing — sang — sung

Semantic Irregularity

The same morpheme could have different functions depending on the base it attaches to

A kind-ly old man
*a slow-ly old man
MORPHOLOGICAL ANALYSIS

Input: a word

Output: the word’s stem(s)/lemma(s) and grammatical features expressed by the morphemes

Example:

geese → goose + N + Pl

gooses → goose + V + S + 3p

leaves → { leaf + N + Pl, leave + V + S + 3p }

Checkout UniMorph!
SUBWORDS
WHY SUBWORDS?

Is your first name in an English dictionary?

How many word types are there in English?

What about new words?

Solution:

Work with subwords!
Keep a fixed vocab of subwords (including all characters)
Segment every word as needed.

https://twitter.com/nyt_first_said
BYTE PAIR ENCODING

Init:
   a) split corpus into characters
   b) create character vocabulary

For k steps:
    Find most common pair of adjacent symbols
    Merge them

<table>
<thead>
<tr>
<th>Merge</th>
<th>Current Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>(ne, w)</td>
<td>_, d, e, i, l, n, o, r, s, t, w, er, er, ne, new</td>
</tr>
<tr>
<td>(l, o)</td>
<td>_, d, e, i, l, n, o, r, s, t, w, er, er, ne, new, lo</td>
</tr>
<tr>
<td>(lo, w)</td>
<td>_, d, e, i, l, n, o, r, s, t, w, er, er, ne, new, lo, low</td>
</tr>
<tr>
<td>(new, er)</td>
<td>_, d, e, i, l, n, o, r, s, t, w, er, er, ne, new, lo, low, newer</td>
</tr>
<tr>
<td>(low, _)</td>
<td>_, d, e, i, l, n, o, r, s, t, w, er, er, ne, new, lo, low, newer, low</td>
</tr>
</tbody>
</table>

[Sennrich et al. 2016]
Language Modeling and Smoothing