ANTONIS ANASTASOPOULOS

CS499 INTRODUCTION TO NLP

WORDS AND MORPHOLOGY

https://cs.gmu.edu/~antonis/course/cs499-spring21/

STRUCTURE OF THIS LECTURE

1. Definitions
2. Morphology
3. Working with Subwords

DEFINITIONS

SOME TERMINOLOGY

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

SOME TERMINOLOGY

A rose is a rose is a rose.

#Types: 4

Vocabulary: {a, rose, is, .}

#Tokens: 9

SOME TERMINOLOGY

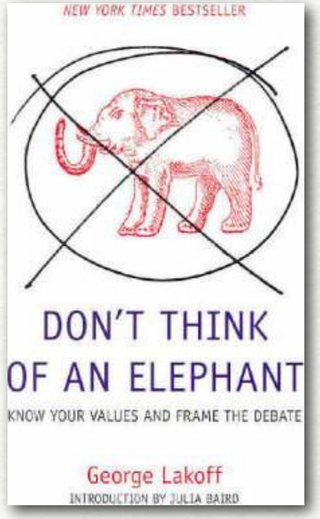
Corpus: a computer-readable collection of text or speech

TEXT NORMALIZATION

“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.

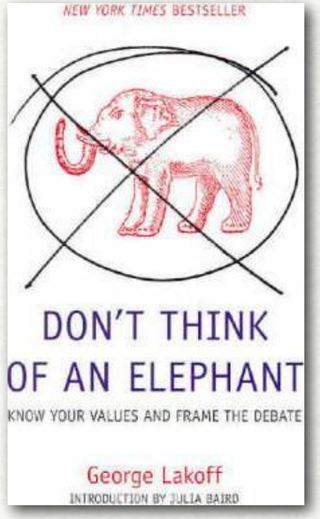


SEGMENTATION

“ Do n’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 .

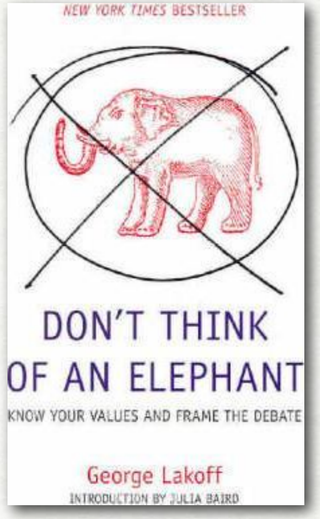


TRUE CASING

“ do n’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/)
- Moses tools (http://www.statmt.org/moses/?n=Moses.SupportTools)

MORPHOLOGY

WORDS

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

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

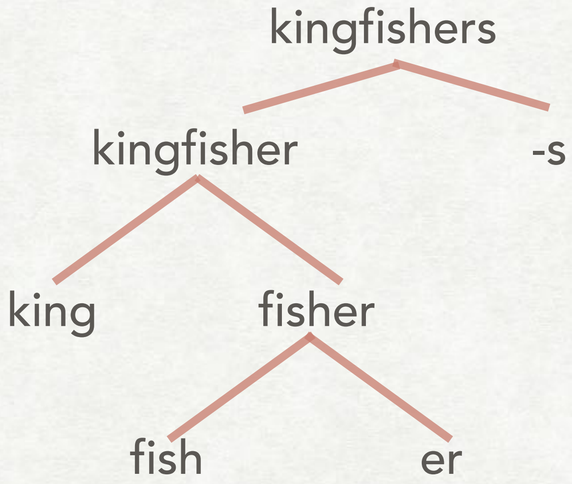
同志们 (tong-zhi-men) *comrades*

The minimal **meaningful** units are called **morphemes.**

Morphology is the study of the structure of words

HIERARCHICAL STRUCTURE

Words are not necessarily sequences of morphemes



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 (l*istemes*) 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

	Perfect		Imperfect		Participle	
	Active	Passive	Active	Passive	Active	Passive
I	katab	kutib	ktub	ktab	kaatib	ktuub
II	kattab	kuttib	kattib	kattab	kattib	kattab
III	kaatab	kuutib	kaatib	kaatab	kaatib	kaatab
IV	ʔaktab	ʔuktib	ktib	ktab	ktib	ktab

EXAMPLE IN TAGALOG

Tagalog, the basis of Filipino, makes extensive use of both infixation and reduplication in its grammar:

Stem	Perfective	Contemplative	Imperfective	Gloss
kain	kumain	kakain	kumakain	'eat'
sulat	sumulat	susulat	sumusulat	'write'
hanap				'seek'

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

corpus		vocabulary
5	l o w _	_, d, e, i, l, n, o, r, s, t, w
2	l o w e s t _	
6	n e w e r _	
3	w i d e r _	
2	n e w _	

For k steps:

Find most common pair of adjacent symbols

Merge them

corpus		vocabulary
5	l o w _	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne
2	l o w e s t _	
6	ne w er_	
3	w i d er_	
2	ne w _	

Merge	Current Vocabulary
(ne, w)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new
(l, o)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo
(lo, w)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low
(new, er_)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low, newer_
(low, _)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low, newer_, low_

[Sennrich et al. 2016]

NEXT CLASS PREVIEW

Language Modeling and Smoothing