TTIC 31190: Natural Language Processing

Lecture 2: Words

Fall 2023

Announcements

- Prerequisite Quiz Papers

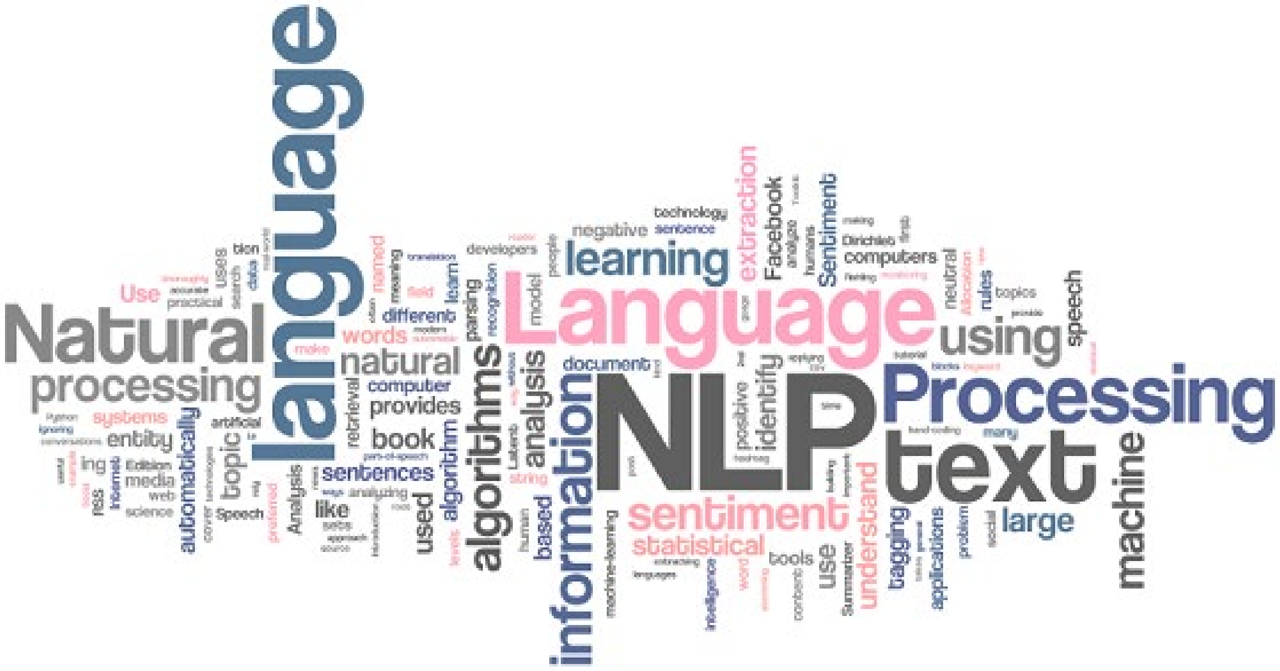
- TA (Jiamin Yang) Tutorial Sessions & Office Hours

  Fridays 3 pm – 4 pm; OH afterwards 4 pm – 5 pm; TTIC Room 530
- Instructor Office Hours

  Tuesdays 1:30 pm – 2:30 pm; location: TTIC 4th floor lobby

- Assignment 1 to be released this week; due in two weeks

Words


language
Natural
processing
Language
NLP
Processing
text
information
algorithms
sentiment
learning
natural
using
machine
extraction
analysis
Facebook
Sentiment
statistical
understand
large
speech
negative
technology
sentence
developers
computers
document
computer
provides
book
sentences
entity
systems
automatically
topic
words
different
identify
positive
tagging
applications
tools
use
used
like
based
media
science
web
Speech
Analysis
retrieval
artificial
Use
practical
uses
search
data
many
topics
rules
neutral
Dirichlet
humans
model
recognition
parsing
analyzing
Edition
Internet
ing
intelligence
machine-learning
Summarizer

What is a word?

“A single distinct meaningful element of speech or writing, used with others (or sometimes alone) to form a sentence and typically shown with a space on either side when written or printed.”

(Oxford Languages)

What is a word?

- The things that are in the dictionary?

What makes a *word* a real word?

The word *word* has a wide range of meanings and uses in English. Yet one of the most often looked for pieces of information regarding *word* is not something that would be found in its definition. Instead, it is some variant of the question, What makes a word a real word?

One of the most prolific areas of change and variation in English is vocabulary; new words are constantly being coined to name or describe new inventions or innovations, or to better identify aspects of our rapidly changing world. Constraints of time, money, and staff would make it impossible for any dictionary, no matter how large, to capture a fully comprehensive account of all the words in the language. And even if such a leviathan reference was somehow fashioned, the dictionary would be obsolete the instant it was published as speakers and writers continued generating new terms to meet their constantly changing needs.

[Src: Merriam-Webster]

What is a word?

- The things that are in the dictionary?

Most general English dictionaries are designed to include only those words that meet certain criteria of usage across wide areas and over extended periods of time (for more details about how words are chosen for dictionary entry, read "How does a word get into a Merriam-Webster dictionary?" in our FAQ). As a result, they may omit words that are still in the process of becoming established, those that are too highly specialized, or those that are so informal that they are rarely documented in professionally edited writing. But the words left out are as real as those that gain

[Src: Merriam-Webster]

What is a word?

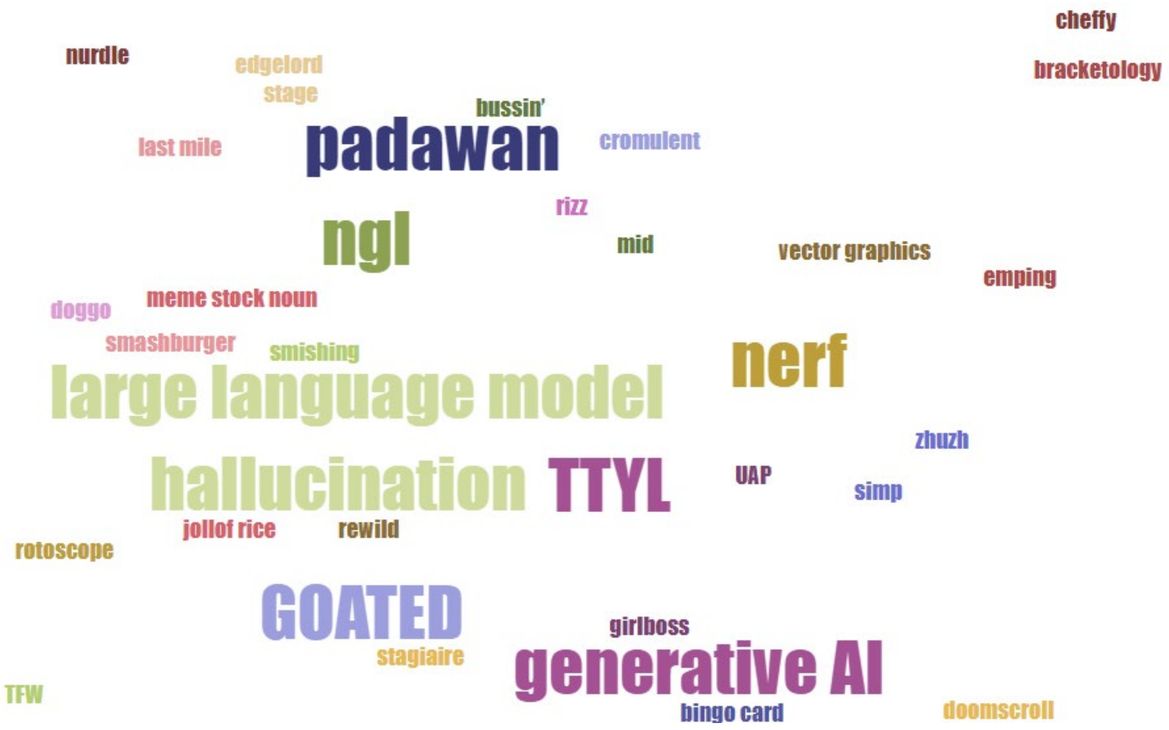
- The things that are in the dictionary?

Merriam-Webster Adds Adds 690 New Words to the Dictionary (September 2023 Update)

September 28, 2023 by Gary Price

From the M-W Website:

Signs of a healthy language include words being created, words being borrowed from other languages, and new meanings being given to existing words. Based on our most recent research, we are pleased to inform you that English is very (very!) healthy.



[Src: Merriam-Webster]

What is a word?

- The things that are in the dictionary?

- The things between spaces and punctuation?

I wish I didn’t have to work.
I wish I did not have to work.
He’s given up his job.
Jimmy’s Pizza Café is expensive!
Did you achieve the state-of-the-art results?

Asian Language Writing Systems

This is Korean: 안녕하세요

This is Chinese: 你好

This is Thai: สวัสดีครับ

This is Japanese: こんにちは

This is also Japanese: グッドモーニング

And this is also Japanese: 猛烈宇宙交響曲

Tokenization

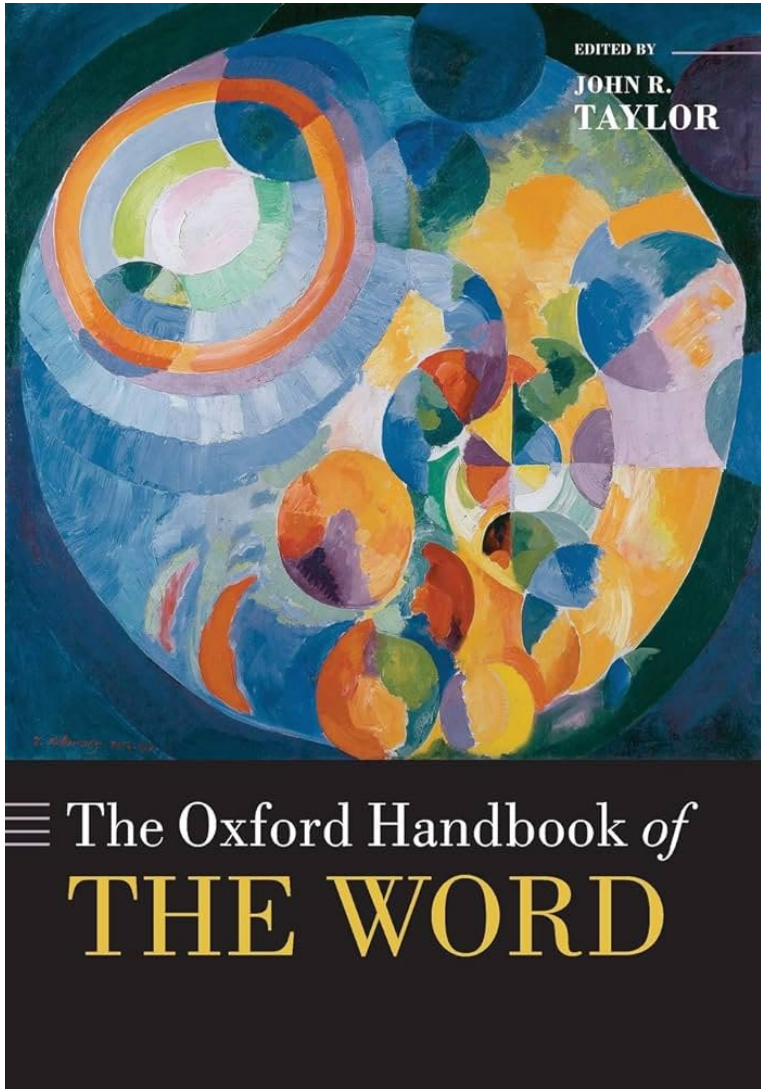
What is a word?

- The things that are in the dictionary?

- The things between spaces and punctuation?

- The smallest unit that can be uttered in isolation?
    - You could say this word in isolation: *Unimpressively*
    - This one too: *impress*
    - But probably not these in isolation, unless you were talking about morphology:
        - *un*
        - *ive*
        - *ly*

What is a word?



- 42 chapters
- Nearly 900 pages
- Covers a lot of different aspects of what makes a word word, “to anyone who shares a fascination with words”

This Lecture

- Linguistic Morphology

  --- The study of internal structures of words

- Lexical Semantics

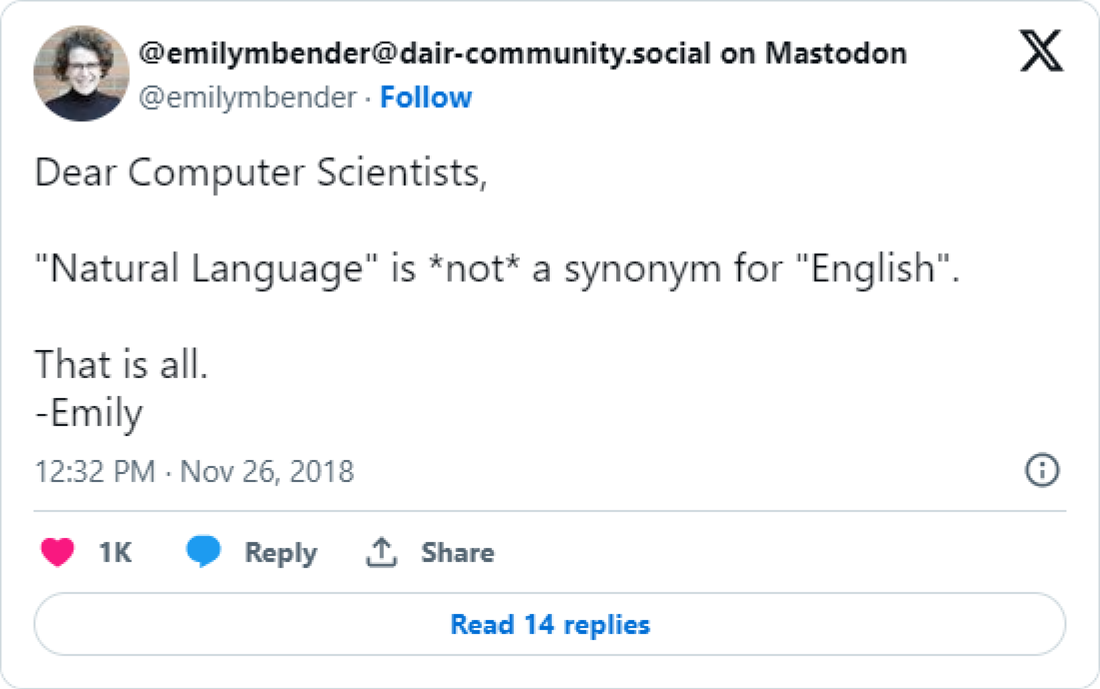
  --- The study of meanings of words

- Word Tokenization

  --- The process of splitting texts into “words” (tokens)

“Bender Rule”

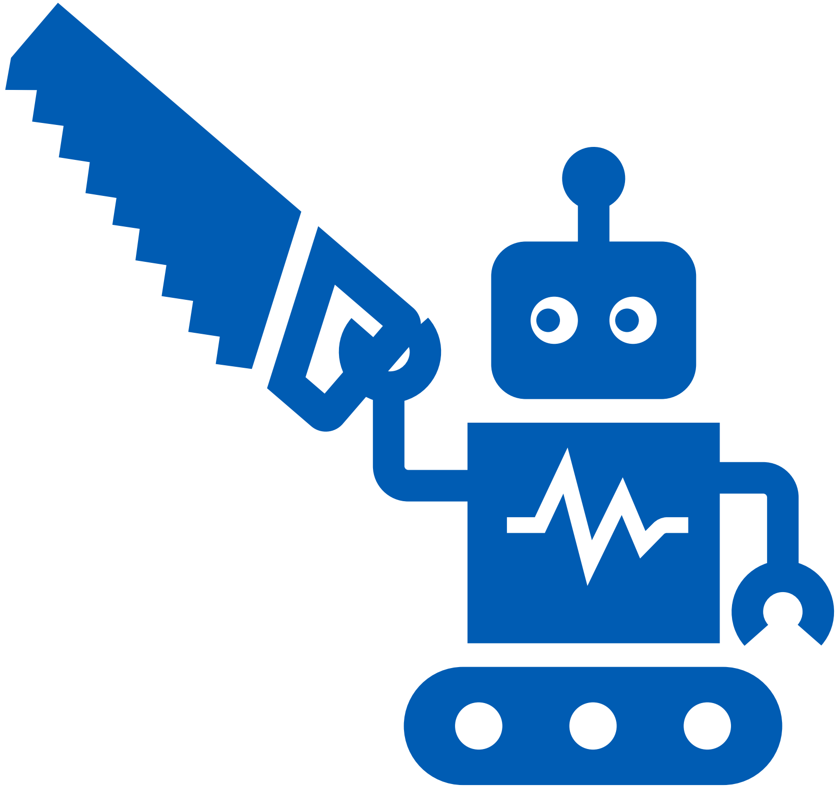
“Always name the language(s) you’re working on.”



colder

replayed

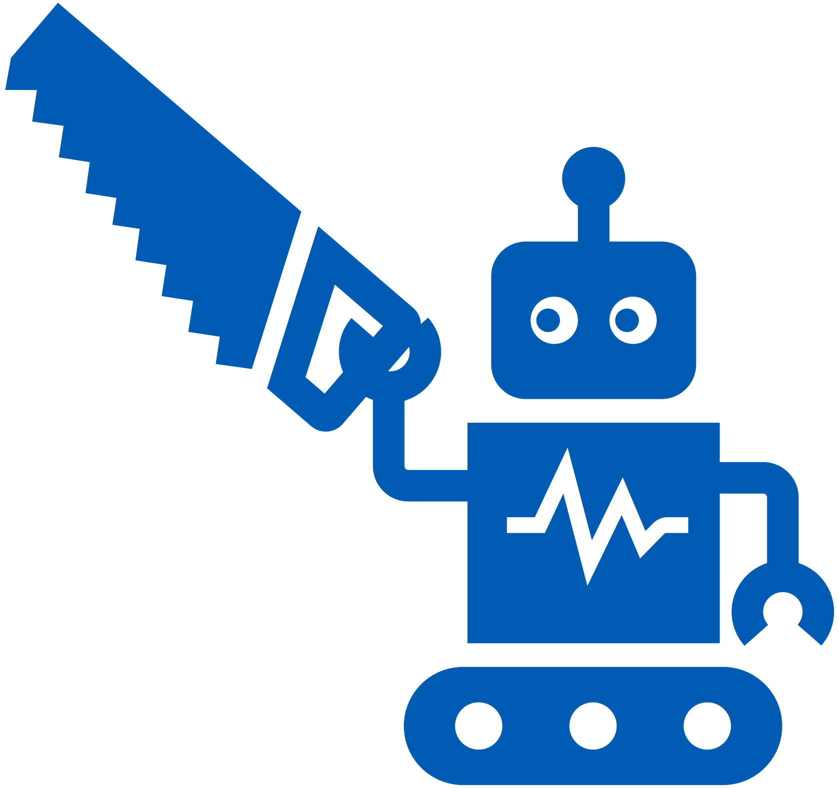
gameplay



cold|er

re|play|ed

game|play



Morphology

The study of how words are built from smaller meaning-bearing units

morpheme: the smallest meaning-bearing unit in a language

types of morphemes:

- **stem**: a core meaning-bearing unit
- **affix**: a piece that attaches to a stem, adding some function or meaning
  - (prefix, suffix, infix, and circumfix)

	stem	affixes
`colder`	`cold`	`-er`
`cats`	`cat`	`-s`
`replayed`	`play`	`re-, -ed`

Sec. 2.4.4 (J&M)

Kinds of Word Formation

- **inflection**: modifying a word with an affix to change its grammatical function (tense, number, etc.)
  - result is another form of the same word

  `book → books` `walk → walked`

- **derivation**: adding an affix to a word to create a new word

  `great → greatly` `great → greatness`

- **compounding**: combining two words

  `lawsuit, keyboard, bookcase`

Kinds of Word Formation

Chinese: **isolating language** (Vietnamese, Thai language, etc.)

- Each word form consists typically of a single morpheme
- Little morphology other than **compounding**

➢Inflection

们： 我们， 你们， 他们

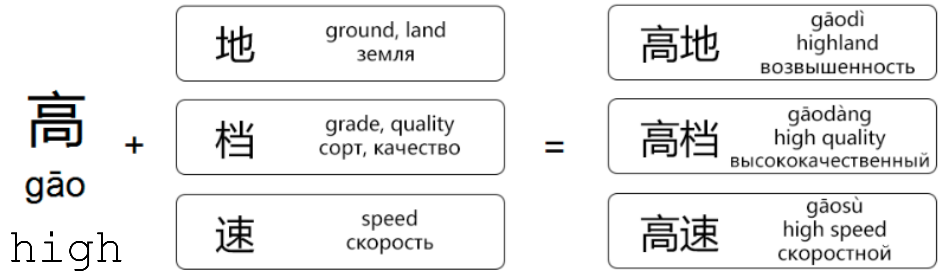
mén: wǒmén, nǐmén, tāmén

plural: we, you (pl.), they

➢Derivation

艺术家 yì shù jiā, artist

➢ Chinese is a champion in the realm of compounding --- up to 80% of Chinese words are actually compounds



Morphological Decomposition

- usually, morphological decomposition is simply splitting a word into its morphemes:

```
walked = walk + ed
greatness = great + ness
```

- but it can actually be a hierarchical structure

```
unbreakable = un + (break + able)
```

Morphological Decomposition

- ambiguity in hierarchical morphological decomposition?

  rare, but it does happen

  example: `unlockable`

  `(un + lock) + able`: “able to be unlocked”

  `un + (lock + able)`: “not able to be locked”

Morphology in NLP

- NLP problems that address morphology:

  lemmatization

  stemming

Word Normalization: putting words/tokens in a standard format

Terminology

- **lemma**
  - canonical/dictionary form of a word
  - words with same lemma have same stem, part of speech, rough semantics
- **wordform**
  - full inflected or derived form of a word as it appears in text

wordform	lemma
run	run
ran	run
running	run

Sec. 2.2 (J&M)

Lemmatization

- **lemmatization**: convert wordform to lemma

  `am, is, are → be`

  `car, cars, car's, cars' → car`

  `the boy's cars are different colors`

  `↓`

  `the boy car be different color`

- mostly about finding the correct dictionary entry, but this may depend on the context

Sec. 2.4.4 (J&M)

Stemming

- **stemming**: reduce words to their stems (approximately) by removing affixes
  - usually implemented with language-specific, manually-designed rules
  - commonly used in information retrieval
  - example:

```
Caillou is an average, imaginative four-year-old boy with a love for
forms of transportive machinery such as rocket ships and airplanes.
```

↓

Caillou is an **averag** **imagin** four year old **boi** with a love for
form of **transport** **machineri** such as rocket **ship** and **airplan**

Sec. 2.4.4 (J&M); en.wikipedia.org/wiki/Caillou

Porter's Algorithm
(the most common English stemmer)

Step 1a

```
sses → ss  caresses → caress
ies  → i   ponies   → poni
ss   → ss  caress   → caress
s    → ø   cats     → cat
```

Step 1b

```
(*v*)ing → ø  walking   → walk
              sing      → sing
(*v*)ed  → ø  plastered → plaster
…
```

Sec. 2.4.4 (J&M)

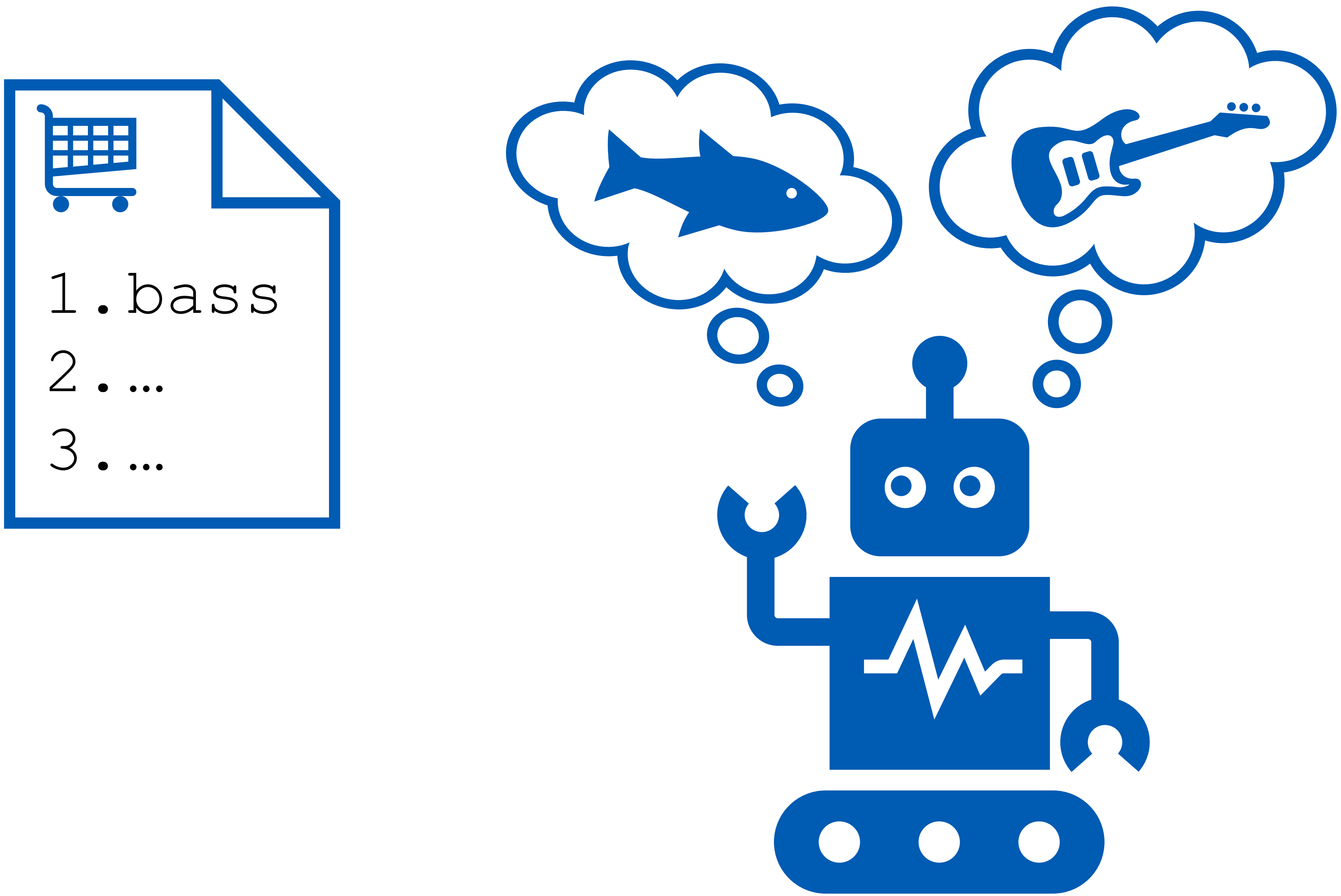
Idiosyncrasies of the Porter Stemmer

$$\left.\begin{array}{l}\texttt{sever}\\ \texttt{severed}\\ \texttt{severing}\\ \texttt{several}\\ \texttt{severe}\\ \texttt{severely}\\ \texttt{severity}\end{array}\right\} \rightarrow \texttt{sever}$$

$$\left.\begin{array}{l}\texttt{wit}\\ \texttt{wits}\\ \texttt{witness}\\ \texttt{witnesses}\\ \texttt{witnessing}\end{array}\right\} \rightarrow \texttt{wit}$$

Lemmatization vs. Stemming

- Lemmatization
  - viewed as an NLP task
  - solved with dictionary look-up, possibly with machine learning
- Stemming uses manually-defined rules
  - simple and fast, but limited due to reliance on rules
  - may conflate words erroneously
- Both may remove information
  - To mitigate, combine lemma/stem form with original form, use both!


1.bass
2.…
3.…

Lexical Semantics

The study of meanings of words

Variability — multiple forms, similar meaning

sing, sang, sung, singing

↓

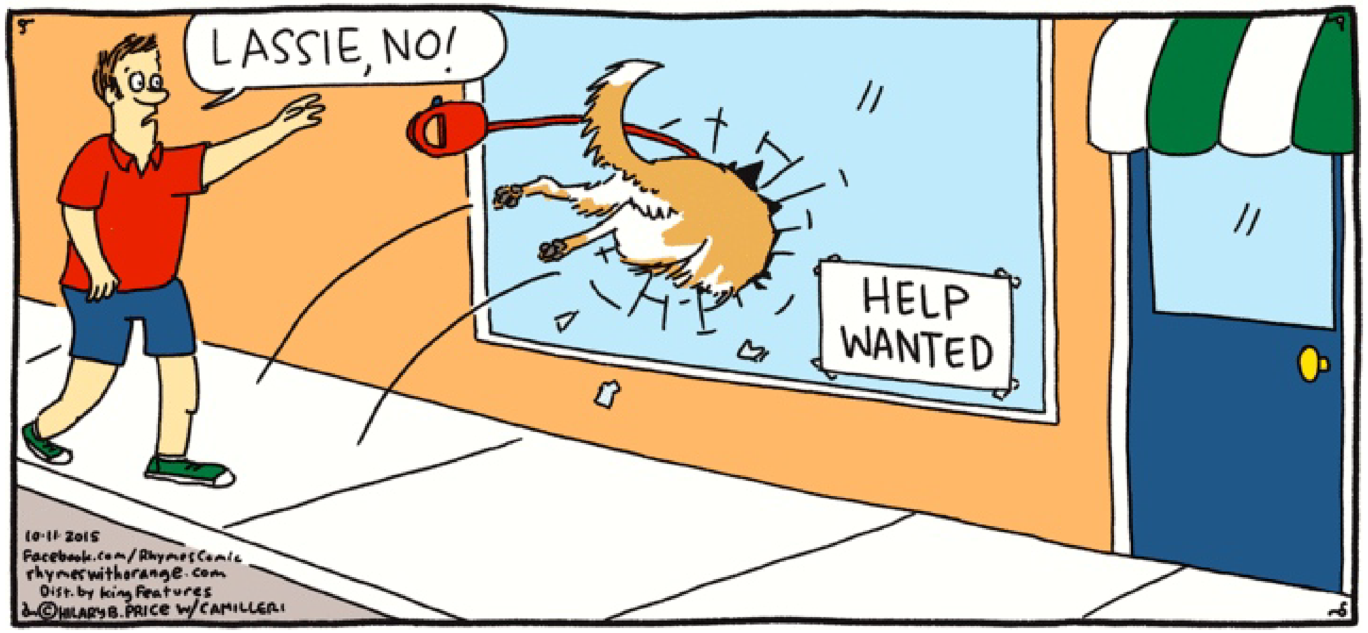
lemmatizer

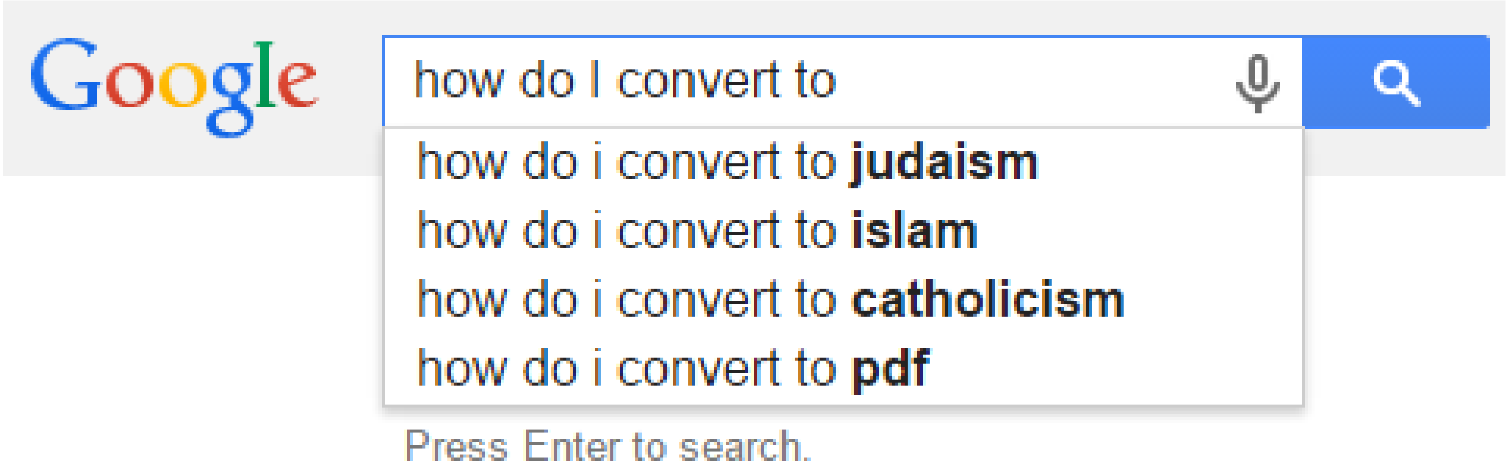
↓

sing

Ambiguity

one form,
multiple meanings





large language model noun

plural **large language models**

: a language model that utilizes deep (see DEEP entry 1 sense 8) methods on an extremely large data set as a basis for predicting and constructing natural-sounding text

> GPT-3 was a *large language model* built by OpenAI that could write impressively human-like poems, sonnets, jokes, and even code samples.
> – Dale Markowitz

> About five years ago, companies like Google, Microsoft and OpenAI began building neural networks that learned from huge amounts of digital text called *large language models* …
> – Cade Metz

→ abbreviation *LLM*

[Src: Merriam-Webster]

deep 1 of 3 adjective

ˈdēp

Synonyms of *deep* ›

1 : extending far from some surface or area: such as

a : extending far downward

a *deep* well

a *deep* chasm

b (1) : extending well inward from an outer surface

a *deep* gash

a *deep*-chested animal

(2) : not located superficially within the body

deep pressure receptors in muscles

c : extending well back from a surface accepted as front

a *deep* closet

d : extending far laterally from the center

deep borders of lace

e sports : occurring or located near the outer limits of the playing area

hit to *deep* right field

2 : having a specified extension in an implied direction usually downward or backward

a shelf 20 inches *deep*

cars parked three-*deep*

3 a : difficult to penetrate or comprehend : RECONDITE

deep mathematical problems

deep discussions on the meaning of life

b : MYSTERIOUS, OBSCURE

a *deep* dark secret

c : grave or lamentable in nature or effect

in *deepest* disgrace

d : of penetrating intellect : WISE

a *deep* thinker

e : intensely engrossed or immersed

She was *deep* in her book.

f : characterized by profundity of feeling or quality

a *deep* sleep

also : DEEP-SEATED

deep religious beliefs

4 a of color : high in saturation and low in lightness

a *deep* red

5 a : situated well within the boundaries

a house *deep* in the woods

b : remote in time or space

found *deep* in rural England

c : being below the level of consciousness

deep neuroses

d : covered, enclosed, or filled to a specified degree → usually used in combination

ankle-*deep* in mud

6 : LARGE

deep discounts

7 : having many good players

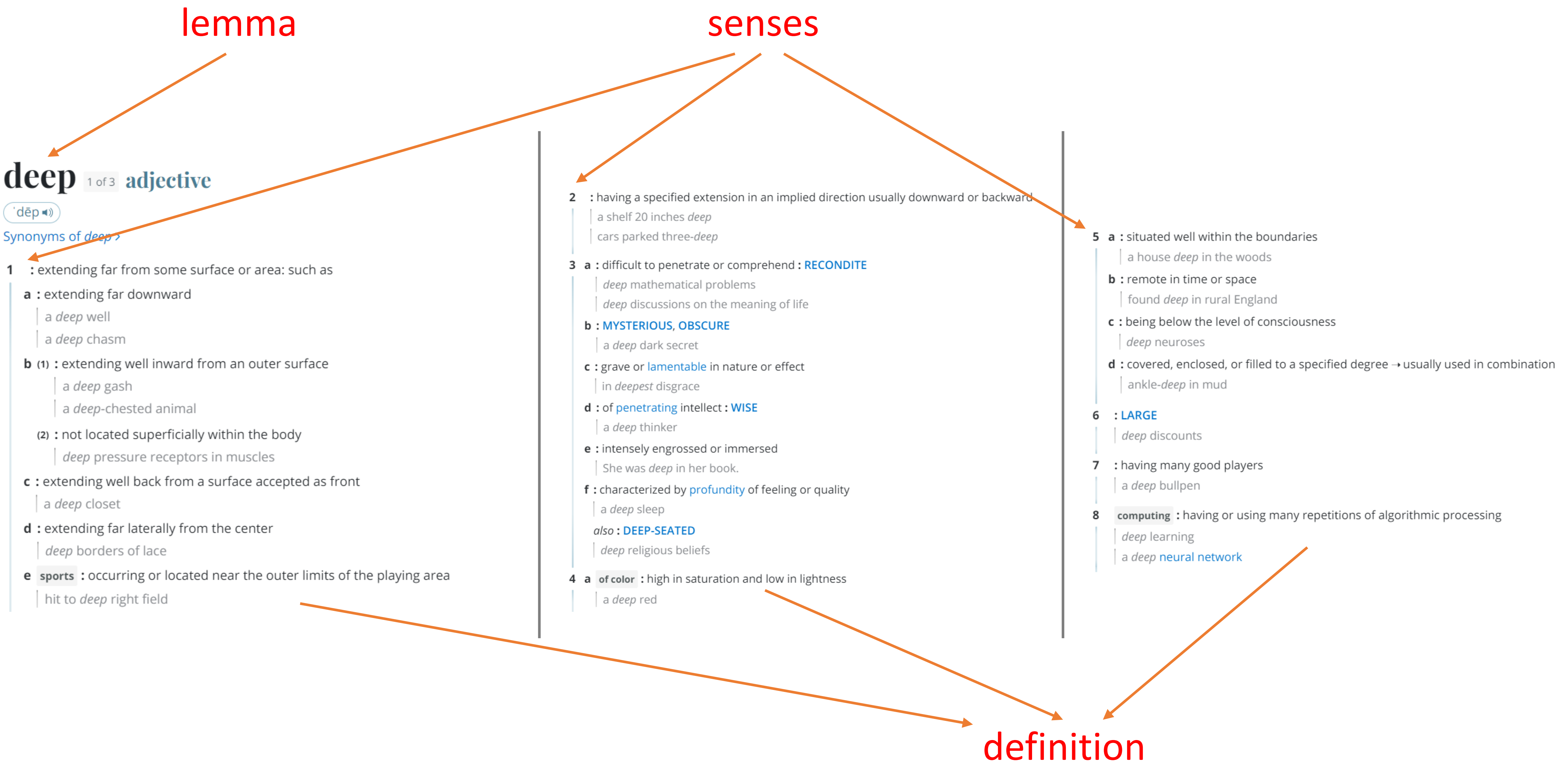
a *deep* bullpen

8 computing : having or using many repetitions of algorithmic processing

deep learning

a *deep* neural network

[Src: Merriam-Webster]



deep 1 of 3 *adjective*

ˈdēp

Synonyms of *deep* ›

1 : extending far from some surface or area: such as

a : extending far downward

a *deep* well

a *deep* chasm

b (1) : extending well inward from an outer surface

a *deep* gash

a *deep*-chested animal

(2) : not located superficially within the body

deep pressure receptors in muscles

c : extending well back from a surface accepted as front

a *deep* closet

d : extending far laterally from the center

deep borders of lace

e sports : occurring or located near the outer limits of the playing area

hit to *deep* right field

2 : having a specified extension in an implied direction usually downward or backward

a shelf 20 inches *deep*

cars parked three-*deep*

3 **a** : difficult to penetrate or comprehend : RECONDITE

deep mathematical problems

deep discussions on the meaning of life

b : MYSTERIOUS, OBSCURE

a *deep* dark secret

c : grave or lamentable in nature or effect

in *deepest* disgrace

d : of penetrating intellect : WISE

a *deep* thinker

e : intensely engrossed or immersed

She was *deep* in her book.

f : characterized by profundity of feeling or quality

a *deep* sleep

also : DEEP-SEATED

deep religious beliefs

4 **a** of color : high in saturation and low in lightness

a *deep* red

5 **a** : situated well within the boundaries

a house *deep* in the woods

b : remote in time or space

found *deep* in rural England

c : being below the level of consciousness

deep neuroses

d : covered, enclosed, or filled to a specified degree → usually used in combination

ankle-*deep* in mud

6 : LARGE

deep discounts

7 : having many good players

a *deep* bullpen

8 computing : having or using many repetitions of algorithmic processing

deep learning

a *deep* neural network

[Src: Merriam-Webster]

Word Sense

- **sense** (or **word sense**): a discrete representation of an aspect of a word's meaning

- one lemma `bank` can have multiple senses:

sense 1: …a **bank**$_1$ can hold the investments in a custodial account

sense 2: …as agriculture burgeons on the east **bank**$_2$ the river will shrink even more

Sec. 19 (J&M)

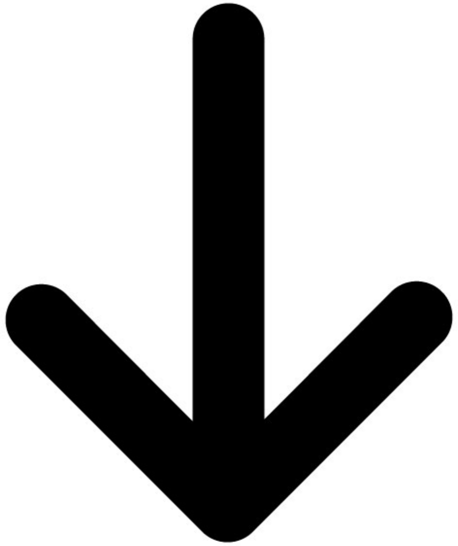
- ways to categorize the patterns of multiple meanings of words:

homonymy: the multiple meanings are unrelated (coincidental?)

down

soft fine feathers

being or moving lower in position or less in some value

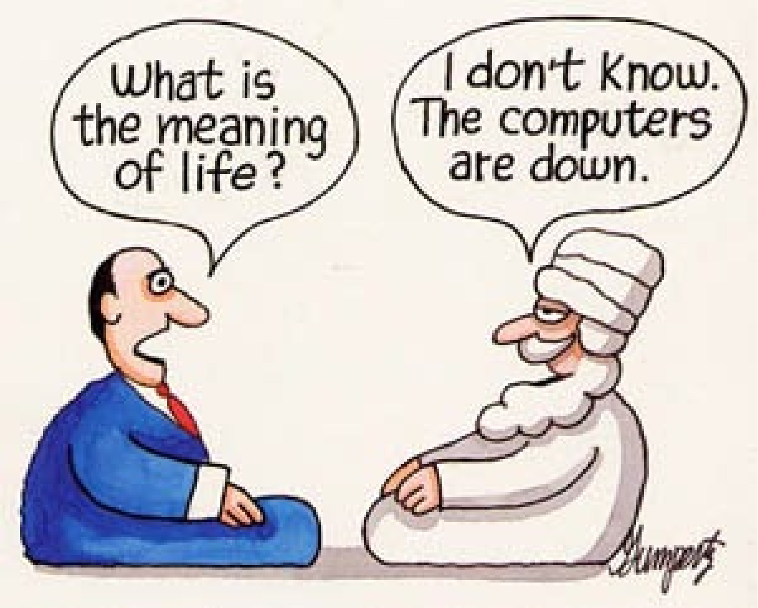


- ways to categorize the patterns of multiple meanings of words:

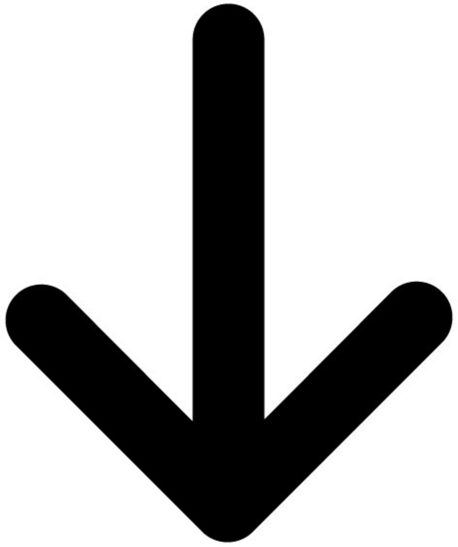
homonymy: the multiple meanings are unrelated (coincidental?)
polysemy: the multiple meanings are related

down

in an inactive or inoperative state



being or moving lower in position or less in some value



Related Senses

1: `The` **bank**$_1$ `was constructed in 1875 out of local red brick.`

2: `I withdrew money from the` **bank**$_2$`.`

- are these the same sense?
  - sense 2: “a financial institution”
  - sense 1: “the building belonging to a financial institution”

- many non-rare words have multiple senses, but sometimes the senses are very similar

Sec. 19 (J&M)

Synonyms

- informally: words with same meaning in some or all contexts
  - `filbert / hazelnut`
  - `couch / sofa`
  - `big / large`
  - `water /` H_2O
- two words are synonyms if they can be substituted for each other in all situations

Sec. 19.2 (J&M)

Synonyms

- few (or no) examples of perfect synonymy
  - even if many aspects of meaning are identical
  - still may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.
- examples:
  - water / H_2O
  - big / large
  - brave / courageous

Sec. 19.2 (J&M)

Synonymy is a relation between **senses** rather than words

- consider the words `big` and `large`

- are they synonyms?

  `How **big** is that plane?`
  `Would I be flying on a **large** or small plane?`

- how about here:

  `Miss Nelson became a kind of **big** sister to Benjamin.`
  `?Miss Nelson became a kind of **large** sister to Benjamin.`

Sec. 19.2 (J&M)

Antonym

- **antonyms**: senses that are opposites with respect to one feature of meaning

```
dark/light    short/long
fast/slow     rise/fall
hot/cold      up/down
in/out
```

- otherwise, they are very similar!

- can be difficult to distinguish synonyms and antonyms with data-driven methods (e.g., distributional word vectors)

Sec. 19.2 (J&M)

Hyponymy and Hypernymy

- sense A is a **hyponym** of sense B if A is more specific, denoting a subclass of B
  - `car` is a hyponym of `vehicle`
  - `mango` is a hyponym of `fruit`
- conversely: **hypernym** (“hyper is super”)
  - `vehicle` is a hypernym of `car`
  - `fruit` is a hypernym of `mango`

Sec. 19.2 (J&M)

Word Sense Disambiguation (WSD)

- given:
  - an ambiguous word in context
  - a fixed inventory of potential word senses

- choose the correct sense based on the context

All-Words WSD

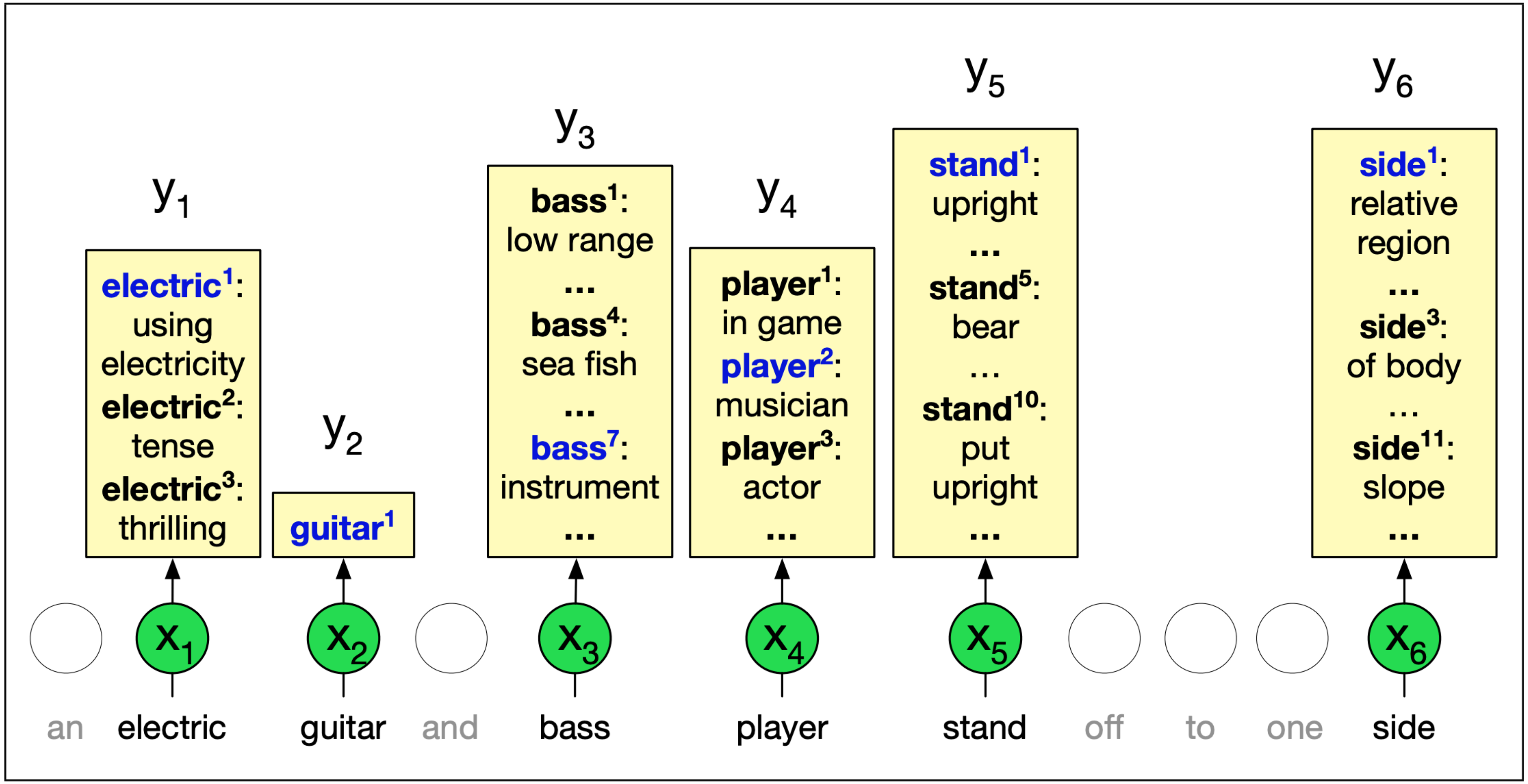


Figure 19.8 The all-words WSD task, mapping from input words (x) to WordNet senses (y). Only nouns, verbs, adjectives, and adverbs are mapped, and note that some words (like *guitar* in the example) only have one sense in WordNet. Figure inspired by Chaplot and Salakhutdinov (2018).

Sec. 19.4.1 (J&M)

How to solve WSD?

Intuition from Warren Weaver (1955):

"If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine… the meaning of the words…

But if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word"

Ch. 19 (J&M)

Example

- using a window of size 3 around the central word ***bass***:

  *An electric guitar and **bass** player stand off to one side not really part of the scene*

Role of WSD?

- many WSD systems have been developed since the 1990s

- researchers hoped that WSD systems would be useful for tasks like machine translation, question answering, sentiment analysis, etc.

- unclear if a separate system is needed

- trend today: end-to-end modeling that disambiguates word sense as part of translation, question answering, etc.

What is a word?

“Oh!” said Lydia stoutly, “I am
not afraid; for though I _am_
the youngest, I’m the tallest.”

(Austen, 1813)

“Oh!”
said
Lydia
stoutly,
“I
am

not
afraid;
for
though
I
_am_

the
youngest,
I’m
the
tallest.”

Word Tokenization

Segmenting running text into “words” (tokens)

“Oh!” said Lydia stoutly, “I am
not afraid; for though I _am_
the youngest, I’m the tallest.”

tokenization: convert sequence of characters into sequence of tokens

tokenizer

“ Oh ! ” said Lydia stoutly , “ I
am not afraid ; for though I _ am _
the youngest , I ’m the tallest . ”

- most tokenizers are rule-based
- several conventions:

	Penn Treebank	Moses
`don’t`	`do n’t`	`don ’t`
`aren’t`	`are n’t`	`aren ’t`
`can’t`	`ca n’t`	`can ’t`
`won’t`	`wo n’t`	`won ’t`

- important to be consistent across NLP systems, match tokenization of external tools/resources
- see `nltk.tokenize` (also for sentence tokenization)

- Chinese, Japanese, Thai: no spaces between words
- Tokenization becomes highly non-trivial!

姚明进入总决赛
"Yao Ming reaches the finals"

- Multiple conventions:

姚明	进入	总决赛	Chinese Treebank
"YaoMing	reaches	finals"	

姚	明	进入	总	决赛	Peking University
"Yao	Ming	reaches	overall	finals"	

Example from Chen et al. (2017), cited in Jurafsky & Martin (SLP3)

- tokenization usually only *adds* whitespace

- might we also want to remove whitespace?

names:

`New York` → `NewYork` ?

non-compositional compounds:

`hot dog` → `hotdog` ?

“ Oh ! ” said Lydia stoutly , “ I am not afraid ; for though I _ am _ the youngest , I ’m the tallest . ”

3	i	1	!	1	oh
2	,	1	.	1	said
2	_	1	;	1	stoutly
2	am	1	afraid	1	tallest
2	the	1	for	1	though
2	“	1	lydia	1	youngest
2	”	1	not	1	’m

3	i	1	!	1	oh
2	,	1	.	1	said
2	_	1	;	1	stoutly
2	am	1	afraid	1	tallest
2	the	1	for	1	though
2	“	1	lydia	1	youngest
2	”	1	not	1	’m

type: a unique word (an entry in a vocabulary or dictionary)

token: an instance of a type in the text

3	i	1	!	1	oh
2	,	1	.	1	said
2	_	1	;	1	stoutly
2	am	1	afraid	1	tallest
2	the	1	for	1	though
2	“	1	lydia	1	youngest
2	”	1	not	1	’m

two types of counts:

type count = 21

token count = 29

useful statistic: **type/token ratio**

(here, 21/29 = 0.724)

How does the type/token ratio change when adding more data?

Corpora

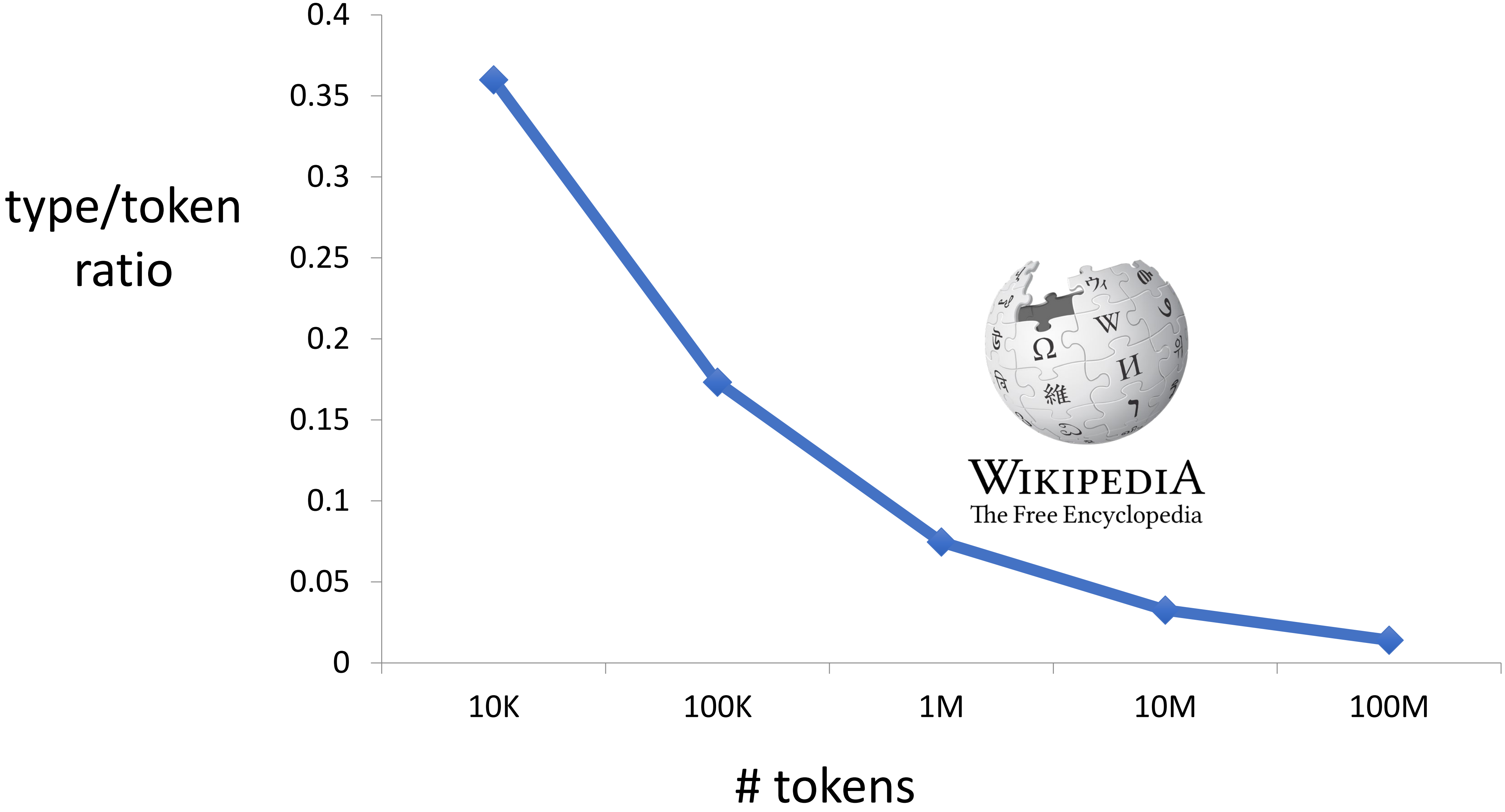
Words don’t appear out of nowhere ---

corpus (plural corpora): a computer-readable collection of text or speech.

A text produced by

- One or more specific writers
- At a specific time
- In a specific place
- Of a specific language
- For a specific function

more data → lower type/token ratio

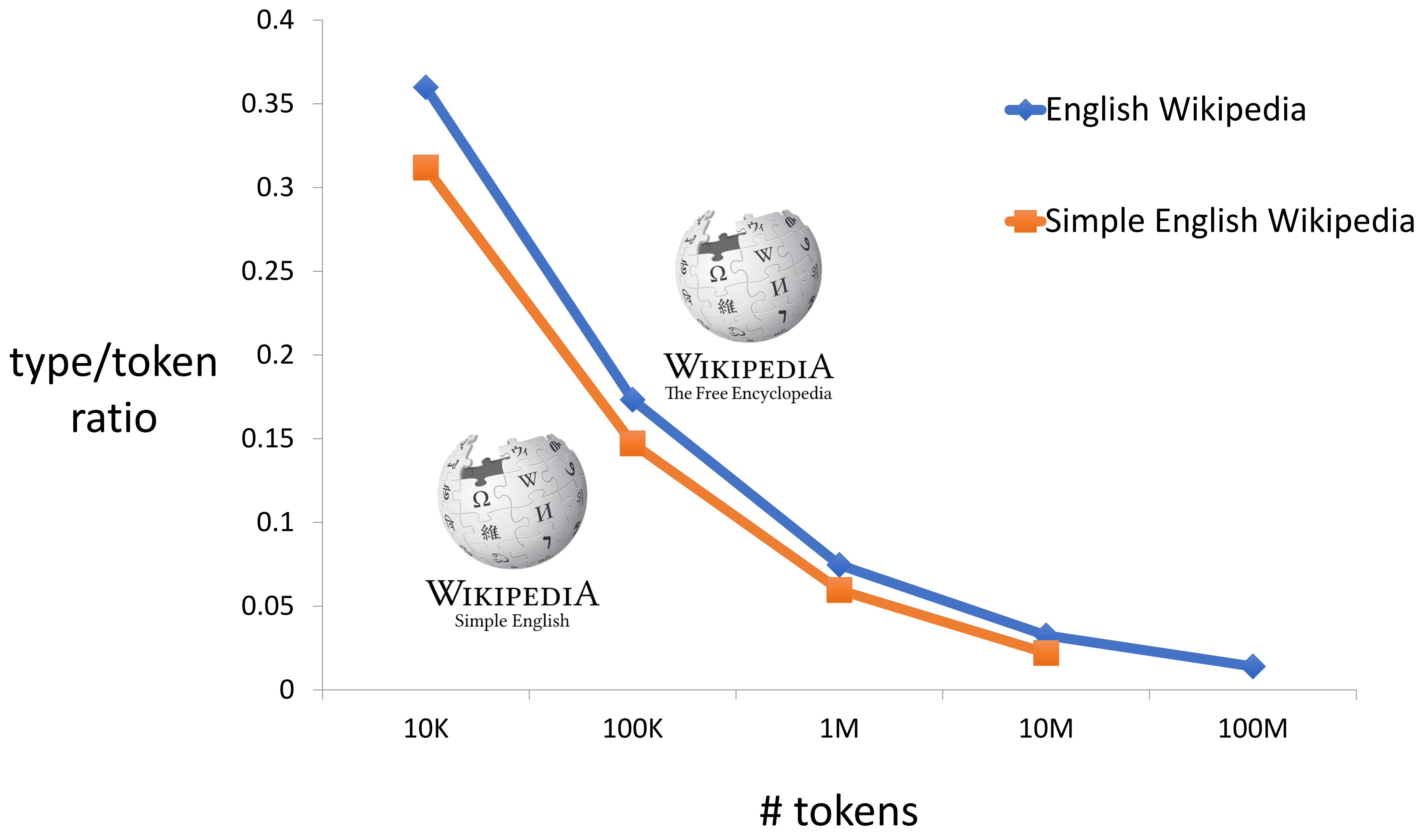


WIKIPEDIA
The Free Encyclopedia

vs.

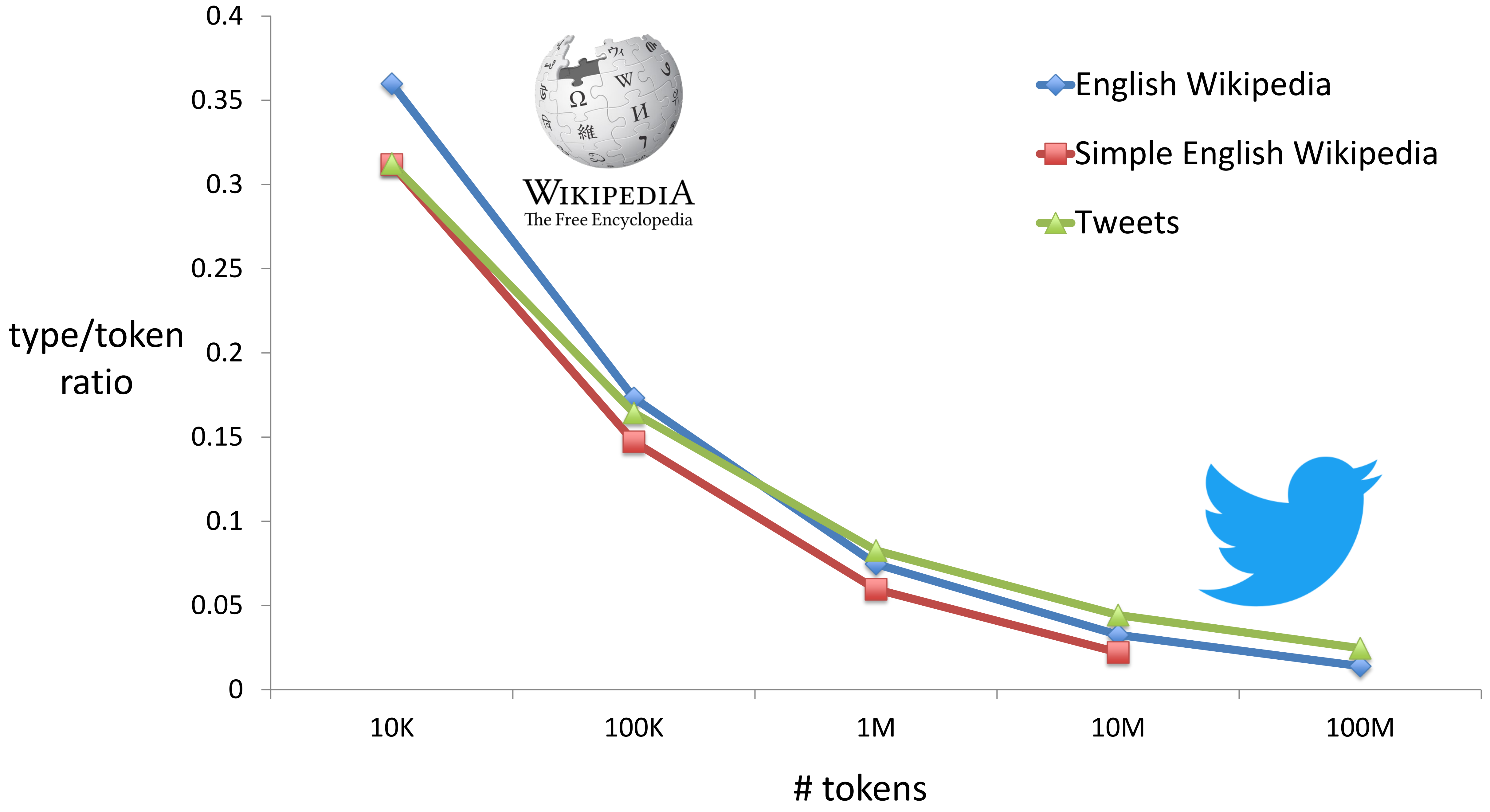
WIKIPEDIA
Simple English

Which has a higher type/token ratio?


type/token ratio
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
10K
100K
1M
10M
100M
tokens
English Wikipedia
Simple English Wikipedia
Wikipedia
The Free Encyclopedia
Wikipedia
Simple English

WIKIPEDIA
The Free Encyclopedia

vs.



type/token ratio
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
10K
100K
1M
10M
100M
tokens
English Wikipedia
Simple English Wikipedia
Tweets
WIKIPEDIA
The Free Encyclopedia

224571 really
1189 rly
1119 realy
731 rlly
590 reallly
234 realllly
216 reallyy
156 relly
146 reallllly
132 rily
104 reallyyy
89 realllllly
89 reeeally
84 reaaally
82 reaally
72 reeeeally
65 reaaaally
57 reallyyyy
53 rilly
50 reallllllly
48 reeeeeally
41 reeally
38 really2
37 reaaaaally
35 reallyyyyy
31 reely
30 realllyyy
27 reaaly
27 realllyy
26 realllyyyy
25 reallllllllly
22 reaaallly
21 really-
19 reeaally
18 reallllyyy
16 reaaaallly
15 reaallly
15 realllllllllly
15 reallllyy
15 reallyreally
15 realyy
14 reallllyyyy
14 reeeeeeally
13 reeeaaally
12 reaaaaaally
12 rreally
11 reaallyy
11 realllllyyy
11 reeeallly
11 reeeeallly
10 reaaaly
10 reallyreallyreally
9 r)eally
9 really-really
9 reallys
9 reeeeeeeally
8 realky
8 reallyyyyyy
8 reallyyyyyyy
8 reeeaally
7 r3ally
7 raelly
7 reaaaaaaally
7 reallllllllllllly
7 realllllllyyy
7 reeeeaaally

7 reeeealy
7 reeeeeeeeeally
7 relaly
6 r-e-a-l-l-y
6 r-really
6 reaaaaaallly
6 realllllllllly
6 realllyyyyy
6 realyl
6 reeeaaaally
6 reeeaaallly
6 reeeaaalllyyy
5 reaaaaallly
5 reaaaallllly
5 reaalllyy
5 realllllllllllllly
5 realllllllllllly
5 reeallyyy
5 reeeeaaallly
5 reeeeaally
5 reeeeeeeeally
5 rellly
5 rrly
5 rrrreally
4 reaaaaly
4 reaaallllly
4 reaaalllyy
4 reaallllly
4 reaalllyyy
4 realllllllyyyy
4 realllllyyyy
4 reeaaaally
4 reeealy
4 reeeeeeeeeeally
4 rllly
3 r34lly
3 r]eally
3 reaaaaaaaally
3 reaaaaaly
3 reaaaallllly
3 reaaaallyy
3 reaaallyy
3 reaallllly
3 reaallyyyy
3 realiy
3 realllllllllllllllly
3 reallllllyy
3 reallllllyyyy
3 reallllllyyyyyyy
3 realllllyyyyy
3 realluy
3 really)
3 reallyl
3 reallyyyyyyyyy
3 reeaaallly
3 reeaallllly
3 reeaalllyyy
3 reeaaly
3 reeallly
3 reealy
3 reeeaaallllyyy
3 reeeaallly
3 reeeeaaaaally
3 reeeeallllly
3 reeeeeallllly
2 reaaaaaaaaaally

2 reaaaaaaaaally
2 reaaaaaaaallly
2 reaaaaaalllllyyy
2 reaaaaalllllly
2 reaaaaallllly
2 reaaalllllyyy
2 reaaallllyyy
2 reaaalllyyy
2 reaallllyy
2 reaallllyyy
2 reaallyyy
2 reaalyy
2 realllllllllllllllllly
2 realllllllllllllllly
2 realllllllyy
2 realllllllyyyyy
2 reallllllyyyyy
2 realllllyy
2 realllllyyyyyy
2 reallllyyyyyy
2 realllyyyyyy
2 really*
2 really/
2 reallyyyyyyyy
2 reallyyyyyyyyyyyy
2 realyyy
2 reaqlly
2 reeaaally
2 reeaallly
2 reeaalllyy
2 reeaallyy
2 reeallyy
2 reeeallyy
2 reeeeaaaalllyyy
2 reeeeaaaally
2 reeeeaaallllly
2 reeeeaaallllyyyy
2 reeeeallllyyy
2 reeeeallyyy
2 reeeeeaaalllllly
2 reeeeeaaally
2 reeeeeaally
2 reeeeeallly
2 reeeeeealy
2 reeely
2 rellys
2 rellyy
2 reqally
2 rlyyy
2 rlyyyy
2 rreeaallyy
2 rrreally
1 r-r-r-really
1 r3aly
1 r3ly
1 raaahhhlllaaayyyy
1 raeally
1 re-e-e-eally
1 re-eaaaaaaaly
1 re-he-he-he-ealy
1 re-he-he-heeeeally
1 re3ally
1 rea(l)ly
1 reaaaaaaaaaaaaaaaaaaally
1 reaaaaaaaaaaaaaaaally
1 reaaaaaaaaaaaaaaalllllly

1 reaaaaaaaaaaaaaally
1 reaaaaaaaaaaaally
1 reaaaaaaaaaaallllly
1 reaaaaaaaaaaallllly
1 reaaaaaaaaaaally
1 reaaaaaaaaaaallllllyyyyyyyyy
1 reaaaaaaaaallly
1 reaaaaaaaallllllyyyyy
1 reaaaaaaaallllly
1 reaaaaaaaallllyyyy
1 reaaaaaallllly
1 reaaaaaalllyyyy
1 reaaaaalllllllyyyy
1 reaaaaalllllyyy
1 reaaaaalllllyyyy
1 reaaaaalllllyyyyy
1 reaaaaalllllyyyyyyy
1 reaaaaallllyy
1 reaaaaallllyyyyy
1 reaaaaallyy
1 reaaaallllllly
1 reaaaallllllyy
1 reaaaalllllyyy
1 reaaaalllllyyyyy
1 reaaaallllyyy
1 reaaaallllyyyyy
1 reaaaalllyy
1 reaaaalllyyy
1 reaaaalllyyyy
1 reaaaallyyy
1 reaaallllllly
1 reaaalllllllyyy
1 reaaallllly
1 reaaallllyyyyy
1 reaaalllllyyyyy
1 reaaalllyyyyy
1 reaaallyyy
1 reaaallyyyy
1 reaallllllly
1 reaalllllyyyy
1 reaallllyyyy
1 reaalllyyyy
1 reaalllyyyyy
1 reaalllyyyyyy
1 real(ly
1 realaaay
1 realkly
1 reall(y
1 reallhy
1 reallllllllllllllllllllllllllllllllll
llllllllllllllllllllllllllllllllllly
1 reallllllllllllllllllllllllllllllllll
lllllllllllllllllllllly
1 reallllllllllllllllllllllllllllllllll
llllllllllly
1 realllllllllllllllllllllllllllllllllly
1 realllllllllllllllllllllllllllly
1 reallllllllllllllllllllllllllly
1 realllllllllllllllllllllllyyyyy
1 reallllllllllllllllllllly
1 realllllllllllllllllllly
1 reallllllllllllyyyyyy
1 realllllllllllyyyyy
1 reallllllllllyyy
1 reallllllllllyyyy
1 reallllllllllyyyyy
1 reallllllllllyyyyyy
1 realllllllllyyyy
1 realllllllllyyyyyyy
1 realllllllllyz
1 realllllyly
1 realllllyyyyy
1 reallllyyyyyyy
1 reallllyyyyyyyyy
1 realllylyyy
1 realllyyyyyyy
1 realllyyyyyyyyy
1 really(really
1 really//
1 really/really/really
1 really100
1 really2x
1 really:')
1 really^^
1 really_
1 reallyii
1 reallyreallyreallyreallyreallyr33lly
1 reallyreallyreallyreallyreallyreally
reallyreallyreallyreallyreallyreally
reallyreallyreallyreallyreally

1 reallyyyyyyyyyyy
1 really😒
1 realoly
1 realys
1 realyyyyy
1 real•ly
1 reawly
1 ree-hee-heally
1 reeaaaaaaaaaaaaaaaaalllllllllyyy
1 reeaaaaaaaallllllllly
1 reeaaaaallllllly
1 reeaaaallly
1 reeaaalllly
1 reeaaallllyyy
1 reeaaallyyy
1 reeaaaly
1 reeaalllllyyy
1 reeaallllyyy
1 reeaallyyy
1 reealllly
1 reealllyyy
1 reealllyyyyy
1 reeeaaaaaaaly
1 reeeaaaaally
1 reeeaaaallllly
1 reeeaaaallllyyy
1 reeeaaaallllyyyy
1 reeeaaaallly
1 reeeaaaalllyyy
1 reeeaaaaly
1 reeeaaalllllyyyy
1 reeeaaallllyyyy
1 reeeaaallyy
1 reeeaaallyyy
1 reeeaaaly
1 reeeaalllly
1 reeeaallyy
1 reeeaallyyy
1 reeeallllly
1 reeealllly
1 reeeallllyy
1 reeeallllyyy
1 reeeallllyyyy
1 reeealllyyy
1 reeealllyyyy
1 reeeallys
1 reeeeaaaaaaallllllyyyyyyy
1 reeeeaaaaaallllllyyyyy
1 reeeeaaaaallllllly
1 reeeeaaaaalllly
1 reeeeaaaaallly
1 reeeeaaaaalllyyy
1 reeeeaaaalllly
1 reeeeaaaallllyyyyyy
1 reeeeaaallllyy
1 reeeeaaalllyy
1 reeeeaaallyyy
1 reeeeaaaly
1 reeeeaallly
1 reeeeaalllyy
1 reeeeaaly
1 reeeeallllly
1 reeeeallllyy
1 reeeealllyy
1 reeeealllyyy
1 reeeealllyyyy
1 reeeeeaaaalllllllly
1 reeeeeaaaaally
1 reeeeeaaalllly
1 reeeeeaaallllyyy
1 reeeeeaaallly
1 reeeeeallllyyy
1 reeeeealy
1 reeeeeeaaaaaally
1 reeeeeeaaaaalllllyyyy
1 reeeeeeaaaaally
1 reeeeeeaaaalllly
1 reeeeeeaally
1 reeeeeeaaly
1 reeeeeeallllly
1 reeeeeealllyyyy
1 reeeeeeeaaaallly
1 reeeeeeeaally
1 reeeeeeeallyy
1 reeeeeeealy
1 reeeeeeeeaaaaaaaalllllyyyyyy
1 reeeeeeeeeaaaaaaaaalllllllllyyyyyyyy
1 reeeeeeeeeaaaaaaallllllllyyyyyyyy
1 reeeeeeeeeaaaaaalllyyy
1 reeeeeeeeeaaally
1 reeeeeeeeeaally
1 reeeeeeeeeallllly
1 reeeeeeeeeeaaally

1 reeeeeeeeeeaally
1 reeeeeeeeeeeally
1 reeeeeeeeeeeeally
1 reeeeeeeeeeeeeaaally
1 reeeeeeeeeeeeeeeaaaallly
1 reeeeeeeeeeeeeeeeally
1 reeeeeeeeeeeeeeeeeeeeeesallllllllllllllllllllllllllly
1 reeeeeeelly
1 reeeeeely
1 reeeeelly
1 reeeeely
1 reeeelllllyyy
1 reeellllyy
1 reellly
1 reelllyy
1 reheheally
1 relally
1 rellllllly
1 relllly
1 rellyrell
1 rellzy
1 rieeely
1 rlllllly
1 rllllllyy
1 rlllly
1 rllyrlly
1 rllyy
1 rllyyy
1 rlyrlyrly
1 rlyy
1 rraarreellyy
1 rreaalllyyy
1 rreaally
1 rreeaaallllyyyy
1 rreeaallllyy
1 rreeaallyyy
1 rreealy
1 rreeeaaaaallllyyyy
1 rreeeaaallllyyy
1 rreeeeeeaaaaallllllllyyyyyy
1 rreeeeeeallly
1 rreeeeeely
1 rrreallyyy
1 rrreeeaaalllyyy
1 rrreeealllyyy
1 rrreeeeaalllllyy
1 rrreeeeallly
1 rrrlyyy
1 rrrreeeally
1 rrrreeeeeeaaaaallllllllyyyyyy
1 rrrrreeeeaaalllly
1 rrrrreeeeaaalllyy
1 rrrrrreally
1 rrrrrrealy
1 rrrrrrreeeeeeaaaalllllyyyyyyy
1 rrrrrrrrrreally
1 rrrrrrrrrrrrrrreeeeeeeeeeeeaaaaaaallllllly yyyyy

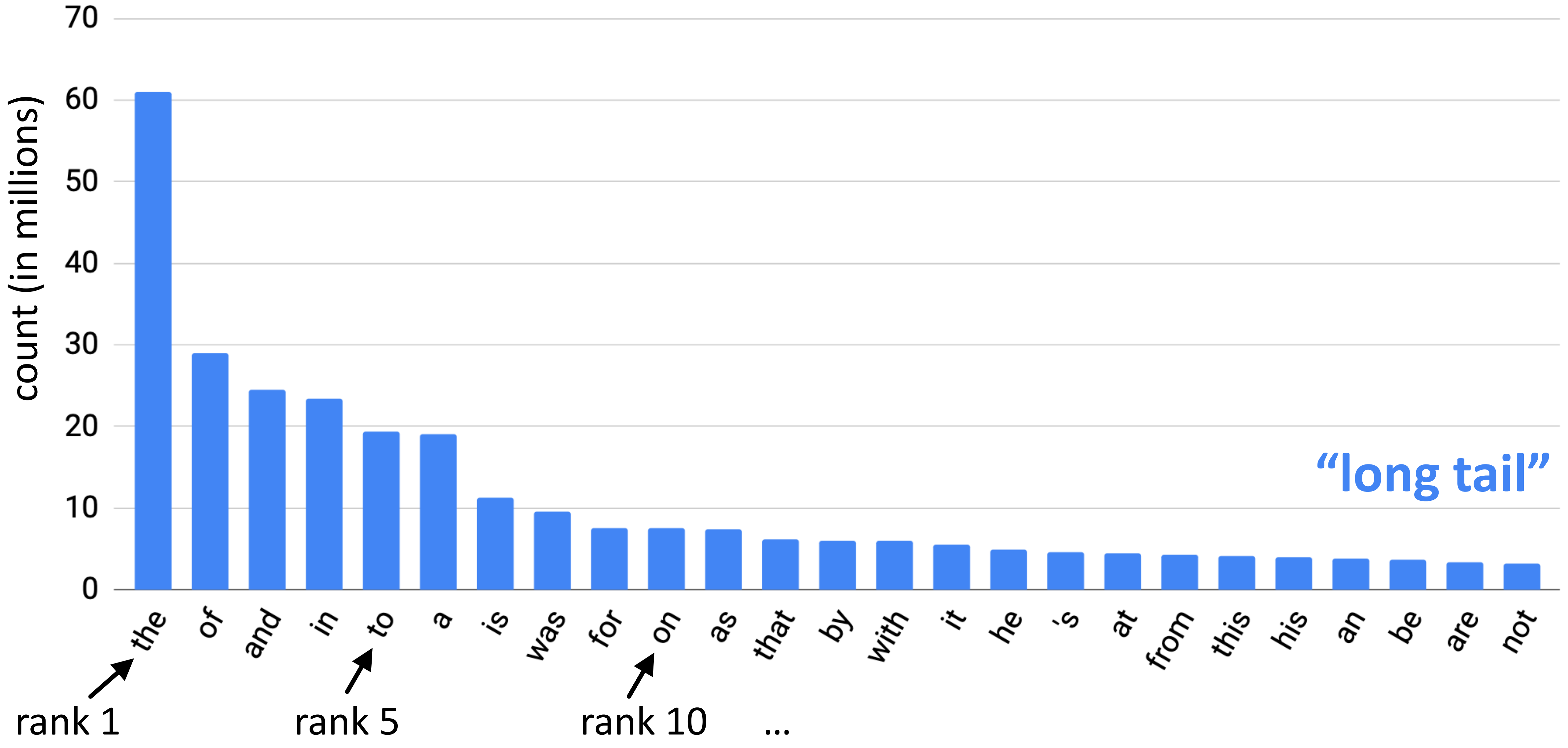
How many words are there?

- size of vocabulary continues to grow as you collect more data
- you’ll never find all the words

How are words distributed?

224571	really	1	rreeeeeeaaaaallllllllyyyyyy
1189	rly	1	rreeeeeeallly
1119	realy	1	rreeeeeely
731	rlly	1	rrreallyyy
590	reallly	1	rrreeeaaalllyyy
234	realllly	1	rrreeealllyyy
216	reallyy	1	rrreeeeaalllllyy
156	relly	1	rrreeeeallly
146	reallllly	1	rrrlyyy
132	rily	1	rrrreeeally
104	reallyyy	1	rrrreeeeeeaaaaallllllllyyyyyy
89	realllllly	1	rrrrreeeeaaallllly
89	reeeally	1	rrrrreeeeaaalllyy
84	reaaally	1	rrrrrreally
82	reaally	1	rrrrrrealy
72	reeeeally	1	rrrrrrreeeeeeaaaalllllyyyyyyy
65	reaaaally	1	rrrrrrrrrreally
…		1	rrrrrrrrrrrrrrrreeeeeeeeeeeaaaaaaalllllllyyyyyy

Zipf's law: frequency of a word is inversely proportional to its rank in the word frequency list



The Long Tail

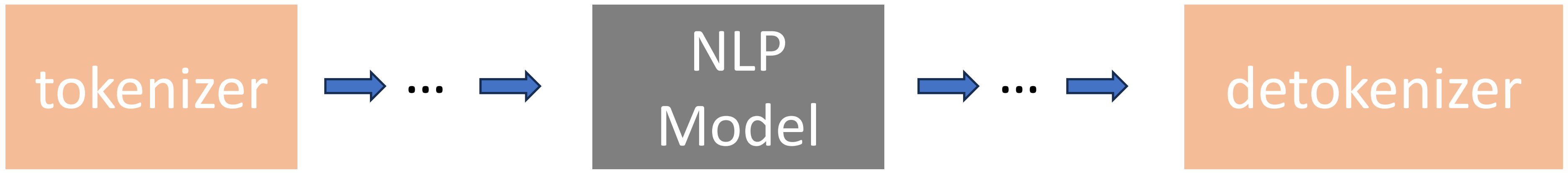
- there are so many word types!
- but words have **internal structures** and **semantic relationships**

```
really
rly
realy
rlly
reallly
realllly
reallyy
relly
reallllly
rily
```

```
play
plays
played
playing
player
players
replay
replays
gameplay
horseplay
```

Tokenization in NLP Systems

Raw text input



Raw text output

a token is the basic unit of text for NLP models

Issues in Tokenization

- White space tokenizer

  a lot of word types (`I`, `I’m`, `I.`)

- Can’t just blindly remove punctuations

  `Ph.D.` `AT&T` numbers (`$45.55`)

- Rule based tokenizers

  Complexity of rules to cover all use cases

Data-Driven Tokenizers

- segment words into pieces (subword units or wordpieces) based on common character sequences in a dataset

- most popular methods:
  - Byte pair encoding (BPE)
  - SentencePiece’s unigram language model (LM)

- these are efficient and effective, but they don’t necessarily correspond to morphology; splits may be arbitrary

- very popular when using neural networks (BERT, GPT, etc)

Sec. 2.4.3 (J&M)

Byte Pair Encoding (BPE)
(Gage, 1994)

- simple data compression technique
- iteratively replaces most frequent pair of bytes in a sequence with a single, unused byte
- Sennrich et al. (2016) adapted BPE for segmenting words

Byte Pair Encoding for Words

- "merge": operation that combines two consecutive units into a single unit
  - initially, units are characters (e.g., `s` or `t`)
  - after merges, units become character sequences (e.g., `st` or `books`)
- greedy algorithm:
  - merge 2 units with the largest 2-unit sequence count, produce merged unit
  - replace all instances of that 2-unit sequence with the merged unit, recompute counts

Sennrich et al. (2016): *Neural Machine Translation of Rare Words with Subword Units*

Example from movie review dataset (Stanford Sentiment Treebank):

word that was not in training set:

`writer/director/producer`

BPE segmenter based on training set

`writ@@ er@@ /@@ direct@@ or@@ /@@ producer`

(recover original text by removing "`@@ `")

```
It wouldn't be my preferred way of spending 100
minutes or $ 7.00 .
```

likely good: “prefer” is the lemma of “preferred”

```
It wouldn't be my prefer@@ red way of sp@@ ending 100
minutes or $ 7@@ .@@ 00 .
```

maybe bad: “spending” is not related to “ending”

Byte Pair Encoding (BPE)

function BYTE-PAIR ENCODING(strings C, number of merges k) **returns** vocab V

$V \leftarrow$ all unique characters in C # initial set of tokens is characters
for $i = 1$ **to** k **do** # merge tokens til k times
  $t_L, t_R \leftarrow$ Most frequent pair of adjacent tokens in C
  $t_{NEW} \leftarrow t_L + t_R$ # make new token by concatenating
  $V \leftarrow V + t_{NEW}$ # update the vocabulary
  Replace each occurrence of t_L, t_R in C with t_{NEW} # and update the corpus
return V

Extension

How does ChatGPT (GPT-2 etc.) tokenize texts from different languages, with a unified tokenizer and fixed vocabulary size?