CMSC 473/673
Natural Language Processing

Instructor: Lara J. Martin (she/they)

TA: Duong Ta (he)

Slides modified from Dr. Frank Ferraro

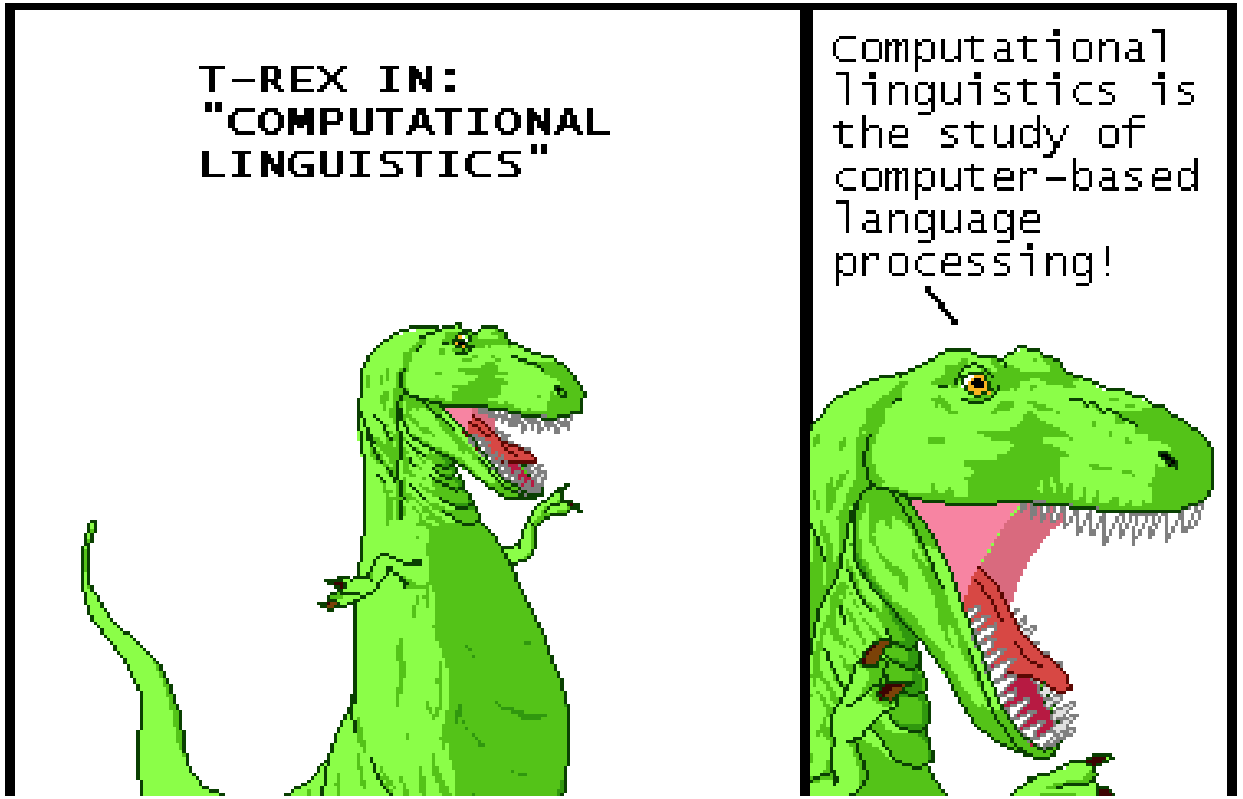
Learning Objectives

Develop a working vocabulary of terms in the field

Recognize sub areas of linguistics

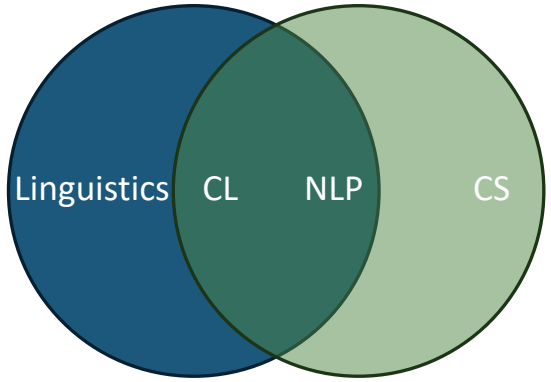
Distinguish between types and tokens

Computational Linguistics



https://qwantz.com/index.php?comic=170

Computational Linguistics
=?
Natural Language Processing



The computational **study** of language

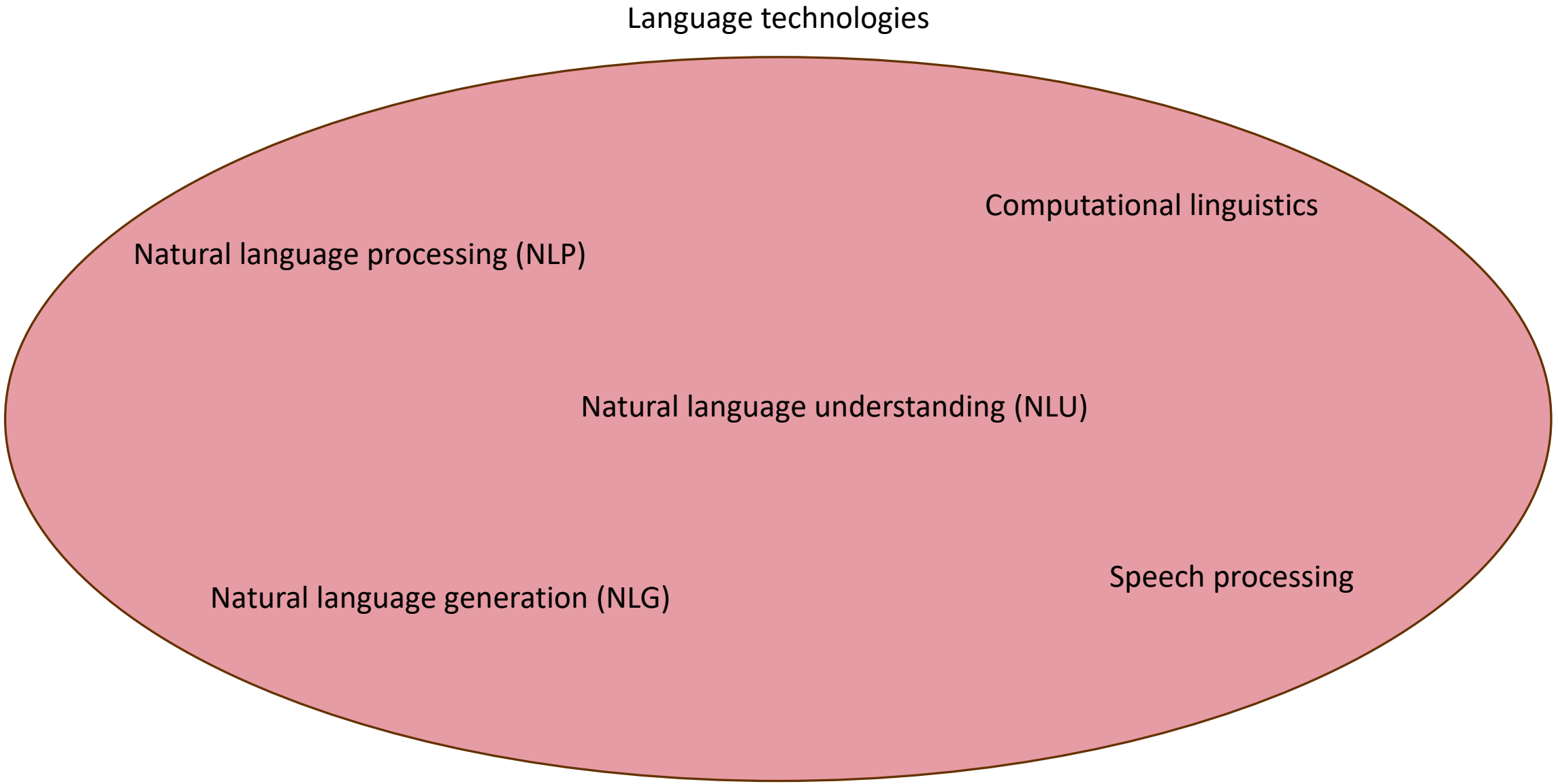
Computational Linguistics

≈

Natural Language Processing

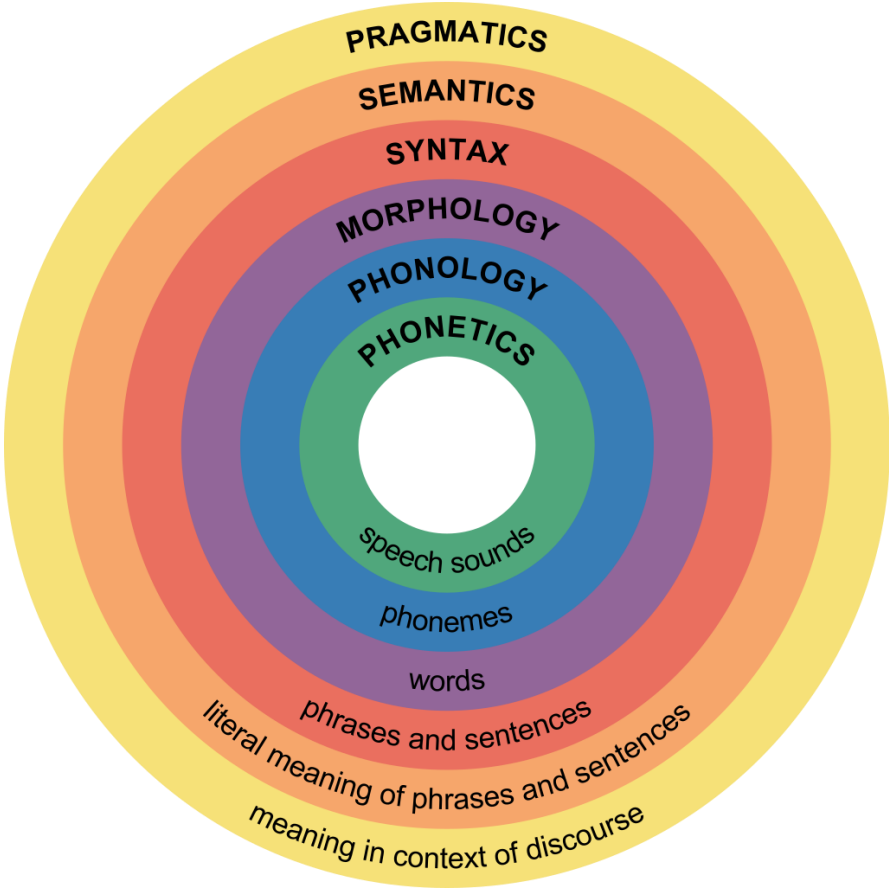
The computational **use** of language

Association for Computational Linguistics


Language technologies
Computational linguistics
Natural language processing (NLP)
Natural language understanding (NLU)
Speech processing
Natural language generation (NLG)

Linguistics

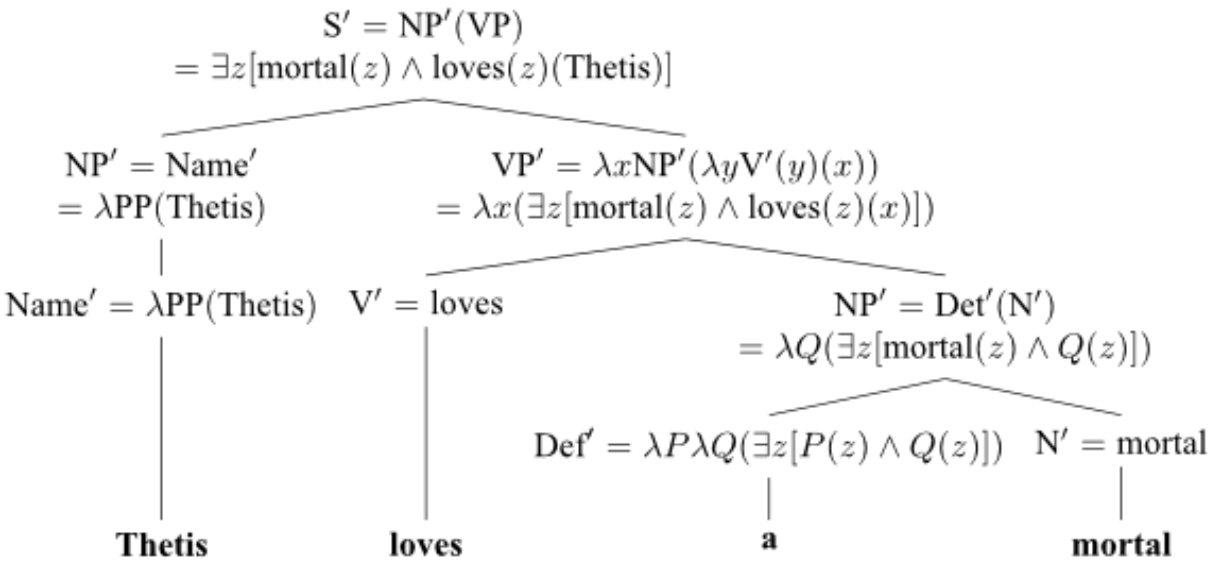
The study of language



https://en.wikipedia.org/wiki/Morphology_(linguistics)#/media/File:Major_levels_of_linguistic_structure.svg

Semantics

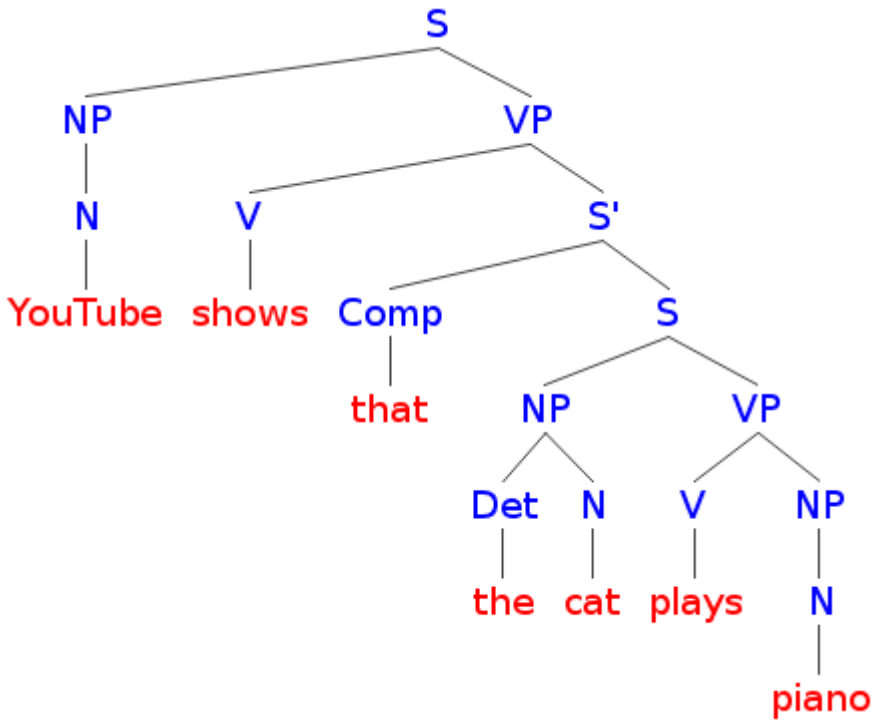
Meaning



https://plato.stanford.edu/entries/computational-linguistics/

Syntax

Grammar



https://allthingslinguistic.com/post/100617668093/how-to-draw-syntax-trees-part-3-type-1-a

Phonology

Processing of sounds

https://upload.wikimedia.org/wikipedia/commons/a/a5/Tsunami_by_hokusai_19th_century.jpg

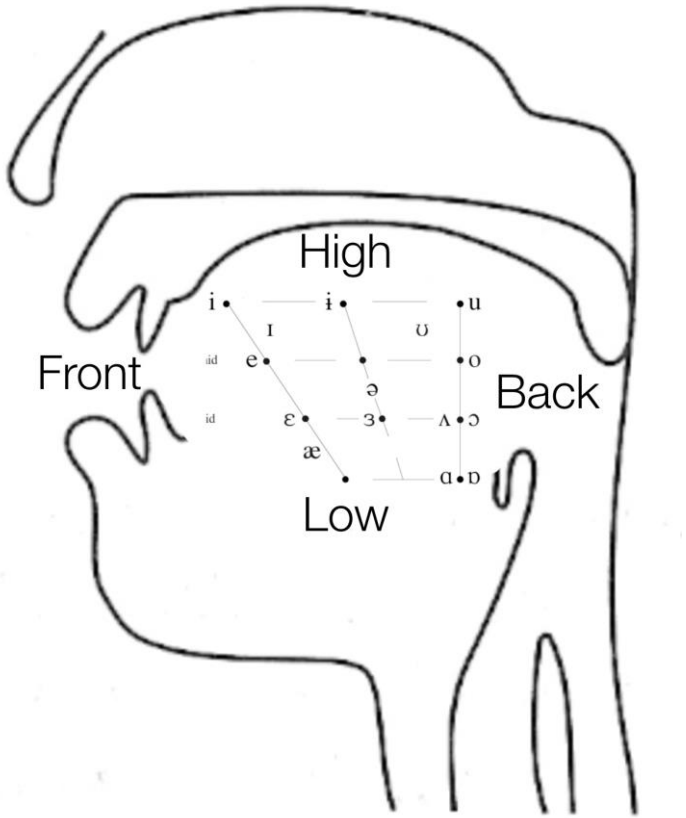
tsunami

↓

sunami

Phonetics

Physical production/understanding of sounds



https://wstyler.ucsd.edu/talks/l111_3_phonetics_review_handout.html

IPA

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2020)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

https://medium.com/@bwsmith.linguist/a-guide-to-the-international-phonetic-alphabet-part-i-d16e127c2ca4

Back to CL vs NLP

Computational linguistics: Using computers to solve linguistic questions

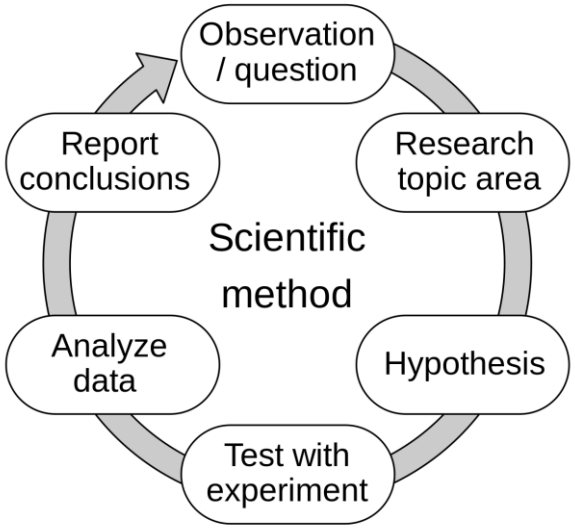
- E.g., How does language X order their sentences? SVO, SOV, VOS…?

And this can inform NLP work

- E.g., How can we create a system that generates text in language X?

Or not…

- E.g., Let's feed a model a bunch of text so that it can generate text in language X.



How do we solve any of these problems?

Data!

https://upload.wikimedia.org/wikipedia/commons/thumb/8/82/The_Scientific_Method.svg/1200px-The_Scientific_Method.svg.png

Where does the data come from?

Corpus (plural: corpora)

- Literally a “body” of text

Languages with few corpora are called “low-resource languages”

- This might not mean the language is endangered!

We can collect corpora in a few different ways:

- Curation: data tagged & organized by experts
- Internet: data “scraped” from open-access sources (Wikipedia, Reddit)
  - Or data collected with permission from closed sources (Facebook, texts) – more rare
- Elicitation: carefully getting participants to produce language (lab studies, crowdsourcing, field studies)
- Pre-existing corpora

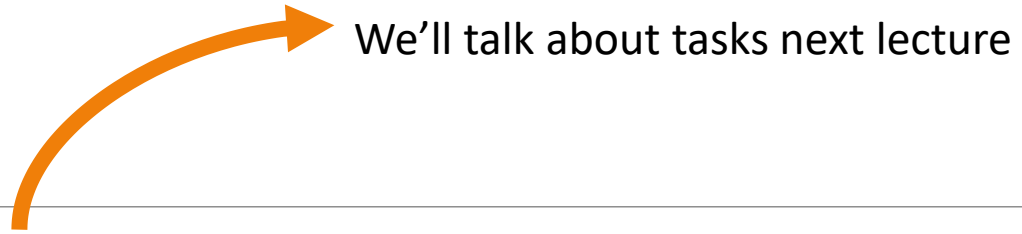
! Facebook has gotten into trouble several times for using data or manipulating people’s feeds without their permission

Benchmarking

Collecting & publishing corpora is helpful for…

- Replication
- Improving performance

Benchmarking



If you want people to work on your problem, make it easy for them to get started and to measure their progress. Provide:

- Test data, for evaluating the final systems
- Development data, for measuring whether a change to the system helps, and for tuning parameters
- An evaluation metric (formula for measuring how well a system does on the dev or test data)
- A program for computing the evaluation metric
- Labeled training data and other data resources
- A prize? – with clear rules on what data can be used

What does the data look like?

Curated data (and some collected data) are usually labeled, especially when made for a particular **task**

- E.g., Universal dependencies (https://universaldependencies.org/)

Current UD Languages

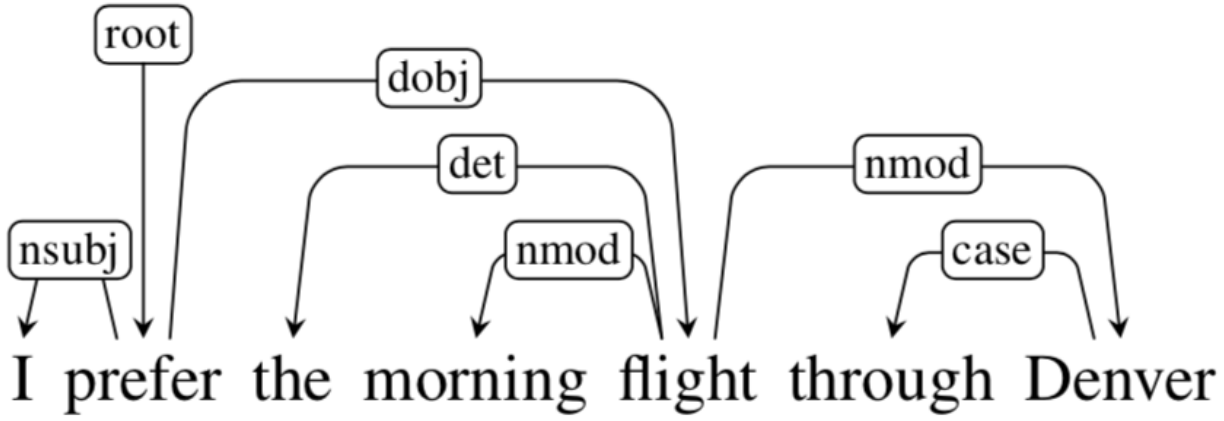
Information about language families (and genera for families with multiple branches) is mostly taken from WALS Online (IE = Indo-European).

Abaza	1	<1K		Northwest Caucasian
Afrikaans	1	49K		IE, Germanic
Akkadian	2	25K		Afro-Asiatic, Semitic
Akuntsu	1	1K		Tupian, Tupari
Albanian	1	<1K		IE, Albanian
Amharic	1	10K		Afro-Asiatic, Semitic
Ancient Greek	3	456K		IE, Greek
Ancient Hebrew	1	39K		Afro-Asiatic, Semitic
Apurina	1	<1K		Arawakan
Arabic	3	1,042K		Afro-Asiatic, Semitic
Armenian	2	94K		IE, Armenian
Assyrian	1	<1K		Afro-Asiatic, Semitic
Bambara	1	13K		Mande
Basque	1	121K		Basque
Beja	1	1K		Afro-Asiatic, Cushitic
Belarusian	1	305K		IE, Slavic
Bengali	1	<1K		IE, Indic
Bhojpuri	1	6K		IE, Indic
Bororo	1	1K		Bororoan
Breton	1	10K		IE, Celtic
Bulgarian	1	156K		IE, Slavic
Buryat	1	10K		Mongolic
Cantonese	1	13K		Sino-Tibetan
Catalan	1	553K		IE, Romance
Cebuano	1	1K		Austronesian, Central Philippine
Chinese	7	309K		Sino-Tibetan
Chukchi	1	6K		Chukotko-Kamchatkan
Classical Armenian	1	13K		IE, Armenian
Classical Chinese	1	433K		Sino-Tibetan
Coptic	1	57K		Afro-Asiatic, Egyptian
Croatian	1	199K		IE, Slavic
Czech	6	2,253K		IE, Slavic
Danish	1	100K		IE, Germanic
Dutch	2	306K		IE, Germanic
English	10	726K		IE, Germanic
Erzya	1	20K		Uralic, Mordvin
Estonian	2	529K		Uralic, Finnic
Faroese	2	50K		IE, Germanic
Finnish	4	397K		Uralic, Finnic
French	7	635K		IE, Romance

What does the data look like?

Curated data (and some collected data) are usually labeled, especially when made for a particular **task**

- E.g., Universal dependencies (https://universaldependencies.org/)



https://medium.com/data-science-in-your-pocket/dependency-parsing-associated-algorithms-in-nlp-96d65dd95d3e

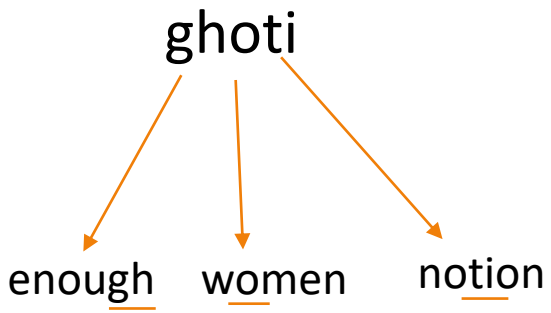
Modalities

Text → TTS isn't straight forward. Unless you have information on how text is pronounced, an orthography (a writing system) by itself can be misleading.

Audio (speech)

Video (closed captioning, sign languages)

Pictures (handwriting recognition, image captioning)



Any of these can be labeled

What’s in a word?

bat

https://www.freepngimg.com/download/bat/9-2-bat-png-hd.png

What’s in a word?

bats

https://www.freepngimg.com/download/bat/9-2-bat-png-hd.png

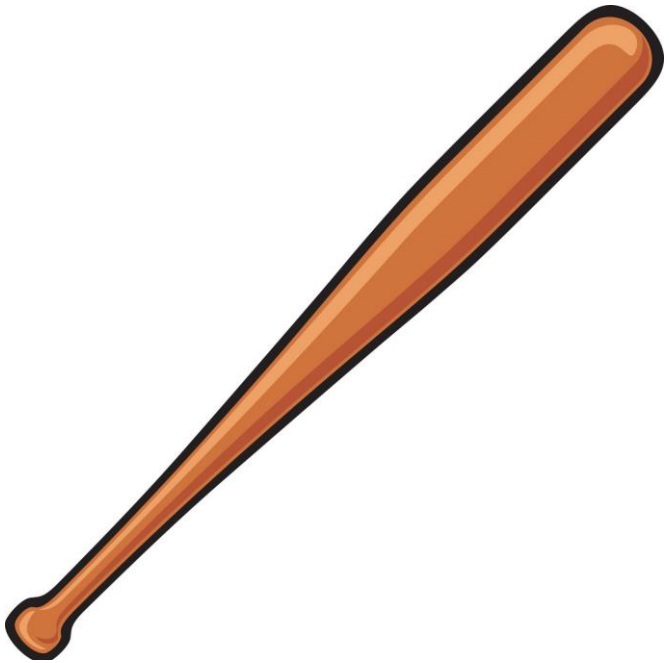
What's in a word?

Fledermaus

flutter mouse

https://www.freepngimg.com/download/bat/9-2-bat-png-hd.png

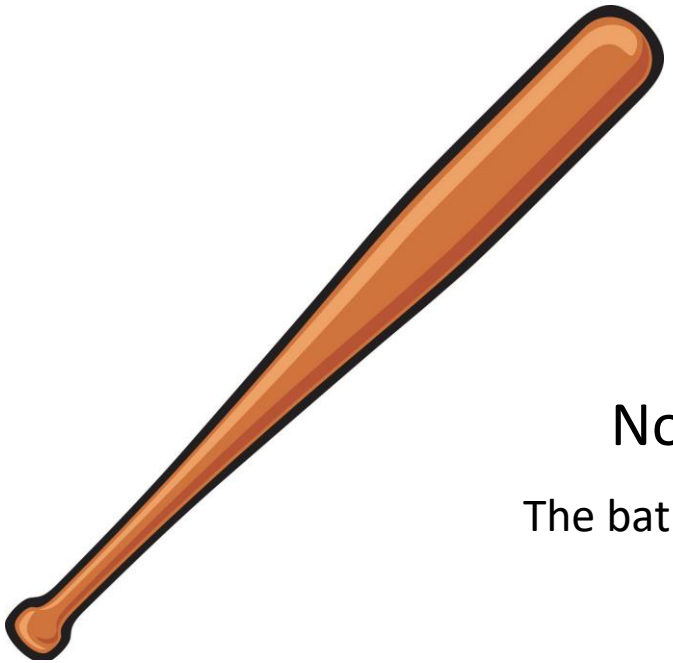
What’s in a word?



bat

https://www.vectorstock.com/royalty-free-vector/baseball-bat-vector-1448799

What’s in a word?



bat

Noun?

The bat was heavy.

Verb?

They bat 1000.

https://www.vectorstock.com/royalty-free-vector/baseball-bat-vector-1448799

What’s in a word?

):

What’s in a word?

my leg is hurting nasty):

What’s in a word?

add two cups (a pint): bring to a boil

Tokens vs Types

The film got a great opening and the film went on to become a hit .

Vocabulary: the words (items) you know

Type: an element of the vocabulary.

Token: an instance of that type in running text.

How many of types & tokens appear in the above sentence?

Tokens vs Types

Types

- The
- film
- got
- a
- great
- opening
- and
- the
- went
- on
- to
- become
- hit
- .

Tokens

- The
- film
- got
- a
- great
- opening
- and
- the
- ~~film~~
- went
- on
- to
- become
- ~~a~~
- hit
- .