What is NLP?

CMSC 473/673 - NATURAL LANGUAGE PROCESSING

Slides modified from Dr. Frank Ferraro

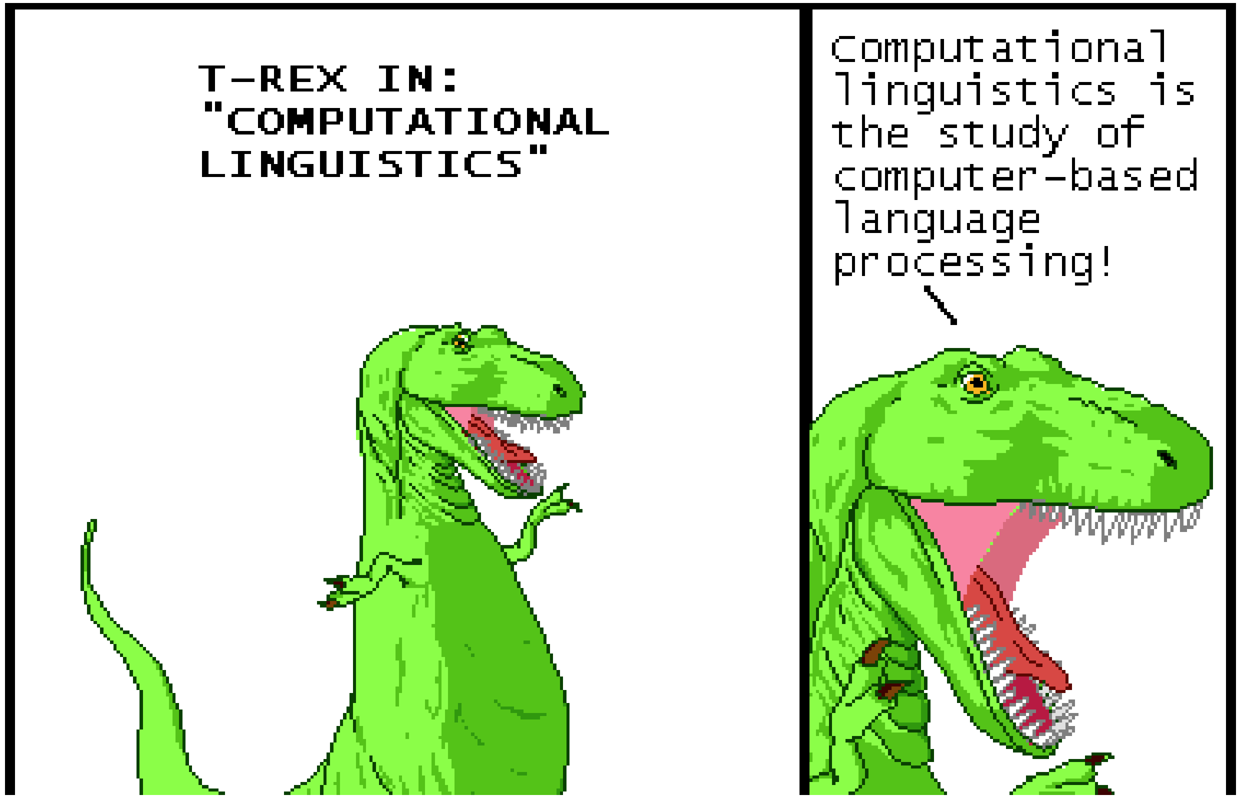
Learning Objectives

Develop a working vocabulary of terms in the field

Recognize sub areas of linguistics

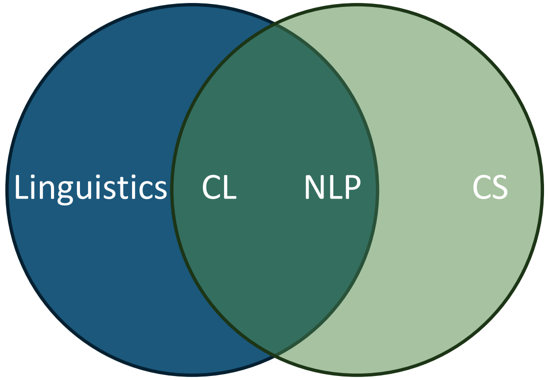
Distinguish between types and tokens

Computational Linguistics


T-REX IN: "COMPUTATIONAL LINGUISTICS"
Computational linguistics is the study of computer-based language processing!

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

Computational Linguistics
=?
Natural Language Processing



The computational **study** of language

Computational Linguistics

≈

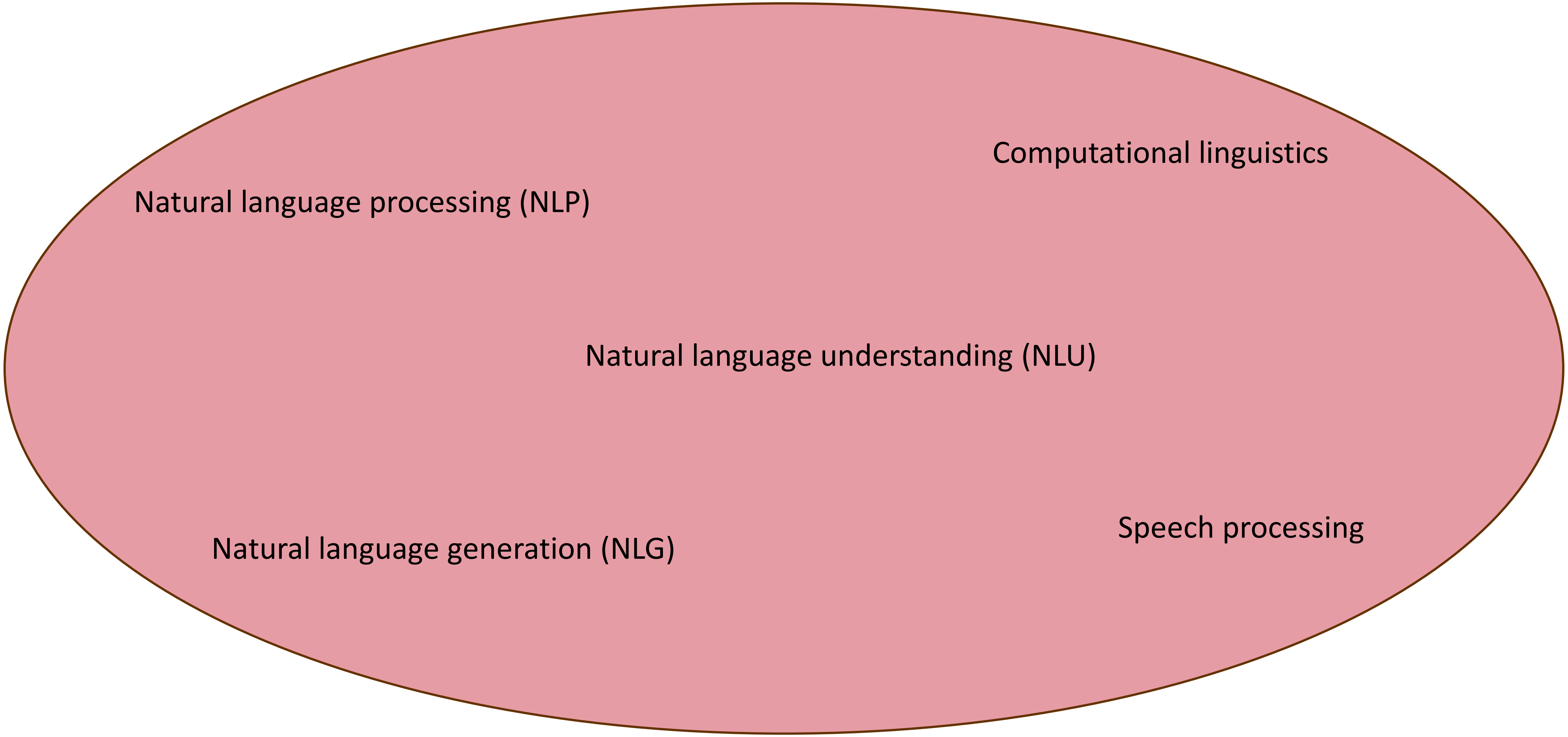
Natural Language Processing

The computational **use** of language

Association for Computational Linguistics

😅

Language technologies



Computational linguistics

Natural language processing (NLP)

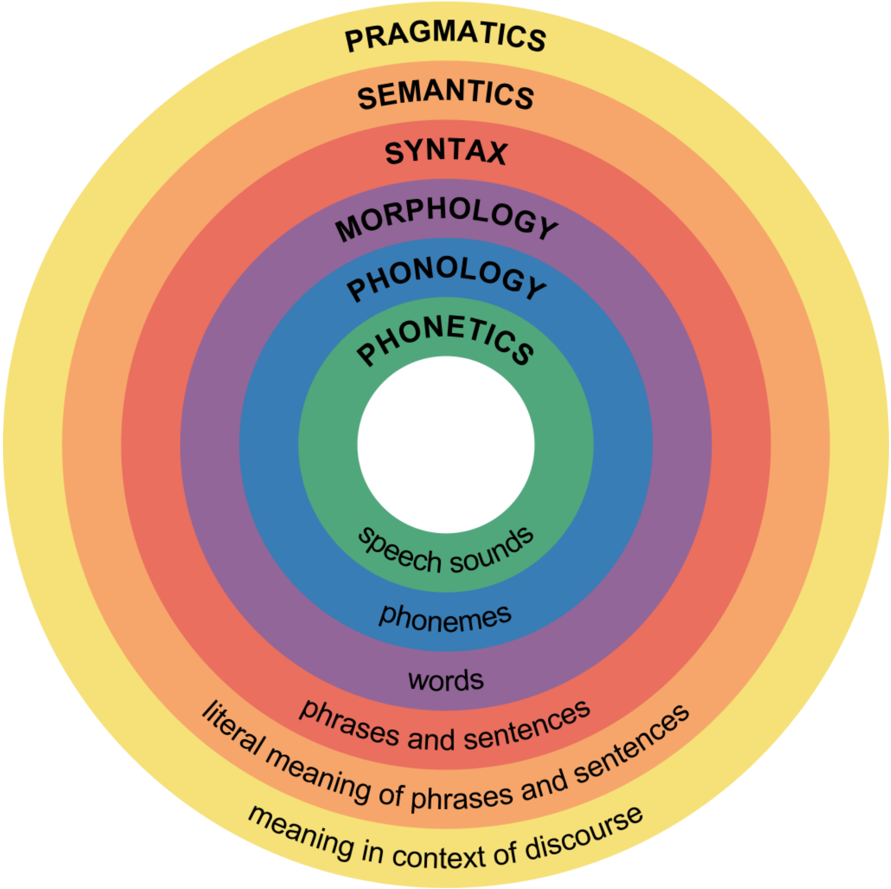
Natural language understanding (NLU)

Speech processing

Natural language generation (NLG)

Linguistics

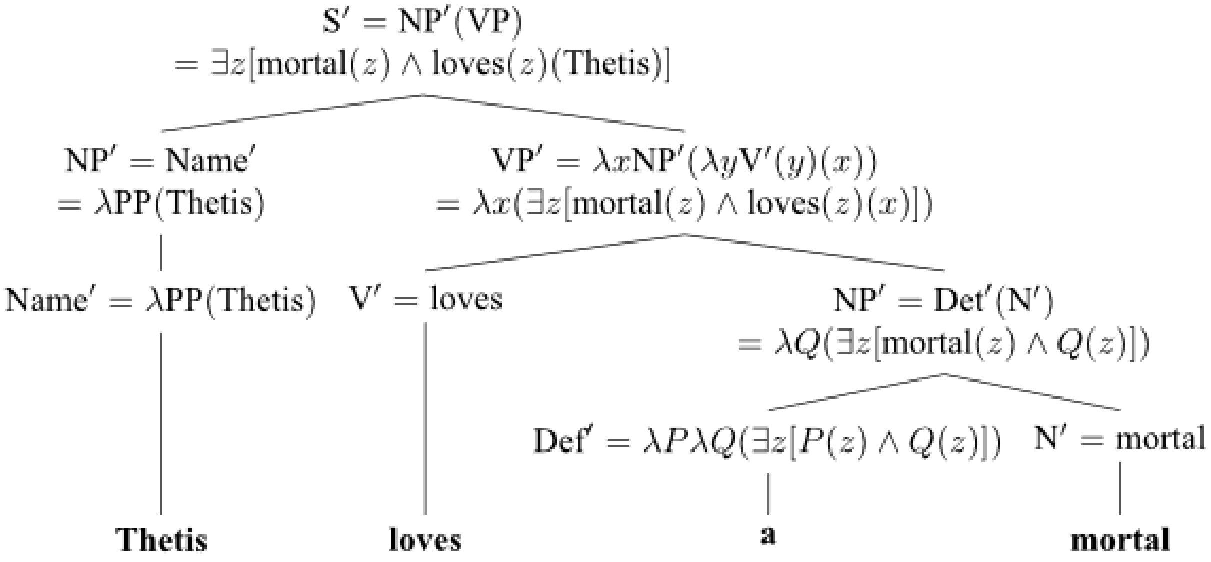
The study of language



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

Semantics

Meaning



$\mathrm{S}' = \mathrm{NP}'(\mathrm{VP})$
$= \exists z[\mathrm{mortal}(z) \wedge \mathrm{loves}(z)(\mathrm{Thetis})]$

$\mathrm{NP}' = \mathrm{Name}'$
$= \lambda \mathrm{PP}(\mathrm{Thetis})$

$\mathrm{VP}' = \lambda x \mathrm{NP}'(\lambda y \mathrm{V}'(y)(x))$
$= \lambda x(\exists z[\mathrm{mortal}(z) \wedge \mathrm{loves}(z)(x)])$

$\mathrm{Name}' = \lambda \mathrm{PP}(\mathrm{Thetis})$

$\mathrm{V}' = \mathrm{loves}$

$\mathrm{NP}' = \mathrm{Det}'(\mathrm{N}')$
$= \lambda Q(\exists z[\mathrm{mortal}(z) \wedge Q(z)])$

$\mathrm{Def}' = \lambda P \lambda Q(\exists z[P(z) \wedge Q(z)])$

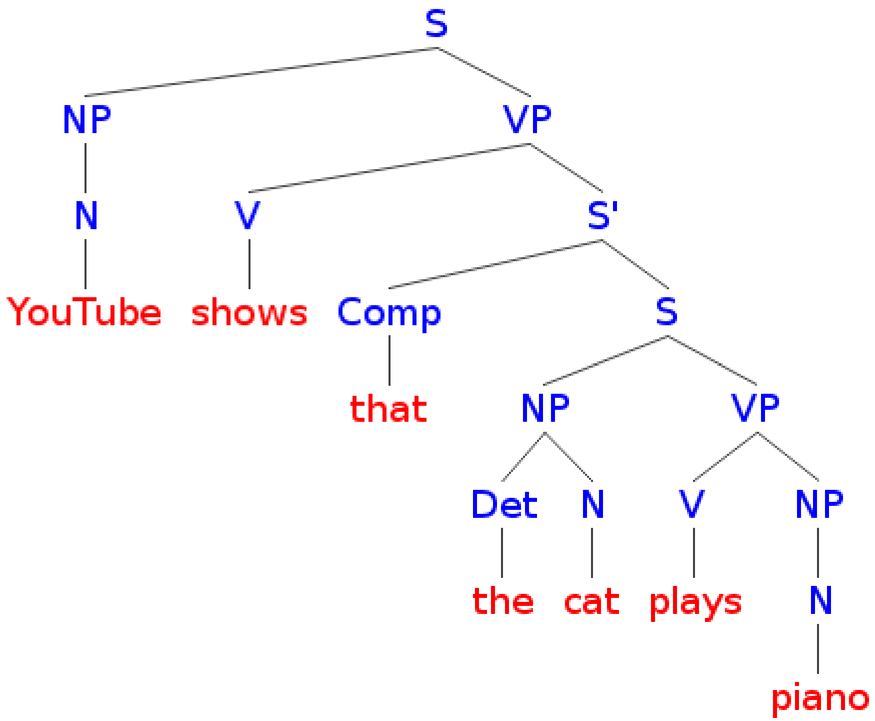
$\mathrm{N}' = \mathrm{mortal}$

Thetis **loves** **a** **mortal**

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

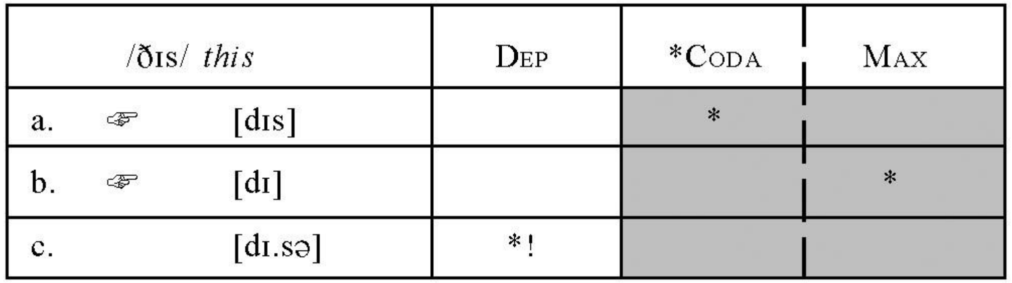
Processing of sounds

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

tsunami

↓

sunami

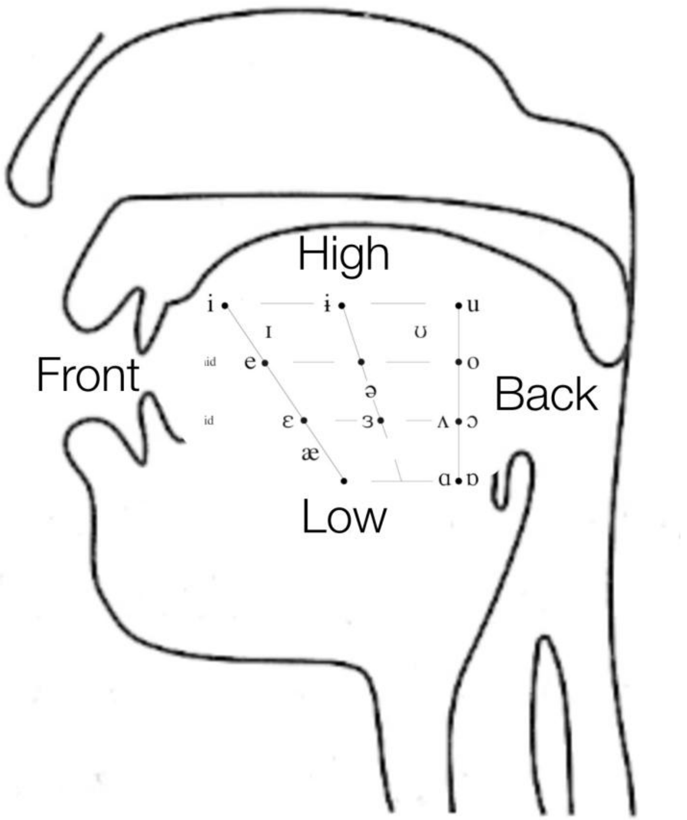


/ðɪs/ *this*	DEP	*CODA	MAX
a. ☞ [dɪs]		*	
b. ☞ [dɪ]			*
c. [dɪ.sə]	*!		

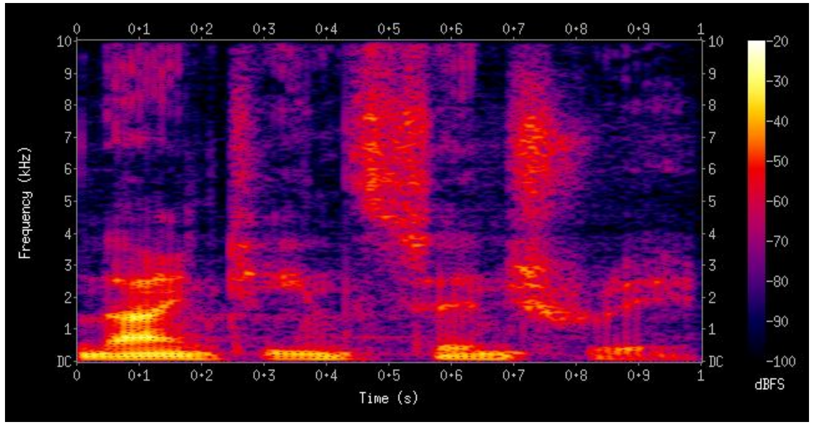
https://pubs.asha.org/doi/10.1044/0161-1461%282001/022%29

Phonetics

Physical production/understanding of sounds



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



https://en.wikipedia.org/wiki/Spectrogram#/media/File:Spectrogram-19thC.png

Back to CL vs NLP

Computational linguistics: Using computers to solve linguistic questions

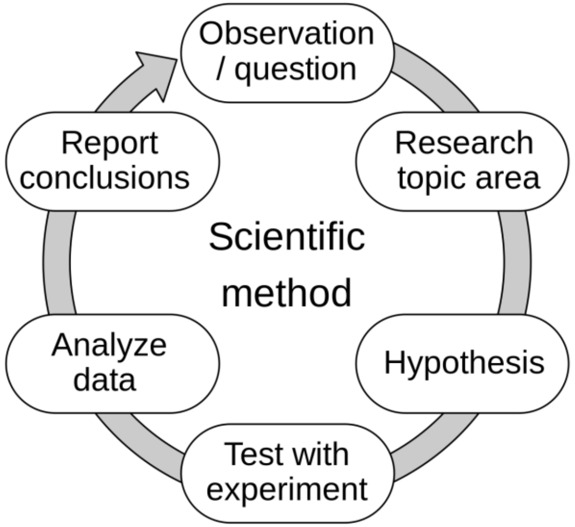
- E.g., How does language X order their sentences? SVO, SOV, VOS…?

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

We'll talk all about tasks next lecture

Benchmarking

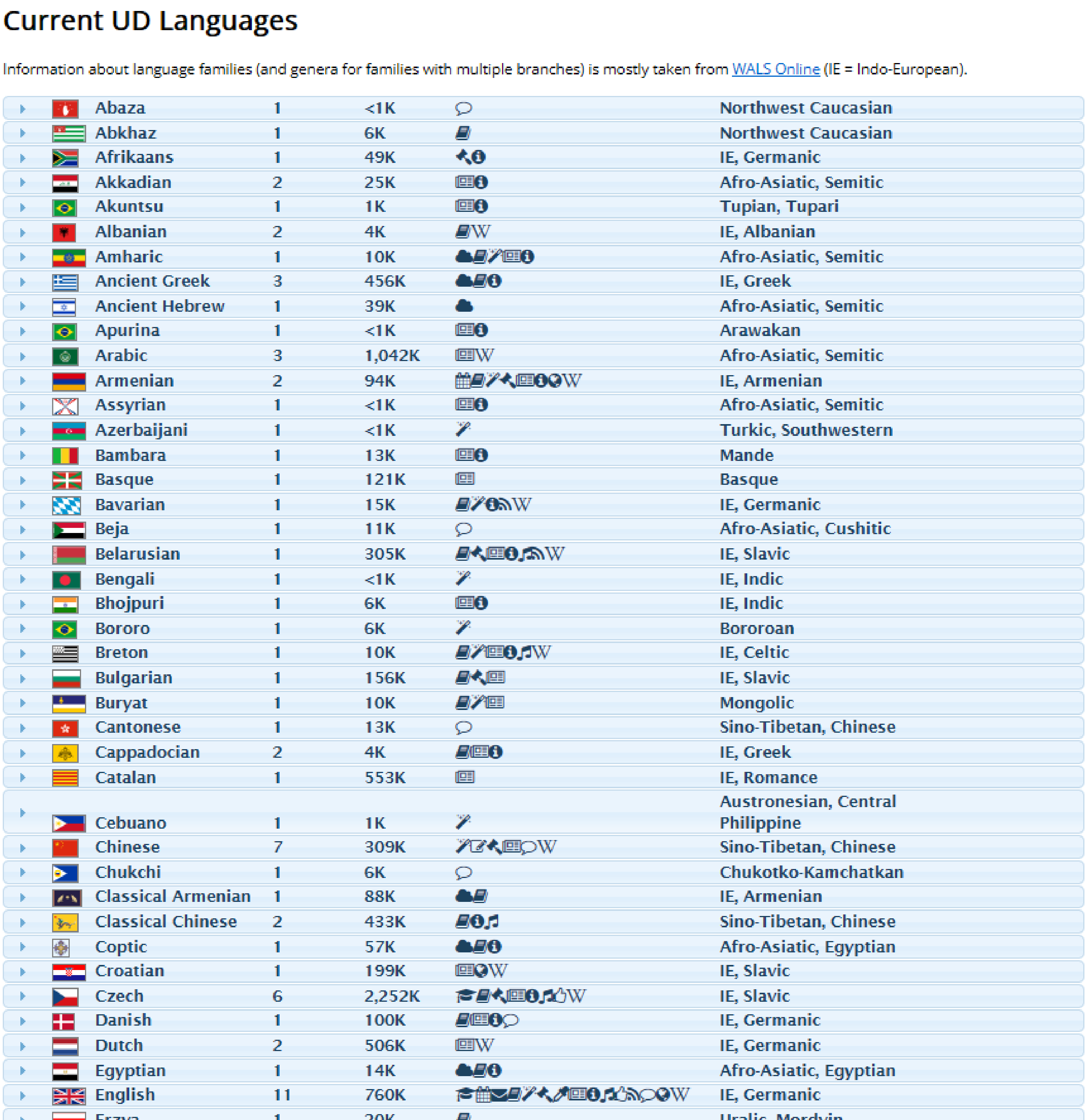
If you want people to work on your problem, make it easy for them to get started and to measure their progress. Provide:

- 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

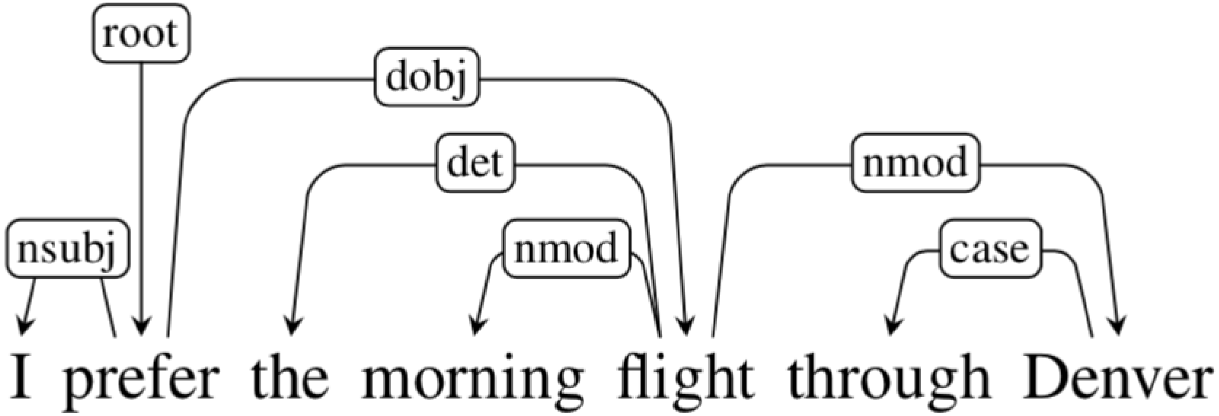
Information about language families (and genera for families with multiple branches) is mostly taken from WALS Online (IE = Indo-European).

Abaza	1	<1K	💬	Northwest Caucasian
Abkhaz	1	6K	📖	Northwest Caucasian
Afrikaans	1	49K	⚖ℹ	IE, Germanic
Akkadian	2	25K	📰ℹ	Afro-Asiatic, Semitic
Akuntsu	1	1K	📰ℹ	Tupian, Tupari
Albanian	2	4K	📖W	IE, Albanian
Amharic	1	10K	☁📖🪄📰ℹ	Afro-Asiatic, Semitic
Ancient Greek	3	456K	☁📖ℹ	IE, Greek
Ancient Hebrew	1	39K	☁	Afro-Asiatic, Semitic
Apurina	1	<1K	📰ℹ	Arawakan
Arabic	3	1,042K	📰W	Afro-Asiatic, Semitic
Armenian	2	94K	📅📖🪄⚖📰ℹ🌐W	IE, Armenian
Assyrian	1	<1K	📰ℹ	Afro-Asiatic, Semitic
Azerbaijani	1	<1K	🪄	Turkic, Southwestern
Bambara	1	13K	📰ℹ	Mande
Basque	1	121K	📰	Basque
Bavarian	1	15K	📖🪄ℹ📶W	IE, Germanic
Beja	1	11K	💬	Afro-Asiatic, Cushitic
Belarusian	1	305K	📖⚖📰ℹ♫📶W	IE, Slavic
Bengali	1	<1K	🪄	IE, Indic
Bhojpuri	1	6K	📰ℹ	IE, Indic
Bororo	1	6K	🪄	Bororoan
Breton	1	10K	📖🪄📰ℹ♫W	IE, Celtic
Bulgarian	1	156K	📖⚖📰	IE, Slavic
Buryat	1	10K	📖🪄📰	Mongolic
Cantonese	1	13K	💬	Sino-Tibetan, Chinese
Cappadocian	2	4K	📖📰ℹ	IE, Greek
Catalan	1	553K	📰	IE, Romance
Cebuano	1	1K	🪄	Austronesian, Central Philippine
Chinese	7	309K	🪄✎⚖📰💬W	Sino-Tibetan, Chinese
Chukchi	1	6K	💬	Chukotko-Kamchatkan
Classical Armenian	1	88K	☁📖	IE, Armenian
Classical Chinese	2	433K	📖ℹ♫	Sino-Tibetan, Chinese
Coptic	1	57K	☁📖ℹ	Afro-Asiatic, Egyptian
Croatian	1	199K	📰🌐W	IE, Slavic
Czech	6	2,252K	🎓📖⚖📰ℹ♫👍W	IE, Slavic
Danish	1	100K	📖📰ℹ💬	IE, Germanic
Dutch	2	506K	📰W	IE, Germanic
Egyptian	1	14K	☁📖ℹ	Afro-Asiatic, Egyptian
English	11	760K	🎓📅✉📖🪄⚖🖉📰ℹ♫👍📶💬🌐W	IE, Germanic
Erzya	1	20K	📖	Uralic, Mordvin

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

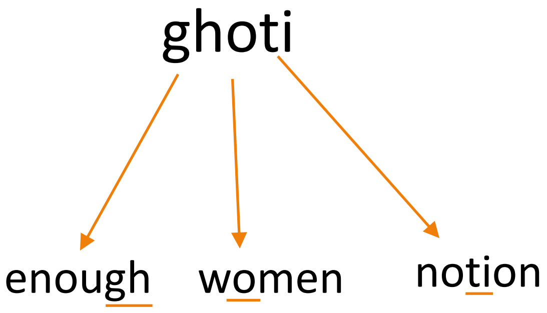
Text

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)



ghoti

enough

women

notion

Any of these can be labeled

What’s in a word?

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

bat

What’s in a word?

bats

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

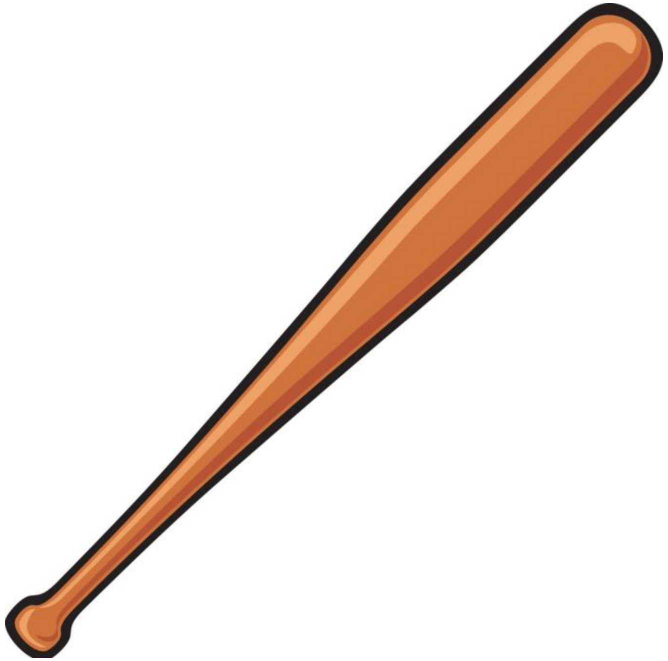
What’s in a word?

Fledermaus

flutter mouse

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

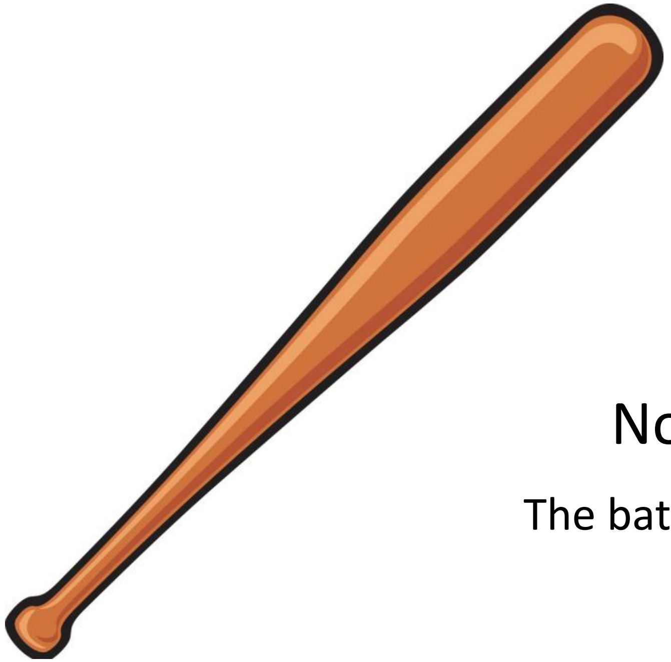
What’s in a word?



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
- .

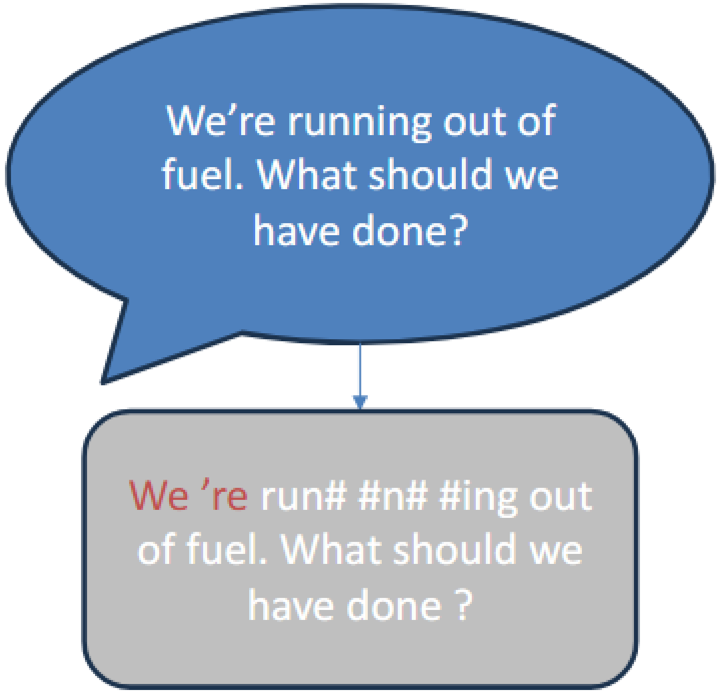
Tokens

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

For your {task} how do you define tokens?

Sometimes:

1. They're defined for you by the *dataset creator*



We're running out of fuel. What should we have done?

We 're run# #n# #ing out of fuel. What should we have done ?

What usually happens when you input a word that your writing/texting program doesn't recognize?

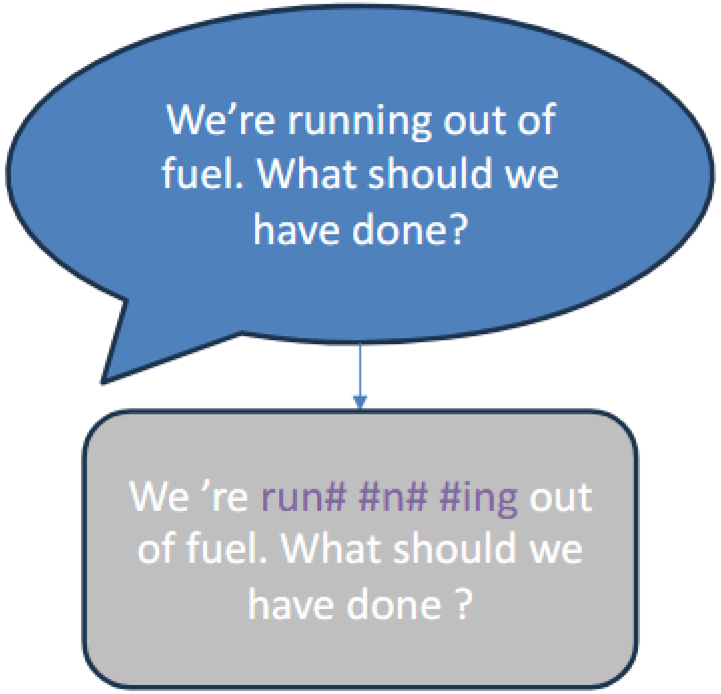
why?

- *scaleably handling novel words*
- *linguistic reasons*
- *historical reasons / technical debt*

For your {task} how do you define tokens?

Sometimes:

1. They’re defined for you by the *dataset creator*
2. They’re defined by the *model*



We’re running out of fuel. What should we have done?

We ’re run# #n# #ing out of fuel. What should we have done ?

(why? scaleably handling novel words)

For your {task} how do you define tokens?

Sometimes:

1. They’re defined for you by the *dataset creator*
2. They’re defined by the *model*
3. It might be part of the *research problem itself*

pişirdiler
They cooked it.

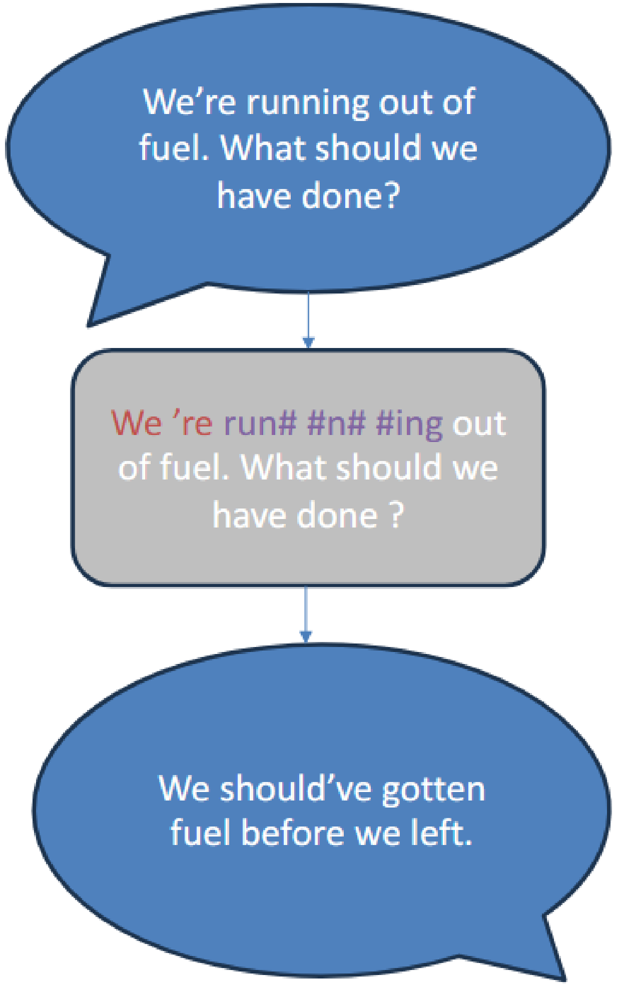
vs.

pişmişlermişlerdi
They had it cooked it.

For your {task} how do you define tokens?

Sometimes:

1. They’re defined for you by the *dataset creator*
2. They’re defined by the *model*
3. It might be part of the *research problem itself*
4. They’re defined by the *end user*
    1. You’ll need to handle points 1 and/or 2 on-the-backend…
    2. and then reversing the process to present output to the user



We’re running out of fuel. What should we have done?

We ’re run# #n# #ing out of fuel. What should we have done ?

We should’ve gotten fuel before we left.