

# Introduction to Machine Learning for Language Processing

LIAD MAGEN



# hə'lou, wɜrl̩d

## THE INTERNATIONAL PHONETIC ALPHABET (revised to 2020)

### CONSONANTS (PULMONIC)

© 2020 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b		t d		t̪ d̪	c j	k g	q ɣ		?	
Nasal	m	nj		n		ɳ	ɲ	ɳ	N		
Trill		B		r					R		
Tap or Flap		v̪		f		l̪					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ɟ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative						ɬ ɺ					
Approximant		v̪		x̪		ɻ ɻ̪	j ɻ̪	w̪			
Lateral approximant				l̪		ɭ ɭ̪	ɻ̪ ɻ̪̪	L			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

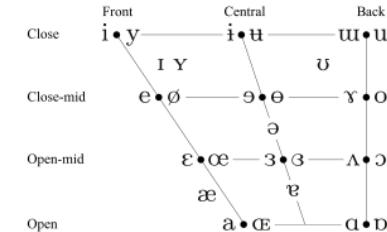
### CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
○ Bilabial	ɓ Bilabial	,
Dental	ɗ Dental/alveolar	Examples:
! (Post)alveolar	ʄ Palatal	p'
‡ Palatoalveolar	ɠ Velar	t'
Alveolar lateral	ʄ' Uvular	k'
		s'

### OTHER SYMBOLS

ʬ Voiceless labial-velar fricative	ʬ Z Alveolo-palatal fricatives
ʬ Voiced labial-velar approximant	ʬ ɬ Voiced alveolar lateral flap
ʬ Voiced labial-palatal approximant	ʬ ɬ Simultaneous ʃ and X
ʬ Voiceless epiglottal fricative	
ʬ Voiced epiglottal fricative	Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.
ʬ Epiglottal plosive	

### VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

### SUPRASEGMENTALS

- ! Primary stress      ,founeɪ tʃən
- ! Secondary stress
- Long      e•
- Half-long      e•
- Extra-short      ē
- | Minor (foot) group
- || Major (intonation) group
- Syllable break      .i.a.ekt
- Linking (absence of a break)

### DIACRITICS

o Voiceless	ɳ ɳ̪	.. Breathy voiced	b ɳ	Dental	t ɳ
~ Voiced	ʂ ʂ̪	~ Creaky voiced	b ɳ	Apical	t ɳ
h Aspirated	tʰ dʰ	~ Linguolabial	t ɳ	Laminar	t ɳ
› More rounded	ɔ	ʷ Labialized	tʷ dʷ	~ Nasalized	ɛ
⟨ Less rounded	ɔ̪	{j Palatalized	t{j d{j	ᵑ Nasal release	d^n
+ Advanced	u	˥ Velarized	t˥ d˥	˥ Lateral release	d˥
- Retracted	e	˥ Pharyngealized	t˥ d˥	˥ No audible release	d˥
.. Centralized	ĕ	~ Velarized or pharyngealized	˥		
× Mid-centralized	ĕ	Raised	e	(.I = voiced alveolar fricative)	

LEVEL	CONTOUR
é or ĕ	Extra high
é	High
ē	Mid
é	Low

# Organizational Info



## Missing classes

Allowed to miss one class



## Late

Up to 15 minutes



## Grading

Homework assignments (45%)  
Final presentation (55%)



## Late assignments submission

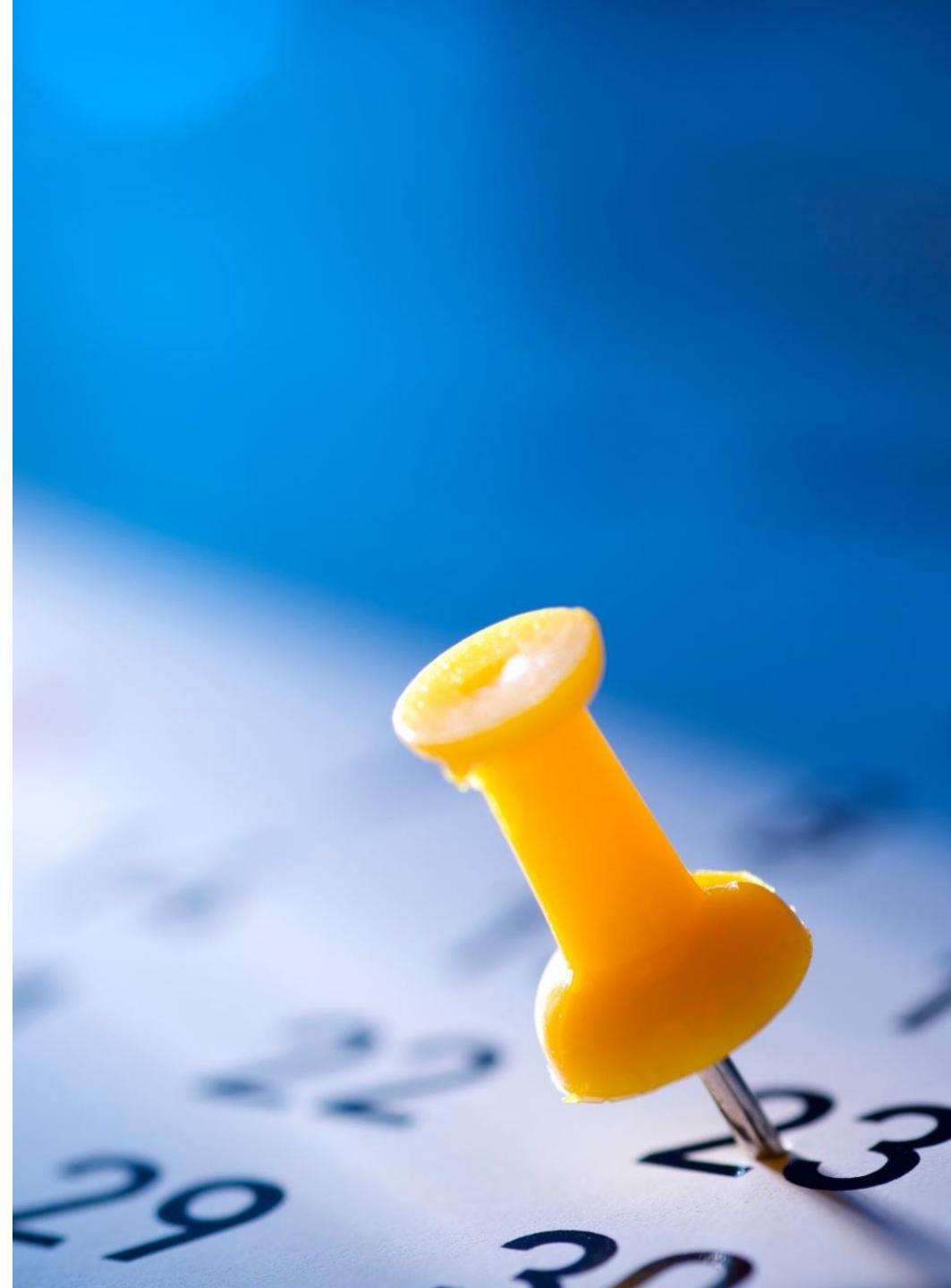
5 days overall late  
10 points penalty per day afterward

# Agenda Today

- > NLP – Motivation, Applications & Challenges
- > 5 Levels of Linguistics
- > Basic Units of Text Processing

If we have time:

- > ML Models Definition & Types
- > First Classification Model



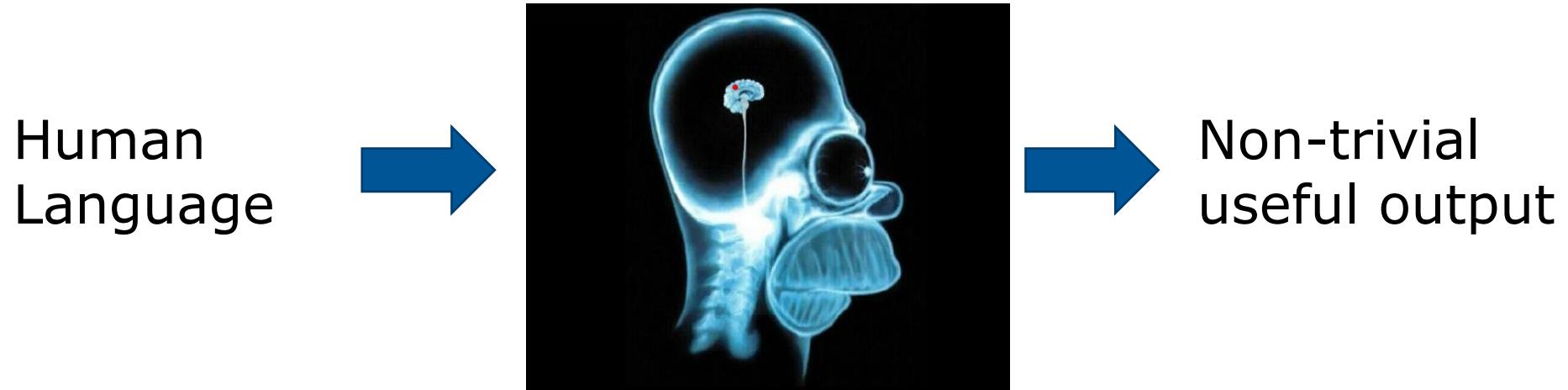
# True or False?

Read the statements and decide if they are **correct** or **not**.

# Motivation

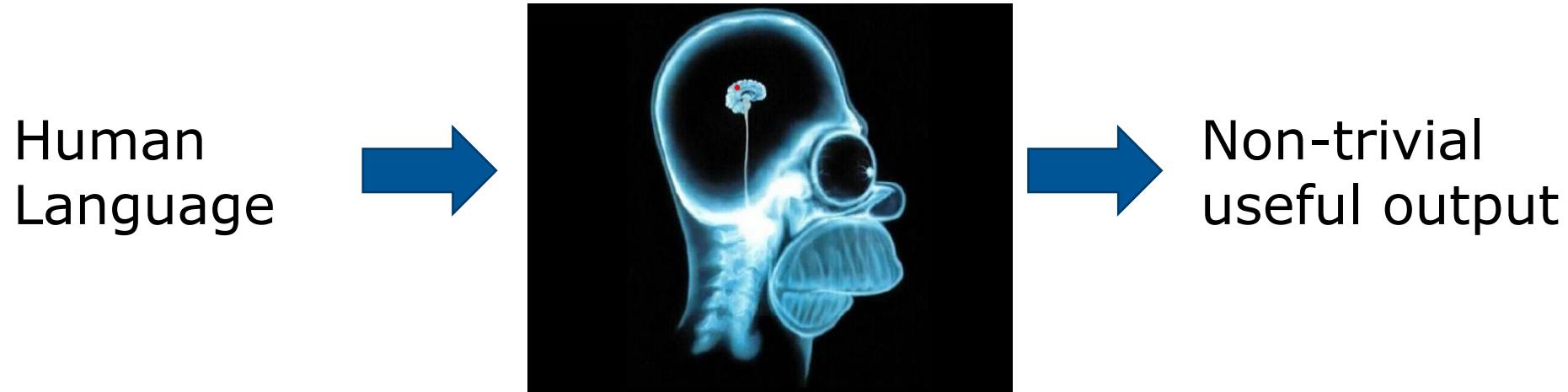
Applications, Challenges, 5 levels of Linguistics

# What is NLP?



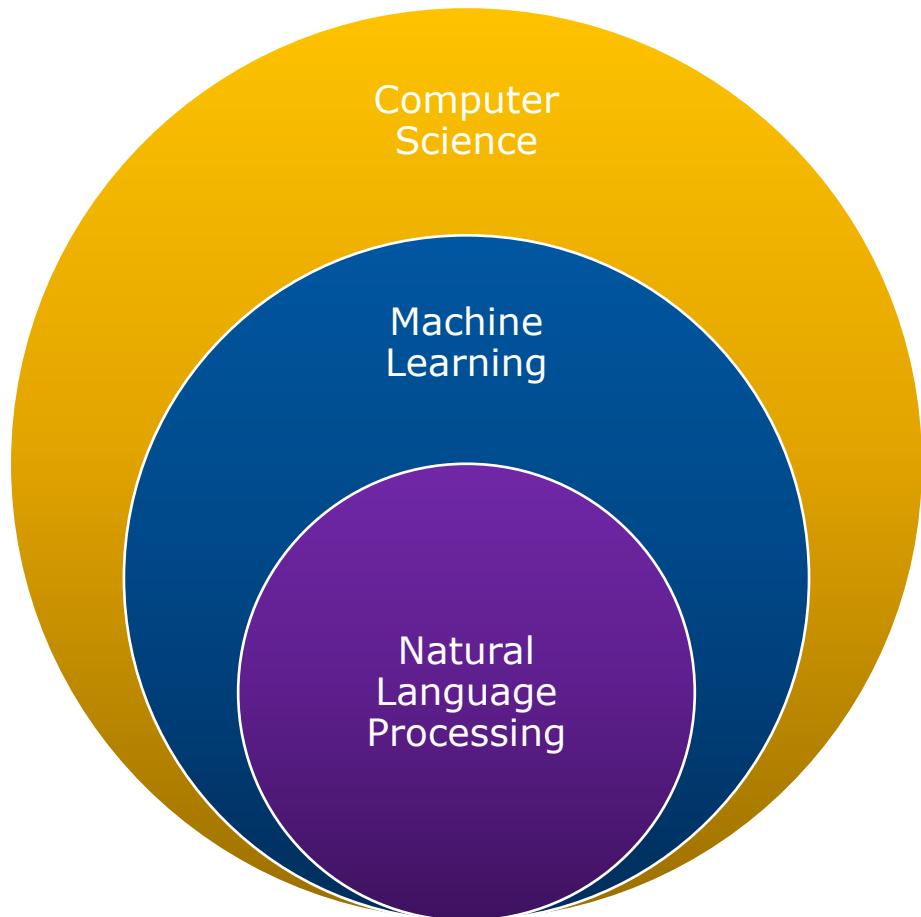
> **Algorithms** that operate on **Human Language**

# What is NLP?

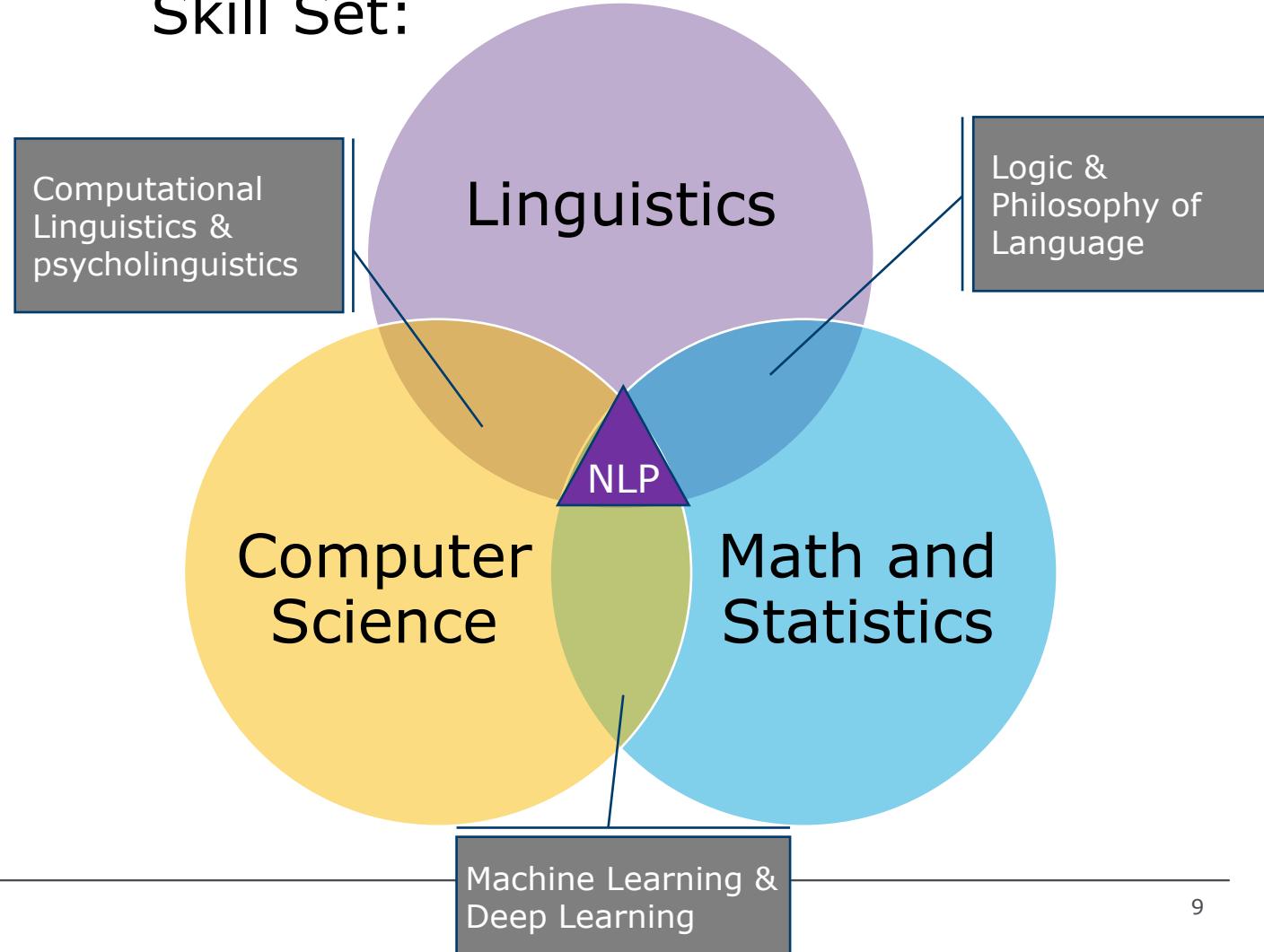


- > Process input text in human language, in a way that suggests an intelligence was involved

# What is NLP?

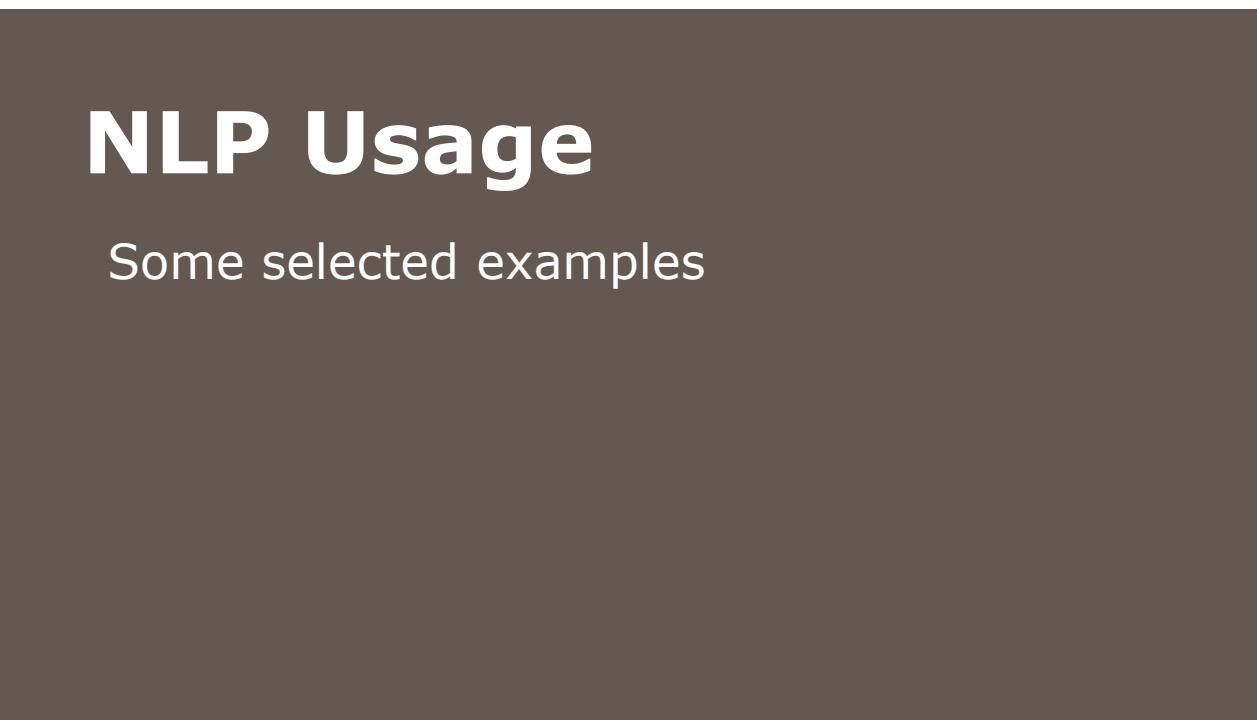
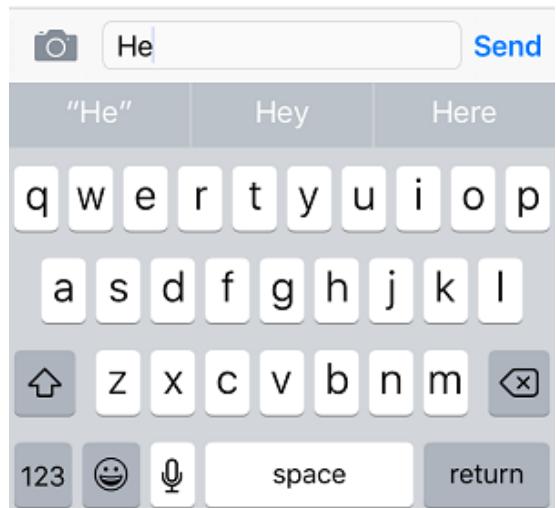


Done With an Interdisciplinary Skill Set:





# Where is NLP being Used?



# Search Engines

- Exact words
- Synonyms

Google assisting elderly

aginginplace.org > Caregiving ▾  
**Helping The Elderly At Home | Updated for 2020 ...**  
Tips for Aging Well – At Home · Eating · Functional Mobility (moving from one place to another while performing activities such getting in and out of bed, in and out ...)

www.helpguide.org > articles > senior-housing > home-... ▾  
**Home Care Services for Seniors - HelpGuide.org**  
Help with the activities of daily living, such as dressing, bathing, or meal preparation, is called personal or custodial care. Home health aides can provide personal care services that range from a few hours a day to around-the-clock live-in care.

www.helpinghandshomecare.co.uk > home-care-services ▾  
**Elderly Care - Book Care & Support Of The Elderly | Helping ...**  
Aug 28, 2020 — Introductory senior care agencies can't enforce the training of their agency staff. However, with **Helping Hands**, each of our in-house carers are ...

What are the types of elderly care? ▾

What does care of the elderly mean? ▾

Why is elderly care important? ▾

londonlive-incare.com > what-does-a-personal-assistant... ▾  
**What Does a Personal Assistant For The Elderly Actually Do ...**  
Feb 18, 2019 — A personal assistant for the **elderly** helps an older person with activities related to general life, e.g. managing money, paying bills on time, making appointments, arranging transport or actually catching the transport, and tasks around the house which the person might otherwise struggle with.

en.wikipedia.org > wiki > Elderly\_care ▾  
**Elderly care - Wikipedia**  
Their adult children often find it challenging to **help** their **elderly** parents make the right choices. Assisted living is one option for the **elderly** who need **assistance** ...

www.agincare.com > ... > Home Care > Articles ▾  
**In-Home Services That **Help** Seniors Continue to Live at ...**  
Apr 16, 2020 — There are various sources of **assistance** for **seniors** living alone. The following list includes some common things elders need a **helping hand** ...

About 939.000 results (0,78 seconds)

Ludwig Boltzmann / Date of birth

# February 20, 1844

Born in Vienna on **February 20, 1844** to an Austrian government official, Boltzmann studied physics at the University of Vienna. He received his doctorate in 1866 and in 1869 was appointed to the chair of theoretical physics at the University of Graz.

<https://depts.washington.edu/boltzmann/boltzmannbio> ::

[Boltzmann's Biography](#)



People also search for



James Clerk  
Maxwell  
June 13, 1831



Paul Ehrenfest  
January 18, 1880



Josef Stefan  
March 24, 1835

Feedback

[https://en.wikipedia.org/wiki/Ludwig\\_Boltzmann](https://en.wikipedia.org/wiki/Ludwig_Boltzmann) ::

[Ludwig Boltzmann - Wikipedia](#)

Ludwig Eduard Boltzmann (German pronunciation: [lu:tviç 'bołtsman]; **20 February 1844 – 5 September 1906**) was an Austrian physicist and philosopher.

**Ludwig Boltzmann**

Austrian physicist ::



Ludwig Eduard Boltzmann was an Austrian physicist and philosopher. His greatest achievements were the

# Word Prediction & Spelling Correction

Hey hope you had a good day! Do you want to go ruini?

108/1

ruining ruini running

1 2 3 4 5 6 7 8 9 0

q w e r t y u i o p

@ # & \* - + = ( )

a s d f g h j k l

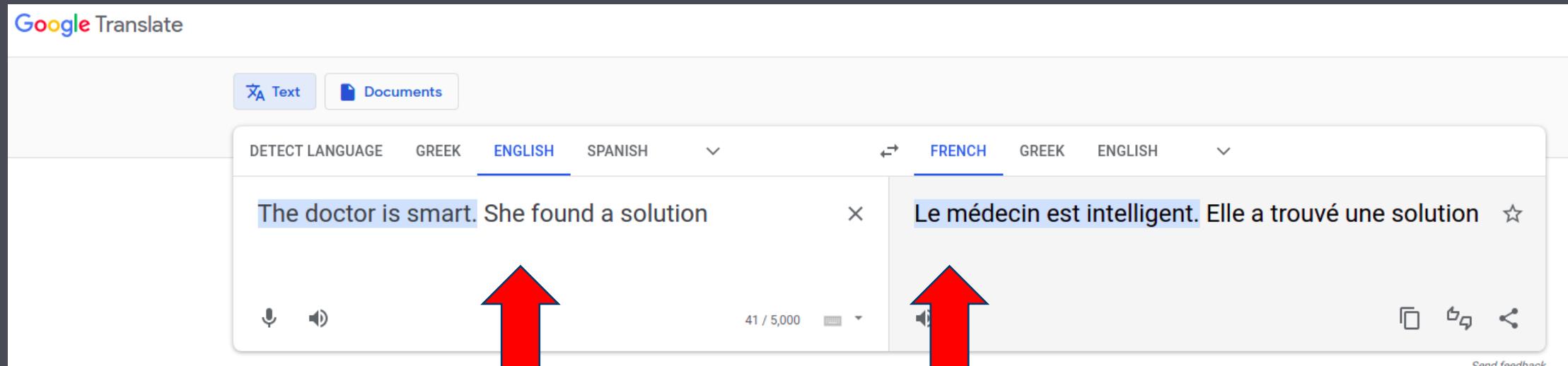
z x c v b n m

123 , . !? .

SwiftKey

The image shows a smartphone screen displaying a messaging interface. At the top, there's a message from a user with a plus sign icon and a green circular send button with a white arrow. The message text is "Hey hope you had a good day! Do you want to go ruini?". In the top right corner, it says "108/1". Below the message is a virtual keyboard. The word "ruini" is being typed, with each letter appearing sequentially. A red arrow points specifically to the letter "i" on the keyboard, highlighting the misspelling. The keyboard has a standard QWERTY layout with additional symbols like punctuation and numbers. The SwiftKey logo is visible at the bottom of the keyboard area.

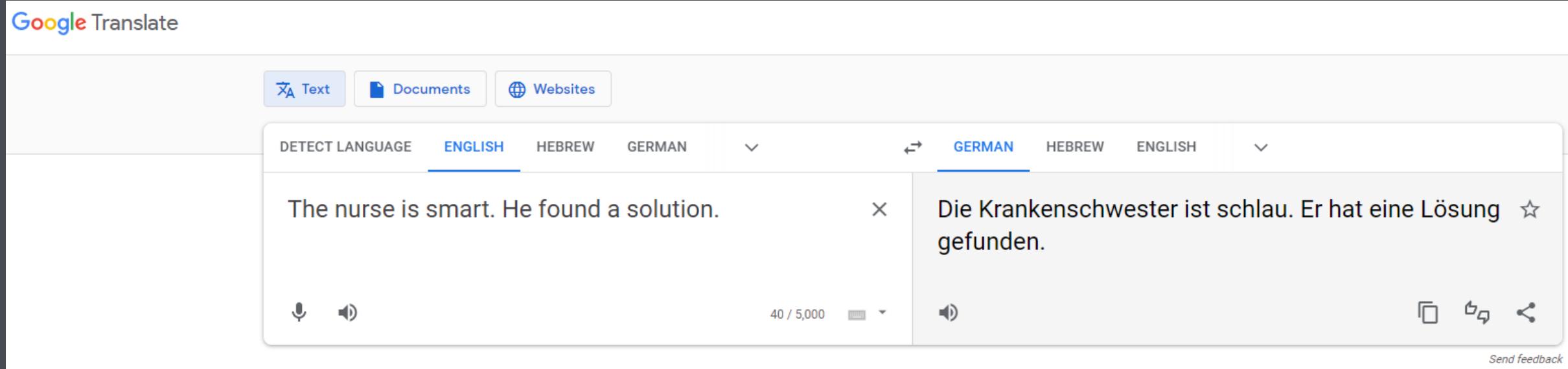
# Machine Translation



... How would you translate this passage to French / German / Spanish / Italian / Portuguese / Arabic?

# Hidden Model Bias

Google Translate



The interface shows two input fields. The left field has "Text" selected and contains the English sentence "The nurse is smart. He found a solution." The right field has "Websites" selected and contains the German translation "Die Krankenschwester ist schlau. Er hat eine Lösung gefunden." Below the input fields are language detection dropdowns set to "DETECT LANGUAGE" and "GERMAN" respectively, and a bidirectional arrow icon. At the bottom are various interaction icons like microphone, speaker, and share.

Text Documents Websites

DETECT LANGUAGE ENGLISH HEBREW GERMAN ↕ GERMAN HEBREW ENGLISH

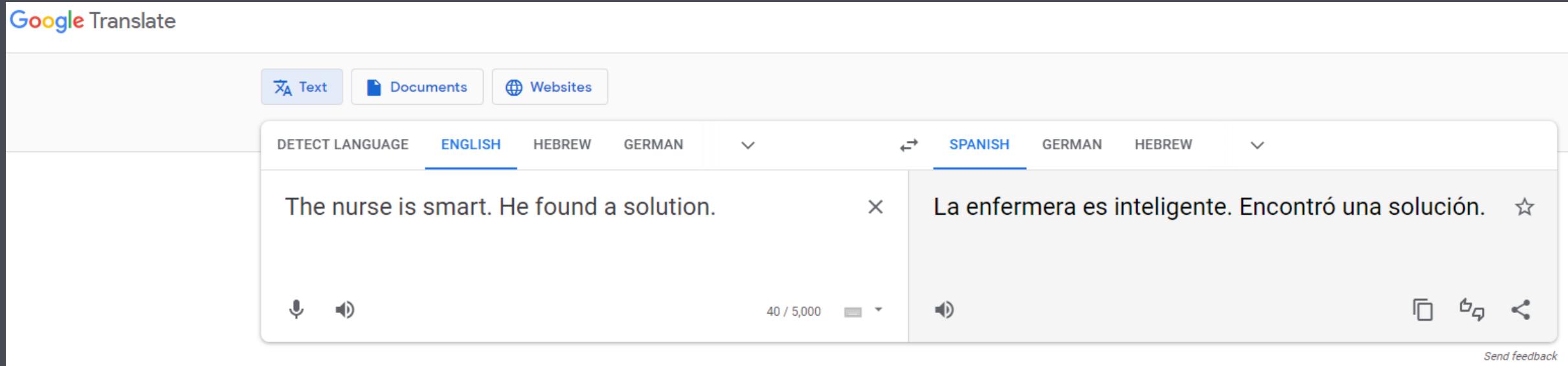
The nurse is smart. He found a solution. × Die Krankenschwester ist schlau. Er hat eine Lösung gefunden. ☆

40 / 5,000

Send feedback

# Hidden Model Bias

Google Translate



The screenshot shows the Google Translate interface. At the top, there are three tabs: 'Text' (selected), 'Documents', and 'Websites'. Below the tabs, the 'DETECT LANGUAGE' dropdown is set to 'ENGLISH', and the target language dropdown is set to 'SPANISH'. The input text 'The nurse is smart. He found a solution.' is on the left, and the translated text 'La enfermera es inteligente. Encontró una solución.' is on the right. A red star icon is visible next to the Spanish translation. At the bottom of the interface, there are icons for microphone, speaker, and sharing, along with a 'Send feedback' link.

Text

Documents

Websites

DETECT LANGUAGE ENGLISH HEBREW GERMAN ▾

SPANISH GERMAN HEBREW ▾

The nurse is smart. He found a solution. × La enfermera es inteligente. Encontró una solución. ☆

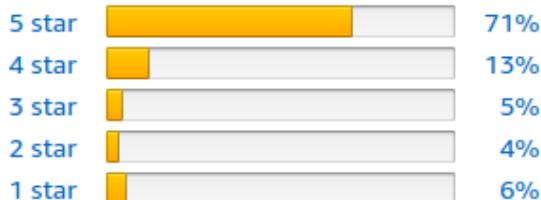
40 / 5,000

Send feedback

## Customer reviews

★★★★★ 4.4 out of 5

11,242 global ratings



[How are ratings calculated?](#)

## By feature

Quality of material	★★★★★ 4.5
Battery life	★★★★★ 4.5
Ergonomic	★★★★★ 4.4

[See more](#)

## Review this product

Share your thoughts with other customers

[Write a customer review](#)

## Customer images



## Read reviews that mention

battery life   stopped working   wireless mouse   great mouse  
programmable buttons   thumb buttons   highly recommend   extra buttons  
thumb rest   wireless gaming   buttons on the side   ever owned

Top reviews ▾

## Top reviews from the United States

Showing 1-8 of 161 reviews with "highly recommend". [Clear filter](#)



Smitty83

★★★★★ Excellent mouse for general use! Can't comment on gaming experience

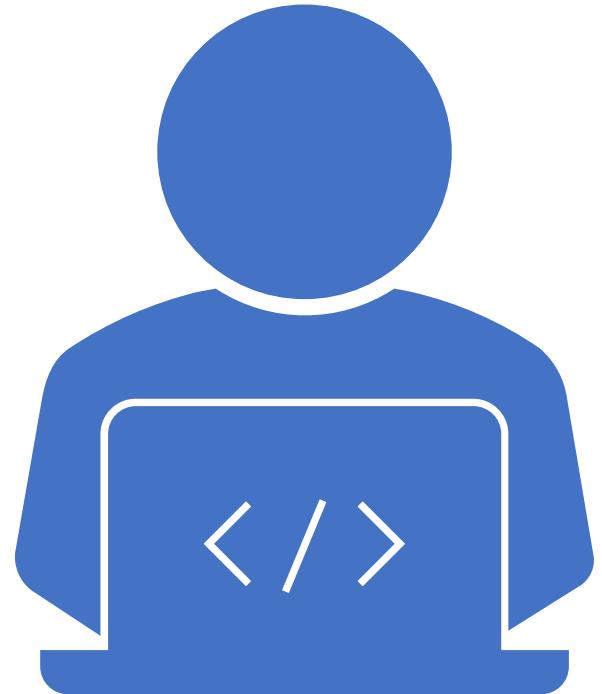
Reviewed in the United States on December 9, 2019

Verified Purchase

...are all very good. I wish I had bought one of these for the office years ago. Highly recommended for \$25!  
G1 - Left Mouse G2 - Right Mouse G3 - Enter G4 - Backwards G5 - Redo G6 - Paste G7 - Forward G8 - Undo  
G9 - Copy G10 - DPI Shift (2200 ->... [Read more](#) >

# True NLP is AI-Complete

- > Turing Test: is a computer program intelligent? (1954)
- > Would a human find out that he speaks (chats) with a computer?





# True NLP

True NLP isn't keyword matching but rather  
**real understanding of language.**

AKA **Computational Linguistics.**

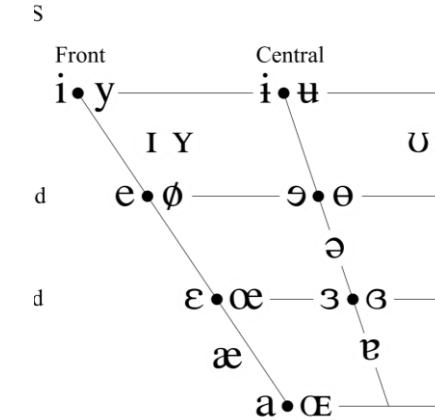
... But unfortunately, this is not yet  
feasible...

# Layers of Linguistic Knowledge



# Phonology: sound combination

- Frequent sounds combination
- Language differences



Where symbols appear in pairs, the one to the right represents a rounded vowel

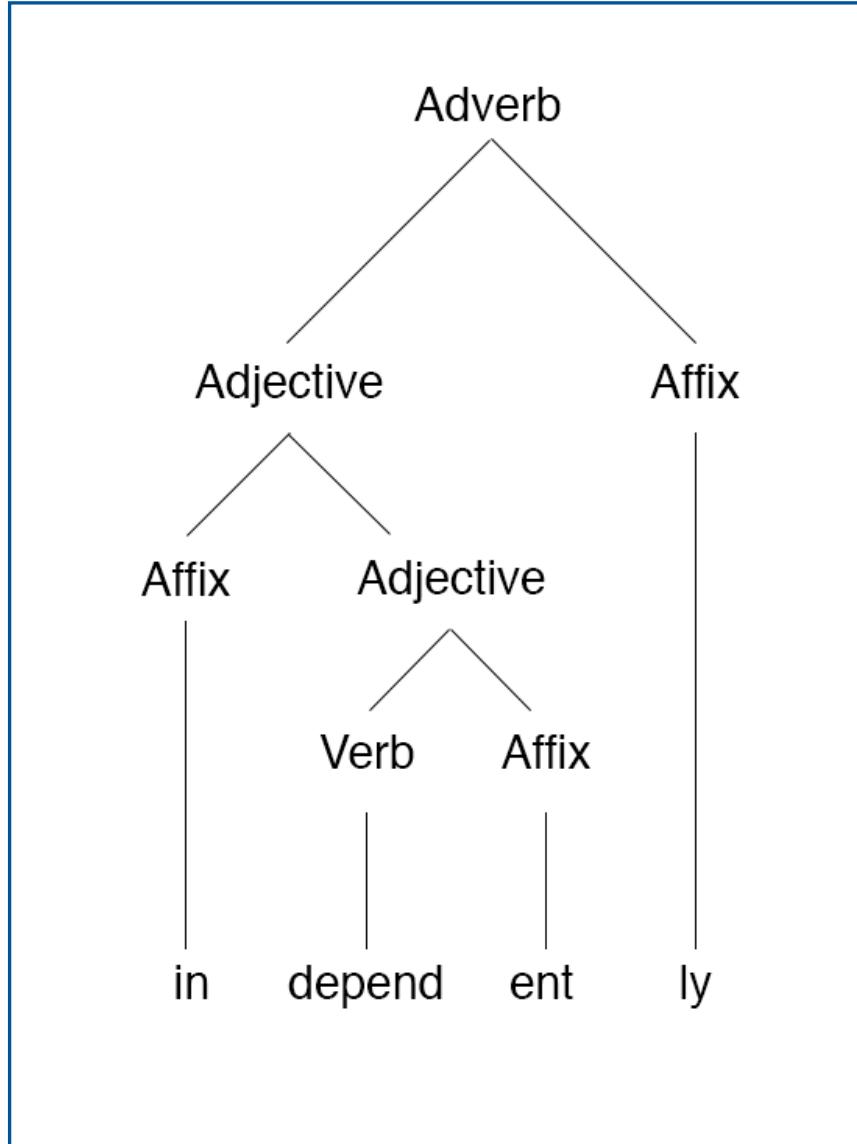


# Morphology: Word Structure

Independently: in | depend | ent | ly

Tanıştırılamazlar (Turkish)  
they can not be introduced

Lebensversicherungsgesellschaftsang  
esteller



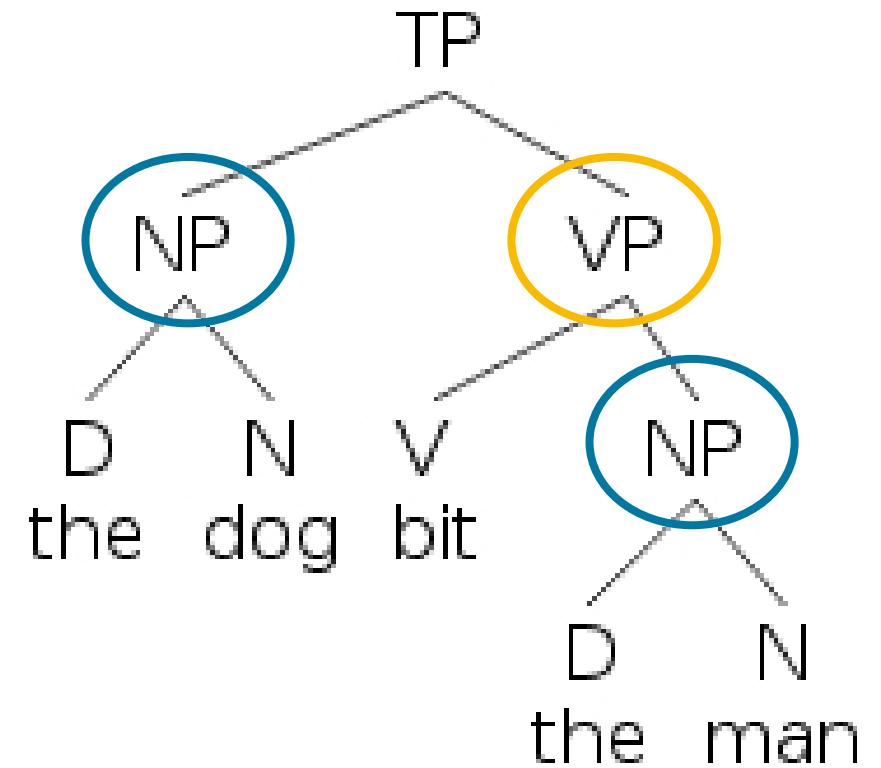
# Syntax: Sentence Structure

- > Categorization - Part of Speech (POS)
- > Phrases (more than one word)
- > Sentence Hierarchy

Good for:

- > Question Answering
- > Entity & Information Extraction

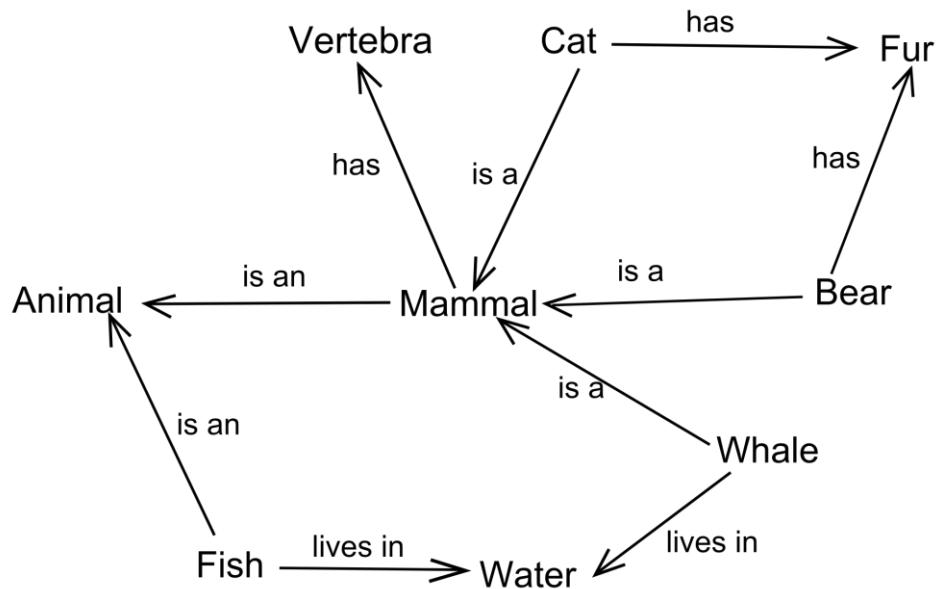
**The dog bit the man**



# Semantics: Meaning

\exists x,y: Dog(x) & Man(y) & BITE(x,y)

Semantic Graph:



The dog bit the man



# Pragmatics: Intention / Act

- > Aspects beyond the text
  - > Social relationship / Context
  - > Time & Place
- 
- I **heart** you!
  - Gosh, would you look at the time!
  - Will you **crack** open the door? It's hot in here...
  - It's cold in here, isn't it?



# Why is NLP hard?

Why Linguistics isn't  
enough?



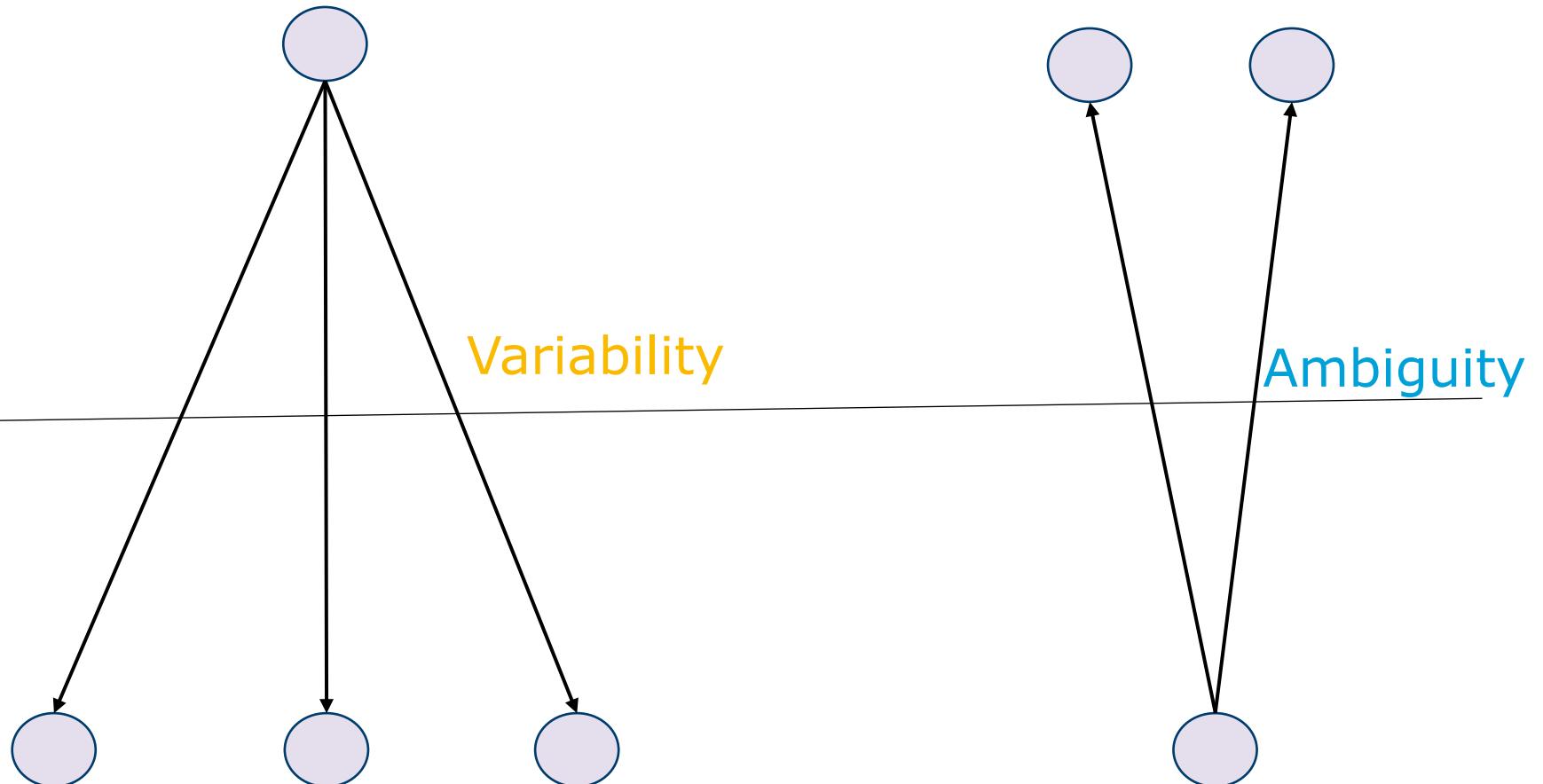
# Why is NLP hard?

Meaning

Language

Variability

Ambiguity



# Variability

## Semantical / Pragmatical

He acquired it  
He purchased it  
He bought it  
It was bought by him  
It was sold to him  
She sold it to him  
She sold him that

## Lexical \*

I loooooove it!  
everytime → every time  
oscar nom'd doc → Oscar-nominated  
documentary  
is bananas → is great

\* Wei Xu:  
<https://cocoxu.github.io>

## Negativity and Restrictivity

Martin is a Data Scientist  
Martin is **not** a data Scientist

Martin is a great Data Scientist  
Martin is **not** a great data scientist

Martin is **not** a Data Scientist  
Martin is **not** a Mad Scientist

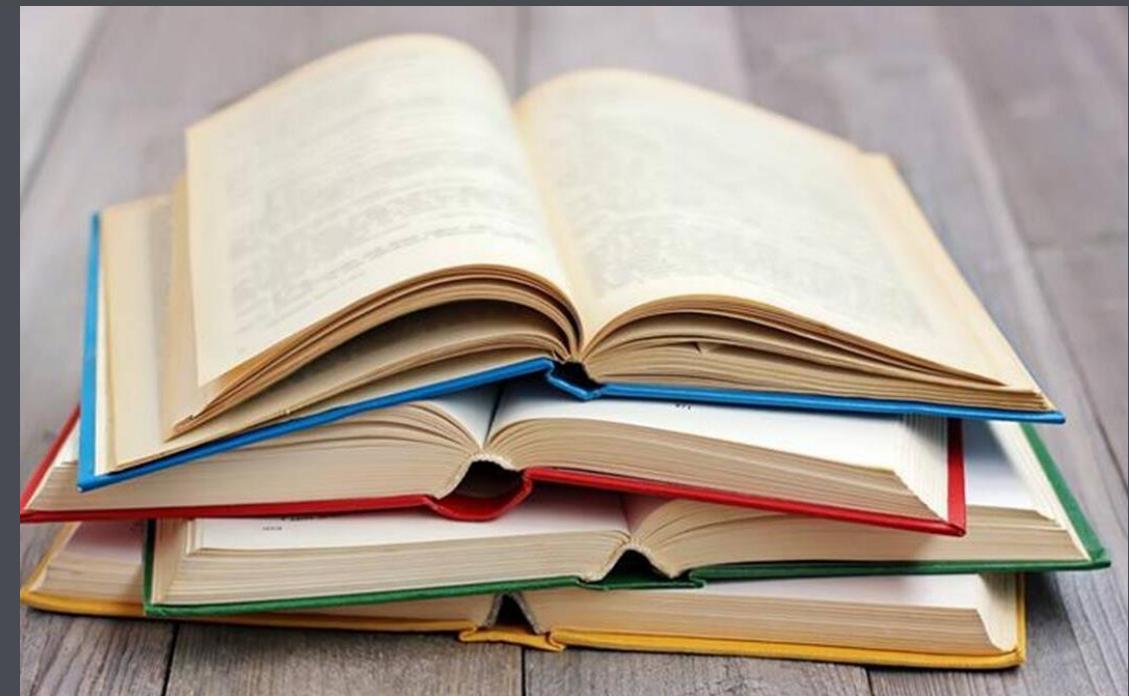


# Ambiguity: Across all levels of Description

## Phonology:

Ship // Sheep  
eat // it

Seite (page) // Saite (string)  
(isst // ist)



## Ambiguity: Across all levels of Description

**Morphology** (polysemy): books, parks  
Verb Phrase // Noun Phrase



# Ambiguity: Across all levels of Description

**Morphology (Lexical):** Bank

Noun // Noun

Der/Das Spektakel

# Ambiguity:

Across all levels of Description

## Syntactical:

I ate pizza **with olives**

I ate pizza **with fork & knife**



# Ambiguity: Across all levels of Description

## Semantical:

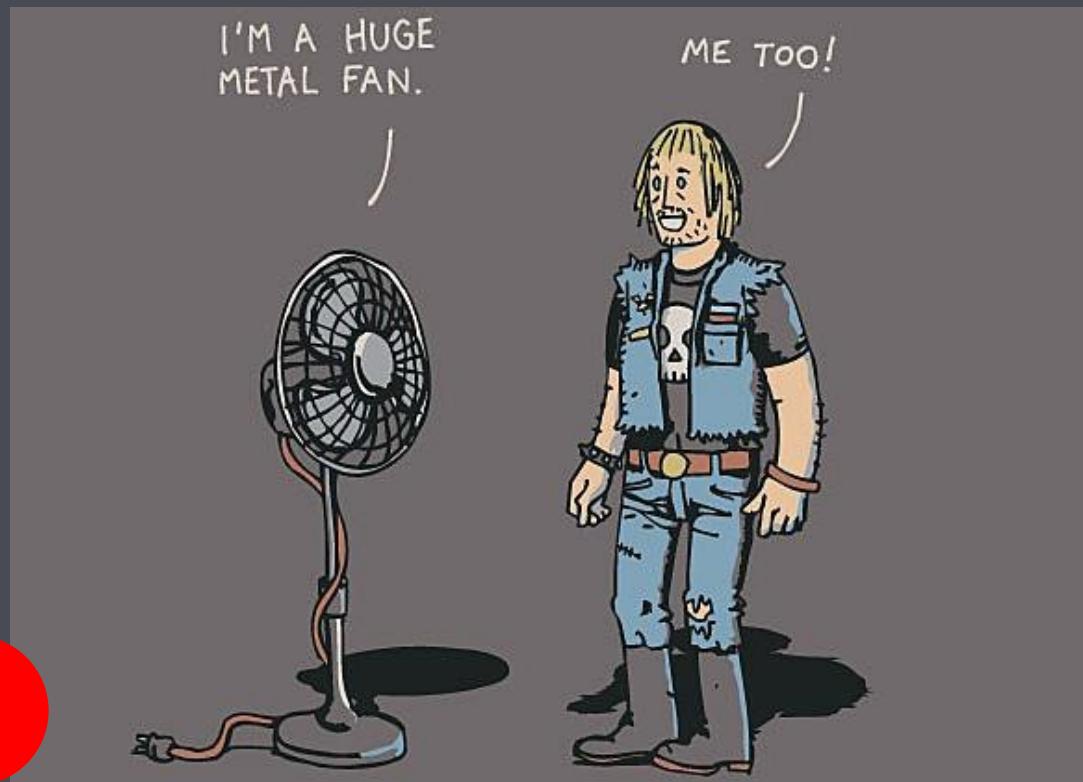
- > The chicken is ready to eat
- > Jeder Mann liebt eine Frau



# Ambiguity:

Across all levels of Description

## Pragmatics



5



# Extreme Ambiguity (Real Headlines!)

- > Scientists count whales from space
- > The pope's baby steps on gays
- > Girl hit by car in the hospital
- > Iraqi head seeks arms
- > Lung cancer in women mushrooms
- > Students get first hand job experience
- > Squad helps dog bite victim
- > Miners refuse to work after death
- > Stolen painting found by tree
- > Actor sent to jail for not finishing sentence
- > Stadium air conditioning fails — Fans protest

# Pop Quiz!

Which linguistic level is this ambiguity example?

What can I help you with?

“ Siri I'm bleeding really bad can you call me an ambulance ”

From now on, I'll call you ‘An Ambulance’. OK?

Cancel

Yes

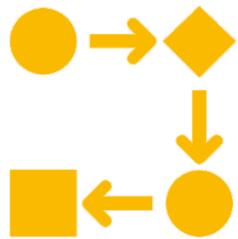


# Modeling

How is NLP done?



# NLP – How is it done?



## Rule-based modeling

Many rules (with many exceptions)



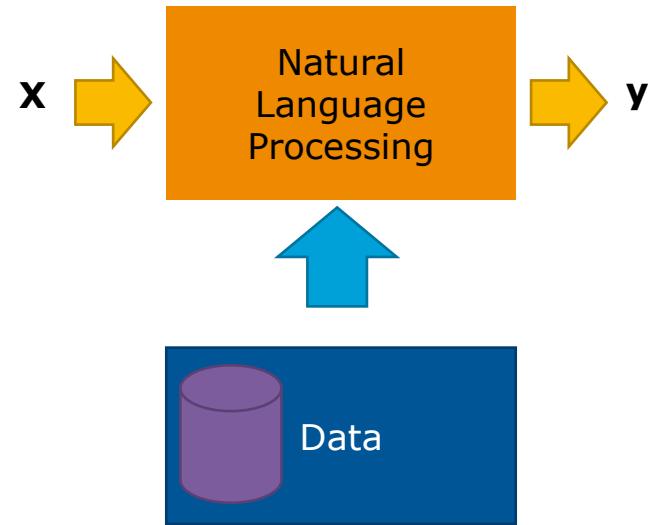
## Statistical Modeling (Data Driven)

Like children who learn by listening and trying  
Input sets of 'correct' samples (X, Y)

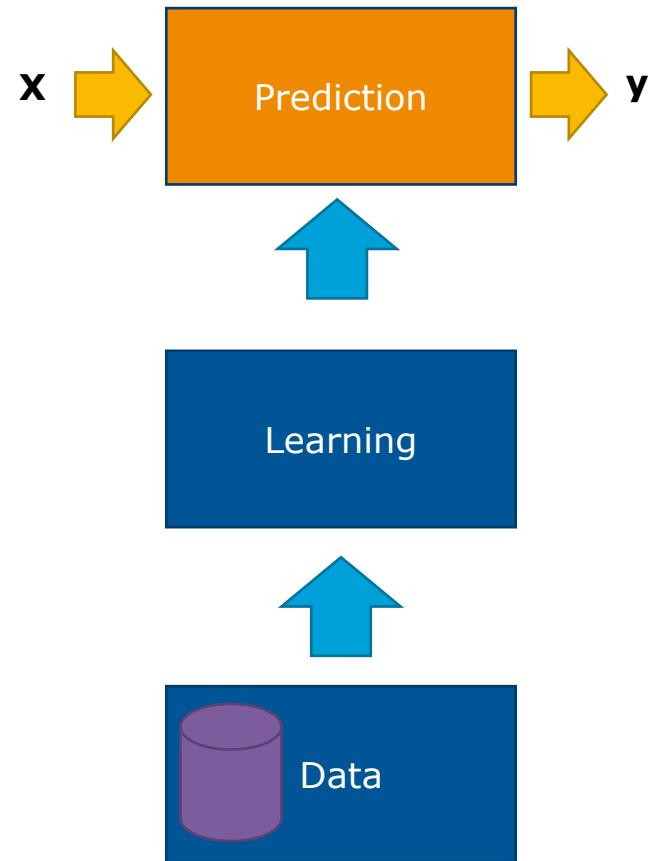
## Rule-Based Modeling

- > The soup was cold.
- > The pudding was warm.
- > The waiter was rude.
  
- > Rule: The <X> was <Y>
- > Rule: If <Bank> && <Bill> in **text** → bank statement

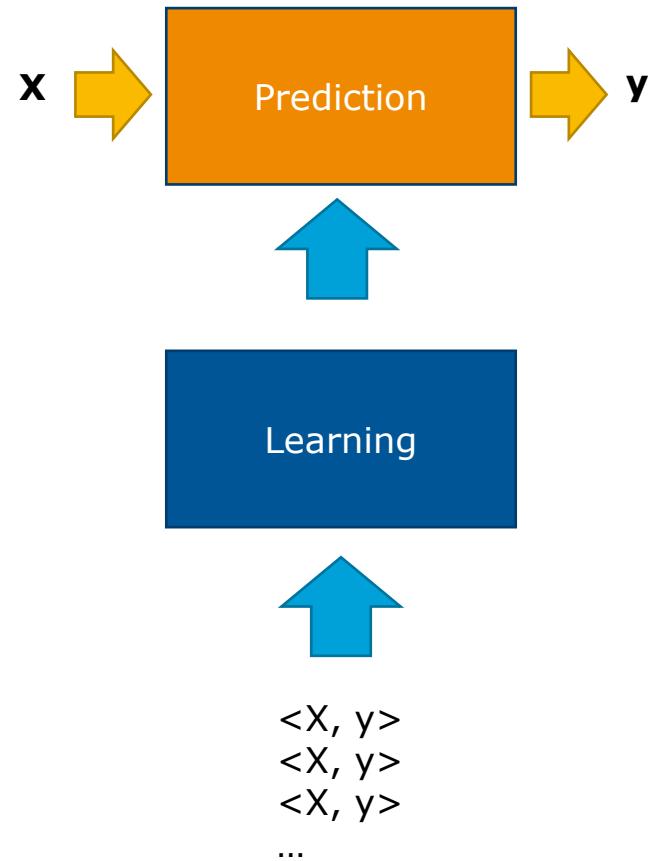
# Statistical Modeling



# Statistical Modeling



# Statistical Modeling



# Statistical Modeling

- > Define  $x$
- > Define  $y$
- > Curate Data  $\{(x, y)\}$
- > Devise Learning
- > Devise Prediction
- > Evaluate  $y^* =? y$

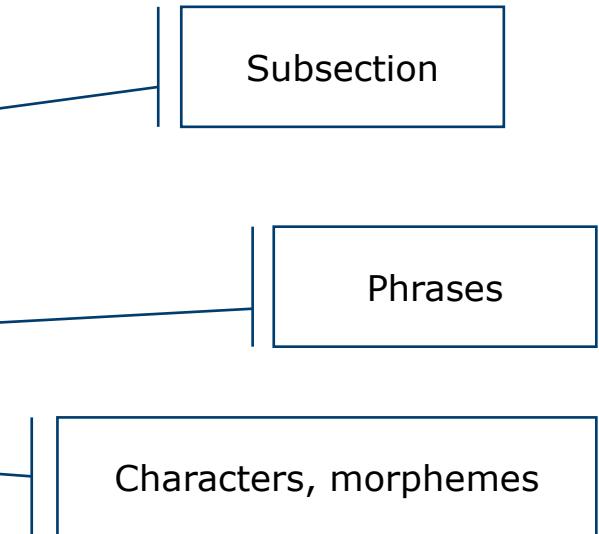


# Basic Units of Text Processing

What is this X?  
What do we process?

# What do we process?

- > A Collection of Documents
- > Document
- > Section
- > Paragraph
- > Sentence
- > Word



By Date  
By Topic  
By Web Domain  
...

# What do we process?

- > A Collection of Documents
- > **Document**
- > Section
- > Paragraph
- > Sentence
- > Word

What is a  
Document?



# Document Processing

The screenshot shows the DeepL Translator website. On the left, under 'Spanish (detected)', there is a text box containing a paragraph about globalisation. On the right, under 'Translate into English', the same paragraph is shown in English. The English text discusses communication not only between speakers of different languages but also of different varieties of the same language, such as Neutral Spanish.

at his touch of a certain icy pang along my blood. "Come, sir," said I. "You forget that I have not yet the pleasure of your acquaintance. Be seated, if you please." And I showed him an example, and sat down myself in my customary seat and with as fair an imitation of my ordinary manner to a patient, as the lateness of the hour, the nature of my preoccupations, and the horror I had of my visitor, would suffer me to muster.

"I beg your pardon, Dr. Lanyon," he replied civilly enough. "What you say is very well founded; and my impatience has shown its heels to my politeness. I come here at the instance of your colleague, Dr. Henry Jekyll, on a piece of business of some moment; and I understood..." He paused and put his hand to his throat, and I could see, in spite of his collected manner, that he was wrestling against the approaches of the hysteria—"I understood, a drawer..."

But here I took pity on my visitor's suspense, and some perhaps on my own growing curiosity.

"There it is, sir," said I, pointing to the drawer, where it lay on the floor behind a table and still covered with the sheet.

He sprang to it, and then paused, and laid his hand upon his heart: I could hear his teeth grate with the convulsive action of his jaws; and his face was so ghastly to see that I grew alarmed both for his life and reason.

"Compose yourself," said I.

He turned a dreadful smile to me, and as if with the decision of despair, plucked away the sheet. At sight of the contents, he uttered one loud sob of such immense relief that I sat petrified. And the next moment, in a voice that was already fairly well under control, "Have you a graduated glass?" he asked.

I rose from my place with something of an effort and gave him what he asked.  
He thanked me with a smiling nod, measured out a few minims of the red tincture and added one of the powders. The mixture, which was at first of a reddish hue, began, in proportion as the

## Ministerien melden 25.972 Neuinfektionen

Das Gesundheits- und das Innenministerium meldeten 25.972 neu registrierte Coronavirus-Fälle innerhalb der letzten 24 Stunden (Stand: heute, 9.30 Uhr). Diese Zahlen nannten die Bundesländer dem nationalen Krisenstab.

Landesweit starben laut Ministerien bisher 14.762 Personen an oder mit Covid-19. Derzeit befinden sich 2.412 Personen aufgrund des Coronavirus in Spitalsbehandlung, davon 190 auf Intensivstationen.

Daten des Krisenstabs in [ORF.at/corona/daten](#)

## 7-Tage-Inzidenz laut AGES bei 2.092,1

Die 7-Tage-Inzidenz, also die Zahl der Neuinfektionen mit dem Coronavirus in den abgelaufenen sieben Tagen je 100.000 Einwohnerinnen und Einwohner, liegt laut Agentur für Gesundheit und Ernährungssicherheit (AGES) bei 2.092,1 (Stand: gestern, 14.00 Uhr).

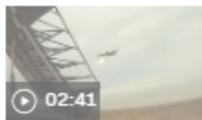
Am höchsten ist die Zahl in Tirol (2.309,2) und Vorarlberg (2.302,6). Am niedrigsten ist der Wert in Wien (1.932,9).

Das Berichtsschema der AGES zeigt die Zahlen vom Vortag – wie die Meldung der Ministerien. Wesentlicher Unterschied zu den Ministeriumsmeldungen: Laborbestätigte Fälle, Tote etc. werden nicht zum Meldezeitpunkt dargestellt, sondern zum Diagnose- bzw. Sterbedatum. In der Darstellung in ORF.at wird auch täglich transparent gemacht, welchen Tagen die neu gemeldeten Fälle zugeordnet werden.

Karten, Grafiken und Informationen zu aktuellen Fällen und zum Epidemieverlauf in [ORF.at/corona/daten](#)



At least 440 unmarked graves found in recently liberated...



Russia steps up 'kamikaze' drone attacks on Ukraine



Video reveals a major problem for new Russian soldiers



**Kyiv, Ukraine (CNN)** — Mykhailo Yatsentiuk left the basement to make tea for his granddaughter just as the bomb struck. When he came to his senses a half hour later the entire middle section of his apartment block had been destroyed; the basement where he had been sheltering with his family and neighbors was engulfed in flames.

The Ukrainian government says 54 people died at the apartment complex on 2 Pershotravneva street in Izium, eastern Ukraine, on March 9, almost half of the building's residents. Entire families were killed in the attack, including the Yatsentiuks, Kravchenkos and Stolpakovas.

Their fates remained largely unknown until a few weeks ago when Ukrainian forces pursuing a counter-offensive reclaimed Izium after six months of Russian occupation, revealing a mass burial site on the outskirts of the city.

Most of the residents of 2 Pershotravneva were buried there among more than 400 graves, few with identifying marks other than numbers daubed on rough wooden crosses.

After speaking to a survivor, ex-residents and family members, and reviewing photos and video taken in the aftermath of the attack and following the town's liberation, CNN can now tell the story of what happened at 2 Pershotravneva on that day.

### News & Buzz



Putin says war could continue until "last Ukrainian is left standing"



Injured, alone and destined for a Russian orphanage, a 12-year-old Ukrainian girl is recruited fo...

# What do we process?

- > A Collection of Documents
- > Document
- > Section
- > Paragraph
- > Sentence
- > **Word**



What is a Word?

# What is a word?

- > A sequence of characters?
- > A basic unit of meaning?
- > white-space tokenized?

# What is a word?

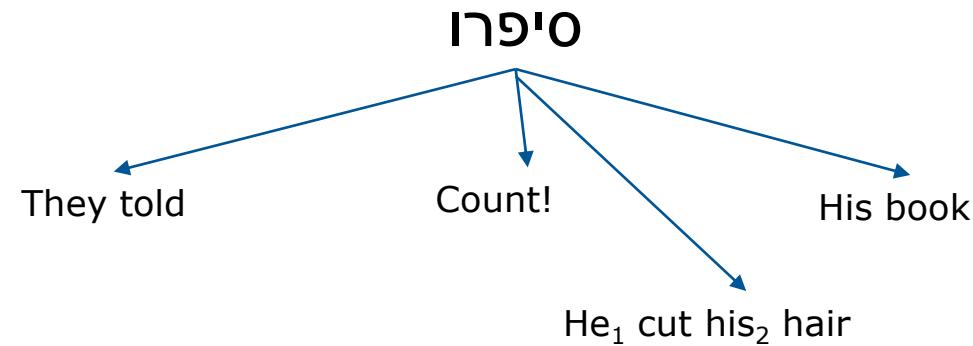
Doesn't

John's

Unlucky

Freundschaftsbeziehungen

# What is a word?



## What is a word?

> Whitespace isn't enough...

我开始写小说 = 我 开始 写 小说  
*I start(ed) writing novel(s)*

# What is a word?

Ice cream

Web site

New York

New York-Based

# What is a word?

Gave up

Made sense

Took a picture

Took apart

Took the toy apart

brechen ab

steigen ein/aus

---

# Reflexive verbs // Verb conjunction

> **French:**

Tu te llaves

> **Portuguese:**

Sente-se

> **Spanish:**

Recuérdame / le dijeron

> **Albanian:**

Laj duart

Noun conjunction: duar -> duart

# What do we process?

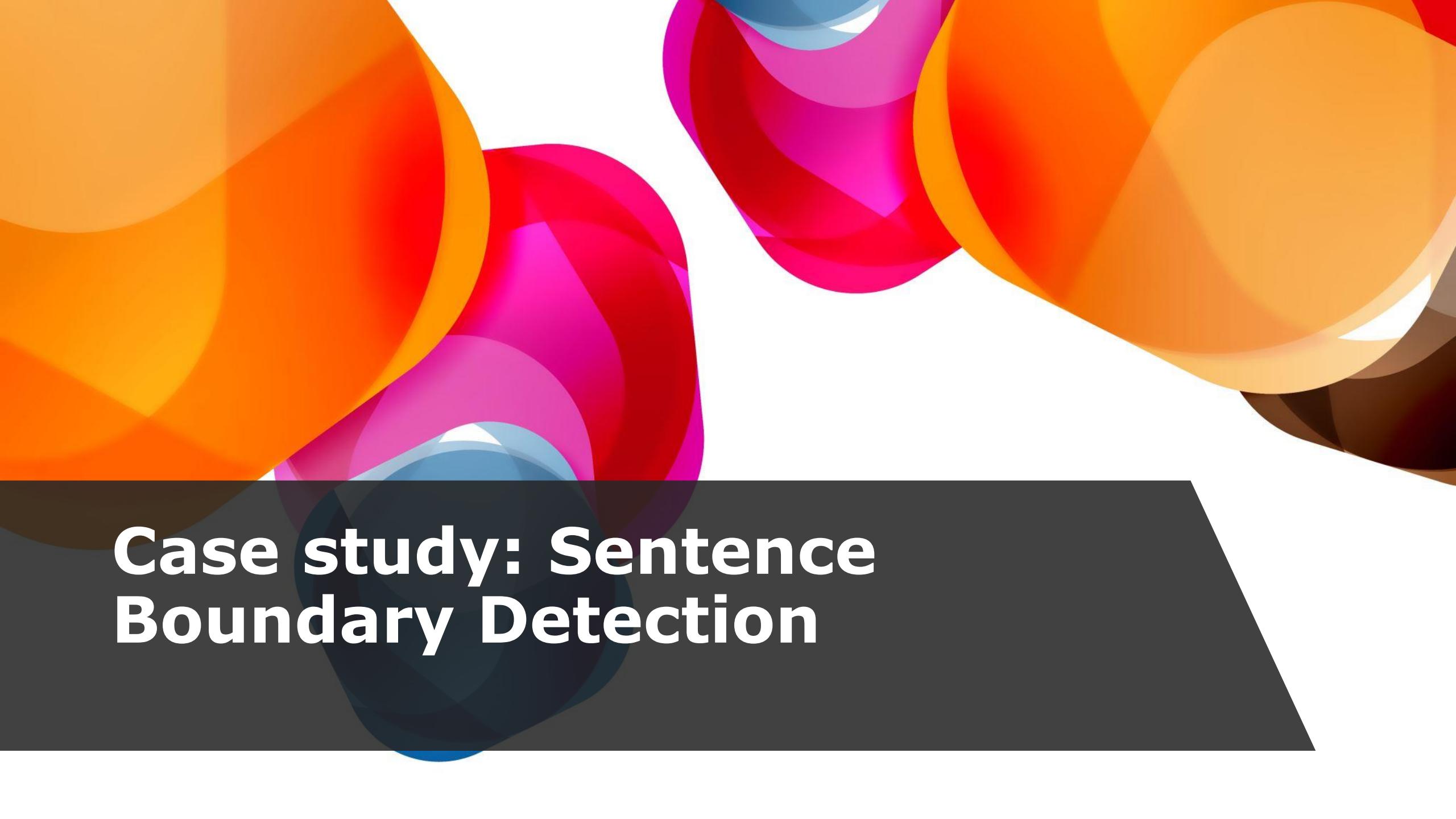
- > A Collection of Documents
- > Document
- > Section
- > Paragraph
- > **Sentence**
- > Word



What is a Sentence?

# What is a Sentence?

```
sentences = text.split(".")?
```

The background of the slide features a dynamic, abstract design composed of numerous overlapping circles in various colors. These colors include shades of orange, yellow, red, pink, purple, and blue, creating a sense of depth and movement. The circles are primarily located in the upper half of the slide, while the lower half is occupied by a solid dark gray rectangular area.

# Case study: Sentence Boundary Detection

# Question:

---

When do we need (or don't need) to divide a document into sentences?

Not needed:  
if it's already divided; Search  
Query; Working on a document-  
level (e.g., document classification)

What about translation – needed or  
not?

---



# Anaphora / Cataphora / Co-reference resolution

*Human:* What do we want?

*Computer:* Natural-Language Processing!

*Human:* When do we want it?

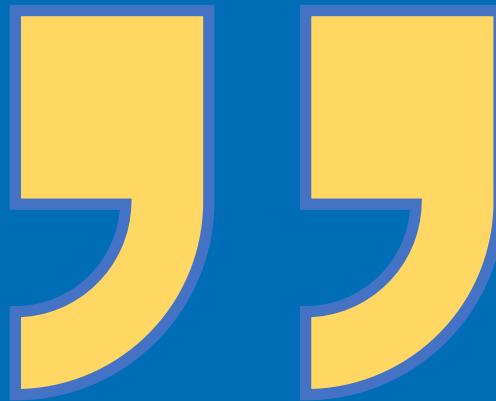
*Computer:* When do we want what?



# Sentence Boundary Detection

Python:

```
sentences = text.split(".")
```



## Sentence Boundary Detection

However, if you picked up a certain book from the library, you can opt to start up your own biplane instead of boarding the zeppelin and forgo the hassle of having to navigate around the blimp's infrastructure. Flying the plane requires you to utilize the number pad (in a slightly less frustrating way than the fights) to dodge the Luftwaffe assault while Henry blasts the enemy planes. Depending on how many planes you destroy before you get shot down, you will have to face fewer roadblocks on the way to Alexandretta when you continue the rest of the journey by car. The guards you stop you at each roadblock act as in Brunwald; i.e., find out a way to talk or bribe your way around them, or just get to punching their faces in. A certain item can even help you skip over every single guard just by showing it to them!

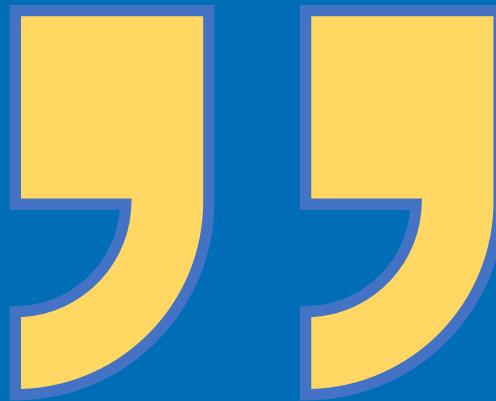
## Sentence Boundary Detection

However, if you picked up a certain book from the library, you can opt to start up your own biplane instead of boarding the zeppelin and forgo the hassle of having to navigate around the blimp's infrastructure. Flying the plane requires you to utilize the number pad (in a slightly less frustrating way than the fights) to dodge the Luftwaffe assault while Henry blasts the enemy planes. Depending on how many planes you destroy before you get shot down, you will have to face fewer roadblocks on the way to Alexandretta when you continue the rest of the journey by car. The guards you stop you at each roadblock act as in Brunwald; **i.e.**, find out a way to talk or bribe your way around them, or just get to punching their faces in. A certain item can even help you skip over every single guard just by showing it to them!

# Sentence Boundary Detection

Python:

```
sentences = text.split(". ")
```



# Sentence Boundary Detection

WASHINGTON — Former President **George W. Bush** called on Americans on Saturday to put aside partisan differences, heed the guidance of medical professionals and show empathy for those stricken by the coronavirus and the resulting economic devastation.

In a three-minute video message, **Mr. Bush**, who rarely speaks out on current events, struck a tone of unity that contrasted with the more combative approach taken at times by President Trump as the former president evoked the sense of national solidarity in the wake of the attacks of **Sept. 11, 2001**.

# Sentence Boundary Detection

Learning system, that **learns**, according to text features, which ones are attributed to sentence ends.

- Q: How to get annotated data?
- Q: Which kind of data to use?
- Q: Which clues could hint at the end-of-sentences?



## Clues to identify end-of-sentence

- > A dictionary of safe-words:  
*Mr. Dr. Ltd. Inc. U.S. e.g. i.e.*
- > Hinting words in a windowed neighborhood
- > An ensemble or a combination of clues:
  - > *Capital letter after ". "*  
Though not perfect: *George W. Bush*
- > Word length: short words are probably part of a word

Statistical (ML Task): Weighting the hints as a learnable solution.

# Weighting the hints as a learnable solution

- > Collect the data
- > Annotate
- > Define the 'hints' (a.k.a. *feature engineering*)
- > Identify the problem: Binary classification
  - > Decide which of the hints to use to determine end-of-sentence
- > Train
- > Evaluate

# Worth Reading

## Sentence Boundary Detection and the Problem with the U.S.

**Dan Gillick**

Computer Science Division  
University of California, Berkeley  
[dgillick@cs.berkeley.edu](mailto:dgillick@cs.berkeley.edu)

# SOTA is still not 100% solved

It's not a good time. (It's never a good time).

In a quiet voice, he said  
"this will not work. I am quitting",  
and then he left the room.



## CODE REVIEW

Home Questions Tags Users Unanswered Jobs

### Retrieving a substring from an exponentially growing string

Asked today Active today Viewed 68 times

For integer A, the base level of 1 is "XYZ". For subsequent levels, the levels become "X" + Level (A - 1) + "Y" + Level (A - 1) + "Z". So for level 2, the string would be "XXYZXYZZZ".

8 The objective is to return the substring from the Level string using the Start and End Index.

Example: If entering 2 3 7, it would Level 2, 3rd character to the 7th character and the result would be "YZYXY" from "XXYZXYZZZ".

1 The following constraints are given:

- $1 \leq \text{Level } K \leq 50$ ,
- $1 \leq \text{start} \leq \text{end} \leq \text{length of Level } K \text{ String}$ ,
- $1 \leq \text{end} - \text{start} + 1 \leq 100$ .

I have written a brute force approach for this problem in Python as follows

```
def my_substring():
    level = int(input())
    start_index = int(input()) - 1
    end_index = int(input())
    strings_list = [None] * level
```

# Real world example



## Sentence Boundary Detection

- Probably the most basic task...
- ...but still non-trivial
- Requires considering:
  - which corpus to use
  - features
  - annotation procedure
  - potential biases
- Depends on the use case:
  - choose a sentence definition, methods, and trade-offs.

---

# Natural Language is Hard

Two main takeaways from this course (and maybe the degree...):

- > Natural Language Processing is hard
  - Why is it hard?
  - (hence, why is it interesting!)
  - How to cope with it?
- > NLP is about compromising
  - Perform a reasonable compromise
  - Understand that you compromised
  - Understand what the compromise was
  - Try to improve it, if needed, as much as possible.