

# Multilingual Natural Language Processing

Yulia Tsvetkov & Chan Young Park

<[ytsvetko@cs.cmu.edu](mailto:ytsvetko@cs.cmu.edu)>

<[chanyoun@andrew.cmu.edu](mailto:chanyoun@andrew.cmu.edu)>



# Plan for our class

- What is multilingual NLP
- Why it can be difficult to build good NLP tools for many languages
- Why multilingual NLP is important
- A “recipe” of decisions in building NLP tools for new languages
- An exercise: multilingual sentiment analysis



# Communication with machines

- ~50s-70s



# Communication with machines

- ~80s

```
File Edit Edit_Settings Menu Utilities Compilers Test Help
EDIT      BS9U.DEVT3.CLIBPAU(TIMMIES) - 01.31          Columns 00001 000
Command ==> █
***** **** Top of Data ****
000001 /* REXX EXEC ****
000002 /*
000003 /* TIMMIES FACTOR - COMPOUND INTEREST CALCULATOR
000004 /*
000005 /* AUTHOR: PAUL GAMBLE
000006 /* DATE: OCT 1/2007
000007 /*
000008 /*
000009 ****
000010
000011
000012 say '*****'
000013 say 'Welcome Coffee drinker.'
000014 say '*****'
000015 DO WHILE DATATYPE(CoffeeAmt) \= 'NUM'
000016   say ""
000017   say "What is the price of your coffee?","
000018   "(e.g. 1.58 = $1.58)"
000019   parse pull CoffeeAmt
000020 END
000021
000022 DO WHILE DATATYPE(CoffeeWk) \= 'NUM'
000023   say ""
000024   say "How many coffees a week do you have?"
000025   parse pull CoffeeWk
000026 END
000027
000028 DO WHILE DATATYPE(Rate) \= 'NUM'
000029   say ""
000030   say "What annual interest rate would you like to see on that money?","
000031   "(e.g. 8 = 8%)"
000032   parse pull Rate
000033 END
000034 Rate = Rate * 0.01 /* CHG TO DECIMAL NUMBER */
000035
```



# Communication with machines

- Today



# Language technologies: conversational agents

- A conversational agent contains

- Speech recognition
- Language analysis
- Dialogue processing
- Information retrieval
- Text to speech



# Natural Language Processing

- A conversational agent contains
  - Speech recognition
  - Language analysis
  - Dialogue processing
  - Information retrieval
  - Text to speech



# Natural Language Processing

- Applications
  - Machine Translation
  - Information Extraction
  - Question Answering
  - Dialogue Systems
  - Summarization
  - Sentiment Analysis
  - ...
- Core technologies
  - Language modelling
  - Part-of-speech tagging
  - Syntactic parsing
  - Named-entity recognition
  - Coreference resolution
  - Word sense disambiguation
  - Semantic role labelling
  - ....



# Machine Translation

Google Translate™ BETA

Text and Web Translated Search Dictionary Tools

Translate Text

Original text:

Istota instytucji wyłączenia organu podatkowego od załatwienia sprawy dotyczącej zobowiązania podatkowego lub innej sprawy normowanej przepisami prawa podatkowego jest utrata właściwości danego organu do załatwienia danej sprawy.

Translation: Polish (automatically detected language: Finnish)

Pelkät vapautusta veron käsittelyvälle viranomaiselle tapauksissa, joissa verovelar muita aineita, normowanej vero-oikeuden menetys kiinteäkyseisen viranomaisen ratkaiserityinen veronmaksajille.

[Detect language](#) » [Finnish](#) [Translate](#)

[Suggest a better translation](#)

usage	English	Telugu	Swahili	Translate		
Detect language	Corsican	Gujarati	Kazakh	Marathi	Shona	Urdu
Afrikaans	Croatian	Haitian Creole	Khmer	Mongolian	Sindhi	Uzbek
Albanian	Czech	Hausa	Korean	Myanmar (Burmese)	Sinhala	Vietnamese
Amharic	Danish	Hawaiian	Kurdish (Kurmanji)	Nepali	Slovak	Welsh
Arabic	Dutch	Hebrew	Kyrgyz	Norwegian	Slovenian	Xhosa
Armenian	English	Hindi	Lao	Pashto	Somali	Yiddish
Azerbaijani	Esperanto	Hmong	Latin	Persian	Spanish	Yoruba
Basque	Estonian	Hungarian	Latvian	Polish	Sundanese	Zulu
Belarusian	Filipino	Icelandic	Lithuanian	Portuguese	Swahili	Swedish
Bengali	Finnish	Igbo	Luxembourgish	Punjabi	Tajik	Tamil
Bosnian	French	Indonesian	Macedonian	Romanian	Telugu	Thai
Bulgarian	Frisian	Irish	Malagasy	Russian	Turkish	Ukrainian
Catalan	Galician	Italian	Malay	Samoan		
Cebuano	Georgian	Japanese	Malayalam	Scots Gaelic		
Chichewa	German	Javanese	Maltese	Serbian		
Chinese	Greek	Kannada	Maori	Sesotho		



# Sentiment Analysis



**HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner**

\$89 online, \$100 nearby    377 reviews

September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sh

## Reviews

**Summary** - Based on 377 reviews



### What people are saying

ease of use	A horizontal green progress bar with a small red segment at the beginning, indicating a low rating.	"This was very easy to setup to four computers."
value	A horizontal green progress bar with a small red segment at the beginning, indicating a low rating.	"Appreciate good quality at a fair price."
setup	A horizontal green progress bar with a small red segment at the beginning, indicating a low rating.	"Overall pretty easy setup."
customer service	A horizontal red progress bar, indicating a low rating.	"I DO like honest tech support people."
size	A horizontal green progress bar with a small red segment at the beginning, indicating a low rating.	"Pretty Paper weight."
mode	A horizontal green progress bar with a small red segment at the beginning, indicating a low rating.	"Photos were fair on the high quality mode."
colors	A horizontal green progress bar with a small red segment at the beginning, indicating a low rating.	"Full color prints came out with great quality."



# Text Summarization

text summarization

All Images News Videos Maps More Settings Tools

About 2,960,000 results (0.36 seconds)

**Text summarization** is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks).

Aug 6, 2018

**Unsupervised Text Summarization using Sentence Embeddings**  
<https://medium.com/.../unsupervised-text-summarization-using-sentence-embeddings-ad...>

>About this result | Feedback

```
graph LR; SD[Single Document] --> IT[Input Text]; MD[Multi Document] --> IT; IT --> TS[Text Summarization]; TS --> E[Extractive]; TS --> A[Abstractive]; E --> DS[Document Summary]; E --> SB[Story-based]; A --> S[Summary]
```

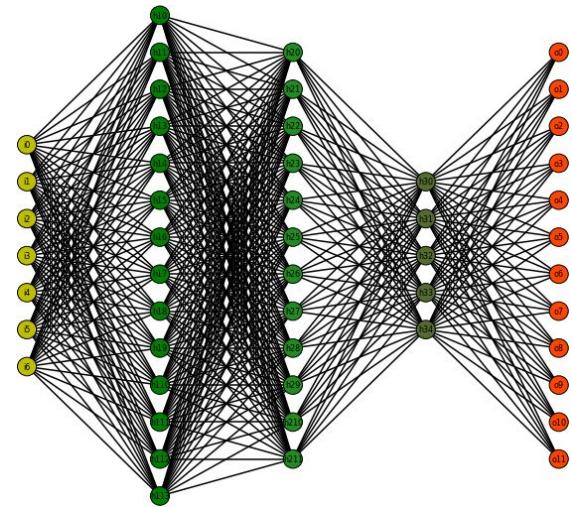
## People also ask

What is Abstractive text summarization?

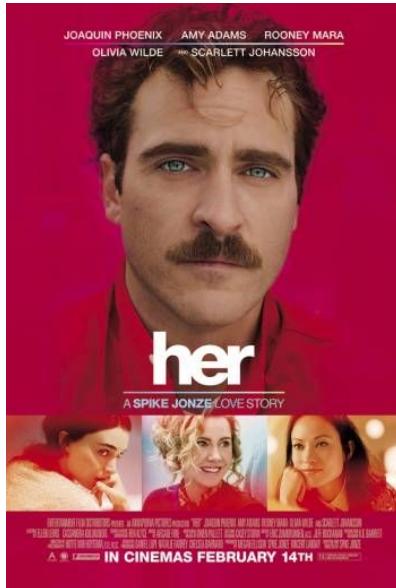
Why do we need text summarization?







# Still, NLP is not solved yet



## Baseline mutual information model (Li et al. 2015)

A: Where are you going? (1)

B: I'm going to the restroom. (2)

A: See you later. (3)

B: See you later. (4)

A: See you later. (5)

B: See you later. (6)

...

...

A: how old are you? (1)

B: I'm 16. (2)

A: 16? (3)

B: I don't know what you are talking about. (4)

A: You don't know what you are saying. (5)

B: I don't know what you are talking about . (6)

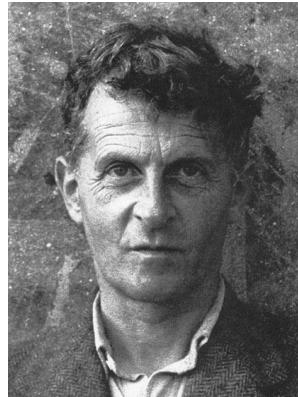
A: You don't know what you are saying. (7)

...

Li et al. (2016), "Deep Reinforcement Learning for Dialogue Generation" *EMNLP*

# Why NLP is hard?

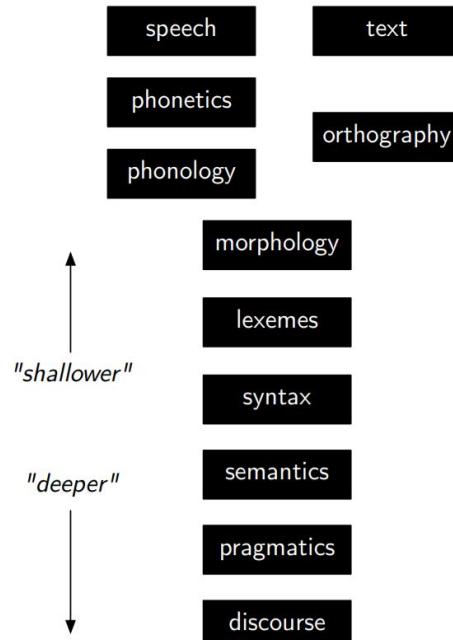
*"the limits of my language are the limits of my world"*



Ludwig Wittgenstein

# Why NLP is hard?

- Language consists of many levels of structure
- Humans fluently integrate all of these to produce and understand language
- Ideally, so would a computer!



# Phonetics, phonology

- Pronunciation modeling

**SOUNDS**

Th i a si e n



# Words

- Language modeling
- Tokenization
- Spelling correction

**WORDS**

This is a simple sentence



# Morphology

- Morphological analysis
- Tokenization
- Lemmatization

WORDS	This is a simple sentence
MORPHOLOGY	be 3sg present



# Syntax

- Part-of-speech tagging

PART OF SPEECH	DT	VBZ	DT	JJ	NN
WORDS	This	is	a	simple	sentence
MORPHOLOGY				be 3sg present	



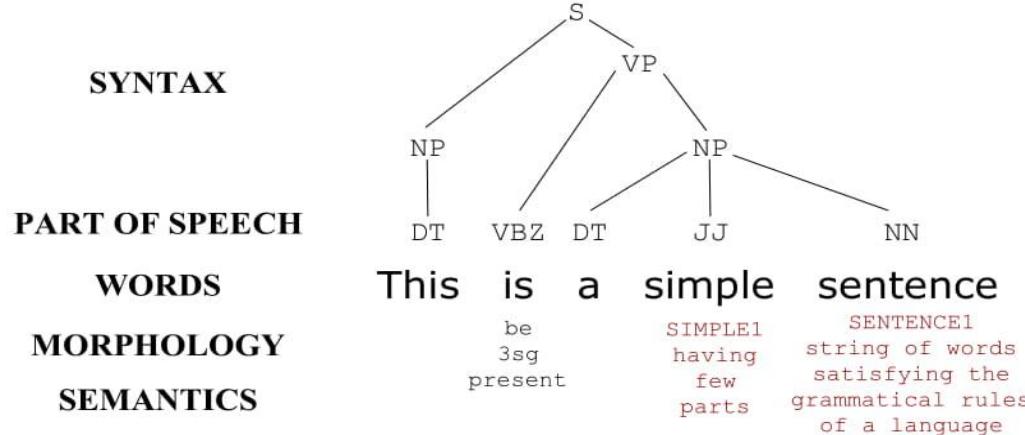
# Syntax

- Syntactic parsing



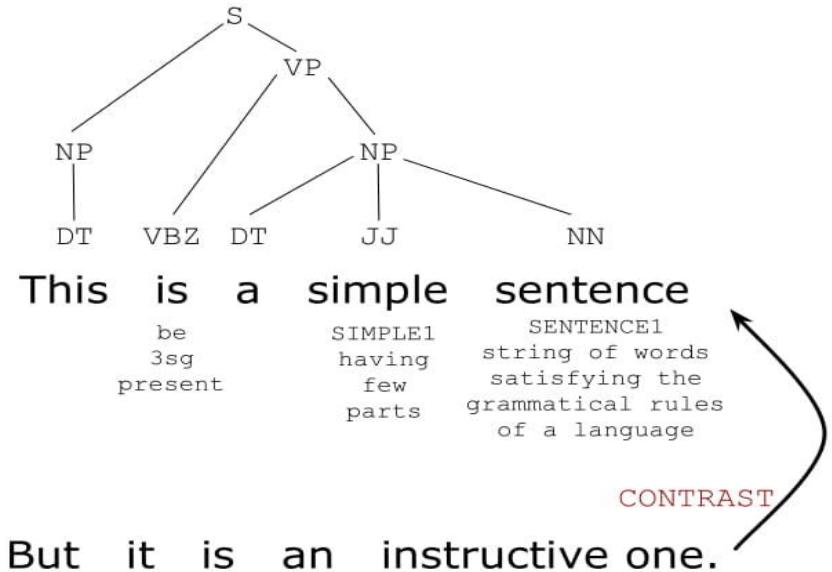
# Semantics

- Named entity recognition
- Word sense disambiguation
- Semantic role labelling



# Discourse

- Reference resolution
- Discourse parsing



# Why NLP is hard?

Ambiguity at multiple levels:

- Word senses: **bank** (finance or river?)
- Part of speech: **chair** (noun or verb?)
- Syntactic structure: **I can see a man with a telescope**
- Multiple: **I saw her duck**



# Why NLP is hard?

**Ambiguity at multiple levels**

**Linguistic diversity**



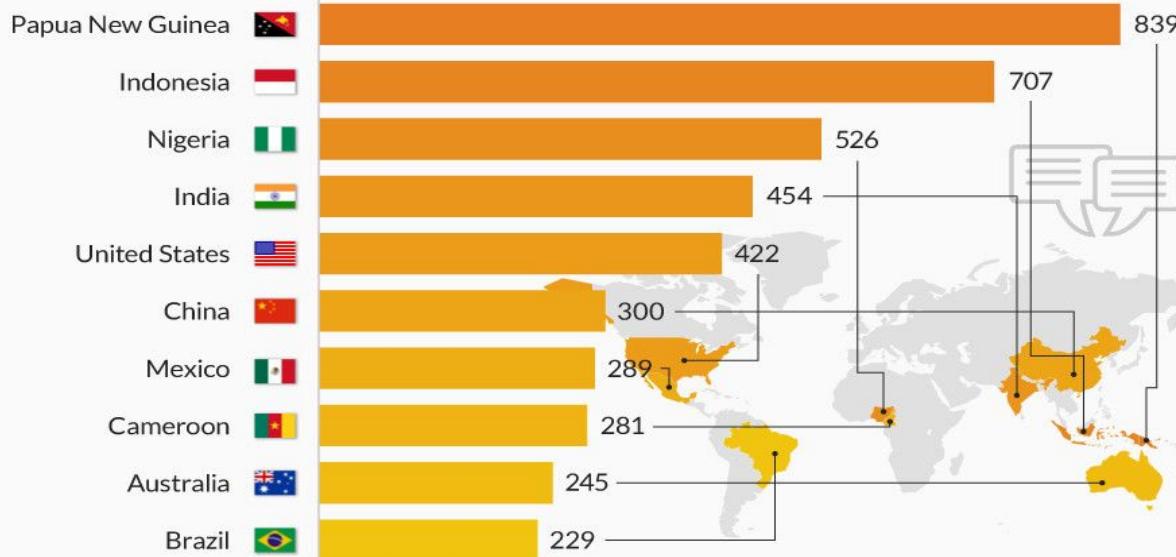
# Linguistic diversity: ~6,000 languages

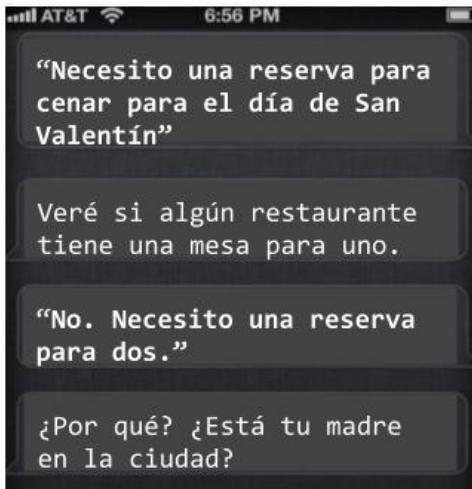


# The multilingual world

## The Countries With The Most Spoken Languages

Number of living languages spoken per country in 2015

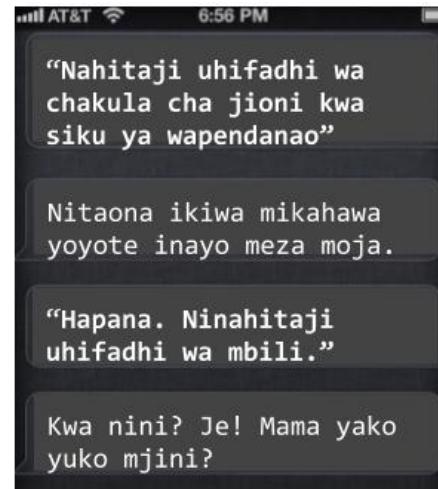




Spanish  
534 million speakers



Hindi  
615 million speakers



Swahili  
100 million speakers

# Language diversity: language families

[www.enthologue.com](http://www.enthologue.com)

- Niger-Congo (1538 languages) (20.6%)
- Austronesian (1257 languages) (16.8%)
- Trans-New Guinea (480 languages) (6.4%)
- Sino-Tibetan (457 languages) (6.1%)
- Indo-European (444 languages) (5.9%)
- Australian (378 languages) (5.1%)
- Afro-Asiatic (375 languages) (5.0%)
- Nilo-Saharan (205 languages) (2.7%)
- Oto-Manguean (177 languages) (2.4%)
- Austroasiatic (169 languages) (2.3%)
- Volta Congo (108 languages) (1.5%)
- Tai-Kadai (95 languages) (1.3%)
- Dravidian (85 languages) (1.1%)
- Tupian (76 languages) (1.0%)



# Linguistic diversity: words

- Tokenization is different for different languages

这是一个简单的句子

**WORDS**

This is a simple sentence

זה משפט פשוט



# Linguistic diversity: Hebrew words

- Most vowels are left unspecified

in tea

her daughter

בָתָה



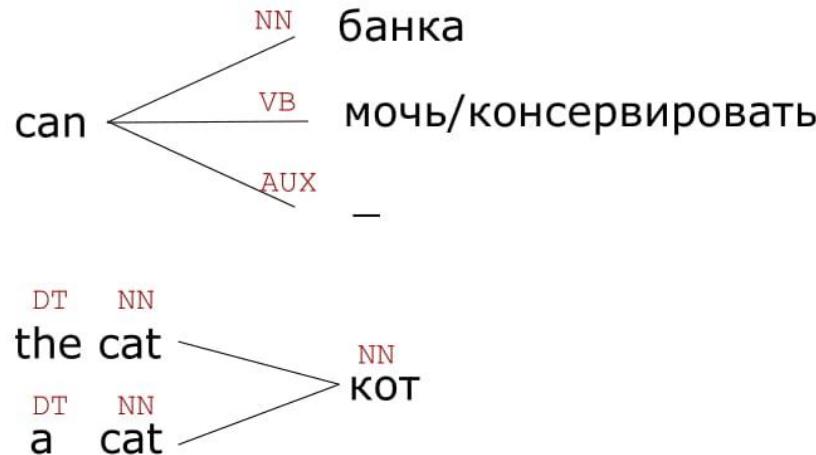
# Linguistic diversity: Hebrew words

- Most vowels are left unspecified
- Particles, prepositions, definite article, conjunctions attach to the words which follow them
- Tokenization + word sense disambiguation are much harder

in tea	ב תה
in the tea	ב ה ת ה
that in tea	ש ב ת ה
that in the tea	ש ב ה ת ה
and that in the tea	ו ש ב ה ת ה
	ו ש ב ת ה
and her saturday	ו ש ב ת ה +ה
and that in tea	ו ש +ב+ת ה
and that her daughter	ו ש +ב ת ה +ה



# Linguistic diversity: parts of speech



# Linguistic diversity: English morphology

**WORDS**  
**MORPHOLOGY**

This is a simple sentence

be  
3sg  
present



# Linguistic diversity: English morphology

unfriend, Obamacare



# Linguistic diversity: Russian morphology

	<b>Singular+neut</b>	<b>Plural+neut</b>	
<b>Nominative</b>	предложение	предложения	sentence (s)
<b>Genitive</b>	предложения	предложений	(of) sentence (s)
<b>Dative</b>	предложению	предложениям	(to) sentence (s)
<b>Accusative</b>	предложение	предложения	sentence (s)
<b>Instrumental</b>	предложением	предложениями	(by) sentence (s)
<b>Prepositional</b>	предложении	предложениях	(in/at) sentence (s)



# Linguistic diversity: Quechua morphology

Much'ananayakapushasqakupuniñataqsunamá

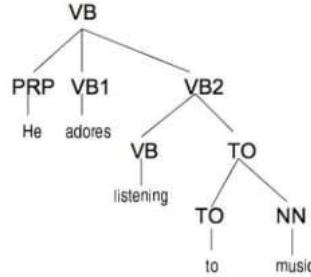
Much'a -na -naya -ka -pu -sha -sqa -ku -puni -ña -taq -suna -má

*"So they really always have been kissing each other then"*

Much'a	to kiss
-na	expresses obligation, lost in translation
-naya	expresses desire
-ka	diminutive
-pu	reflexive (kiss *eachother*)
-sha	progressive (kiss*ing*)
-sqa	declaring something the speaker has not personally witnessed
-ku	3rd person plural (they kiss)
-puni	definitive (really*)
-ña	always
-taq	statement of contrast (...then)
-suna	expressing uncertainty (So...)
-má	expressing that the speaker is surprised

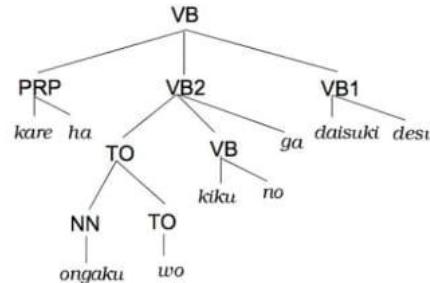
# Linguistic diversity: syntax

**svo**



he adores listening to music

**sov**



かれは おんがくを きくのが だいすき です  
kare ha ongaku wo kiku no ga daisuki desu



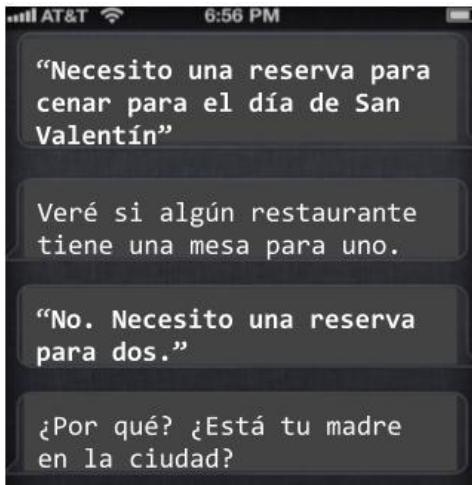
# Linguistic diversity: semantics

- Every language sees the world in a different way
- For example, it could depend on cultural or historical conditions



- Russian has very few words for colors, Japanese has hundreds
- Multiword expressions, e.g. **it's raining cats and dogs** or **wake up** and metaphors, e.g. **love is a journey** are very different across languages

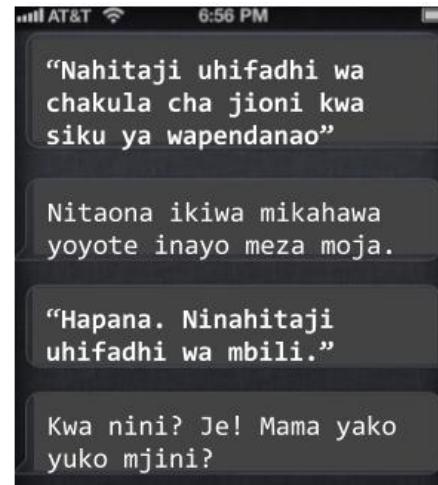
# ~6,000 languages



Spanish  
534 million speakers

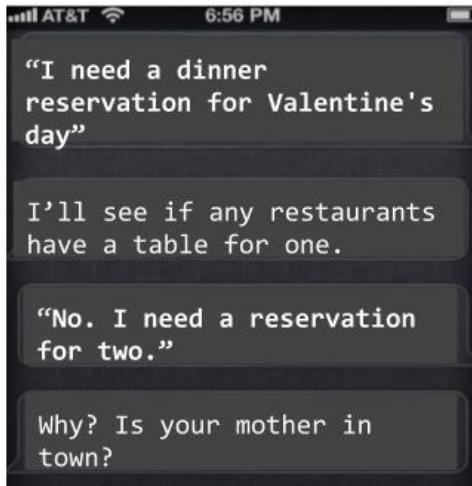


Hindi  
615 million speakers

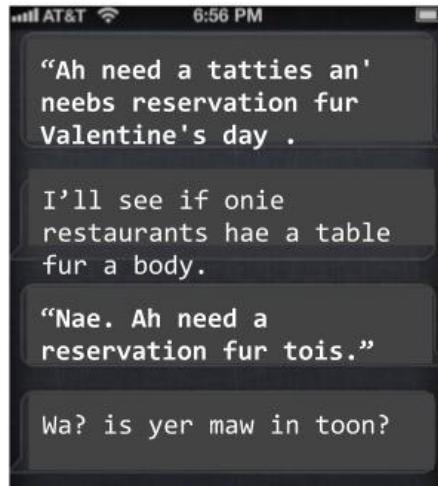


Swahili  
100 million speakers

# Thousands of language varieties



American English

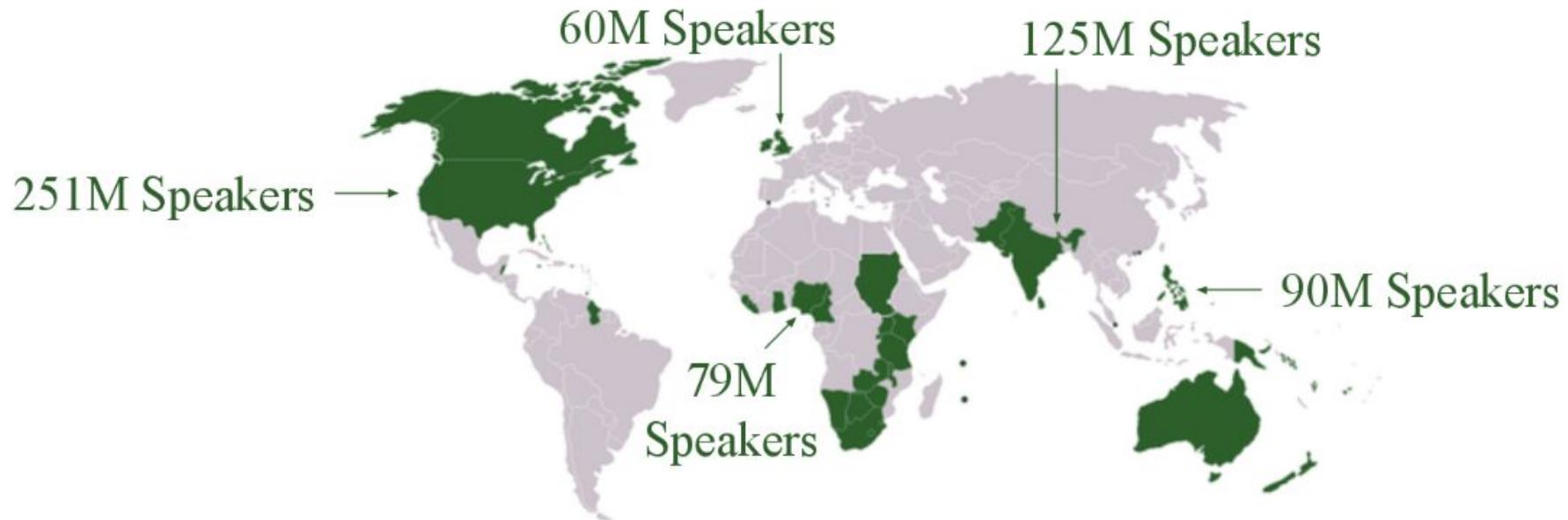


Scottish English



Hinglish

# World Englishes





The Royal Family

@RoyalFamily

Follow



da'Rah-zingSun

@TIME7SS

Follow

Taking place this week on the river Thames is 'Swan Upping' – the annual census of the swan population on the Thames.



@kimguilfoyle prblm I hve wit ur reportng is its 2 literal, evry1 knos pple tlk diffrrnt evrywhere, u kno wut she means jus like we do!



Mooktar

@bossmukky

Follow



Ebenezer·

@Physique\_cian

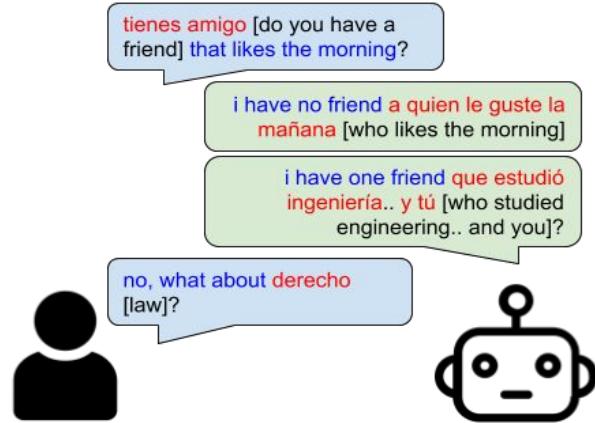
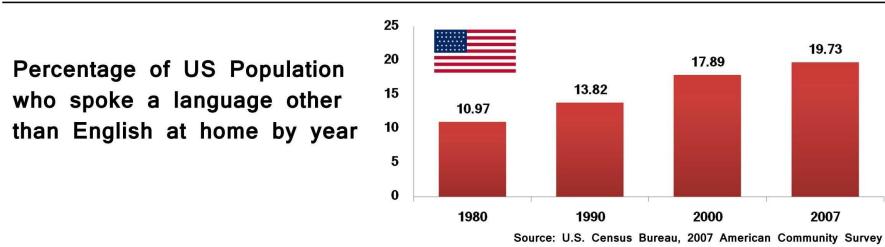
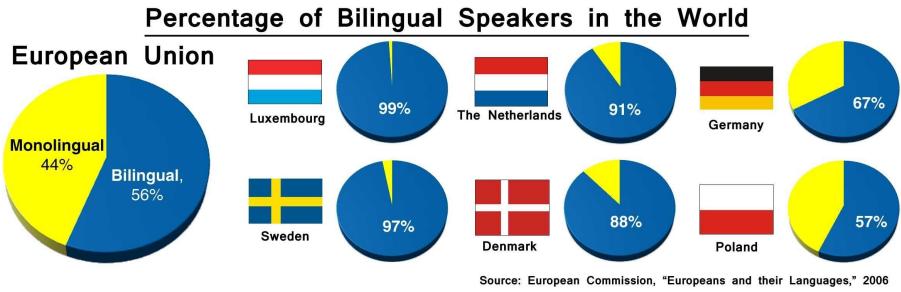
Follow

"@Ecstatic\_Mi: @bossmukky Ebi like say I wan dey sick sef wlh 'Flu' my whole body dey weak"uw gee...

@Tblazeen R u a wizard or wat gan sef : in d mornin- u tweet, afternoon - u tweet, nyt gan u dey tweet.beta get ur IT placement wiv twitter

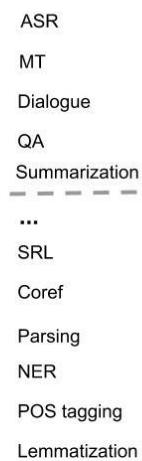


# Code-switching



Source: US Census Bureau

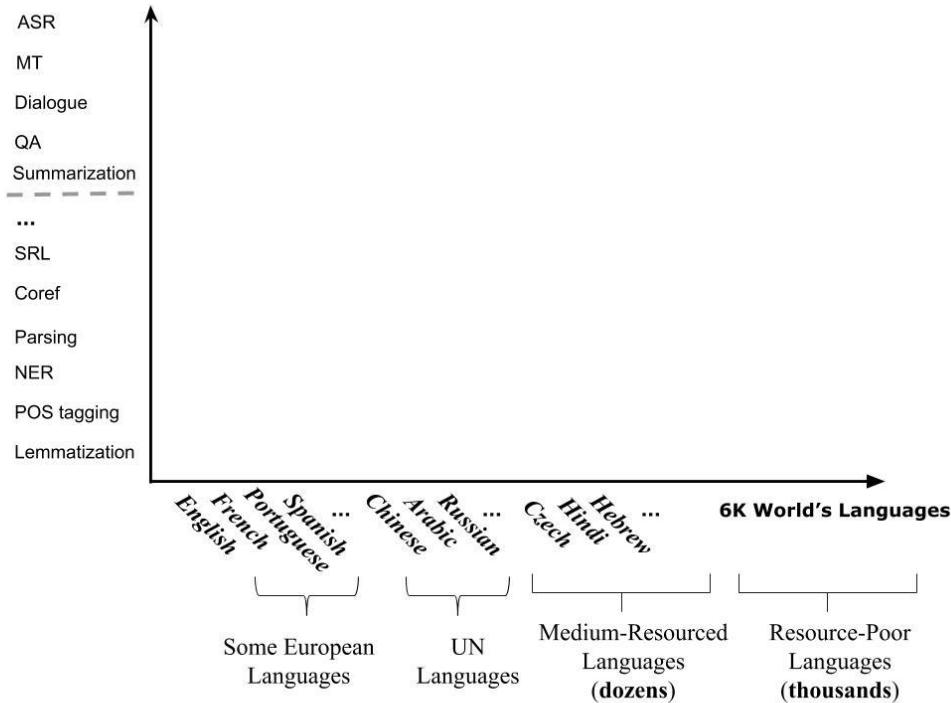
## NLP Technologies/Applications

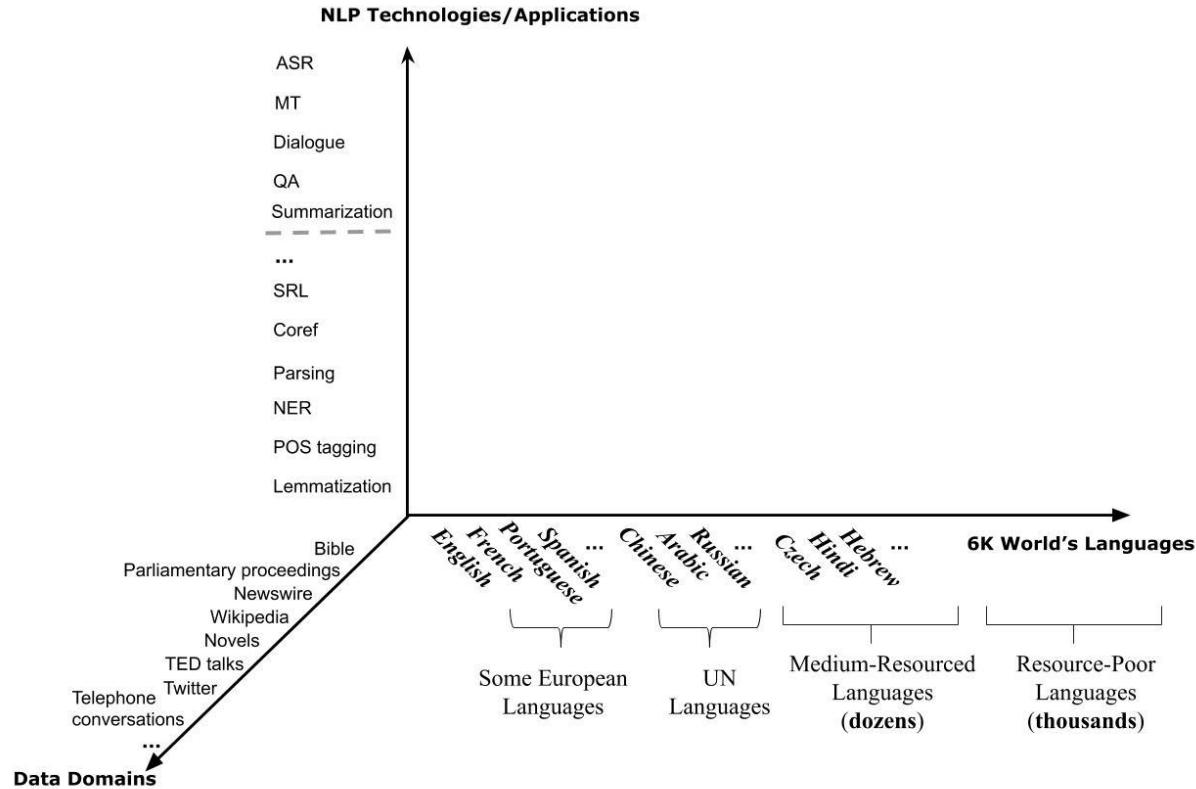


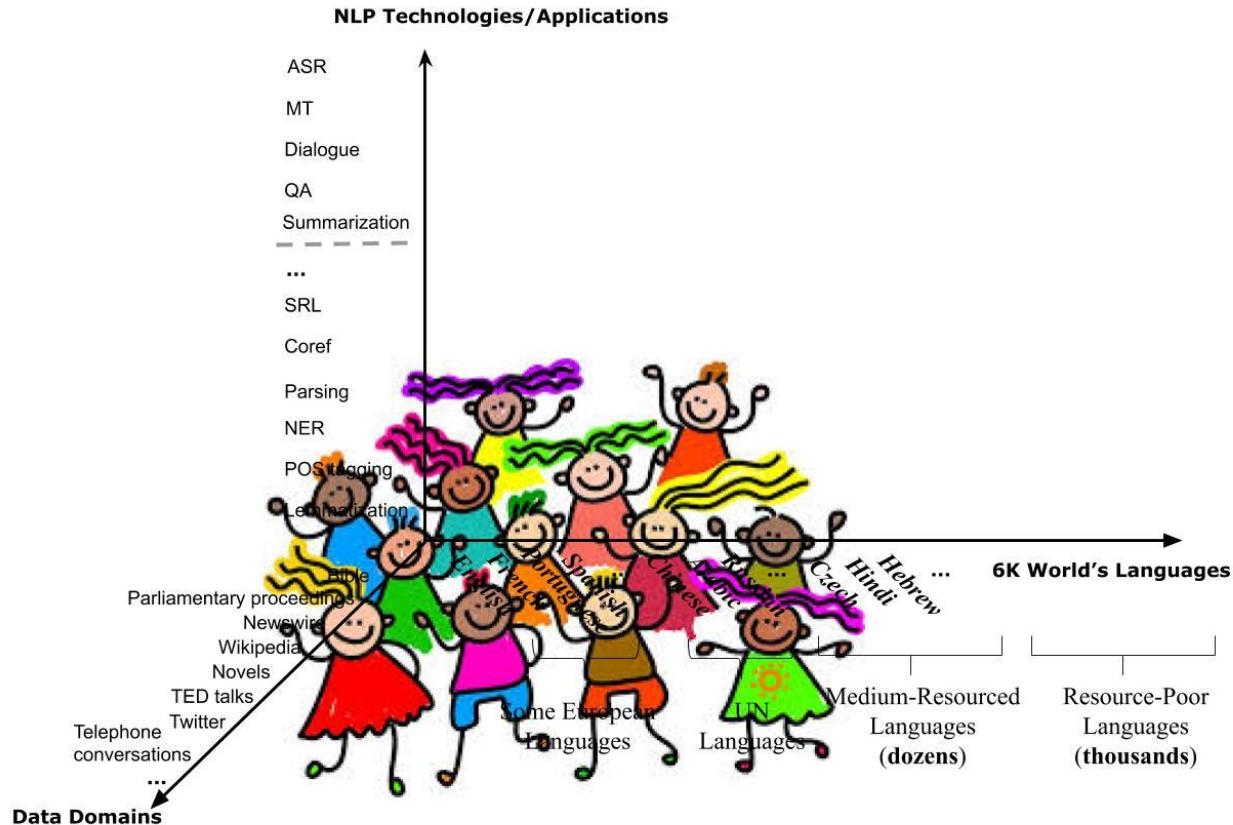
ASR  
MT  
Dialogue  
QA  
Summarization  
...  
SRL  
Coref  
Parsing  
NER  
POS tagging  
Lemmatization



### NLP Technologies/Applications







# Low-resource NLP

Languages and problems lacking large monolingual or parallel corpora and/or manually crafted linguistic resources sufficient for building statistical NLP tools



# Resource-rich vs. resource-poor NLP: machine translation

- Training data for machine translation: Parallel corpus

Nenhum deles reparou na janela , através da qual teria podido ver uma enorme coruja amarelada , esvoaçando em grande alvoroço .

assim , não viu as corujas descondo rapidamente em plena luz do dia , apesar de todos os transeuntes apontarem estarrecidos e de boca aberta enquanto coruja após coruja lhes passavam A grande velocidade sobre as cabeças .

Queira enviar-nos A sua coruja até dia 31 de Julho , sem falta .

- O que é que quer dizer esperarem A minha coruja ?

Hagrid Hagrid enrolou A nota , deu-a à coruja que A agarrou com O bico e , dirigindo-se à porta , soltou A ave no meio da tempestade .

O próprio Hagrid adormecera no sofá totalmente destruído e , bicando no vidro da janela , estava uma coruja que segurava um jornal .

A coruja entrou e depôs O jornal em cima de Hagrid

None of them noticed a large , tawny owl flutter past the window .

He didn ' t see the owls swoop ing past in broad daylight , though people down in the street did ; They pointed and gazed open-mouthed as owl after owl sped overhead .

We await your owl by no later than July 31 .

after a few minutes He stammered , " what does it mean , They await My owl ?

Hagrid Hagrid rolled up the note , gave it to the owl , which clamped it in its beak , went to the door , and threw the owl out into the Storm .

the hut was full of sunlight , the Storm was over , Hagrid himself was asleep on the collapsed sofa , and there was an owl rapping its claw on the window , a newspaper held in its beak .

the owl swooped in and dropped the newspaper on top of Hagrid , who didn ' t

- Resource-rich language pair - few millions of parallel sentences

Resource-poor language pair - few thousands of parallel sentences



# Resource-rich vs. resource-poor NLP: machine translation

English → French

Translate

Turn off instant translation 

Russian English French Detect language  English Spanish French  Translate

You will just have to find a way of getting over it.  Vous devrez trouver un moyen de le surmonter.

   52/5000    Suggest an edit

French → English

Translate

Turn off instant translation 

Russian English French Detect language  English Spanish French  Translate

Vous devrez trouver un moyen de le surmonter.  You will have to find a way to overcome it.

   45/5000    Suggest an edit

Did you mean: Vous **devez** trouver un moyen de le surmonter.



# Resource-rich vs. resource-poor NLP: machine translation

English → Swahili

Translate Turn off instant translation 

Russian English French Detect language  English Swahili French 

You will just have to find a way of getting over it.  Utakuwa tu kupata njia ya kupata juu yake.

   53/5000     Suggest an edit

Swahili → English

Translate Turn off instant translation 

Swahili English French Detect language  English Swahili French 

Utakuwa tu kupata njia ya kupata juu yake.  You will just find the way to get on it.

   42/5000     Suggest an edit



# Resource-rich vs. resource-poor NLP: machine translation

English → Hindi → English

The screenshot shows the Google Translate interface with three language tabs: Hindi, English, and Yoruba. The "Detect language" tab is also present. The input field contains the Hindi sentence "आपको इसे खत्म करने का एक तरीका मिलना होगा।". The output field shows the English translation "You have to find a way to eliminate it." Below the input field, there are small icons for text direction, font size, and a dropdown menu. The character count is 42/5000. On the right side, there are edit and suggestion options.

English → Telugu → English

The screenshot shows the Google Translate interface with three language tabs: Uzbek, English, and Telugu. The "Detect language" tab is also present. The input field contains the Telugu sentence "మీరు దాని ప్రైకి రావడానికి ఒక మార్గాన్ని కనుగొనవలని ఉంటుంది." The output field shows the English translation "You have to find a way to get it up." Below the input field, there are small icons for text direction, font size, and a dropdown menu. The character count is 59/5000. On the right side, there are edit and suggestion options.

English → Uzbek → English

The screenshot shows the Google Translate interface with three language tabs: Pashto, English, and Uzbek. The "Detect language" tab is also present. The input field contains the Uzbek sentence "Buning ustiga faqatgina bir usulni topish kerak." The output field shows the English translation "On top of that, you just have to find a way out." Below the input field, there are small icons for text direction, font size, and a dropdown menu. The character count is 48/5000. On the right side, there are edit and suggestion options.



# Resource-rich vs. resource-poor NLP: machine translation

English → Swahili

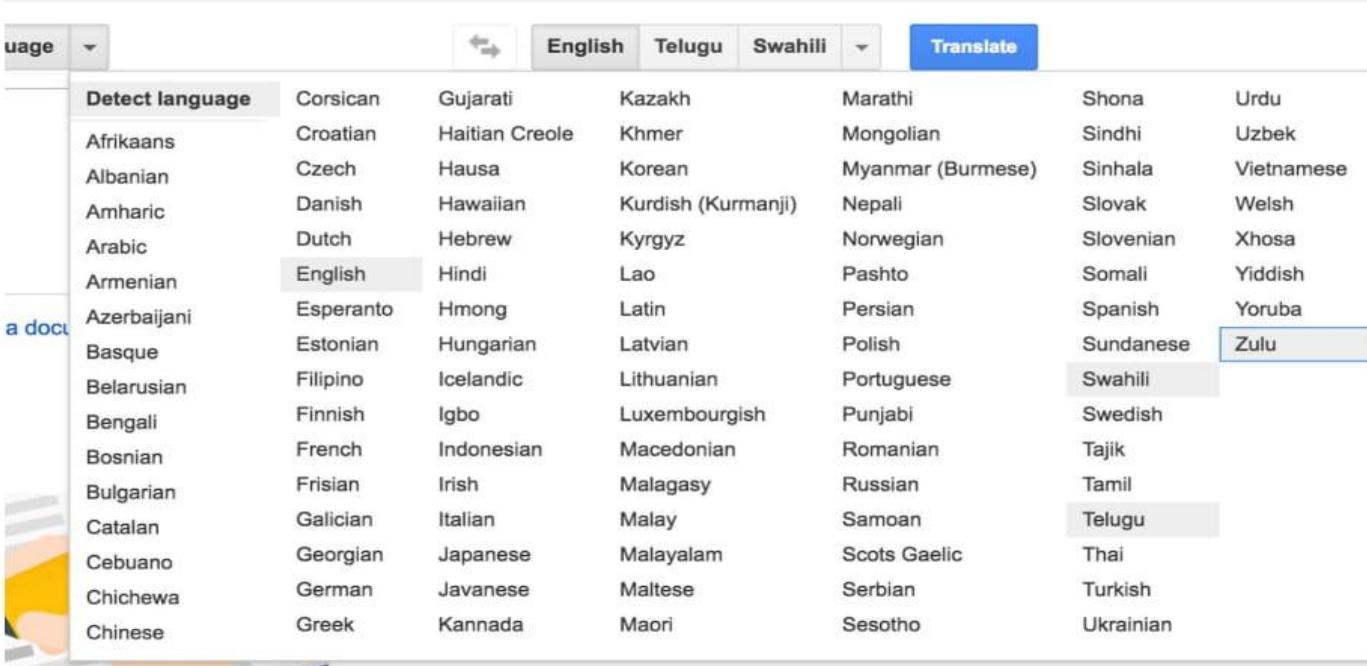
The screenshot shows a machine translation interface with two panels. The left panel has a "Swahili" button highlighted, followed by "English", "Telugu", and "Detect language". The right panel has "English" and "Swahili" buttons highlighted, followed by "Telugu" and a "Translate" button. Both panels have a "Detected language" dropdown set to "Swahili". The English input is: "The summer school is meant to be an introduction to the state-of-the-art research in the speech and language technology area for graduate and undergraduate students." The Swahili output is: "Shule ya majira ya joto ina maana ya kuanzishwa kwa utafiti wa hali ya sanaa katika eneo la teknolojia na lugha ya wanafunzi kwa wanafunzi wahitimu na wahitimu." Below the input and output are toolbar icons and character counts (166/5000 and 160/5000).

Swahili → English

The screenshot shows a machine translation interface with two panels. The left panel has a "Swahili" button highlighted, followed by "English", "Telugu", and "Detect language". The right panel has "English" and "Swahili" buttons highlighted, followed by "Telugu" and a "Translate" button. Both panels have a "Detected language" dropdown set to "English". The Swahili input is: "Shule ya majira ya joto ina maana ya kuanzishwa kwa utafiti wa hali ya sanaa katika eneo la teknolojia na lugha ya wanafunzi kwa wanafunzi wahitimu na wahitimu." The English output is: "Summer school means the establishment of a state-of-the-art arts research technology and pupil language for graduate students and graduates." Below the input and output are toolbar icons and character counts (160/5000 and 160/5000).



# Resource-rich vs. resource-poor NLP: machine translation

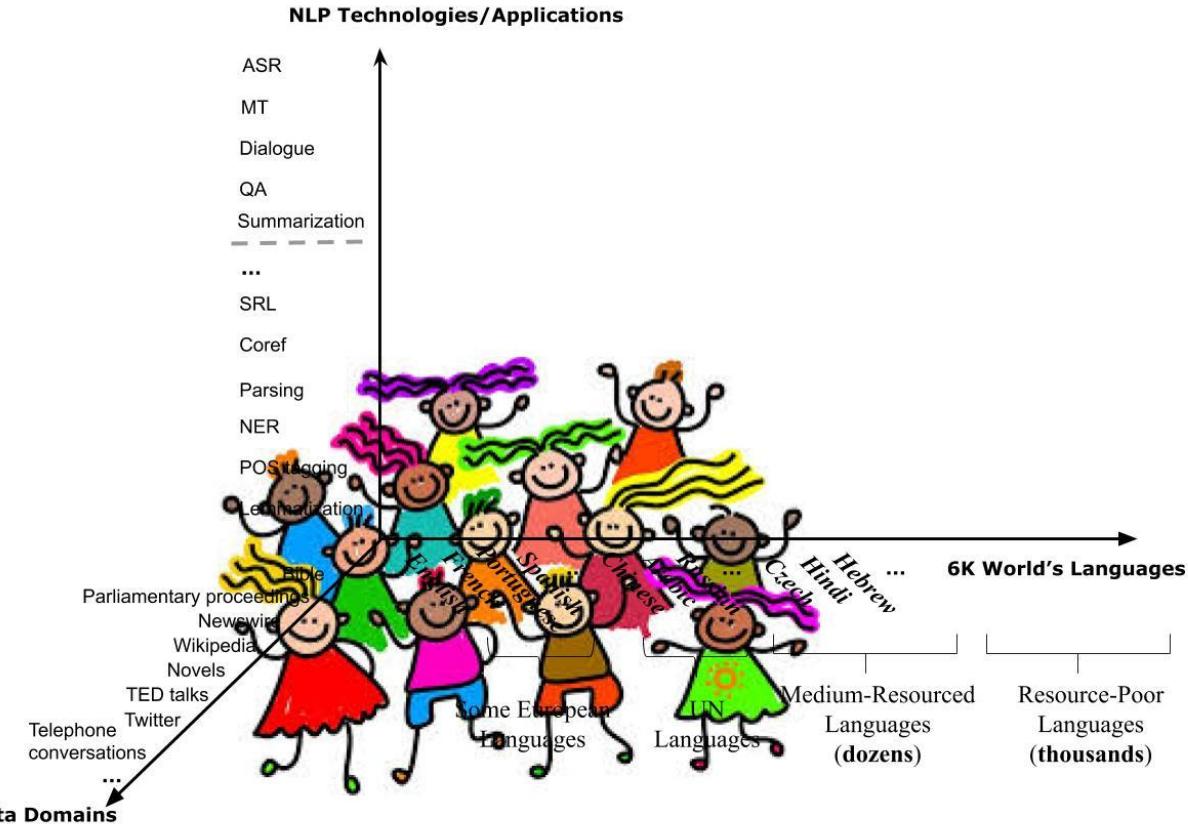


The screenshot shows a user interface for machine translation. At the top, there is a navigation bar with a dropdown menu labeled "Usage", a double-headed arrow icon, and three language selection buttons: "English", "Telugu", and "Swahili". To the right of these buttons is a dropdown menu and a blue "Translate" button. Below this header is a large table containing a grid of language names.

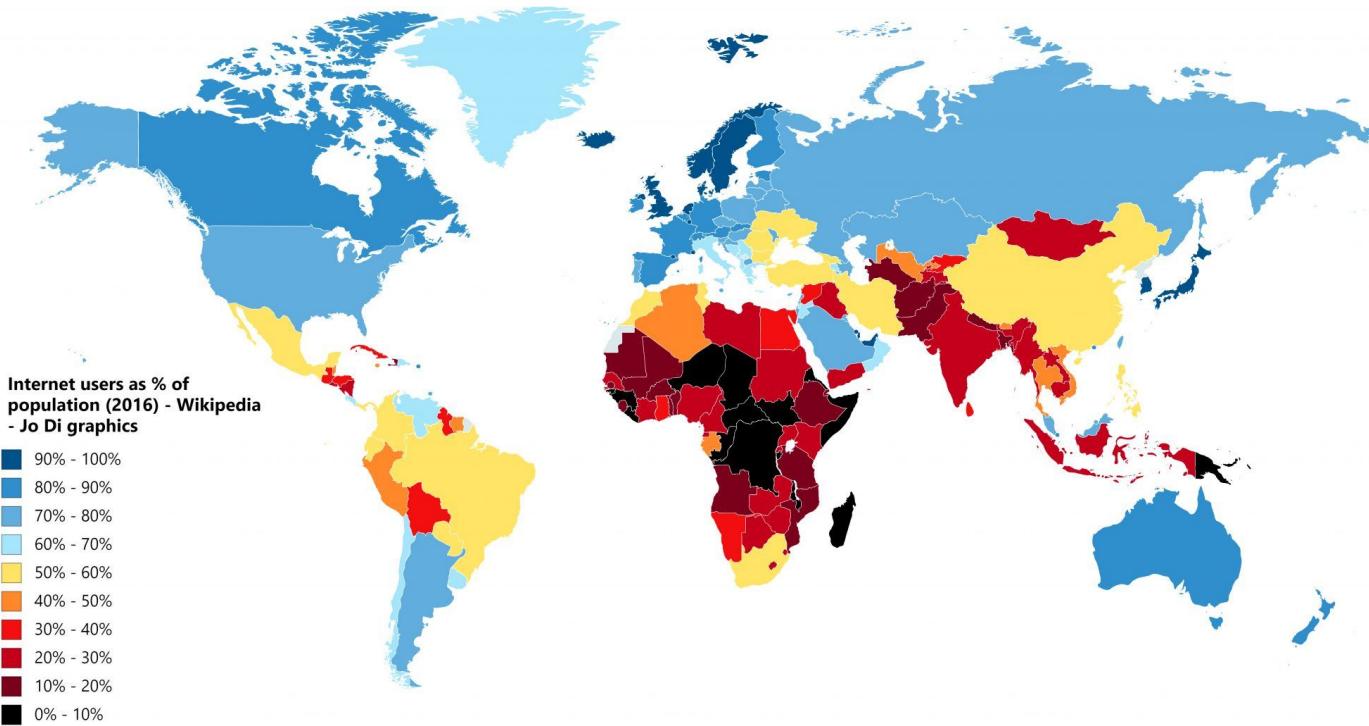
Language	Corsican	Gujarati	Kazakh	Marathi	Shona	Urdu
Afrikaans	Croatian	Haitian Creole	Khmer	Mongolian	Sindhi	Uzbek
Albanian	Czech	Hausa	Korean	Myanmar (Burmese)	Sinhala	Vietnamese
Amharic	Danish	Hawaiian	Kurdish (Kurmanji)	Nepali	Slovak	Welsh
Arabic	Dutch	Hebrew	Kyrgyz	Norwegian	Slovenian	Xhosa
Armenian	English	Hindi	Lao	Pashto	Somali	Yiddish
Azerbaijani	Esperanto	Hmong	Latin	Persian	Spanish	Yoruba
Basque	Estonian	Hungarian	Latvian	Polish	Sundanese	Zulu
Belarusian	Filipino	Icelandic	Lithuanian	Portuguese	Swahili	
Bengali	Finnish	Igbo	Luxembourgish	Punjabi	Swedish	
Bosnian	French	Indonesian	Macedonian	Romanian	Tajik	
Bulgarian	Frisian	Irish	Malagasy	Russian	Tamil	
Catalan	Galician	Italian	Malay	Samoan	Telugu	
Cebuano	Georgian	Japanese	Malayalam	Scots Gaelic	Thai	
Chichewa	German	Javanese	Maltese	Serbian	Turkish	
Chinese	Greek	Kannada	Maori	Sesotho	Ukrainian	



# Why low-resource/multilingual NLP?



# Low-resource/multilingual NLP



<https://jodi.graphics/2018/05/11/internet-users-as-of-population/>

40% of world's population: South Asia - 1.75 billion, Africa - 1.3 billion, etc.



# Who are future users of low-resource/multilingual NLP?

Africa is a continent with a very high linguistic diversity:  
there are an estimated 1.5-2K African languages from 6 language families.  
**1.33 billion people**



# Who are future users of low-resource/multilingual NLP?

There are about 460 languages in India.

1.38 billion people



# Why working on low-resource/multilingual NLP is important?

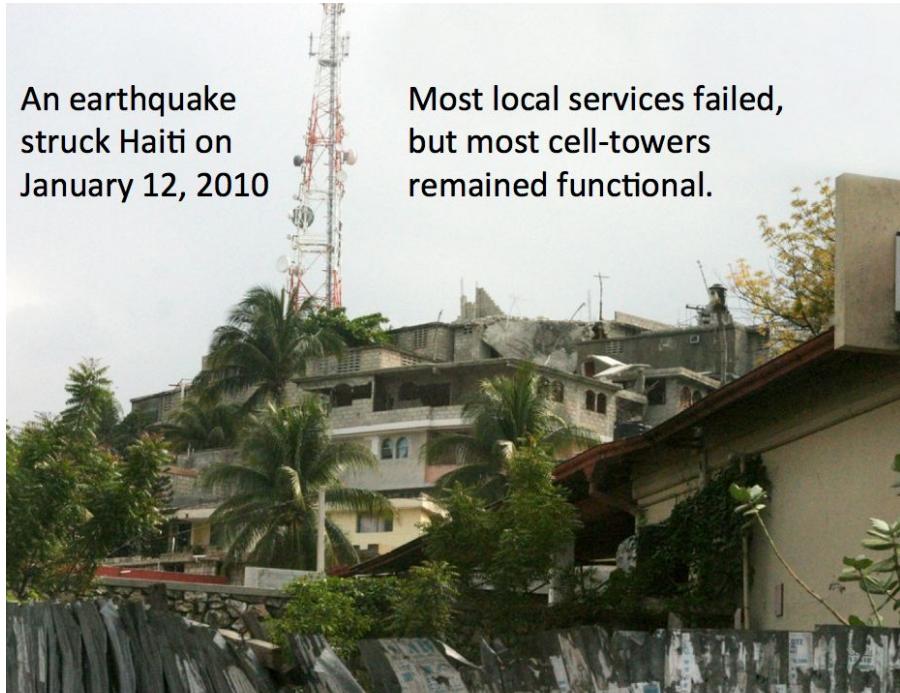
NLP provides opportunities and access to:

- Translation systems, Speech interfaces, Dialogue systems
- Educational applications
- Emergency response applications
- Online communities and communication
- Journalism, monitoring democratic processes
- Opportunity to save and nurture language diversity
- Documenting cultural history



# Example: emergency response applications

- About 3 million people were affected by the quake in Haiti



Example and slides by  
Rob Munro

# Example: emergency response applications

## Messages start streaming in

- Fanmi mwen nan  
Kafou, 24 Cote Plage,  
41A bezwen manje  
ak dlo
- Moun kwense nan  
Sakre Kè nan  
Pòtoprens
- Ti ekipman Lopital  
General genyen yo  
paka minm fè 24 è
- Fanm gen tranche  
pou fè yon pitit nan  
Delmas 31

iDIBON



# Example: emergency response applications

## Messages start streaming in

- Fanmi mwen nan Kafou, 24 Cote Plage, 41A bezwen manje ak dlo
- Moun kwense nan Sakre Kè nan Pòtoprens
- Ti ekipman Lopital General genyen yo paka minm fè 24 è
- Fanm gen tranche pou fè yon pitit nan Delmas 31
- My family in Carrefour, 24 Cote Plage, 41A needs food and water
- People trapped in Sacred Heart Church, PauP
- General Hospital has less than 24 hrs. supplies
- Undergoing children delivery Delmas 31

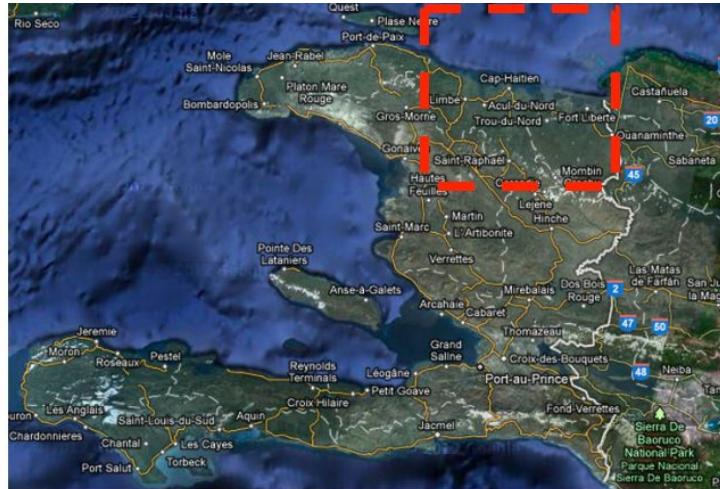
iDIBON



# Example: emergency response applications

Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

“Sacre-Coeur Hospital which located in this village of **Okap** is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.”



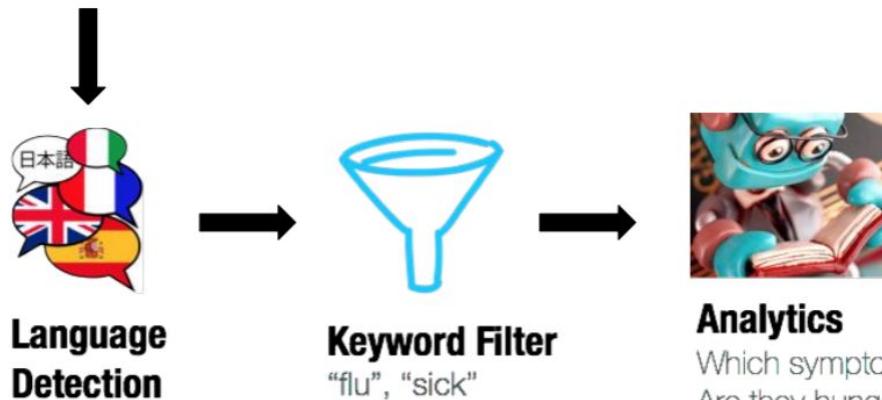
idibon

# Example: Identifying disease outbreaks

 Brooke  
@Brooklepooc134

Follow

got the flu over the weekend and I didn't know until today, & I somehow managed to give it to FIVE of my friends!!!!!!

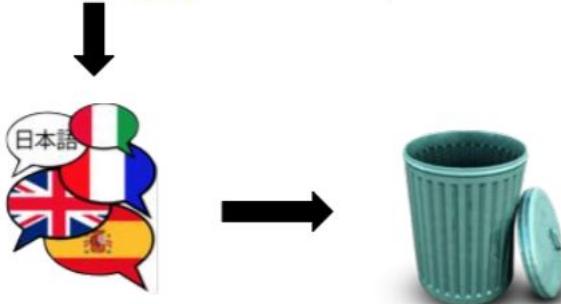


Slides on LID by  
David Jurgens  
(Jurgens et al. ACL'17)

# Example: Identifying disease outbreaks

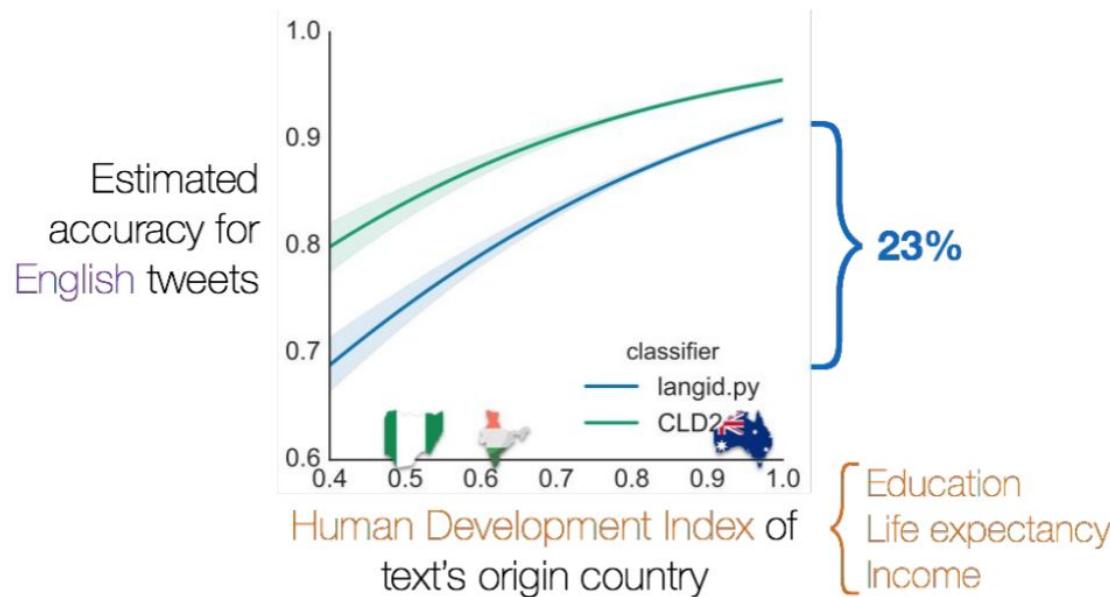
Nana Rayne  
@Nana\_Rayne  
Like serious dis flu nor dey wan go oooo.... Sick

Venus  
@christinedarvin  
 @\_rkptrnte hindi ko alam babe eh, absent ako kanina I'm sick rn hahaha 😊💪

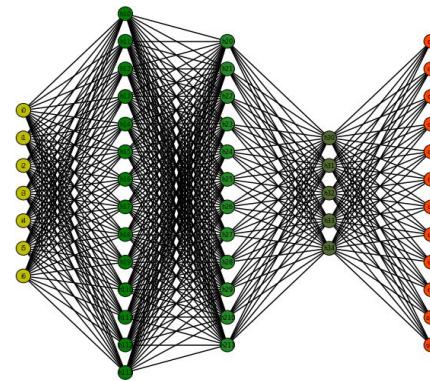


**Language  
Detection**

# Example: Identifying disease outbreaks



# Why standard techniques used in NLP cannot simply be applied to low-resource languages?



# Why standard techniques used in NLP cannot simply be applied to low-resource languages?

- State-of-the-art NLP models require large amounts of training data and/or sophisticated language-specific engineering
- Large amounts of training data are unavailable for most languages
  - an extreme case is languages that don't have a written form, e.g. Shanghaiese spoken by 14 million people
  - or languages that just don't have online presence, e.g. Chichewa, a Bantu language spoken by 12 million people
- Language-specific engineering is expensive, requires linguistically trained speakers of the language

# Build a NLP application for these languages:

×

×

×

×

×

×

×

×

×

×

×

×

×

×

×

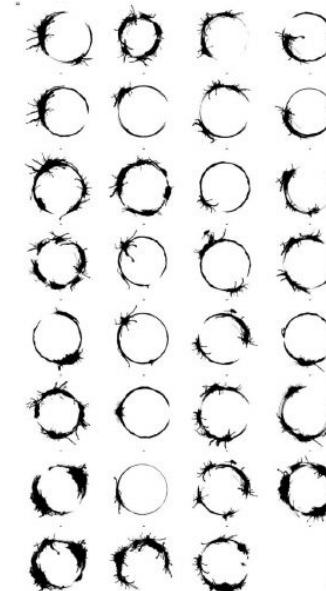
×

×

×

×

•



# What language is it?

[www.ethnologue.com](http://www.ethnologue.com), <http://wals.info>

- Words
  - Does it require segmentation?
  - Is it a phonographic language?
- Morphology
  - Does it require lemmatization?
  - Does it require splitting tokens to bound morphemes?
- Syntax
  - How sentences are structures?
- Typology
  - What other languages are similar to this one?



# Are there sufficient resources?

Is there training data for the task?

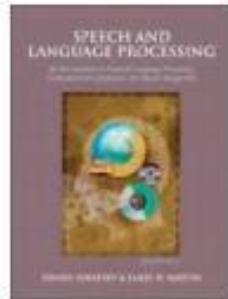


[corpora-request@uib.no](mailto:corpora-request@uib.no)



# Are there sufficient training data for the task?

- Use supervised approaches



ACL  
NAACL  
EMNLP

# Can we use resources from other languages?

- Are there relevant training data from other languages?
- If yes, build cross-lingual bridges
- Use transfer learning or joint multilingual approaches



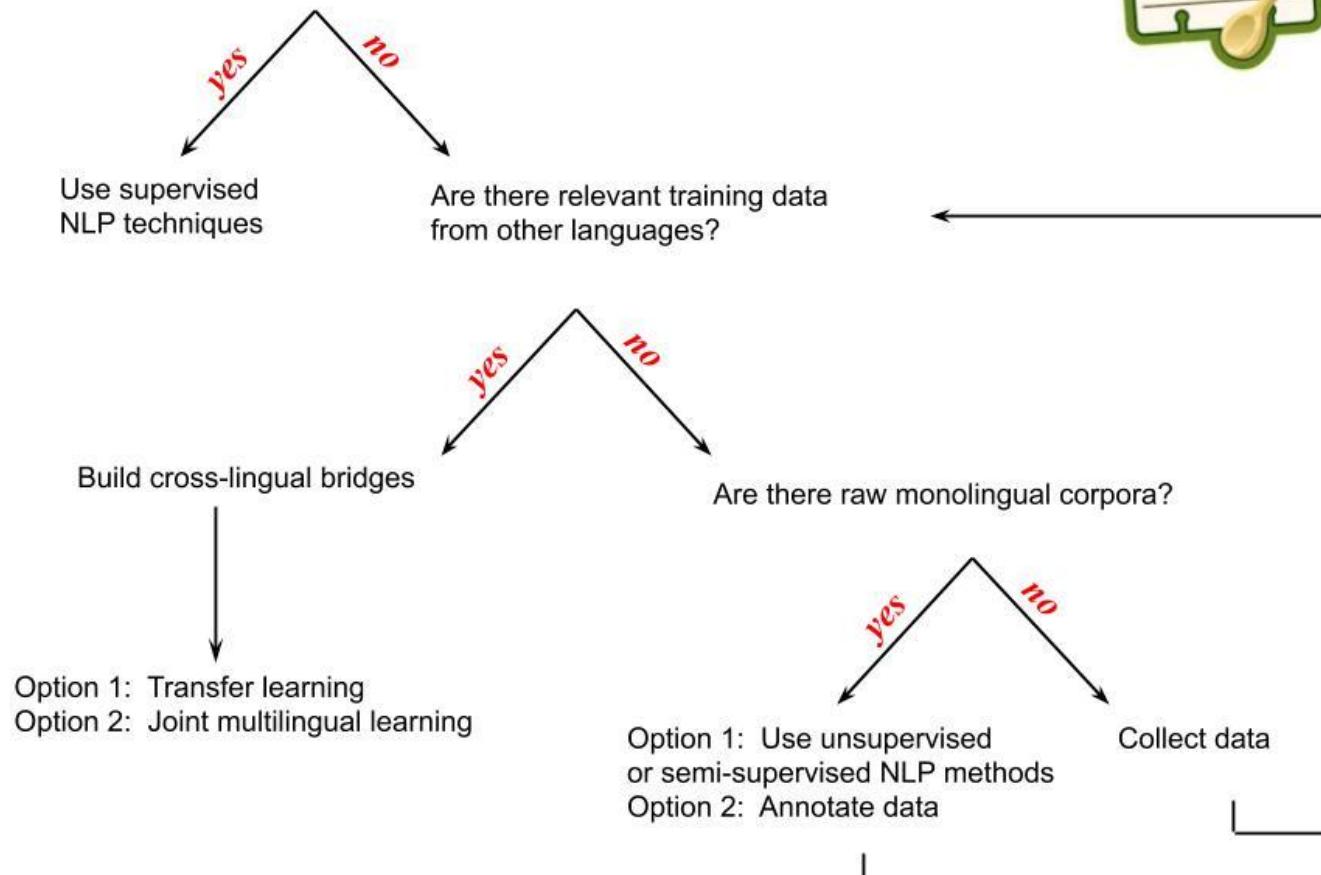
# Are there sufficient raw monolingual corpora?

- Use unsupervised approaches





Are there (sufficient) training data for the task?



# Exercise today: multilingual sentiment analysis



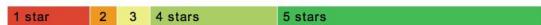
HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner

\$89 online, \$100 nearby ★★★★☆ 377 reviews

September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 shi

## Reviews

Summary - Based on 377 reviews



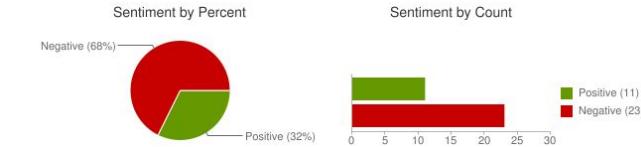
### What people are saying

ease of use	■	"This was very easy to setup to four computers."
value	■	"Appreciate good quality at a fair price."
setup	■	"Overall pretty easy setup."
customer service	■	"I DO like honest tech support people."
size	■	"Pretty Paper weight."
mode	■	"Photos were fair on the high quality mode."
colors	■	"Full color prints came out with great quality."

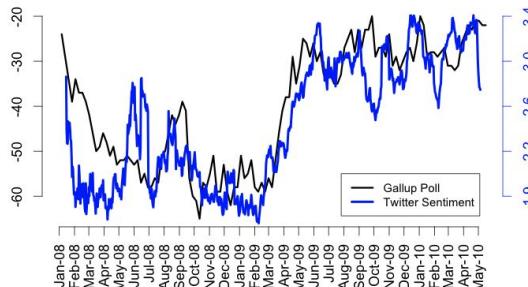
Type in a word and we'll highlight the good and the bad

"united airlines"  [Save this search](#)

### Sentiment analysis for "united airlines"



window = 15, r = 0.804



# Positive or negative movie review?



- unbelievably disappointing



- Full of zany characters and richly applied satire, and some great plot twists



- this is the greatest screwball comedy ever filmed



- It was pathetic. The worst part about it was the boxing scenes.

# Exercise



Chan Young Park

[chanyoun@andrew.cmu.edu](mailto:chanyoun@andrew.cmu.edu)

A large word cloud centered on the word "thank you" in various languages. The word "thank you" is repeated in many different scripts and colors, including red, green, blue, and yellow. Surrounding it are numerous other words, mostly in Asian scripts, representing the same concept in different languages. The background is white, and the overall effect is a dense, colorful collage of language.

Language