Multilingual Natural Language Processing

Yulia Tsvetkov & Chan Young Park

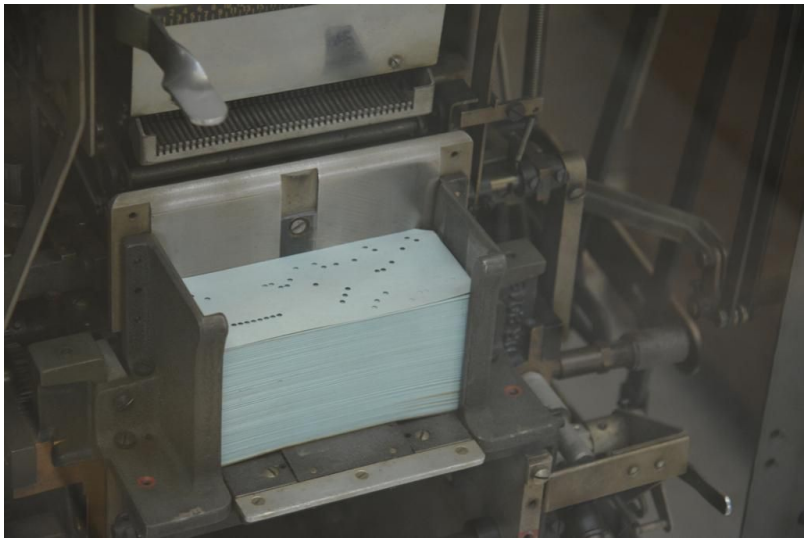
<ytsvetko@cs.cmu.edu> <chanyoun@andrew.cmu.edu>

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

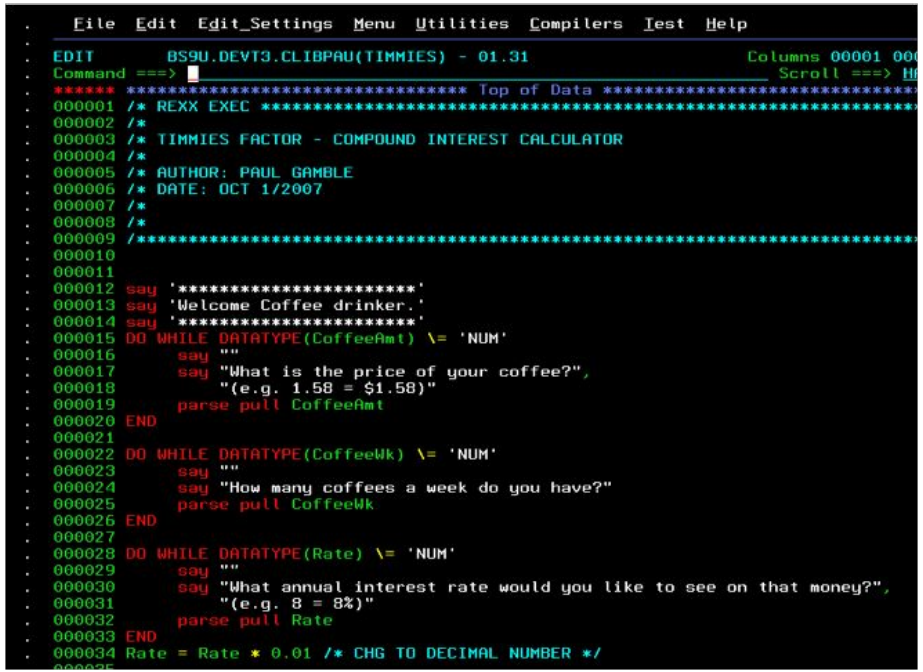
Communication with machines

- ~50s-70s



Communication with machines

- ~80s



Communication with machines

- Today

Language technologies: conversational agents

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

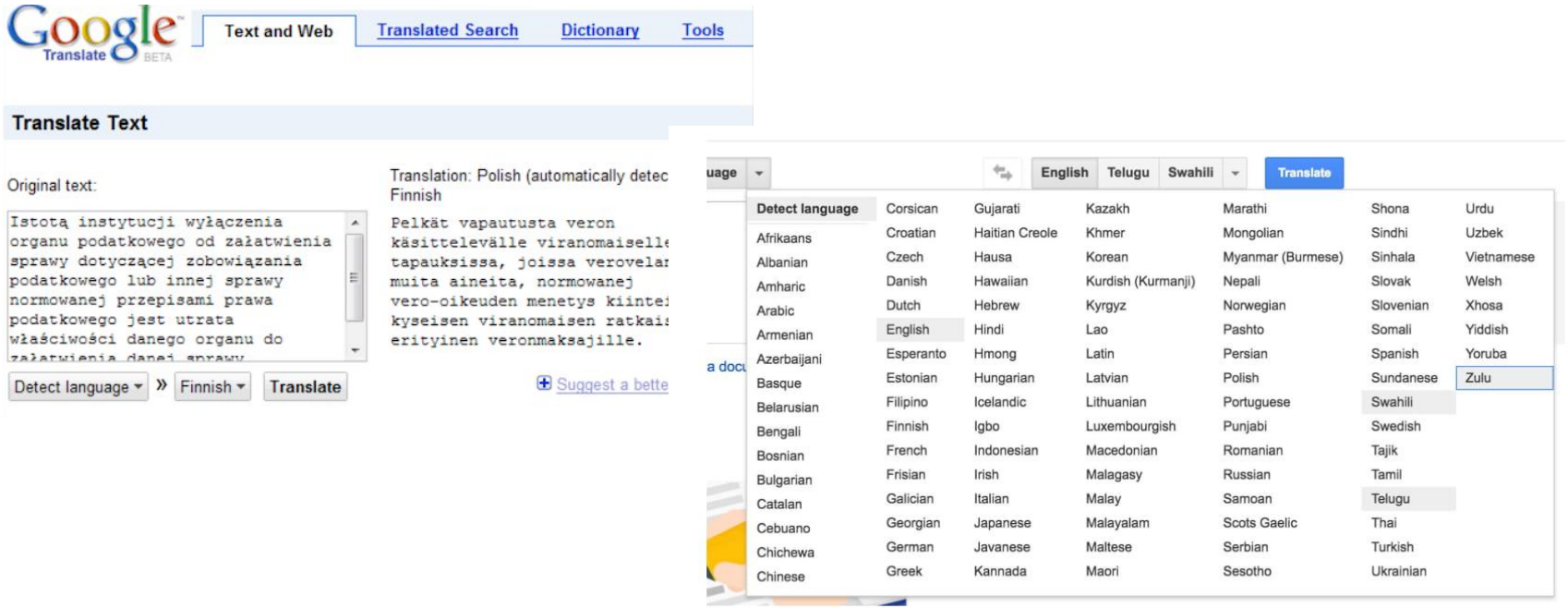
Natural Language Processing

- 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

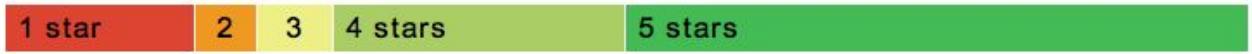


Sentiment Analysis

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	"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."

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

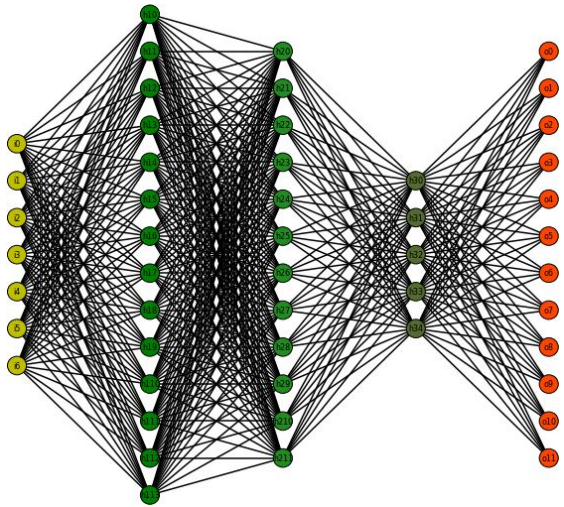
People also ask

What is Abstractive text summarization?

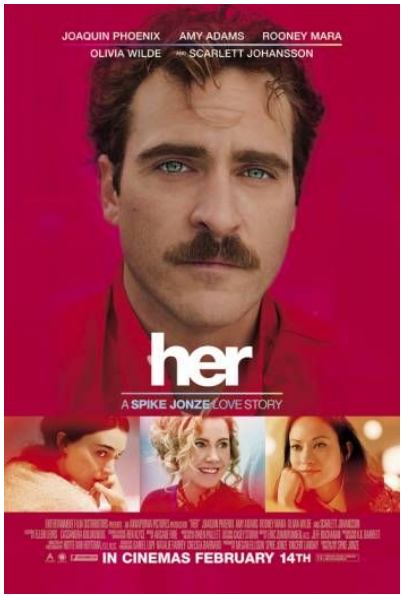
Why do we need text summarization?


21:25
73 %
“ I need a dinner reservation for Valentine's Day. ”
I'll see if any restaurants have a table for one.
“ No, I need a reservation for two. ”
Why? Is your mother in town?


21:25
73 %
“I need a dinner reservation for Valentine's Day.”
I'll see if any restaurants have a table for one.
“No, I need a reservation for two.”
Why? Is your mother in town?



Still, NLP is not solved yet

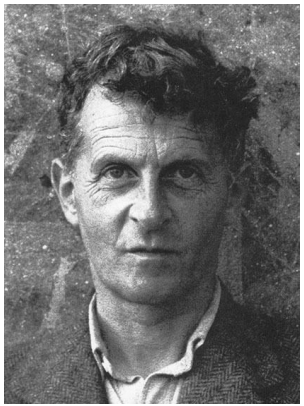


Baseline mutual information model (Li et al. 2015)
A: Where are you going? (1)
B: I'm going to the restroom. (2)
A: See you later. (3)
B: See you later. (4)
A: See you later. (5)
B: See you later. (6)
...
...
A: how old are you? (1)
B: I'm 16. (2)
A: 16? (3)
B: I don't know what you are talking about. (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?

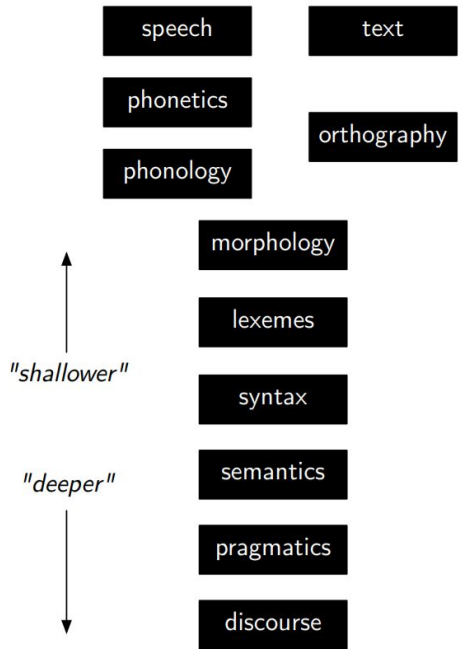
"the limits of my language are the limits of my world"



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!



Drawing by Noah Smith

Phonetics, phonology

- Pronunciation modeling

SOUNDS	Th	i	a	si	e	n

Words

- Language modeling
- Tokenization
- Spelling correction

Morphology

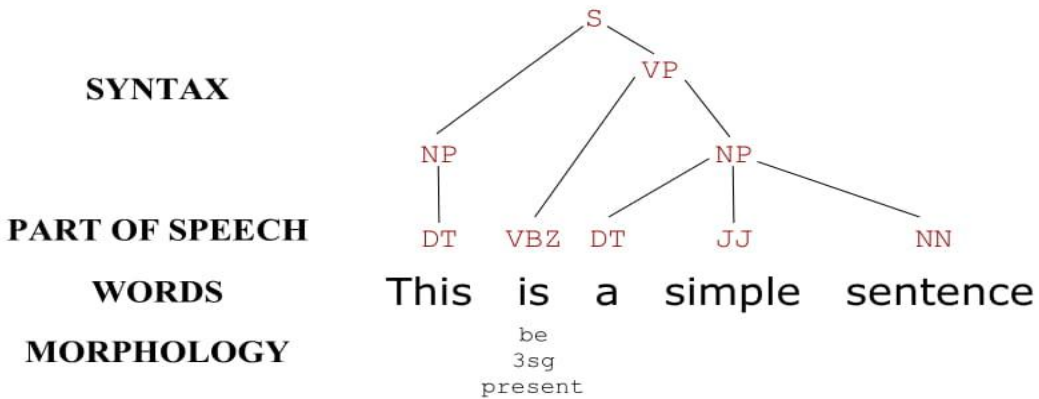
- Morphological analysis
- Tokenization
- Lemmatization

Syntax

- Part-of-speech tagging

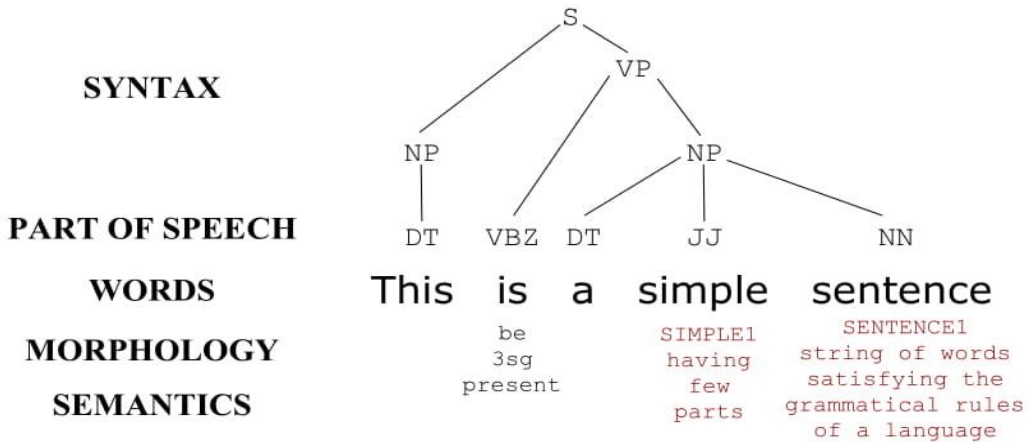
Syntax

- Syntactic parsing



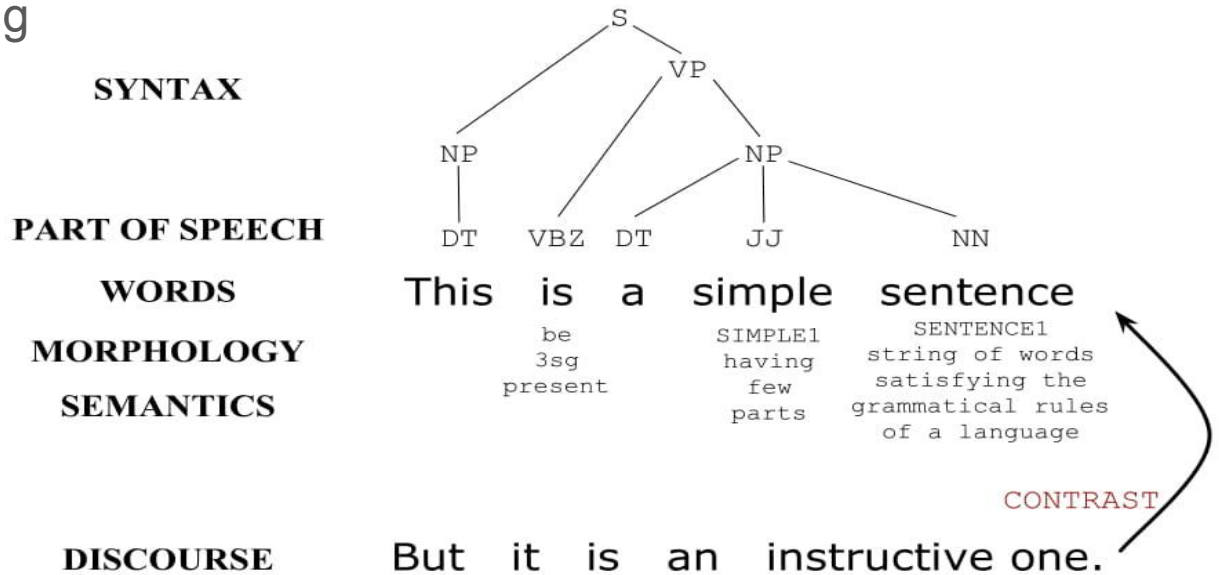
Semantics

- 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**

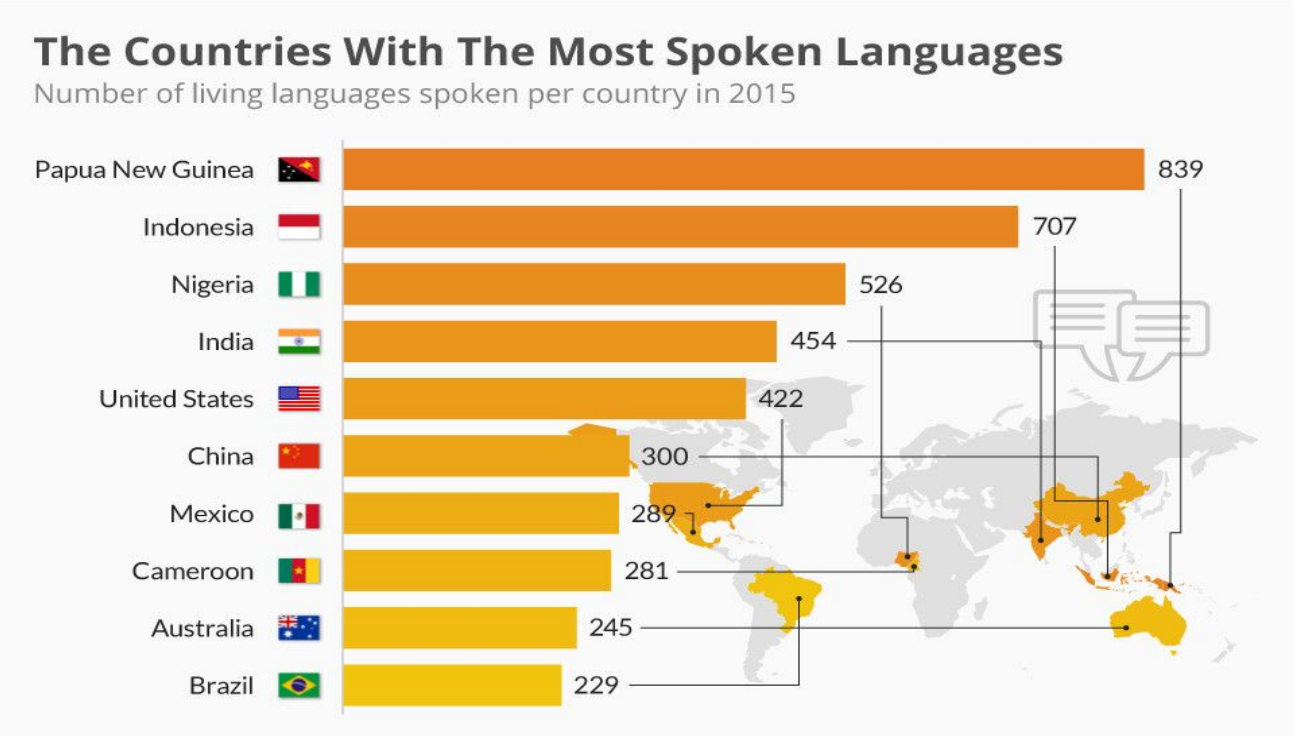
Why NLP is hard?

Ambiguity at multiple levels

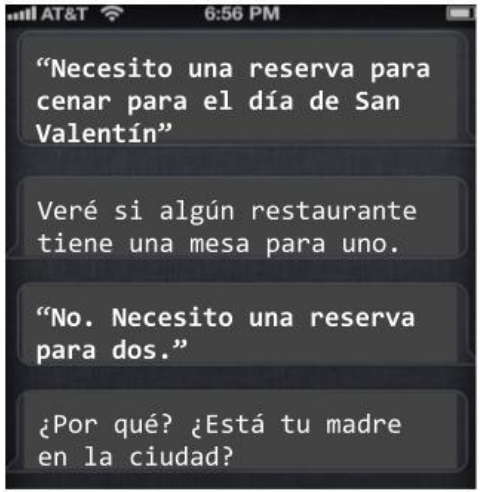
Linguistic diversity

Linguistic diversity: ~6,000 languages

The multilingual world

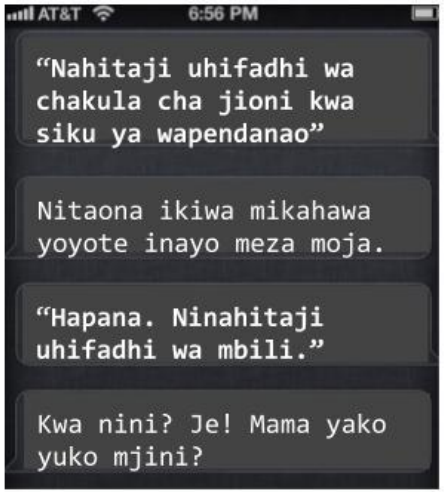


Source: Ethnologue



Spanish
534 million speakers

Hindi
615 million speakers



Swahili
100 million speakers

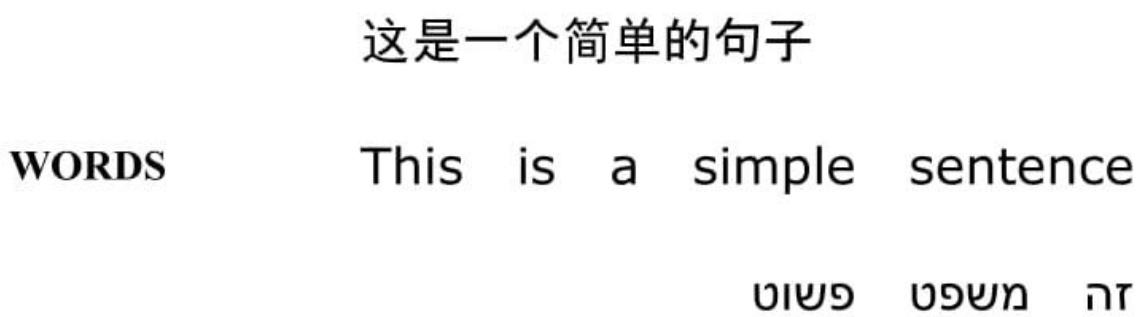
Language diversity: language families

www. enthologue.com

- Niger-Congo (1538 languages) (20.6%)
- Austronesian (1257 languages) (16.8%)
- Trains-New Guinea (480 languages) (6.4%)
- Sino-Tibetian (457 languages) (6.1%)
- Indo-European (444 languages) (5.9%)
- Australian (378 languages) (5.1%)
- Afro-Asiatic (375 languages) (5.0%)
- Nilo-Saharan (205 languauges) (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



Linguistic diversity: Hebrew words

- Most vowels are left unspecified

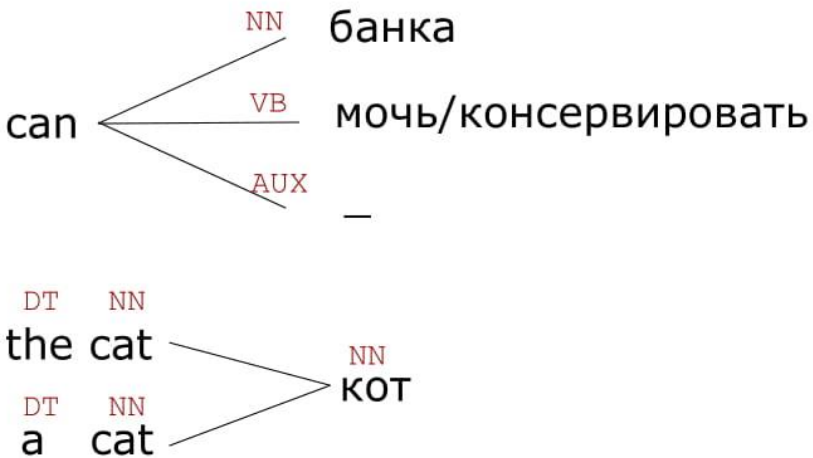
in tea her daughter	בתה

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	בתה
in the tea	בהתה
that in tea	שבתה
that in the tea	שבהתה
and that in the tea	ושבהתה
	ושבתה
and her saturday	ו+שבת+ה
and that in tea	ו+ש+ב+תה
and that her daughter	ו+ש+בת+ה

Linguistic diversity: parts of speech



Linguistic diversity: English morphology

Linguistic diversity: English morphology

unfriend, Obamacare

Linguistic diversity: Russian morphology

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

Linguistic diversity: Quechua morphology

Much'ananayakapushasqakupuniñataqsunamá

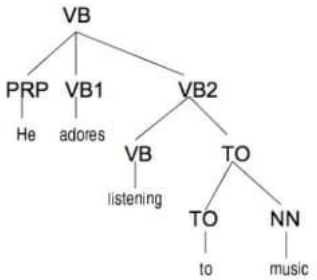
Much'a -na -naya -ka -pu -sha -sqa -ku -puni -ña -taq -suna -má

"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

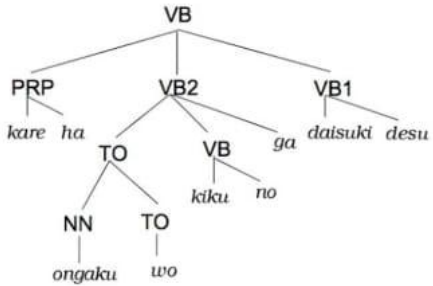
Linguistic diversity: syntax

SVO



he adores listening to music

SOV

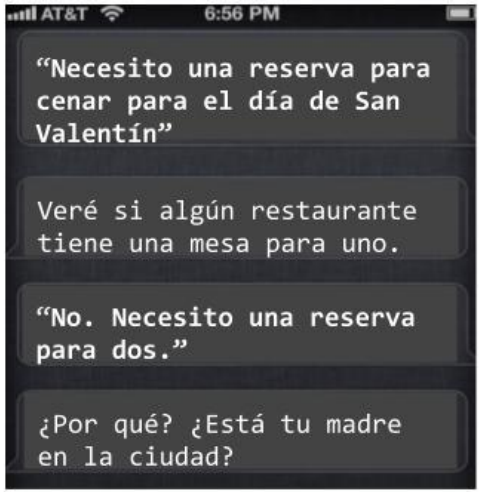


Linguistic diversity: semantics

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

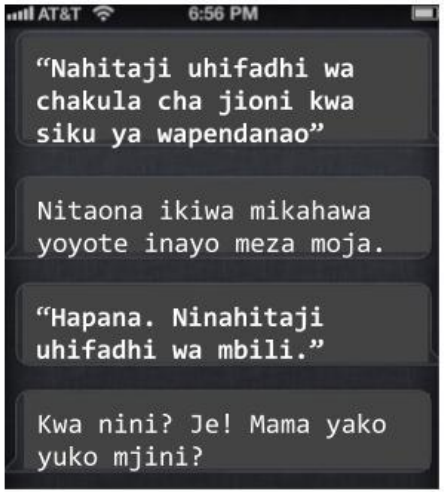
- Russian has very few words for colors, Japanese has hundreds
- Multiword expressions, e.g. it's raining cats and dogs or wake up and metaphors, e.g. love is a journey are very different across languages

~6,000 languages



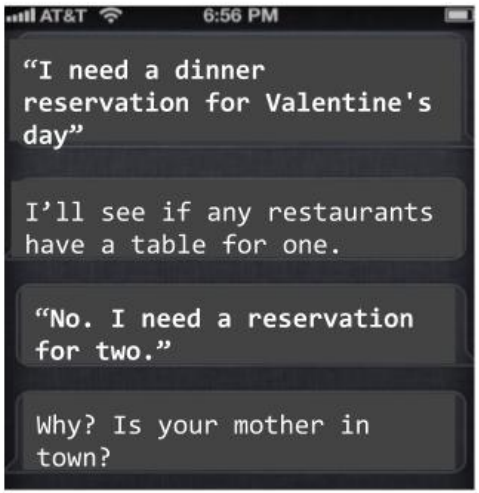
Spanish
534 million speakers

Hindi
615 million speakers



Swahili
100 million speakers

Thousands of language varieties

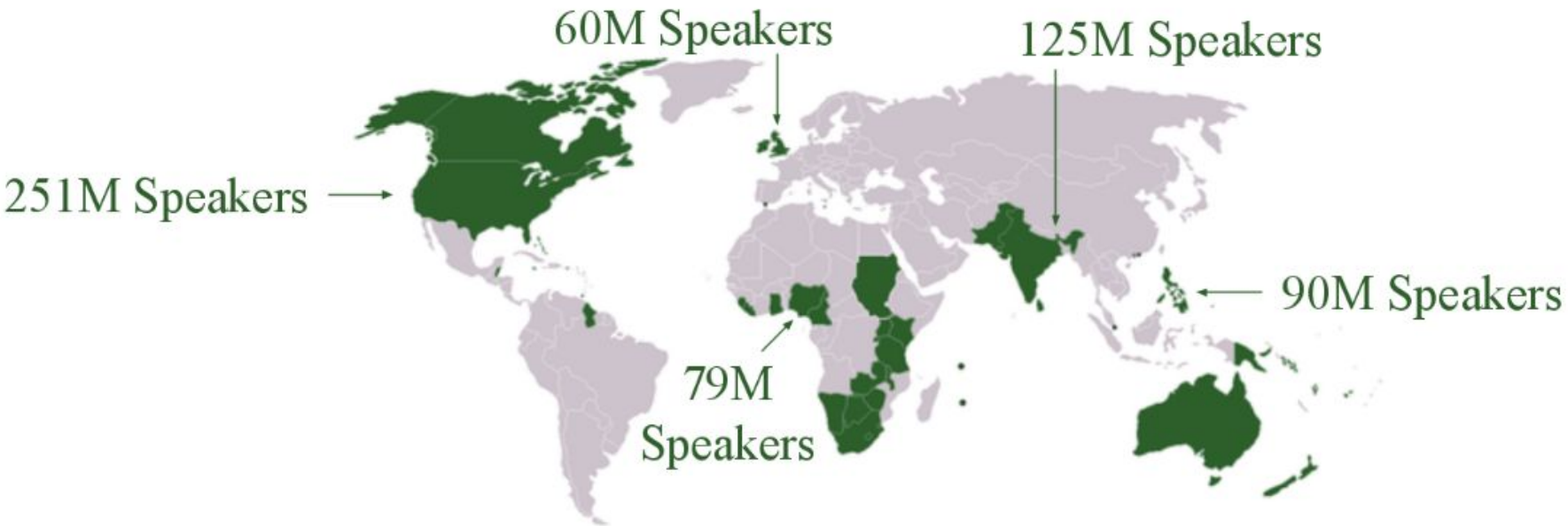


American English

Scottish English

Hinglish

World Englishes



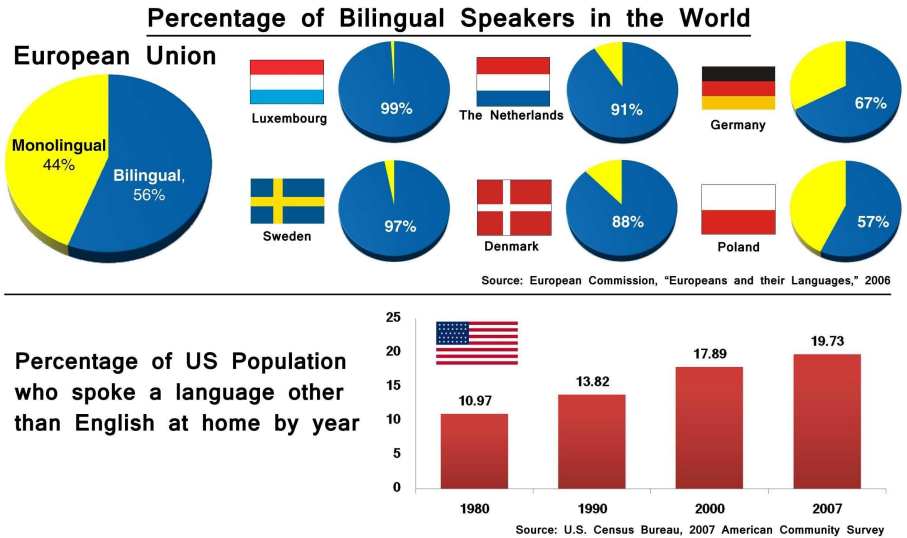

The Royal Family
@RoyalFamily
Follow
Taking place this week on the river Thames is 'Swan Upping' – the annual census of the swan population on the Thames.

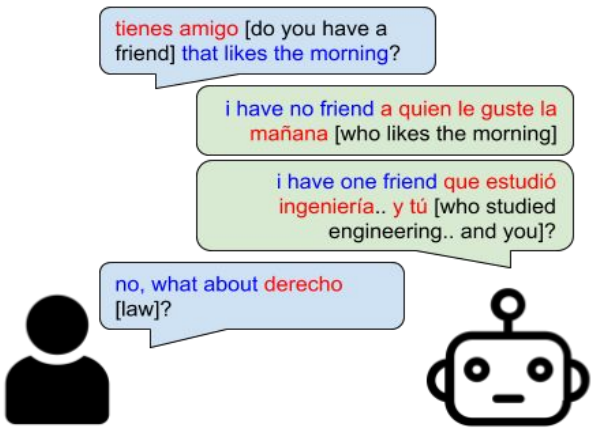

da'Rah-zingSun
@TIME7SS
Follow
@kimguilfoyle prblm I hve wit ur reportng is its 2 literal, evry1 knos pple tlk diffrnt evrywhere, u kno wut she means jus like we do!


Mooktar
@bossmukky
Follow
"@Ecstatic_Mi: @bossmukky Ebi like say I wan dey sick sef wlh 'Flu' my whole body dey weak"uw gee...


Ebenezer·
@Physique_cian
Follow
@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

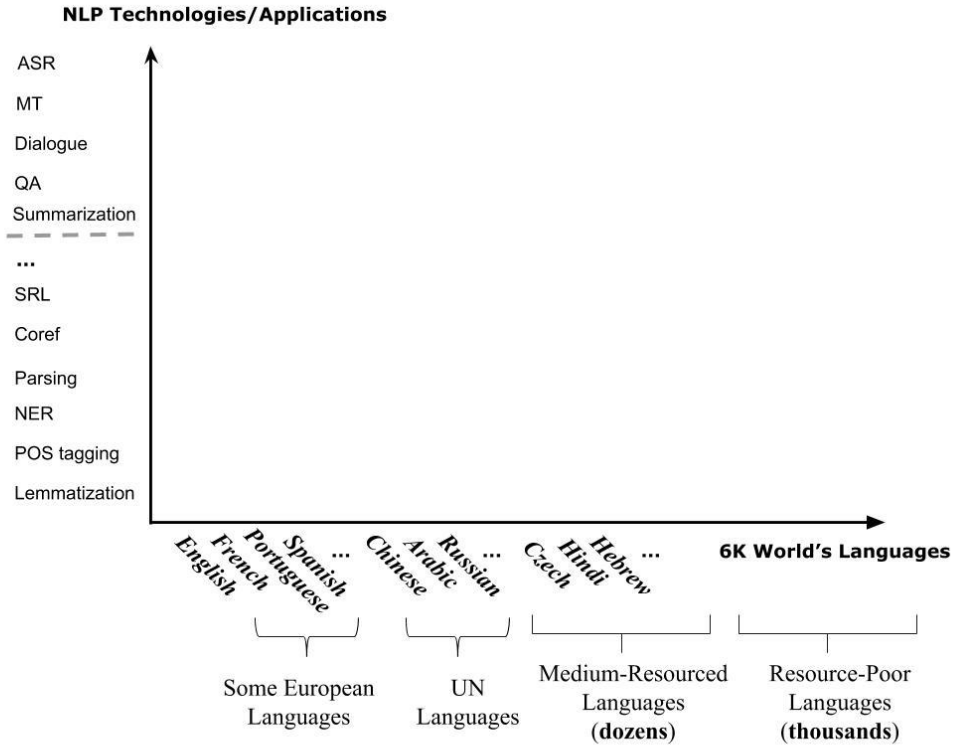
Code-switching

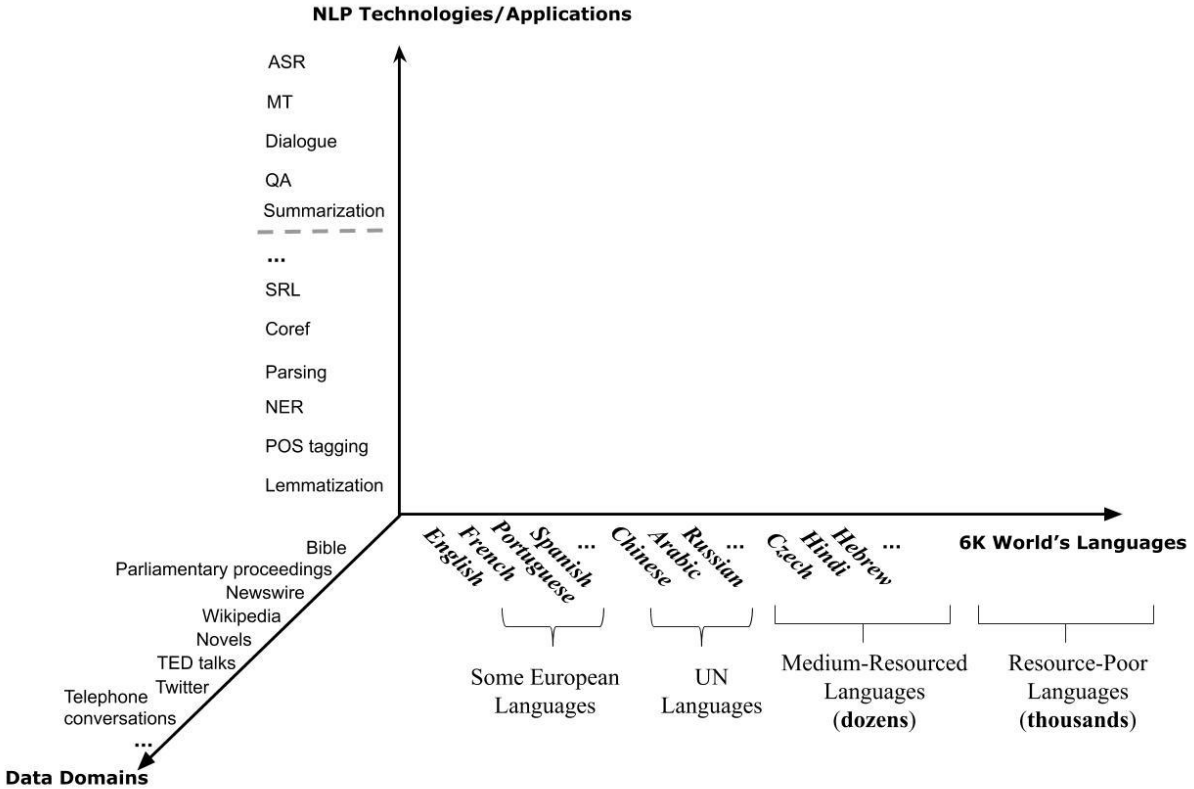


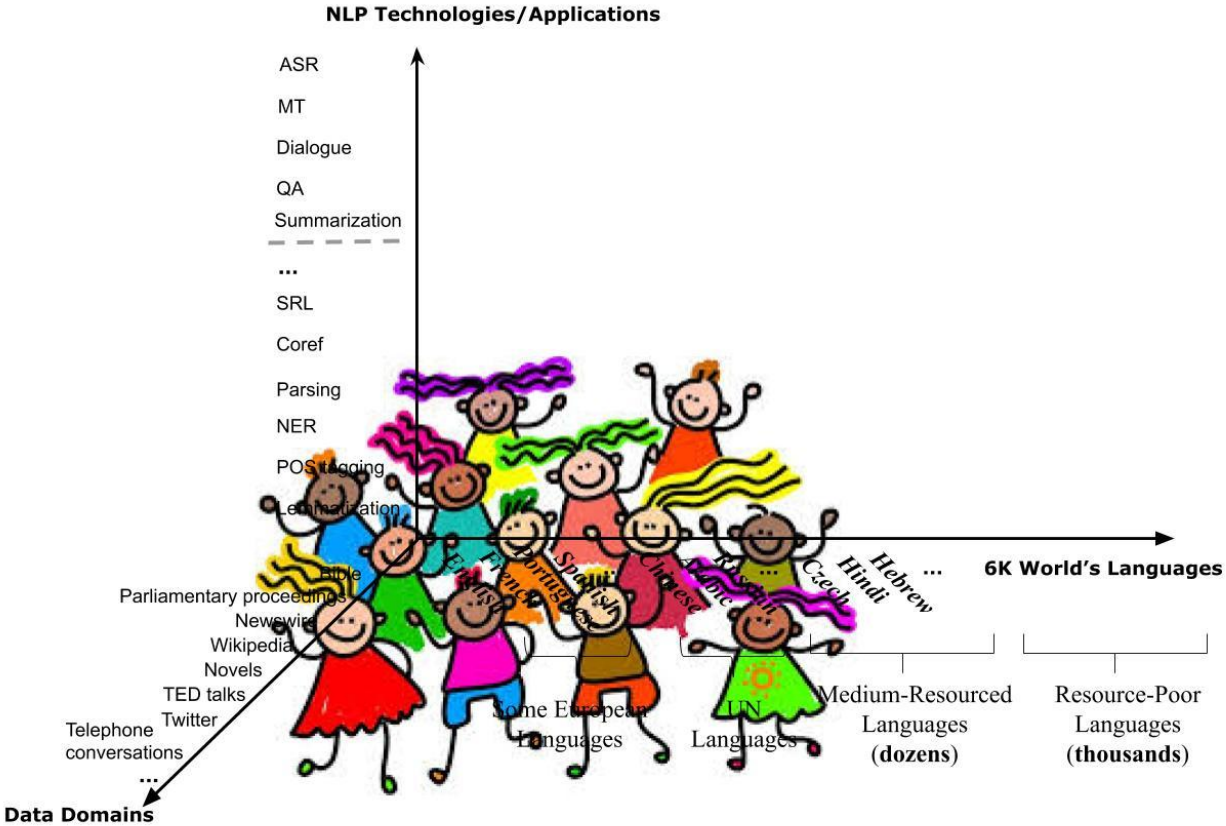


Source: US Census Bureau


NLP Technologies/Applications
ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization


NLP Technologies/Applications
ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization
English
French
Portuguese
Spanish
...
Chinese
Arabic
Russian
...
Czech
Hindi
Hebrew
...
6K World's Languages
Some European Languages
UN Languages
Medium-Resourced Languages (dozens)
Resource-Poor Languages (thousands)


NLP Technologies/Applications
ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization
Bible
Parliamentary proceedings
Newswire
Wikipedia
Novels
TED talks
Twitter
Telephone conversations
...
Data Domains
English
French
Portuguese
Spanish
...
Chinese
Arabic
Russian
...
Czech
Hindi
Hebrew
...
6K World's Languages
Some European Languages
UN Languages
Medium-Resourced Languages (dozens)
Resource-Poor Languages (thousands)


NLP Technologies/Applications
ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization
Bible
Parliamentary proceedings
Newswire
Wikipedia
Novels
TED talks
Twitter
Telephone conversations
...
Data Domains
Some European Languages
UN Languages
Medium-Resourced Languages (dozens)
Resource-Poor Languages (thousands)
6K World's Languages

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 .	None of them noticed a large , tawny owl flutter past the window .
assim , não viu as corujas descendo 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 .	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 .
Queira enviar-nos A sua coruja até dia 31 de Julho , sem falta .	We await your owl by no later than July 31 .
- O que é que quer dizer esperarem A minha coruja ?	after a few minutes He stammered , " what does it mean , They await My owl ?
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 .	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 .
O próprio Hagrid adormecera no sofá totalmente destruído e , bicando no vidro da janela , estava uma coruja que segurava um jornal .	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 .
A coruja entrou e depôs O jornal em cima de Hagrid	the owl swooped in and dropped the newspaper on top of Hagrid , who didn ' t

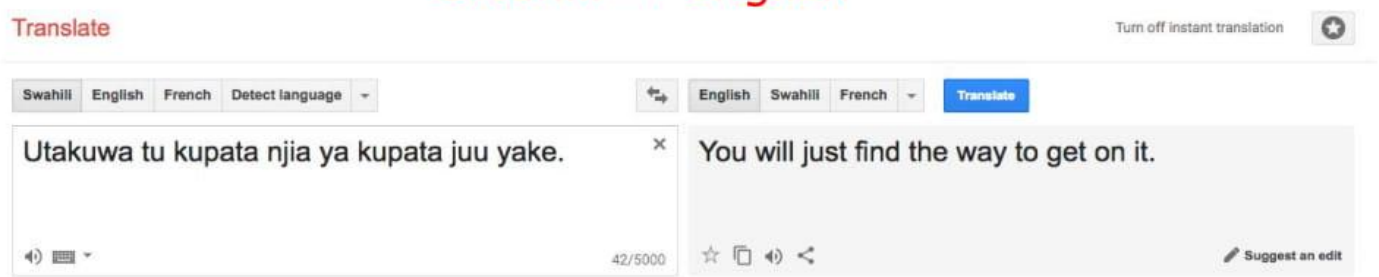
- Resource-rich language pair - few millions of parallel sentences
  Resource-poor language pair - few thousands of parallel sentences

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

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

English → Swahili

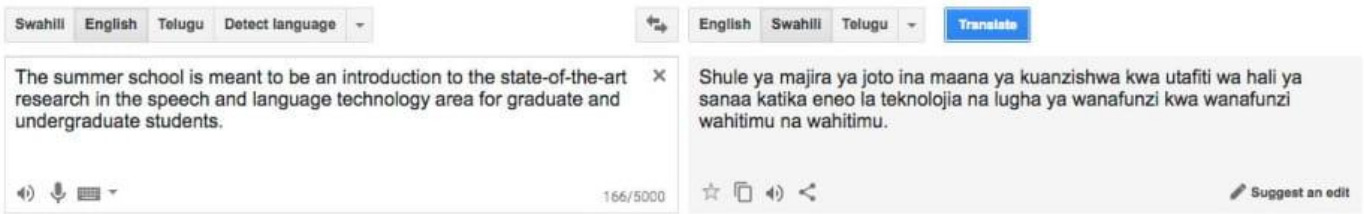
Swahili → English



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

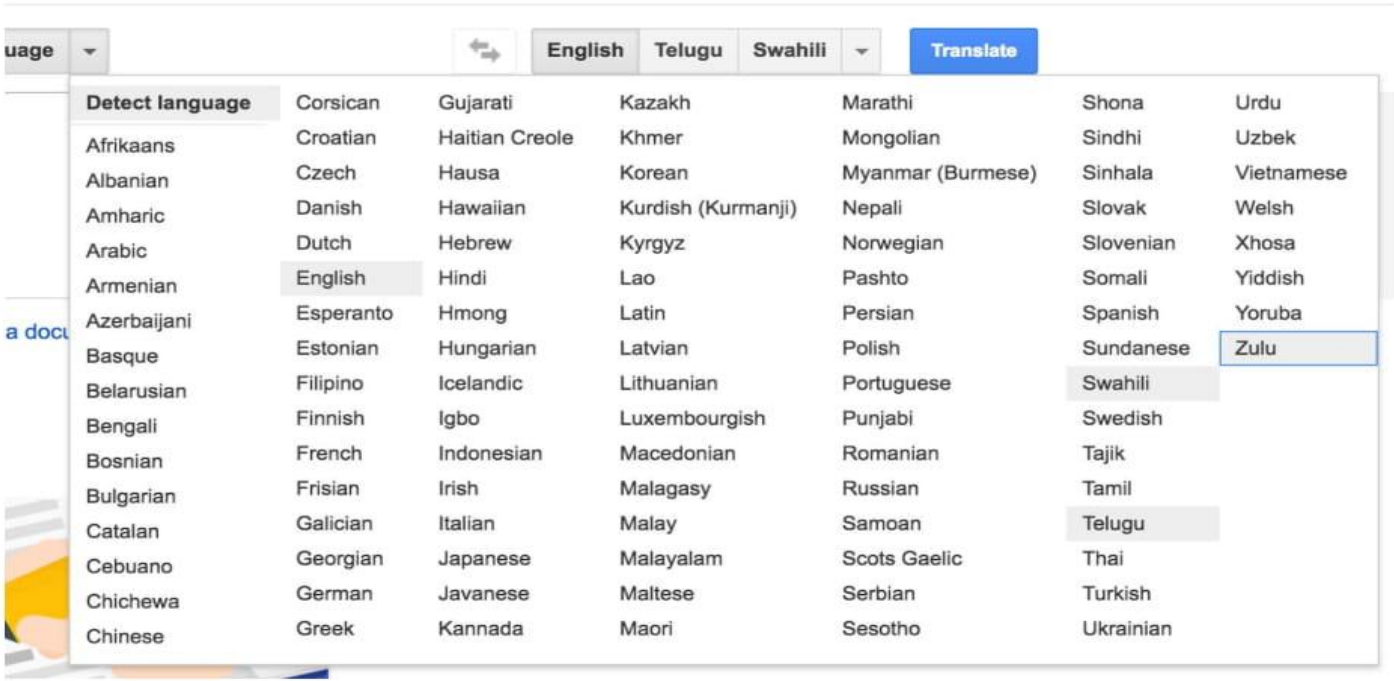
Resource-rich vs. resource-poor NLP: machine translation

English → Swahili

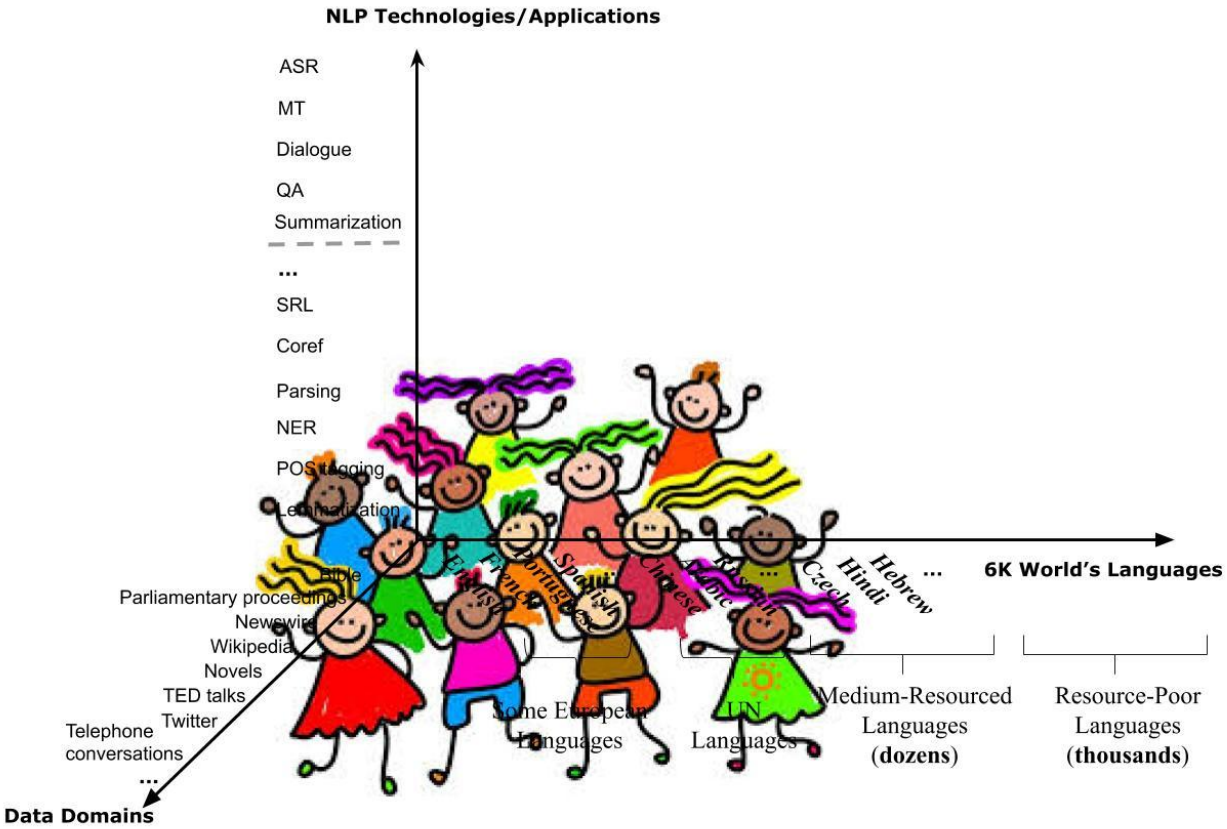


Swahili → English

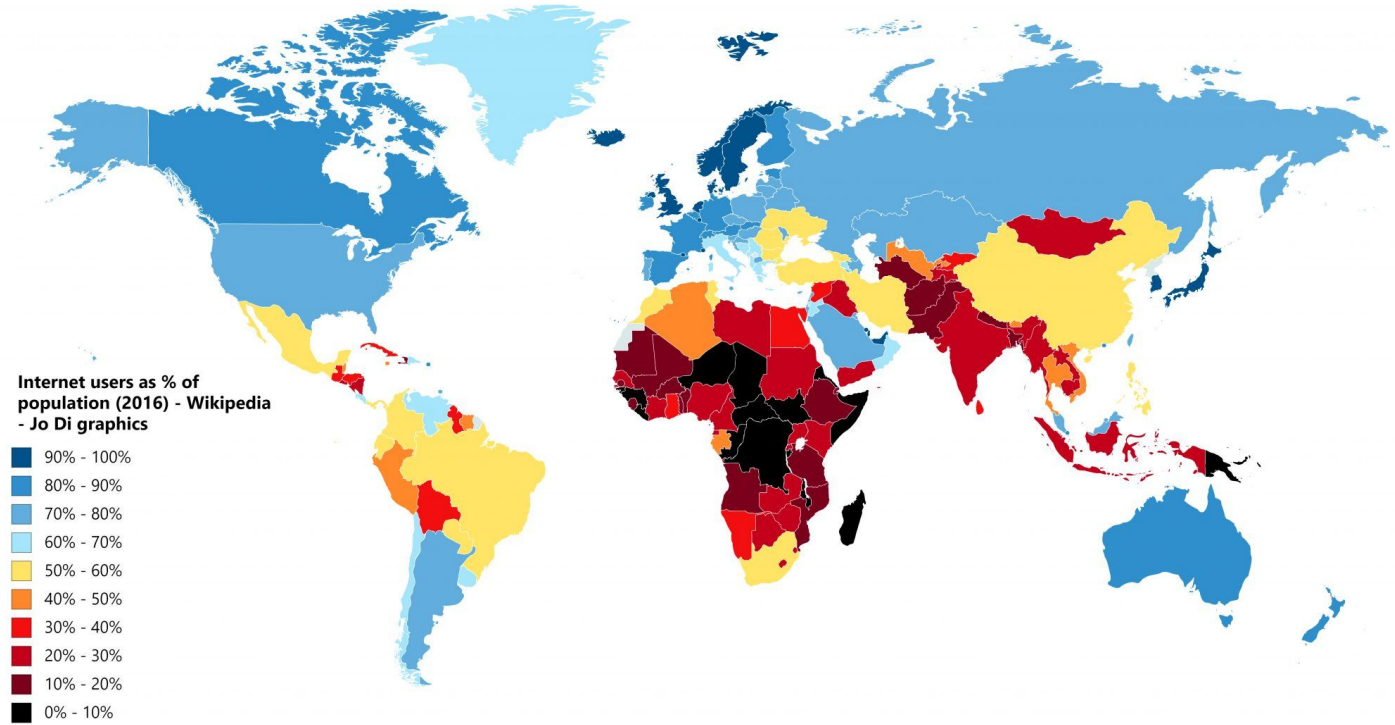
Resource-rich vs. resource-poor NLP: machine translation



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:

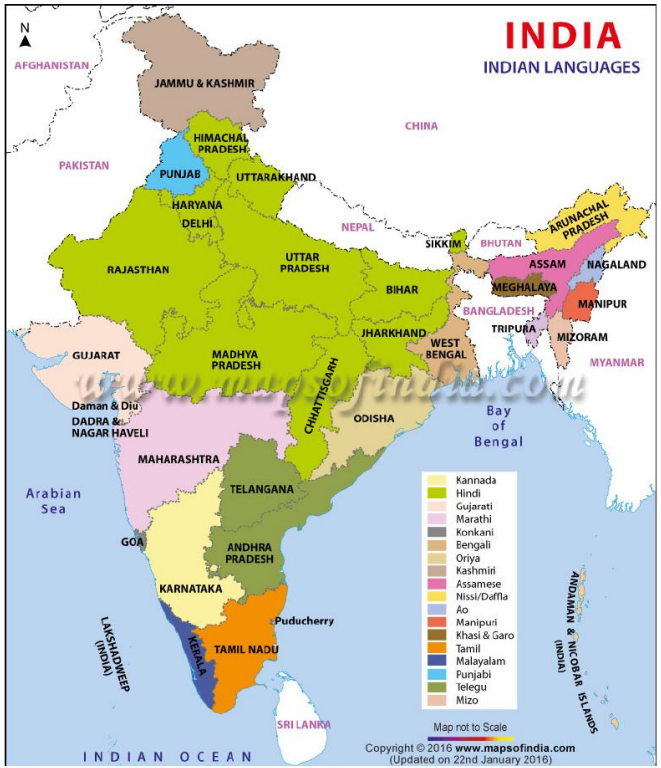
there are an estimated 1.5-2K African languages from 6 language families.

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

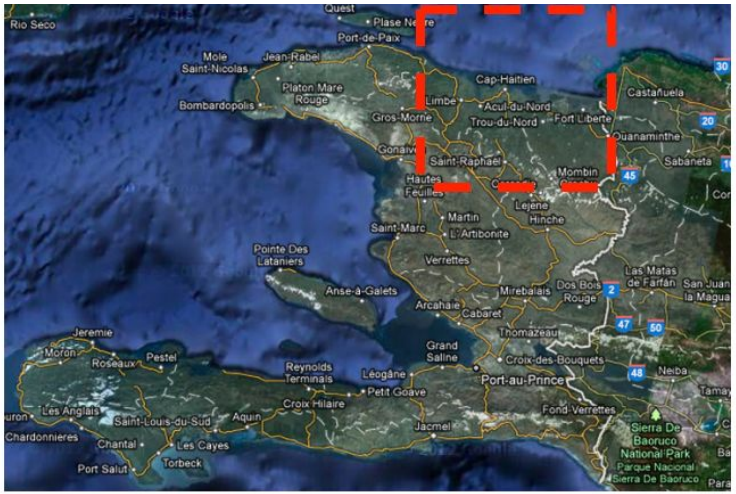
- My family in Carrefour, 24 Cote Plage,41A needs food and water
- People trapped in Sacred Heart Church, PauP
- General Hospital has less than 24 hrs. supplies
- Undergoing children delivery Delmas 31

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.”



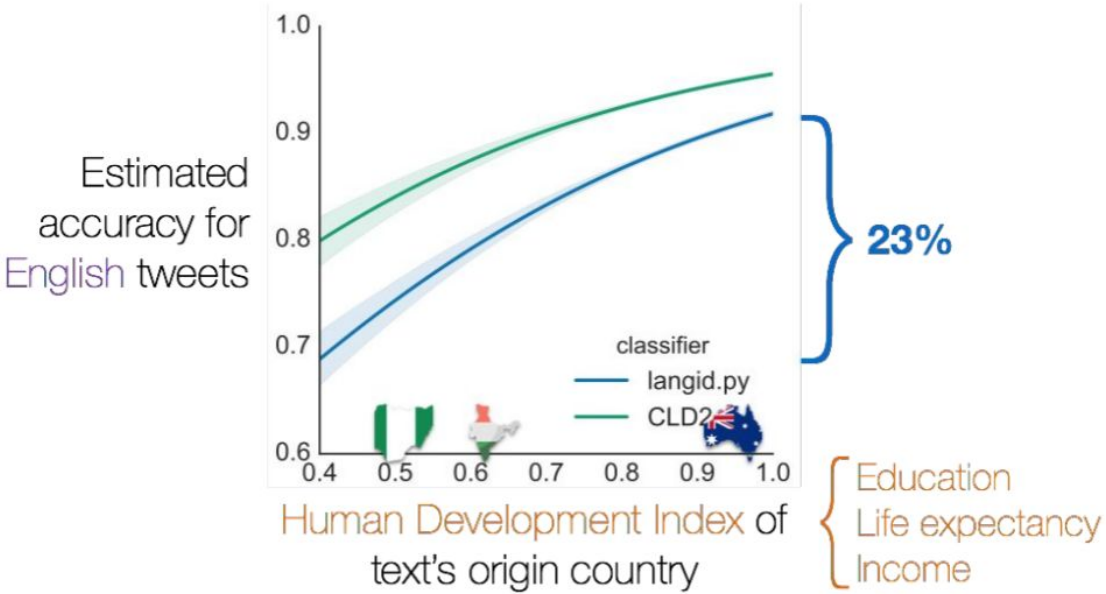
IDIBON

Example: Identifying disease outbreaks

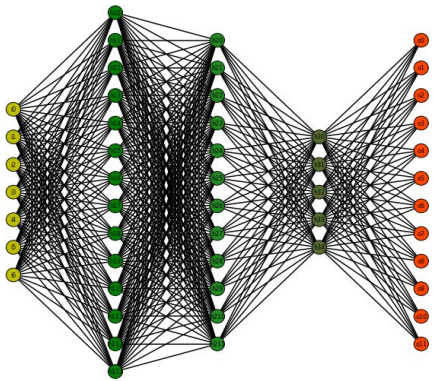
Slides on LID by David Jurgens (Jurgens et al. ACL'17)

Example: Identifying disease outbreaks

Example: Identifying disease outbreaks



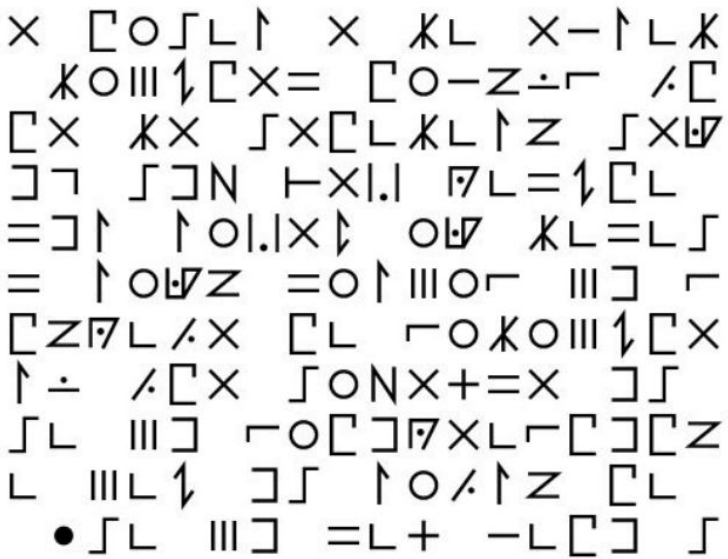
Why standard techniques used in NLP cannot simply be applied to low-resource languages?

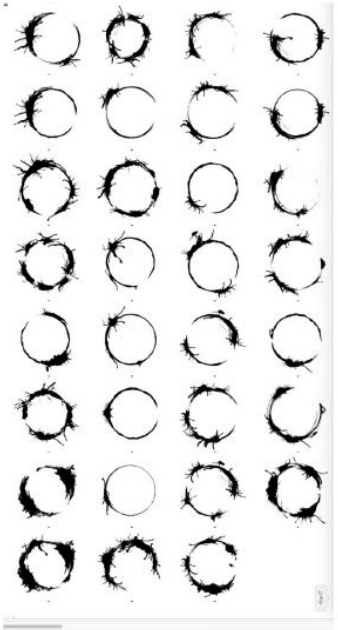


Why standard techniques used in NLP cannot simply be applied to low-resource languages?

- 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. Shanghainese 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://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

Are there sufficient training data for the task?

- Use supervised approaches

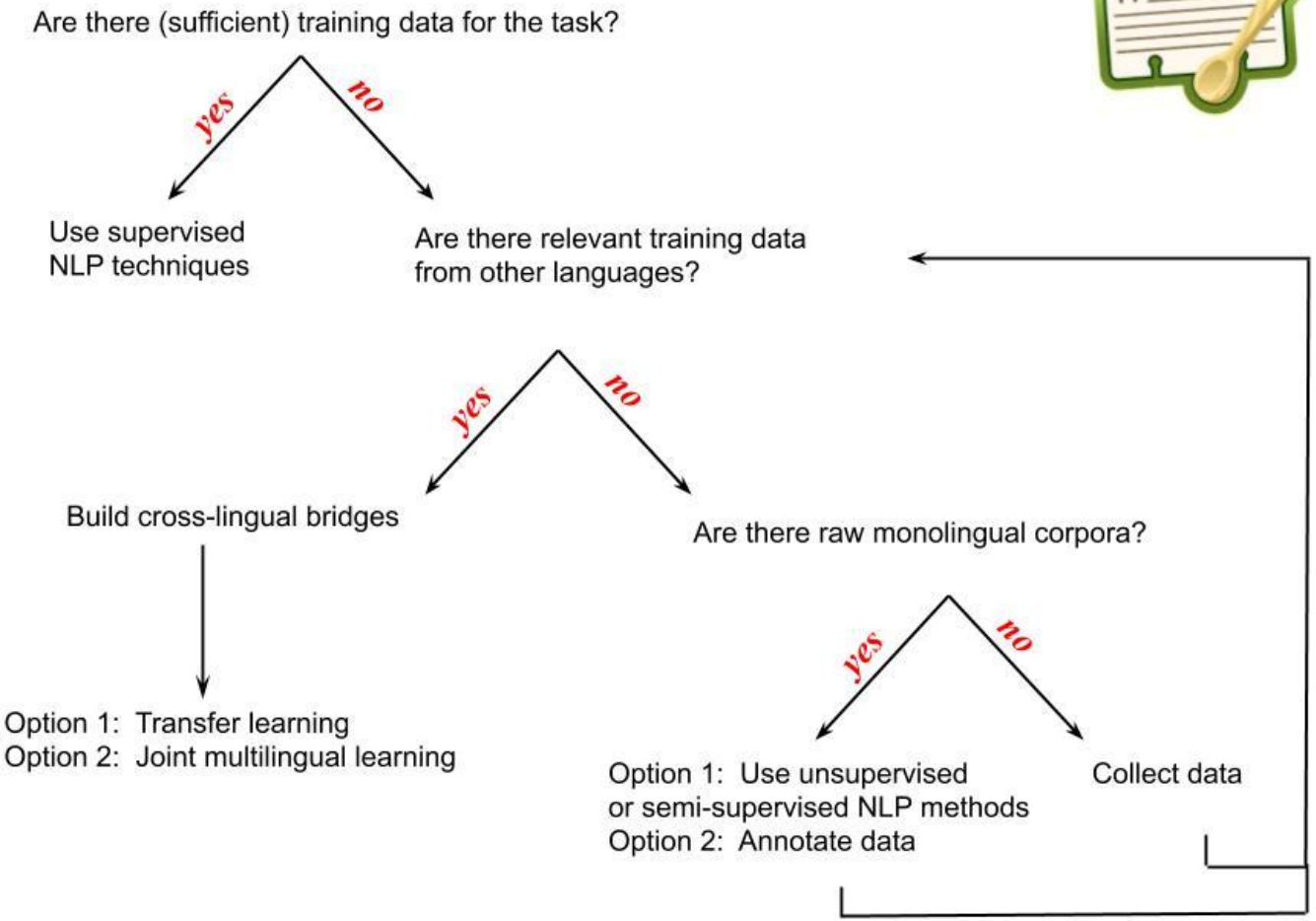
ACL
NAACL
EMNLP

Can we use resources from other languages?

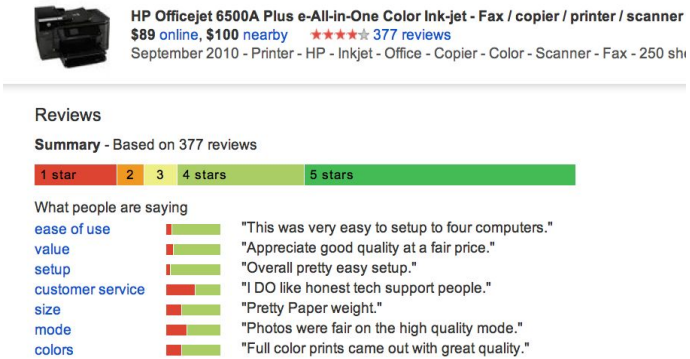
- Are there relevant training data from other languages?
- If yes, build cross-lingual bridges
- Use transfer learning or joint multilingual approaches

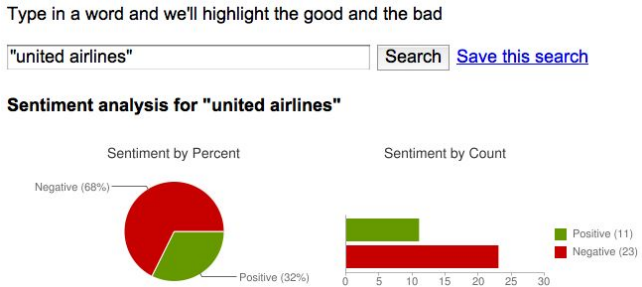
Are there sufficient raw monolingual corpora?

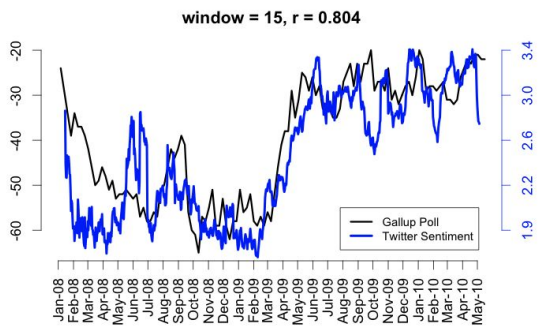
- Use unsupervised approaches


Are there (sufficient) training data for the task?
yes
no
Use supervised NLP techniques
Are there relevant training data from other languages?
yes
no
Build cross-lingual bridges
Are there raw monolingual corpora?
Option 1: Transfer learning
Option 2: Joint multilingual learning
yes
no
Option 1: Use unsupervised or semi-supervised NLP methods
Option 2: Annotate data
Collect data

Exercise today: multilingual sentiment analysis







Positive or negative movie review?

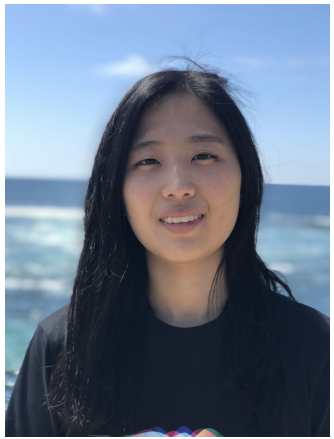
- unbelievably disappointing

- 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


thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
merci
obrigado
bedankt
dziękuje
sukriya
kop khun krap
go raibh maith agat
mochchakkeram
спасибо
рахмат
Баярлалаа
arigatô
dakujem
grazie
terima kasih
감사합니다
ありがとう
мерси
tapadh leat
dank je
maururu
hvala
sagolun
takk
taiku
ευχαριστώ
mersi
kia ora
barka
welalin
tack
spas
mahalo
faafetai lava
vinaka
спасибі
blagodaram
misaotra
matondo
paldies
grazzi
хвала
asante
manana
obrigada
tenki
murakoze
chokrane
mamnun
дякую
djiere dieuf
tau
kiitos
dankie
dhanyavad
nanni
nandri
bayarlalaa
gràcie
köszönöm
enkosi
sobodi
děkuji
mèsi
didi madloba
najis tuke
kam sah hamnida
rahmat
tanemirt
rahmet
xièxie
diolch
dhanyavadagalu
shukriya
mercé
trugarez
chnorakaloutioun
gratias ago
gràcies
sulpáy
akun
dankon
ačiū