Linguistic Diversity and NLP

Ximena Gutierrez-Vasques

LXAI

November, 2020

$WhoAmI

Currently

- *Postdoctoral researcher, Language and Space Lab (Text group). University of Zürich*
- Some of my research interests:

  - Natural language Processing (NLP) *- Quantitative linguistics* *- Low-resource languages*

- I currently work with approaches for quantifying morphological diversity/complexity in languages:
  - "Non-randomness in Morphological Diversity: A Computational Approach Based on Multilingual Corpora" (lead by Tanja Samardžić)

$WhoAmI

Before

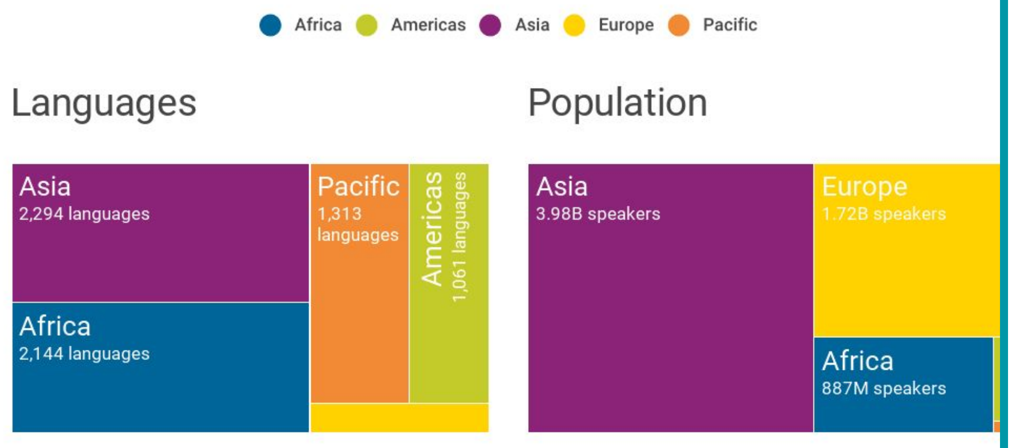
- *PhD Computational Linguistics (UNAM, Mexico)*. Working with bilingual lexicon extraction for Spanish-Nahuatl (an indigenous language of Mexico)

Outline

- Linguistic diversity and NLP
- Challenges
  - Dealing with *non-homogeneous text
  - Lack of corpus/datasets
  - How to adapt current methods?
- Final remarks

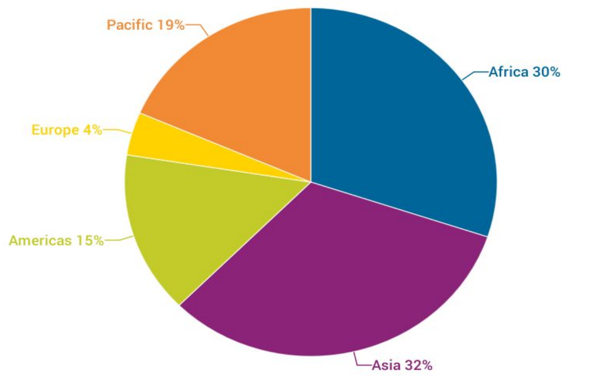
Linguistic Diversity

~Around 7K languages spoken in the world



For each region of the world, this graphic compares the number of languages (left) with how many people speak those languages (right). The population data isn't concerned with where people actually live, but rather, where their language is based. So, for instance, an English-speaking man living in China would be categorized under Europe.

Percentage of the world's languages, by region



Ethnologue

The case of Mexico

```
 ______________________________
|                              |
| 68 languages                 |
| 364 dialectal variations     |
| 11 linguistic families       |
| ___________________ ___      |
(\__/) ||
(•ㅅ•) ||
/    づ
```

United Nations
Educational, Scientific and
Cultural Organization

2019 International Year
of Indigenous Languages

Linguistic Diversity

- International Year of Indigenous Languages (2019):

  International Conference **Language Technologies** for All (LT4All) UNESCO, HeadQuarters, 2019

  *Languages represent **complex systems** of knowledge and communication and should be recognized as a strategic national resource for development, peace building and reconciliation[...]They also foster and promote unique local cultures, customs and values which have endured for thousands of years. Indigenous languages add to the rich tapestry of global **cultural diversity**. Without them, the world would be a poorer place.*

- Importance of **enabling** the use of indigenous languages in justice systems, the media, labour and health programmes.

UNESCO

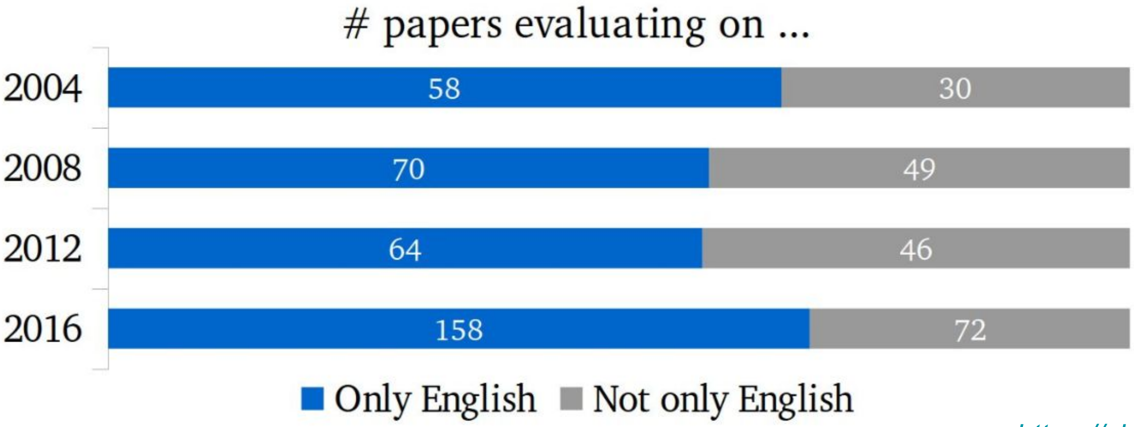
United Nations
Educational, Scientific and
Cultural Organization

2019 International Year
of Indigenous Languages

Linguistic Diversity and NLP

NLP does not necessarily reflect this diversity:

- ~60% of ACL papers use **English**
- They often do not even mention the language, assuming that English is some sort of “default”



https://sjmielke.com/acl-language-diversity.htm

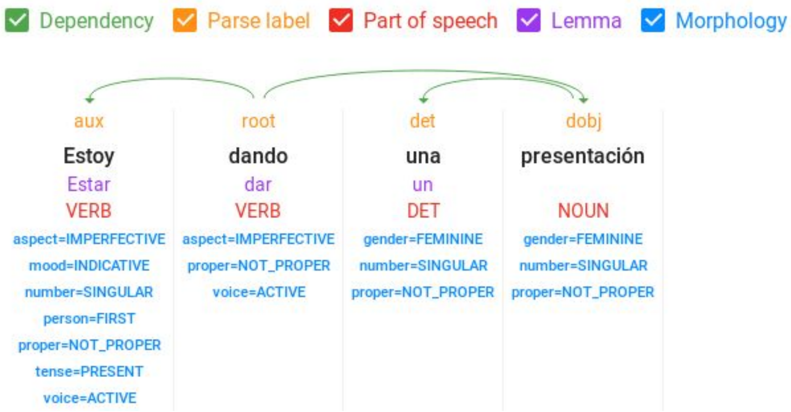
Linguistic Diversity and NLP

Many of the languages of the world **lack of**:

- **Pre-processing** tools: tokenizers Lemmatizers, spell checkers, taggers

- **Corpora/datasets**: raw text, annotated data, evaluation datasets

State-of-the-art (SOTA) methods do not necessarily work well under low-resource scenarios



☑ Dependency ☑ Parse label ☑ Part of speech ☑ Lemma ☑ Morphology

aux	root	det	dobj
Estoy	**dando**	**una**	**presentación**
Estar	dar	un	
VERB	VERB	DET	NOUN

aspect=IMPERFECTIVE
mood=INDICATIVE
number=SINGULAR
person=FIRST
proper=NOT_PROPER
tense=PRESENT
voice=ACTIVE
	aspect=IMPERFECTIVE
	proper=NOT_PROPER
	voice=ACTIVE
		gender=FEMININE
		number=SINGULAR
		proper=NOT_PROPER
			gender=FEMININE
			number=SINGULAR
			proper=NOT_PROPER

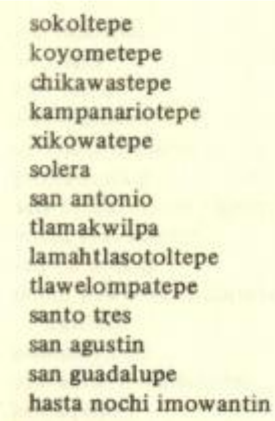
*Example generated using Google Cloud Natural Language API

Linguistic Diversity and NLP

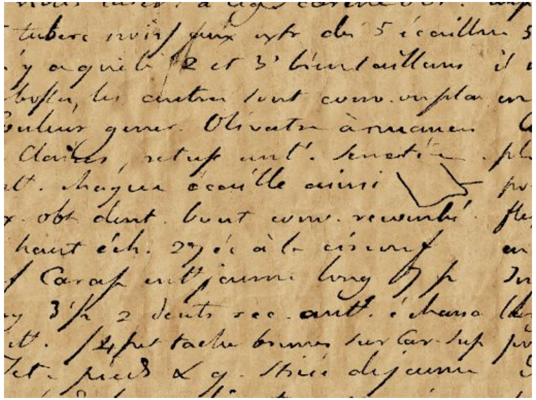
- The great diversity of languages posses interesting **scientific challenges**, e.g.,
  - Adapting well established approaches
  - Creation of new methods
  - Collecting new data.

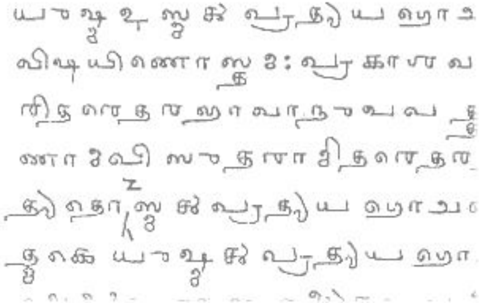
- Tackling these challenges **contributes** to building more general computational models of language, and to get a **deeper insight into human language** understanding

Challenge 1. Dealing with *non-homogeneous text



sokoltepe
koyometepe
chikawastepe
kampanariotepe
xikowatepe
solera
san antonio
tlamakwilpa
lamahtlasotoltepe
tlawelompatepe
santo tres
san agustin
san guadalupe
hasta nochi imowantin





Panorama:

- Not all languages have a strong orthographic tradition
- Lack of orthographic standardization
- Low production of digital texts
- Wide dialectal variation

Challenge 2. Lack of corpus/datasets

- SOTA models in NLP often require **big amounts** of training **data.** Examples:

  - GPT-2 (trained with 8 million web pages, 1.5 billion parameters)

  - Machine translation (~ from 35k to 2 billion parallel sentences)

- **Low-resource languages** do not have big amounts of digital text, readily available

  - Sometimes it is necessary to go to physical books (OCR)

  - Work with language communities to create small text corpora.

  - Crowdsourcing

Challenge 2. Some works

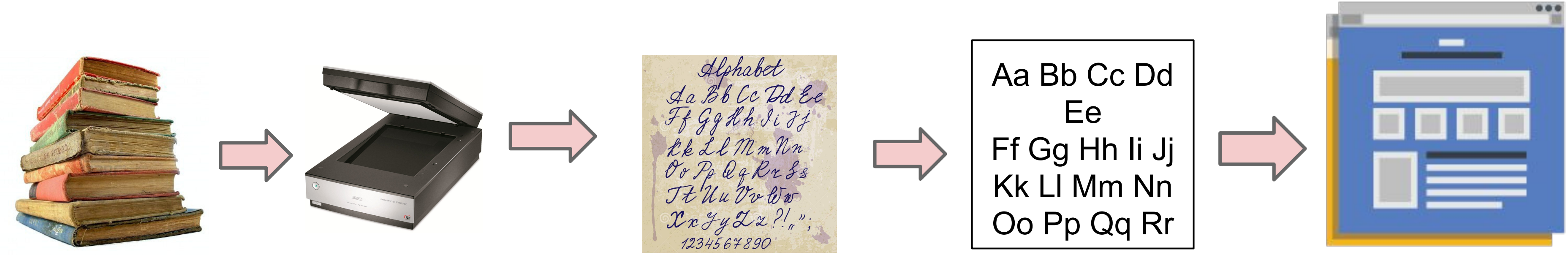
Extract bilingual and monolingual text from **different sources,** e.g. physical books, PDFs.

Peru

No data to crawl? Monolingual corpus creation from PDF files of truly low-resource languages in Peru *(Bustamante et al., 2020)*

Mexico

Axolotl: a Web Accessible Parallel Corpus for Spanish-Nahuatl *(Gutierrez-Vasques et al., 2016)*



Aa Bb Cc Dd
Ee
Ff Gg Hh Ii Jj
Kk Ll Mm Nn
Oo Pp Qq Rr

Challenge 2. Some works

Increasing interest in making truly typological **diverse datasets** for NLP tasks.

- *PBC corpus*. Parallel Bible Corpus, 1593 languages
- *OPUS* (an open source parallel corpus)
- *Sigmorphon, Unimorph*. Morphological datasets available in typological diverse languages

- *Universal Dependencies (UD)* framework

- *LC100.* Based on WALS 100-language sample, which aims to maximize both genealogical and areal diversity (in progress, URPP Language and Space, UZH)

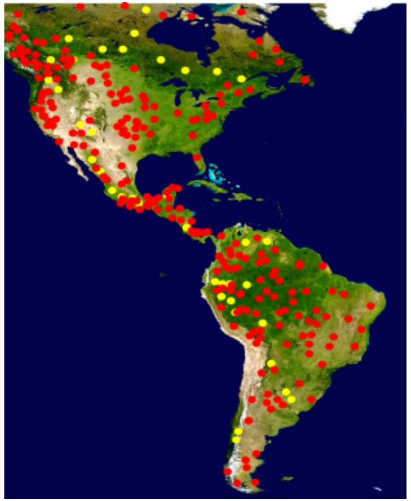
Challenge 2. Some works

Increasing interest within the NLP community. Examples:

- ACL special interest group on **multilinguality and linguistic typology** (SIGTYP)
- ACL ComputEL. Use of Computational Methods in the Study of **Endangered Languages**
- "First Workshop on NLP for Indigenous Languages of the Americas" **(upcoming NAACL 2021)**

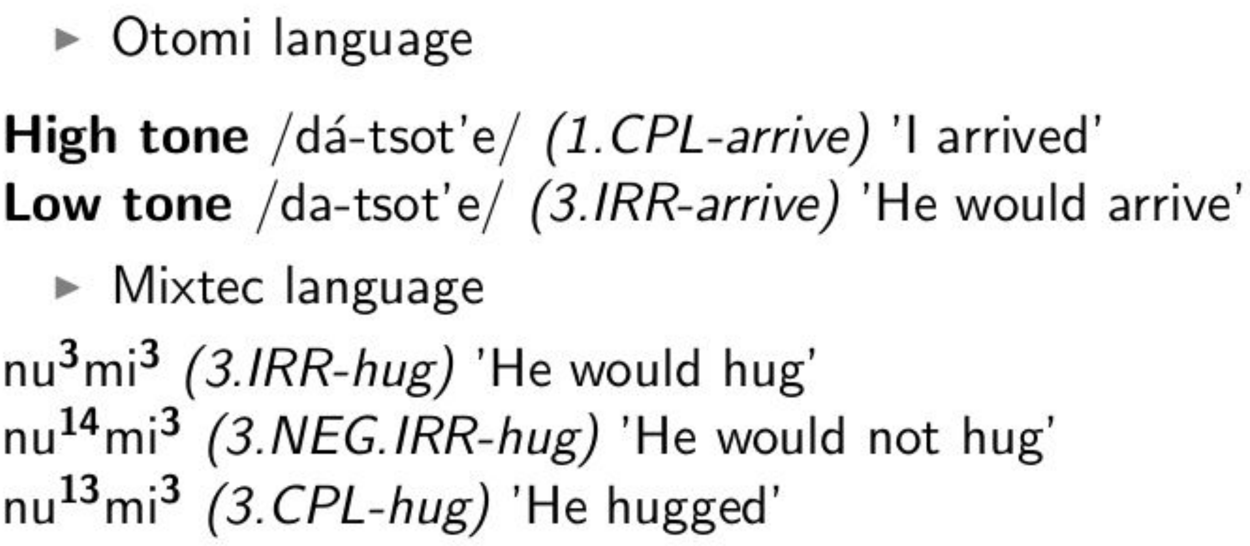
Challenge 3. How to adapt current methods?

- Languages of the world may exhibit **linguistic phenomena** that are **different** from the languages usually studied in Natural Language Processing(NLP)



Challenge 3. How to adapt current methods?

- Example. **Tonal** languages



▸ Otomi language

High tone /dá-tsot'e/ *(1.CPL-arrive)* 'I arrived'
Low tone /da-tsot'e/ *(3.IRR-arrive)* 'He would arrive'

▸ Mixtec language

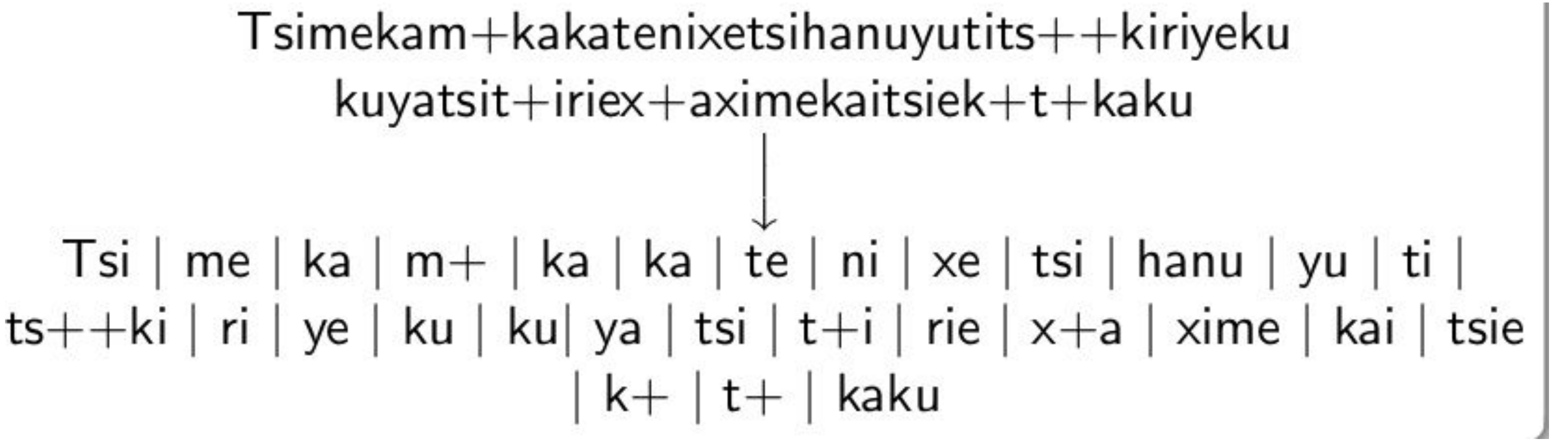
nu^{3}mi^{3} *(3.IRR-hug)* 'He would hug'
nu^{14}mi^{3} *(3.NEG.IRR-hug)* 'He would not hug'
nu^{13}mi^{3} *(3.CPL-hug)* 'He hugged'

** Mager, M., Gutierrez-Vasques, X., Sierra, G., & Meza, I. (2018). Challenges of language technologies for the indigenous languages of the Americas. COLING 2018*

Challenge 3. How to adapt current methods?

- Example. **Polysynthetic** languages

Wirrarika language:



Tsimekam+kakatenixetsihanuyutits++kiriyeku
kuyatsit+iriex+aximekaitsiek+t+kaku

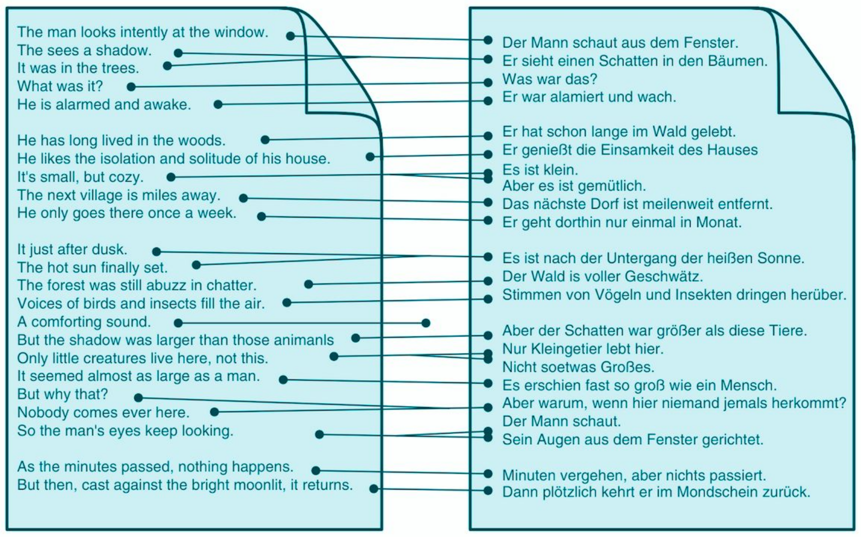
↓

Tsi | me | ka | m+ | ka | ka | te | ni | xe | tsi | hanu | yu | ti |
ts++ki | ri | ye | ku | ku| ya | tsi | t+i | rie | x+a | xime | kai | tsie
| k+ | t+ | kaku

** Mager, M., Gutierrez-Vasques, X., Sierra, G., & Meza, I. (2018). Challenges of language technologies for the indigenous languages of the Americas. COLING 2018*

Challenge 3. Machine translation

- Heavily **affected** by training **data size**
- … And also by the typological **distance between languages**



The man looks intently at the window.
The sees a shadow.
It was in the trees.
What was it?
He is alarmed and awake.

He has long lived in the woods.
He likes the isolation and solitude of his house.
It's small, but cozy.
The next village is miles away.
He only goes there once a week.

It just after dusk.
The hot sun finally set.
The forest was still abuzz in chatter.
Voices of birds and insects fill the air.
A comforting sound.
But the shadow was larger than those animanls
Only little creatures live here, not this.
It seemed almost as large as a man.
But why that?
Nobody comes ever here.
So the man's eyes keep looking.

As the minutes passed, nothing happens.
But then, cast against the bright moonlit, it returns.

Der Mann schaut aus dem Fenster.
Er sieht einen Schatten in den Bäumen.
Was war das?
Er war alamiert und wach.

Er hat schon lange im Wald gelebt.
Er genießt die Einsamkeit des Hauses
Es ist klein.
Aber es ist gemütlich.
Das nächste Dorf ist meilenweit entfernt.
Er geht dorthin nur einmal in Monat.

Es ist nach der Untergang der heißen Sonne.
Der Wald is voller Geschwätz.
Stimmen von Vögeln und Insekten dringen herüber.

Aber der Schatten war größer als diese Tiere.
Nur Kleingetier lebt hier.
Nicht soetwas Großes.
Es erschien fast so groß wie ein Mensch.
Aber warum, wenn hier niemand jemals herkommt?
Der Mann schaut.
Sein Augen aus dem Fenster gerichtet.

Minuten vergehen, aber nichts passiert.
Dann plötzlich kehrt er im Mondschein zurück.

- Training dataset:

  Parallel corpus

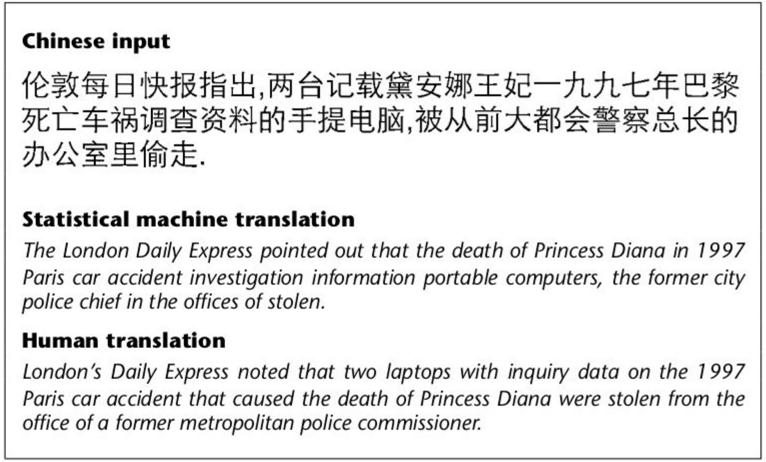
** Koehn, P. (2009). Statistical machine translation. Cambridge University Press.*

Challenge 3. Machine translation

Dataset size and languages distance

Language pair	Training corpus (words)
French-English	40 M
Arabic-English	200 M
Chinese-English	200 M

SMT system



Chinese input

伦敦每日快报指出,两台记载黛安娜王妃一九九七年巴黎死亡车祸调查资料的手提电脑,被从前大都会警察总长的办公室里偷走.

Statistical machine translation

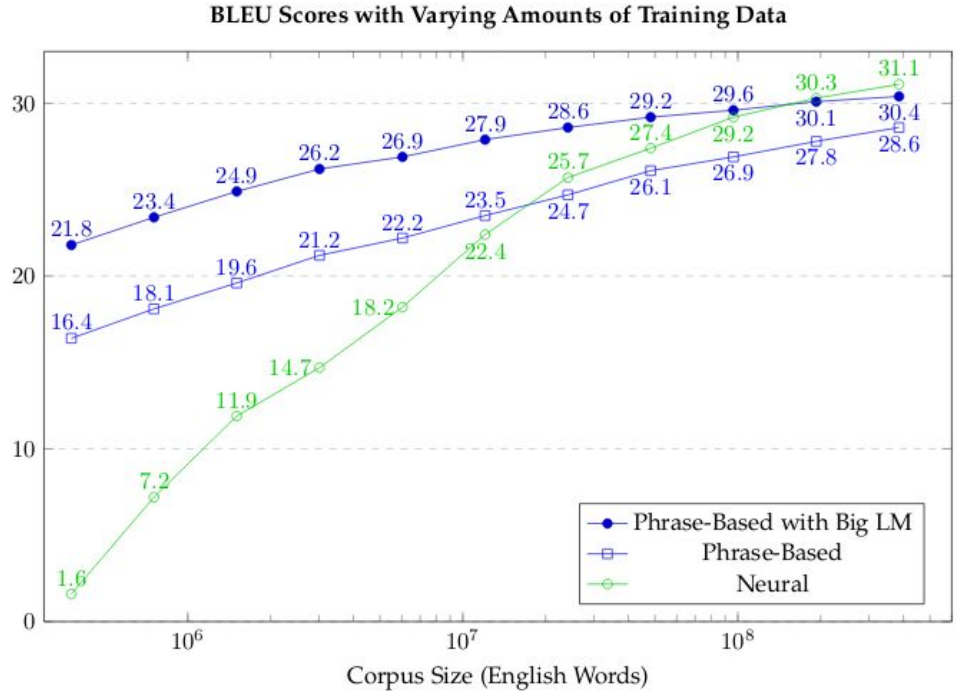
The London Daily Express pointed out that the death of Princess Diana in 1997 Paris car accident investigation information portable computers, the former city police chief in the offices of stolen.

Human translation

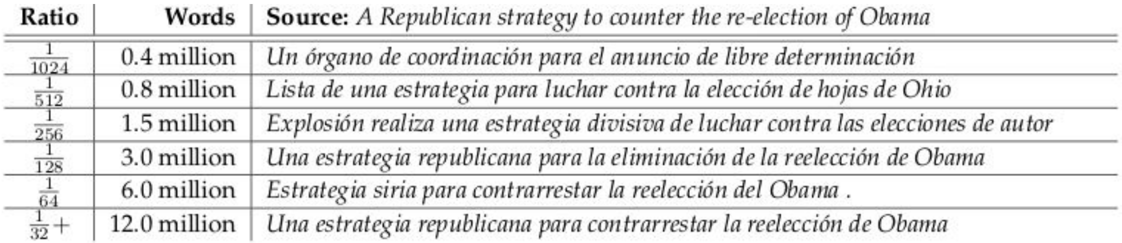
London's Daily Express noted that two laptops with inquiry data on the 1997 Paris car accident that caused the death of Princess Diana were stolen from the office of a former metropolitan police commissioner.

** Koehn, P. (2009). Statistical machine translation. Cambridge University Press.*

SMT and NMT under “low-resource” conditions



BLEU Scores with Varying Amounts of Training Data



Ratio	Words	**Source:** *A Republican strategy to counter the re-election of Obama*
$\frac{1}{1024}$	0.4 million	*Un órgano de coordinación para el anuncio de libre determinación*
$\frac{1}{512}$	0.8 million	*Lista de una estrategia para luchar contra la elección de hojas de Ohio*
$\frac{1}{256}$	1.5 million	*Explosión realiza una estrategia divisiva de luchar contra las elecciones de autor*
$\frac{1}{128}$	3.0 million	*Una estrategia republicana para la eliminación de la reelección de Obama*
$\frac{1}{64}$	6.0 million	*Estrategia siria para contrarrestar la reelección del Obama .*
$\frac{1}{32}$+	12.0 million	*Una estrategia republicana para contrarrestar la reelección de Obama*

** Koehn, P. (2017). Statistical Machine Translation. Draft of Chapter 13: Neural Machine Translation. Statistical Machine Translation.*

Challenge 3. How to adapt current methods?

In general low resource settings can benefit from ML and NLP advances that are able to **generalize better with less data**. Some promising directions:

- Multi-task learning
- Zero shot learning/few shot learning
- Transfer learning
- Meta learning

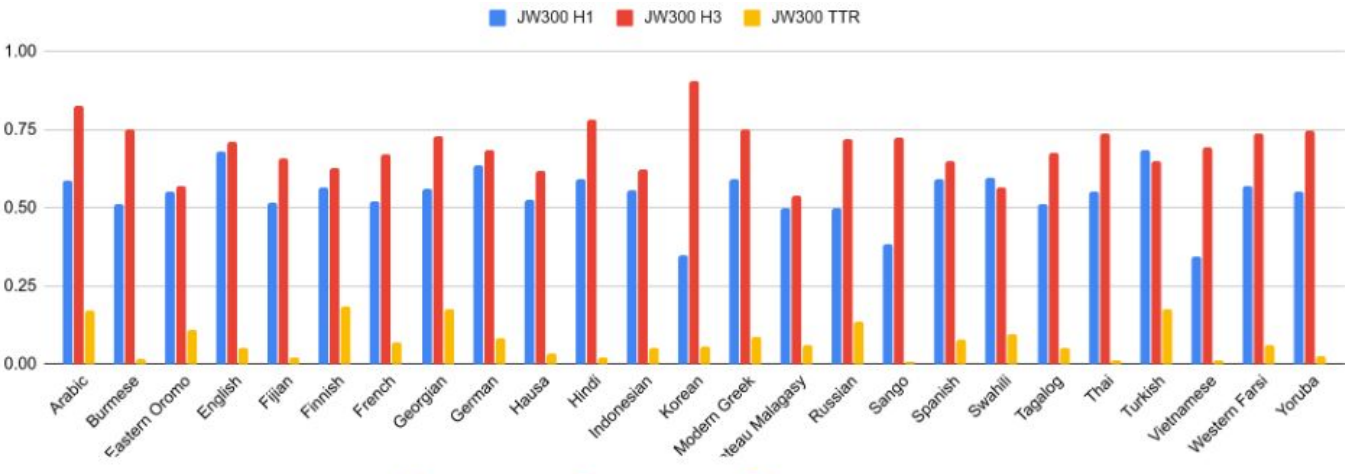
Leverage a set of high resource tasks that are already mastered, to improve the performance on a new (predominantly) low resource task *(Zoph et al., 2016)*

- Data augmentation techniques

Challenge 3. How to adapt current methods?

- **Linguistic knowledge** is important to be able to interpret current models and to inspire creative new methods

Challenge 3. How to adapt current methods?



TTR: Word level type-token ratio

H3: Entropy rate of a char trigram language model

H1: Entropy rate of a char unigram language model

Morphological complexity based on text

**Gutierrez-Vasques, X., & Mijangos, V. (2020). Productivity and Predictability for Measuring Morphological Complexity. Entropy, 22(1)*

Final remarks

- When working with languages, we have to think in the **communities** of people that speak those languages and their necessities
- **Low-resource** language speakers should be **included** in the development of language technologies for their own communities

> *“Technology is never neutral, it's made by humans. If we don't assure truly diverse work groups, we are not really creating technology for all”*

Dorothy Gordon, Ghana (Technology activist)

Some resources

- Masakhane.io *“A grassroots NLP community for Africa, by Africans”*
- Comunidad Elotl. NLP Community focused on Mexico’s indigenous languages
- https://github.com/pywirrarika/naki List of research and engineering of NLP for American Native/Indigenous Languages.

Gracias
Thank you
Tlasohkamati

Questions?