

---

# Taller básico de PLN

## Escuela de verano 2020

---

Ana Valeria González  
Ximena Gutiérrez-Vasques

# Información del taller

## → Parte 1:

- ◆ Intro a PLN/diversidad lingüística (~30 min)
- ◆ Sesión práctica (~50 min)

## → Parte 2:

- ◆ Introducción a representaciones vectoriales ( ~35 min)
- ◆ Sesión práctica ( ~ 45 min )

## → Parte 3:

- ◆ Sesgos e interpretabilidad (~25 minutos)
- ◆ Comentarios finales (~5 min)

# Procesando el lenguaje humano

- Gran reto:
  - Modelar el lenguaje humano desde una perspectiva computacional
  - El lenguaje es ambiguo y complejo
- Procesamiento del lenguaje natural (PLN)/ Lingüística computacional
  - Buscamos hacer modelos que sean capaces de procesar y generar lenguaje natural
  - Aplicación en las tecnologías del lenguaje
  - Área interdisciplinaria

# Ambigüedad: Nivel léxico

Luis compró una **planta** para decorar su nuevo hogar.

Paola vive en la **planta** baja de su edificio.

Me duele la **planta** del pie.

Mario trabaja en una **planta** industrial.

# Ambigüedad: Nivel sintáctico/pragmático

## Nivel Sintáctico

Golpeó el armario con el bastón y lo rompió

## Nivel Pragmático

¿Me puedes pasar la sal?

# Ambigüedad: Nivel referencial

**Michelle Obama** está atravesando un momento personal difícil. **La ex primera dama** ha revelado en el segundo episodio de su *podcast* personal que se emite en Spotify (*The Michelle Obama Podcast*) que padece depresión. “He tenido altibajos emocionales en los últimos cinco meses. Esos momentos en que no te sientes como tú misma eres”, ha contado en conversación con la periodista Michele Norris **la esposa de Barack Obama**, aclarando que se trata de una depresión leve o de bajo grado.

nota de El País, 2020

# Tecnologías del lenguaje



# Traducción automática

The image shows a screenshot of a web-based translation tool. At the top, there are two dropdown menus for language selection: 'English' on the left and 'Spanish' on the right, separated by a double-headed arrow icon. Below these, the input text in English is 'I would like to learn everything about Natural Language Processing'. To the right of this text is a small 'X' icon. The translated text in Spanish is 'Me gustaría aprender todo sobre el procesamiento del lenguaje natural.' Below the English text is a speaker icon for audio playback. Below the Spanish text are a speaker icon and a copy icon.

English	Spanish
I would like to learn everything about Natural Language Processing	Me gustaría aprender todo sobre el procesamiento del lenguaje natural.



# Búsqueda y Recuperación de información

Q cuanto tiempo

- Q cuanto tiempo **dura el coronavirus**
- Q cuanto tiempo **vive una mosca**
- Q cuanto tiempo **duro la peste negra**
- Q cuanto tiempo **duro la gripe española**
- Q cuanto tiempo **dura el coronavirus en la ropa**
- Q cuanto tiempo
- Q cuanto tiempo **vive un gato**
- Q cuanto tiempo **vive el coronavirus**
- Q cuanto tiempo **dura el covid en la ropa**
- Q cuanto tiempo **dura el coronavirus en las superficies**



# Otras aplicaciones:

**Análisis de sentimientos**

**Reconocimiento de entidades  
nombradas (NER)**

**Tecnologías de voz**

**Resumen Automático**

**Etiquetado lingüístico automático:  
morfológico, sintáctico, semántico**

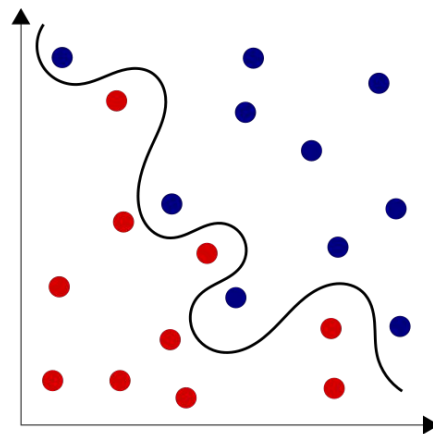
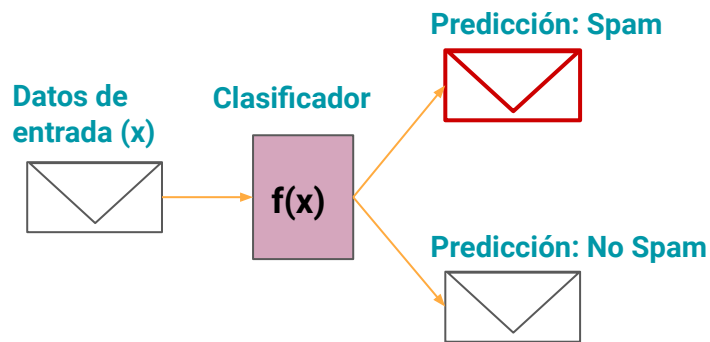
**Sistemas pregunta-respuesta  
(Question answering)**

**Clasificación de textos**

**Generación del lenguaje**

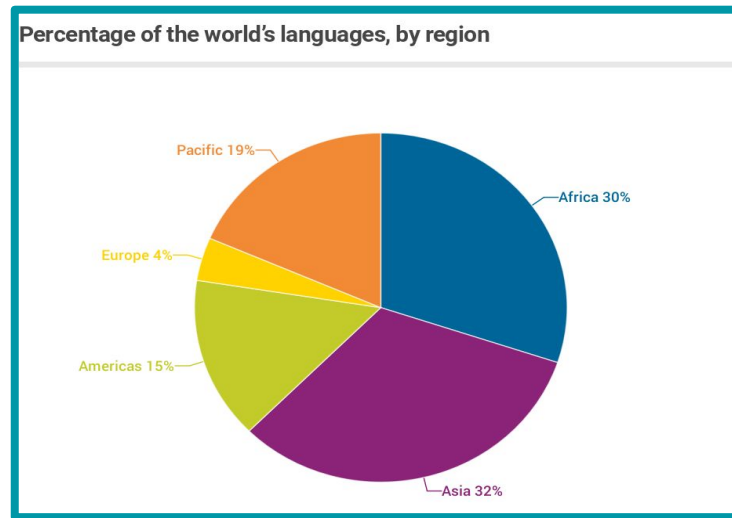
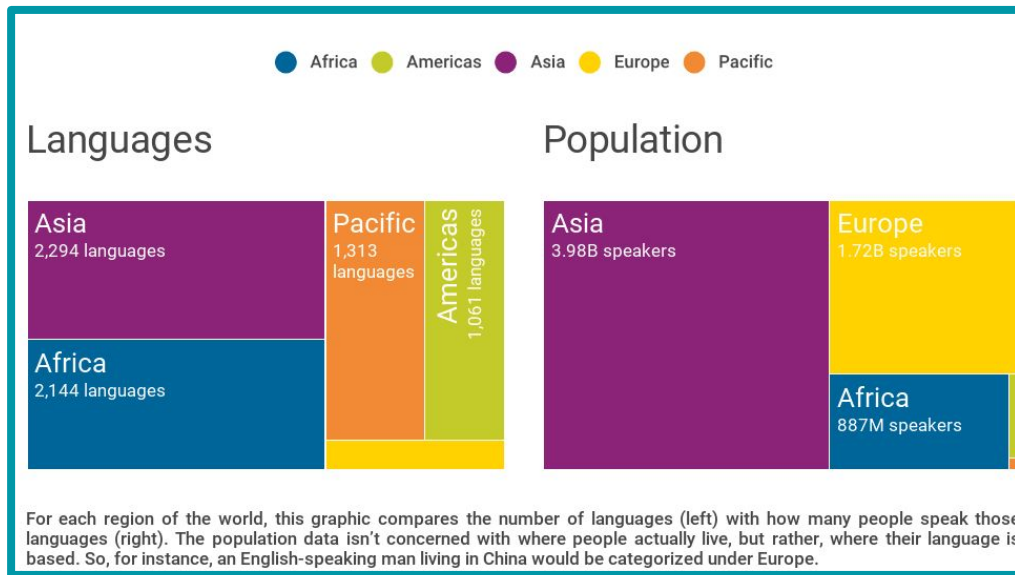
# PLN+Aprendizaje de máquina

- Actualmente, muchas de las tareas en PLN se plantean como un problema de **clasificación**
- Se necesita un dataset con **ejemplos etiquetados** y, a partir de esto, se aprende un modelo que pueda discriminar/clasificar automáticamente



# Diversidad lingüística

~Alrededor de 7,000 lenguas se hablan en el mundo



# El caso de México

68 Agrupaciones lingüísticas

364 Variantes

11 familias lingüísticas

(\\_/) ||  
(•̀•́) ||  
/ づ



United Nations  
Educational, Scientific and  
Cultural Organization



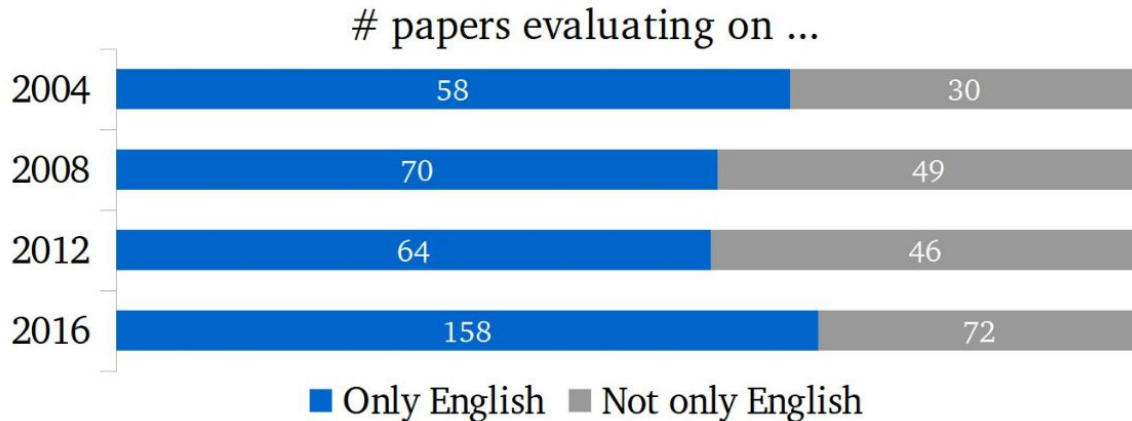
2019 International Year  
of Indigenous Languages



# Diversidad lingüística y NLP

NLP no refleja necesariamente esta diversidad:

- ~60% de los artículos publicados (en ACL) usan **Inglés**
- A veces ni siquiera se especifica la lengua, asumiendo que el inglés es una especie de “default”




# Diversidad lingüística y NLP

Muchas de las lenguas del mundo **carecen de:**

- Herramientas de **preprocesamiento**: tokenizadores, lematizadores, correctores ortográficos, etiquetadores,...
- **Corpus/datasets**: texto plano, corpus anotados, datasets de evaluación

**Los métodos estado del arte (SOTA)** no necesariamente funcionan bien en los escenarios de bajos recursos

✓ 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

\*Ejemplo generado usando Google  
Cloud Natural Language API

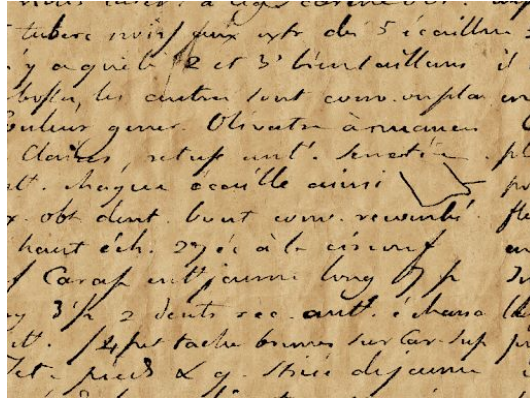
# Principales retos

- ◆ Lidar con texto \*no homogéneo\*
- ◆ Falta de recursos digitales (corpus/datasets/herramientas)
- ◆ ¿Cómo adaptar los métodos actuales?



# Lidiando con texto no-homogéneo

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



யுஷு உஷு க்ஷு ப்ஷு யுஷு  
விஷு யிஷு ணாஷு க்ஷு ப்ஷு  
நிஷு ரெஷு நுஷு வாஷு உஷு  
ணாஷு விஷு ணாஷு நிஷு ரெஷு  
க்ஷு ப்ஷு க்ஷு ப்ஷு யுஷு  
க்ஷு ப்ஷு க்ஷு ப்ஷு யுஷு

## Panorama:

- No todas las lenguas tienen una tradición ortográfica
- Falta de estandarización ortográfica
- Gran variación dialectal
- Poca producción de textos digitales

# Lidiando con texto no-homogéneo

- Tareas como corrección ortográfica, predicción de la siguiente palabra, etc. necesitan de **modelos de lenguaje** estadísticos/neuronales --->

...que usualmente necesitan **grandes cantidades de texto** para ser entrenados

- La normalización ortográfica se convierte en un paso necesario

$$p(w|w_{t-k}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+k})$$

Hola	,	Cómo	estás	?
------	---	------	-------	---

# Falta de recursos digitales

- Los modelos actuales en NLP requieren **grandes cantidades** de corpus de entrenamiento. Ejemplos:
  - GPT-2 (entrenado con 8 millones de páginas web, 1.5 billones de parámetros)
  - Traducción automática (~ 35k a 2 billones de oraciones paralelas)
- **Las lenguas de bajos recursos** no poseen grandes cantidades de texto digital fácilmente accesible
  - A veces es necesario ir a libros físicos (OCR)
  - Trabajar con comunidades de hablantes para generar pequeños corpus
  - Crowdsourcing

# Falta de recursos digitales. Algunos trabajos

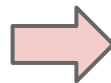
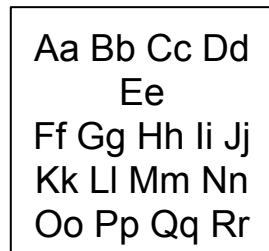
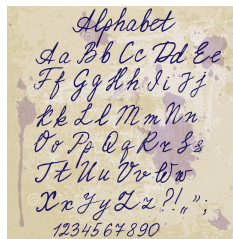
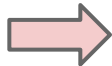
**Extraer** texto monolingüe y bilingüe de **diferentes fuentes**: Libros físicos, PDFs, etc.

## 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*)



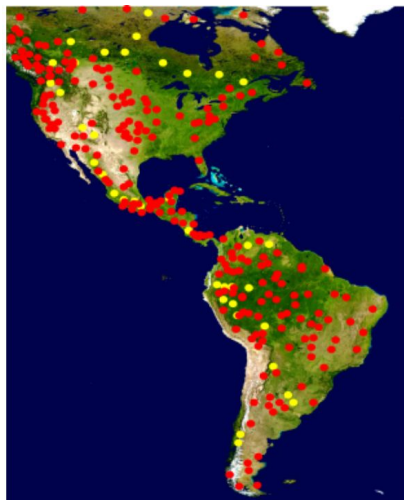
# Falta de recursos digitales. Algunos trabajos

Interés creciente en hacer datasets tipológicamente diversos para diferentes tareas de NLP. Ejemplos:

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

# ¿Cómo adaptar los métodos actuales?

- Las diferentes lenguas del mundo pueden exhibir **fenómenos lingüísticos** que son muy distintos de los que usualmente se estudian en NLP



# Ejemplo. Riqueza morfológica

**Tinehcakisneki** (Nahuatl)

*Me quieres escuchar*

*You want to hear me*

**Siebentausendzweihundertvierundfünfzig** (German)

**Siete mil doscientos cincuenta y cuatro**

**7,254**

**tsă** (Otomi)

*Comer una sola cosa*

*To eat a single thing*

# Ejemplo. Lenguajes tonales

- ▶ 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<sup>3</sup>mi<sup>3</sup> (3.IRR-hug) 'He would hug'

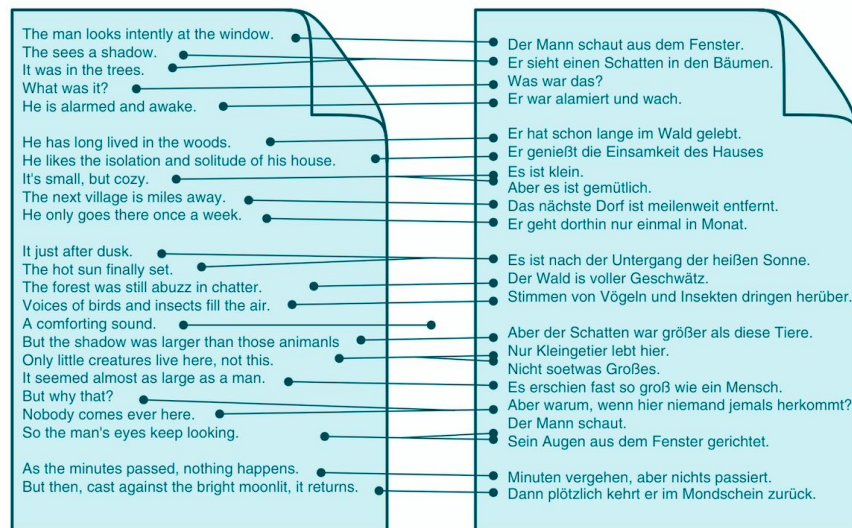
nu<sup>14</sup>mi<sup>3</sup> (3.NEG.IRR-hug) 'He would not hug'

nu<sup>13</sup>mi<sup>3</sup> (3.CPL-hug) 'He hugged'



# Ejemplo. Traducción automática

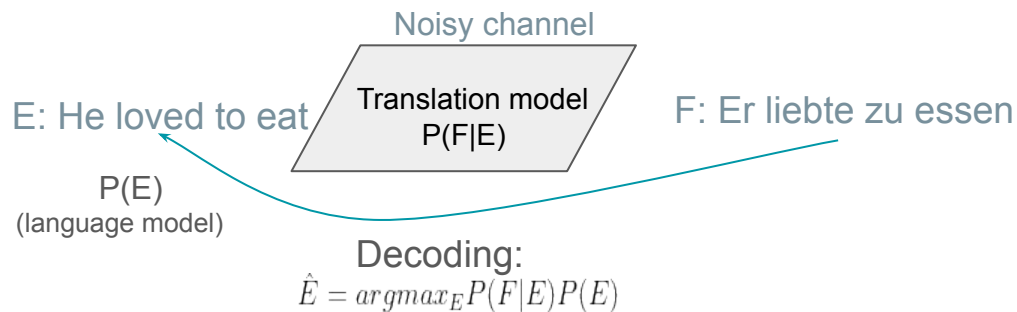
- Desempeño altamente **afectado** por el **tamaño** de corpus de entrenamiento
- ...y también por la **distancia tipológica** entre lenguas



- Dataset de entrenamiento:

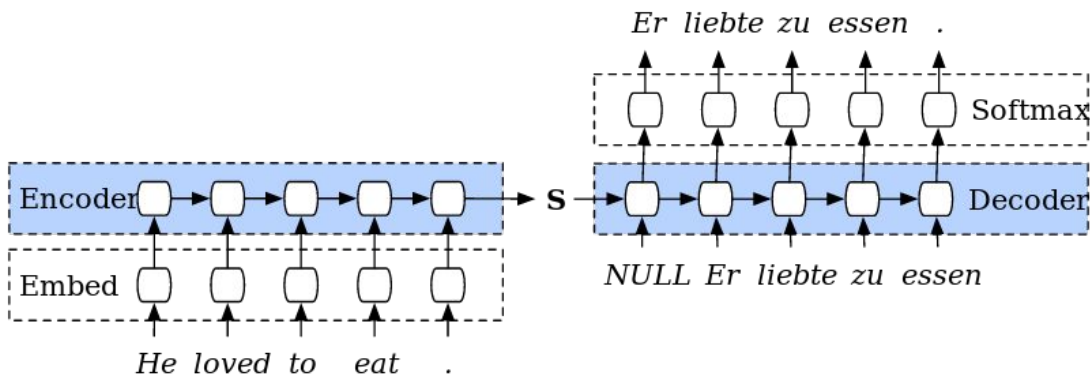
Corpus paralelo

# Ejemplo. Traducción automática



## Statistical machine translation (SMT)

- Inició en los 90's
- IBM models



## Neural machine translation (NMT)

- ~ Inició en 2015
- Representaciones vectoriales (embeddings)
- Ya no hay modelos separados, solo un modelo secuencial que predice una palabra a la vez
- RNN, LSTMS, Transformers

# Ejemplo. Traducción automática

## Tamaño del corpus y “distancia” entre lenguas

Language pair	Training corpus (words)
<b>French-English</b>	<b>40 M</b>
<b>Arabic-English</b>	<b>200 M</b>
<b>Chinese-English</b>	<b>200 M</b>

### 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

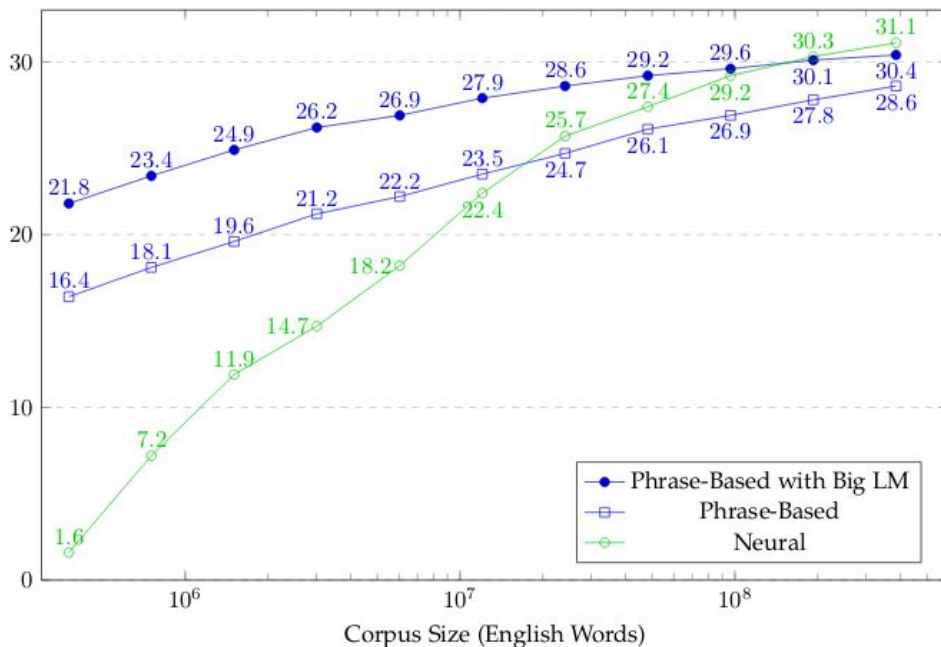
*The London Daily Express pointed out that the death of Princess Diana in 1997 Paris car accident investigation information portable computers, the former city police chief in the offices of stolen.*

### 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.*

## SMT system

BLEU Scores with Varying Amounts of Training Data



## SMT y NMT bajo condiciones de “bajos recursos”

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

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

# Ejemplo. Traducción automática

$En \rightarrow Any$	High 25	Med. 52	Low 25
Bilingual	29.34	17.50	11.72
400M	28.03	16.91	12.75
1.3B Wide	28.36	16.66	11.14
1.3B Deep	29.46	17.67	12.52
$Any \rightarrow En$	High 25	Med. 52	Low 25
Bilingual	37.61	31.41	21.63
400M	33.85	30.25	26.96
1.3B Wide	37.13	33.21	27.75
1.3B Deep	37.47	34.63	31.21

## Massively Multilingual NMT

\*Avg. translation quality (BLEU) of multilingual models with increasing capacity.

\*High 25 refers to the top 25 languages by dataset size, while low 25 refers to the bottom 25.

*\* Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., ... & Macherey, W. (2019). Massively multilingual neural machine translation in the wild: Findings and challenges.*

# ¿Cómo adaptar los métodos actuales?

Los entornos de bajos recursos pueden beneficiarse de avances en aprendizaje de máquina que son capaces de **generalizar mejor con menos datos de entrenamiento**. Algunas direcciones:

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

Aprovechar un conjunto de tareas con muchos recursos de entrenamiento para mejorar el desempeño de tareas nuevas con pocos recursos de entrenamiento (*Zoph et al., 2016*)

- Data augmentation techniques



# Práctica