

# Consideraciones de NLP para lenguas minorizadas. El caso de México

Ximena Gutiérrez-Vasques

URPP Language and Space, Text group



University of  
Zurich <sup>UZH</sup>



## 1. Un poco sobre mi

## 2. Low-resource languages?

- Alentar grupos diversos en la comunidad de NLP
- Disponibilidad de los recursos
- “Abrazar” la variación

## 3. Extra. Algunos recursos

# Mi perfil. Líneas de investigación

→ PLN Multilingüe

Francés

Nawat

Mazateco

Japonés

Hñahñu

Alemán Suizo

Tepehua

Español

Popoloca

Mixteco

Totonaco

## Mi perfil. Líneas de investigación

→ PLN Multilingüe

→ Morfología Computacional

Tinechcakisneki

2.SG.S-1.S.O-‘escuchar’-FUT-‘querer’

Rasca|cielos

## Mi perfil. Líneas de investigación

→ PLN Multilingüe

→ Morfología Computacional

→ Teoría de la información  
aplicada a la lingüística



## Mi perfil. Líneas de investigación

→ PLN Multilingüe

→ Morfología Computacional

→ Teoría de la información  
aplicada a la lingüística

→ Tecnologías del lenguaje  
para lenguas minorizadas

## Mi perfil. Líneas de investigación

→ PLN Multilingüe

→ Morfología Computacional

→ Teoría de la información  
aplicada a la lingüística

→ Tecnologías del lenguaje  
para lenguas minorizadas

→ PLN y aprendizaje de máquina en  
entornos de bajos recursos



## Mi perfil. Actualidad

→ Investigadora postdoctoral. Language and Space Lab, Text Group. University of Zurich



Proyecto: *Non-randomness in Morphological Diversity: A Computational Approach Based on Multilingual Corpora*

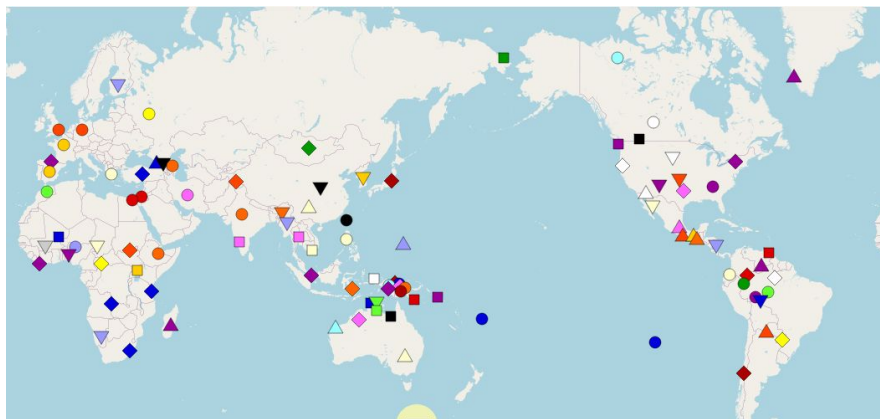
Coordinado por: Dra. Tanja Samardžić y Dr. Christian Bentz



## Mi perfil. Actualidad

Proyecto: *Non-randomness in Morphological Diversity: A Computational Approach Based on Multilingual Corpora*

- Aplicar teoría de la información, modelado estadístico y aprendizaje de máquina enfocados al estudio de la diversidad lingüística
- Auxiliar en la creación y mantenimiento de infraestructura para un corpus de 100 lenguas tipológicamente diversas.



\*WALS 100-language Sample

## Mi perfil. Actividades

- Adicional

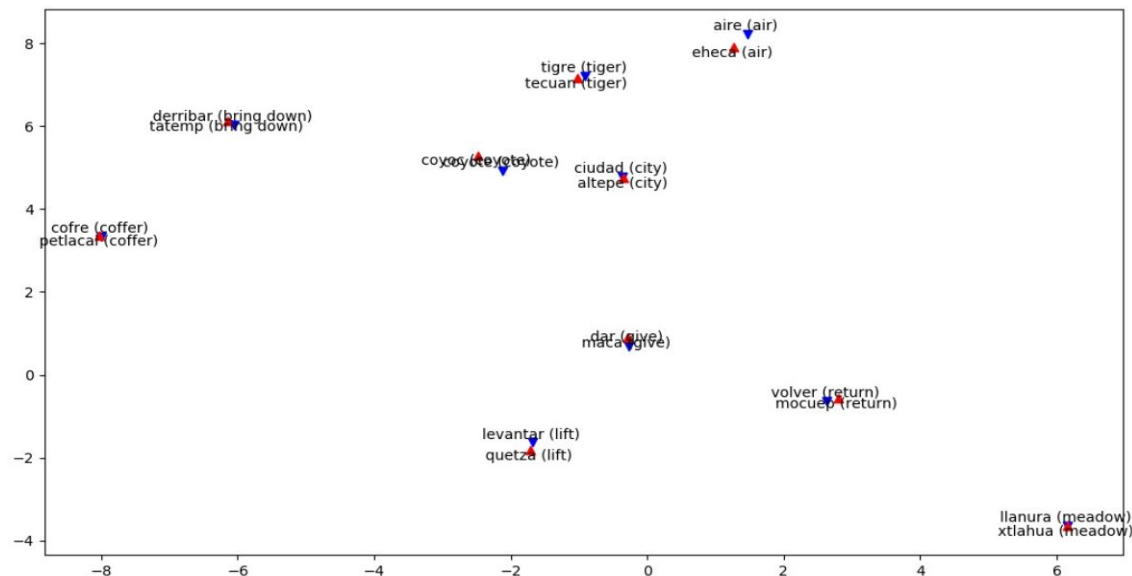


[elotl.mx](http://elotl.mx)

- Tecnología para las lenguas habladas en México
- Difusión
- Software libre, APIs, corpus, taggers

## Mi perfil. Formación previa

→ Doctorado en Ciencias de la Computación (UNAM)



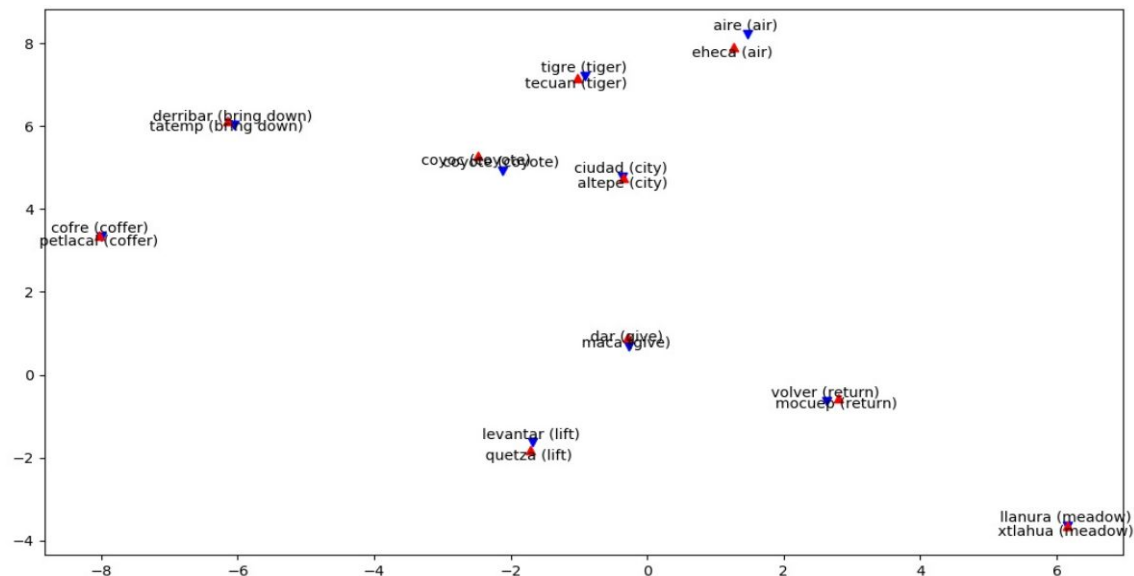
Vectores node2vec  
español-náhuatl

Gutierrez-Vasques, M. X. *Extracción léxica bilingüe automática para lenguas de bajos recursos digitales*. UNAM (2018)

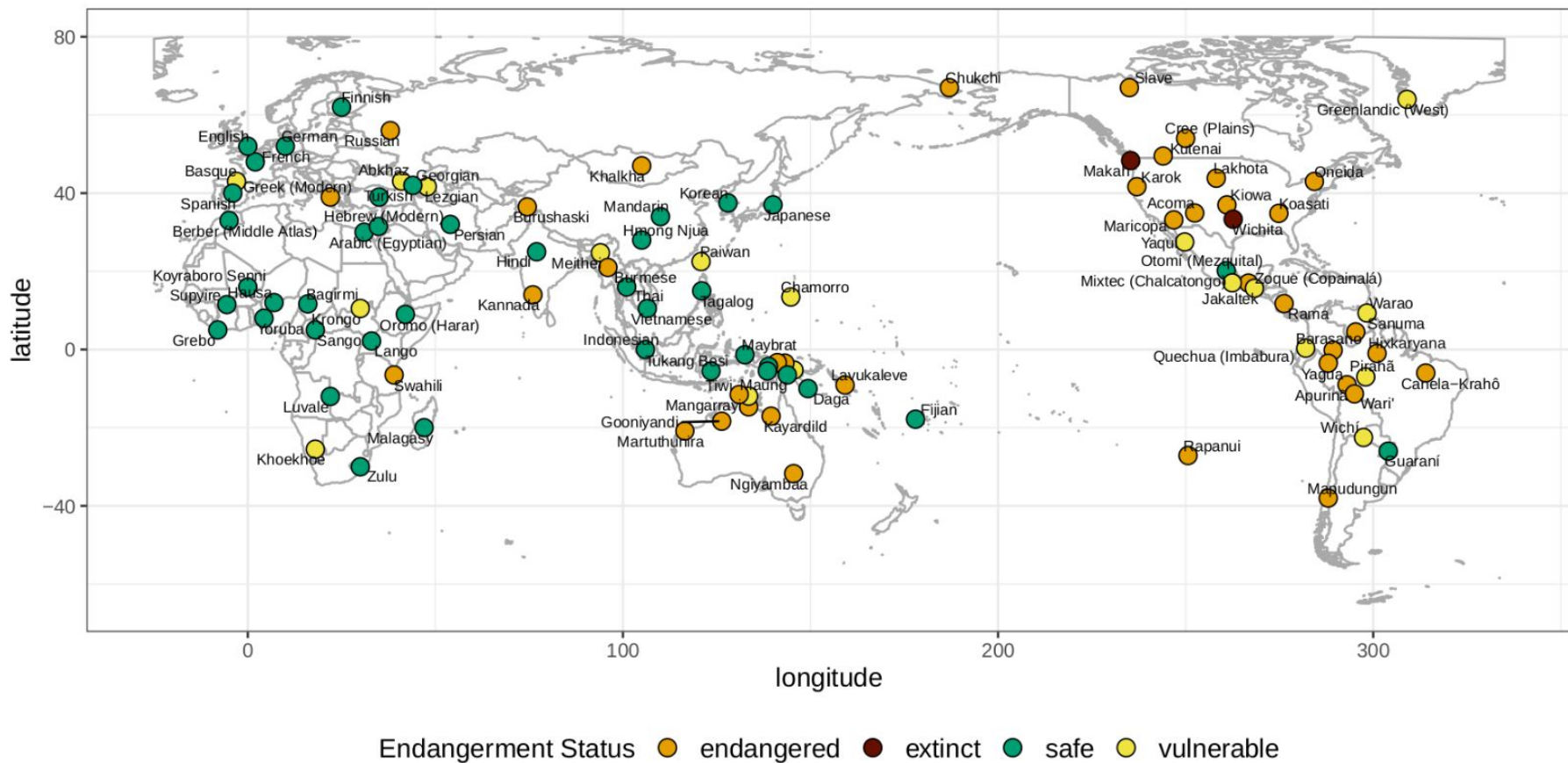
Gutierrez-Vasques, X., Medina-Urrea, A., & Sierra, G. (2019). *Morphological segmentation for extracting Spanish-Nahuatl bilingual lexicon*.

## Mi perfil. Formación previa

→ Doctorado en Ciencias de la Computación (UNAM)



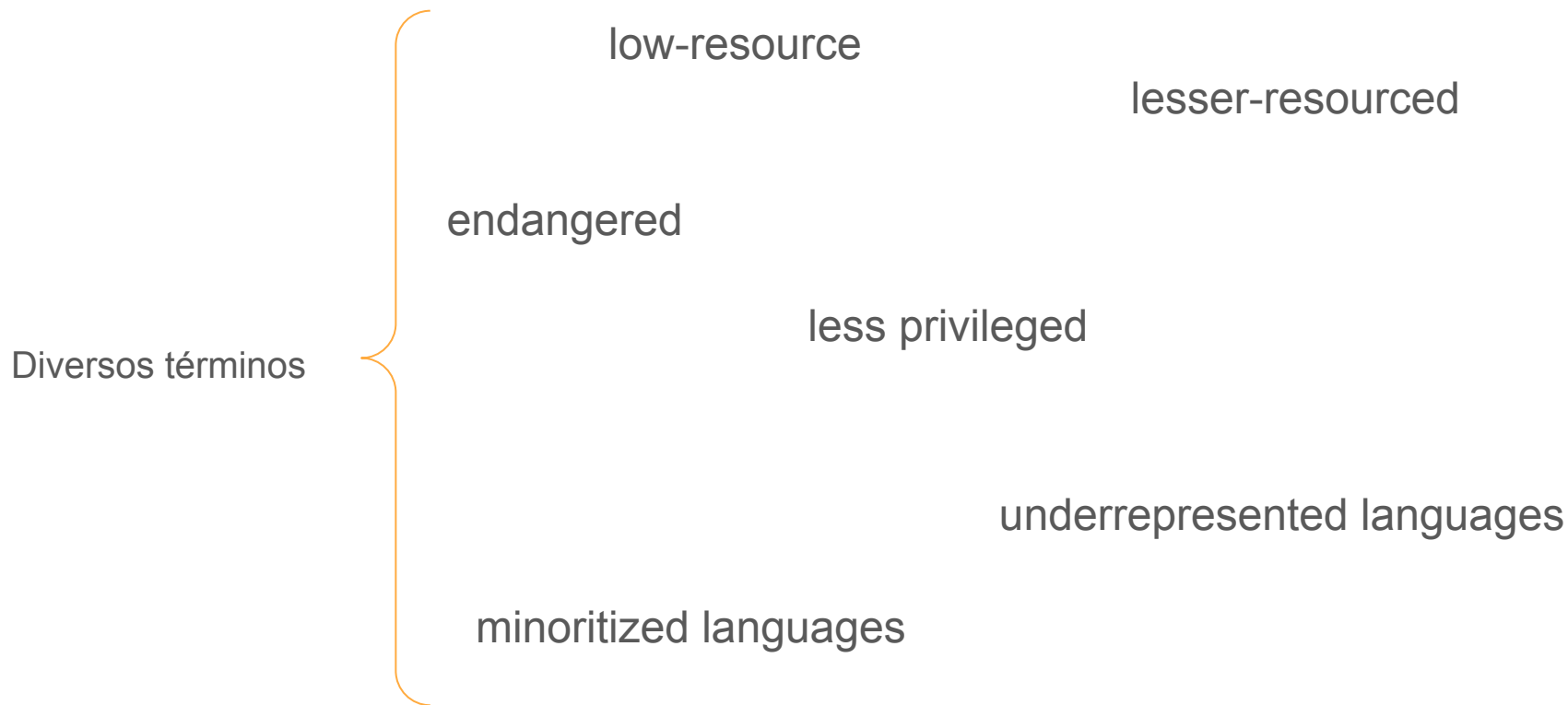
Gutierrez-Vasques, X., Sierra, G., & Pompa, I. H. (2016, May). Axolotl: a web accessible parallel corpus for Spanish-Nahuatl.



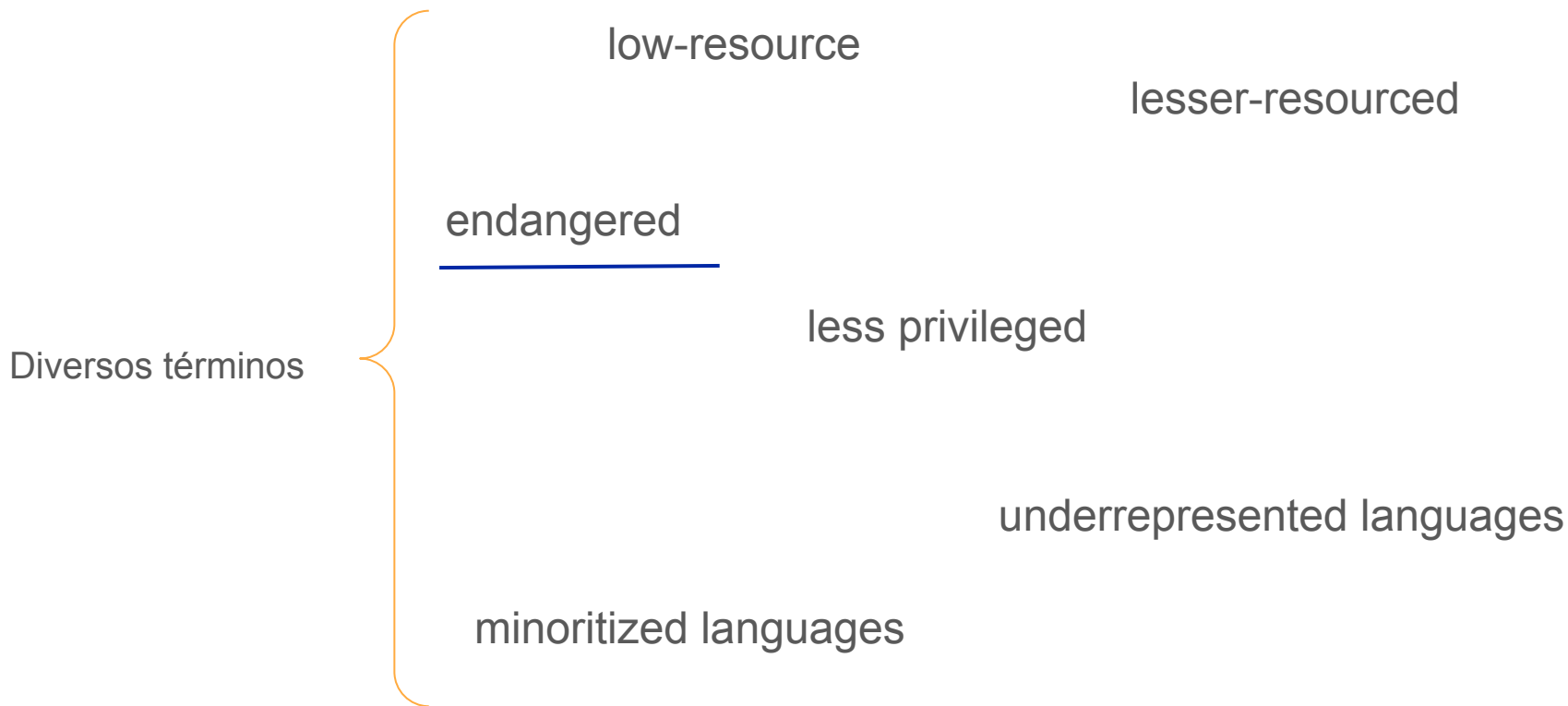
*Non-randomness in Morphological Diversity: A Computational Approach Based on Multilingual Corpora*

<https://www.spur.uzh.ch/en/departments/research/textgroup/MorphDiv.html>

## Low-resource languages?



## Low-resource languages?



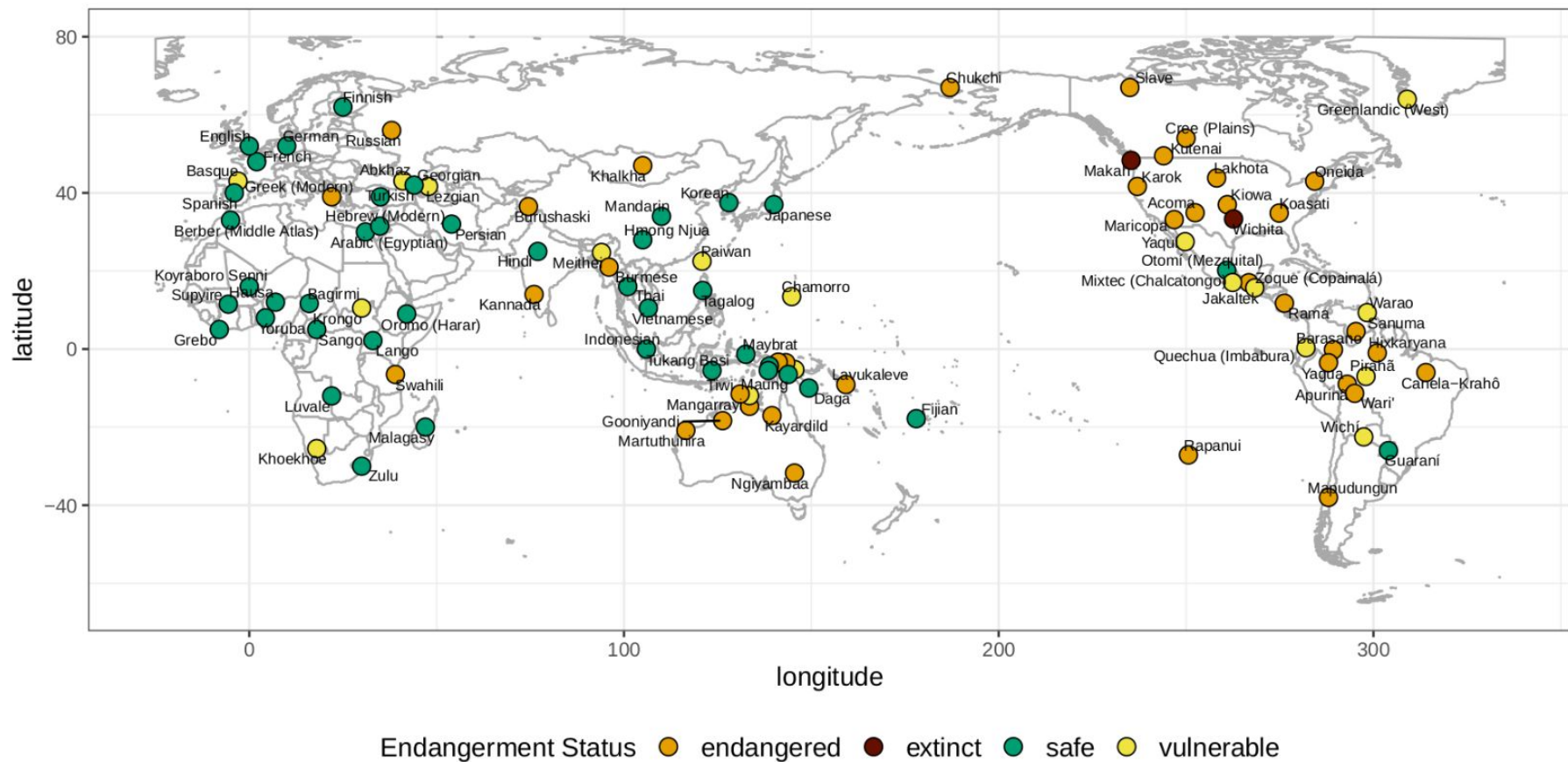
## Low-resource languages?

Endangered/  
lenguas en peligro  
de extinción

Algunos factores a considerar según UNESCO (2003):

- 1) Transmisión **intergeneracional** de la lengua
- 2) Número absoluto de **hablantes**
- 3) Proporción de hablantes dentro de la **población** total
- 5) Respuesta a nuevos **dominios y medios**
- 6) Materiales para **enseñanza** de la lengua y alfabetización
- 7) Actitud y **políticas** para el lenguaje por parte de Gobierno e instituciones
- 8) **Actitud** de miembros de la comunidad hacia su propia lengua





*Mapa de Non-randomness in Morphological Diversity: A Computational Approach Based on Multilingual Corpora*

<https://www.spur.uzh.ch/en/departments/research/textgroup/MorphDiv.html>

## Low-resource languages?

A veces los términos/adjetivos reflejan posiciones de poder que pueden **perpetuar narrativas de exclusión**, sin cuestionar realmente las razones por las que estas lenguas poseen un estatus menos privilegiado (*Cynthia Montaña, University of California, Berkeley*)

## Low-resource languages?

A veces los términos/adjetivos reflejan posiciones de poder que pueden **perpetuar narrativas de exclusión**, sin cuestionar realmente las razones por las que estas lenguas poseen un estatus menos privilegiado (*Cynthia Montaña, University of California, Berkeley*)

✓ Tener en mente:

- **No solo es un problema de escasez de datos**, del tamaño de un dataset o de cubrir el mayor de lenguas posible
- Detrás hay desigualdades sistémicas, disparidades, marginalización, etc
- Considerar el contexto de los hablantes cuando queremos hacer un acercamiento a través de las tecnologías del lenguaje.

## 1. Alentar grupos diversos en la comunidad de NLP

*“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)

## 1. Alentar grupos diversos en la comunidad de NLP

**Hugging Face:** Un paso hacia la democratización del NLP y el Machine Learning



## 1. Alentar grupos diversos en la comunidad de NLP

El caso de México

68 Agrupaciones lingüísticas	
364 Variantes	
11 familias lingüísticas	
(\_/)	
(•^•)	
/	づ

## 1. Alentar grupos diversos en la comunidad de NLP



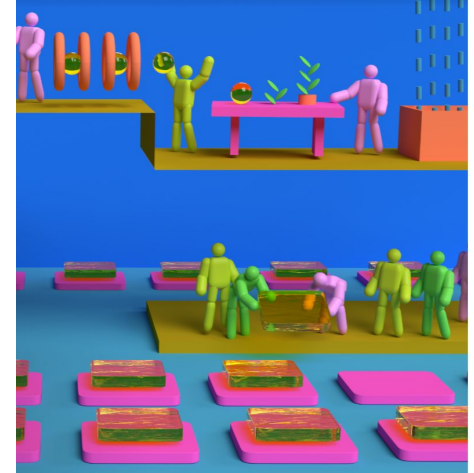
- No es un asunto de grupos vulnerables recibiendo tecnología **pasivamente**, sino de **fortalecer un diálogo intercultural** sobre cómo hacemos tecnología
- **Prácticas de cooperación** (que las comunidades indígenas tienen muy arraigadas como método de supervivencia) pueden influenciar la forma en que hacemos tecnología

*Yasnaya Aguilar, lingüista y activista digital Mixe*

## 1. Alentar grupos diversos en la comunidad de NLP

### Tequio-logías

*“La tecnología vista como tequio, la creación tecnológica como un bien común y de código abierto del que podemos participar [...]”*



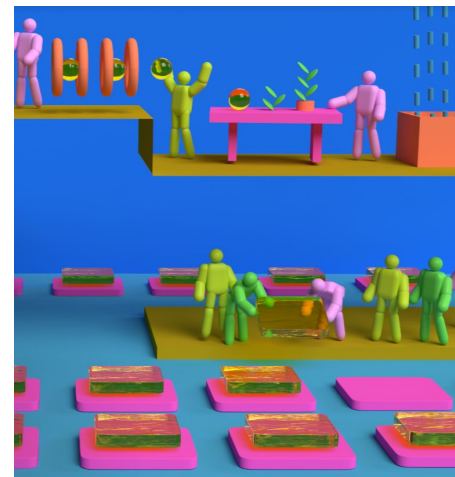


## 1. Alentar grupos diversos en la comunidad de NLP

### Tequio-logías

*“La tecnología vista como tequio, la creación tecnológica como un bien común y de código abierto del que podemos participar [...]”*

✓ Compatible con la filosofía del software libre



*Una propuesta modesta para salvar al mundo* <https://restofworld.org/2020/tecnologia-tequio-cambio-climatico/>

## 2. Disponibilidad de los recursos

- **Liberar** el código, modelos y conjuntos de datos es considerado una **buena práctica en NLP** para asegurar la reproducibilidad

## 2. Disponibilidad de los recursos

- **Liberar** el código, modelos y conjuntos de datos es considerado una **buena práctica en NLP** para asegurar la reproducibilidad
- ...pero esto se vuelve **crucial** cuando se trabaja con lenguas minorizadas o en peligro de extinción. Implicaciones éticas:
  - Riesgo de apropiación lingüística/cultural con el fin de obtener mérito académico ([Hämäläinen, 2021](#))
  - Algunos incluso señalan la necesidad de adaptar el licenciamiento para reflejar el compromiso con la comunidad. Ejemplo “Guelaguetza clause” (Reciprocidad con la comunidad) ([Washington et al., 2021](#))

*Washington, J., Lopez, F., and Lillehaugen, B. (2021). Towards a morphological transducer and orthography converter for western tlacolula valley zapotec*  
*Hämäläinen, M. (2021). Endangered languages are not low-resourced!*

### 3. “Abrazar” la variación

- La **gran variación** de las lenguas puede ser concebida como algo “**atípico**” desde el NLP

### 3. “Abrazar” la variación

- La **gran variación** de las lenguas puede ser concebida como algo “**atípico**” desde el NLP
- Sin embargo, el concepto de **estandarización** es relativamente reciente (Europa siglo XIX) → esfuerzos por unificar/estandarizar las expresiones lingüísticas

	Standard	Non-standard
Good	<ul style="list-style-type: none"><li>• Broader unity</li><li>• Clarity</li><li>• Easy to process</li></ul>	<ul style="list-style-type: none"><li>• Expressiveness</li><li>• Fun</li><li>• Local identity</li></ul>
Bad	<ul style="list-style-type: none"><li>• Pressure on minorities</li><li>• Flatness</li><li>• Hard to maintain</li></ul>	<ul style="list-style-type: none"><li>• Hard to process</li><li>• Hard to process!!</li><li>• Hard to process!!!</li></ul>

*Language (de)standardisation and NLP*  
<https://github.com/tsamardzic/nonstandard>  
*Tanja Samardžić, Tutorial at the Mexican NLP Summer School, 2 June 2021*

### 3. “Abrazar” la variación

- Aunque la variación es perfectamente **normal**, ciertamente representa un **gran reto** para la mayor parte de pipelines de NLP
- Nuestro rol **no es imponer una estandarización**, más bien entender el contexto sociolingüístico y desarrollar herramientas que puedan lidiar con esta variación

### 3. “Abrazar” la variación

Ejemplo. Módulo para ortografías del náhuatl:

```
import elotl.corpus
import elotl.nahuatl.orthography
```

- **sep**

*yujki*

- **inali**

*yuhki*

- **ack**

*yuhqui*

\*Representación  
fonológica

yuʔki

### 3. “Abrazar” la variación

Ejemplo. Módulo para ortografías del náhuatl:

```
import elotl.corpus
import elotl.nahuatl.orthography
```

Python package for Natural Language Processing (NLP), focused on low-resource languages spoken in Mexico.

This is a project of [Comunidad Elotl](#).

Developed by:

- Paul Aguilar [@penserbjorne](#), [paul.aguilar.enriquez@hotmail.com](mailto:paul.aguilar.enriquez@hotmail.com) ✉
- Robert Pugh [@Lguyogiro](#), [robertpugh408@gmail.com](mailto:robertpugh408@gmail.com) ✉

- **sep**

*yujki*

- **inali**

*yuhki*

- **ack**

*yuhqui*

\*Representación  
fonológica

yu?ki



## Extra. Algunos recursos

### Shared task: Open Machine translation



✓ Pares de lenguas:

Quechua–Spanish

Wixarika–Spanish,

Shipibo-Konibo–Spanish

Asháninka–Spanish

Raramuri–Spanish

Nahuatl–Spanish

Otomí–Spanish,

Aymara–Spanish,

Guarani–Spanish

Bribri–Spanish

Data:

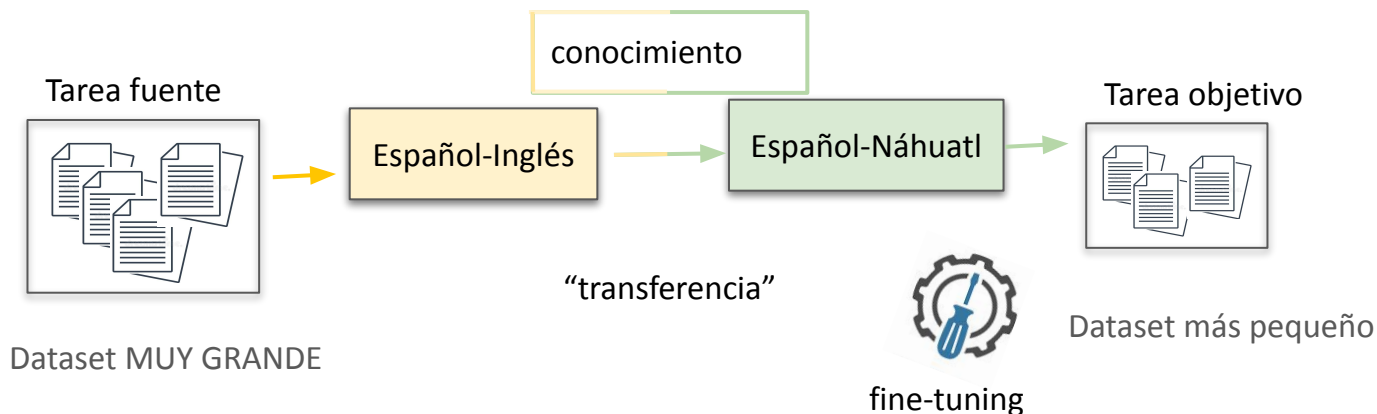
<https://github.com/AmericasNLP/americasnlp2021>



## Extra. Algunos recursos

Sistema ganador (Open Machine Translation): Helsinki Team ([Vazquez et., al, 2021](#))

- Transformer multilingüe, **pre-entrenado Español-Inglés** y *fine-tuned* en las 10 lenguas indomaericanas
- **Preprocesamiento** extensivo: recolección de más datos, normalización ortográfica, estimación de ruido/calidad de los datasets




## Extra. Algunos recursos

Sistema ganador (Open Machine Translation): Helsinki Team ([Vazquez et., al, 2021](#))

Demo

Spanish ▾

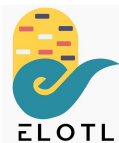
Aymara ▾



<https://translate.ling.helsinki.fi/ui/americanlp>

## Extra. Algunos recursos

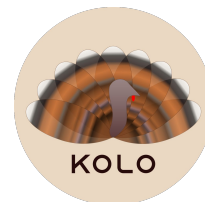
Comunidad Elotl



- Corpus paralelo: español-otomi
- Hablantes: ~300,000
- <https://tsunkua.elotl.mx>



- Corpus paralelo: español-mixteco
- Hablantes: ~600,000
- <https://kolo.elotl.mx>



- Python library
- <https://pypi.org/project/elotl/>

- ESQUITE: Framework abierto para corpus paralelos
- <https://github.com/ElotlMX/Esquite>

## Extra. Algunos recursos

Common voice



Localización mozilla



## Extra. Algunos recursos

### Story Weaver

- Traducir colaborativamente cuentos a muchas lenguas del mundo
- Recientemente se organizó un gran maratón multilingüe en México
- Software libre
- Los libros tienen licencia Creative Commons CC BY 4.0



¡Piyali! Kwalli xiahsikan nochan. Nonana, notata  
iwan Xochitl nemih nowan.



¡Hola! Bienvenido a mi casa. Mi mamá, mi papá y Cheena viven conmigo.

## Extra. Algunos recursos

- [Masakhane.io](https://Masakhane.io) *“A grassroots NLP community for Africa, by Africans”*
  - *An open-source, continent-wide, distributed, online research effort for machine translation for African languages*



## Extra. Algunos recursos

- Recopilación sobre la investigación y herramientas de NLP para lenguas habladas en América (continente)
- <https://github.com/pywirrarika/naki>



## Comentarios finales

- **Diversidad** no solo en los datasets... también en los grupos que pueden acceder a este conocimiento especializado
- **Inspiración** en prácticas de cooperación y reciprocidad para desarrollar tecnología
- La **variación es normal**... y un reto muy interesante para aplicar nuestra creatividad e innovación científica!

NLP realmente **multilingüe**

**¡Gracias!**

**Tlasohkamati**