Comunidad Elotl
Language Technologies for Mexico's Indigenous Languages

elotl.mx ✉contacto@elotl.mx @elotl_

2019 | INTERNATIONAL YEAR OF
Indigenous Languages

Language diversity in Mexico

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

- Very few (or none) **language technologies** have been developed for these languages

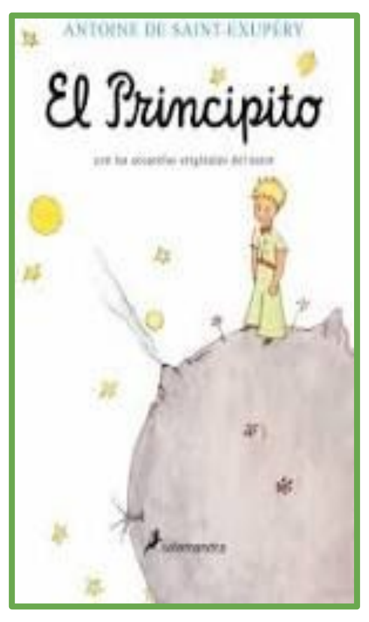
About us

- Community interested in **Natural Language Processing** (NLP) applied to Mexico's languages (mainly text based)
- **Make visible** these topics in **Government** and **Scientific** agendas
- We **disseminate** cultural, linguistic and technological information related to Mexico's languages
- Our Community is formed by **volunteers, students, researchers**
- In constant search for **collaborators, donations**

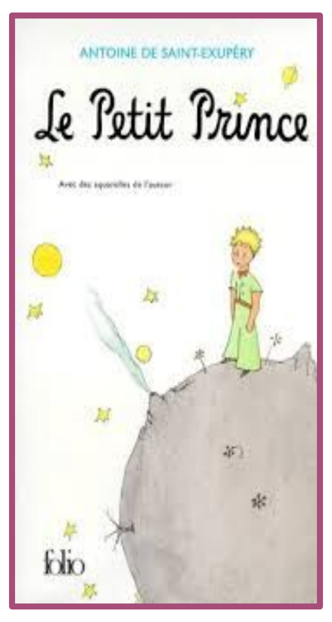
Technological challenges

- **Difficult to find digital content on the web**: Government, touristic, education websites rarely translate their content to the national languages
- Big **dialectal and orthographic variation** in the texts. Lack of orthographic standardization
- Scarcity of **pre-processing language tools** (taggers, morphological analyzers, linguistic annotated datasets)
- **Scarcity of data:** Popular NLP/ML methods do not work well in low-resource settings

→ Research & Development for these languages can have a positive **social impact**, but it is also a **technological and scientific challenge**

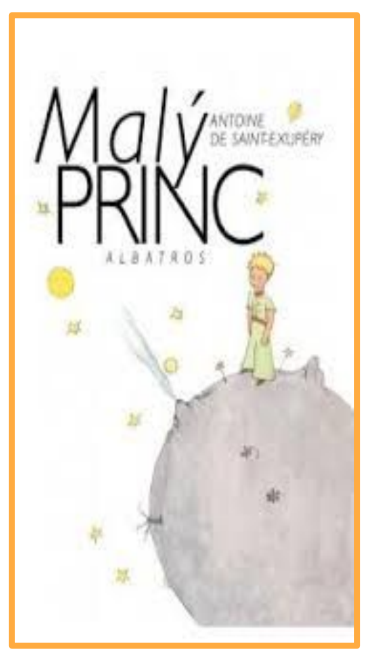


(Spanish)



(French)

(Nahuatl)



(Czech)

(Otomi)

Example of a parallel corpus. Valuable resource for:

- Building multilingual technologies: machine translation, bilingual lexicon extraction
- Linguistic studies, second language learning

We developed **digital parallel corpora** for several **language pairs** (and their **search interfaces**)

Otomi-Spanish parallel corpus

- Otomi: **Oto-Manguean** language family
- Speakers: ~**300,000**
- Parallel sentences: ~**5,000**

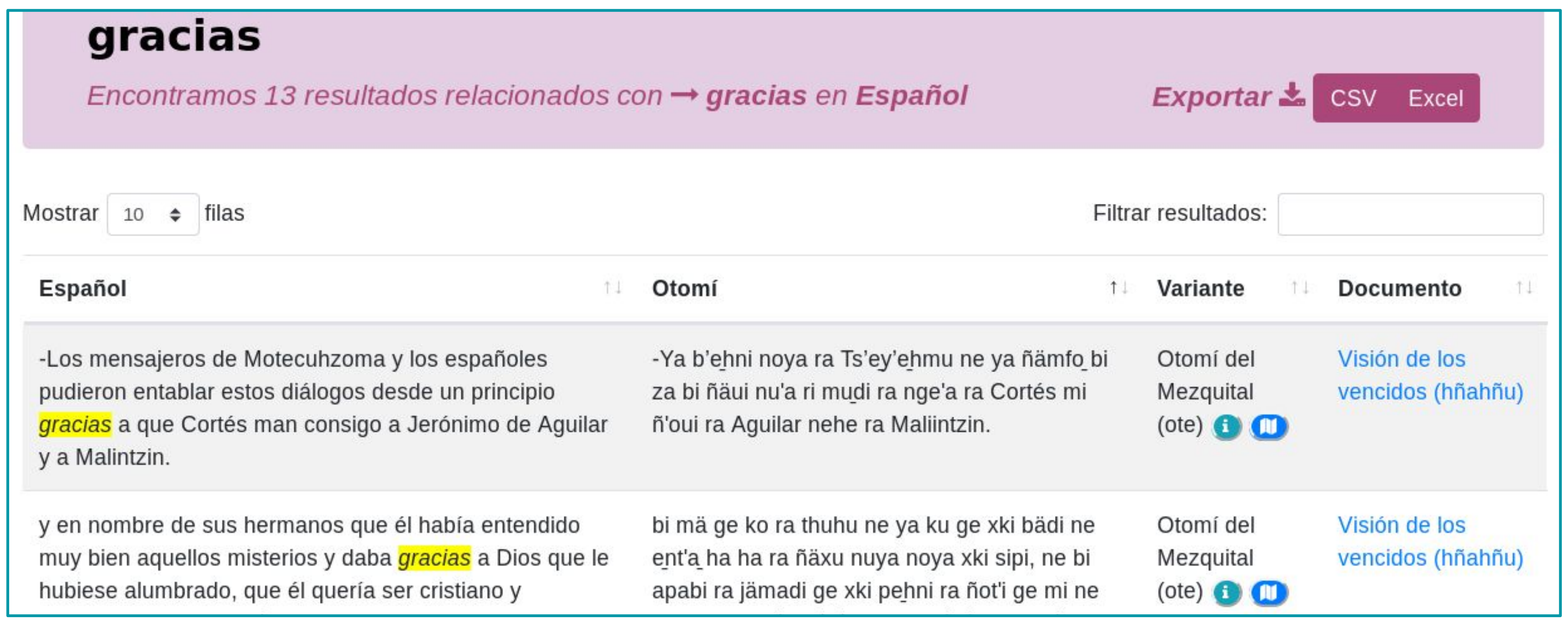
https://tsunkua.elotl.mx/

Nahuatl-Spanish parallel corpus

- Otomi: **Uto-Aztecan** language family
- Speakers: ~**1.7 M**
- Parallel sentences: ~**18,000**

http://www.corpus.unam.mx/axolotl

Our web systems allow to search within the corpora for words or phrases:



gracias

*Encontramos 13 resultados relacionados con → **gracias** en **Español***

Español	Otomí	Variante	Documento
-Los mensajeros de Motecuhzoma y los españoles pudieron entablar estos diálogos desde un principio *gracias* a que Cortés man consigo a Jerónimo de Aguilar y a Malintzin.	-Ya b'ehni noya ra Ts'ey'ehmu ne ya ñämfo̱ bi za bi ñäui nu'a ri mu̱di ra nge'a ra Cortés mi ñ'oui ra Aguilar nehe ra Maliintzin.	Otomí del Mezquital (ote)	Visión de los vencidos (hñahñu)
y en nombre de sus hermanos que él había entendido muy bien aquellos misterios y daba *gracias* a Dios que le hubiese alumbrado, que él quería ser cristiano y	bi mä ge ko ra thuhu ne ya ku ge xki bädi ne ent'a̱ ha ha ra ñäxu nuya noya xki sipi, ne bi apabi ra jämadi ge xki pehni ra ñot'i ge mi ne	Otomí del Mezquital (ote)	Visión de los vencidos (hñahñu)

Ongoing projects

- O'odam-Spanish parallel corpus (only 15, 000 speakers)
- Morphological gloss tagger for Otomi (using Conditional Random Fields)
- Python libraries for pre-processing these languages

Ximena Gutierrez-Vasques, Víctor Mijangos, Diego Alberto Barriga, Javier Santillán, Elesban Landero, Inocencia Arellanes, Cynthia Montaño, José Luis Olivares Castillo, Paul Aguilar