Mexican Indigenous Corpora

Ivan Vladimir Meza Ruiz
IIMAS-UNAM

Close collaborators

- Jesús Mager, soon in University of Stuttgart
- Jorge García, CNRS LIPN-Paris 13
- Juan de Jesús Amador, UAEM CNRS LIPN-Paris 13
- Nadi Tomeh, CNRS LIPN-Paris 13
- Miguel Soriano, Centro Universitario de los Lagos
- Students at LIPN-Paris 13

Collaborators

- Ximena Gutierrez & Gerardo Sierra, II/UNAM
- Hinrich Schütze, LMU Munich
- Katharina Kann, NYU
- Dionicio Carrillo, Translator

Americas native languages

- 28 million speakers
- 900~1060 languages
- Cultural heritage
- Linguistic richness

This talk focus on Mexico

¡Welcome to México!

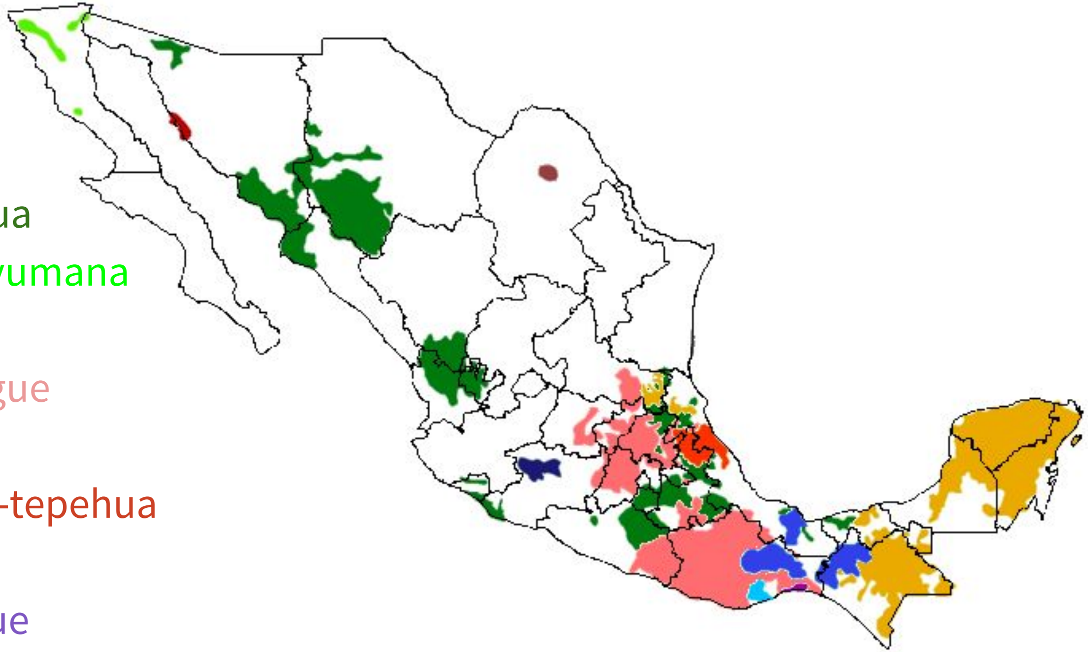
68 official language

Akateko, Amuzgas, Awakateco, Ayapaneco, Chatinas, Chichimeco jonas, Chinantecas, Chocholtecas, Ch’oles, Chontales de Oaxaca, Chontales de Tabasco, Chuj, Coras, Cucapá, Cuicatecas, Guarijías, Huastecas, Huaves, Huicolas, Ixcateco, Ixil, Jakalteco, Kaqchikel, K’iche’s, Kickapoo, Kiliwa, Ku’ahl, Kumiai, Lacandón, Mam, Matlatzinca, Maya, Mayo, Mazahuas, Mazatecas, Mixes, Mixtecas, Nahuas, Oluteco, Otomíes, Paipai, Pames, Pápago, Pimas, Popolocas, Popoluca de la sierra, Purépecha, Q’anjob’al, Qato’k, Q’echi’, Sayulteco, Seri, Tarahumaras, Teko, Tepehuas, Tepehuano del norte, Tepehuanas del sur, Texistepequeño, Tlahuica Tlapanecas, Tojolabal, Totonacas, Triquis, Tseltal, Tsoltsil, Yaqui, Zapotecas and Zoques

350 dialects

11 families

1. Álgica
2. Yuto-nahua
3. Cochimí-yumana
4. Seri
5. Oto-mangue
6. Maya
7. Totonaco-tepehua
8. Tarasca
9. Mixe-zoque
10. Chontal de Oaxaca
11. Huave



Populations

- 7,382,785 speakers (51% women, 49% man)

Nahuatl	1,725,620
Maya	859,607
Tseltal	556,720
Mixteco	517,665
Tzotzil	487,898
...	...
Ayapaneco	24
Aguacateco (Awakateko)	17

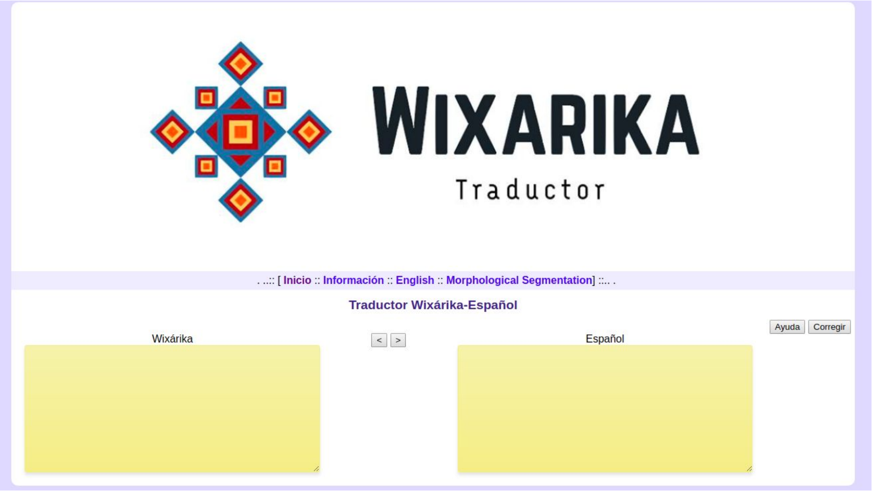
- 13% monolingue
- 45.3% younger than 30yo

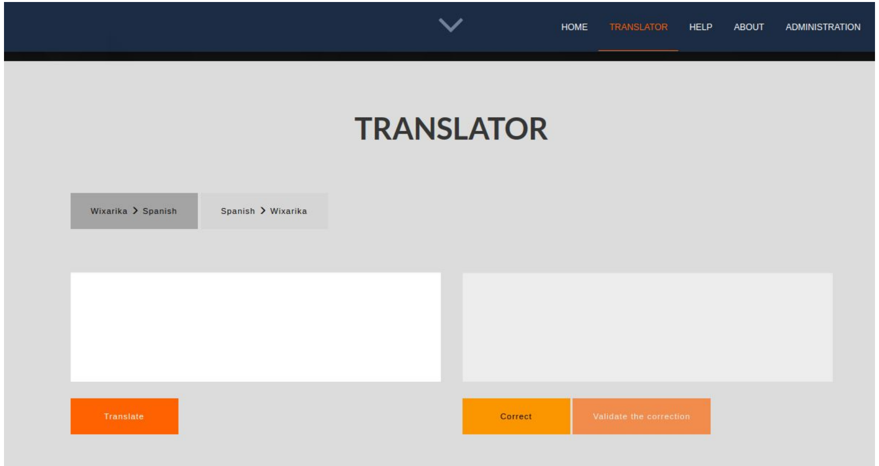
Little NLP

- Great tradition of lingüístic studies of Native Méxican languages … not so much from the technological point of view
- Research focus on: Morphology and Machine translation
- Mainly focused on: Yuto-nahua, Oto-mangue and Maya

Machine translation (MT)

- MT, Mayan and Ñañú (Otomi) from Microsoft
- MT, Wixarika-Spanish
- MT, Purepecha-Spanish
- Apps: Yalam (16 Oto-mange), Didxazapp (Diidx Zah [Zapoteco]), Doulingo (Nahuatl)*
- In development MT: Nahuatl-Spanish, Yorem Nokki-Spanish, Mexicanero-Spanish


WIXARIKA
Traductor
. ..:: [Inicio :: Información :: English :: Morphological Segmentation] ::.. .
Traductor Wixárika-Español
Ayuda
Corregir
Wixárika
<
>
Español



Our work

Wixarika-spanish, two versions

NLP Morphology

- Affixes discovery Huasteco, Chuj, Tojolabal, Yucateco,
- *Chachalaca* for Náhuatl
- WixNLP, FST for Wixarika
- GRU encoder-decorer, Wixarika, Mexicanero, Nahuatl and Yorem Nokki

Others

- Codeswitching, OCR; Latin, Spanish and Nahuatl
- OCR, Tzental-Spanish
- Language identification, 30 languages includes Nahuatl
- Tense behaviour, Bible corpus
- Speech synthesis, Raramuri

NLP Corpora

Type	Language	Dimension	Cite
MT	Nahuatl-Spanish	18K sentences	***Axolotl***, Ximena et al. (2016)
MT	Wixarika-Spanish	8k sentences	**Wixarikacorpus**, Mager et al. (2018)
MT	Mixteco-Spanish	1K sentences	***Ve'e Savi***, Santiago et al. (2017)
Dictionary	Nahuatl		Palancar and Feist (2015)
Morphological	20 Oto-Manguean languages	13, 000 verbal entries	Palancar and Feist (2015)
Morphological	Uto-Aztecan	4,468 word	Kann et al. (2018)

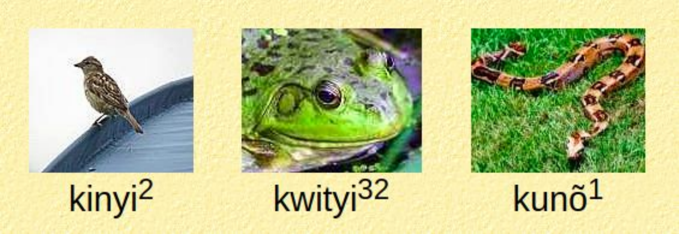
Our lab corpora

- Wixarika corpus
  - *segcorpus.wixes* high quality paired corpus with , with Wixarika word segmentation.
  - *corpora.wixes* Wixarika - Spanish paired corpus
  - *social.wix* Social networks mined wixarika text collection
  - *mixed.txt* Social networks mined text that contains wixarika, spanish and other languages in the same phrase.
  - *dictionary.wixes* Wixarika-spanish vocabulary
- Morphological segmentations
  - Basic 4k segmented: Mayo, Nahuatl, Mexicanero and Wixarika
- Parallel corpus
  - Mayo, Nahuatl, Mexicanero, Wixarika, Purepecha

Why has been difficult?

- Cultural discrimination
- Poor digitalization
  - Twitter: http://indigenoustweets.com/
  - Facebook: several communities
  - Whatsapp, youtube
- Technology lacking support
- Very little previous work

Challenges



Tomado de: http://www.native-languages.org/

- Low resources → big impact
  - Techniques that work for several languages?
- Rich linguistic variance, e.g., no orthographic standardization
  - Techniques that work for several dialects?
- Rich linguistic phenomena
  - Rich morphology, agglutination
  - Compositional semantics
  - Tonal languages
  - Whistling: amuzgo, chinanteco, chol, kickapoo, mazateco, náhuatl, otomí, tepehua, totonaca, zapoteca
  - Codeswitching

Our strategy

- Minimal corpus
  - Colegio de México, Yolanda Lastra, same sentences 35 books (5)
- The web as a corpus
  - Maya in collaboration with LIPN-Paris 13: 87,806
- Exploiting linguistic studies
  - OCR, done in Axolotl and Ve'e Savi
- Crowdsourcing a community
  - Wide spectrum telecommunications

Future of the effort

- ¡Collect data!
- Children short stories
- Translation between indigenous languages, business oriented
- Medical, education and law domains
- Government domain
- ¡Collect data!

Official references and advertisement

- Diario Oficial de la Federación. *Catálogo de las Lenguas Indígenas Nacionales: Variantes Lingüísticas de México con sus autodeterminaciones y referencias geoestadísticas*. 2013.

- Diario Oficial de la Federación. LEY GENERAL DE DERECHOS LINGÜÍSTICOS DE LOS PUEBLOS INDÍGENAS. 2003.

- INEGI. *ESTADÍSTICAS A PROPÓSITO DEL DÍA INTERNACIONAL DE LOS PUEBLOS INDÍGENAS*. 2016.

- Mager, J., Gutierrez, X., Meza, I. and Sierra, G.: Challenges of language technologies for the Americas Indigenous Languages, COLING 2018, 2018, To appear

Gracias

Thank you