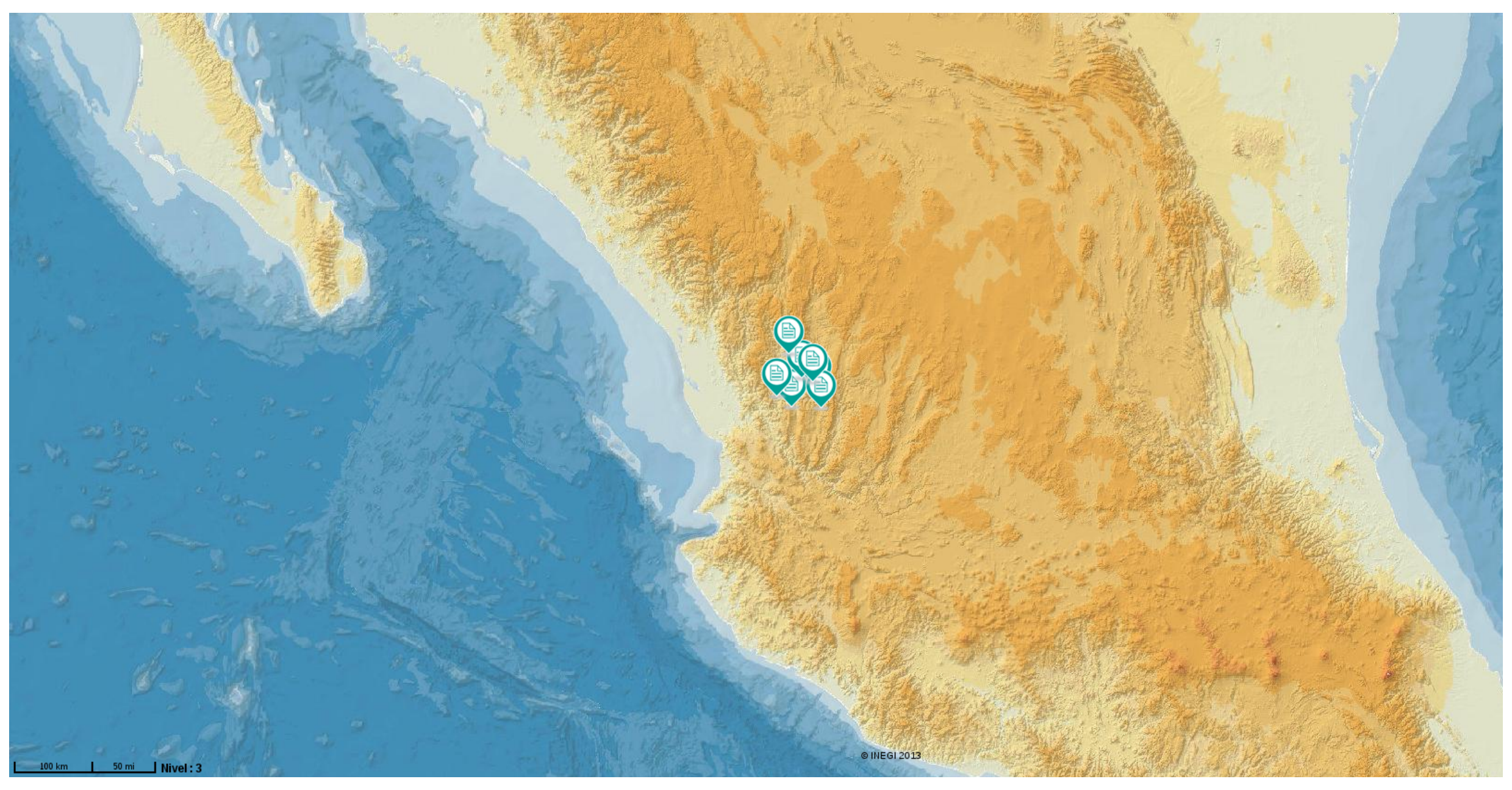
nlpwix

# Probabilistic Finite-State morphological analyzer for Wixarika language

Manuel Mager*, Dionico Carrillo, Ivan Meza**

*Universidad Autónoma Metropolitana (UAM), Unidad Azcapotzalco
**Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS)
Universidad Nacional Autónoma de México (UNAM)

We present the first morphological analyzer for the Mexican indigenous language Wixarika, also known as Huichol. Wixarika is a yutonahua language which has a complex agglutinative verbal morphology. The low amount of resources and the lack of a orthographic standard among dialects add to the challenge. Our proposal is based on a probabilistic finite-state approach that exploits regular agglutinative patterns and requires little linguistic knowledge.



Localization of the main Wixaritari communities in Mexico

## Wixarika Language

Wixarika is a language spoken in the Mexican states of Jalisco, Nayarit, Durango and Zacatecas (in central west Mexico) by approximately fifty thousand people and has complex verbal morphology, e.g.:

**nep+ka'ukats+k+**
*Wixarika word*

**ne | p+ | ka | 'u | ka | ts+k+**
*Wegmented wixarika word*

**I don't have a dog**
*English translation*

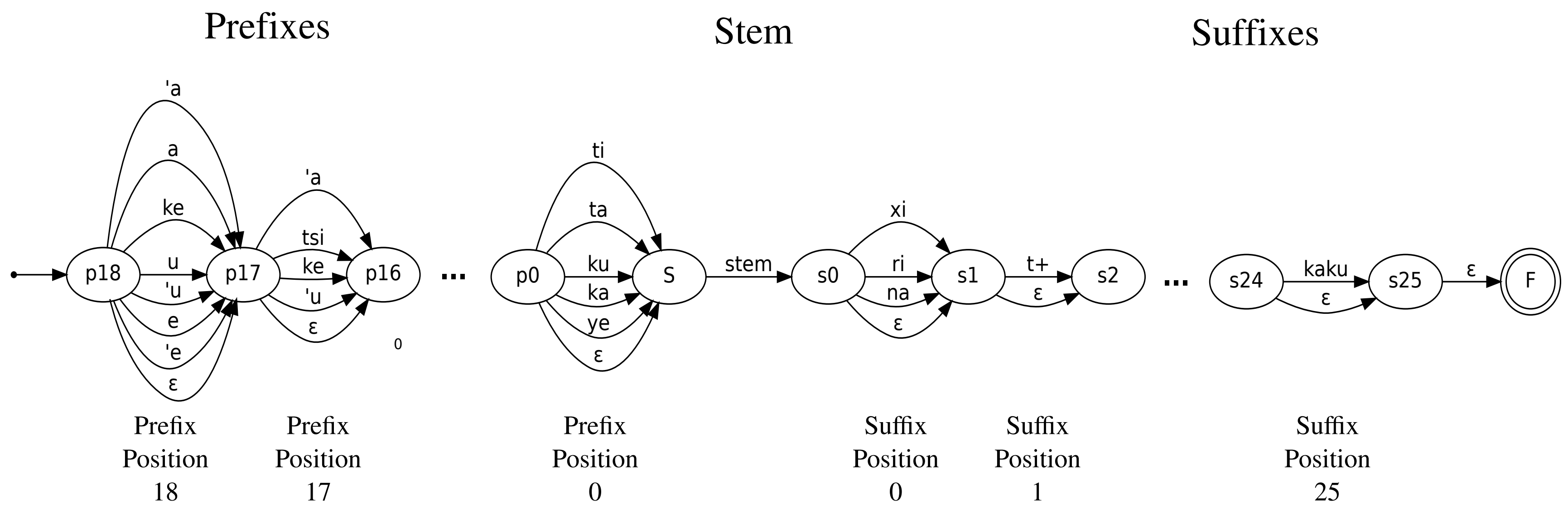
Its spelling is still not standardized. The most common spelling in practice by native speakers is an alphabet of 18 symbols:

**a, e, h, i, +, k, m, n, p, r, t, s, u, w, x, y, ’**

## Method

There are 8 such prefix positions and 23 suffix positions where each position allows a certain set of morphemes (or can be left empty). This can be used to construct a FSM from a list of legal morphemes at each position. We will assume that the only condition is that each position allows only morphemes from its list. The errors introduced by this assumption will be corrected later by the *n*-gram model
.

In practice, there are few analyses that we can enumerate all ofthem. To choose the most probable analysis from among these, we used a simple *n*-gram model with Kneser-Ney smoothing, where each gram is a morph.



Extract of the FSA for Wixarika verbs. The ``stem'' arc stands for a collection of 374 arcs representing different stems.

## Results

We use a high-quality segmented text containing 1079 unique words, which we used as our gold standard. We randomly extracted 400 words from this collection, to be used as a test set, and the rest were used for the training of a semi-supervised Morfessor model and our *n*-gram model, and a translation of Hans Christian Andersen's to Wixarika containing an estimation of 47,131 segmentable words, used for the training of the unsupervised Morfessor model.

| Method | ED | 1-best |
|---|---|---|
| **Morfessor** | 64.95 | 0.213 |
| **Morfessor SS** | 49.93 | 0.355 |
| **WixNLP** | 41.77 | 0.477 |
| **WixNLP 2-grams** | 39.16 | 0.485 |
| **WixNLP 3-grams** | 32.48 | 0.579 |
| **Hybrid 2-grams** | 31.48 | 0.562 |
| **Hybrid 3-grams** | 27.85 | 0.599 |

Results for the morphological segmentation task on Wixarika using direct comparison to the gold segmentation: Edit distance (ED) and error rate (1-best).

| Method | P | R | F |
|---|---|---|---|
| **Morfessor** | 0.508 | 0.48 | 0.493 |
| **Morfessor SS** | 0.648 | 0.626 | 0.637 |
| **WixNLP** | 0.666 | 0.724 | 0.694 |
| **WixNLP 2-grams** | 0.697 | 0.733 | 0.71 |
| **WixNLP 3-grams** | 0.726 | 0.757 | 0.742 |
| **Hybrid 2-grams** | 0.739 | 0.773 | 0.756 |
| **Hybrid 3-grams** | 0.78 | 0.805 | 0.792 |

Results for the morphological segmentation task on Wixarika using EMMA metric. P stands for precision, R for recall and F for the F-measure.

## Conclusions

Morphological segmentation is an important task for language processing of indigenous languages. In this work we presented the first Wixarika morphology analyzer, a finite-state method. We showed that for Wixarika our method improves on the Morfessor baselines. We also created and publicly released a parallel Wixarika-Spanish dataset to encourage the community to study this language further.

## Acknowledgement

We thank the guidance, feedback and support of Professor Jason Eisner as a mentor of this work.Also to Armando Martínez how designed the wixnlp logo.