Lemmatizer for a Mayan Language

C. Bocionek, L. Heuschkel, A. Piwowarek
bocionek@coli.uni-saarland.de heuschkelliesa@gmail.com aleksandrapiwowarek@gmail.com

UNIVERSITÄT DES SAARLANDES

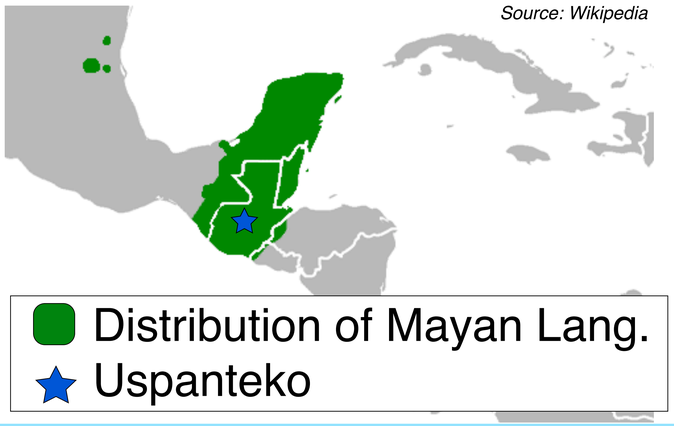
Objective

- develop a lemmatizer for a low-resource language

Language

- Uspanteko
- spoken in Guatemala
- ~ 3,000 native speakers
- endangered language

Kwand xink'uli'k' (When i married)



Def. Lemmatization

- determine the lemma for a given word
- lemma = dictionary form of a set of words

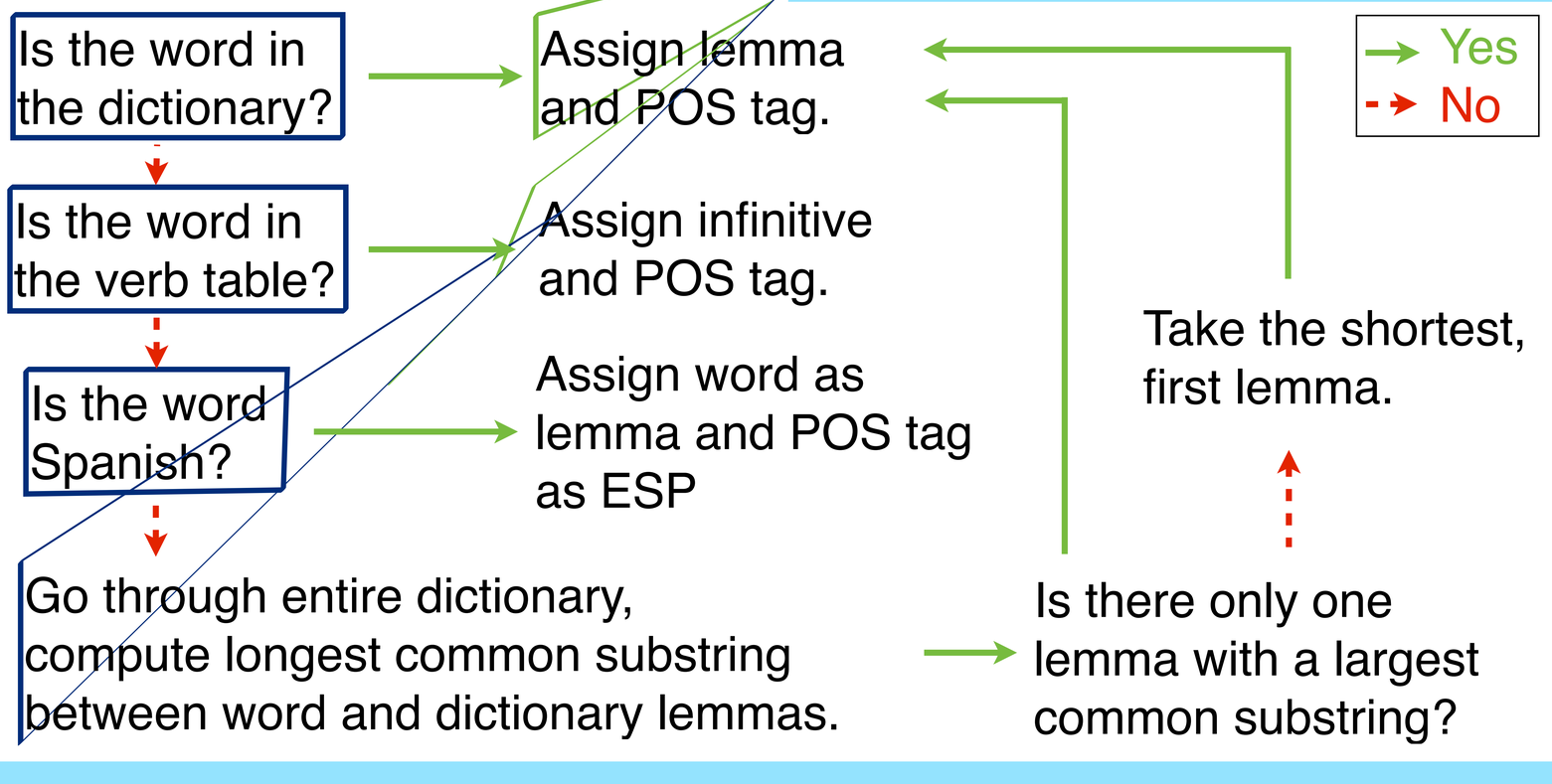
run, runs, running –> run

Resources

- corpus of 50,000 words (23 stories of transcribed audio)
  - ¾ training corpus
  - ¼ evaluation (planned)
- **dictionary** of 4,456 lemmas

Lemmatizer

- generation of **verb table** by retrieval of corpus verbs
- retrieval of **spanish words**
- lemmatization: decision graph



Evaluation

- manual annotation of 100 sentences (466 words)
- word by word comparison of lemmatized output

Best result: