Lemmatization of a Mayan Language

Liesa Heuschkel

The Java Rejects
Softwareproject: NLP tools for low resource languages

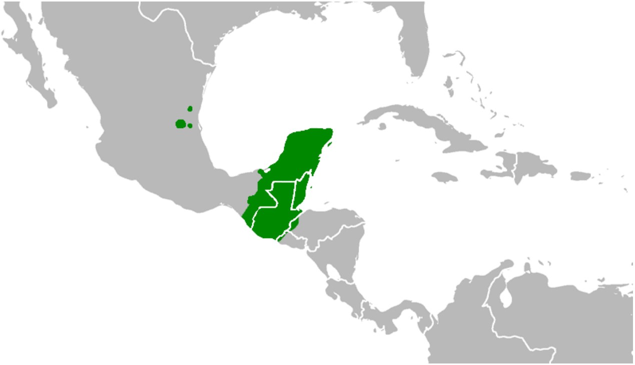
12.11.2013

Overview

- Analyze a corpus in a Mayan language
- Develop a lemmatizer
- Programming language: Python

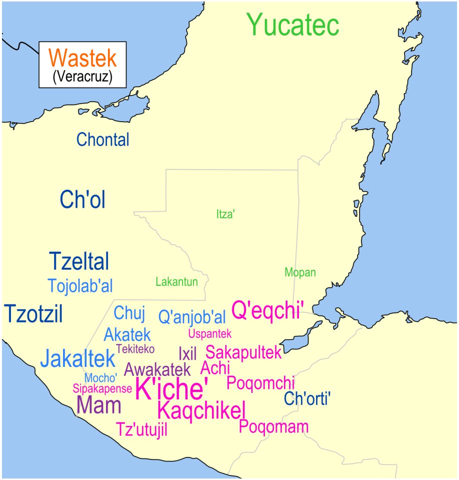
Mayan Languages

- Spoken in Mesoamerica and northern Central America
- 31 recognized languages
- Approximately 6 million speakers
- Subject of numerous studies



Uspanteko

- Mayan language of Guatemala
- 3,000 native speakers



Grammar of Mayan Languages

- Agglutinating: create new words by joining morphemes
- Polysynthetic: words are composed of many morphemes
- Use of relational nouns (instead of prepositions) to indicate spatial relationships
  - on (top of) the mountain –> its head the mountain
- Use of ergativity in grammatical treatment of verbs and their subjects and objects
  - I moved her/She moved –> I moved her/Her moved
- Specific inflectional categories on verbs
- Special word class of positionals, which reflect the position or shape of a given item
  - xoyan: curled up like a rope or snake

Motivation

- Mayan languages are endangered
- Projects for preservation
- Lack of NLP tools
- Agglutinating and polysynthetic properties make Mayan languages perfect candidates for lemmatization

Uspanteko Corpus

- Annotated Corpus of transcribed audio of native Uspanteko speakers
- Contains 23 stories in Uspanteko
- Total of 50,435 words

```
\ref trtex004Usp03 001
\t Kwand xink’uli’k’,+
\m kwand x- in- k’uli’ -ik
\g cuando COM A1S casarse SC
\c ADV TAM PERS VI SUF
\l Cuando me casé.
\p ADV COM#TAM A1S#PERS VI SC#SUF
```

\ref corpus-specific ref number

\t transcription

\m morphol. segmentation of \t

\g combination of grammatical glosses (like 'DEM') and stem translations

\c part-of-speech category line

\l translation into Spanish

\p combines \g and \c

Lemmatization

- Lemmatization = determining the **lemma** for a given word
- Lemma = dictionary form of a set of words, base form
  - *run, runs, ran, running –> run*
- Often used as a pre-step in text analysis
- Supports searches by reducing a set of words to their base form
- Building a lemmatizer for a new language is challenging
- Requires:
  - understanding of the context of the sentence
  - POS-tags for the words
  - dictionary for the given language.

Lemmatization

- Examples:
  - walking –> walk+**ing** –> walk/VB
  - desks –> desk+**s** –> desk/NN
- Problems
  - Homographs (words that share the same spelling, regardless of their pronunciation)
    - dove (*bird* or *past of dive*) –> dove/NN or dive/VP
    - dishes (*plural of dish* or *3rd person singular present tense of dish*)
  - Irregular Forms
    - dove –> dive
    - went –> go

Training a Lemmatizer

- Analyze corpus
- Find different prefixes, suffixes and stems
- Store data in files

Step-by-step lemmatization

- Go through corpus
- Tokenize line to words
- Look up morphemes and split word into morphemes
- Classify prefixes and suffixes to find stem
- Look up stem in dictionary to find lemma
- Look up word tag for this lemma
- Print lemma + POS tag
- Evaluation: Running lemmatizer on part of corpus, compare results with annotated data

Outlook

- Mayan languages have a similar grammar
- Why just develop a tool for one language?
- All Mayan languages are low resource languages
  - –> Make the lemmatizer adaptable for any Mayan language!
- How do we do this?
  - No set rule system *(e.g. Verb, 3rd Pers Sg, Present –> take off 's')*
  - Automatically recognize affixes
  - Use this knowledge for lemmatization

Where are we now?

- Implemented corpus reading methods
- Extracted all of the prefixes, suffixes and stems from the training corpus
- Convert Pdf Uspanteko Dictionary to Machine Readable txt File
- TO DO:
  - Learn more about the grammar
  - Learn more about lemmatization
  - Find out how to implement a structure for affixes (Suffix trees)

Discussion

Thank you for your attention!
Questions?

References

- Picture on slide 3
  http://en.wikipedia.org/wiki/File:Distribution-myn2.png

- Picture on slide 4
  http://en.wikipedia.org/wiki/File:Mayan_Language_Map.png

- Mayan Languages
  http://en.wikipedia.org/wiki/Mayan_languages

- Uspanteko
  http://de.wikipedia.org/wiki/Uspanteco

- Lemmatization
  http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

- M. Covington: How to make a lemmatizer
  http://www.ai.uga.edu/mc/8570/Lemmatizer.pdf