Analysis of Language Relatedness for the Development of Multilingual Automatic Speech Recognition for Ethiopian Languages

Martha Yifiru Tachbelie, Solomon Teferra Abate, Tanja Schultz

{Marthayifiru,abate,tanja.schultz}@uni-bremen.de

Abstract

In this poster, we present the analysis of GlobalPhone (GP) and speech corpora of Ethiopian languages (Amharic, Tigrigna, Oromo and Wolaytta). The aim is to select speech data from GP for the development of multilingual Automatic Speech Recognition (ASR) system for the Ethiopian languages. To this end, the phonetic overlaps among GP and Ethiopian languages have been analyzed. Moreover, morphological complexity of the GP and Ethiopian languages, reflected with high out of vocabulary rate and type to token ration, has been analyzed using training transcriptions. We also present baseline ASR performances for each of the GP and four Ethiopian languages.

Intro: Ethiopia

- More than 80 languages
- Almost all are under-resourced
- Poor to develop language resources for all languages
- Solution: *use resources of other languages to develop Speech and Language Processing Application*
  - *Develop multilingual system*

Aim

- Develop Multilingual Speech recognition systems for Ethiopian languages using resources of other languages

Existing Resources:

GlobalPhone

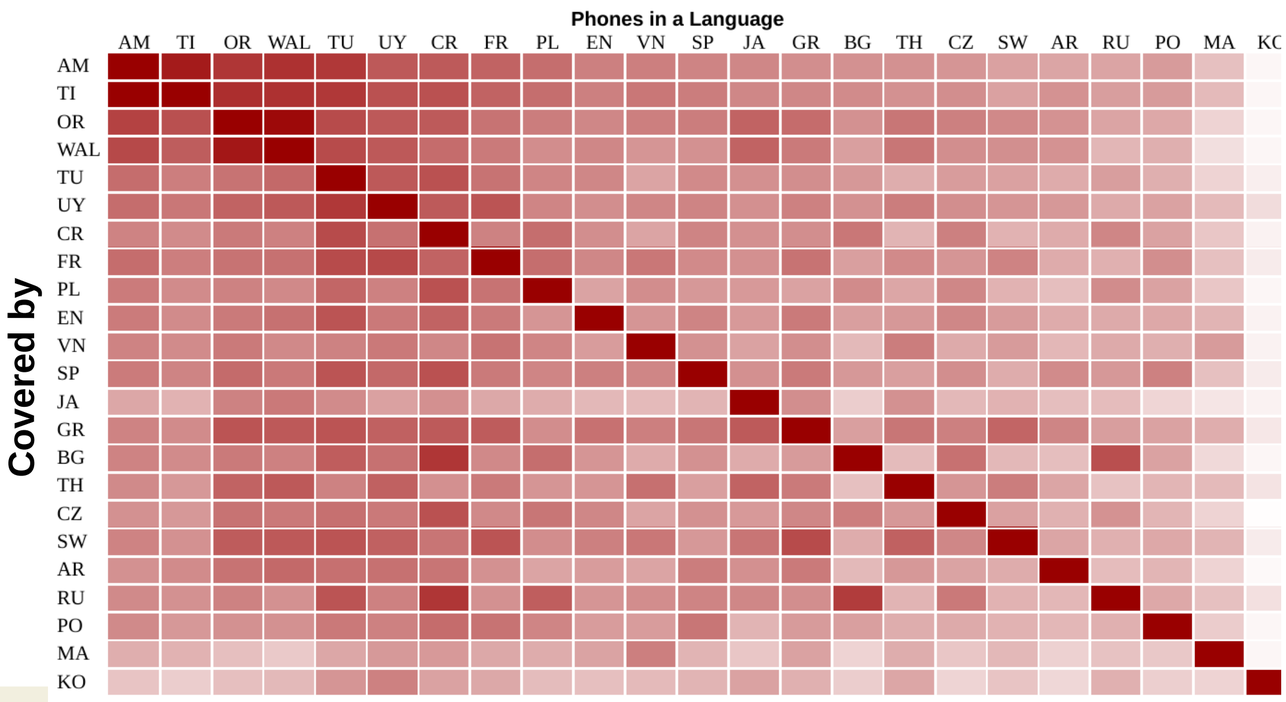
- A multilingual database of high-quality read speech with corresponding transcriptions and pronunciation dictionaries in more than 20 languages

Corpora of Ethiopian Languages

- Read Speech Corpus of four Ethiopian Languages: Amharic, Tigrigna, Oromo and Wolaytta

Which Language Resource to use?

- We have analyzed language relatedness among Globalphone and Ethiopian Languages

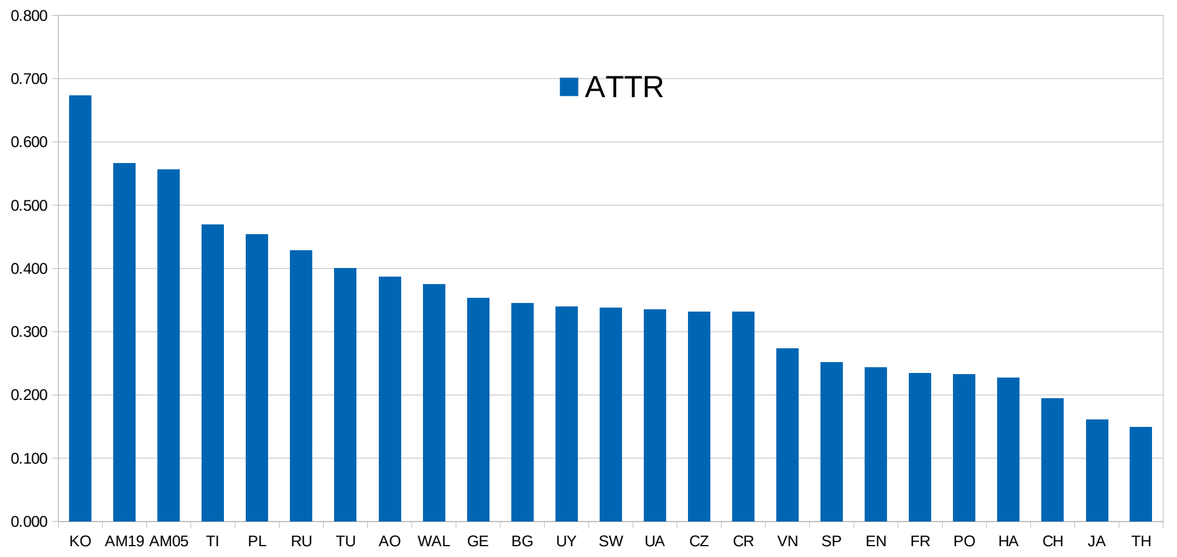


Phonetic Overlap Among languages

- Analyzed how many of the phones of one language are covered by another languages.
- Done based on International Phonetic Association (IPA) representation.

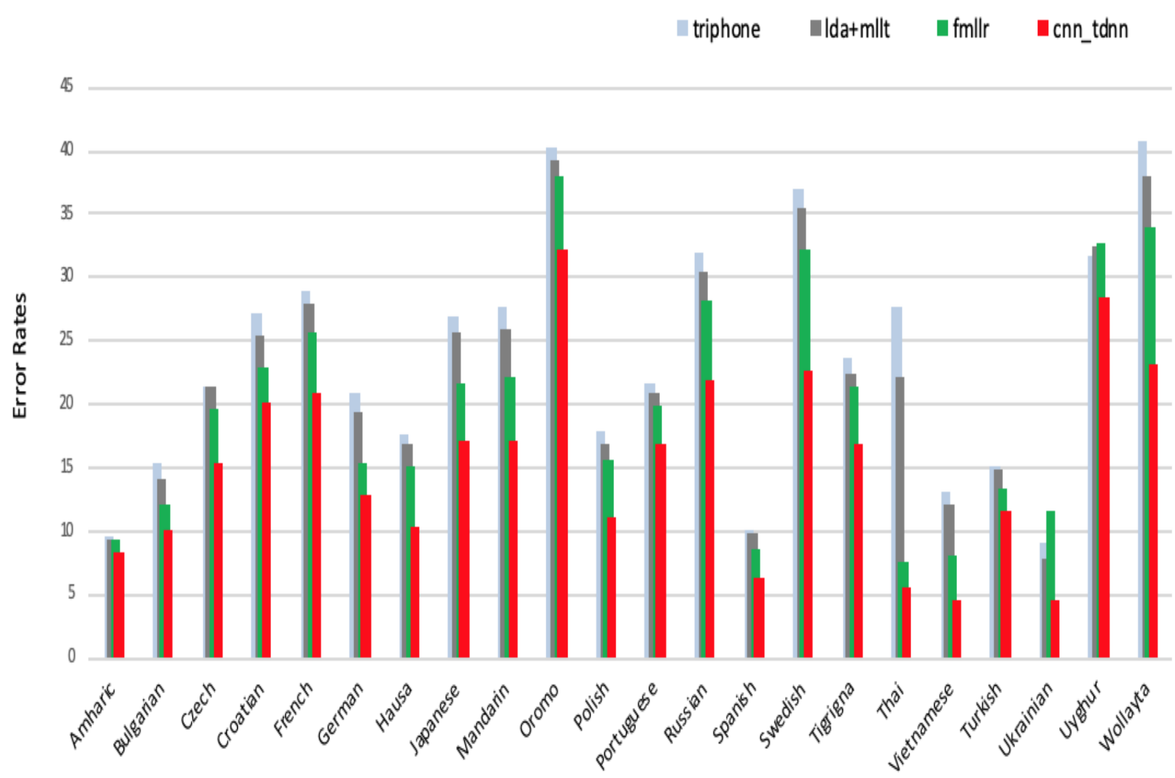
Morphological Complexity

- Type to Token ration, which crudely indicates the morphological complexity of a language, has been computed based on the training transcription of each language



Baseline Monolingual ASR Results

- We have developed GMM and DNN based monolingual ASR systems for each Globalphone and Ethiopian language



Universität Bremen