Web Tool for Phonemes Week-1

University at Buffalo The State University of New York

Agenda

- Data Sources
- Phonetics and Linguistics
- Choosing a Search Engine / Database
- Selecting a Backend

Phonetics and Linguistics

Learning the Concepts:

- Phonemes : Sound Unit

  Eg: University → ˌjunəˈvɝsəri (Phoneme)

- Graphemes: Written linguistic unit

  Eg : University → uni·ver·si·ty(Grapheme)

- ArpaBet → representation of each phoneme by one or two capital letters(ASCII)

  University → [Y UW2 N AH0 V ER1 S IH0 T IY0]

- Articulation Place and Manner : https://icspeech.com/consonant-sounds.html
- Semantics: Parts of speech.
- Morphology - Structure and Formation of words - Suffix, Prefix,etc

Phonetics and Linguistics

- Articulation Place and Manner :

  1.Place of Articulation : Bilabial, labiodental, dental, Alveolar, Palato-Alveolar, Palatal, Velar, Glottal

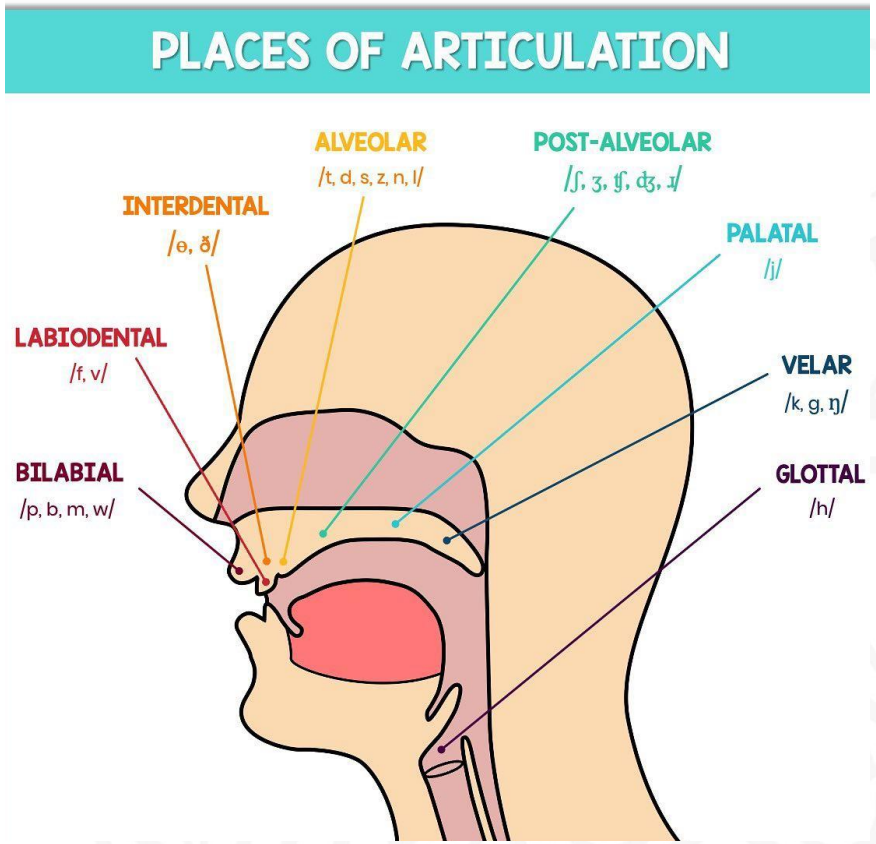
  2. Manner of Articulation : Nasal, Plosive, Fricative, Approximant, Tap/Flap, Trill, Lateral Fricative, Lateral Approximant, Lateral Flap

  3. Voiced or Voiceless

  https://icspeech.com/consonant-sounds.html

- Semantics: Parts of Speech

- Morphology - Structure and Formation of words - Suffix, Prefix,etc



Data Sources

- Graphemes → (Wiktionary has Data we can web scrape)
- Phoneme Data → (NLTK contains around 100k words)
- Articulation: Can go by some Rules
  https://github.com/AdamSteffanick/ipa-data/blob/master/guid-o-matic/ipa-data/ipa-data.csv
- Categories : WordNet, SemCor , BabelNet, FrameNet
- Semantics: WordNet

Database / Search Engine:

Requirements:

1. Ease of Dataset Integration
2. User Query Processing
3. Relevance of Results
4. Efficiency and Speed of Results

Options:

Elastic Search

Apache Solr

Frontend and Backend

Requirements :

We are using Python. We have 3 Framework options

1. Flask
2. Django
3. FastApi

FrontEnd:

1. Html
2. CSS
    a. BootStrap
    b. Bulma
3. React

Data Organization

1. We are thinking to store data in the Json Files
    a. Easy to load data to python, easy access
    b. Iterate on every word collect the data from different sources
    c. Easy to integrate with solr.
    d. can be easy transferred among
    e. can be easily sent as json (API response)

Question

1. How many categories are needed?
2. Any suggestion on databases? for the metadata
3. Any ml modes that give us score based on complexity?