Tools for cross-corpus linguistic analysis

Elias Stengel-Eskin
BA&Sc, Cognitive Science

Introduction

- Linguistic corpora consist of annotated speech data
  - usually timing information,
  - transcription (how something is said, transcribed in an alphabet representing sounds)
- Vast number of corpora
- Used mainly for phonetics (study of speech sounds) and phonology (study of systems of sounds, relationships of sounds)
- Speech Corpus Tools developed by Montreal Language Modeling Lab (MLML) to make searching these corpora user-friendly and fast
  - relies on PolyglotDB software (also developed by MLML)
  - translates different corpora into database format

Motivation

- Problem: huge amount of data, all in different formats
  - commonly in varying file types
  - often programming knowledge required to reduce data to desired portion
  - even if researcher has programming knowledge, searching can be very slow/tedious (1,000's of files)
- Searching for needle in a haystack, need some sort of unified method
- Better corpus querying saves time, money
- Can help to protect privacy of speakers in corpora by abstracting away from original recordings/only allowing user to view snippets of information.

Methodology

- SCT is the highest level software being used
  - written by MLML in Python
  - most abstracted away from data
  - most user-friendly, requires least programming knowledge
- It is based on the PolgyglotDB software
  - also written in Python by MLML
  - designed for researchers to incorporate into Python scripts
  - requires general programming ability
- PolyglotDB software loads data into databases
  - uses both graphical and relational databases
  - graphical DBs represent data as nodes and edges
  - relational DBs represent data in tables of relation ships
  - both dramatically reduce time to complete query

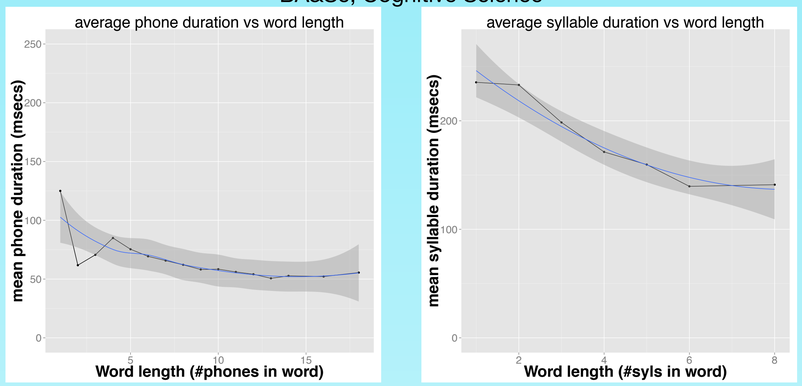

Fig. 1: Plots of results showing downward trend

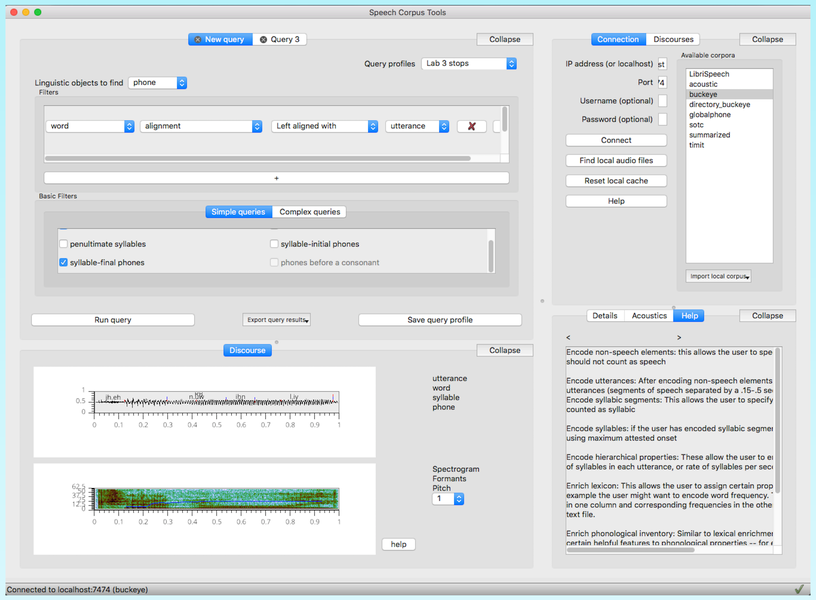

Fig. 2: The SCT application

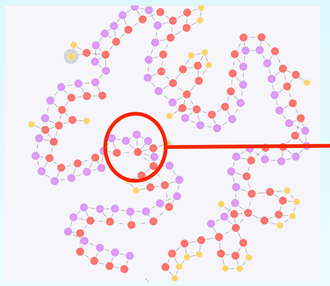

Fig. 3: Neo4j graph representation of a sentence

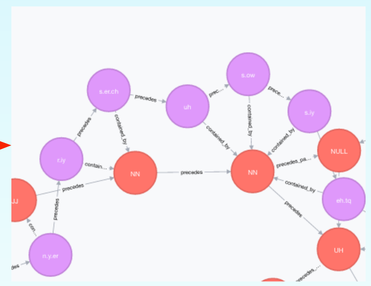

Fig. 4: A subset of that sentence

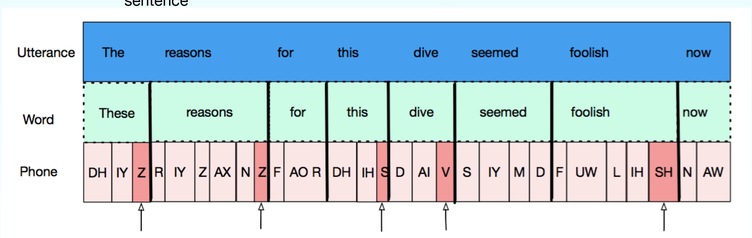

Fig. 5: The hierarchical nature of language shown. Thick black lines represent alignment — sharing a start or end time. The example query was word-final fricatives.

Example: Menzerath's Law

- Menzerath's law states that as syllables in word increase
  - duration of syllables decreases
  - duration of segments in syllables decreases
  - number of segments in syllables decreases
- Normally finding data to support this would be extremely tedious and time-consuming
- With SCT, can be done in minutes
- Querying the LibriSpeech corpus
  - 1,000 hours of read English
- Using filters to limit the data (see fig. 5)
  - Filters are used to select linguistic objects (utterances, words, phones, or syllables) and specify properties about them
  - Enriching (building extra relationships) data necessary to get properties like number of segments in syllable

Results

- Clear downward trend for syllables and segments
  - the more syllables/segments, the shorter the average length
- Normally getting data for these results would have taken much longer
  - 200,000+phones, 70,000+ syllables from 50,000+ words
  - might have taken days (if not weeks) to gather data by hand/write individualized scripts for each corpus
  - once imported into SCT, data exported in matter of seconds
  - gave exact subsets of data that were useful for research question

My Jobs

- Working with both SCT and PolyglotDB
- Testing
- Writing documentation/tutorials
- Adding additional features to SCT/PolyglotDB
  - Help panel
  - Relativized/summary statistics
    - ~ average, median, standard deviation, baseline
  - Enrichment (speaker info, stress, tone)

References

- Michael McAuliffe, Morgan Sonderegger, Michael Wagner. 2016. A system for unified corpus analysis applied to duration compression effects across 12 languages [PowerPoint slides].
- Esposti, M. D., Altmann, E. G., & Pachet, F. (2016). *Creativity and Universality in Language*. Springer Verlag.

Thanks to

- Prof. Morgan Sonderegger
- Michael McAuliffe
- Montreal Language Modelling Lab members
- Arts Internship Office
- Donors to the Arts Undergraduate Research Internship Awards fund