A Client-Side Web Application for Advanced Searching of Linguistic Corpora

Jack Isaac Rabinovitch

Harvard University

https://jackisaacrabinovitch.github.io/projects/lingcorpus.html

Introduction

Client Side Apps:
- Can be run without a server
- Poor internet connection OK
- Download for use offline

Problem:
- Larger corpora difficult to upload and search quickly on client's machine

Solution:
- Limit redundancy in format

smaller files = faster uploads
- Limit redundancy in searches by query language

fewer searches = faster results

Technology:
- Corpus files are in JSON
- Query language is custom
- Application runs in browser on (vanilla) JavaScript

The File Format

Corpus is a JSON object of *dictionaries*: arrays of *entries*.

```
Corpus = {
    "participants":[...],
    "sentences":[...],
    "lexicon":[...],
}
```

Entries have attributes: strings, numbers, or "*reference lists*", objects which contain a reference to a *dictionary* "d" and an array of indices "i" for elements within that *dictionary.*

```
{                          {
    "orth":"bi",               "orth":"ere",
    "gloss":"1sg.nom"          "gloss":"this"
}                          }
```

Words are defined once.

```
{
    "type":"question",
    "translation":"Did I take the rice yesterday?",
    "words":{
        "d":"lexicon",
        "i":[0,2,3,15,102,34]
    }
}
```

When words are contained in a sentence, only their index is referenced, saving space.

The Query Language

The Query Language is based on *filters* which search through a *reference list* and return another *reference list.*

(i) DICT lexicon WHERE gloss IS "think"
reference list *filter*

Search (i) creates a *reference list* of all words whose gloss is "think". This can then be used in another *filter.*

(ii) DICT sentences WHERE words CONTAINS DICT lexicon WHERE gloss IS "think"

Search (ii) creates a reference list of sentences whose speakers are within reference list (i).

Filters can be combined with logical operators (iii).

(iii) gloss IS "think" AND orth IS "ngononggʷ"
filter *filter*

TABLE makes a table of the *reference list,* joining arrays with a delimiter set by the user (iv).

(iv) word DELIMITER " "
TABLE DICT sentences WHERE type IS "question"

Result:

type	translation	words_orth	words_gloss
question	Did I take…	bi ere buda b…	1sg.nom this ric…
question	Where di…	sini deo aibid…	2sg.gen younge…

Adapting Corpora

FLEx corpora (.flextext) can be converted to compatible JSON via python code.

Convert Custom Format:
- Determine dictionaries and attribute types
- "Compress" file by iterating over *entries* within *dictionaries*: consolidate identical *entries*
- Replace nested entries with *reference list.*

Thanks

Thanks to the Ende people and other Pahoturi River language communities of the Paho River, Papua New Guinea, for the language data upon which this project started, as well as Dr. Kate Lindsey, Dr. Catherine O'Connor, Brady Dailey, and Ankana Saha for their help, discussion, and inspiration.