CS486C – Senior Capstone Design in Computer Science

Project Description

Project Title: Gamified Language Data Explorer	
Sponsor Information:	Dr. Okim Kang Department of English (TESL/Applied Linguistics) Northern Arizona University Okim.Kang@nau.edu

Project Overview:

1. **Gamification in language learning**

A lot of people think that unless you are very young, learning a new language is hard. But it's not. Well, not exactly. What it does take is quite a bit of time focused on reading, listening, and finally speaking the language. And that can get boring, so some folks just let it go. However, recent gamification efforts by companies such as Duolingo and Babel demonstrate that motivation can be sustained with gamified elements that give a feeling of competition and reward when goals are met. The result is learners engaging with language tasks more frequently and with more regularity.

Language data exploration

An important part of learning a new language correctly is receiving input in both spoken and written forms, and in extended chunks so that patterns can be noticed and learned. In written forms of language, words are embedded in sentences and paragraphs that provide learners with data about the grammar, patterns, and contexts in which words tend to occur.

When large datasets of naturally produced language are collected (called a corpus), technologies that allow learners to explore the data (i.e., search the dataset and create a concordance table of the preceding and subsequent text) allow learners to see such patterns in a targeted and systematic way. Traditional learning with corpora is most often used with publicly available corpora (an online database of texts, see https://www.english-corpora.org/coca/ or http://phrasesinenglish.org/searchBNC.html) in text-only form, restricting its use to grammar and phraseology.

Innovation is ripe for the inclusion of audio for speaking practice and recently, Dr. Okim Kang and her team have collected an audio corpus that aligns text to recordings of academic experts from different countries in English with the support of the National Science Foundation.

This project

This project would extend data-driven learning by leveraging technologies for audio corpora to allow English language learners to search and listen to words and short phrases, in context and naturally produced, in order for them to incorporate aspects of pronunciation in their data-driven learning efforts. Furthermore, it would motivate learners with gamified elements of points, a tracking of activity, and visualization that keeps learners coming back.

The goal of this capstone project would be to create a user-friendly web interface that allows learners to log in, search for words in an audio corpus, listen to the audio of those words in context, record themselves, and compare their speech to the corpus speech with simple playback tools. It should be delivered in a gamified platform that motivates learner to log in regularly and complete search / listen / record tasks.

2. **Our team**

Our team has been working in Computer-Assisted Language Learning (CALL) for several years and has noticed the lack of tools for foreign language learners that both focuses recent advances in audio corpora and research-informed pedagogy in language learning. We have experience researching second language acquisition in computerized contexts and developing internet-based language learning applications for a wide variety of learners.

3. **Current solutions for Data-driven learning for pronunciation learning**

- There are no known user-friendly online spoken corpora for data-driven learning.
- Current online spoken corpora do not have search features, are not freely available, or have very complex search functions that are designed for linguists (See searchable conversation corpus here https://sla.talkbank.org/TBB/ca/SBCSAE and search manual here https://talkbank.org/DB/).

4. **Solution overview**

We envision a web app that allows a simple interface for:

- Logging in and tracking user activity
- Searching for a word or phrase from the NAU audio corpus
- Viewing a concordance list (i.e., search result list) with the textual content from the corpus (See Fig. 1)
- Sorting and filtering the concordance list with preceding or subsequent text
- Listening to the target word or phrase in context with 2-4 seconds of speech before and after
- A user-friendly tracking to see how frequently and recently learning has been done, with elements of spaced repetition (see Fig. 2)

We can also envision two stretch goals:

- A simple recording widget that allows for playback of the learner repeating the search target on the same screen as the search results
- A transcription error reporting tool in order to improve the transcription accuracy
- An administrator page to create logins, invite users, assign users to groups, and see user progress

5. **Impact of successful product**

The project would allow us to distribute this as a learning tool to learners, teachers, and researchers as a free and beneficial tool for the first known efforts of data-driven learning for pronunciation. We would be able to distribute it to language learning students at NAU and other institutions globally.

An example of working with the language, and an example of gamification in language learning are provided on the next page.

Fig. 1. A screenshot of a concordance from the PFC French corpus. The search term “bonjour” can be listened to in the following files using the blue speaker button. https://public.projet-pfc.net/transcription/?q=bonjour#results

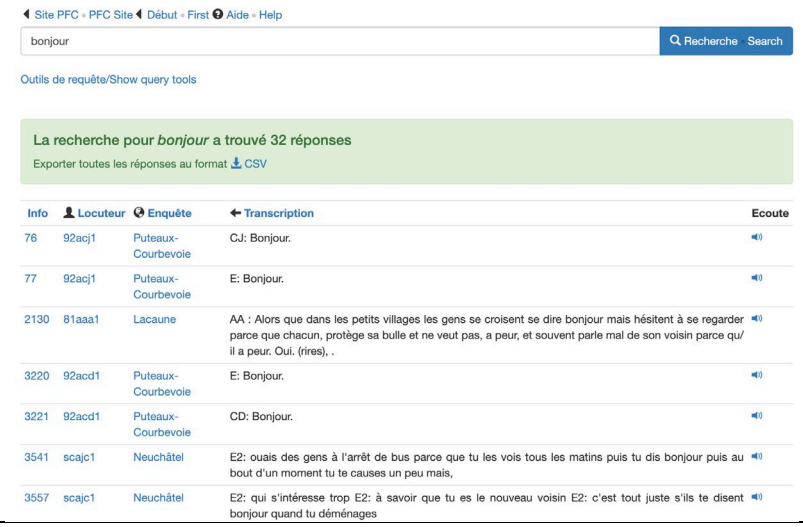
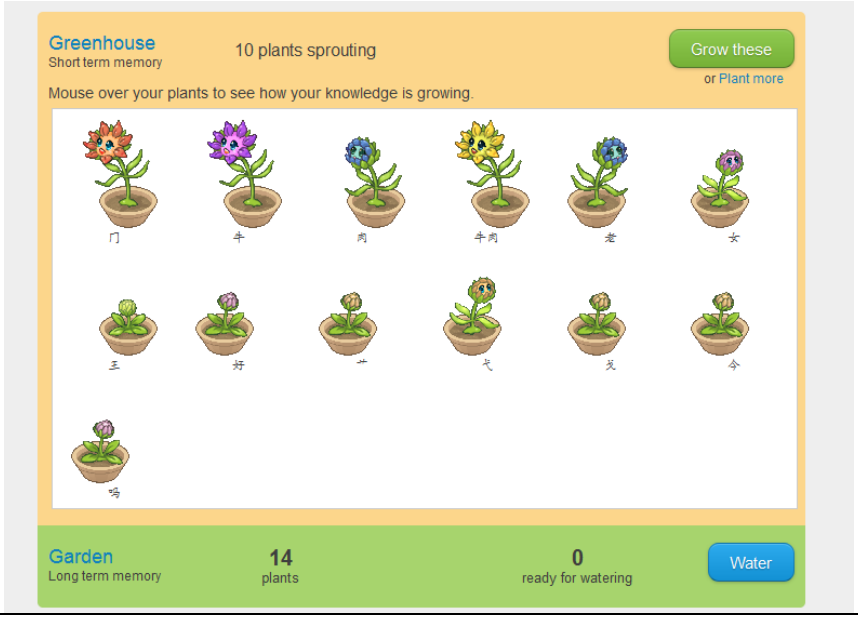


Fig. 2. A screenshot of a Memrise’s spaced repetition tracker. The flowers grow with more practice and shrink with time, reinforcing spaced repetition



Knowledge, skills, and expertise required for this project:

- Familiarity with building databases and efficient search strategies
- Experience in building a user-friendly multimedia web app
- Creativity for gamification elements of the web app

Equipment Requirements:

- There should be no special equipment or software required other than a development platform and software/tools freely available online.
- As we would like the web app to be mobile-friendly, no app store developer accounts will be necessary.

Software and other Deliverables:

- A gamified web app that has a search / concordancing tool for the NAU audio corpus
- An administrator page for managing users and viewing progress
- A strong as-built report detailing the design and implementation of the product in a complete, clear and professional manner. This document should provide a strong basis for future development of the product.
- Complete professionally-documented codebase, delivered both as a repository in GitHub, BitBucket, or some other version control repository; and as a physical archive on a USB drive.