Anki Vocabulary Extractor

Matthew Lochman

CIS201-HYB2 Fall 2018

Project Summary:

The project will focus on allowing the user to extract information from digital flash cards made in the program Anki. In particular, it will concentrate on extracting data from a user-specified field on each card. The extracted data is assumed to be vocabulary and will then be displayed in a table format. The user will have the ability to export the extracted data into a CSV file format for use in other spreadsheet programs or a PDF file for quick reference and searching. Other information will be extracted, such as part of speech. Users will be able to select which Anki decks information will be extracted from.

Project Background:

Anki is a powerful SRS program used to create digital flashcards that can be customized using HTML, LaTeX, and other formats. It is a primary tool for language learners (among other types of users) and has a very active community. Anki groups flashcards into decks. Each card consists of multiple fields, which hold different pieces of information. For example, using a program called subs2srs, a user can create a deck of cards from a video file and accompanying subtitle file. Subs2srs will create cards with subtitle dialogue text, the subtitle time stamp for the start of the dialogue line, a JPEG screen shot from the video taken in the middle of the time interval, an MP3 file containing the audio for that time interval, and an empty field for the user to enter notes.

This program will be geared towards Japanese language learners using subs2srs. It will allow users to extract vocabulary for each card in the deck. This will allow users to create a dictionary of words that they have learned across some or all of their different Anki decks.

IPOS Requirements:

The program will require the user to provide an export file from Anki that corresponds to cards made using the subs2srs template. The program will read through the export file and save the 'notes' field data into an array and display it to the user in a table. While reading the Anki export file, the program will attempt to extract vocabulary using the format "word - (part of speech) definition" or, if no part of speech is included, "word - definition".

The user will be able to save the extracted information into a CSV file or pdf file for easy searching. The user will have the option to load previously created files and add newly extracted entries (the program will check to see if an entry already exists to avoid duplication).

Conclusion:

I have been using Anki to study Japanese for about 3-4 months now. To help organize things, I currently have about 12 different decks in which cards are grouped by their source material. While reviewing, it can be challenging to keep track of whether I have encountered a word before in another deck or if that word has been used with the current meaning. While this program will not necessarily address those issues directly, it will provide the framework for the extraction of information to form a more formal dictionary, which could include example sentences as well as parts of speech and definitions. In addition, it will also provide significant motivation by displaying a total count of all the learned vocabulary. This is especially relevant since it is often hard to discern increases in skill or to decide at what level (beginner, intermediate, advanced, etc.) you may be. Providing a list will at least give evidence of ever-increasing comprehension.

Glossary

Anki An SRS program for creating “Powerful, intelligent flashcards” that can include anything “from images to Scientific markup.”

Card A digital flash card, with information stored in fields. The program uses formatting code to program how the field information is displayed on each side of the card.

Deck A collection of cards in Anki

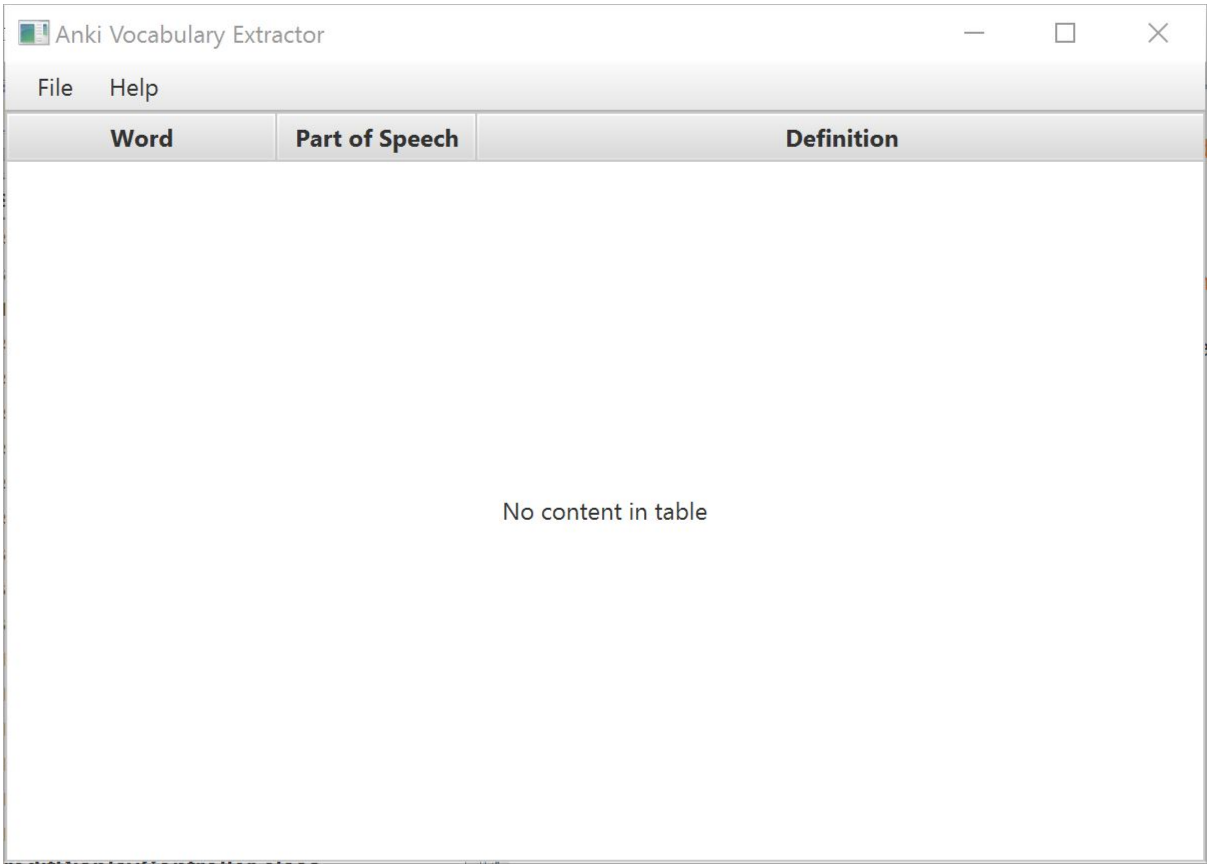
Field A piece of content that is part of a card. Can be content like images, audio, videos and scientific markup (via LaTeX)

LaTeX A document preparation system for high-quality typesetting. LaTeX uses the TeX typesetting program for formatting its output, and is itself written in the TeX macro language.

SRS Spaced Repetition System: a system that uses an algorithm to predict when you need to review a flashcard. Algorithms are based off of the SuperMemo algorithm created by Dr. P.A. Wozniak.

Subs2srs Subs2srs is a program that allows you to create import files for Anki or other Spaced Repetition Systems (SRS) based on your favorite foreign language movies and TV shows to aid in the language learning process.

Subtitle File A text based file that contains subtitles for movies and TV shows.



Word	Part of Speech	Definition

No content in table