Proposal: Grappler Baki Word Analysis

DATA 450 Capstone

Andrew Nemkov

February 13, 2025

1 Introduction

Language is often considered a significant barrier in the study and understanding of global cultures. Two people from different countries often struggle to bond due to a lack of communication. This barrier is especially difficult to breach when comparing the English and Japanese languages. Both the United States and Japan are economic and democratic leaders globally, yet the major linguistic differences between the two types of speech divide both regions more than the vast Pacific Ocean.

The sharing of literature has contributed significantly to cultural intertwining. In particular, Japanese manga has helped to connect both global cultures as a mirror of artistic and compositional expression. This project seeks to analyze the language use in manga by extracting Japanese words with Optical Character Recognition (OCR) software and performing linguistic analysis. In more detail, the analysis will delve in word frequency and type of character usage in the work as well as assess OCR performance for future improvements and project development. As a result, the process aims to provide insights into how language is used in manga for story progression and character development.

2 Dataset

The dataset used in this project is a collection of image panels from three series of the manga Grappler Baki, specifically from “Hanma Baki”, “Baki Dou 2”, and “Baki Rahen”. These images were obtained from the DL-Raw.ac site in the JPG format and amount to a total of around 4000 images spread between the three series. The quality and size of each image varies, but most are around 3500 pixels wide and 2160 pixels in height with resolution of around 144 ppi.

A key aspect of this project involves analysis of Japanese words or phrase and for this reason additional background information is required for the reader. The Japanese writing system

consists of logographic kanji symbols and a pair of kana. These kana consits of hiragana, used for native Japanese words and grammer, and katakana, which is used primarity for foreign words and names. Both hiragana and katakana each contain 46 basic characters while kanji consists of over 50,000 characters. Considering this, Japanese school students are required to learn the 2,136 most common of these kanji characters after finishing college. As a result, most of the word phrases extracted from the manga will consist of a combination of these three types of characters.

From this background informtion, a total of 11 varibles will be used for data analysis. These variables vary from information on words extracted with the OCR process, image information, OCR characteristics, Japanese character information, and others. Further details are shown below:

- *word_JAP:* specific Japanese word of manga image
- word_US: English translation of Japanese word word_JAP
- word_POS: grammatical classification of word_JAP
- img_title: manga image's file title containing specific word_JAP
- img_series: manga image's specific Baki series containing specific word_JAP
- length: length in characters of extracted word_JAP
- confidence: OCR confidence score for extracted Japanese phrase containing word_JAP (formatted as a percent)
- word_freq: frequency of Japanese word word_JAP across all images
- hiragana_ratio: the portion of hiragana characters in the phrase which includes specific Japanese word word_JAP (formatted as percent)
- katakana_ratio: the portion of katakana characters in the phrase which includes specific Japanese word word_JAP (formatted as percent)
- kanji_ratio: the portion of kanji characters in the phrase which includes specific Japanese word word_JAP (formatted as percent)

Itagaki, Keisuke. Hanma Baki, Baki-Dou 2, and Baki Rahen. Shōnen Champion, 2024. https://dl-raw.ac/.

3 Data Acquisition and Processing

The JPG images for this project was downloaded through a publicly accessible website DL-Raw.ac, which provided these pages for free usage. The pages used in this analysis will be used for non-commercial purposes only, all source citation will be provided, and the data will not be redistributed.

The JPG images will be run through a python OCR program, which would take each image and collect the words and phrases from the images into textual format. Specifically, this process will involve using PaddleOCR, a popular python OCR library, as well as Fugashi, a python library developed for the purpose of performing tasks on Japanese text such as cleaning. To

increase accuracy and word recognition from each image, the Paddle OCR parameters will be adjusted accordingly to detect smaller text, improve the model's ability to detect text regions, increase batch size for more character recognition, and others.

After an image's phrases or general unstructured words are extracted with the methods above, this text will be tokenized and separated into words using the Fugashi library. Before actual dataset creation, further metadata on these words will need to be composed.

- Firstly, the Fugashi library will either translate the Japanese word into English with its lookup function or by using a imported python dictionary library such as SudachiPy.
- Secondly, this library will take the separaterd words and identify which type of grammer classification each fits into such as noun; verb; or adjective; or other.
- Additionally, the number of characters in the specific Japanese word will be counted, the word's frequency across all images will be counted, and the proportion of hiragana; katakana; and kanji in the specific phrase containing this word will be calculated.

These steps, as well as further collection of other data, will be gathered and transfered into a single row of a CSV file, now representing 11 points of information on a single extracted Japanese word. The overall CSV dataset will be a collection of many of these Japanese words that were accurately extracted from the raw JPG images of the manga.

4 Research Questions and Methodology

1. What are the most common and least common Japanese words across different manga series in the Baki franchise?

- The variables used for this question will be word_JAP, img_series, and word_freq.
- This question can be answered by first plotting the top 10 most common and top 10 least common Japanese words into a bar graph with the y-axis representing the count of this word across all three series. The answer of most and least common words would be in textual format. These two bar graphs can later be broken down into three separate bar graphs for each series individually. These six values would also be in textual format. Both groups of visualizations would most likely take around four hours to complete, but this could take longer due to visualization styling.

2. How is the usage of different parts of speech distributed across the three manga series?

- The variables used for this question will be word_JAP, word_POS, and img_series.
- This question can be answered by firstly organizing the data into three separate datasets by series. After this, the number of each part of speech would be counted and recorded in textual format using a side by side bar graph visualization of nouns, verbs, and adjectives of each series. This would take around a little more than four hour for creating and styling of the visualization.

3. How does the OCR confidence score correlate with the complexity of the word based on length of phrase and character type of specific word?

- The variables used for this question will be word_JAP, confidence, length, hiragana_ratio, katakana_ratio, and kanji_ratio.
- This question can be answered by firstly identifying what would represent complexity in a word or phrase. This would be using the length and type of character of the word used. To understand the relationship between OCR confidence and complexity, a correlation would be created to understand the relationship between the two variables. Further, a heatmap can be created to visualize this. This question would be answered in around five hours.

5 Work plan

Week 4 (2/10 - 2/16):

- experiment further on OCR (3 hous)
- add more data to work with to make the analysis fairer (4 hours)

Week 5 (2/17 - 2/23):

- take time to improve OCR model to detect more image text (3 hours)
- set up dataset and perform basic testing and overview (4 hours)

Week 6 (2/24 - 3/2):

- clean data and implement Japanese dictionary connection (2 hours)
- begin work on question 1 and create visualization (4 hours)
- spend a bit more time learning about how works OCR (1 hour)

Week 7 (3/3 - 3/9):

- Refine results of question 1 and begin work on question 2 (3 hours)
- Presentation prep and practice (4 hours)

Week 8 (3/10 - 3/16): *Presentations given on Wed-Thu 3/12-3/13.*

- Poster prep (4 hours)
- Presentation peer review (1.5 hours)
- continue work on question 2

Week 9 (3/24 - 3/30): *Poster Draft 1 due Monday morning 3/24 at 9am. Poster Draft 2 due Sunday night 3/30.*

- Peer feedback (2 hours)

- Poster revisions (1.5 hours)
- finish and refine question 2 (4 hours)

Week 10 (3/31 - 4/6): *Final Poster due Sunday 4/6.*

- Peer feedback (1.5 hours)
- Poster revisions (2 hours)
- begin work on question 3 (4 hours)

Week 11 (4/7 - 4/13):

- finish question 3 (2 hours)
- do double-check of previous work and make needed improvements (5 hours)

Week 12 (4/14 - 4/20):

- consider question and results of analysis (3 hours)
- consider need for project expansion (1 hour)
- look over basic info on how to present results (3 hours)

Week 13 (4/21 - 4/27): *Blog post draft 1 due Sunday night 4/28.* [All project work should be done by the end of this week. The remaining time will be used for writing up and presenting your results.]

- Draft blog post (4 hours).
- find images and style on how to put together blog post (3 hours)

Week 14 (4/28 - 5/4):

- Peer feedback (3 hours)
- Blog post revisions (4 hours)
- [Do not schedule any other tasks for this week.]

Week 15 (5/5 - 5/8): *Final blog post due Tues 5/7. Blog post read-throughs during final exam slot, Thursday May 8th, 8:00-11:20am.*

- Blog post revisions (2 hours)
- Peer feedback (2 hours)
- [Do not schedule any other tasks for this week.]

5.1 Some cool Quarto stuff

[You can delete this section from your proposal.]

For your reference, here's an example of a Python code cell in Quarto, along with a figure that gets generated, along with a caption and a label so that it can be referred to automatically as "Figure 1" (or whatever) in the writeup.

For a demonstration of a line plot on a polar axis, see Figure 1.

```
import numpy as np
import matplotlib.pyplot as plt

r = np.arange(0, 2, 0.01)
theta = 2 * np.pi * r
fig, ax = plt.subplots(
  subplot_kw = {'projection': 'polar'}
)
ax.plot(theta, r)
ax.set_rticks([0.5, 1, 1.5, 2])
ax.grid(True)
plt.show()
```

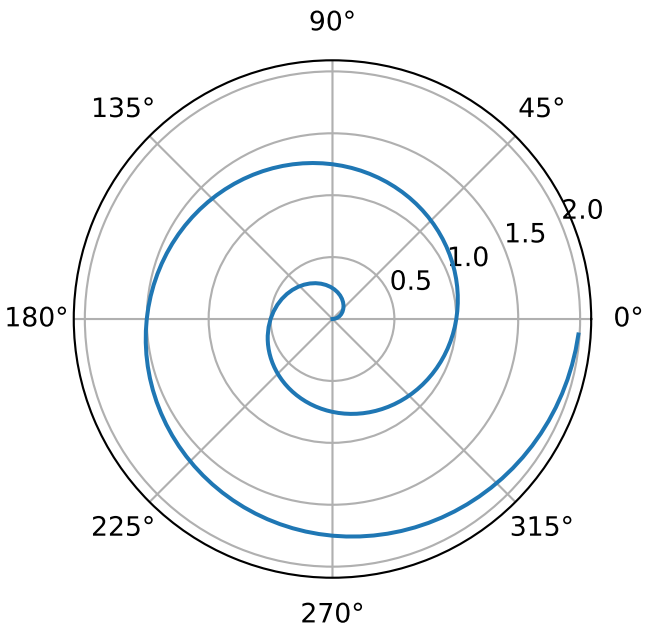

Figure 1: A line plot on a polar axis

Here's an example of citing a source (see Phillips 1999, 33–35). Be sure the source information is entered in "BibTeX" form in the `references.bib` file.

6 References

[The bibliography will automatically get generated. Any sources you cite in the document will be included. Other entries in the `.bib` file will not be included.]

Phillips, T. P. 1999. "Possible Influence of the Magnetosphere on American History." *J. Oddball Res.* 98: 1000–1003.