Proposal: Grappler Baki Word Analysis

DATA 450 Capstone

Andrew Nemkov

February 6, 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 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.

After applying OCR to the manga images, the following variables are planned to be used in the linguistic analysis:

- word_JAP: Japanese word of manga image
- word_US: English word of manga image
- word_POS: grammatical classification of Japanese word
- img_title: manga image's file title
- img_series: manga image's specific Baki series
- length: length in characters of extracted word
- confidence: OCR confidence score for extracted Japanese word (percent)
- word_freq: frequency of Japanese word from all images
- hiragana_ratio: portion of hiragana in phrase with Japanese word (percent)
- katakana_ratio: portion of katakana in phrase with Japanese word (percent)
- kanji_ratio: portion of kanji in phrase with Japanese word (percent)

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

3 Data Acquisition and Processing

[In this section, if applicable, describe how you will obtain the data (if it's anything more complicated than a simple download). Discuss what data processing steps will be needed, such as recoding variables, data cleaning, data tidying, imputing missing values, etc. See sections 1c, 1d, 1e in the "Good Enough Practices" paper.]

4 Research Questions and Methodology

[In this section, list each of the questions you will explore. Following each question, provide a detailed and specific plan for how you plan to answer the question. Include the specific steps you will take, what form the answer will take (a number? table? visualization? model? Give all the specifics), and estimate how many hours each question will take to complete.]

1. Is smoking correlated with diabetes? To answer this, I will create a filled bar plot, with the left bar representing non-smokers, the middle bar representing people who smoke moderately, and the right bar representing heavy smokers. The bars will be the same height, and each bar will be colored two colors based on the proportion of patients in the group who do or do not have diabetes.
2. Question 2? Plan for question 2.
3. Question 3? Plan for question 3.
4. etc.

5 Work plan

[Fill in the list below with a plan for what you will do each week, starting 2/10. You should have around 7 hours worth of work each week. Writing work counts. Several tasks have already been filled in for you.]

Week 4 (2/10 - 2/16): [Just an example:

- Data tidying and recoding (4 hours)
- Question 2 (4 hours).]

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

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

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

- 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)

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)

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

- Peer feedback (1.5 hours)
- Poster revisions (2 hours)

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

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

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).

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()
```

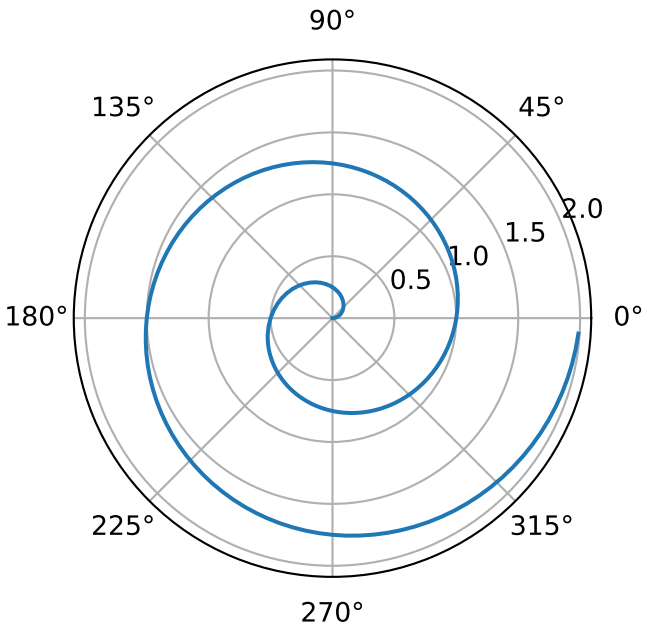

Figure 1: A line plot on a polar axis

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.