Visualizing Zipf's Law in Japanese

Kip Turner*
University of California, Santa Cruz

ABSTRACT

This paper will discuss visualization techniques employed to learn more about the evolution of natural languages. Through InfoVis techniques we can explore the effects of Zipf's law for many different languages. I will be looking at the application of the Japanese language regarding the frequency and rank that words are written in contrast to spoken. By analyzing Japanese under this light we can see if the use of Kanji has had an impact on how the Japanese language is spoken.

1 INTRODUCTION

Zipf's law is a naturally occurring phenomena, named after George Kingsley Zipf, a Harvard Linguistics professor. It is used to explain a power-law relationship between the frequency of a word's usage and the rank, which is the ranking it holds amongst most commonly used words. The relationship states thats given a corpus of natural language utterances, the frequency of any word is inversely proprotional to its rank in the frequency table. Zipf's law obeys a log-log scale, this is to say that the most frequency word (word with rank 1) is an order of magnitude more frequent than the rank 2 word and so on [2]. Zipf's law stems from latin based language analysis, which is inherently biased in word-frequency distributions due to their shared origin. In contrast, Asian languages are relatively independent from latin languages, and therefore are not inherently biased and can be used to shed light on the universality of Zipf's law. If Japanese indeed obeys Zipf's law, we can further probe into the evolution of the language by looking at how well the written complexity of a word obeys Zipf's law in contrast with the spoken complexity of the word. To elaborate: the written length of a word in Japanese and Chinese is not directly proportional to the spoken length of the word. This raises the question: does mankind's intent on the path of least resistance apply to molding both written and spoken forms of a language or is one dominant over the other?

Japanese and Chinese characters, Kanji and Hanzi respectfully, are classified as logographs and stem from ideographs or pictograms, evolving from a visual representation of a concept. This is to suggest that the evolution of a writing system and the pronunciation system in the Japanese language did not originate in such a way that they should have effected each other. The Korean writing system, Hangul, was created in a different manner, the vowels are still pictograms of concepts, but instead of pictorializing the the shape of the mouth required to pronounce the vowel. e.g. The vowel for 'eu' represents the flatness of the lips when pronouncing it. This is relevant because Korean traditionally used the same asian character system as Japanese and Chinese did up until 1443 when Korea switched to Hangul. Korean wasn't the only Asian language to make iterations on the character writing system, Chinese also reduced the complexity of their writing system around 1950 when the simplified Hanzi set was introduced. Japanese's Kanji also has simplifications, while not as extensive as the changes to Hanzi in the 1950s, Kanji has many shared simplifications to that of Hanzi, e.g. the character for country is the same in the new Chinese Hanzi as it is in the Japanese Kanji.

The purpose of my research is to test whether or not Zipf's law extends to Chinese, Japanese, and Korean (CJK) in not only spoken but also written contexts.

*e-mail: katurner@ucsc.edu

2 EXPOSITION

2.1 Does Japanese obey Zipf's Law?

Does Japanese obey the log-log proportionality that is characteristic of Zipf's law? If it does not, then a comparisons between spoken complexity and written complexity may not be a legitimate visualization because the ranking I am comparing it against is not characteristic of Zipf's law.

To test this, I needed access to a large corpus of Japanese phrases and the normalized frequencies of words that occur in it. The corpus should contain colloquial and casual text to be an accurate representation of modern day Japanese. The best resource to obtain such a large corpus is the Internet; the Internet has a near endless supply of colloquial text on forums, mailing lists etc... To obtain this data I looked towards the University of Leeds which maintains a Japanese corpus they have obtained through web crawlers [1]. In the Japanese language, seperating word forms (what constitutes an individual word), is an extremely difficult task and is well beyond the scope of this project. To solve this issue I used the University of Leeds data to obtain a tab separated file that contained the normalized frequency and rankings of individual word forms.

2.1.1 Implementation

For the primary data gathering part, I loaded the file tab separated word form frequency file into my database for potential future processing. I used a scatter plot to visualize the word forms' frequency against its rank. The scatter plots x-axis and y-axis are both in log scale as that is dictated by Zipf's law. If Japanese trully obeys Zipf's law then the slope of the visualization will be roughly equivalent to -1.

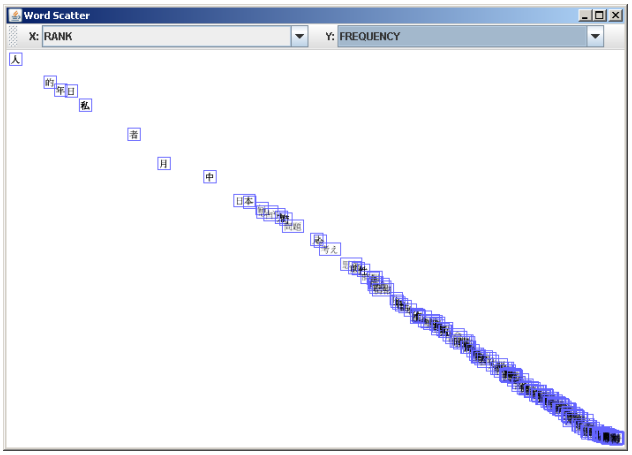

Figure 1: Corpus of rank vs frequency in log-log scale.

It is very apparent in figure 1 that there is a tight correlation between rank and frequency, moreover the slope of the line is almost exactly -1. The implications of this suggest that Japanese obeys Zipf's law and it is reasonable to assume that spoken and written complexity can be appropriately judged against it as a base.

2.2 Spoken Complexity of Japanese and Zipf's Law

Since the distribution of words in Japanese obeys Zipf's law, the next step is to see if the spoken language of Japanese also obeys the path of least resistance theory that is the foundation of Zipf's law.

To do this I will analyze the spoken complexity of Japanese words, the complexity of a word is approximated by the summing its moraic length. A mora in linguistics is the timing of a syllable, in the case of Japanese every mora is the same length. The moraic length is achieved by determining how many hiragana characters the word is composed of, since each hiragana element makes up for one syllable, and each syllable in Japanese is equivalent in vocal length because each syllable makes up one mora, the calculation of spoken complexity can simply be the number of hiragana in a word.

In order to obtain the required data to process this part of the visualization a dictionary had to be used. A mapping of a word to the syllables in it (hiragana) was only possible using a dictionary like look up. To obtain a digital Japanese dictionary I looked towards professor Jim Breen's archive at Monash University [3]. The file is xml formatted and has a complete dictionary with one-to-many mappings from each Japanese word to its syllabic readings. Using those readings, the program can calculate the spoken complexity of any given word. However, since a word can have many readings, and the distribution of the frequency that those readings appear in is unknown, I treated each reading as a separate data item. To reiterate, I only have the data for frequencies that the words appear in, not down to the frequency that each reading of each word appears.

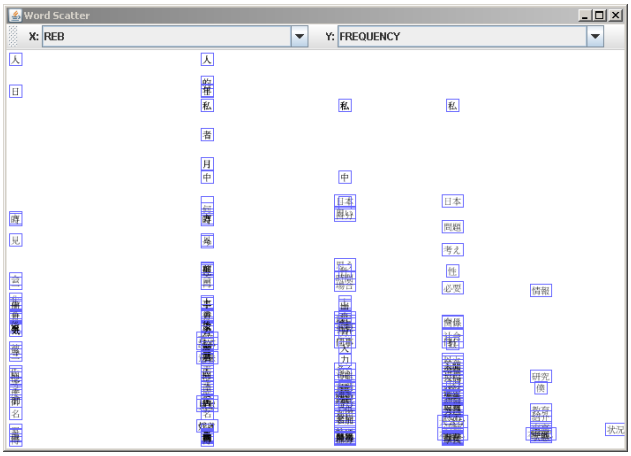

Figure 2: Spoken Complexity.

In figure 2, the spoken complexity is measure on a log-log scale. The axis layout is one where the independent axis is spoken complexity, computed using the above equation, and the dependent axis is the normalized frequency of that word in the colloquial corpus. The results do not correlate as well as they did in figure 1, nonetheless we can see a weak correlation of roughly slope -0.5 by looking at the dense areas in the figure. As spoken complexity increases, the frequency the word is used dies off. Based on that, there is obviously some Zipf's law like correlation between spoken complexity and frequency.

Future research may include a more complex model for approximating spoken complexity, rather than just the time it takes to pronounce a word, an approximation that takes into account guttural stops and voicing sounds that require use of different parts of the mouth may be more indicative of the resistance required for any particular words.

2.3 Written Complexity of Japanese and Zipf's Law

In order to calculate the written complexity of Japanese I used the stroke count of the kanji. The stroke count is how many times the writer must lift up their writing instrument to complete the character, this seems like a reasonable approximation of the difficulty to write a kanji. Calculating written complexity using words is not an effective approximation, a word can contain a kanji that takes 1 stroke to write or 30 strokes to write. Thus counting the number of kanji in a word is a misleading statistic on how difficult that word is to write. Therefore, I decided to calculate on a per kanji basis.

To calculate on a per kanji basis, I needed to obtain another corpus with analysis of the rank of each kanji. A corpus to this effect was also located on Monash University's ftp website [3]. The data that was provided has a kanji and it relative ranking based on its frequency in Japanese newspapers. The use of kanji in Japanese newspapers is a common metric of a person's literacy in Japan. As well, the Japanese government regulates what kanji can appear in a newspaper so that a high school level education is sufficient to read the news.

Figure 3: Written Complexity.

In figure 3, the written complexity on the x-axis is measured in strokes it takes to complete a character, and in the y-axis is the rank that character occurs in newspapers. We can see a general trend in the direction of a higher higher stroke count (positive x-axis) leads to a high rank (positive y-axis) in the frequency table, which means it is used less often. Despite the top-left and bottom-right corners being especially sparse, the correlation of written complexity and rank in the frequency table is weak at best.

In figure 4, the fish-eye, or magic lens is used so that the user may look at each individual kanji. Since the data set is large, singling out a particular kanji is impossible, so, the implementation of a fish-eye lens enabled picking of data items. When a data item is picked a tool-tip will come appear at the mouse location that explains information about the particular kanji's complexity and rank.

2.4 Tools

2.4.1 Java

I chose Java as the language of implementation for my project due to its advanced Unicode support. Since almost every character and string in my program was outside of the ASCII range, it was tremendously advantages to have a language the managed unicode code points behind the scenes.

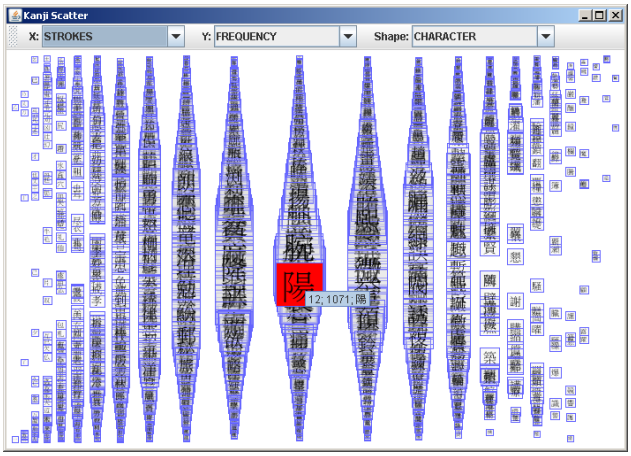

Figure 4: Written Complexity with lens.

2.4.2 Prefuse

A datavisualization toolkit that helped me speed up the presentation of my data. Various tools such as glyph rendering and automatic axis resizing spared me some of the grunt work of setting up a basic render. It also featured a JDBC connection feature that allowed me to easily hook up my database to the visualization part of the program.

2.4.3 HyperSQL Database

A fast java-based implementation of a database system that allowed me to run SQL queries locally. It complies with the JDBC and seamlessly integrates with the Java runtime environment.

3 Conclusion

From section 2.1 it is apparent that the Japanese language obeys Zipf's power-law distribution when observing word frequencies from a colloquial context. Visualizations of whether the spoken or written complexity of the Japanese language could be determined by Zipf's law provided less perfect data. As expected, the spoken complexity had a near-strong correlation as a best fit line could be imagined in figure 2 to extend from middle-left to bottom-right of the scatter plot indicating a slope characteristic of Zipf's law. However, the written complexity had a much weaker correlation to Zipf's law in figure 3. There was some rounding at the edges where common word had very complicated writings, or where seldom used words had a very simplistic writing. The results are inconclusive, the results hint at a relationship between word frequency and the complexity of the word but the findings were too rough to prove a strong correlation. In future research a better approximation of complexity could be used that may produce cleaner results. However, in the current state it looks like spoke complexity has a stronger correlation than written complexity towards Zipf's law.

Acknowledgements

The author wish to thank Professor Alex Pang. The work was made possible by Professor Jim Breen at the Monash University for the use of his dictionary files.

References

[1] Large corpora used in cts. University of Leeds, UK.

[2] L. A. Adamic. Zipf, power-laws, and pareto - a ranking tutorial. *Information Dynamics Lab, HP Labs Graphics*, Mar. 2002.

[3] J. Breen. Japanese dictionary. Monash University.