INSTITUTO TECNOLÓGICO DE AERONÁUTICA

Márcio Valença Ramos

RELATE KANJI: A STATISTICAL ANALYSIS OF THE JAPANESE LANGUAGE AND CONSEQUENCES FOR TEACHING METHODS

Final Paper
2016

Course of Computer Engineering

CDU 681.3:809.56-5

Márcio Valença Ramos

RELATE KANJI: A STATISTICAL ANALYSIS OF THE JAPANESE LANGUAGE AND CONSEQUENCES FOR TEACHING METHODS

Advisor

Prof. Dr. Carlos Henrique Costa Ribeiro (ITA)

COMPUTER ENGINEERING

São José dos Campos
Instituto Tecnológico de Aeronáutica

2016

Cataloging-in Publication Data
Documentation and Information Division

Ramos, Márcio Valença
Relate Kanji: A Statistical Analysis Of The Japanese Language And Consequences For Teaching Methods / Márcio Valença Ramos.
São José dos Campos, 2016.
99f.

Final paper (Undergraduation study) – Course of Computer Engineering– Instituto Tecnológico de Aeronáutica, 2016. Advisor: Prof. Dr. Carlos Henrique Costa Ribeiro.

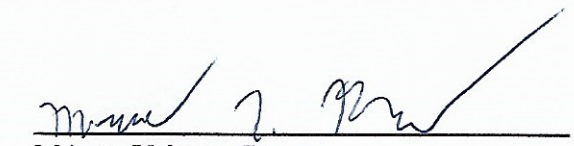
1. Kanji. 2. Language. 3. Linguistics. 4. Statistics. 5. Pedagogy. 6. Andragogy. I. Instituto Tecnológico de Aeronáutica. II. Title.

BIBLIOGRAPHIC REFERENCE

RAMOS, Márcio Valença. **Relate Kanji: A Statistical Analysis Of The Japanese Language And Consequences For Teaching Methods.** 2016. 99f. Final paper (Undergraduation study) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

CESSION OF RIGHTS

AUTHOR'S NAME: Márcio Valença Ramos
PUBLICATION TITLE: Relate Kanji: A Statistical Analysis Of The Japanese Language And Consequences For Teaching Methods.
PUBLICATION KIND/YEAR: Final paper (Undergraduation study) / 2016



Márcio Valença Ramos
Rua H8B, 239
12228-461 – São José dos Campos–SP

RELATE KANJI: A STATISTICAL ANALYSIS OF THE JAPANESE LANGUAGE AND CONSEQUENCES FOR TEACHING METHODS

This publication was accepted as Final Course Work

Márcio Valença Ramos

Author

Carlos Henrique Costa Ribeiro (ITA)

Advisor

Prof.Dr. Cecília de Azevedo Castro César
Course Coordinator of Computer Engineering

São José dos Campos: DECEMBER 21, 2016.

I dedicate this thesis to the nameless and long exploded suns that fused together the atoms that make up my body and all of that which is around me, including the ones I love. Without those long dead and anonymous suns, nothing as I know would ever come to existence, hence I am forever grateful to them.

Acknowledgments

First and foremost, I would like to thank the relentless work of my adviser, Carlos Henrique Costa Ribeiro, for the support on all steps of the project, from the delimitation of the theme to the review of the written report. The quality of this work was greatly elevated thanks to his suggestions.

The second person I would like to thank is my life companion, Luciana Yumi Ieiri, who helped me keep my dedication and calm even on the most stressful moments.

Lastly, I would like to thank all my Japanese learning friends that somehow gave me validations about the progress and motivation to work on this project, including most importantly my senior student, friend and dance partner Lara Diniz.

"Understanding is a kind of ecstasy."
— Carl Sagan

Abstract

The task to learn the written form of the Japanese language is a remarkably daunting one. Japanese is composed of three different scripts, and albeit two of them are relatively simple, with only 46 unique syllabic graphemes each, the third – the Jouyou Kanji – is an ideographic script with 2,136 official ideograms, each with a range of different associated meanings and a number of different readings. Learning this complex script is essential to learning the Japanese language, however there are few teaching methods that rely strongly in statistics for ensuring the efficiency of the teaching method. This academic work proposes better learning methods through the estimation of prevalence of Kanji in the Japanese language, the construction of graphs that relate different letters together and the application of a Personalized PageRank algorithm to estimate the best order a student should follow to learn these ideograms. Finally, we gather the results achieved in the earlier parts of the work to propose aspects of a web app for the teaching of Japanese Jouyou Kanji.

List of Figures

List of Tables

Contents

1 Introduction

The written form of the Japanese Language is composed of three scripts, being two of them syllabic[1] and one of them ideographic.[2] The syllabic scripts are both composed of 46 unique letters each and pose no big challenge to be learned, being thus studied in the primary school stage for native Japanese students. On the other hand, the ideographic script is composed of 2,136 official ideograms that are called the "Kanji" – more specifically "Jouyou Kanji" in the case of the official letters. These ideograms are the complete representation for the great majority of conceptual ideas in the Japanese language. As so, each Kanji carry a number of related meanings, some Japanese or "natural" readings, some Chinese or "radical" readings and finally some special readings to be used in the case of names. Provided that a linguistic structure of such high complexity is part of the foundation of the Japanese written form, it becomes interesting to create efficient teaching tools to aid the learning process of the Japanese Kanji, a study that lasts to the end of high school to the native Japanese speakers themselves.

1.1 Short History of the Japanese Writing System

Although it is debated if the core base of the Japanese language is Korean or Altaic, it is accepted that the writing system of Japan was introduced from China. That is to say, prior to the introduction of writing from China, Japan did not have its own coherent writing System. This introduction came with the spread of Buddhism and is believed to have happened before the 5th century. The earliest found text dates from the early 8th century (the Kojiki).

At first, the Japanese language was expressed solely through Chinese characters. Several issues arose from this approach, since Japanese uses concepts of particles and conjugation that are different from Chinese, making it to require some few sounds repeatedly. At this stage of writing of Old Japanese (the Man'yōgana writing system), Kanji were

[1] In a syllabic script every character represents a complete phoneme which is not associated with any specific meaning. More precisely, those two scripts are considered to be moraic, a concept that is further explained in the Linguistic Appendix, at section A.1

[2] every character of an ideographic script expresses a concept, rather than a sound

used both for semantic and phonetic functions. In the early 9th century, a writing system was invented by the Japanese to represent solely sounds. This type of script was called "Kana" and have two modern representatives, Hiragana and Katakana, which are further explained in Appendix B.2. Figure 1.1 illustrates this adaptation of Kanji to pure syllabic graphemes in Hiragana.

FIGURE 1.1 – Adaptation from kanji to Hiragana, with the original Kanji at the top, an intermediary form in the middle and Hiragana at the bottom

Prior to World War II, thousands of Kanji characters were in use in various writing systems and styles, leading to great difficulties to learning and teaching Japanese. To ease this problem, in 1946 the Japanese Ministry of Education created a list of the most commonly used 1,850 characters and declared them to be the official Kanji of the Japanese language. This list of official Kanji was called the Tōyō Kanji (当用漢字, literally "Chinese characters for general use"). This set of Kanji was proposed as an intermediary step for a plan that aimed at completely abolishing Kanji from the Japanese language, an ambitious plan that received strong opposition from the Japanese public.

In 1981 an updated list that added 95 additional Kanji to the Tōyō Kanji was forwarded, and was named the Jōyō Kanji, or Jouyou Kanji (常用漢字, literally "Chinese characters for regular use"). In 2010 the official list was again reviewed, including addi-

tional 196 characters, while removing 5, forming the current Jouyou Kanji list with 2,136 official characters. Additionally, a second list of Kanji was created, the called Jinmeiyō Kanji (人名用漢字, literally "Chinese characters for use in personal names"). This second list supplements the Jouyou Kanji list, so that all proper names in Japanese should have all characters present in one of those two lists.

1.2 Kanji Complexity Overview

There are a number of factors that introduce complexity to the study of Japanese Kanji. Below, we list a a number of the most challenging issues.

1.2.1 Various representations for the same Kanji

For 212 Kanji in the Jōyō Kanji there are two forms of the same character: the Old form (Kyūjitai) and the Modern form (Shinjitai). Additionally, there are 4 characters listed in the Jouyou Kanji list that are not presented in Japan's basic character set, the JIS X 0208, and are thus commonly replaced by similar characters that are in this list. Therefore, in any proper statistical analysis it is required that both forms of these Kanji are mapped to the same concept. In Table 1.1 we list 7 of these 216 variations.

TABLE 1.1 – Comparison of variants for 7 Jōyō Kanji

New form/ Proper form	**Old form/ Adapted form**
慎	愼
竜	龍
涼	凉
円	圓
尽	盡
礼	禮
頬	頰

1.2.2 Multiple meanings associated to a single Kanji

A further complication in the case of Japanese is that its roots are very apart from our own, so concepts we take as dissimilar in Romantic or Anglo-Saxonic languages may be

considered to be related in Japanese, and therefore be grouped under the same ideogram. One such example is the ideogram 気 that simultaneously means: spirit (as in soul), atmosphere (as in atmospheric pressure), mood/feeling, the mind and air.

1.2.3 Multiple Chinese readings (on'yomi)

As explained in Section 1.1, the Japanese writing system was heavily influenced by Chinese. In this system, Chinese came to be the etymological root of compound words in Japanese. An analogy with the Anglo-phonic case would be the concept of Water, that can be worded as *Hydro* (as in *Hydraulics*) or *Aqua* (as in *Aquaphobia*). This kind of reading came to be known as the On'yomi. One situation that arose from this borrowing is that Chinese and Japanese use very different phonotactics. While in Japanese the only kind of different voicing is done by increasing the length of the voicing of a vowel (as in the proper name Tooru, or Tōru), Chinese has five different kinds of intonations for the vowel parts of phonemes (for example: ma, mā, mà, má and mă are five different words in Chinese, differentiated solely through intonation), and also includes many consonant sounds that are not existent in Japanese. This incompatibility caused the Japanese mimics of the Chinese sound to be radically different than the original. Also, since the Chinese language entered in Japanese firstly through scholars and this was a process that lasted centuries, it occurred that miscopying was occasionally carried into Japanese as an accepted form or multiple steps of the phonetic evolution of the same word entered in Japan and multiple of those became accepted.

One such example of a Kanji with multiple on'yomi reading is 数, that represents the idea of figure or number. The officially recognized Chinese reading of this Kanji is "sū", but four more readings are common: "su", "shu", "soku" and "saku". Note that the current reading in Chinese for this Ideogram is "shŭ".

Figure 1.2 presents a histogram for the number of readings by Kanji. As it can be noted, most Kanji have only one Chinese reading, but about 470 have two readings and more than a hundred present three or more readings.

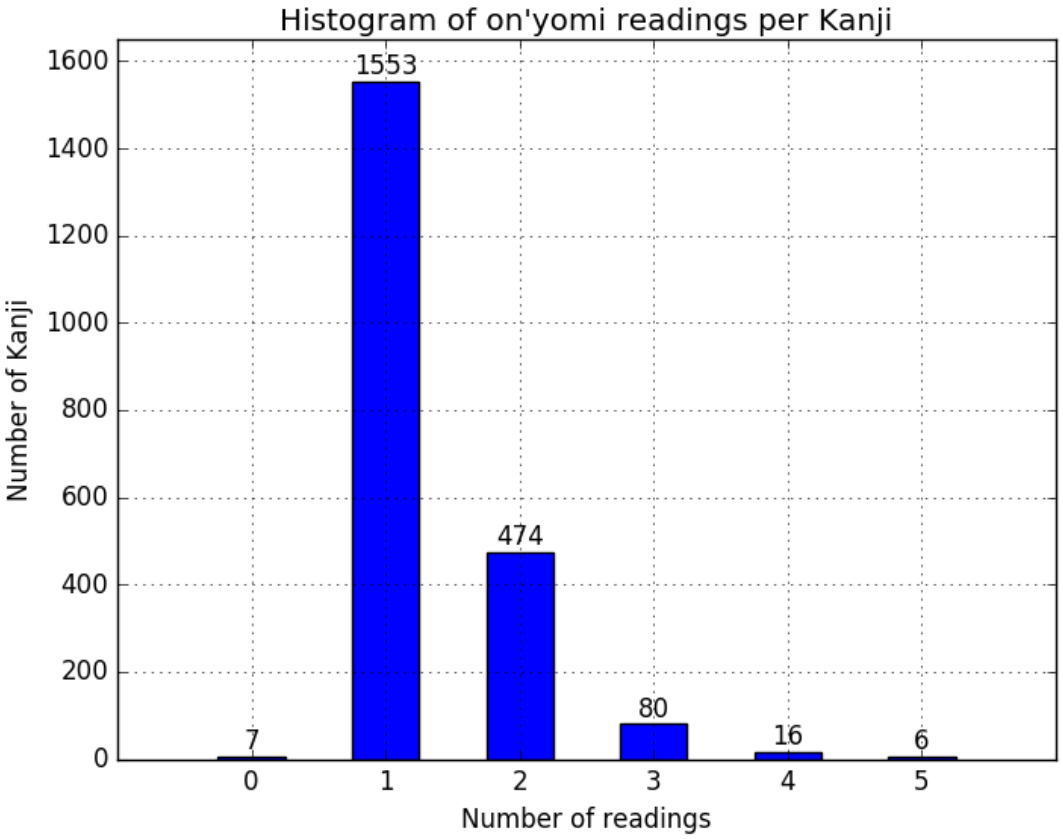


FIGURE 1.2 – A Histogram of the number of on'yomi readings by Kanji

1.2.4 Multiple Japanese readings (kun'yomi)

Even before the introduction of writing through Chinese influence, Japan already had a spoken language that had to be fitted to the imported characters. This kind of reading is mostly used for isolated Kanji, used as substantives, adjectives or verbs. To use again the analogy with Water/Hydro/Aqua, the kun'yomi reading would be the equivalent of the word "Water".

Once more, the Kanji 数 is one that presents multiple readings. The officially recognized Japanese reading of this Kanjis are "kazu" and "kazo", though the readings "se", "wazurawa" and "shibashiba" are also used.

Figure 1.3 presents a histogram for the number of readings by Kanji. As it can be noted, about half of the Jouyou Kanji have only one reading, but about 430 have two readings, 370 have no Japanese readings and almost 300 present three or more readings.

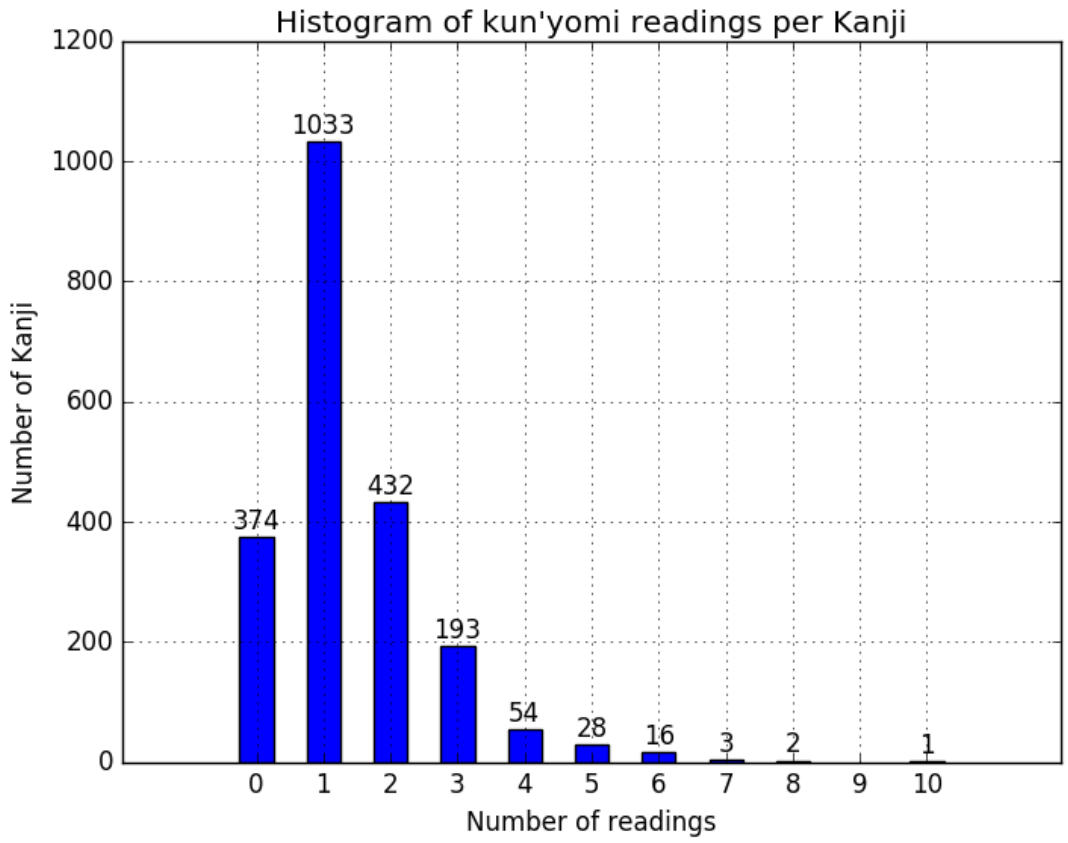


FIGURE 1.3 – A Histogram of the number of kun'yomi readings for Kanji

1.2.5 Readings for use in names (nanori)

There is one more important type of reading for Kanji that should be considered by Japanese learners: the nanori reading. This type of reading represents readings for Kanji that are used in proper names (for example the name of cities, prefectures or people).

The Kanji 希, which represents hope or request is a case with multiple nanori readings. This Kanji has the Chinese readings “ke” or “ki” and Japanese reading “mare”, but in names it can be read as “nozo” or “nozomi” (note that “Nozomi” by itself is a full Japanese name).

Figure 1.4 presents a histogram for the number of readings by Kanji. As it can be noted, about 1200 Jouyou Kanji have no nanori readings, but 360 have one reading, 213 have two Japanese readings and about 350 present three or more readings.

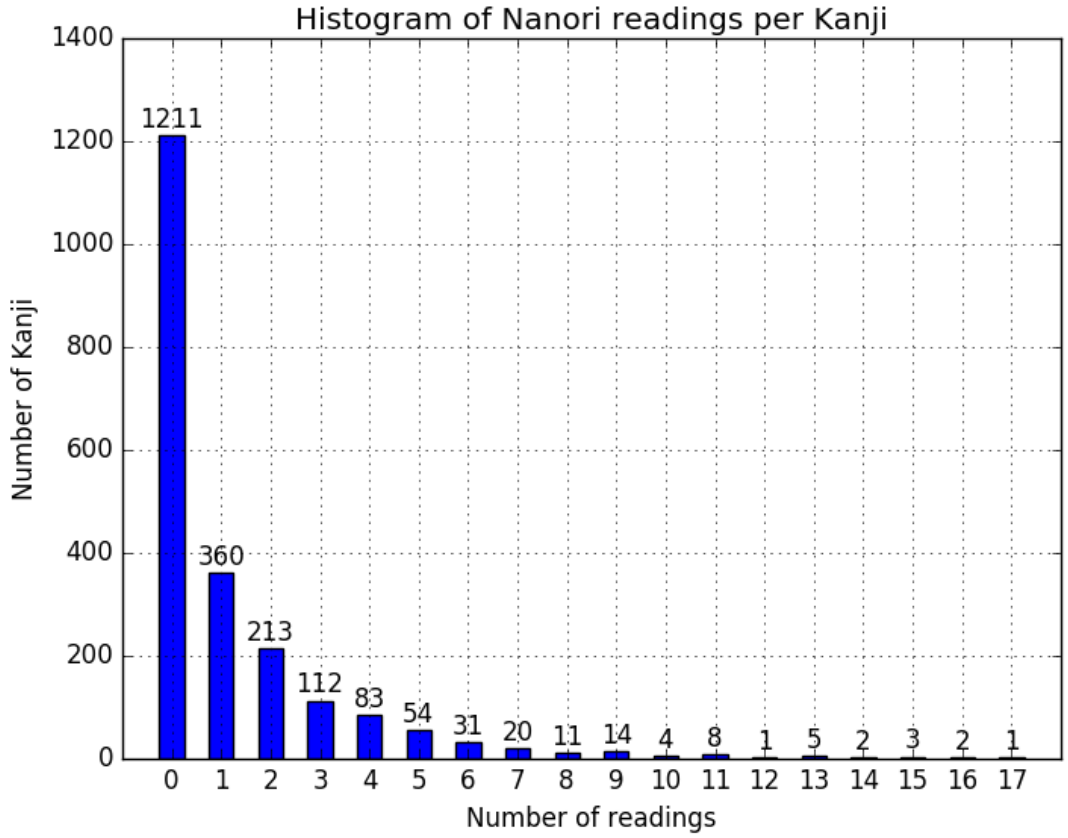


FIGURE 1.4 – A Histogram of the number of nanori readings for Kanji

1.2.6 Irregular readings

As if not enough, Kanji can also be read in a completely different fashion than its usual reading. One such example is the word 一昨日 that can be read in a regular fashion as Issakujitsu (いっさくじつ) where all parts of this word can be matched to its three ideograms: is/saku/jitsu. However, this word can also be read as Ototoi, a reading that have no connection whatsoever to the accepted Kanji readings of this word.

1.2.7 Conclusion

In brief, we can say that the task of teaching one single Kanji passes through teaching:

- How this Kanji is written (including stroke order);
- One or more possibly dissimilar conceptual ideas to associate with it;
- Zero or more Chinese readings;
- Zero or more Japanese readings;
- Zero or more Naming readings;
- Possible words where it appear with an irregular reading;

1.3 Present approaches for learning Kanji

A number of different methods have been proposed to teach the Japanese Kanji. For Japanese natives, the Japanese Ministry of Education have selected a list of 1,006 Kanji, the Kyōiku Kanji (教育漢字, literally the "Educational Kanji"). This list divides those 1,006 Kanji that are expected to be learnt by a native Japanese from grades one through six. Note that among Kanji of the same grade there's no order established by the Ministry of Education, and thus each school can choose the order according to its own preferences. The remaining 1,130 Kanji are expected to be studied through the three years of Junior High School. The number of Kanji that are selected for each grade is presented in Table 1.2.

This list of Kanji was developed through the combination of multiple different approaches, and do not directly reflect the frequency that these characters are present in Japanese media. Additionally, this list was put together with consideration to the age of the child in question. Considering that a First Grade child should be around 6 years old, the first grade Kanji refer to concepts that should be familiar to children that age, such as

TABLE 1.2 – Number of Kanji to be taught per grade

Grade	Number of Kanji
First	80
Second	160
Third	200
Fourth	200
Fifth	185
Sixth	181

sun, moon, th ordinal numbers (1 through 10, 100 and 1,000), left, right, big, small, male and female. One example that shows how this list may not be the ideal for non-native speakers to follow is the Kanji 関, which means "to be related" or "connected". This Kanji is only taught in the fourth grade to Japanese students, but it is the 56th most common Kanji that is used in the Japanese version of Wikipedia.[3]

Another Source of order to study Japanese characters is the list that was provided by the JLPT – the Japanese Language Proficiency Test (日本語能力試験). The organization responsible for this exam provided Kanji lists for each of its 4 levels of certificates: 103 Kanji for level four, 181 additional Kanji for level three, 739 additional for level two and finally 903 additional for level one, totalizing 1926 Kanji. Since its renewal in 2009, the JLPT became a five levels exam and stopped providing Kanji list. Since those official lists take in consideration non-native Japanese students, it is still a widely used guide to the order of Kanji to be learnt, although no order withing this levels was predetermined by the JLTP organization.

With this and other issues presented in section 1.2, it is evident that non-native speakers can benefit much with learning media that are targeted specifically to their needs. Below are listed a number of different media that seeks to teach Japanese Kanji to non-native speakers, each followed by a critique of pros and cons.

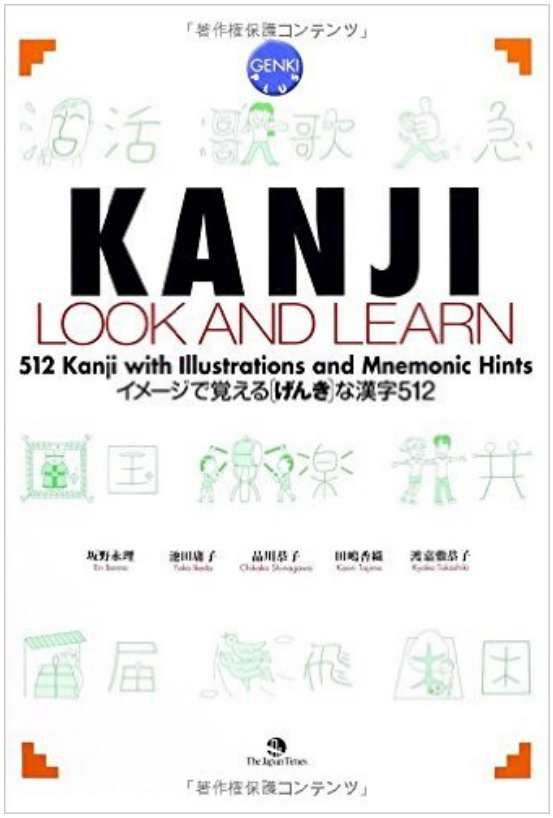
1.3.1 Learning Books

1.3.1.1 Kanji Look and Learn

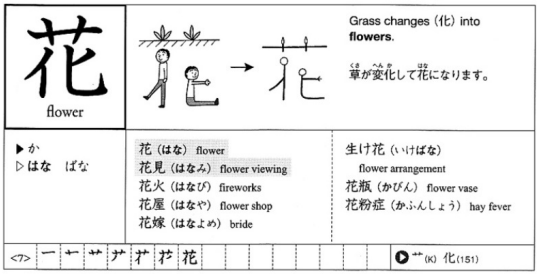
Kanji Look and Learn(BANNO, 2010) is a book by a collaboration of 5 Japanese teachers and is one of the most famous books for learning Japanese Kanji. In fact, it is a series of books that divides Kanji according to its grade on the previous lists issued by the JLPT organization. Figure 1.5 presents the book cover and a small excerpt of the book.

Like all the other examples of this section, this is a great book. Some advantages it has are:

[3]This data is the result from a study made in this work, which can be found in chapter 2.



(a) Cover of Kanji Look and Learn.



(b) Excerpt from the book.

FIGURE 1.5 – Cover and Excerpt from Kanji Look and Learn.

- Shows visual representations for Kanji using real life objects;
- Shows Japanese and Chinese readings;
- Presents multiple examples per Kanji;
- Displays stroke order;
- Separates Kanji among morphological radical parts;
- It has a simple clean interface for each Kanji;

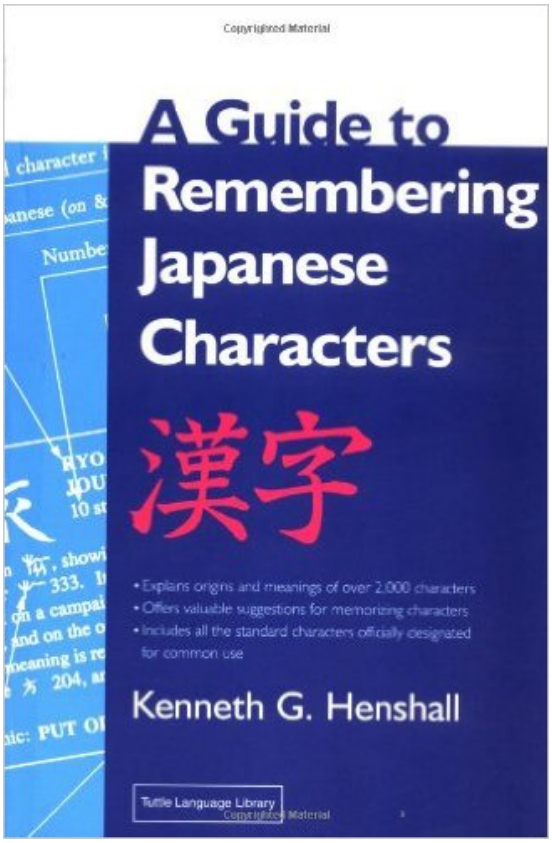
Some limitations it faces are:

- As any book, finding an information can be hard, encouraging a passive form of study (indexing problem);
- No Nanori readings are presented;
- Just a small number of Japanese and Chinese readings are presented;

- Even if radicals are broken up, it is hard to answer simple questions as "which other Kanji shares this same radical" or "is this Kanji visually similar to any other Kanji" or "is this Kanji a sub-component of any other Kanji";
- Definitions for words are short, making it possibly hard to understand the concept;
- Example words are not sorted by frequency, and a relatively small number of examples is shown;

1.3.1.2 A guide to Remembering Japanese Kanji

This Kanji study book by Kenneth G. Henshall(HENSHALL, 1988) is a classic and takes a very different approach to most other books. Firstly it develops from a serious and well structured history research that tries to uncover the true formation factors of the Kanji. It presents Kanji by the order established for grades in the Kyōiku Kanji, and sorts Kanji alphabetically through their Chinese readings. Figure 1.6 presents the book cover and a small excerpt.



(a) Cover of A Guide to Remembering Japanese Characters.

1733 避 **HI, sa*keru***
AVOID
16 strokes

回避 KAIHI avoidance
不可避 FUKAHI unavoidable
避妊 HININ contraception

辟 is an NGU character with a wide range of meanings, such as false, punish, crime, law, and ruler, while in Chinese (even after discounting the obvious borrowings) it can also mean punish, castrate, execute, wail, perverse, specious, flattery, decadent, remove, twist, open, develop, summon, and appoint. It comprises **buttocks** 尸 236, **opening/ hole** 口 20, and **needle/ sharp** 辛 1432 q.v. Buttocks 尸 and hole 口 clearly combine to give **anus**, as in 屁 858 q.v. Needle 辛 is used in its sense of **pierce/ penetrate**, to give **anal penetration** (see also vaginal penetration 商 317). This core meaning gave rise on the one hand to a range of meanings associated with **torture/ punishment**, which also symbolised **law** and **authority**, and on the other to meanings associated with **sodomy**, which when used in relation to a male partner was also a symbol of **flattery**. (Note that when combined with woman 女 35 it gives an NGU character 壁, which in Chinese means sexual partner/ lecherous/ depraved [though in Japanese it is listed with the euphemistic meaning of agreeable person]. When 壁 is itself combined with child 童 363 it gives in Chinese a compound term meaning catamite.) In the case of 1733 辟 acts phonetically to express **evade**, though its semantic role is a matter of conjecture, and combines with **movement** 辶 129 to give **evasive movement**.

Mnemonic: **MOVE TO AVOID NEEDLE IN ANUS**

(b) Excerpt from the book.

FIGURE 1.6 – Cover and Excerpt from A Guide to Remembering Japanese Characters.

Some advantages it has are:

- Presents a very accurate description of the character morphological history, also defining which of its radicals are thought to have phonetic influence and which are thought to have conceptual influence;
- Shows On and Kun readings;
- Displays stroke number;
- Presents example words;
- Presents a mnemonic;
- Numbers Kanji in a fashion that is recognized by other important Japanese learning media, making it easy to refer to this book from ;
- cross references Kanji with other information that can be looked up through entry numbers;

Some limitations it faces are:

- Once again, it presents indexing problems for looking up Kanji and relations between multiple Kanji;
- No Nanori readings are presented;
- Just a small number of Japanese and Chinese readings are presented;
- Does not always breaks a Kanji in its radical components;
- Definitions for words are short, making it possibly hard to understand the concept;
- Example words are not sorted by frequency, and a relatively small number of examples is shown;
- A great number of Kanji have very obscure history, provided that many of these have millennium of history, making this kind of study frustrating at times;
- Each entry presents a great deal of information which can not be minimized to browse only the essential pieces of information and connections;

1.3.2 Websites

1.3.2.1 Jisho.org

The Denshi Jisho project(AHLSTRöM, 2016) describes itself as: “Jisho is a powerful Japanese-English dictionary. It lets you find words, Kanji, example sentences and more

quickly and easily". It is a free and open project that serves many functions, being in essence a very powerful and clean interactive dictionary in English for the Japanese language. As written in the website footnote, it was "lovingly crafted by Kim, Miwa and Andrew". Figure 1.7 presents an example entry of the Kanji dictionary functionality of this website.

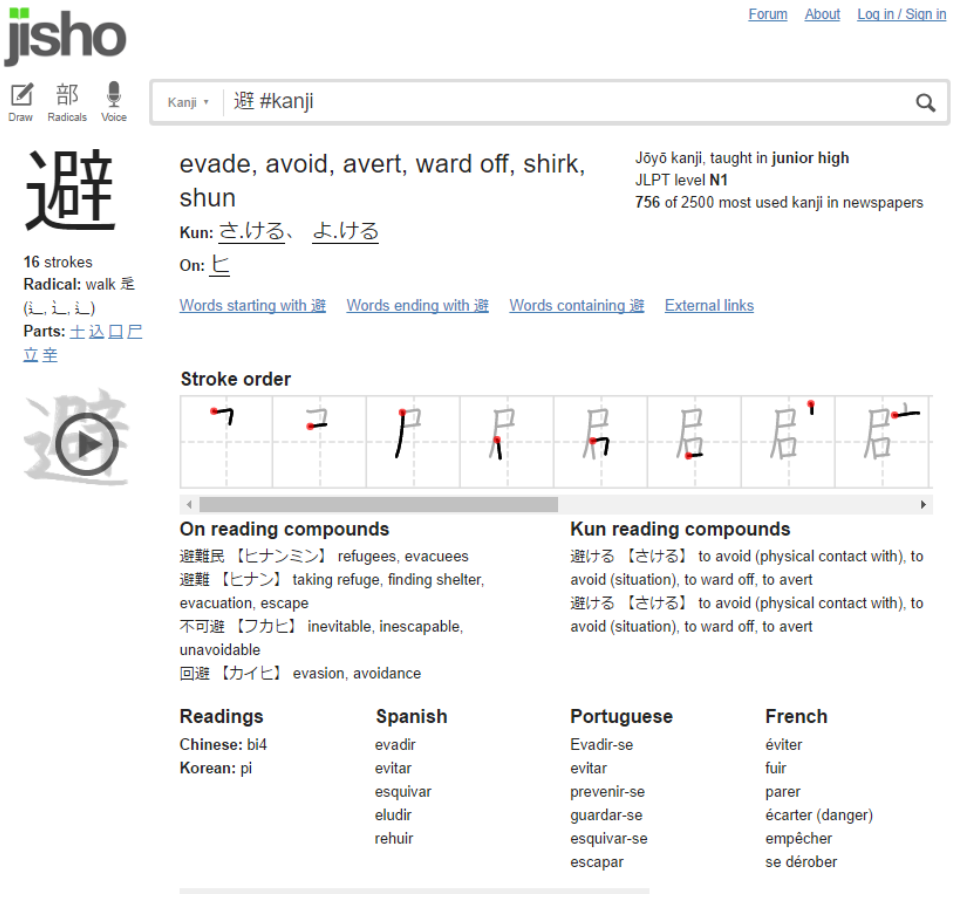


FIGURE 1.7 – An example entry from Jisho.org

It is a very useful website and has the following perceived advantages:

- It has a very complete search capability, being able to look for Kanji from hand drawings, its radical parts (refer to figure 1.8) and even through voice;
- Presents multiple links that can be followed for more information about a part, including some component parts;
- Provides a very complete set of definitions each Kanji can assume;
- Separates examples were it can be seen with its Chinese and Japanese Readings;
- Presents Nanori readings, when they exist;
- Displays definitions in multiple languages;

FIGURE 1.8 – Radical look-up feature from jisho.org

- Has animated and step by step stroke order guides;
- Displays information of frequency, JLPT level and Kyōiku grade;
- It is free, which increases its reach;

Unfortunately, it has a few shortcomings:

- The list of parts it is composed of is confusing, and does not follow any clear rule;
- Does not present simple relationships among characters, like which characters are composed of which other Kanji. For example, the Kanji in figure 1.7 is 避 and has one big component that is not shown in Jisho's list: 辟. The provided list falls somewhat between trying to show all components to showing some that are apparently disconnected from this Kanji (such as 込). Following through the analysis of 辟, we note that it appears in four and four only Jōyō Kanji: 避, 壁, 壁, 癖. This relationship cannot be attained in any easy way from Jisho's website, since the radical 辟 is not one of its supported at its radical look-up screen (refer to figure 1.8) either;
- Does not have a learning platform, only a browsing feature;
- Does not present a feature where characters are listed in ascending or descending order with their frequency;

1.3.2.2 Wani-Kani

Wani-Kani(TOFUGU, 2016) is a paid web-app from where the "parts" and radicals used in Jisho.org were granted. It has a very complete learning feature, with a dedicated team to hand proof its data and designers to make elaborate custom mnemonic drawings. Its merits, according to its own description, are listed in figure 1.9 and also explained below.

- Mnemonics: Has one mnemonic for every Kanji;
- Radicals: Approaches every Kanji from a building perspective, always teaching radicals prior to a Kanji;
- Kanji: Attempts to teach all 2,136 Jouyou Kanji, in a way that is "cleverly ordered";
- Vocabulary: It uses example words, as all previous books;
- Lessons & Reviews: The website does not provide a browse feature (at least not until a character is considered to be learned), but only a Lessons and Reviews board;
- Spaced Repetition System: uses SRS theory(BATURAY *et al.*, 2009) to make spacing custom to each student;

Although all of these and other merits can rightfully be associated with Wani-Kani, it also has a number of perceived issues:

- It is a paid service, which curbs its access rate and therefore its teaching potential;
- The website does not adapt well to its user preferences. For example: there is no functionality that lets the user edit and choose its own mnemonics in favor of those proposed unilaterally by Wani-Kani;
- Radicals are not well separated from Kanji, and the approach this website use to define which are the components of a Kanji (a process that is not always so straightforward) is not clear, sometimes referring to radicals for their visual appearance (As Kanji Look and Learn does) and some times for technical correctness (as Henshsall's Guide does);
- The "cleverly ordered" form does not follow any apparent technical guideline, just humans' subjective opinions;
- Shows a very limited number of words for each Kanji (a mean of three).
- Does not provide a leveling method, following the same bottom-up approach for every student;
- Its system does not facilitate browsing through Kanji, nor does it show important relations also missing from Jisho.org;
- Its Spaced Repetition System algorithm appears to treat Kanji as independent learning subjects, when in fact they have very rich relational structures among themselves;

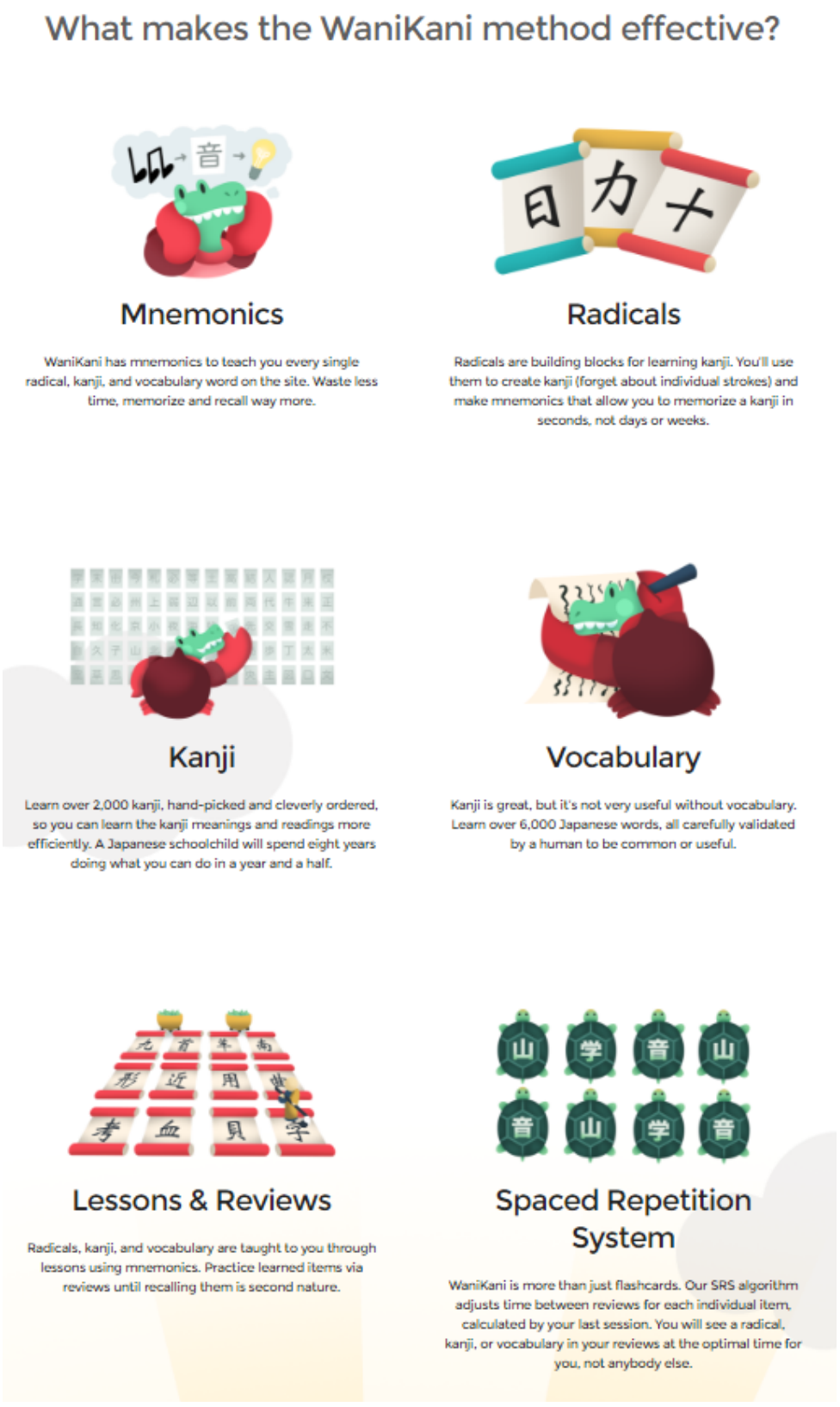


FIGURE 1.9 – Wani-Kani advertised advantages

1.4 The need for a new way of learning

Taking in consideration all the advantages and disadvantages listed for the provided examples, we clearly notice that a new way of learning would greatly benefit Japanese learners. Firstly, this new way of learning should be a website, since it inherently holds many advantages in comparison with books, such as a bigger space for information, interactivity, indexing functions, browse and learning aids.

On a second note, it can be noted that Jisho.org and Wani-Kani are already great learning aids, but each of them still holds some limitations. The work developed in this Academic Work is a stepping stone in the direction of creating a website that takes advantages of most of the functions that are considered to be advantages on those two platforms, as well as addressing their shortcomings.

Chapter 2 elaborates on the issue of taking Kanji frequency in Japanese media more seriously, so that the a student can have maximum success in learning frequent Kanji in minimal time.

Chapter 3 elaborates the concept of teaching order further, elaborating two graphs that relate Kanji among themselves (a morphological graph and a co-occurrence graph), as well as providing predictions of behaviors of students in analogy with random walk models with random restarts.

Finally, Chapter 4 uses the knowledge obtained in previous chapters to propose specific details of the future website functionality in order to create an efficient teaching aid for the study of Japanese Kanji.

2 Linguistic Frequency Distribution

As explained in Section 1.4, one aspect that deserves special attention when teaching Japanese is the **importance** of a word or Kanji. Importance can be defined in a number of ways, such as the definition implied by the grade lists brought forward by the Japanese Ministry of Education, the Kyōiku Kanji, which was introduced in Section 1.3. This grading system uses a constructivist approach, where Kanji that hold more simple and concrete meanings are taught earlier than Kanji that represent more complex or abstract ideas. Additionally, the Kyōiku Kanji grades also reflect the frequency which characters are expected to be present in texts compatible with the age of the student in question[1]. Although this specific order proves itself useful when teaching native Japanese children, it is not necessarily the best order to teach foreign adults that are seeking to learn Japanese as an additional language.

Taking these cases in mind, we propose that for adults, a much more concerning factor in deciding which Kanji should be tackled first and with more seriousness[2] would be the relative frequency this Kanji is used in media that the student is likely to be interested in. It is important to note that the order proposed in this chapter is not the final order that would be advised to be used in the study of Japanese Kanji, but only a stepping stone in the direction of obtaining this better list. Chapter 3 elaborates more on other factors that should be considered when estimating the importance of a Kanji character.

In the next sections, we describe the process of seeking an appropriate source of data, compare the total word distribution and Kanji-only distribution to a power log distribution (also known as Zipf's distribution) and finally explore the implications of the relationship of the specific format of the frequency distribution curves and the learning process of a student.

[1]For example, a student learning first grade Kanji should be around six years old, and therefore be presented more often with text where the concepts, and therefore the Kanji, are simple

[2]In here, we assume that the study technique will be similar to a factorial method: the student will learn one topic and progress to learn other topics returning from time to time to check if the first concepts are clear. If this is the method used in a web app to teach Japanese, the first Kanji to be taught would also coincide with those that are expected to be the best understood

2.1 Frequency estimation of Japanese text

Japanese has a number of particularities that make the frequency estimation of words (or lemmas) to be unique. Those will now be explained in order to understand the complexity and steps of the process.

2.1.1 Parsing the Japanese Language

The Japanese language does not use the concept of blank spaces in its writing system, although it does use some punctuation marks such as commas, periods and exclamation marks. That being so, the task of reading Japanese also comprises the task of determining the start and end of words and syntax particles. As an example, let us examine a simple sentence in Japanese: *The cat is sitting on the roof.*

猫が屋根に座っている

Firstly, it can be noted that there are ten graphemes in this sentence, but no spacing marks, as mentioned earlier. Now, let us artificially introduce space marks along with a rōmaji transliteration[3]:

猫　が　屋根　に　座っている
neko ga yane ni suwatteiru

With this, we identify five separate words in this sentence. It was possible to do the parsing of this sentence due to the previous knowledge of the structure of Japanese sentences and the function each word plays in the sentence. Initially, we read the Kanji 猫, which is read as neko and means "cat", we identify this as an individual word of the sentence and expect the next element to be a syntax particle, designing the function of this element in the sentence. In succession, we find the Hiragana が, ga, the subject marker, which indicates that "cat" is the subject. We then proceed to look for the predicate of this sentence and find the word with two Kanji 屋根, yane, which means "roof". We then expect to see a direct or indirect object particle marker, and thus we find the Hiragana に, ni, the indirect object marker. Finally we expect to find the verb of the sentence [4], and read the Kanji accompanied by some Hiragana stem 座っている, suwatteiru, the verb "to sit" in the present progressive form.

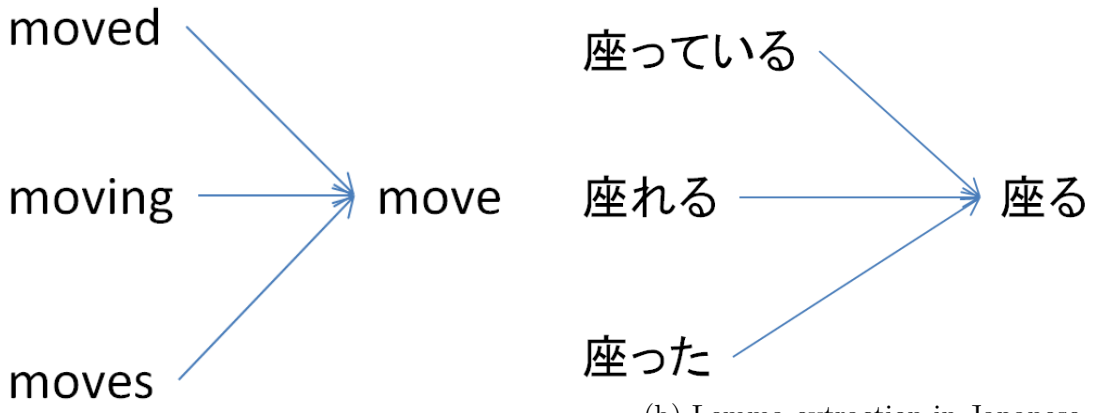
[3] refer to Section B.1 in the Appendix for more information about the romanization process

[4] Japanese is structured around Subject/Object/Verb, instead of the form Subject/Verb/Object, which is more common in western languages

This task of identifying word boundaries is much facilitated by the intermixing of Kanji, Hiragana and occasionally Katakana in sentences, which is one of the reasons that makes Kanji so useful for the Japanese language.

2.1.2 Optional Lemma Extraction

Optionally, the occurrence of elements can be done around lemmas instead of words. Lemmas are the unification of multiple words under a same root such as the uninflected forms of words. With this process, different representations of the same concept can be condensed to a unique entry, instead of having it spread in multiple entries. In English, the process would be similar to unifying the words *"moves"*, *"moving"* and *"moved"* to the word *"move"*. Figure 2.1 exemplifies the process in English and Japanese. In the case of the Japanese language, there is an additional issue that may be solved, which is the unification of homographs under a same lemma. For example, both spellings お勧め and お薦め are valid ways to write おすすめ, osusume, which means recommendation.



(a) Lemma extraction in English.

(b) Lemma extraction in Japanese.

FIGURE 2.1 – Comparison of lemma extraction in English and Japanese.

2.2 Projects that estimate word frequencies

A variety of different projects undertake the mission of determining which are the most common words in the Japanese Language. Each of the next three presented projects were taken from a different type of media.

2.2.1 Alexandre Girardi's analysis of newspaper data

This is the oldest of the three projects, dating to 1998, and was taken from parsing 4 years of the newspaper Mainichi Shinbun (毎日新聞, literally "Daily Newspaper"), in

a total of 74,721,217 occurrences spread among 311,049 words(GIRARDI, 1998). This is the most widely used project when estimating Kanji frequency rank, since this rank is already in built in a number of open source projects, including the most widely used open-source dictionary of Japanese words, Jim Breen's WWWJDIC(BREEN, 2000). Alexandre Girardi's project has four main problems: first, the fact that the type of speech used in newspapers in Japanese is radically dissociated from the common day-to-day written form of Japanese. A second issue is that it is skewed to some few topics, such as economy, politics and weather forecast. The third issue is that this project is based around word extraction, instead of lemma extraction, which dilutes different forms of the same concept in multiple examples, that may not be available in a dictionary[5]. Finally, a fourth issue is that the methodology used in this project was not comprehensively published by Alexandre Girardi, limiting itself to some few lines in a readme file.

2.2.2 Wikitionary analysis of a Japanese Wikipedia dump

This project was drawn from a complete dump of the Japanese Wikipedia on April 22th, 2015, using the morphological analyzer MeCab(KUDO, 2005) and cleaning the end results to unify words under lemmas, yielding a total of 669,419,716 occurrences spread among 2,610,776 lemmas(WIKITIONARY, 2015). It is important to note that a great number of lemmas appears only one or two times. A great number of this rare lemmas are attributable to Hiragana transliterations of Kanji words inside parenthesis, a phenomena that is customary for the first sentence of each article. For example, the article about the Japanese language starts with:

日本語（にほんご、にっぽんご）は、

As we can see, this article starts by writing "Japanese Language" in Kanji, followed by two different possible readings for this word. In this case, since a sequence of Hiragana can be usually converted to a number of different words with different meanings, the lemma extraction failed to unify these with the lemmas containing Kanji.

If we limit the examples only to lemmas that appear more than once, the total number of occurrences is reduced to 668,024,047 occurrences over 1,214,107 lemmas. If we limit to only lemmas with at least three samples, the total number of occurrences is then reduced to 667,301,019 occurrences over 853,593 lemmas. As it can be noted, each step in this direction reduces the total number of occurrences by only a small fraction, while the number of words drops quickly in comparison with its scale. This project can be seen as as

[5]A future step of this project is selecting example words and ranking them according to frequency. If word extraction is used instead of lemma extraction, forms existent in dictionaries will be under represented, because of the presence of inflected counter parts.

improvement over Alexandre Girardi's project for its sheer size (over 669 million **lemma** occurrences on this project, while the previous one was restricted to only 74 million **word** occurrences) and the fact that it does lemma unification. However, the written form usually used in Wikipedia is also of a different sort than the form of speech that would be more expected to see in day-to-day texts for its formality, and it also presents an imbalance of terms, this time towards date related Kanji (the Kanji for "year" is specially oversampled), as well as descriptive terms. Also, simple concepts such as the Kanji for "ear" and other very simple Kanji are undersampled, since they represent simple concepts that are not useful in explaining other concepts.

2.2.3 Christopher Brochtrup's analysis of Japanese novels

In 2007, Christopher Brochtrup created a tool called the "Japanese Text Analysis Tool"(BROCHTRUP, 2012a), for the analysis of arbitrary Japanese texts, including an inbuilt capability of doing morphological parsing of texts through the MeCab(KUDO, 2005) analyser followed by a word frequency report, along with other functionalities. As an example of this tool, on May 27th, 2012, he created a report from 5000+ Japanese novels that were in public domain(BROCHTRUP, 2012b). This analysis was done by lemma reduction to root forms of conjugated words, but with no attempt to unify different spellings of words under one type of writing. It counted a total of 366,120,879 occurrences, spread over 193,121 lemmas. It is a matter of course that this project also shows an imbalance toward words more commonly used in novels, but a comparative analysis of Kanji rank versus Japanese grade did not show any significant problems. As for its size, it is composed of almost 5 times the number of occurrences in the newspaper data and about half the number of occurrences of the Wikipedia project. These are spread over only 193 thousand lemmas, a number much more closely related to the expected size of a language vocabulary.

2.2.4 Final choice and considerations

Each of the presented projects have its own implementation and data-cleaning process particularities, but all of them fall victim to the same shortcoming: being single-sourced.

The relative frequency of words holds a strong dependence with the media in which it is published. That being the case, the most appropriate choice would be taking multiple sources and conjoining them to respect their individual scale of total words and to keep the statistical properties of each project, also adjusting each to be centered around the same lemma creating strategy, instead of a word-listing process. An even greater enhancement would be aggregating more sources of data, and classifying those regarding topics,

interests and expected reader age group, so that a different frequency distribution could be generated for each class of user profile, specifically tailored to that profile's tastes and needs.

Since this task of pluralizing the sources of data was considered to be outside the scope of this academic work, it was not executed and is proposed as a possible improvement over the results presented herein. Instead, we choose the media that was the most expected to be useful to an adult learner with broad interests: **Christopher Brochtrup's analysis of Japanese novels**. That being the case, we will be optimizing the learning experience of a student that tries to read Japanese novels. It is important to note that since this decision was in some sense arbitrary and therefore may be expected to change in a future step of this project, the back-end of the developed analysis framework was designed in a way that does not make any assumptions about the origin of the lemma-count pairs, so that this source could be easily replaced with any other project that provides words in the format "count,word(or lemma)" in a .csv file.

2.3 Zipf's Law

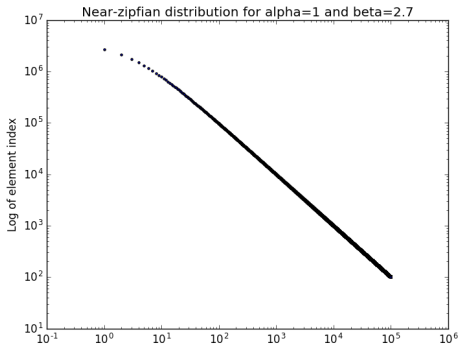
The Zipf law is an empirical law that states that given the evolution of a natural language, the frequency of any given word is approximately proportional to its rank in a biggest-to-lowest frequency table, a rule also known as the power law. It is generally recognized as so since it was popularized by the American linguist George Kingsley Zipf(ZIPF, 1935). His work comprised a hand counting and later on a log-log graphing of James Joyce's book Ulysses(JOYCE; GOYERT, 1926). Zipf noticed that in this log-log analysis the series of points seemed to form an apparent downward straight line, giving rise to the law enunciated above, or in a more technical fashion as Equation 2.1.

$$f(r) \propto 1/r^{\alpha} \tag{2.1}$$

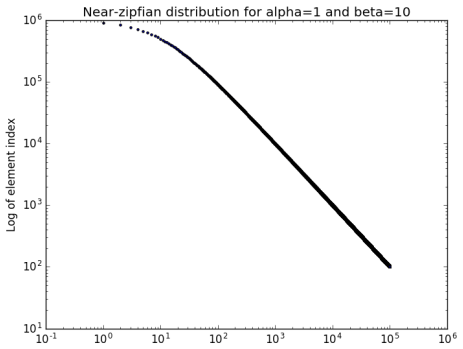
Where $\alpha \approx 1$. A generalization of this law that more closely fits the frequency distribution in languages was proposed by Mandelbrot(MANDELBROT, 1962), and is expressed in Equation 2.2.

$$f(r) \propto 1/(r+\beta)^{\alpha} \tag{2.2}$$

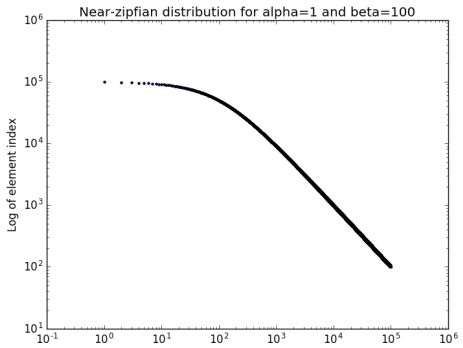
where $\alpha \approx 1$ and $\beta \approx 2.7$. While the original equation portrays a perfectly straight line in the log-log graph, the Zipf-Mandelbrot equation yields a downward facing curve formed by the connection of two approximately straight lines. The behaviour of this β parameter is exemplified in figure 2.2.



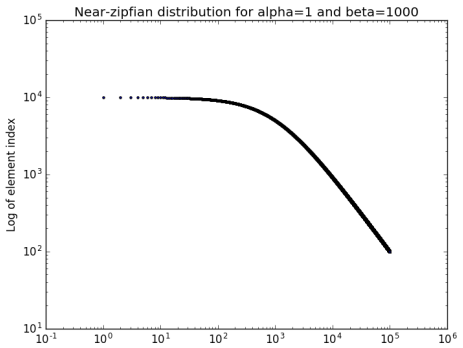
(a) Using $\beta = 2.7$.



(b) Using $\beta = 10$.



(c) Using $\beta = 100$.



(d) Using $\beta = 1000$.

FIGURE 2.2 – Log-Log graphs of data produced according to Zipf-Mandelbrot equation with various parameters.

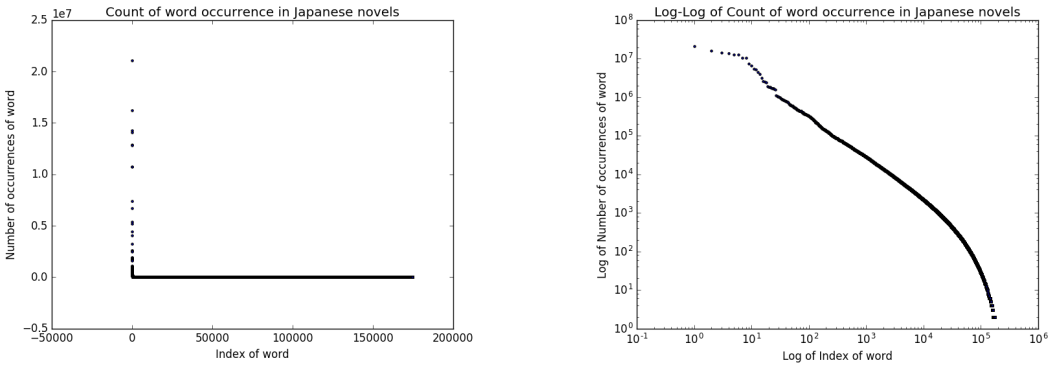
As it can be seen in this sequence of four figures, as the parameter beta increases the fall of the curve is delayed.

Although a number of possible explanations were given about the origin of this phenomenon, with a number of different deductions from base principles arriving at this same formula, very few of these theories focus themselves in giving testable accounts of which should be the right theory for the psychological mechanisms responsible for this effect(PIANTADOSI, 2014).

In the subsequent sections we will analyse a number of frequency distribution and do various analysis of the shape of the Log-Log graphs of these distributions. Although a more rigorous scrutiny over the presence of absence of power log behaviour would be possible through mathematical tools(NEWMAN, 2005), we will focus on a simple visual analysis of the curve, giving more emphasis to the basic insights that these can bring.

2.4 General Word Distribution

Using Christopher Brochtrup report, the process of creating a *occurrence*×*rank* graph is very straight forward. We simply produce the scatter plot of the occurrence of lemmas versus their rank in a sorted list where their counts are decreasing with the increase of their rank. The result of this analysis on linear and log-log scale is presented on Figure 2.3.



(a) Linear distribution of lemmas. (b) Log-Log distribution of lemmas.

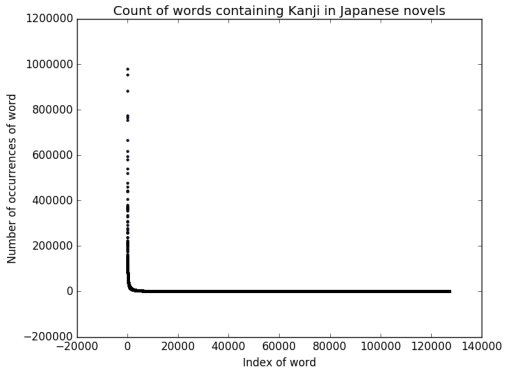
FIGURE 2.3 – Distribution of lemmas used in Japanese Novels.

The linear graph does not bring much insight except for the expected fact that the first few words appear a lot, with a long tail of lemmas that only happen a few times. Upon inspection of the Log-Log graph, we are able to more concretely grasp the exchange rate between the rise of the index and the fall of the number of occurrences of words. In this case, we can visually detect that the curve has a Near-Zipfian behaviour, with most of the core part of the curve closely matching a straight curve. This was to be expected, since most human languages were already observed to be closely related to power log curves.

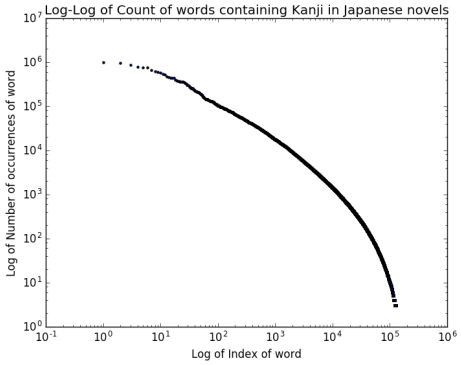
2.5 Distribution of Words with Kanji

We can now refine our earlier data-set so that it is exclusively formed by words with at least one Jōyō Kanji, keeping the ordering as was done in the previous section. The result of this analysis is presented in Figure 2.4.

Once again, we note that the linear plot is only able to bring the expected insight that the occurrences of lemmas are highly concentrated in the fist few ones. Now, upon scrutiny of the Log-Log graph, we observe a slight curving of the plot in the direction of the upper right corner, indicating a slightly different behaviour for this subset. This outward moving of the center of the curve indicates that words containing Kanji tend to be more frequently used than it would be expected of them if they represented the totality



(a) Linear distribution of lemmas.



(b) Log-Log of the distribution of lemmas.

FIGURE 2.4 – Distribution of lemmas used in Japanese Novels that contain at least on Jouyou Kanji character.

of components from a language that follows a power log curve with great fidelity.

2.6 Jouyou Kanji Distribution

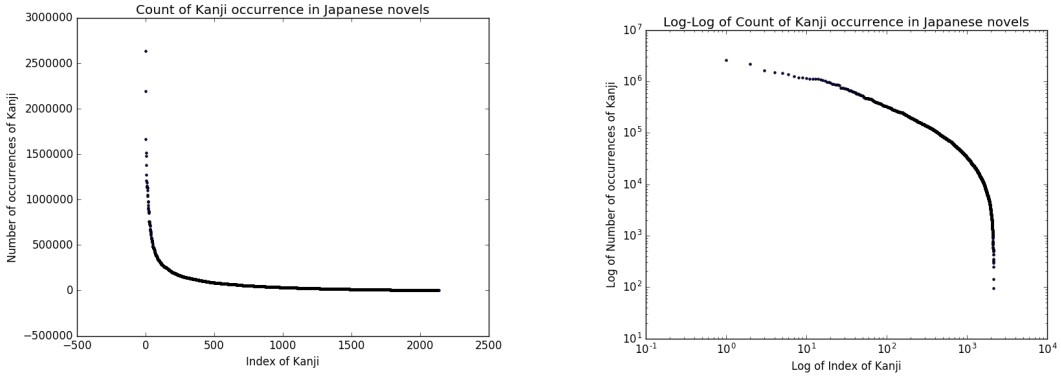
Finally, we can take the subset of words in the last chapter and proceed to invert the center of counting to be around Kanji, instead of being around lemmas. Since we already start the process knowing which are the 2,136 Kanji that should be counted, the task resumes itself on creating a python Counter object (a hash map of string to int) and adding the number of times its host word occurs. This task was performed while simultaneously gathering example words for each Kanji, but for simplicity we present a fictitious code that would perform only the task of inverting this count.

```
import collections
import IS  # Important structures defined for this project.
import IP  # A collection of relevant paths for this project.
import toolbox  # A toolbox to perform useful tasks.

def create_kanji_counter():
    # count_words is a tuple (count, word)
    count_words = toolbox.load_data(IP.WORDS_FILTERED)
    kanji_counter = collections.Counter()
    for c_w in count_words:
        for character in c_w[1]:
            # Checks if the word is in an alternative form
            if character in IS.equiv_to_jouyou:
                character = IS.equiv_to_jouyou[character]
```

```
            # a set containing the Jouyou Kanji
            if character in IS.jk_set
                kanji_counter[character] += int(c_w[0])
    return kanji_counter
```

Using the generated *kanji_counter* object we can now analyse the relationship between the frequency of a Kanji and the total number of times it occurred. These graphs are presented in Figure 2.5.



(a) Linear distribution of Kanji. (b) Log-Log of the distribution of Kanji.

FIGURE 2.5 – Distribution of Jouyou Kanji usage in Japanese Novels.

Once more, the linear graph is just auxiliary to the conclusion that a few of the elements concentrate most of the occurrences, but now we observe a smoother transition from the initial number of examples to the tail of the curve. This behaviour can be even more easily noticed in the Log-Log graph, where a curve that distances radically from a power log curve was created. In the specific case of Kanji, we note that the number of occurrences supports itself high even with an increase of rank from the Kanji approximately until the 10^3 mark. After this point, the distribution abruptly falls in number of occurrences for small logarithmic increments in the index.

This characteristic could be explained by the fact that the list of Jouyou Kanji was artificially created by the Ministry of Education in Japan, instead of forming over the pure necessity of users, as is the evolution process of a language and as it is still observed for general Japanese words. The Kanji chosen by the Japanese Government followed basically two rules, which could explain the two different sections of the occurrence graph:

1. Kanji that represented frequently used concepts.

2. Kanji that were already in use in legal documents from the government. This need arose from the strict rule that all public documents should be written exclusively with Jouyou Kanji. To accommodate this rule, Kanji that were obscure but were used in the legal jargon were introduced in the Jouyou Kanji list.

To better elucidate this concentration of uses in a few Kanji, we generated an alternative view for this case. By dividing the number of occurrences by the total number of Kanji surveyed, we were able estimate the frequency of each Kanji as a fraction between 0 and 1. Furthermore, by adding each element with its previous frequency, we generated a cumulative distribution function of the use of Kanji. This CDF is presented in Figure 2.6.

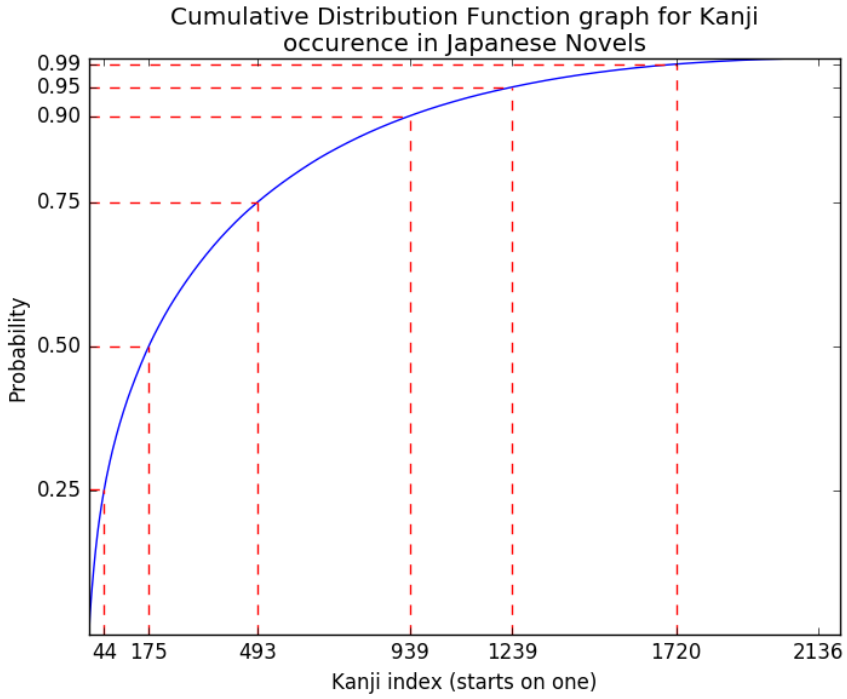


FIGURE 2.6 – Cumulative Distribution Function for the use of Jouyou Kanji in Japanese Novels

The dashed reference lines show the number of Kanji to be learned to reach some certain statistical level. As an example, it can be noted that 175 of the 2,136 Jouyou Kanji account for 50% of the total occurrences of Jouyou Kanji in the surveyed novels. A deeper interpretation of this numbers will be presented on Section 2.8.

2.7 Comparing study efforts on pure frequency versus Kyouiku Kanji order

As a base of reference, we will now proceed to study the distribution of the grades assigned by the Japanese Government for the Kyouiku Kanji list. Since no order is defined inside each of these six grades, we proceed to analyse the study of a student under the

best-case scenario[6] and the worst-case scenario[7]. A representation of this curve on the best-case and worst-case scenarios is depicted in Figure 2.7.

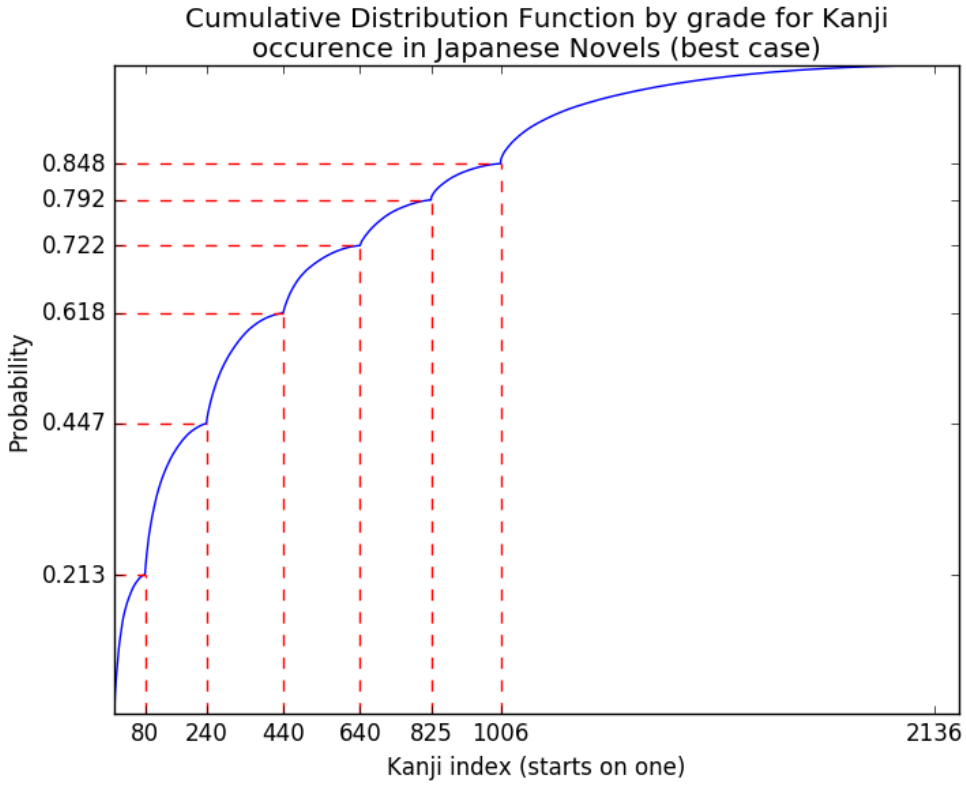
The reference lines in this case represent the number of Kanji to be studied in each grade and the corresponding expected value for the cumulative frequency up to that point. An example of this interpretation would be to state that after learning all the 1006 Kanji from primary school, a Japanese student is expected to recognize 84.8% of the Kanji characters in a Japanese novel. As it could be expected, between the best and worst case scenarios the inflection points remain stable, representing the cumulative probability gain for each grade. The curves differ in each other inside grades and after the end of primary school in its concavity.

To better compare the study of a native Japanese following the grade lists and a student that uses the order implied by the relative frequencies of Kanji characters, those two curves were plotted together in Figure 2.8, again separating in best and worst case scenarios.

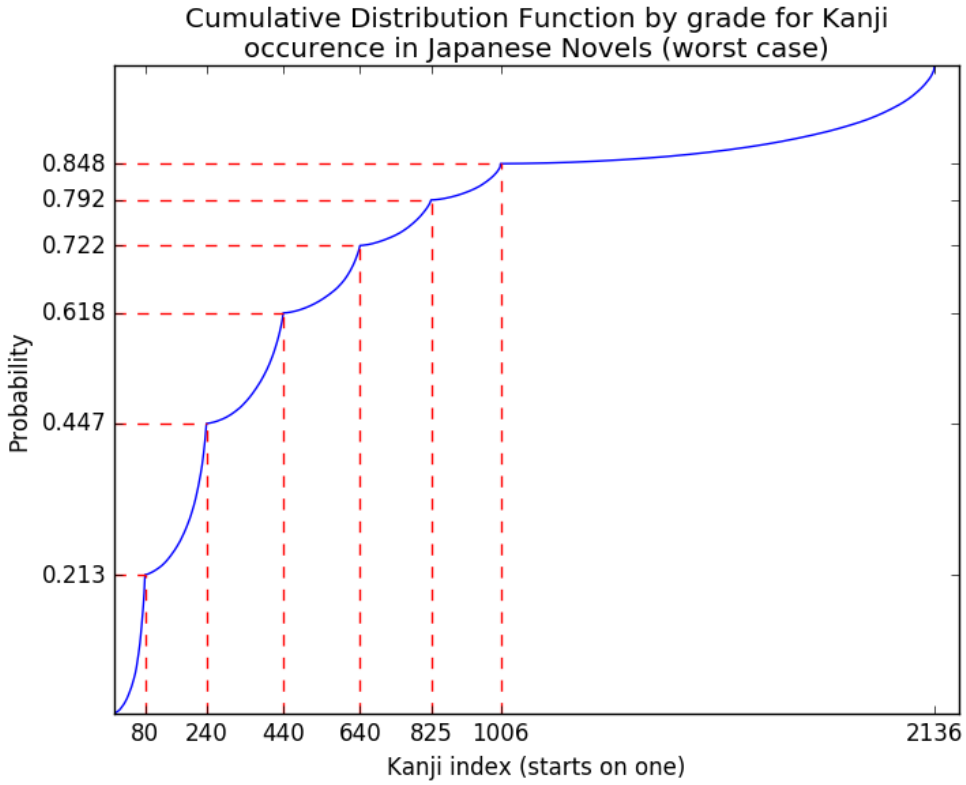
In this case, the reference marks are used to depict the same reference probabilities as before, so that an effective comparison can be drawn between these two methods. The interpretation of these discrepancies are evaluated in the next Section.

[6]although the student limits itself to proceed studying Kanji one grade at a time, he does so studying the most common Kanji first.

[7]In this case, the student limits itself on the grades and study the least frequent Kanji first

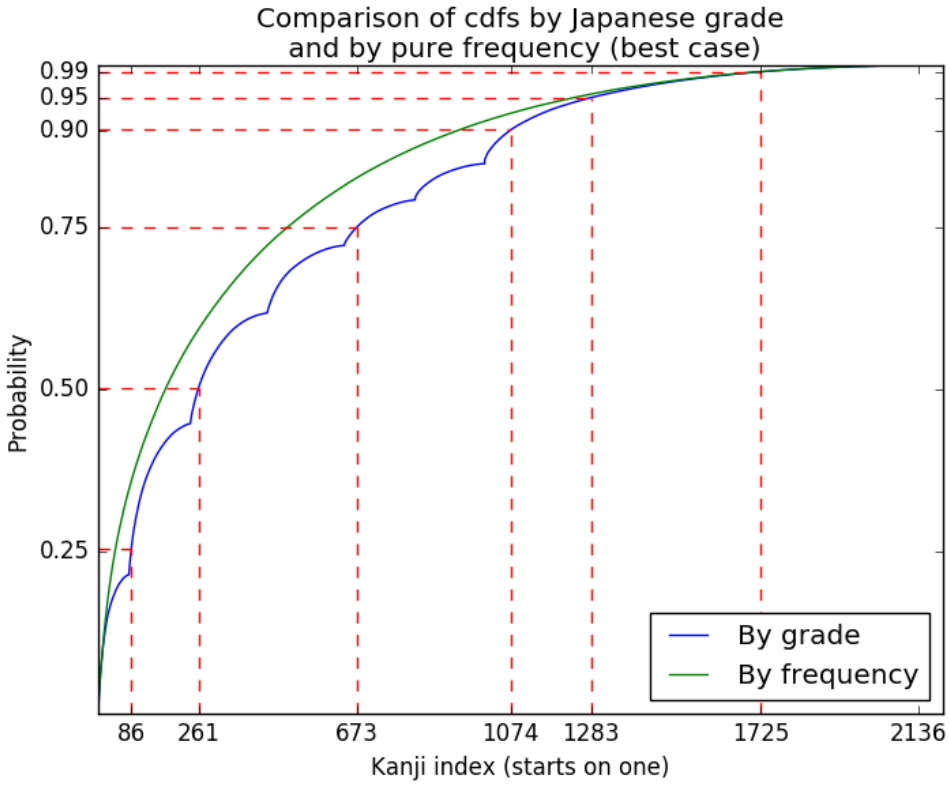


(a) Best case scenario.

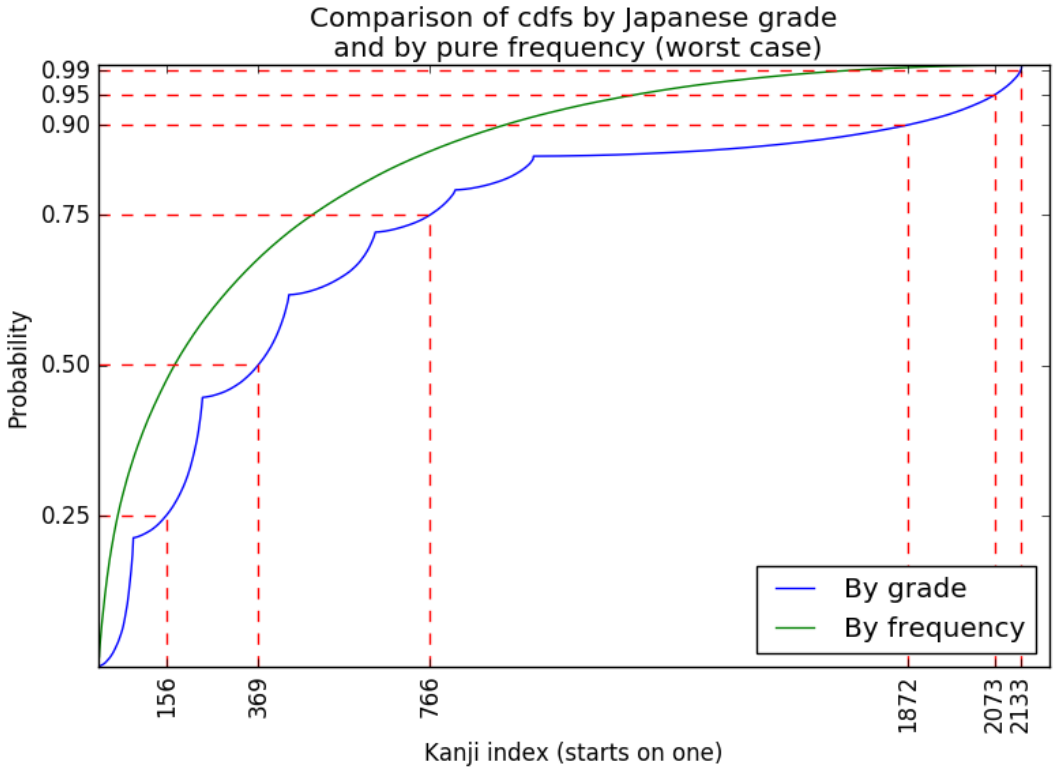


(b) Worst case scenario.

FIGURE 2.7 – Cumulative Distribution Function limited by the grades determined in the Kyouiku Kanji.



(a) Best case scenario.



(b) Worst case scenario.

FIGURE 2.8 – Comparison of the cumulative distribution functions bounded by grade and ordered solely through frequency.

2.8 Kanji Probability Distribution Interpretation

To better understand the information that can be drawn Figure 2.8, first we need to understand what this cumulative probability distribution represents. An intuitive statement of the properties that follow from this analysis is:

"If I follow a certain strategy S and have already studied N Kanji, I will a have P probability of having already studied a Kanji chosen arbitrarily in a random Japanese novel"

With this intuition, we can interpret the markings under the reference lines as the number of Kanji that should be studied if we want to know a certain percentage of characters that could be shown to us in a Japanese novel. A summary of the information presented in Figures 2.6 and 2.8 is put forward in Table 2.1.

TABLE 2.1 – Comparison of Kanjis to be studied to the same cumulative probabilities under different strategies

Probability of having studied	**Number of Kanji studied on Strategy**		
	Frequency	**By Grades (best case)**	**By Grades (worst case)**
25%	44	86	156
50%	175	261	369
75%	493	673	766
90%	939	1074	1872
95%	1239	1283	2073
99%	1720	1725	2133

Firstly, we should appreciate the difference in proportions for the most efficient way of study by studying the first two columns of the table. For example, we note that by studying a total of 8.2% (175) of the total number of Kanji we are already able to recognize half of all the Kanji occurrences in a Japanese novel. If we wish to be able to recognize 90% of the Kanji, we only need to study a total of 44% (939) of these. Furthermore, 99% of occurrences are concentrated in 1,720 Kanji, leaving more than 400 Kanji sharing the last 1% of occurrences. As we already commented in Section 2.6, this distribution is not exactly Zipfian, but it does still follow a weak form of a power log curve, in a manner that a concentrated effort on the first few characters is much more justified than an indiscriminate study of Kanji.

By comparing columns two and three of Table 2.1, we note that to reach 25% of progress while following the Japanese grades but ordering them in the best case scenario, we will need to study roughly double the Kanji needed for the same probability in a study purely guided by frequency, which at this point are a measly 42 extra characters. This

gap increases for 50% and 75% of probability, but closes down close to the end. This behaviour is attributable to the fact that the last 1,133 are not divided in specific grades denominations, so that the gap between these two curves falls asymptotically after the inflection located at 84.8% probability and 1,003 characters.

Now, under the perspective of an unlucky student that follows the grade denominations but studies Kanji in the reverse order of their frequency, results are much more grim. To reach the mark of 25% probability of having studied an arbitrary character, this student needs to study 3.5 times more Kanji, or some extra 112 characters. This gap increases to almost 200 characters on 50% probability, 273 for 75%, 939 characters for 90%, 834 characters for 95% and 413 characters for 99%. Note that at this point, the unlucky student would have studied 2,133 characters of the 2,136 total, leaving the last 1% to be seen in the last three characters, a percentage that was dissipated among 416 characters in the best case.

A list with the ordered Kanji in decreasing frequency obtained from the data collected on this chapter is presented in the Appendix, on Table C.1.

3 A Graph-based Methodology for Kanji Learning

In the previous chapter we explored the concept of Kanji **importance**. At that point of the process, we associated the importance of a Kanji solely with its frequency in the Japanese language, in a way that more frequent Kanji are seem as more important than less frequent Kanji. In this chapter we will explore relations between Kanji to try reaching a more reasonable measure of importance.

There is in Computer Science a field of study that concerns itself with the study of mathematical structures that model pairwise relations between objects, Graph Theory. In this chapter we will bring together our existent need to better rank the importance of Kanji with the existent solutions found in Graph Theory.

3.1 The PageRank Algorithm

In 1998, Larry Page et al. published a paper that radically changed the scene for web search(PAGE *et al.*, 1999). In this paper, they describe an algorithm devised to rank web pages objectively and mechanically, measuring the human interest and attention devoted to each page. In this paper they also compare the PageRank algorithm with an idealized random Web surfer, a comparison that will be useful for our application.

This algorithm estimates the rank of web pages based on the structure of hyperlinks that exist between these pages. Figure 3.1 shows an example graph and the solution to the ranking algorithm in that case.

We can think of this algorithm as a flow for influence from pages, one that does not just count incoming links and outgoing links, but rather takes in consideration the influence of the page where this link is coming from, which is also calculated by this same algorithm, recursively. For example, in this graph we note that the rank for node “E” is smaller than the rank of node “C”, even though the first has many more incoming links than the second. In this case, it can be seen that the pages that refer to “E” are of minor importance, while

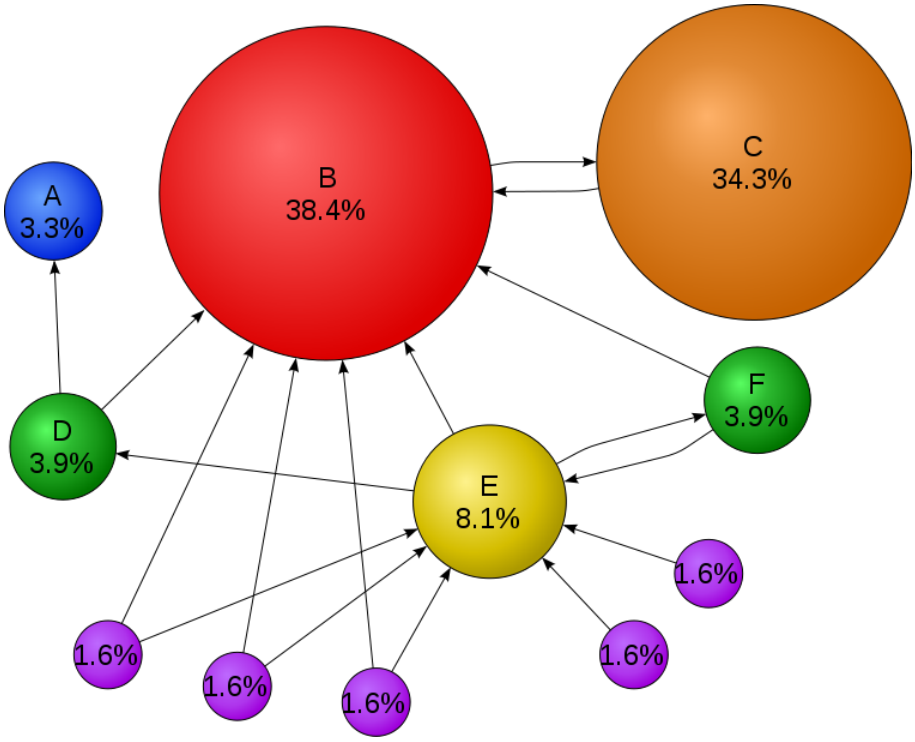


FIGURE 3.1 – Example of the ranking done by PageRank

"C" is the only page hyperlinked from "B", the most important page in this ranking.

The second form of abstraction that was described for this algorithm is the random surfer picture. PageRank behaves as an estimator of the probability of the node where the random surfer could be found in the graph after an arbitrary long time since the start starting from any node. Its behavior would be to start at an arbitrary page on the web and then start a game of arbitrarily following hyperlinks to other pages. This representation allows us to find some problems that should be addressed at this point of the description, which are based on where the surfer can or can not go at any given point: the rank sinks.

3.1.1 The Rank Sinks

3.1.1.1 Dead Ends

A first type of rank sink are the dead ends, pages that have no outgoing links. At this point in the algorithm description, our random surfer would get stuck in that page forever. That way, after a long enough time in the graph our imaginary surfer would get stuck in the dead end and the probability to find it there would be one (if only one such page existed). In the mathematical representation of this problem, it can be noted that the importance of the graph will "leak out" through such nodes, since those pages have no other pages to share its importance with. In the example picture, node "A" is a dead end.

3.1.1.2 Spider Traps

Spider traps are a similar kind of threat to our calculation, but it is a more subtle one. Instead of being points where the user can't go anywhere else, spider traps are regions of the graph where the user would get stuck between pages from that region, with no outgoing links to other regions of the graph. One such case in the example Figure 3.1 are nodes "B" and "C", that although both of them are not dead ends, once any user enters through "B" he would be stuck going back and forth between "B" and "C", and another case would happen if the "A" link had a link to itself. In these cases, all of the influence from the web would be drained by these regions.

3.1.2 Random Teleports

To solve issues raised by rank sinks, Page et al.(PAGE *et al.*, 1999) proposed the use of a vector parameter that would be responsible for random jumps in the graph. According to their description, the web surfer would get "bored" from time to time and jump to an arbitrary web page with probability $1 - \beta$. Furthermore, in the case of dead end nodes the chance to jump to another random node would be one.

3.1.3 Non-Homogeneous Random Teleports

Though this approach is sufficient for a basic rating of the web[1], it may be altered to be used for other objectives. This type of customization is commonly referred as Personalized PageRank(PAGE *et al.*, 1999). Although it is important to have random teleports to solve problems of dead ends and spider traps, there's no mathematical restriction that requires this teleport to be over **all** nodes of the graph or that this teleport should pick a jump node with equal probability for every node. The only requirement is that the jump probabilities make up for a total of 100%.

One example of application that personalizes the jump vector is the called Topic-Sensitive PageRank(HAVELIWALA, 2002). This approach forces the random jump to be done to a page that is identified as being of a certain topic, so that the random restarts do not jump to any point of the web, but to a very specific restart set. With this approach, it is possible to rank pages that are identified as being of a certain topic as well as pages that can be found after a small random walk from those pages.

This academic work focuses in another personalization of the jump vector: the non-homogeneous random teleports. In this approach, instead of giving an equal chance to all

[1] Albeit this was the original algorithm for Google's search engine, many other papers were forwarded creating a better understanding of the topic, including an algorithm to counter-attack spam attempts: TrustRank.

the nodes to be the targets of the random jump, we can skew the probabilities so that some nodes are more probable than others. For example, imagine a simple graph with three nodes. With homogeneous random restarts we would have a probability of 1/3 to jumping to any node of the graph. If we use non-homogeneous random restarts however, it is possible to define a custom jump vector such as:

$$\begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix}$$

With this application we can arbitrarily favour nodes in the graph as we see fit to the application at hand.

3.1.4 Mathematical Representation

Now that we've explained all the base intuition to the PageRank algorithm and the abstraction of the random web surfer, we may also delve in to the mathematical and computational base that can be used to implement these concepts.

3.1.4.1 Flow Formulation

Firstly, we can mathematically define the rank of a page j as:

$$r_j = \sum_{i \to j} \frac{r_i}{d_i}$$

Where j is a page, $i \to j$ represents all pages that point to j and d_i is the degree of i, measured as the number of outgoing links of i. That is to say, the rank of a page j is the summation of the rank contributed by each of its incoming links, where each page contributes to the rank of each of its pointed pages by a fraction of its own rank and the number of pages it points. That being so, we consolidate our perspective that pages that point to many other pages will fan out their influence among these, while pages that point to just a few will concentrate its influence on those. Since we start the process without knowing any of the ranks of the nodes, we end up with a system of N variables and N equations. To better solve this problem, we can condense all those equations in a matrix.

3.1.4.2 Matrix Formulation

The matrix that can be created to represent the mentioned flow equations is called an stochastic adjacency matrix. It is named as so because each column sums to 1. To

illustrate this process and the resulting matrix, we introduce an example graph to be solved, which is presented in Figure 3.2.

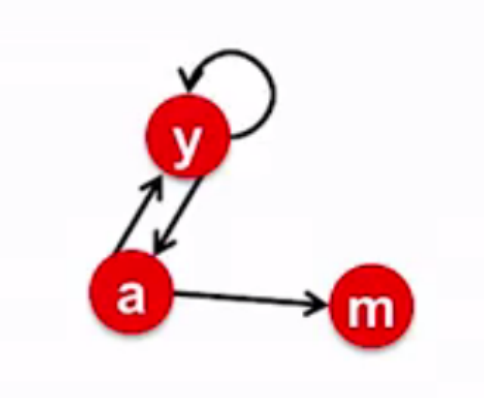


FIGURE 3.2 – An example of a small graph

In this graph, we can write the flow equations as:

$$r_y = \frac{1}{2}r_y + \frac{1}{2}r_a$$
$$r_a = \frac{1}{2}r_y$$
$$r_m = \frac{1}{2}r_a$$

At this point r_m only appears once, since it is a dead end.

To turn this into a matrix we can write:

$$\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix}$$

That is, given that page i has d_i out-links, we set $M_{ji} = \frac{1}{d_i}$ if page i points to page j, otherwise $M_{ji} = 0$.

3.1.4.3 Teleport Vector

The matrix we've seen so far is similar to a Markov transition matrix. We can benefit from this proximity to use a result known for Markov chains:

For any start vector, the power method applied to a Markov transition matrix $\boldsymbol{P}$ *will converge to a unique positive stationary vector as long as* $\boldsymbol{P}$ *is stochastic, irreducible and aperiodic.*(PAKES, 1969)

So, as discussed earlier, our transitions matrix can be made stochastic, irreducible and aperiodic by adding a stochastic column vector to all columns of this matrix[2]. Also, it is necessary that all columns that sum to zero (dead-ends) have 100% probability to jump to a node from the restart set. If we use the homogeneous case we can do a pre-processing step to fix the A matrix as:

$$\begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1/2 & 1/2 & 1/3 \\ 1/2 & 0 & 1/3 \\ 0 & 1/2 & 1/3 \end{bmatrix}$$

And also rewrite our base equation as:

$$\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix} = \left(\beta \cdot \begin{bmatrix} 1/2 & 1/2 & 1/3 \\ 1/2 & 0 & 1/3 \\ 0 & 1/2 & 1/3 \end{bmatrix} + (1-\beta) \cdot \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \right) \begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix}$$

If we now name the rank vector as r, our stochastic adjacency matrix as M and the teleport matrix as H we can rewrite the equation as:

$$r = (\beta \cdot M + (1-\beta) \cdot H) \cdot r$$

And we can proceed in renaming $A = \beta \cdot M + (1-\beta) \cdot H$ to yield our final equation:

$$r = A \cdot r$$

3.1.4.4 Solution Through Eigenvector calculation

If we closely inspect the final equation for the PageRank calculation, we note that our equation is in the form

$$\lambda x = A \cdot x$$

That is, our equation represents the calculation of the eigenvector correspondent to the eigenvalue where $\lambda = 1$.

Since each column of M sums to one and r is a stochastic vector, it follows that $Mr \leq 1$, so the sought after eigenvector is in fact the first or principal eigenvector.

[2]This can be any stochastic vector as long as all elements are different from zero. If some elements are equal to zero, we can not guarantee that the final matrix has these properties.

3.1.4.5 Solution Through Power Iteration

Although we could use a linear algebra package to solve our final equation, we have in hands a more efficient method to calculate a single eigenvector which has its eigenvalue known: the power iteration method. With this method, we start initializing $r^{(0)} = [1/N, ..., 1/N]^T$, and then iterate $r^{(t+1)} = M \cdot r^{(t)}$ as long as $|r^{(t+1)} - r^{(t)}|_1 > \epsilon$, where $|x|_1$ represents the L_1 norm of x(ARASU *et al.*, 2002).

Furthermore, it can be noted that matrix M is generally sparse, and matrix H consists in the horizontal repetition of a column vector. Using this fact, we can rewrite our power iteration as:

$$r^{(t+1)} = \beta \cdot M \cdot r^{(t)} + (1 - \beta) \cdot H'$$

Where H' is a single column of H. This step can be achieved since all the elements in the same row of H are equal, following from the quality that this is a column vector repeated horizontally. So, the multiplication of $H \cdot r$ can be written as:

$$H \cdot r = \sum_i r_i \cdot H'$$

However, r is a stochastic vector, meaning that the sum of all of its elements is always one.

Although other approaches designed to work at the scale of millions of nodes exist, this method is efficient enough for the calculation of a problem as small as the determination of ranks of the 2,136 Jouyou Kanji.

3.2 Modeling the study of Japanese Kanji as a PageRank Problem

Let us now imagine a foreign student of Japanese. Once he starts learning, he will be gradually faced with random Kanji, according to what type of media he is using as a base for his study. Also, as will be explained in more detail in the subsequent sections, Kanji may present relations to one another. In this framework, our imagined student will start learning an arbitrary Kanji, chosen with probability proportional to the occurrence rate of that Kanji in media of his interest and then he may start studying any Kanji that is related to that first Kanji, in a way to reinforce his knowledge. After that he may keep navigating that network of related Kanji for some hops, until he falls back to the first step of his routine: he will randomly jump to a new Kanji, with a probability proportional to the occurrence rate of that Kanji.

Classic learning methods mostly fail to mimic this learning pattern, since they focus solely either on the frequency this Kanji is taught or in the order proposed in the Kyouiku Kanji list, not having a statistical framework to justify importance estimation.

It is clear that the behaviour of this imaginary student is very similar to the abstraction of the random web surfer idealized for the PageRank algorithm. With the understanding of graph algorithms and our initial measure of importance for Kanji, we can now look upon the relations between different characters as an important source of information to rank Kanji. Our approach will be a slightly modified version of PageRank, one where the jump vector is non-homogeneous and scaled with the frequency which this character appears in written media.

3.2.1 The Morphological Graph

The first graph that was considered is the morphological graph, a graph that takes advantage of the fact that most Kanji reuse the same few base components, which are on occasion Kanji by themselves. Figure 3.3 helps to better illustrate this concept, through the break down of a single Kanji.

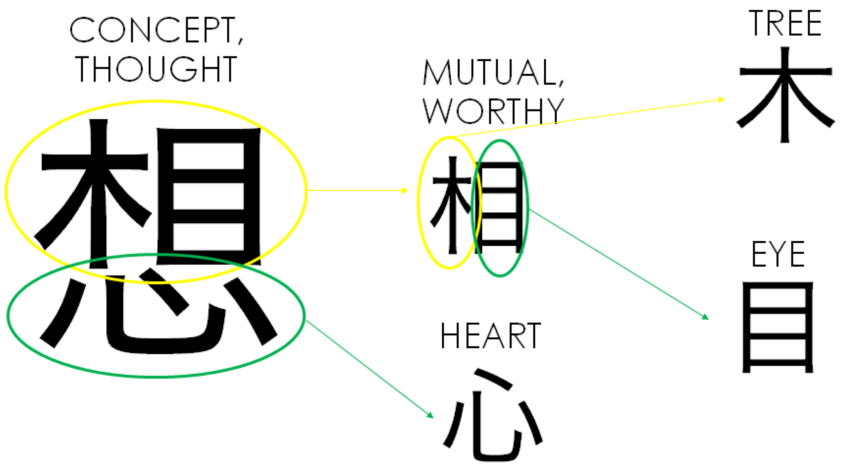


FIGURE 3.3 – Example decomposition of a Kanji in multiple parts, each of them an individual Kanji.

As it can be seen in Figure 3.3, the character 想 can be separated initially into two parts, 相 and 心, where each of those two are also Jouyou Kanji. The first part can be further decomposed into two other parts: 木 and 目. In this case, solely through the decomposition of one Kanji (想) we were able to create implicit connections to four different Kanji, two on a first level and two after one additional level (decomposition of 相). Furthermore, the Kanji for *heart*, 心, is a component of 49 other Kanji, and the Kanji for *mutual*, 相, two other. In that manner, although the Kanji for *heart* does not create a very tight group, since it associates with many other Kanji, the character for *mutual*

associates with only three Kanji, creating a small and coherent group that would do well to be studied together. Figure 3.4 presents this coherent group.

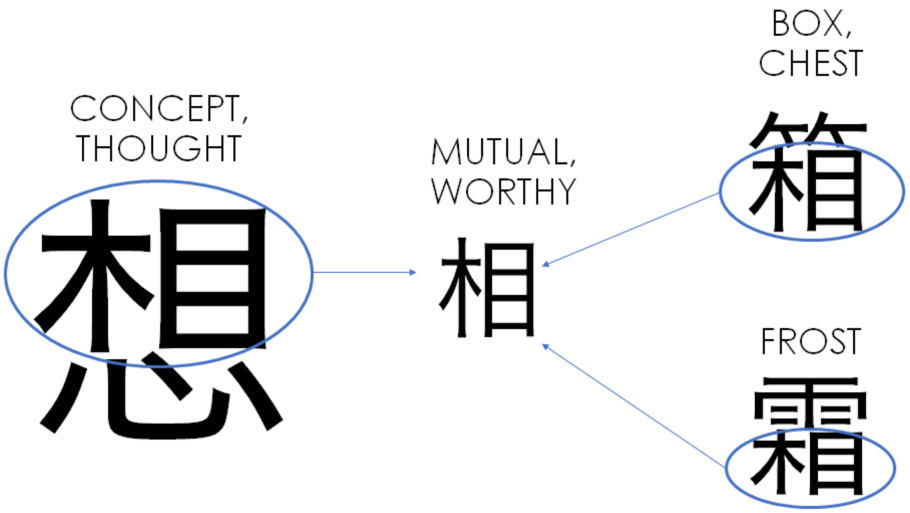


FIGURE 3.4 – Example of multiple Kanji that share the same radical.

Most Kanji are composed of two parts, though there are Kanji composed of three, four, five or six parts. Additionally, 110 of the Jouyou Kanji are "pure", presenting no components. The break up of the number of components of Jouyou Kanji is depicted in Figure 3.5.

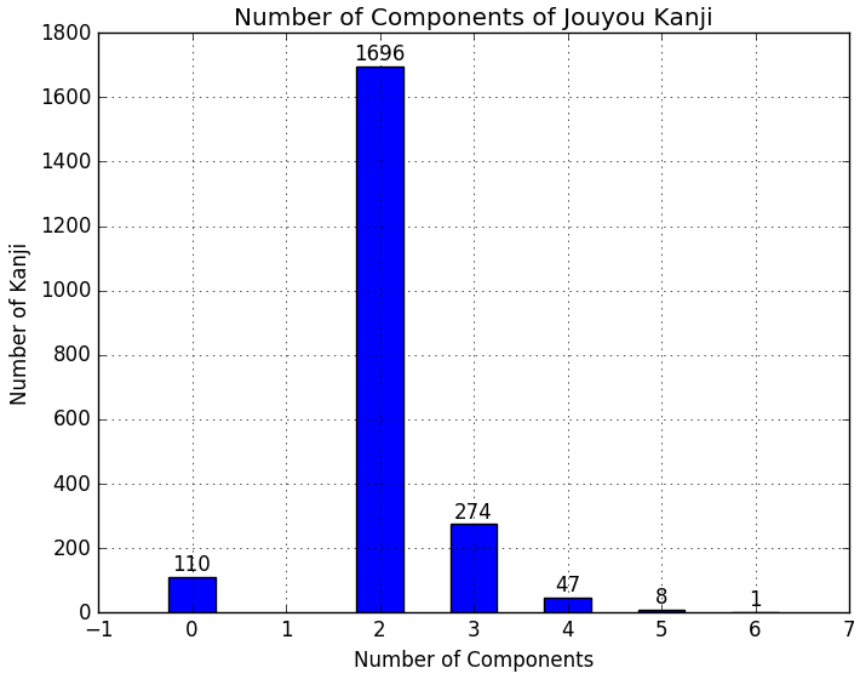


FIGURE 3.5 – Number of Components in Jouyou Kanji

Furthermore, we can break each Kanji in its minimum parts, to count the number of smallest elements it is composed of. In the case of 想, the Kanji used in Figure 3.3, it has

2 components (相 and 心), but three minimum parts (木, 目 and 心). The break up of the number of minimum components of Jouyou Kanji is presented in Figure 3.6.

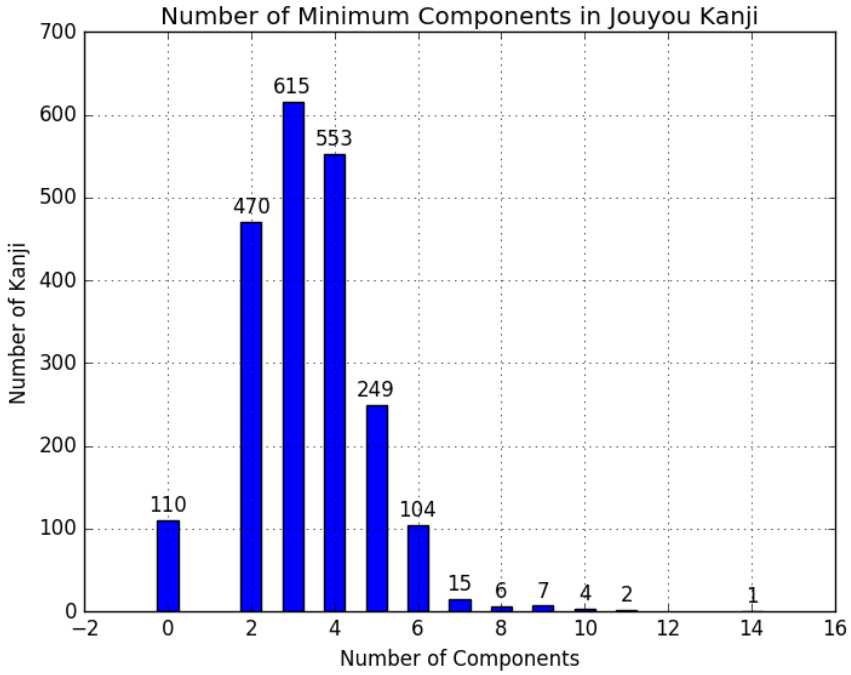


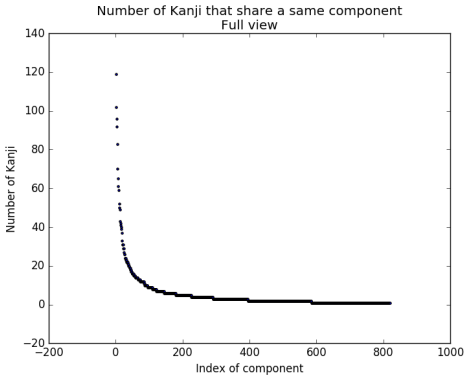
FIGURE 3.6 – Number of Minimum Components in Jouyou Kanji

The usage of components presents approximately the same behaviour as Kanji themselves, where only a few represents most of the decomposition elements. In the morphological graph proposed herein, we identify 836 components, where 484 of them are Jōyō Kanji themselves and 352 are a mixture of variations of Kanji, uncommon Kanji and pure radicals, which for the sake of simplicity will be simply called as *radicals* hereafter. Figure 3.7 illustrates the distribution of use of these components.

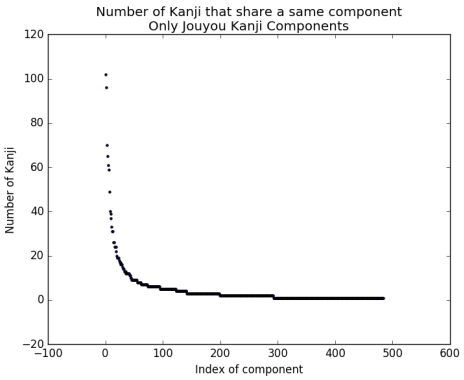
3.2.1.1 Creation of the Morphological Graph

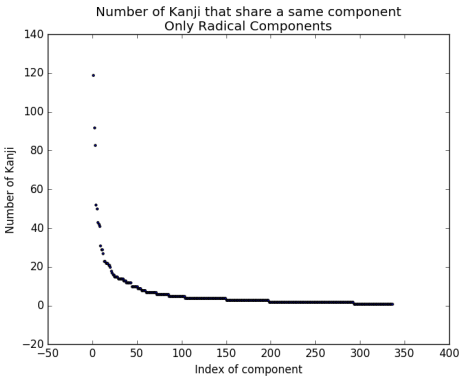
Although research groups have proposed different forms of decompositions of Kanji, most projects present significant issues. For example, the methodology used by Wani-Kanji(TOFUGU, 2016) mostly decomposes characters in its smallest components, ignoring the intermediate Kanji that could be used to form groups. One such example is the character 想, which is decomposed as 木, 目 and 心, missing the opportunity to relate 相 with the coherent group present in Figure 3.4. Other issues include ignoring characters that have no easy representation in unicode and separating characters according to the Chinese writing of characters[3]. For this reason, the morphological connections used

[3]For example, though both characters represent the same idea, Chinese uses the form 诚 while Japanese uses the form 誠 for the concept of "honesty". It is notable that the spelling of these characters is very different, and it is not rare that the decomposed format is taken from Chinese.



(a) Distribution of use of all components.





(b) Distribution of use of Kanji components. (c) Distribution of use of radical components.

FIGURE 3.7 – Distribution of use of components.

by this project were created by hand, using as reference Wikitionary's pages of Kanji definition(WIKITIONARY, 2016) and Henshall's book *A guide to Remembering Japanese Characters*(HENSHALL, 1988). The decomposition was extensively validated so that the number of radicals remain small, and at the same time assuring that each of the additional radicals should be useful, connecting to at least one other Kanji or radical.

3.2.1.2 Special types of radicals

To better represent decompositions, two novel concepts were used for the representation of data.

The first one is the concept of using variants, and relating them to the original characters. For example, in the left side of 飢 or 飼 we see a different form of 食, with one less stroke in the lower right part of it. Another different form of this same Kanji is found in the left part of 餌 and 餅. On some cases, there exists a character to denote this variant in Unicode, in others there are no characters for it and it is denoted by adding an asterisk to a close example. One such case is the variant *生, which is part of 青. Though the recognition of variants is not novel by itself, we took special care to link each variant to its

origin character and separating the use of the original character to its variants, creating better clustered groups.

Another differentiation of this project is the recognition of compositions as radicals. One such example is the composition ヨ一又, which although has a specific meaning associated with it ("A hand holding another hand") and is a part of 浸, 侵 and 寝, has no character to represent it. On most Kanji decomposition projects, this composition is not recognized. Since 又 is present in 19 Kanji and 一 21 Kanji, the small group formed by the three characters 浸, 侵 and 寝 would not be so clearly related without the consideration of this component group as a radical itself.

3.2.1.3 Practical implementation

In practice, each Kanji was defined in JSON (Javascript Object Notation), with a number of different fields. Part of the edition of the information for this character was done by editing by hand the .json file, while other information was written using Python. One such line can be seen in Figure 3.8.

{"k": "右", "c": ["ナ", "口"], "l": ["石"], "s": ["□"], "on": ["う", "ゆう"],
"kun": {"みぎ": []}, "nanori": ["あき", "すけ"], "m": "right (direction)",
"extra_m": "right", "strokes": "5", "grade": "1", "short_m": "right
(direction)", "ic": ["若"], "r": "A rock to the right (右), a constructon (工) to
the left", "ic+": ["匿", "諾"], "cc": []}

FIGURE 3.8 – An example JSON line that represents a Kanji.

In this figure, we have a number of different fields:

- k: The official character that represents this Kanji.
- c: The components that make up this Kanji.
- l: Other character that are visually similar to this Kanji. This field is fed by hand, but its symmetry is maintained automatically (If B is similar to A, A should also be similar to B).
- s: The Ideographic Description Sequences that link these Kanji. In this example, it represents that the composition is made by the first component enveloping on the top left the second element to form 右.
- ic: A list of characters where this Kanji is contained in. Created automatically by the inversion of the "c" (contains) field.

- c+: If the components of this Kanji can be further decomposed, this field lists the smallest components this Kanji can be written in. Created automatically through a recursive break down of the "c" field.

- c+: If the components of this Kanji can be further decomposed, this field lists the Ideographic Description Sequences between components to form the Kanji.

- ic+: A list of characters that contain this character, though not directly. In Figure 3.8 we note that 右 is directly contained by 若 and indirectly contained by 匿 and 諾. Generated automatically through a recursive composition of the "ic" field.

- cc: A list of characters that are components of this character, though not directly and are not the minimum components. For example, 諾 is composed of 言 and 若. Though 右 is not a minimum component of 諾, it is an intermediary component, thus listed under cc.

- alt: A list of alternative interpretations of the composition of Kanji. For example, the Kanji for chapter, 章 can be decomposed in two different forms. It can be seen as 立 on top of 早 or 音 on top of 十. The first interpretation is listed in the "c" field, while the second is listed in the "alt" field.

- o/v: Denotes the historical origin of a radical character and the proper variation it represents today. For example, the Kanji 月 means "moon", but it is in many cases used as a variation of "meat", 肉, for example in "skin" 肌. For this case we created a variation character, *月, to be used when the radical 月 means "meat", not "moon". In this case, the origin for this character is the character for "meat" and the variation is marked as "moon".

- on: A list of on'yomi readings.

- kun: A map of kun readings, where the keys are possible readings for the Kanji itself while the values are lists of possible stems for that reading. For example, one of the readings of character 食 is "ku", and in those cases it is written as 食う, "kuu" or 食らう, "kurau". These cases are written as a key-value pair as ku: [u, rau] (in Hiragana).

- nanori: A list of naming readings.

- m: A small string of meanings.

- extra_m: Additional meanings.

- short_m: The unique meaning that best defines this Kanji for English.

- grade: The grade in which this character is is taught for Japanese children.

- r: A mnemonic phrase.

Though this large number of fields will be useful in future steps of this project, for the development of the morphological graph we will only use the fields for "contains" (c), "is contained" (ic), "looks like" (l), "alternative interpretations" (alt) and "variation" (v). All connections will be bi-directional. All the other listed connections, such as "recursive contains" (c+), "recursive is contained" (ic+) and "sub-component" (cc) can be reached with hops on the previously listed connections, so there is no need to add them again.

This graph will not exclusively contain Kanji, but will also include the 352 radical components, which will bridge connections between Kanji. Also, there is a wild-card element '?', which will not be used as a node in the graph, since it would connect dissimilar Kanji.[4]

To fill in the stochastic matrix of relations, we will initially create a hash map of Japanese characters (Kanji or radical) to a map that relates each related character to the number of times it is linked.[5] This can be achieved by the following python code:

```
from collections import defaultdict
import itertools
import toolbox  # A toolbox of utilities.
from structures import IP, IS  # Important Paths and Structures.

def create_morphological_graph():
    # Set up the base structure for the relations dictionary
    morpho_dict = {k: defaultdict(int)
                   for k in (IS.jk_set | IS.rads_set - {'?'})}
    for json_line in itertools.chain(IS.jk, IS.rads):
        current_char = json_line['k']
        # Skip wildcards
        if current_char == '?':
            continue
        # Add the variation field bi-directionally
        if 'v' in json_line and json_line['v']:
            morpho_dict[current_char][json_line['v']] += 1
            morpho_dict[json_line['v']][current_char] += 1
```

[4]The wild-card element serves when only a part of the Kanji can be decomposed in parts that create new relations and are present or close to characters in the unicode table of characters. For example, the Kanji for Horse, 馬, clearly contains the radical 灬, but the upper right part of it is unique and makes no relation to any other Kanji present in the Jouyou Kanji list.

[5]For example, the character 歌, which is composed of a doubling of 可 and a single 欠 is represented as: 歌: {欠: 1, 可: 2}

```
            # Add fields for contains, is contained, looks like and
            # alternative contains.
            for other_node in itertools.chain(json_line['c'],
                                              json_line['ic'],
                                              json_line['l'],
                                              json_line.get('alt', [])):
                if other_node != '?':
                    morpho_dict[current_char][other_node] += 1
    # Order Kanji by their frequency
    kfg = toolbox.load_data('../data/kanji_freq_grade_novels.csv')
    Kanji, Freq, Grade = zip(*kfg)
    id_to_char = [k for k in Kanji]
    # Also add radicals, in the order they appear on file
    id_to_char.extend([rad['k'] for rad in IS.rads if rad['k'] != '?'])
    char_to_id = {c: i for i, c in enumerate(id_to_char)}
    # Create a non-homogeneous jump vector, that only jumps back to Kanji
    freqs = np.zeros(len(morpho_dict))
    freqs[:len(Freq)] = np.array(Freq)
    return morpho_dict, id_to_char, char_to_id, freqs
```

This form of representation is appropriate since the graph is very sparse. It contains 10,542 connections, whereas the total possible would be $(2136 + 352)^2 = 6,190,144$, that is, only about 0.17% of the fields of the stochastic matrix will be non-zero. The number of connections per character is plotted in Figure 3.9.

Only one Kanji/radical is not related to any other: 飛, the Kanji for "fly".

3.2.1.4 Creation of the Relationships Stochastic Matrix

From this point, all we need to do is to use the morphological dictionary created earlier to create a relationship matrix. To do so, we will transform the form *character* → *map_of_characters* to three lists: *pointing_char_ids* (the column coordinate), *pointed_char_ids* (the row coordinate) and *values*, so that we can create a numpy sparse matrix. The code used to perform this task is transcribed below:

```
import numpy as np
from scipy.sparse import csr_matrix
from collections import defaultdict

def build_sparse_matrix(elem_dict, char_to_id, filler=None, fill_deadends=True):
```

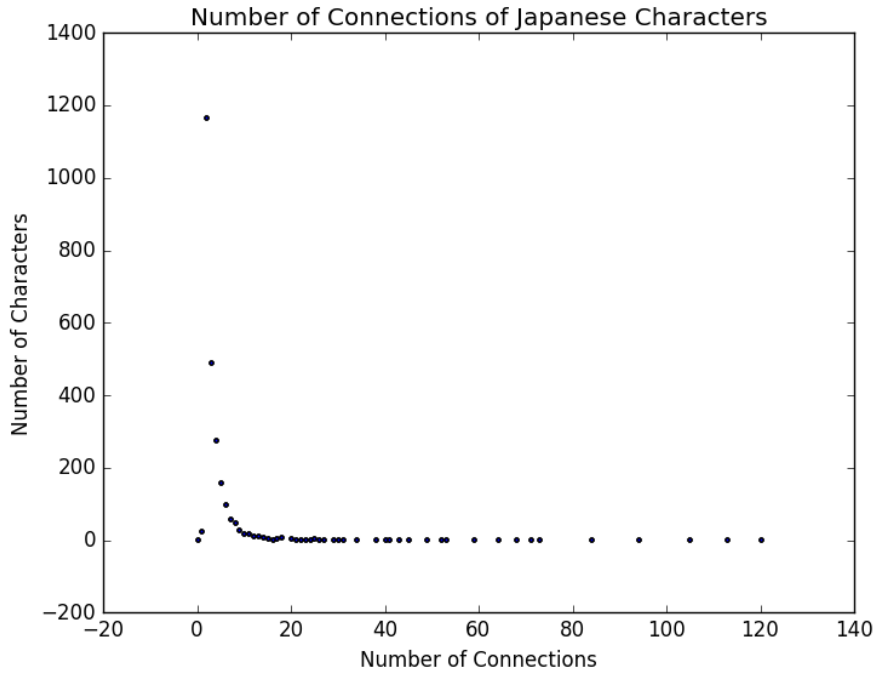


FIGURE 3.9 – A scatter plot of the number of connections of Japanese chracters

```
# A dictionary that maps (id1, id2) -> probability
coord_to_val = dict()
# If no filling array is given, adopt a homogeneous one.
filler = (1/len(elem_dict))*np.ones(len(elem_dict)) if filler is None else fill
filler = filler.reshape(len(elem_dict), 1)
empty_cols = set()
for kanji, relations_dict in elem_dict.items():
    # Keep dead-ends separated, so they can be filled later.
    if not relations_dict:
        empty_cols.add(char_to_id[kanji])
    total = sum(relations_dict.values())
    for kanji2, count in relations_dict.items():
        # Scale each probability by the relative weight of the
        # number of connections.
        coord_to_val[(char_to_id[kanji2], char_to_id[kanji])] = count/total
coords, values = zip(*coord_to_val.items())
i_index, j_index = zip(*coords)
if empty_cols and fill_deadends:
    M = csr_matrix((values, (i_index, j_index)),
                   shape=(len(elem_dict), len(elem_dict)))
    M = M.tolil()
```

```
        for row in empty_cols:
            M[:, row] += filler
        return M.tocsr(), empty_cols
    else:
        return csr_matrix((values, (i_index, j_index)),
                          shape=(len(elem_dict), len(elem_dict))), empty_cols
```

The sparsity pattern of the matrix generated by the morphological graph can be visualized in Figure 3.10. In this graph, black points represent a relationship from the element of that column to the element of that row. The relative size of points is exaggerated, so that the pattern can be seen (only 0.17% of the relations exist). Furthermore, the dashed lines represent the division between the Jouyou Kanji and radicals.

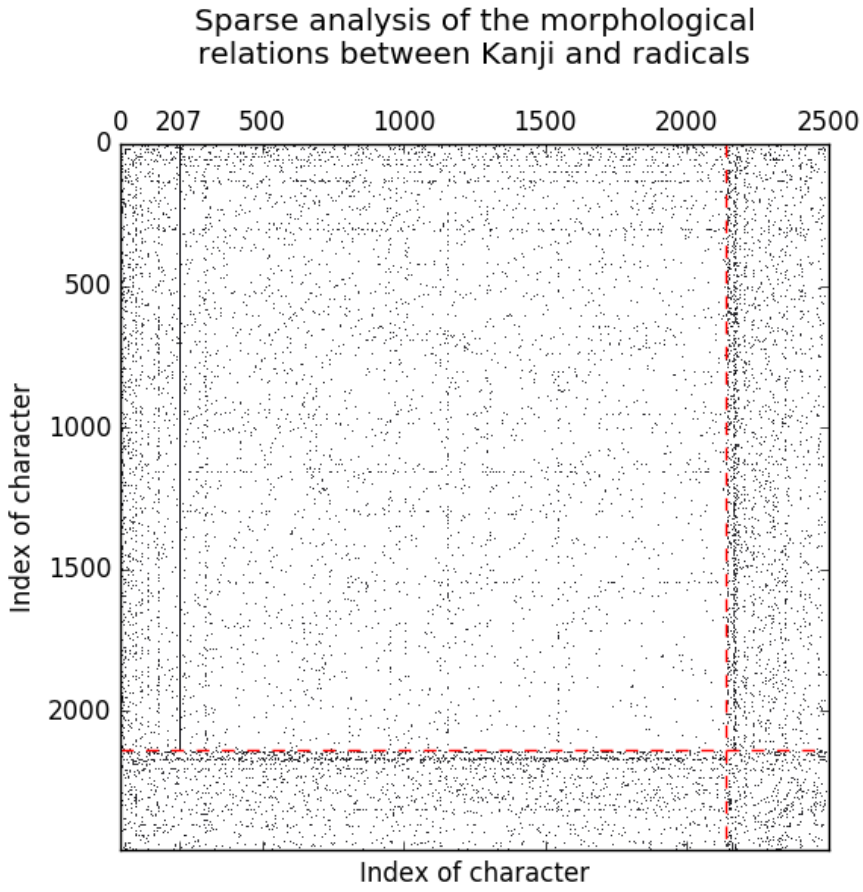


FIGURE 3.10 – Sparsity pattern of the stochastic relations matrix of the morphological graph

Some points of interest that can be noted in this matrix:

- The relatively bigger density of use of radicals in relation to the use of Kanji characters.

- The first few characters also have a bigger density of relations, indicating that they are not only common as Kanji, but they are also common as components for other Kanji.
- The black solid line at index 207: the Kanji 飛, that has no connections with other Kanji, being thus a dead-end, which is filled for the use in the PageRank algorithm.
- This matrix is almost symmetric, except for the black line at index 207. This is from construction, since all links are bidirectional.

3.2.1.5 Application of PageRank

Finally, we can use the sparse matrix and the non-homogeneous jump vector to apply PageRank, as described in Section 3.1.4.5. The following code implement the proposed power iteration.

```
import numpy as np

def page_rank(sparse_matrix, difference_threshold=1e-12, beta=0.85,
              jump=None):
    # Set the initial difference to infinity.
    curr_dif = np.inf
    size = sparse_matrix.shape[0]
    # If no jump vector is given, adopt a homogeneous one.
    jump = np.ones(size)/size if jump is None else jump
    # Start assuming all importances are equal
    importance = np.ones(size)/size
    start = time.clock()
    while curr_dif > difference_threshold:
        new_importance = beta * sparse_matrix * importance + (1 - beta) * jump
        curr_dif = abs(new_importance - importance).sum()
        importance = new_importance
    print('Total elapsed time: %g seconds' % (time.clock() - start))
    return importance
```

The result of applying this algorithm to the morphological problem is a vector of 2,488 ranks. By construction, the order of the first 2,136 points follow the prevalence of those characters in the Japanese language, and the subsequent radicals are numbered according to the order they appear in the JSON file where they were defined. Figure 3.11 depicts the distribution of ranks according to the rank of the character, and the red dashed line represents the division between Jouyou Kanji and radicals.

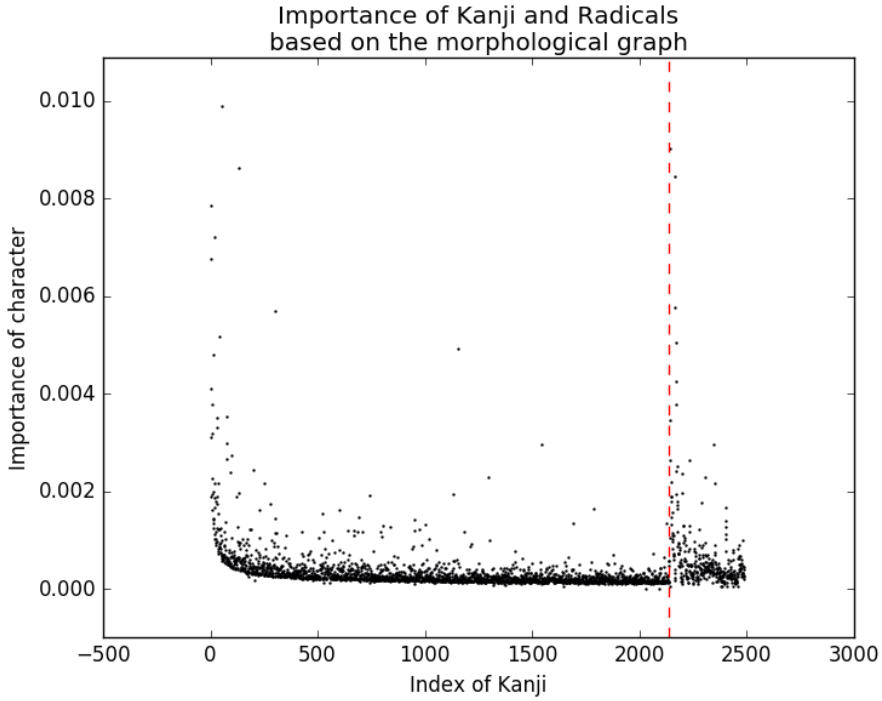


FIGURE 3.11 – Importance of Kanji and Radicals according to the Morphological Graph

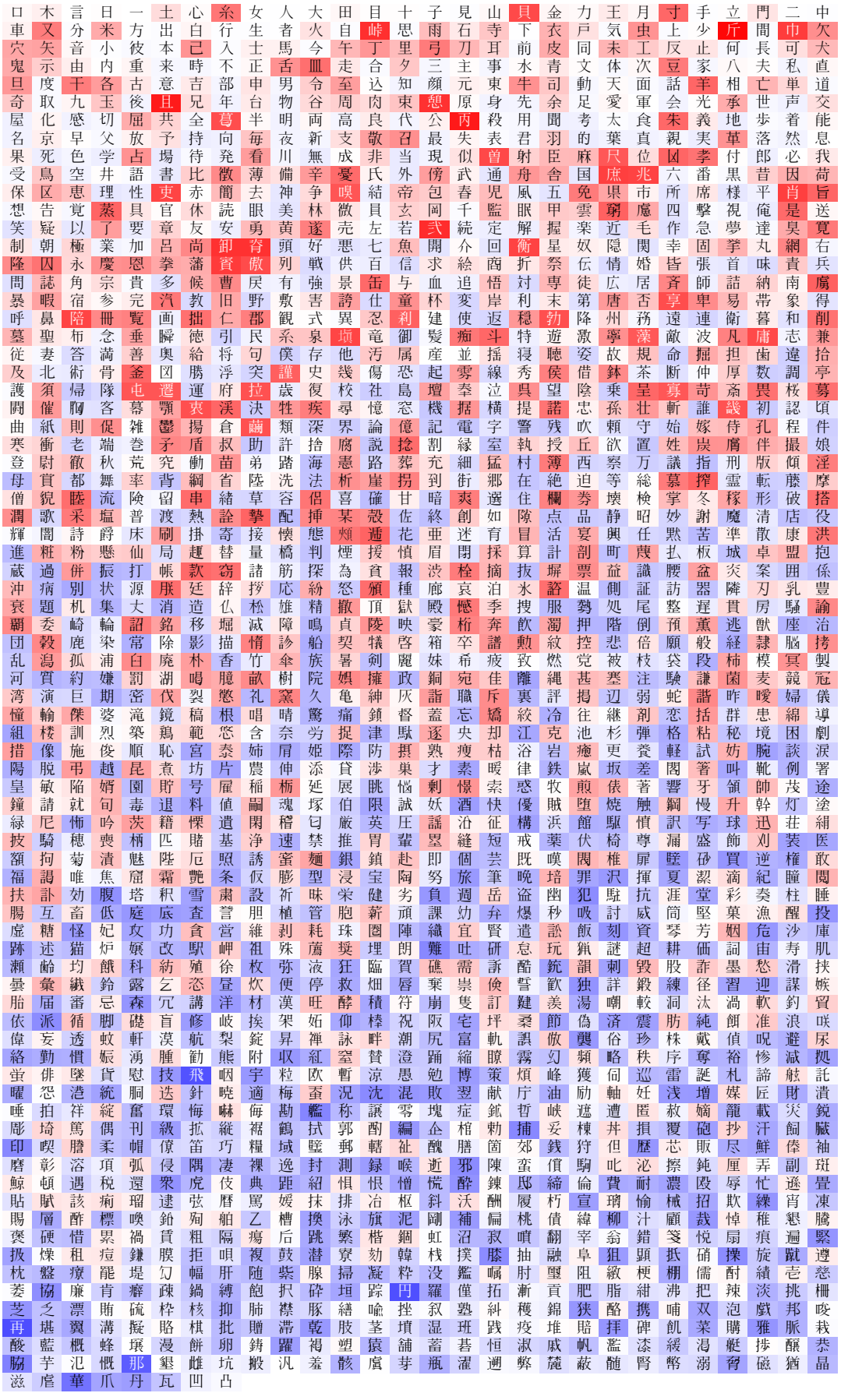
Referring to the picture, we note that the rough outline of the frequency distribution can still be seen, which was to be expected since we used a non-homogeneous jump vector that gave preference to characters according to their relative frequencies. However, it is notable that several points escape the norm and appear separated from the main line. Furthermore, we note that radicals themselves have a rank which follows a pattern very different by that followed by Kanji.

To better visualize this result, an alternate method is used: first we eliminate the radicals in the rank vector. Then, we sort Kanji from greater importance to lower importance. Lastly, we compare the new position it reached with the old position it was in. To visualize this difference in positions, we present this new order on Table 3.1, and manipulate the colors so that characters that shifted upwards have their backgrounds proportionally tinted red[6] and characters that shifted downwards have their backgrounds proportionally tinted blue[7], while a white background represents no shift in position. Finally, font colors were swapped to white in cases that the background was too dark.

[6]For example, if we found that the least frequent Kanji had the highest rank, it would be placed first on the table with a strong red background. If we found that the least frequent Kanji had a median value of rank it would be written near the middle of the table with a background color midway through red and white.

[7]Similarly as it happened with red, if the most frequent character was found to have the lowest rank among Kanji, it would be written in the last position of the table with a strong blue background

TABLE 3.1 – Color representation of the relative shift in ranks after Morphological PageRank



One of the earliest strong examples is 糸, which represents *thread*. By itself it is not a very important Kanji: it places as the 1,154th more frequent, about the middle of the distribution. However, one of its components is the Kanji 小, which is very frequent (the 45th most frequent). Additionally, the concept of *thread* is a very recurrent topic in the morphological origin of Kanji, and it is a direct component of 61 other Kanji. These connections promote 糸 1,144 positions upward, making it the 10th most important Kanji in this morphological perspective. Interestingly, even though it is a somewhat rare Kanji, 糸 is taught in the first grade, showing that its importance is also recognized by the Japanese Ministry of Education. Another interesting case is 斤: by itself it is one of the least frequently used Kanji, ranking in position 2,127. And although it has no components by itself and is the component of only 11 other Kanji, among those Kanji there are some very important concepts such as *new* (新), *close*[8] (近) and *place*[9] (所). Following the expected behaviour for the PageRank algorithm, where the importance of neighbors is more relevant than the number of incoming links, this Kanji ascend to be the 69th most important Kanji, registering a upward movement of 2,058 positions.

3.2.2 The Co-occurrence Graph

There is still a second source of relationships that can be explored to estimate Kanji importance. We can turn our attention to the words that contain each Kanji, and what other characters it co-occurs with, at which rate.

For example, in the English language we could observe with what other concepts the concept for water appears. Some words that use this concept are: *Hydraulics*, *Aquaphobia* and *Hydrophilic*. We expect the word Hydraulics to occur much more than the other two, meaning that this should indicate a stronger relation between the concept of water and mechanism (*aulics*) than between the concept of water and fear (*phobia*) or water and affinity (*philic*). Whereas in other languages the identification of base concepts might be a daunting task, in Japanese it is a trivial one: concepts are represented by Kanji.

While in the case of the morphological graph the relationships were very sparse, the co-occurrence of Kanji yields a much more interconnected graph, where some nodes connect to more than a thousand other Kanji. The number of co-occurring Kanji can be visualized in Figure 3.12.

[8] As in the opposite of far.
[9] As in location.

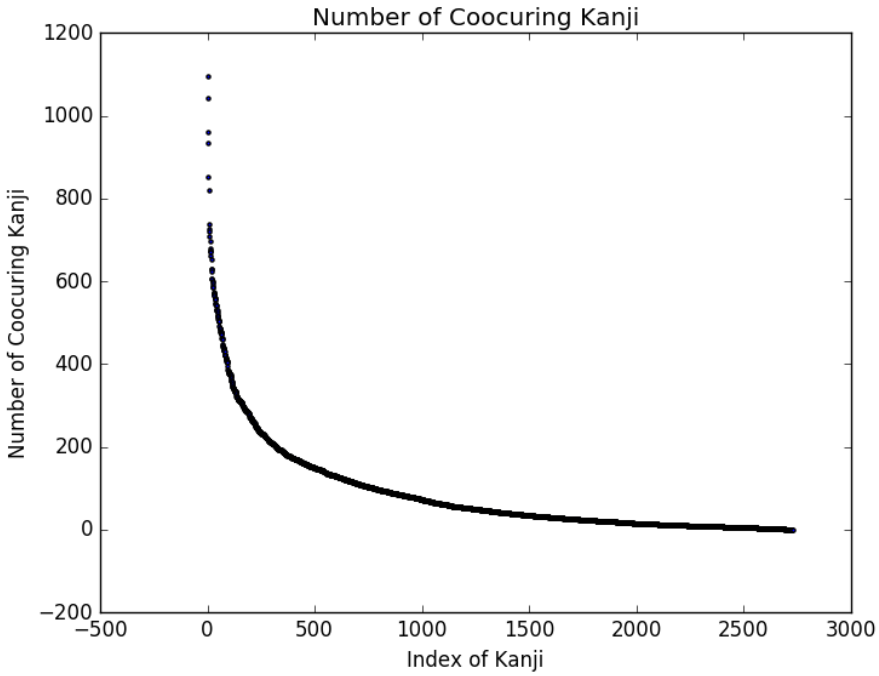


FIGURE 3.12 – Number of relations each Kanji has with other co-occurring Kanji

3.2.3 Creation of the Co-occurrence Graph

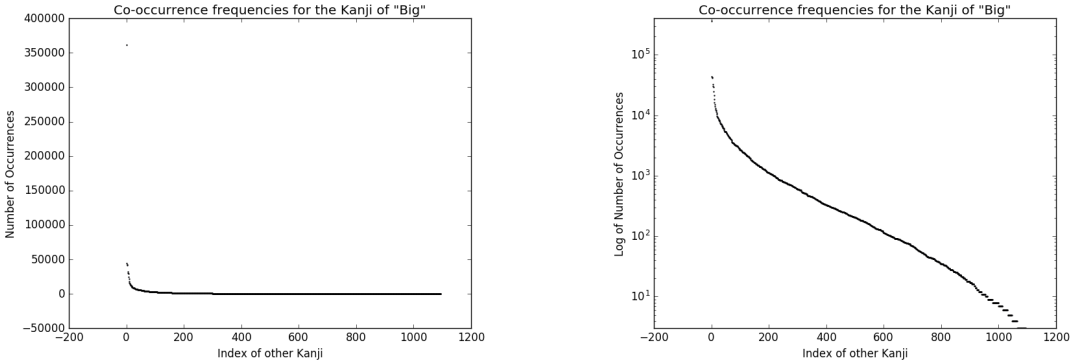
The Co-occurrence graph is much more straightforward to create than the morphological graph, since much fewer assumptions and novel concepts are needed. Essentially, we filter a subset of the word-count pairs of Christopher Brochtrup's analysis of Japanese novels that contain at least one Jouyou Kanji and where the words are composed solely of characters that are either Jouyou Kanji, Jinmeiyou Kanji (Kanji used in names), equivalent forms of Jouyou Kanji, the rogue character 々[10] and Hiragana or Katakana. Afterwards, all that is needed is to create a similar dict as done previously, where each Kanji maps to a hash map where each other Kanji links to the number of times those two occurred together, or if the concept is used by itself we create a link to itself. For example, the five most frequent relations for the Kanji that represents big (大) are:

{"大": {"大": 361592}, {"夫", 43969}, {"学", 42170}, {"丈", 41799}, {"変", 32371} ... }

However, there is an abrupt change between usages: while the first relation of 大, which is with itself, occurs more than 350 thousand times, the second greatest occurs only less than 50 thousand times. When these numbers are used to form a stochastic adjacency matrix, it will be much more probable that this transition will go to the first element than to any other, since it dominates the group. This does not seem similar to what

[10]The character 々 is not properly a Kanji, but it is a character that means "Repeat the last used character". For example, in the word 我々 it takes the role of 我. 我我 is also an accepted form.

a student would do: if the student sees two different usages of the same word, he will think of them with comparable probability, even if the first appears much more frequently than the second. To soften this dominance of terms, we may observe the profile of the curve each co-occurrence Hashmap follows and scale relations to the *log* of the number of times they co-occurred. To illustrate the reason why this approach is more meaningful, we present the full list of co-occurrences of 大 with linear and log scales in Figure 3.13.



(a) Co-occurrences with linear scale. (b) Co-occurrences with log scale.

FIGURE 3.13 – Co-occurrences of “Big” under linear and log scales.

Since the obtained map is in the same format as the one used in Section 3.2.1.3 and the presented codes make no assumptions about the origin of the relations, we will be able to use the code presented in Sections 3.2.1.4 and 3.2.1.5 to create the Stochastic Relationships Matrix and apply the PageRank algorithm in this new graph.

To create this structure we used the code transcribed below:

```
def create_coocurrence_graph(use_log=True):
    count_words = toolbox.load_data(IP.WORDS_TEACHABLE)
    # Kanji that are equivalent to jouyou will be equiparated with jouyou,
    # So they will not be individual nodes in the graph.
    graph_nodes = IS.acceptable_set - IS.equivalents_set
    coocurrence_dict = {k: defaultdict(int) for k in graph_nodes}
    # Take each word and keep only the Kanji characters of it,
    # eliminating Hiragana and Katakana.
    for count, word in count_words:
        kanji_chars = [IS.equiv_to_jouyou.get(k, k)
                       for k in word
                       if k in IS.acceptable_set]
        add_val = np.log(int(count)) if use_log else int(count)
        # If it appears by itself / only accompanied by Hiragana or Katakana,
        # create a link to itself proportional to the number of times this
```

```
            # word occurred.
            if len(kanji_chars) == 1:
                k = kanji_chars[0]
                coocurrence_dict[k][k] += add_val
            else:
                # Otherwise, create a relationship for every possible pair
                # of Kanji.
                for origin, sink in itertools.permutations(kanji_chars, 2):
                    coocurrence_dict[origin][sink] += add_val
    # Clean up Kanji that are not official and don't create relations.
    to_pop = []
    for key, values in coocurrence_dict.items():
        if key not in IS.jk_set and len(values) == 0:
            to_pop.append(key)
    for key in to_pop:
        coocurrence_dict.pop(key, None)
    # Create a mapping following the frequency of Kanji and then
    # inserting the other used characters in a well defined order.
    kfg = toolbox.load_data('../data/kanji_freq_grade_novels.csv')
    Kanji, Freq, Grade = zip(*kfg)
    id_to_char = [k for k in Kanji]
    jinmeiyou = toolbox.load_data(IP.JINMEIYOU_PATH)[0]
    id_to_char.extend([j for j in jinmeiyou if j in coocurrence_dict])
    char_to_id = {c: i for i, c in enumerate(id_to_char)}
    # Create the non-homogeneous jump vector.
    freqs = np.zeros(len(coocurrence_dict))
    freqs[:len(Freq)] = np.array(Freq)
    return coocurrence_dict, id_to_char, char_to_id, freqs
```

The application of this code yields a python dictionary of length 2,727, in which 2,136 are Jouyou Kanji, 590 are Jinmeiyou Kanji and 1 character is the character for repetition 々. As it was expected by the higher number of relations this matrix is much less sparse, containing 240,723 elements of the total possible of $(2136 + 590 + 1)^2 = 7,436,529$ slots. This value represents about 3.2% of the elements, a number 19 timers bigger than the 0.17% density of the morphological matrix.

3.2.3.1 Creation of the Stochastic Relationships Matrix

Applying the same code from Section 3.2.1.4, we obtain the matrix that is presented in Figure 3.14, where each black point represents a connection between the element of that column to the element of that row and the dashed red lines delimit the boundary between Jouyou Kanji and other Kanji.

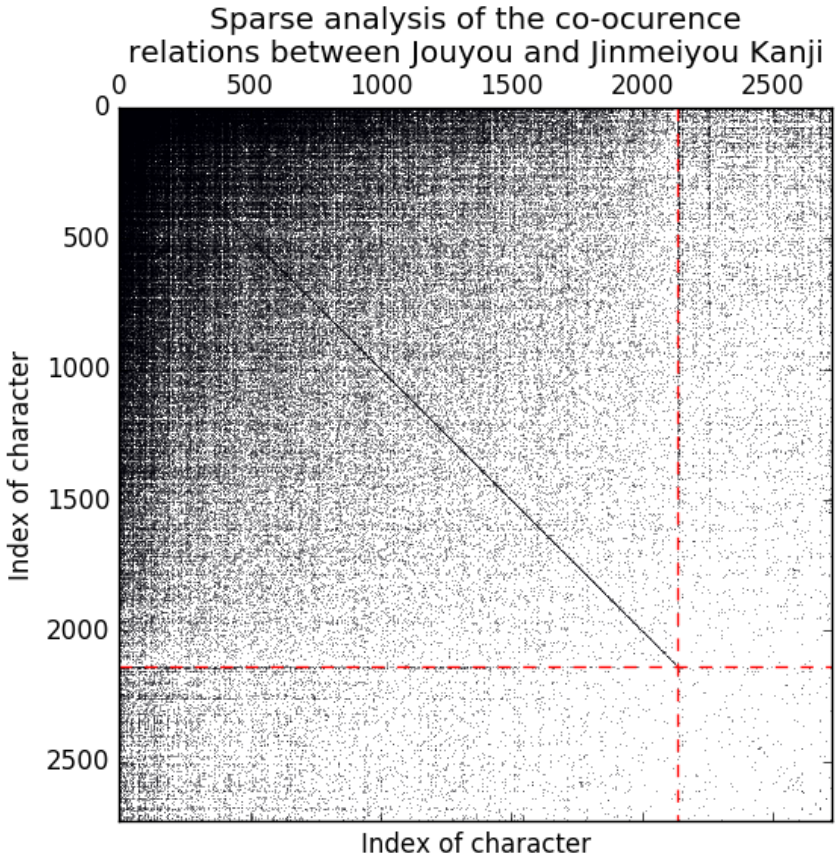


FIGURE 3.14 – Sparsity pattern of the stochastic relations matrix of the co-occurrence graph

Some points of interest that can be noted in this visualization:

- The density in the left superior corner indicates that the most frequent Kanji appear very frequently in composite words that combinate them.
- The black solid line at the diagonal indicate that most Kanji can be found used by themselves.
- The black solid line does not continue after the red dashed lines, since this would mean extra Kanji appearing by themselves. Those were filtered away since our analysis is restricted to words that present at least one Jouyou Kanji.

- This matrix is symmetric. This follows from construction, since all links are bidirectional and no dead ends are present.

3.2.3.2 Application of PageRank

Finally, we can once again use the sparse matrix and the non-homogeneous jump vector and perform PageRank, as described in Section 3.1.4.5, using the same code transcribed in Section 3.2.1.5.

The result of applying this algorithm to the co-occurrences graph is a vector of 2,727 ranks. By construction, the order of the first 2,136 points follow the prevalence of those characters in the Japanese language, and the subsequent Kanji are numbered based on the order they appear in the JSON file where they are defined. Figure 3.15 depicts the distribution of ranks according to the rank of the character, and the red dashed line represents the division between Jouyou Kanji and other Kanji.

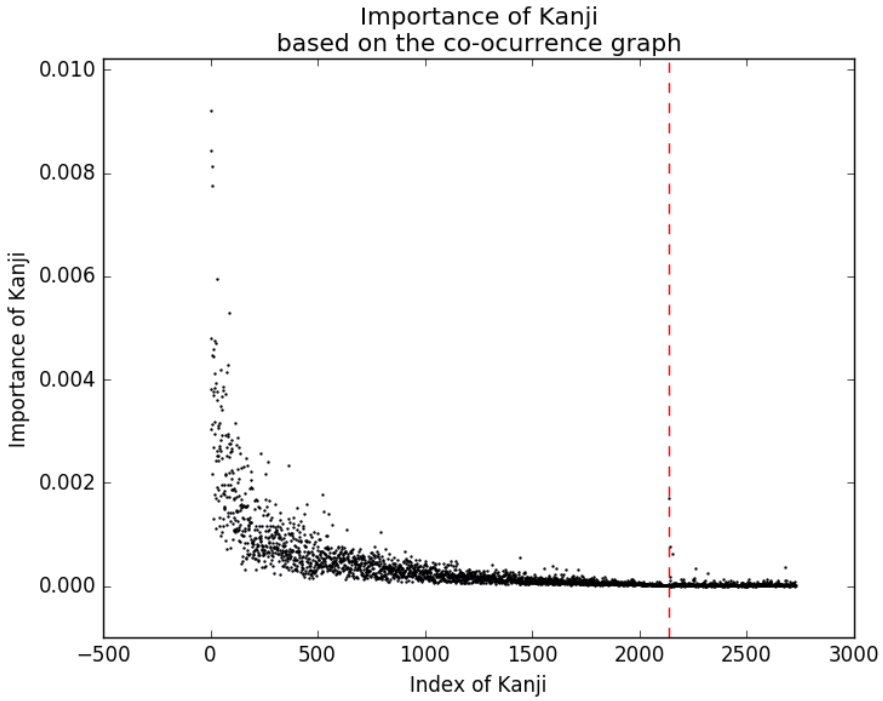


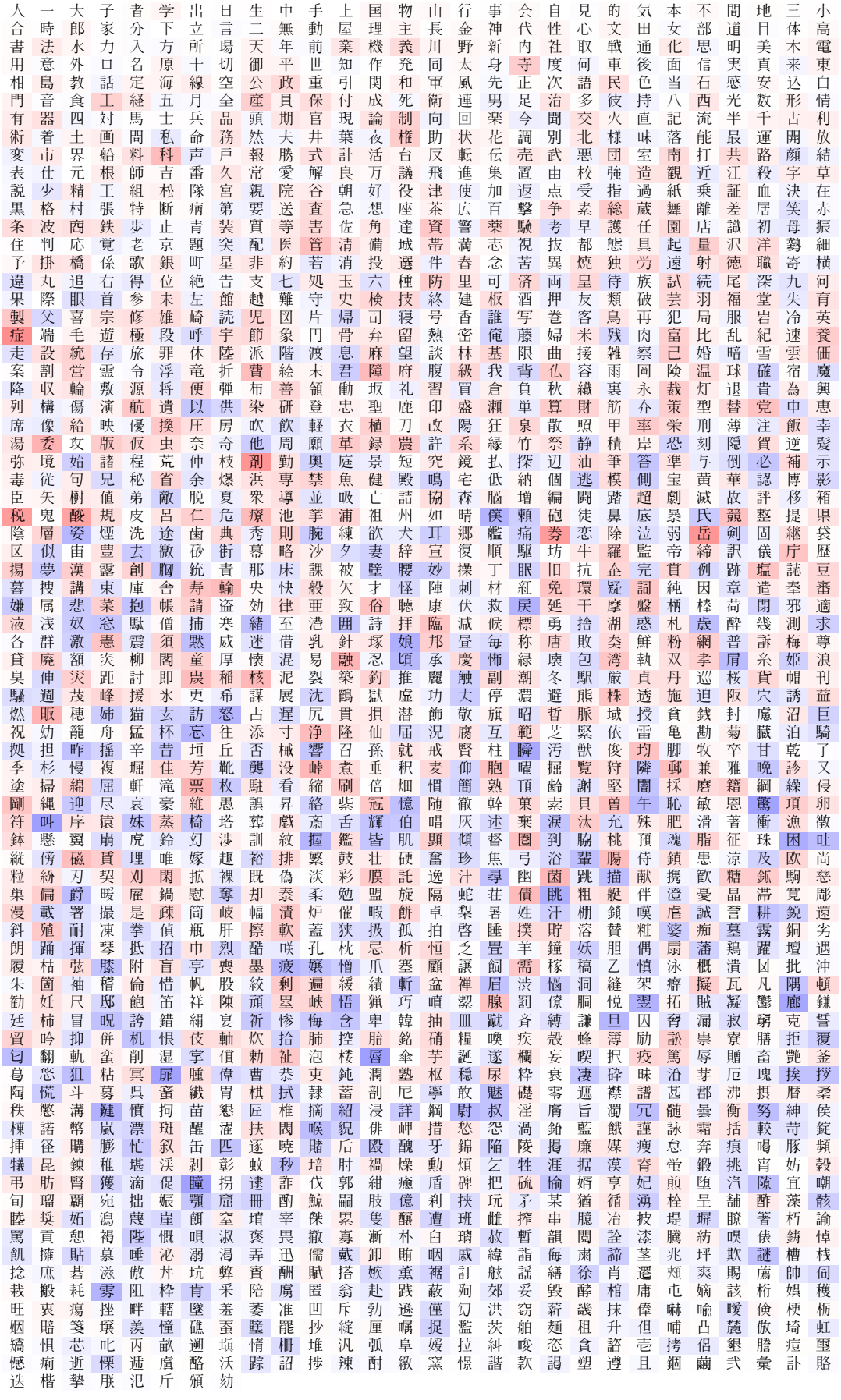
FIGURE 3.15 – Importance of Kanji and Radicals according to the Morphological Graph

Analysis of the figure evidences that the rough outline of the frequency slope is maintained, while some points gain a small shift upward or downward. This is to be expected, since the relations present in the co-occurrence matrix are strongly related to the relative frequency each character displays.

Once again, we use the same style of color table to represent the new order, where clear backgrounds represent small or no shifts in position, red represents upward shifts and blue represent downward shifts. The result is displayed in Table 3.2.

One good example of the shifts caused by this graph is the case with the biggest jump: 症 that means symptom/illness. In terms of frequency it is only the 1,442th most frequent word, but in a rank-wise analysis it shifts to position 505, representing a shift of 937 positions. This shift can be attributed to the fact that 症 is a very versatile Kanji, that appears in conjunction with 193 other Kanji, making it the 341th Kanji with more relations. But, as we noted before, the number of relationships is not decisive to the rank calculated on PageRank. In conjunction with this fact, we note that this Kanji is related to some important concepts such as 状 (condition, the 332th most frequent), 恐 (fear, the 337th most frequent) and 不 (negation, the 71th most frequent).

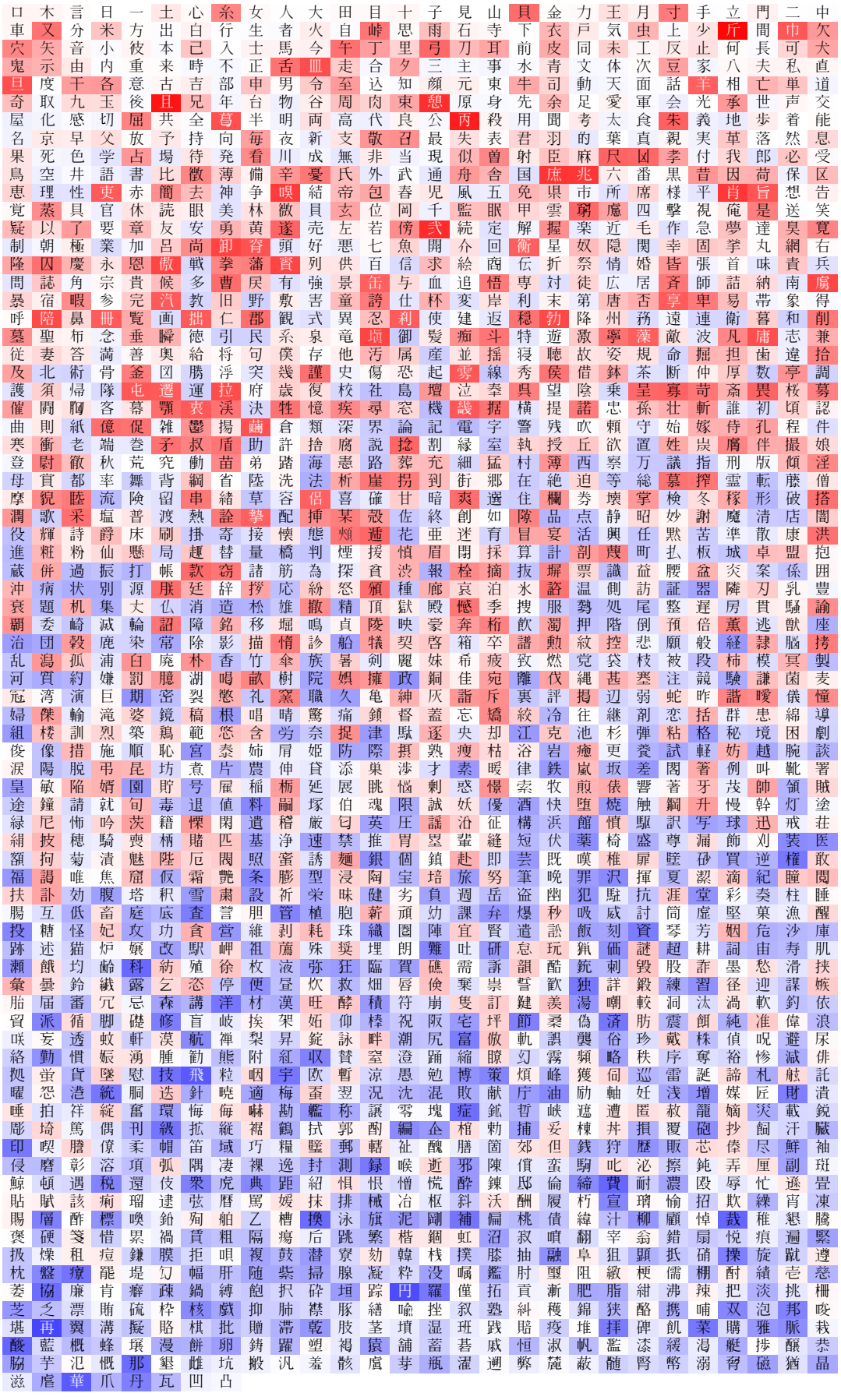
TABLE 3.2 – Color representation of the relative shift in ranks after co-occurrence PageRank using log proportionality



3.2.4 The Morphological Graph

Another interesting comparison that can be made is the relative shifts between elements under different graphs. To do this comparison, we ordered Kanji by their morphological importance and calculated how much they shifted in comparison to the ordering for the co-occurrence graph. As expected, the color table is much similar to the obtained by comparing the morphological importance and the frequency ordering. This result is displayed on Table 3.3, following the same color scheme from the two previous tables.

TABLE 3.3 – Color representation of the relative shift in ranks under morphological and co-occurrence PageRank



3.3 Customizing Results

The results obtained so far for morphological and co-occurrence graphs takes as inputs the natural graph for morphological relations and the frequency estimation to form the co-occurrence graph and feed the jump vector. Since the distribution on the jump vector strongly influences the outcome of the frequency estimation, it is fair to say that the obtained result is particularly meaningful if the objective of the student is reading Japanese light novels or relatively similar media.

Though this might be the case for a large number of students more interested in the casual or cultural use of Japanese, it is not a good approach for students strictly interested in business or academic use of the Japanese language, for example, since in those domains the distribution of Kanji should be radically different. With this in mind, the whole process this far have been made modular, as to not take any premises about the origin of the word-count pairs. That being so, to create a new importance rating for a new domain, say, the content of Japanese Computer Science books, we can create a new word-count pair using Christopher Brochtrup's semantic and statistics analyser and feed that new source to a pipeline that apply all the statistic and graph methods described so far.

4 Applying Statistical Knowledge To Teaching Methods

Using the results from Chapters 2 and 3 we can now propose future directions and techniques that can be applied upon the obtained knowledge to create an improved learning tool, which is the final objective of this project.

4.1 Applications to Problems Related to Leveling Students

The stochastic adjacency models proposed in Chapter 3 are not only useful for the importance estimation of Kanji. That models can be used to generate a Bayesian Inference model (BERNARDO; SMITH, 2001), which represents probabilistic networks. Using this model we can infer what Kanji a user may have learned from other Kanji he has shown to have learned already.

The guideline to the use of this technique in leveling students implies first studying which is the most appropriate stochastic adjacency matrix to be used when estimating the relations that exist in the mind of an unknown student. This matrix can be first estimated as a weighted sum of the morphological and the co-occurrence stochastic adjacency matrices:

$$M_{leveling} = \alpha \cdot M_{Morphological} + (1 - \alpha) \cdot M_{Co-occurrences}$$

where $0 \leq \alpha \leq 1$. Since each matrix is already stochastic, the resulting matrix is also guaranteed to be. Additionally, if $0 < \alpha < 1$, dead-ends will only be needed to be filled by frequency in cases they are dead-ends in both matrices. Otherwise, they may be filled with the value of the full value of the corresponding column on the other matrix, instead of using $alpha$ or $(1 - \alpha)$.

Initially, an arbitrary α can be chosen, like for example 0.7. To research which would be a better value to mimic the thought process of most users, an Online Controlled Experiment such as A/B testing can be used (KOHAVI *et al.*, 2013), assigning different

sensible values for α to different groups of statistical meaningful size and accompanying their progress.

The adjacency matrix can then be used to make Bayesian inferences about the knowledge of a student on multiple different Kanji from other Kanji we expect him to know. Practically, the process would work as making questions to maximize a function of confidence in the knowledge level of the student weighted by the importance of that letter, so that we take care to know with more precision how much he knows about important Kanji, while not taking much attention to how much he knows unimportant Kanji. Doing so, we can explore the graph with increasing levels of confidence, tackling multiple different regions, so as to have knowledge about his learning of a great number of Kanji only from a small number of questions.

4.2 Modifying Spaced Repetition Systems

A number of web platforms dedicated to teaching languages already use learning theory(YELLE, 1979) to create their products, such as Anki(ANKI, 2016) and Duolingo(DUOLINGO, 2016). More specifically, these apps benefit from an algorithm that consolidates the findings of the learn-forget curve(JABER; KHER, 2004) into exercise drills: the Spaced Repetition System – SRS(BATURAY *et al.*, 2009). This algorithm teaches a concept and attempts to refresh this knowledge periodically, with a frequency governed by the learn-forget curve. Concepts that are successfully absorbed are asked with ever increasing intervals, as to refresh that concept just before it has been forgot. Concepts that are more challenging to the student will be asked more frequently.

An issue existent on most SRS algorithms is treating all concepts equally and failing to perceive the relations between those. Classically, SRS presents an ordered list to a student and attempts to assure that the student learns all of these with equal excellency, distributing study time indistinguishably among important and unimportant topics. Using the result from Section 3.2, we may modify those systems, so as to give more emphasis on the learning of important concepts, instead of learning unimportant ones. For example, if a certain student partially remembers a very important Kanji and is about to forget an unimportant Kanji, we may give priority to asking a question about the more important concept, even if he is more liable to forget the unimportant one. This way, we assure that the student is more fluent on things considered to be important. There is not much benefit on applying this technique on domains where the importance of concepts are not much dissimilar, but on fields such as linguistic the frequency and other metrics of importance are bound to be very dissimilar, as we have seen on Figures 3.11 and 3.15.

The other proposed improvement is to use a stochastic adjacency graph to infer re-

lations between concepts, as already proposed in Section 4.1, and use these relations to maximize knowledge on regions of the graph, instead of maximizing knowledge on individual concepts. For example, picture the case of a student learning through a classic SRS algorithm. This student does not yet have the concept of axe (斤) learned. This classic system would insist on trying to make this student remember this specific concept, while an improved system could bring teaching of related Kanji, such as close (近), earlier on the learn or refresh lists. This way we may refresh individual concepts at the same time we teach/refresh related ones, so as to teach more with less exercises.

4.3 Modeling Learning Exercises for Kanji

The proposed improved SRS algorithm is just the scheduler for the sequence concepts that should be taught or reviewed, but the exercises themselves still need to be modeled. Our understanding is that there are two very distinct things a Japanese learner may be interested in studying. The first is learning the meaning individual Kanji may assume; the second is to learn how to read Kanji on different word compositions. For the first task we propose the use of the Morphological Graph to rank importance and multidimensional flashcards to model interaction. For the second task we propose the use of the Co-occurrence Graph to rank importance and reading exercises using sentences to model the user interaction.

4.3.1 Multidimensional Flashcards

Our first focus is to teach individual meanings of Kanji. To do this, we propose that there are two important resources that may be shown: related Kanji (components, Kanji that use this Kanji as components and visually similar Kanji) and words where this Kanji appears. These information can be accessed in a simple card layout as seen on Figure 4.1.

This card will contain, in a compact fashion, information that can help the student learn Kanji through the morphological relations between them, with links to the neighbor Kanji and a compiled list of examples sorted by their frequency. One final proposed field would be a mnemonic field, that could be edited by the user according to its needs. This field is useful since the morphological relations between Kanji are not always clear. For example, the Kanji shown in Figure 4.1 is composed of the radical for movement and the Kanji for neck, and means street or journey. Different users can use different mechanisms to relate those concepts ($movement + neck = street$), and this should be encouraged by letting each user choose a small sentence to be associated with that Kanji.

This cards would be freely accessible through a browse view of the website. Another

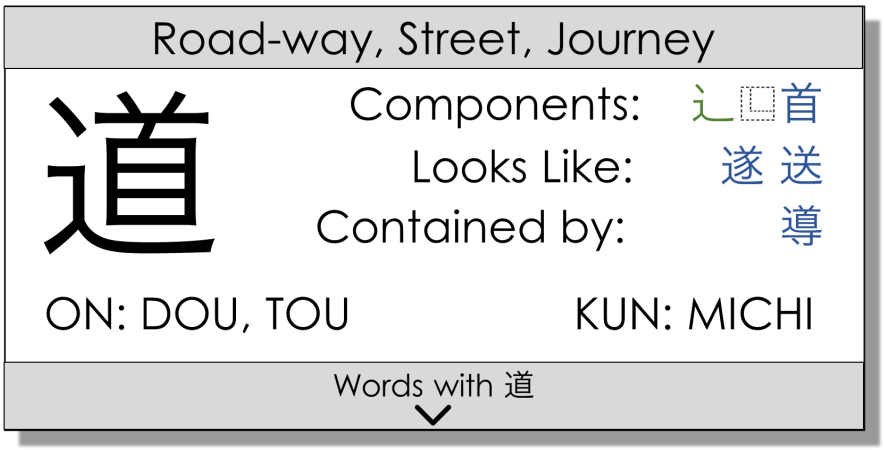


FIGURE 4.1 – Example of a multidimensional browsing card.

view would be the multidimensional flashcards. They are called multidimensional following the property that they will have links to navigate through these learning cards at the back of the card. Figure 4.2 illustrates this concept.

The front side of the card represents what would be seen first: a single Kanji. The implicit question on this exercise is to remember the meanings that Kanji can assume. After probing the knowledge of that Kanji, the user would then see the back side, which would contain a link to the summary card of that Kanji, so that it can study that individual Kanji and related Kanji to create an effective relation in its mind, together with the meaning of that Kanji. The user would then mark how hard or easy was it to remember that concept, and the SRS algorithm would then schedule the refreshment of that card earlier if it was a hard concept, or later if it was considered to be easy.

This interaction model could benefit from gamification(DETERDING *et al.*, 2011), where the user would be encouraged to study more because of game-like mental rewards, but it is important to take care not to shift the incentives of the student from true learning to receiving the gamification mental rewards. Another models exist, like multiple choice or typing exercises, but those are considered to be inadequate and time-wise inefficient by the author of this academic work, since they may be solved in an elimination fashion or fail to assert how hard the student felt the exercise was.

4.3.2 Reading Exercises

The second target of teaching are the different readings Kanji can assume on sentences. On this case we are more interested on teaching the user how to pronounce whole sentences. This is pictured to be a later step, for students that already grasp the meaning of some Kanji – and therefore may be able to infer the meaning of the words composed by them.

(a) Front side of the flashcard.

(b) Back side of the flashcard.

FIGURE 4.2 – Front and back sides of a flashcard for a Kanji.

The proposed exercise will consist of displaying sentences to the student, which will highlight Kanji sequences he is not familiar with the reading. At this point the reading will appear in a syllabary alphabet, such as Hiragana, above the word. An example exercise is presented on Figure 4.3.

FIGURE 4.3 – A mock example of a reading exercise.

The engine that will choose words will also work based on the SRS algorithm proposed on Section 4.2, with the complication that sentences will simultaneously refresh the knowledge of multiple Kanji / reading pairs. The importance measure in this case will be the co-occurrence graph, but the frequency of different reading types will also be ranked in importance. In analogy, when studying the readings that the concept for water may

assume, the readings *Water*, *Aqua* and *Hydro* should also be ranked, in a way as to give more emphasis to teaching readings that are more important. On this case, the relative frequency of readings might be used, since readings do not present a clear a relationship graph model that would enable the use of the tools developed on Section 3.2.

5 Conclusion

This work aggregates statistical insights on the structure of the biggest script of the Japanese language, the Jouyou Kanji. Through the proceedings of this work it is noticeable that there exists ample space for contribution on the task of teaching Japanese Kanji. To delimit the domain of this work, the group of non-native adult speakers was chosen as the main target.

To optimize the learning of these adults, the first approach was to choose a method that would maximize the probability of them knowing a random character taken from the Japanese language. To attain this result, student should learn Kanji in an order consistent with the relative frequency that Kanji appears on the Japanese language on media they are expected to be interested in. Three different word-count projects were reviewed in an attempt to find the set that would best fit these criteria. The chosen project was an extensive word count statistic on Japanese light novels.

Analysis of this light novel data revealed that a student that is ignorant of the best order to learn Kanji as to minimize effort would have to learn significantly more Kanji in the case where it simply follows the Kanji list proposed by the Japanese Ministry of Education to teaching Japanese children, instead of using a sorted list of decreasing importance, pointing to the fact that the use of the list used for Japanese children might not be an efficient resource to foreign adult learners.

To better perceive the importance of Kanji to a student, two different relationship models were proposed. One that perceive as relations between Kanji the component parts they share and their visual resemblance, and another that creates stronger relations between Kanji that co-occurs more frequently in the Japanese language. By applying a PageRank with non-homogeneous restarts proportional to the frequency those Kanji occur in Japanese novels, we were able to create a new importance ordering different than the pure use of frequency. Also, the whole process was made flexible, so that domains other than novels could easily be considered.

Finally, teaching methods that take advantage of the previous contributions were proposed.

The main contributions of this project are the importance orderings for the different

graphs considering the domain of Japanese novels and the stochastic adjacency matrices that may be used as Bayesian inference models in further advancements of this or other projects.

Bibliography

AHLSTRöM, K. **Jisho.org**. 2016. Available from Internet: <http://jisho.org>. Accessed at: 2016-11-01.

ANKI. **Anki**. 2016. Available from Internet: <http://ankisrs.net/>. Accessed at: 2016-11-24.

ARASU, A.; NOVAK, J.; TOMKINS, A.; TOMLIN, J. Pagerank computation and the structure of the web: Experiments and algorithms. In: **Proceedings of the Eleventh International World Wide Web Conference, Poster Track**. [S.l.: s.n.], 2002. p. 107–117.

BANNO, E. **Kanji look and learn: 512 kanji with illustrations and mnemonic hints**. [S.l.]: Japan Times Limited, 2010.

BATURAY, M.; YILDIRIM, S.; DALOĞLU, A. Effects of web-based spaced repetition on vocabulary retention of foreign language learners. **Eurasian Journal of Educational Research (EJER)**, v. 8, n. 34, p. 17–36, 2009.

BERNARDO, J. M.; SMITH, A. F. **Bayesian theory**. [S.l.]: IOP Publishing, 2001.

BREEN, J. A www japanese dictionary. **Japanese Studies**, Taylor & Francis, v. 20, n. 3, p. 313–317, 2000.

BROCHTRUP, C. **Japanese Text Analysis Tool**. 2012. Available from Internet: <https://sourceforge.net/projects/japanesetextana/>. Accessed at: 2016-11-03.

BROCHTRUP, C. **Word Frequency Analysis for Innocent Novels**. 2012. Available from Internet: <http://forum.koohii.com/post-167827.html>. Accessed at: 2016-11-03.

DETERDING, S.; DIXON, D.; KHALED, R.; NACKE, L. From game design elements to gamefulness: defining gamification. In: ACM. **Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments**. [S.l.], 2011. p. 9–15.

DUOLINGO. **Duolingo**. 2016. Available from Internet: <http://www.duolingo.com>. Accessed at: 2016-11-24.

GIRARDI, A. **Alexandre Girardi's word frequency list**. 1998. Available from Internet: <http://ftp.monash.edu.au/pub/nihongo/00INDEX.html>. Accessed at: 2016-11-03.

HAVELIWALA, T. H. Topic-sensitive pagerank. In: ACM. **Proceedings of the 11th international conference on World Wide Web**. [S.l.], 2002. p. 517–526.

HENSHALL, K. G. **A guide to remembering Japanese characters**. [S.l.]: Tuttle Publishing, 1988.

JABER, M. Y.; KHER, H. V. Variant versus invariant time to total forgetting: the learn–forget curve model revisited. **Computers & Industrial Engineering**, Elsevier, v. 46, n. 4, p. 697–705, 2004.

JOYCE, J.; GOYERT, G. **Ulysses**. [S.l.]: Rhein-Verlag, 1926.

KOHAVI, R.; DENG, A.; FRASCA, B.; WALKER, T.; XU, Y.; POHLMANN, N. Online controlled experiments at large scale. In: ACM. **Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.], 2013. p. 1168–1176.

KUDO, T. Mecab: Yet another part-of-speech and morphological analyzer. **http://mecab. sourceforge. net/**, 2005.

MANDELBROT, B. Paretian distributions and income maximization. **The Quarterly Journal of Economics**, JSTOR, p. 57–85, 1962.

MCCAWLEY, J. D. **The phonological component of a grammar of Japanese**. [S.l.]: Mouton, 1968.

NEWMAN, M. E. Power laws, pareto distributions and zipf's law. **Contemporary physics**, Taylor & Francis, v. 46, n. 5, p. 323–351, 2005.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. The pagerank citation ranking: bringing order to the web. Stanford InfoLab, 1999.

PAKES, A. Some conditions for ergodicity and recurrence of markov chains. **Operations Research**, INFORMS, v. 17, n. 6, p. 1058–1061, 1969.

PIANTADOSI, S. T. Zipf's word frequency law in natural language: A critical review and future directions. **Psychonomic bulletin & review**, Springer, v. 21, n. 5, p. 1112–1130, 2014.

TOFUGU. **Wani-Kani**. 2016. Available from Internet: <http://www.wanikani.com>. Accessed at: 2016-11-01.

WIKITIONARY. **Frequency List from 2015 Wikipedia Dump**. 2015. Available from Internet: <https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Japanese2015_10000>. Accessed at: 2016-11-03.

WIKITIONARY. **Description of a Kanji Character**. 2016. Available from Internet: <https://en.wiktionary.org/wiki/\%E5\%8E\%9F>. Accessed at: 2016-11-08.

YELLE, L. E. The learning curve: Historical review and comprehensive survey. **Decision Sciences**, Wiley Online Library, v. 10, n. 2, p. 302–328, 1979.

ZIPF, G. K. The psycho-biology of language. Houghton, Mifflin, 1935.

Appendix A - Linguistic Definitions and Explanations

This annex serves to clarify technical terms and concepts from the field of linguistics that are relevant to the development of this academic project.

A.1 Japanese as a Moraic System

A Mora is a unit in phonology that can be defined as "something of which a long syllable consists of two and a short syllable consists of one"(MCCAWLEY, 1968). As per this definition, the Japanese is said to be moraic, since two of its base scripts (Hiragana and Katakana) translate strongly the concept of mora to its characters.

For example, we can refer to the word Ōsaka, which could also be spelled as Oosaka (with a double "o") or in Hiragana おおさか. This word would be be broken in three syllables: ō/sa/ka, but as denoted by the macron over the "o" letter, this is a long vowel. As can be seen in the Hiragana representation of this word, we have four graphemes, exactly the same as the number of morae. Though this is not always the case in Japanese, it is so in the great majority of cases, representing the importance of morae in the understanding of Japanese.

Appendix B - Japanese Language Details

This annex is dedicated to the explanation of details of the Japanese language that are relevant to the context of this work.

B.1 Romanization

The concept of romanization is the application of Latin letters to transcribe the Japanese language. This form of writing is sometimes referred as rōmaji (ローマ字, literally "roman letter"). There are a number of different romanization methods, but one of them is by far the most widely used: the Revised Hepburn romanization. This system generally follows English phonology to transcribe Japanese in a way that is intuitive to English speakers. Most of the English versions of cities names are derived from slight modifications of this method. Table B.1.

TABLE B.1 – Comparison of city names to its Hepbrun Romanizations

Hiragana	**Hepburn Romanization**	**English Name**
おおさか	Ōsaka	Osaka
とうきょう	Tōkyō	Tokyo
なごや	Nagoya	Nagoya

B.2 Hiragana and Katakana

Hiragana and Katakana are syllabic components of the Japanese writing system, and are referred as Kana scripts. Those two alphabets serve to fulfill syntactical roles in Japanese, as well as to spell some words and as a sort of subtitle of Kanji characters considered to be difficult to read by the reader. Most substantives, root of verbs, adjectives and adverbs are written in Kanji, while the conjugations of these verbs, adjectives and adverbs, indicative particles and some few words are written solely in some Kana script.

Hiragana is the most commonly used script, while Katakana is used on in special cases, such as describing a word alien to Japanese (a borrow word) or to give textual emphasis. Having two forms to write the same grapheme is already something common in the Latin alphabet, which is a Bicameral script[1], but in Japanese this second script is not intermixed within the same word. One analogy that could be made to our writing system would be writing words to be emphasised at all caps. To clarify this concept, let us look at an example.

ネコ が ラ イス を食べている
NEKO ga RAISU wo tabeteiru
the CAT is eating RICE.

In this example, the enlarged characters are in Katakana, while the rest is in Hiragana or Kanji. Although “Neko” (cat) is a Japanese word, it is written in Katakana here to give stress to this word. The second word in Katakana is “Raisu” which is an English borrow word for rice.

In modern Japanese, each of these scripts is formed by 46 unique graphemes, to be added to 25 additional variation graphemes[2] and one consonant voicing mark, っ, the small “tsu”. Both these scripts are displayed in figure B.2.

[1] A type of script that have two forms for every grapheme, one that is said to be the lower case and the other said to be the upper case.

[2] These variations are created through the addition of a voicing mark called the dakuten. For example, we have the pure letter は(ha) that can also be voiced as ば(ba) and ぱ(pa).

	Hiragana					Katakana				
	a	i	u	e	o	a	i	u	e	o
	あ [a]	い [i]	う [u]	え [e]	お [o]	ア [a]	イ [i]	ウ [u]	エ [e]	オ [o]
k	か [ka]	き [ki]	く [ku]	け [ke]	こ [ko]	カ [ka]	キ [ki]	ク [ku]	ケ [ke]	コ [ko]
s	さ [sa]	し [shi]	す [su]	せ [se]	そ [so]	サ [sa]	シ [shi]	ス [su]	セ [se]	ソ [so]
t	た [ta]	ち [chi]	つ [tsu]	て [te]	と [to]	タ [ta]	チ [chi]	ツ [tsu]	テ [te]	ト [to]
n	な [na]	に [ni]	ぬ [nu]	ね [ne]	の [no]	ナ [na]	ニ [ni]	ヌ [nu]	ネ [ne]	ノ [no]
h	は [ha]	ひ [hi]	ふ [fu]	へ [he]	ほ [ho]	ハ [ha]	ヒ [hi]	フ [fu]	ヘ [he]	ホ [ho]
m	ま [ma]	み [mi]	む [mu]	め [me]	も [mo]	マ [ma]	ミ [mi]	ム [mu]	メ [me]	モ [mo]
y	や [ya]		ゆ [yu]		よ [yo]	ヤ [ya]		ユ [yu]		ヨ [yo]
r	ら [ra]	り [ri]	る [ru]	れ [re]	ろ [ro]	ラ [ra]	リ [ri]	ル [ru]	レ [re]	ロ [ro]
w	わ [wa]				を [wo]	ワ [wa]				ヲ [wo]
ń	ん [n]					ン [n]				
g	が [ga]	ぎ [gi]	ぐ [gu]	げ [ge]	ご [go]	ガ [ga]	ギ [gi]	グ [gu]	ゲ [ge]	ゴ [go]
z	ざ [za]	じ [ji]	ず [zu]	ぜ [ze]	ぞ [zo]	ザ [za]	ジ [ji]	ズ [zu]	ゼ [ze]	ゾ [zo]
d	だ [da]	ぢ [ji]	づ [zu]	で [de]	ど [do]	ダ [da]	ヂ [ji]	ヅ [zu]	デ [de]	ド [do]
b	ば [ba]	び [bi]	ぶ [bu]	べ [be]	ぼ [bo]	バ [ba]	ビ [bi]	ブ [bu]	ベ [be]	ボ [bo]
p	ぱ [pa]	ぴ [pi]	ぷ [pu]	ぺ [pe]	ぽ [po]	パ [pa]	ピ [pi]	プ [pu]	ペ [pe]	ポ [po]

FIGURE B.1 – Hiragana and Katakana Scripts

Appendix C - Ordered lists for Kanji Study

This Appendix is dedicated will be dedicated to the listing of the ordered Kanji lists produced throughout this work.

C.1 Order by frequency on Japanese Novels

TABLE C.1 – Kanjis ordered by their frequency on Japanese novels.

1: 人	2: 一	3: 見	4: 言	5: 出	6: 子	7: 大	8: 思
9: 手	10: 分	11: 彼	12: 中	13: 女	14: 気	15: 上	16: 間
17: 日	18: 二	19: 自	20: 行	21: 時	22: 何	23: 事	24: 生
25: 私	26: 前	27: 来	28: 方	29: 本	30: 目	31: 十	32: 三
33: 者	34: 下	35: 年	36: 話	37: 知	38: 入	39: 立	40: 部
41: 家	42: 心	43: 今	44: 男	45: 小	46: 後	47: 顔	48: 長
49: 体	50: 物	51: 合	52: 声	53: 地	54: 屋	55: 場	56: 口
57: 身	58: 聞	59: 的	60: 持	61: 会	62: 意	63: 度	64: 当
65: 同	66: 感	67: 先	68: 動	69: 明	70: 少	71: 不	72: 通
73: 力	74: 考	75: 山	76: 向	77: 実	78: 田	79: 無	80: 国
81: 理	82: 五	83: 死	84: 名	85: 神	86: 所	87: 学	88: 取
89: 面	90: 書	91: 笑	92: 外	93: 代	94: 四	95: 月	96: 道
97: 高	98: 金	99: 美	100: 内	101: 世	102: 頭	103: 戦	104: 全
105: 作	106: 近	107: 様	108: 切	109: 相	110: 足	111: 真	112: 葉
113: 最	114: 郎	115: 情	116: 発	117: 変	118: 夜	119: 食	120: 白
121: 性	122: 然	123: 信	124: 主	125: 味	126: 用	127: 新	128: 音
129: 対	130: 夫	131: 文	132: 木	133: 殺	134: 着	135: 車	136: 引
137: 水	138: 現	139: 込	140: 親	141: 悪	142: 以	143: 色	144: 野
145: 空	146: 教	147: 父	148: 結	149: 君	150: 開	151: 正	152: 連
153: 関	154: 川	155: 仕	156: 八	157: 歩	158: 返	159: 落	160: 問
161: 太	162: 風	163: 僕	164: 帰	165: 次	166: 軍	167: 俺	168: 天

169: 違	170: 重	171: 光	172: 機	173: 母	174: 使	175: 別	176: 数
177: 待	178: 直	179: 命	180: 回	181: 六	182: 多	183: 表	184: 馬
185: 平	186: 社	187: 強	188: 原	189: 兵	190: 電	191: 士	192: 要
193: 御	194: 受	195: 誰	196: 好	197: 指	198: 眼	199: 決	200: 王
201: 起	202: 七	203: 初	204: 海	205: 安	206: 残	207: 調	208: 飛
209: 解	210: 流	211: 東	212: 姿	213: 呼	214: 早	215: 助	216: 定
217: 第	218: 突	219: 乗	220: 村	221: 若	222: 息	223: 九	224: 確
225: 語	226: 石	227: 愛	228: 首	229: 半	230: 界	231: 由	232: 説
233: 置	234: 朝	235: 法	236: 室	237: 黒	238: 警	239: 答	240: 覚
241: 張	242: 記	243: 続	244: 想	245: 両	246: 遠	247: 成	248: 得
249: 門	250: 振	251: 背	252: 必	253: 配	254: 他	255: 島	256: 化
257: 元	258: 運	259: 在	260: 苦	261: 和	262: 楽	263: 付	264: 百
265: 視	266: 終	267: 能	268: 進	269: 業	270: 員	271: 件	272: 線
273: 隊	274: 過	275: 井	276: 番	277: 形	278: 居	279: 供	280: 火
281: 急	282: 衛	283: 打	284: 店	285: 伝	286: 深	287: 走	288: 集
289: 始	290: 点	291: 千	292: 消	293: 失	294: 赤	295: 青	296: 戻
297: 吉	298: 撃	299: 公	300: 横	301: 戸	302: 土	303: 船	304: 反
305: 止	306: 加	307: 花	308: 広	309: 達	310: 放	311: 勝	312: 離
313: 老	314: 古	315: 右	316: 万	317: 追	318: 師	319: 魔	320: 血
321: 台	322: 題	323: 経	324: 常	325: 有	326: 娘	327: 利	328: 校
329: 左	330: 町	331: 読	332: 状	333: 京	334: 寄	335: 断	336: 存
337: 恐	338: 官	339: 奥	340: 座	341: 良	342: 押	343: 紙	344: 申
345: 階	346: 態	347: 転	348: 抜	349: 病	350: 果	351: 役	352: 判
353: 可	354: 側	355: 絶	356: 浮	357: 逃	358: 画	359: 藤	360: 寝
361: 守	362: 城	363: 活	364: 氏	365: 義	366: 西	367: 佐	368: 客
369: 武	370: 段	371: 品	372: 告	373: 鳴	374: 報	375: 友	376: 勢
377: 組	378: 係	379: 応	380: 胸	381: 計	382: 住	383: 路	384: 治
385: 望	386: 貴	387: 議	388: 期	389: 術	390: 特	391: 腕	392: 服
393: 暗	394: 飲	395: 敵	396: 頼	397: 江	398: 渡	399: 送	400: 都
401: 民	402: 細	403: 静	404: 察	405: 去	406: 北	407: 松	408: 論
409: 質	410: 念	411: 識	412: 片	413: 族	414: 市	415: 根	416: 満
417: 将	418: 降	419: 器	420: 宮	421: 抱	422: 草	423: 兄	424: 幸
425: 耳	426: 介	427: 我	428: 犯	429: 頃	430: 酒	431: 影	432: 緒
433: 交	434: 務	435: 保	436: 瞬	437: 格	438: 売	439: 志	440: 字
441: 冷	442: 茶	443: 歳	444: 夢	445: 素	446: 院	447: 政	448: 際
449: 倒	450: 腹	451: 精	452: 予	453: 買	454: 黙	455: 異	456: 肉
457: 熱	458: 妙	459: 谷	460: 叫	461: 髪	462: 約	463: 肩	464: 団
465: 痛	466: 単	467: 隠	468: 礼	469: 久	470: 歌	471: 怒	472: 腰
473: 窓	474: 婚	475: 驚	476: 難	477: 差	478: 妻	479: 支	480: 囲

481: 料 482: 証 483: 嫌 484: 似 485: 認 486: 悲 487: 像 488: 津
489: 壁 490: 容 491: 裏 492: 構 493: 剣 494: 建 495: 破 496: 忘
497: 式 498: 軽 499: 蔵 500: 探 501: 任 502: 香 503: 限 504: 造
505: 注 506: 示 507: 旅 508: 観 509: 席 510: 参 511: 陽 512: 沢
513: 種 514: 具 515: 敷 516: 余 517: 端 518: 奴 519: 負 520: 疑
521: 眠 522: 備 523: 寺 524: 許 525: 願 526: 工 527: 円 528: 談
529: 婦 530: 投 531: 位 532: 令 533: 傷 534: 丸 535: 等 536: 装
537: 庭 538: 共 539: 仲 540: 越 541: 号 542: 舞 543: 南 544: 波
545: 並 546: 微 547: 角 548: 類 549: 周 550: 産 551: 喜 552: 護
553: 星 554: 途 555: 吹 556: 泣 557: 巻 558: 再 559: 景 560: 刻
561: 与 562: 乱 563: 鉄 564: 密 565: 殿 566: 優 567: 秋 568: 夕
569: 制 570: 銀 571: 霊 572: 例 573: 興 574: 春 575: 局 576: 図
577: 案 578: 奇 579: 危 580: 底 581: 選 582: 医 583: 怖 584: 坂
585: 接 586: 退 587: 街 588: 求 589: 映 590: 刑 591: 閉 592: 写
593: 尾 594: 争 595: 商 596: 河 597: 辺 598: 怪 599: 激 600: 雨
601: 完 602: 処 603: 司 604: 弾 605: 派 606: 折 607: 非 608: 遊
609: 橋 610: 雪 611: 査 612: 故 613: 験 614: 宿 615: 毛 616: 従
617: 館 618: 玉 619: 床 620: 休 621: 恋 622: 罪 623: 散 624: 条
625: 速 626: 射 627: 困 628: 程 629: 掛 630: 独 631: 清 632: 弟
633: 夏 634: 崎 635: 昨 636: 権 637: 吸 638: 里 639: 秘 640: 羽
641: 象 642: 姫 643: 雑 644: 布 645: 岡 646: 束 647: 丈 648: 駆
649: 帯 650: 暮 651: 欲 652: 移 653: 職 654: 森 655: 末 656: 焼
657: 園 658: 薄 659: 割 660: 絵 661: 惑 662: 岩 663: 触 664: 害
665: 姉 666: 険 667: 曲 668: 午 669: 働 670: 犬 671: 堂 672: 比
673: 鹿 674: 竜 675: 払 676: 攻 677: 紀 678: 秀 679: 響 680: 登
681: 球 682: 昔 683: 刀 684: 英 685: 踏 686: 握 687: 鼻 688: 史
689: 照 690: 低 691: 衣 692: 洋 693: 骨 694: 福 695: 薬 696: 逆
697: 房 698: 煙 699: 艦 700: 聖 701: 毎 702: 章 703: 宗 704: 留
705: 領 706: 玄 707: 極 708: 捨 709: 未 710: 弁 711: 眺 712: 弱
713: 鳥 714: 沈 715: 育 716: 短 717: 林 718: 晩 719: 型 720: 憶
721: 永 722: 演 723: 刺 724: 訪 725: 跡 726: 伸 727: 闘 728: 銃
729: 個 730: 包 731: 倉 732: 徒 733: 闇 734: 板 735: 雲 736: 縁
737: 快 738: 訳 739: 荒 740: 米 741: 織 742: 那 743: 固 744: 巨
745: 源 746: 徳 747: 忠 748: 狂 749: 修 750: 襲 751: 陸 752: 労
753: 奈 754: 普 755: 筋 756: 脱 757: 捕 758: 迷 759: 皇 760: 収
761: 亡 762: 昼 763: 総 764: 枚 765: 試 766: 雄 767: 印 768: 吐
769: 騒 770: 灯 771: 柄 772: 為 773: 列 774: 迎 775: 芸 776: 隣
777: 習 778: 量 779: 済 780: 輝 781: 捜 782: 甲 783: 技 784: 涙
785: 儀 786: 宇 787: 妹 788: 枝 789: 府 790: 暴 791: 皮 792: 科

793: 恵 794: 震 795: 揺 796: 椅 797: 州 798: 輩 799: 遅 800: 検
801: 鬼 802: 宅 803: 設 804: 岸 805: 穴 806: 輪 807: 週 808: 救
809: 遺 810: 滅 811: 帝 812: 駅 813: 混 814: 描 815: 積 816: 杯
817: 黄 818: 資 819: 詰 820: 皆 821: 鏡 822: 伯 823: 爵 824: 駄
825: 納 826: 節 827: 臣 828: 博 829: 唇 830: 境 831: 廊 832: 砂
833: 改 834: 復 835: 防 836: 毒 837: 尋 838: 邪 839: 竹 840: 矢
841: 麻 842: 晴 843: 箱 844: 宙 845: 祖 846: 伏 847: 盛 848: 宝
849: 洗 850: 敗 851: 如 852: 仰 853: 増 854: 簡 855: 研 856: 児
857: 承 858: 坊 859: 染 860: 浅 861: 群 862: 弥 863: 迫 864: 袋
865: 陰 866: 責 867: 桜 868: 提 869: 届 870: 昭 871: 屈 872: 基
873: 疲 874: 爆 875: 給 876: 絡 877: 練 878: 盗 879: 究 880: 編
881: 諸 882: 湯 883: 詩 884: 陣 885: 純 886: 拠 887: 句 888: 扉
889: 壊 890: 歯 891: 誘 892: 才 893: 評 894: 騎 895: 額 896: 猫
897: 脳 898: 温 899: 借 900: 厳 901: 便 902: 荷 903: 歴 904: 管
905: 甘 906: 被 907: 幕 908: 抗 909: 統 910: 恥 911: 瀬 912: 富
913: 翌 914: 況 915: 衆 916: 値 917: 健 918: 級 919: 希 920: 酔
921: 営 922: 圧 923: 勇 924: 耕 925: 推 926: 価 927: 整 928: 監
929: 授 930: 庫 931: 魚 932: 燃 933: 善 934: 祭 935: 飯 936: 製
937: 棒 938: 紅 939: 勤 940: 至 941: 互 942: 悟 943: 劇 944: 樹
945: 因 946: 功 947: 各 948: 忍 949: 華 950: 砲 951: 丁 952: 虫
953: 寒 954: 央 955: 養 956: 材 957: 慣 958: 栄 959: 冬 960: 郷
961: 傾 962: 替 963: 筆 964: 述 965: 除 966: 準 967: 及 968: 嫉
969: 避 970: 充 971: 己 972: 順 973: 頂 974: 否 975: 幾 976: 賀
977: 録 978: 泊 979: 到 980: 属 981: 繰 982: 厚 983: 旦 984: 奪
985: 辞 986: 区 987: 膝 988: 舎 989: 裂 990: 効 991: 浜 992: 魂
993: 沙 994: 汚 995: 仮 996: 曜 997: 操 998: 威 999: 池 1000: 泉
1001: 財 1002: 謝 1003: 欠 1004: 仏 1005: 僧 1006: 系 1007: 課 1008: 敬
1009: 杉 1010: 崩 1011: 露 1012: 誌 1013: 麗 1014: 裕 1015: 致 1016: 眉
1017: 舌 1018: 慢 1019: 呂 1020: 依 1021: 汗 1022: 油 1023: 導 1024: 牽
1025: 更 1026: 康 1027: 減 1028: 衝 1029: 浴 1030: 埋 1031: 超 1032: 担
1033: 裁 1034: 補 1035: 鋭 1036: 婆 1037: 模 1038: 緊 1039: 専 1040: 珍
1041: 瞳 1042: 脇 1043: 鮮 1044: 隅 1045: 協 1046: 挙 1047: 植 1048: 策
1049: 呪 1050: 牛 1051: 靴 1052: 慮 1053: 執 1054: 港 1055: 県 1056: 脚
1057: 健 1058: 宣 1059: 幼 1060: 展 1061: 匂 1062: 規 1063: 易 1064: 算
1065: 看 1066: 懐 1067: 尻 1068: 炎 1069: 停 1070: 濃 1071: 巡 1072: 妖
1073: 適 1074: 肌 1075: 憎 1076: 臭 1077: 聴 1078: 封 1079: 哀 1080: 獣
1081: 豊 1082: 臓 1083: 飾 1084: 尊 1085: 傍 1086: 尽 1087: 討 1088: 斬
1089: 航 1090: 柳 1091: 載 1092: 省 1093: 藩 1094: 継 1095: 略 1096: 亜
1097: 含 1098: 泥 1099: 禁 1100: 候 1101: 机 1102: 緑 1103: 遣 1104: 雷

1105: 悩 1106: 率 1107: 版 1108: 札 1109: 辛 1110: 距 1111: 烈 1112: 籠
1113: 締 1114: 斎 1115: 獄 1116: 祈 1117: 勘 1118: 狙 1119: 換 1120: 謀
1121: 費 1122: 芝 1123: 測 1124: 乳 1125: 孫 1126: 称 1127: 匁 1128: 墓
1129: 又 1130: 鈴 1131: 革 1132: 揮 1133: 招 1134: 拳 1135: 貸 1136: 貧
1137: 畳 1138: 典 1139: 占 1140: 尉 1141: 灰 1142: 障 1143: 農 1144: 悔
1145: 浦 1146: 署 1147: 召 1148: 虚 1149: 党 1150: 扱 1151: 穂 1152: 奮
1153: 舟 1154: 糸 1155: 狭 1156: 霧 1157: 誠 1158: 柱 1159: 釈 1160: 了
1161: 層 1162: 針 1163: 潜 1164: 帳 1165: 挨 1166: 耐 1167: 講 1168: 般
1169: 即 1170: 帽 1171: 拶 1172: 澄 1173: 雅 1174: 亀 1175: 慎 1176: 拝
1177: 築 1178: 滝 1179: 摩 1180: 弓 1181: 熊 1182: 刊 1183: 努 1184: 冗
1185: 域 1186: 踊 1187: 凄 1188: 魅 1189: 裸 1190: 豪 1191: 昇 1192: 訴
1193: 乾 1194: 旧 1195: 銭 1196: 跳 1197: 塚 1198: 滑 1199: 紹 1200: 齢
1201: 嫁 1202: 督 1203: 則 1204: 律 1205: 仁 1206: 喉 1207: 祝 1208: 旗
1209: 干 1210: 虎 1211: 湖 1212: 浪 1213: 透 1214: 豆 1215: 鶴 1216: 叔
1217: 倍 1218: 誤 1219: 徹 1220: 邸 1221: 丘 1222: 氷 1223: 賞 1224: 覆
1225: 亭 1226: 梅 1227: 猛 1228: 軒 1229: 企 1230: 卒 1231: 拾 1232: 審
1233: 延 1234: 侍 1235: 往 1236: 偶 1237: 牧 1238: 創 1239: 勉 1240: 暇
1241: 頑 1242: 垣 1243: 添 1244: 塔 1245: 葬 1246: 蛇 1247: 賦 1248: 羅
1249: 柔 1250: 歓 1251: 刃 1252: 徴 1253: 標 1254: 懸 1255: 焦 1256: 慌
1257: 姓 1258: 貫 1259: 躍 1260: 棚 1261: 液 1262: 俊 1263: 嘆 1264: 匹
1265: 誇 1266: 掘 1267: 茂 1268: 抵 1269: 垂 1270: 菊 1271: 閣 1272: 荘
1273: 揚 1274: 環 1275: 詳 1276: 愚 1277: 涼 1278: 紫 1279: 複 1280: 双
1281: 趣 1282: 秒 1283: 咲 1284: 携 1285: 撮 1286: 戒 1287: 季 1288: 著
1289: 損 1290: 綱 1291: 援 1292: 漢 1293: 仙 1294: 潮 1295: 寸 1296: 斜
1297: 偵 1298: 拭 1299: 酷 1300: 畑 1301: 械 1302: 阪 1303: 皿 1304: 啓
1305: 塗 1306: 唯 1307: 縮 1308: 堀 1309: 忙 1310: 幹 1311: 唐 1312: 沼
1313: 寂 1314: 委 1315: 慶 1316: 丹 1317: 翼 1318: 暖 1319: 粉 1320: 奏
1321: 貞 1322: 敏 1323: 抑 1324: 賢 1325: 孤 1326: 請 1327: 偉 1328: 索
1329: 梨 1330: 隙 1331: 競 1332: 須 1333: 偽 1334: 珠 1335: 洞 1336: 患
1337: 筒 1338: 縛 1339: 掌 1340: 施 1341: 菜 1342: 凝 1343: 幽 1344: 侵
1345: 預 1346: 訓 1347: 副 1348: 殴 1349: 稿 1350: 淡 1351: 縦 1352: 賊
1353: 袖 1354: 恩 1355: 掃 1356: 腐 1357: 惜 1358: 愉 1359: 惨 1360: 暑
1361: 卓 1362: 庁 1363: 穏 1364: 鎖 1365: 哲 1366: 脅 1367: 隆 1368: 貌
1369: 飼 1370: 益 1371: 稲 1372: 爪 1373: 療 1374: 渋 1375: 廃 1376: 雰
1377: 詞 1378: 恨 1379: 尚 1380: 塩 1381: 沿 1382: 盤 1383: 就 1384: 伴
1385: 幅 1386: 顎 1387: 貨 1388: 紋 1389: 随 1390: 繁 1391: 縄 1392: 童
1393: 邦 1394: 硬 1395: 兼 1396: 寿 1397: 蓋 1398: 拒 1399: 没 1400: 壇
1401: 漂 1402: 唱 1403: 拍 1404: 逸 1405: 曹 1406: 峰 1407: 凍 1408: 免
1409: 既 1410: 釣 1411: 鶏 1412: 征 1413: 網 1414: 乙 1415: 罰 1416: 堅

1417: 羊 1418: 脈 1419: 潰 1420: 孝 1421: 駐 1422: 廷 1423: 遂 1424: 絞
1425: 陥 1426: 砕 1427: 誓 1428: 渉 1429: 胆 1430: 悠 1431: 遇 1432: 噴
1433: 拡 1434: 巧 1435: 殊 1436: 癖 1437: 俗 1438: 災 1439: 僚 1440: 湧
1441: 棄 1442: 症 1443: 謎 1444: 卑 1445: 瓶 1446: 範 1447: 肝 1448: 冒
1449: 彫 1450: 採 1451: 枕 1452: 賭 1453: 誉 1454: 嵐 1455: 朱 1456: 診
1457: 恭 1458: 斉 1459: 寧 1460: 控 1461: 臨 1462: 笛 1463: 喫 1464: 骸
1465: 輸 1466: 憲 1467: 猿 1468: 劣 1469: 熟 1470: 億 1471: 鼓 1472: 沖
1473: 煮 1474: 孔 1475: 狩 1476: 獲 1477: 炉 1478: 賛 1479: 泳 1480: 冊
1481: 序 1482: 催 1483: 寛 1484: 陛 1485: 批 1486: 稽 1487: 戯 1488: 犠
1489: 漏 1490: 乏 1491: 架 1492: 牲 1493: 卵 1494: 壮 1495: 芳 1496: 譲
1497: 匠 1498: 憤 1499: 彩 1500: 潔 1501: 憂 1502: 涯 1503: 盟 1504: 契
1505: 慰 1506: 駒 1507: 摘 1508: 湾 1509: 粒 1510: 旋 1511: 麦 1512: 鈍
1513: 佳 1514: 逮 1515: 溶 1516: 据 1517: 琴 1518: 苛 1519: 較 1520: 睡
1521: 堪 1522: 祥 1523: 粧 1524: 股 1525: 諦 1526: 肥 1527: 塊 1528: 桃
1529: 漠 1530: 磨 1531: 錯 1532: 宴 1533: 悦 1534: 紳 1535: 鍋 1536: 覧
1537: 遭 1538: 凶 1539: 稼 1540: 呉 1541: 棟 1542: 漁 1543: 貝 1544: 晶
1545: 巣 1546: 崖 1547: 隔 1548: 辱 1549: 粋 1550: 滞 1551: 績 1552: 妃
1553: 却 1554: 酸 1555: 絹 1556: 慈 1557: 塞 1558: 鉢 1559: 帆 1560: 窟
1561: 醜 1562: 献 1563: 還 1564: 痕 1565: 郡 1566: 囚 1567: 掲 1568: 贈
1569: 挑 1570: 銅 1571: 鑑 1572: 韓 1573: 挟 1574: 塀 1575: 菓 1576: 朗
1577: 株 1578: 牙 1579: 鬱 1580: 是 1581: 誕 1582: 促 1583: 枯 1584: 畜
1585: 維 1586: 凡 1587: 厄 1588: 痩 1589: 附 1590: 縫 1591: 鎮 1592: 紛
1593: 謙 1594: 汽 1595: 剤 1596: 旨 1597: 徐 1598: 概 1599: 渦 1600: 膚
1601: 綿 1602: 貼 1603: 湿 1604: 肯 1605: 泰 1606: 炭 1607: 符 1608: 鐘
1609: 苗 1610: 蒸 1611: 籍 1612: 粗 1613: 税 1614: 胴 1615: 妬 1616: 鎌
1617: 曇 1618: 岐 1619: 汁 1620: 膨 1621: 翻 1622: 蔑 1623: 裾 1624: 喚
1625: 甚 1626: 勧 1627: 排 1628: 履 1629: 隻 1630: 隷 1631: 赦 1632: 雇
1633: 託 1634: 傘 1635: 猟 1636: 融 1637: 乞 1638: 択 1639: 侯 1640: 錠
1641: 膳 1642: 刷 1643: 瓦 1644: 緩 1645: 盾 1646: 遮 1647: 核 1648: 敢
1649: 尺 1650: 嘲 1651: 滴 1652: 胃 1653: 励 1654: 券 1655: 弱 1656: 拓
1657: 痴 1658: 襟 1659: 克 1660: 嫉 1661: 泡 1662: 飽 1663: 稚 1664: 項
1665: 衰 1666: 怨 1667: 葛 1668: 醒 1669: 扇 1670: 析 1671: 脂 1672: 艇
1673: 糧 1674: 班 1675: 径 1676: 顧 1677: 戴 1678: 浸 1679: 捉 1680: 盆
1681: 伺 1682: 削 1683: 拙 1684: 頭 1685: 均 1686: 陳 1687: 唾 1688: 軌
1689: 巾 1690: 欧 1691: 肘 1692: 缶 1693: 浄 1694: 喪 1695: 忌 1696: 妄
1697: 豚 1698: 冠 1699: 矛 1700: 鍛 1701: 楼 1702: 擦 1703: 零 1704: 戚
1705: 倫 1706: 胞 1707: 汰 1708: 拘 1709: 鋼 1710: 寮 1711: 償 1712: 岳
1713: 虐 1714: 陶 1715: 併 1716: 濁 1717: 疾 1718: 把 1719: 嗅 1720: 剛
1721: 尼 1722: 唄 1723: 罵 1724: 漫 1725: 怠 1726: 崇 1727: 飢 1728: 昧

1729: 艶 1730: 餌 1731: 販 1732: 盲 1733: 郵 1734: 潤 1735: 殻 1736: 臆
1737: 僅 1738: 貯 1739: 鉛 1740: 宰 1741: 撲 1742: 疎 1743: 賃 1744: 銘
1745: 箸 1746: 詮 1747: 濯 1748: 墨 1749: 癒 1750: 肢 1751: 閑 1752: 薫
1753: 俳 1754: 淑 1755: 欺 1756: 侮 1757: 璧 1758: 蜂 1759: 紺 1760: 芽
1761: 弦 1762: 咽 1763: 褒 1764: 刈 1765: 郭 1766: 剖 1767: 餅 1768: 虜
1769: 糖 1770: 粘 1771: 吟 1772: 懇 1773: 妊 1774: 宛 1775: 曽 1776: 秩
1777: 慨 1778: 墳 1779: 棺 1780: 偏 1781: 柿 1782: 帥 1783: 恒 1784: 奔
1785: 藍 1786: 鉱 1787: 峠 1788: 蛮 1789: 禅 1790: 慄 1791: 柵 1792: 暫
1793: 閹 1794: 斗 1795: 膜 1796: 畏 1797: 拐 1798: 溝 1799: 炊 1800: 撤
1801: 挿 1802: 諾 1803: 箇 1804: 沸 1805: 曖 1806: 弄 1807: 懲 1808: 淫
1809: 宵 1810: 溺 1811: 堤 1812: 摂 1813: 募 1814: 堕 1815: 暦 1816: 妨
1817: 羨 1818: 喝 1819: 枠 1820: 椎 1821: 欄 1822: 刹 1823: 某 1824: 簿
1825: 圏 1826: 憧 1827: 漆 1828: 抽 1829: 婿 1830: 遍 1831: 肺 1832: 郊
1833: 蓄 1834: 赴 1835: 詠 1836: 傑 1837: 幣 1838: 票 1839: 桑 1840: 朴
1841: 墜 1842: 暁 1843: 漬 1844: 釜 1845: 蛍 1846: 兆 1847: 伐 1848: 頓
1849: 逐 1850: 蚊 1851: 俵 1852: 麓 1853: 宜 1854: 彰 1855: 吏 1856: 慕
1857: 潟 1858: 爽 1859: 胎 1860: 岬 1861: 餓 1862: 披 1863: 蜜 1864: 塾
1865: 渇 1866: 礎 1867: 阻 1868: 繊 1869: 陵 1870: 塁 1871: 磁 1872: 頻
1873: 旬 1874: 譜 1875: 軟 1876: 后 1877: 錬 1878: 憩 1879: 坪 1880: 軸
1881: 酬 1882: 庶 1883: 腸 1884: 娠 1885: 棋 1886: 殖 1887: 呈 1888: 枢
1889: 粛 1890: 舗 1891: 酌 1892: 伎 1893: 擁 1894: 瞭 1895: 翁 1896: 滋
1897: 峡 1898: 釆 1899: 詣 1900: 芋 1901: 煩 1902: 虹 1903: 貢 1904: 踪
1905: 燥 1906: 渓 1907: 錦 1908: 腫 1909: 褐 1910: 傲 1911: 槽 1912: 謹
1913: 緯 1914: 勲 1915: 肖 1916: 硝 1917: 髄 1918: 捻 1919: 妥 1920: 舷
1921: 購 1922: 畔 1923: 薪 1924: 侶 1925: 藻 1926: 霜 1927: 賜 1928: 拷
1929: 桟 1930: 冥 1931: 勅 1932: 蔽 1933: 剥 1934: 謡 1935: 愁 1936: 芯
1937: 迅 1938: 覇 1939: 穫 1940: 漸 1941: 抹 1942: 朽 1943: 但 1944: 擬
1945: 雌 1946: 碑 1947: 玩 1948: 詐 1949: 諭 1950: 篤 1951: 閥 1952: 昆
1953: 衡 1954: 斑 1955: 搭 1956: 措 1957: 匿 1958: 訂 1959: 茎 1960: 萎
1961: 需 1962: 貿 1963: 享 1964: 債 1965: 挫 1966: 勃 1967: 搬 1968: 媛
1969: 菌 1970: 禍 1971: 尿 1972: 腎 1973: 貪 1974: 括 1975: 剰 1976: 洪
1977: 轄 1978: 嘛 1979: 升 1980: 媒 1981: 廉 1982: 耗 1983: 猶 1984: 睦
1985: 臼 1986: 穀 1987: 瑠 1988: 薦 1989: 頬 1990: 栓 1991: 冶 1992: 繕
1993: 培 1994: 羞 1995: 璃 1996: 鯨 1997: 疫 1998: 遜 1999: 碁 2000: 韻
2001: 扶 2002: 串 2003: 坑 2004: 賄 2005: 叙 2006: 桁 2007: 煎 2008: 倣
2009: 弔 2010: 叱 2011: 匁 2012: 閲 2013: 騰 2014: 礁 2015: 謁 2016: 訟
2017: 屯 2018: 遷 2019: 寡 2020: 惧 2021: 殉 2022: 綻 2023: 畝 2024: 喩
2025: 逝 2026: 窒 2027: 娯 2028: 箋 2029: 凹 2030: 弧 2031: 唆 2032: 旺
2033: 紡 2034: 准 2035: 儒 2036: 惰 2037: 丼 2038: 栽 2039: 辣 2040: 賓

2041: 斥 2042: 践 2043: 陪 2044: 酢 2045: 肪 2046: 庸 2047: 該 2048: 悼
2049: 累 2050: 弊 2051: 梗 2052: 阜 2053: 鋳 2054: 奨 2055: 搾 2056: 拉
2057: 嫡 2058: 硫 2059: 醸 2060: 姻 2061: 酎 2062: 且 2063: 濫 2064: 遡
2065: 糾 2066: 憾 2067: 循 2068: 茨 2069: 酵 2070: 緻 2071: 塡 2072: 舶
2073: 詔 2074: 埼 2075: 賦 2076: 窃 2077: 堆 2078: 壌 2079: 腺 2080: 罷
2081: 栃 2082: 麺 2083: 痢 2084: 矯 2085: 毀 2086: 諧 2087: 賂 2088: 泌
2089: 凸 2090: 氾 2091: 摯 2092: 畿 2093: 祉 2094: 俸 2095: 壱 2096: 脊
2097: 抄 2098: 衷 2099: 卸 2100: 嘱 2101: 厘 2102: 虞 2103: 蚕 2104: 瘍
2105: 倹 2106: 窯 2107: 哺 2108: 賠 2109: 憬 2110: 墾 2111: 租 2112: 丙
2113: 朕 2114: 汎 2115: 弐 2116: 恣 2117: 遵 2118: 捗 2119: 繭 2120: 沃
2121: 彙 2122: 璽 2123: 劾 2124: 諮 2125: 謄 2126: 痘 2127: 斤 2128: 迭
2129: 訃 2130: 款 2131: 頒 2132: 逓 2133: 塑 2134: 錮 2135: 酪 2136: 楷

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO	2. DATA	3. DOCUMENTO Nº	4. Nº DE PÁGINAS
TC	22 de Novembro de 2016	DCTA/ITA/TC-069/2016	101

5. TÍTULO E SUBTÍTULO:

Relate Kanji: A Statistical Analysis Of The Japanese Language And Consequences For Teaching Methods

6. AUTOR(ES):

Márcio Valença Ramos

7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES):

Instituto Tecnológico de Aeronáutica – ITA

8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:

Kanji; Language; Linguistics; Statistics; Pedagogy; Andragogy

9. PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

Aprendizagem (inteligência artificial); Língua japonesa; Linguística; Estatística; Ensino; Computação

10. APRESENTAÇÃO: (**X**) **Nacional** ( ) **Internacional**

ITA, São José dos Campos, Curso de Graduação em Engenharia de Computação. Orientador: Prof.Dr. Carlos Henrique Costa Ribeiro. Publicado em 2016.

11. RESUMO:

The task to learn the written form of the Japanese language is a remarkably daunting one. Japanese is composed of three different scripts, and albeit two of them are relatively simple, with only 46 unique syllabic graphemes each, the third – the Jouyou Kanji – is an ideographic script with 2,136 official ideograms, each with a range of different associated meanings and a number of different readings. Learning this complex script is essential to learning the Japanese language, however there are few teaching methods that rely strongly in statistics for ensuring the efficiency of the teaching method. This academic work proposes better learning methods through the estimation of prevalence of Kanji in the Japanese language, the construction of graphs that relate different letters together and the application of a Personalized PageRank algorithm to estimate the best order a student should follow to learn these ideograms. Finally, we gather the results achieved in the earlier parts of the work to propose aspects of a web app for the teaching of Japanese Jouyou Kanji.

12. GRAU DE SIGILO:

(**X**) **OSTENSIVO** ( ) **RESERVADO** ( ) **CONFIDENCIAL** ( ) **SECRETO**