Savelii Lukash,
Kryvyi Rih National University
Nonna Shapovalova,
Senior Lecturer, Kryvyi Rih National University

RESEARCH AND DEVELOPMENT OF A DYNAMIC LINGUISTIC MODEL AS MEANS OF LEARNING JAPANESE KANJI (LOGOGRAPHIC CHARACTERS AND RADICALS) AND VOCABULARY

This work describes an application of data science methods and techniques to determine the best possible way of learning (or teaching) Japanese logographic characters called Kanji. Another task is to analyze the efficiency of methodologies that humans have developed throughout the years, and that have received wide recognition and application.

Education is one of the fields that benefits the most from any kind of novelties and innovation. Some fields and subjects are affected more than others. Math, History and Law, for example, have already reached a plateau on their innovation curve. Au contraire, other areas of expertise flourish with a vast variety of sometimes opposite approaches, which make learners constantly question their choice and teachers debate the right way to introduce the material. One of such domains is language learning.

The field of language learning poses various diverse tasks such as learning (or teaching) grammar, vocabulary, writing system, pronunciation and many other language concepts. Thousands of scientists, linguists, teachers and language learners constantly come up with new best practices and methods which are better and more efficient than the preceding ones. Needless to say, it's hard or even impossible for humans to encompass every possible aspect, piece of context or use case to produce an ideal linguistic model that accounts for every possible factor. But, it's a rather feasible task for a machine. Hence, why not use the power of data science to see the biggest possible picture and thus get a new linguistic model that is better and more efficient than ever? "How close was the human mind to the perfect solution?" and "How comprehensible will the perfect solution be to the human mind?" are two great questions to find answers for.

Nowadays Japanese writing system comprises two syllabaries: hiragana, katakana and over 50,000 kanji. Yet only 2,136 are considered wide-use and form a list of Jōyō kanji (lit. "regular-use Chinese characters"). Still, learning how to write over two thousand characters of various difficulty (sometimes as easy as「一」– “one”, sometimes as difficult as「鬱」 – “depression”) is a great stumbling block for all Japanese language learners. Two main techniques are used: learning kanji in order of their usage frequency and learning kanji by visual similarity and radical occurrence.

Learning kanji by usage frequency is a common straightforward approach that the vast majority of textbooks and teachers use. It implies learning kanji for frequently used terms, things and ideas first. Here is an example of really common kanji along with their frequency rank (out of 24,882 unique characters found) according to the analysis conveyed in this scientific paper: day「日」– #2, expensive「高」– #35, new「新」– #36, school 「校」– #51, tree「木」– #137, electricity「電」– #161. A good example of a textbook using this technique is “Basic Kanji Book” by Chieko Kano, Hiroko Takenaka, Eriko Ishii and Yuri Shimizu. The pitfall is that frequently used terms do not always have the simplest writing (e.g., “day of the week” -「曜」). The second approach “visual similarity” addresses this issue.

Having to learn the kanji for electricity「電」, for instance, beginner language learners may feel intimidated by the need to remember the whole character at once, which introduces an unnecessary complication. The character for electricity「電」consists of such elements as rain「雨」, field「田」and fish hook「乙」which are just placed one on top of the other. This also enables learners to invent some sort of visual story to remember the character. Such an approach is the key aspect of “Remembering the Kanji” by James Heisig. This way of learning kanji is much more learner-friendly, but on the other side requires students to learn extra characters that they might not need in their early steps.

The main objective of this research is to combine both approaches. Step one is to obtain real usage frequency data from various sources close to real life (such as Wikipedia articles, news articles, manga, anime and

TV shows subtitles, Twitter posts and so on). To implement this stage of the study, it is necessary to perform two sequential tasks: data collection and processing. The authors propose to collect data using web scraping – a technology for obtaining web data by extracting it from web pages. This can be done using Beautiful Soup library to extract data from HTML and XML files. Data processing is conveniently done with Pandas Data Analysis Library. Tools such as distribution histograms and box plots illustrate how often a given symbol is used in different sources. This information makes it possible to assess whether the learner needs to study this hieroglyph as soon as possible, or postpone its acquisition for later. Step two is to establish subset elements for each given character. For example, for the kanji 「電」they are「雨」, 「田」and「乙」. Some of these elements can be further recursively broken down into subparts. And the final step – wisely combine data received from both analyses to form the best possible model that accounts for both usage frequency and kanji visual similarity. A hard problem for a linguist but a regular practical task for a data scientist! Stay tuned for further updates!

SUMMARY

The research is aimed at creating a dynamic linguistic model for learning a foreign language. The work is based on the collection, processing, and analysis of natural language data. The research results can be used not only in the field of learning and linguistics but also as a guide to the practical application of data science tools for natural language processing.

LITERATURE

1. Online Japanese-English dictionary. URL: https://jisho.org (accessed on 15.02.2021).