Ordering Japanese sentences by difficulty

TDT4130 - Text Analysis Project

oysteikt

NTNU

Table of Contents

This article aims to explore the use of natural language processing to order Japanese sentences by their linguistic complexity. In this paper, we provide an overview of the Japanese language and related work in the field, followed by a description of the architecture of our system. We detail the datasets used, the methodology employed, and the evaluation of our system's performance.

1 Introduction

The problem we address in this article arose while developing a mobile dictionary app called Jisho-Study-Tool (h7x4, 2023). We faced a challenge when we needed to link example sentences to words in the dictionary, and arrange them in order. To overcome this challenge, we have utilized techniques and algorithms from natural language processing. In this article, we present our approach to solving this problem.

2 Background

2.1 Japanese Language

Japanese is a language that is very different from English. It employs three writing systems, namely, hiragana, katakana, and kanji. While hiragana and katakana has the same set of characters with different scripts, kanji is a logographic system that is heavily influenced by Chinese characters. Hiragana is generally utilized for native Japanese words, grammatical particles, and verb endings, whereas Katakana is used for loanwords from foreign languages, technical terms, and onomatopoeia. The common term for both of these is *kana*. Kanji, on the other hand, tends to be used for words of Chinese origin such as nouns, adjectives, and verbs. Despite each word typically having a canonical way to be written, the language permits alternative ways of using the writing systems, sometimes for practical purposes, as well as for certain nuances or exceptional cases.

Kanji can have multiple pronunciations, which are usually classified into onyomi and kunyomi. While the difference between these are irrelevant for this article, the fact that there are several pronounciations for a single word brings some challenges. Additionally, the language has many homonyms, which are disambiguated through context when speaking and by using kanji when writing. These homonyms presents both advantages and disadvantages for this project. On one hand, it poses challenges when trying to disambiguate words due to the many words with the same pronunciation. On the other hand, it provides some dimensions lacking from english, that we can utilize to disambiguate the words. For example, some datasets include kana on top of some kanji, named *furigana*. These are written in hiragana to aid in reading the kanji. We can use this in combination with a dictionary to further narrow down which sense of the word is being used.

2.2 Word sense disambiguation

Word sense disambiguation refers to the process of determining which specific meaning or usage of a word is being employed in a given context. This provides important semantic information that is useful in various natural language processing applications. In our case, it helps us gather statistics on the frequency of different word senses and identify common words. There are several algorithms for word sense disambiguation, but in this article, we will utilize a more traditional approach.

2.3 TF-IDF

Term frequency-inverse document frequency, commonly known as TF-IDF, is a popular text vectorization technique that converts raw text into a usable vector. This method combines two important concepts - Term Frequency (TF) and Document Frequency (DF) - to produce a comprehensive representation of text data.

Term frequency refers to the number of times a specific term appears in a document, which helps to determine the importance of that term within the document. By considering the term frequency of every word in a corpus, we can represent the text data as a matrix whose rows correspond to the number of documents and columns correspond to the number of distinct terms found across all documents. Document frequency, on the other hand, measures how many documents contain a specific term, providing insight into the commonality of a particular word across the entire corpus. Finally, the inverse document frequency (IDF) is a weight assigned to each term, which aims to reduce the importance of a term if its occurrences are distributed across all documents. IDF can be calculated using a formula that takes into account the total number of documents and the number of documents containing a particular term.

3 Related work

3.1 Automatic Text Difficulty Classifier

This article describes a system designed to assess the complexity of Portuguese texts, which is intended to provide language learners with texts that correspond to their skill level. To accomplish this, the system extracts 52 features that are grouped into seven categories: parts-of-speech (POS), syllables, words, chunks and phrases, averages and frequencies, and additional features. The system combines these features to calculate a value that represents the text's level of difficulty. The approach of using several features of different kinds is similar to the way we do it in this project. (Curto, Mamede, and Baptista, 2015)

3.2 Jisho.org

Jisho is an online Japanese-English dictionary that offers a wide range of features for searching words, kanji, and example sentences. To accomplish this, Jisho integrates various data sources, including the Japanese-Multilingual Dictionary (JMDict) and the Tanaka Corpus, which will be explained further later on. One of the useful features of Jisho is the ability to provide example sentences to illustrate how a word is used in context. To achieve this, Jisho has employed a similar approach to their data aggregation, although not exactly the same. Although the source of their product is closed, some of the tools used in the process are publicly available. During the development of this project, their kana-romaji translator (Ahlstrom, 2023a) has proven to be a valuable tool. Unfortunately, Jisho usually only provides one or two sentences per sense if any, so it is not as useful as a comparison.

3.3 Surrounding Word Sense Model for Japanese All-words Word Sense Disambiguation

This paper proposes a surrounding word sense model (SWSM) that uses the distribution of word senses that appear near ambiguous words for unsupervised all-words word sense disambiguation in Japanese. It is based around the idea that words with the same senses will often appear with the same surrounding words. By utilizing dictionary data in addition to WORDNET-WALK, they have created an engine which is more accurate than existing supervising. models. This could be used in combination with this project to make it more accurate in the future. (Komiya et al., 2015)

4 Architecture

4.1 Datasets

4.1.1 JMDict

JMDict is a publicly available Japanese to multilingual dictionary developed by Jim Breen and his associates at the Electronic Dictionary Research and Development Group (EDRDG). The dictionary has various types of information such as kanji, readings, word senses, and more. It also includes rare information like different newspaper indices for the different word senses, and the origins of loan words. This resource is valuable to us since it provides a predetermined wordlist that we can use to link our sample sentences. Additionally, JMDict can be utilized as a query tool to examine relationships between words and senses. (Breen, 2004)

4.1.2 Tanaka corpus

The Tanaka corpus is a compendium of sentences that includes an English-translated version for most of them. This compilation was put together by Asuhito Tanaka, who is a professor at Hyogo

University. Originally, the corpus was created by assigning the task of collecting 300 sentence pairs to Professor Tanaka's students. After several years, they had collected 212,000 sentence pairs. In 2002, the EDRDG started to work on creating links to the entries in JMDict. In 2006, them maintanership of the corpus was incorporated into the Tatoeba project. The current version of the corpus released by the EDRDG comes preprocessed with lemmatizations, furigana, and other supplementary data. (Tanaka, 2001)

4.1.3 NHK Easy News

JMDict contains a wealth of information on the frequency of words in the Japanese language. However, some of these statistics are derived from Japanese newspapers, which are renowned for being challenging even for learners in the advanced stages.

Fortunately, Japan's state media, NHK, publishes a newspaper that is designed for learners. This is a valuable resource since we can be certain that every word in this corpus is suitable for learners. Therefore, we will utilize this corpus to construct a new index, which we can use to determine whether a word is suitable for learners.

4.2 Methodology

4.2.1 Data ingestion

The first task was to ingest and preprocess the data from the different sources. For this, we chose to use an SQL database, because it provides us with an easy way of storing temporary result and quickly retrieving entries for complex queries. By reading the document type definition (consortium, 2023) of the JMDict XML-file, we were able to construct most of the schema of the database. Some parts of the schema was never used, so there is a bit of data loss in this process.

NHK News publishes an official index of the last articles from the past year at `http://www3.nhk.or.jp/news/easy/news-list.json`. We have used this to be able to download those articles, and then scrape them for content with an HTML parser. Afterwards, this was also put into the SQL database.

The sentences from the Tanaka Corpus were ingested in a similar manner.

4.2.2 Word Sense Disambiguation

Both corpora contain elements that can facilitate the disambiguation process.

The Tatoeba sentences are already partially annotated with lemmatizations, furigana (which denotes the spelling of the kanji), and at times, even the JMDict identifier. However, the sense disambiguation process is limited to a specific entry. Here, we could have used SWSM in an attempt to further disambiguate the word to its senses listed as listed in the dictionary.

The NHK Easy news corpus doesn't have these kinds of annotations. To solve this problem, we use a combination of the furigana from the corpus, MeCab to analyze the words and get POS tags, and a prioritized list of how to search for the correct meaning of the word. We created a mapping from the MeCab part-of-speech tags to the JMDict tags. The first word that fits based on its existing level of commonality data, and which is also the most likely match is chosen as the match. If no matches are found, the word is not added to the list of connected entries.

This approach may have a limitation where it could make some frequently used words appear even more frequent than they actually are. As a result, some words that are commonly used, but not as much as their similar counterparts, could be wrongly classified as very rare because they don't seem to appear in the NHK Corpora.

4.2.3 TF-IDF

TF-IDF is often used as a tool to estimate how meaningful a word is for one document in a corpus. However, here we want the opposite measure. We are not looking for the words that give the documents most of its meaning, but rather the words which are more common across several documents. If a token only has a high frequency in one of the documents, then there is a high chance that the word is field specific to this document only.

Because of this, we are going to change the formula to give us the averaged term frequency times the document frequency.

$$\begin{aligned} AVG(TF) &= \frac{AVG\,(\text{Occurences of term in document})}{\text{Amount of terms in document}} \\ DF &= \frac{\text{Count of documents where term exists}}{\text{Document Count}} \\ \text{TF-DF} &= AVG(TF) \cdot DF \end{aligned}$$

We then went over the NHK Easy News corpus and collected the "TF-DF" values from here. These were then normalized to be in $[0, 1]$

4.2.4 Determining word and sentence difficulty

At this point, there are a lot of potential factors available to work with. To organize sentences properly, we need to determine how hard the words and sentences are to understand by aggregating some of these factors. We have picked a few factors that we believe are useful to determine the difficulty values, but the chosen curvatures and weights are just based on trial and error, and educated guesses.

Figure 2 shows how the different factors contribute to a words difficulty.

The sentence factors are listed in Figure 1.

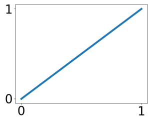
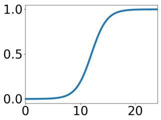
Factor	%	Curve	Notes and reasoning
$\frac{\sum \text{difficulty(word)}}{\text{length of sentence}}$	50%		This is the aggregated value based on the calculation in Figure 2. As the values should be decently curved already, they are left unaffected. We also believe that this should have a lot more effect on the sentence than the other two factors.
max(difficulty(word))	20%		The hardest word in the sentence can be the word that makes the whole sentence useless for a learner. Because of that, we make the hardest word in the sentence its own factor.
Sentence length	30%		Until a sentence reaches around 12 words, it should be regarded as quite easy. But once it surpasses that, it becomes more difficult.

Figure 1: Contributing factors to a sentences difficulty

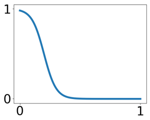
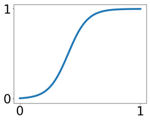
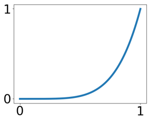
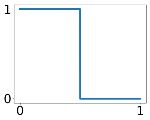
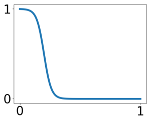
Factor	%	Curve	Notes and reasoning
Common ratings	25%		The different existing ratings of the word are summed together and linearly squished into $[0, 1]$. If the entry is included in more than one or two indices, it can be assumed that it is quite a common word, and should be marked as very easy.
Dialects	10%		This is the sum of all readings which are marked as dialect. If a word has more than roughly 30% dialect readings, we assume that it is a very dialect specific word. This should increase its difficulty.
Most difficult kanji	25%		The input here is the elementary school grade in which the kanji is thaught, where grade 7 is the rest of the $\overset{\text{jouyou}}{\text{常用}}$ kanji (Agency for Cultural Affairs, 2023), and 8 are everything else. Usually, grade 1-6 means that the word is easy, grade 7 is intermediate-difficult, and 8 is extremely difficult. There is an edge case here, where a word has a set of really difficult kanji, but they are usually not used. These come pretagged as such, and are removed from the calculation.
Katakana word	15%		If a word only contains katakana, there is a good chance that it is a loanword from english. This is usually a clear cut case, but some words have alternative kanji that are rarely used. Examples might include $\overset{\text{ぺーじ}}{\text{頁}}$(page) and $\overset{\text{コーヒ}}{\text{珈琲}}$(coffee). Because of this, we use a hard limit at 50% for how many readings are katakana only.
NHK Easy News Frequency Rating	25%		In order to get rid of the words that are document specific, we make the S-curve mark the lower valued words as difficult, but quickly remove their difficulty if they appear more often

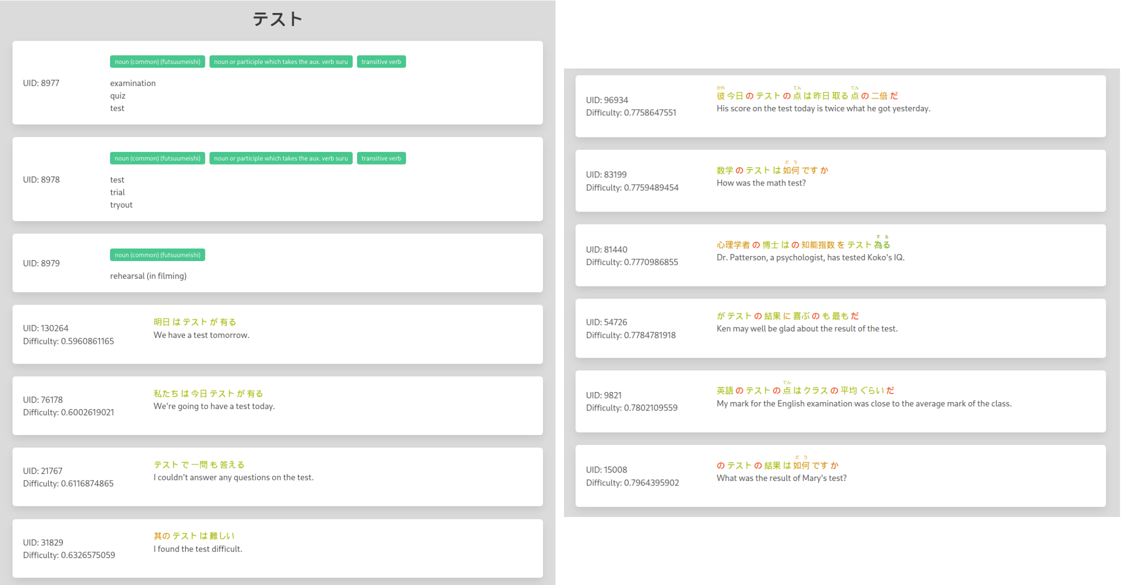
Figure 2: Contributing factors to a words difficulty

5 Evaluation and conclusion

5.1 Evaluation

Despite being unable to measure the accuracy of the results, the first impression was quite good.

Here is an example of the sentences connected to the word テスト (test)



テスト

UID: 8977
noun (common) (futsuumeishi) | noun or participle which takes the aux. verb suru | transitive verb
examination
quiz
test

UID: 8978
noun (common) (futsuumeishi) | noun or participle which takes the aux. verb suru | transitive verb
test
trial
tryout

UID: 8979
noun (common) (futsuumeishi)
rehearsal (in filming)

UID: 130264
Difficulty: 0.5960861165
明日 は テスト が 有る
We have a test tomorrow.

UID: 76178
Difficulty: 0.6002619021
私たち は 今日 テスト が 有る
We're going to have a test today.

UID: 21767
Difficulty: 0.6116874865
テスト で 一問 も 答える
I couldn't answer any questions on the test.

UID: 31829
Difficulty: 0.6326575059
其の テスト は 難しい
I found the test difficult.

UID: 96934
Difficulty: 0.7758647551
$\overset{\text{かれ}}{\text{彼}}$ 今日 の テスト の $\overset{\text{てん}}{\text{点}}$ は 昨日 取る $\overset{\text{てん}}{\text{点}}$ の 二倍 だ
His score on the test today is twice what he got yesterday.

UID: 83199
Difficulty: 0.7759489454
数学 の テスト は $\overset{\text{どう}}{\text{如何}}$ です か
How was the math test?

UID: 81440
Difficulty: 0.7770986855
心理学者 の 博士 は の 知能指数 を テスト $\overset{\text{する}}{\text{為る}}$
Dr. Patterson, a psychologist, has tested Koko's IQ.

UID: 54726
Difficulty: 0.7784781918
が テスト の 結果 に 喜ぶ の も 最も だ
Ken may well be glad about the result of the test.

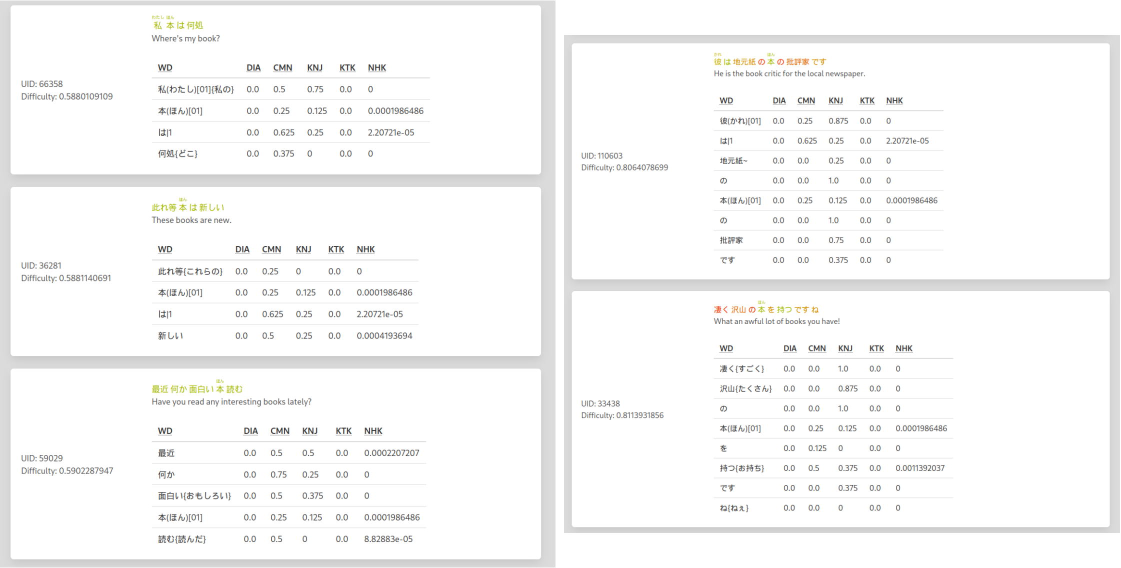
UID: 9821
Difficulty: 0.7802109559
英語 の テスト の $\overset{\text{てん}}{\text{点}}$ は クラス の 平均 ぐらい だ
My mark for the English examination was close to the average mark of the class.

UID: 15008
Difficulty: 0.7964395902
の テスト の 結果 は $\overset{\text{どう}}{\text{如何}}$ です か
What was the result of Mary's test?

Figure 3: Example sentences for the word "test", with the easiest and hardest difficulty levels

From this example, it seems to work quite well, with the one big exception being the particles. These are small suffixes which only exists to indicate the grammatical meaning of the word before it. While these are probably some of the absolutely most common pieces in the japanese language, they have been marked as very difficult. However, the easier words have been marked green, while the harder ones have gotten an orange color.

Here is another example for $\overset{\text{ほん}}{\text{本}}$(book), which has several senses. We have turned on debug information, to see the contributing factors.



UID: 66358
Difficulty: 0.5880109109

私 本 は 何処
Where's my book?

WD	DIA	CMN	KNJ	KTK	NHK
私(わたし)[01]{私の}	0.0	0.5	0.75	0.0	0
本(ほん)[01]	0.0	0.25	0.125	0.0	0.0001986486
は|1	0.0	0.625	0.25	0.0	2.20721e-05
何処{どこ}	0.0	0.375	0	0.0	0

UID: 36281
Difficulty: 0.5881140691

此れ等 本 は 新しい
These books are new.

WD	DIA	CMN	KNJ	KTK	NHK
此れ等{これらの}	0.0	0.25	0	0.0	0
本(ほん)[01]	0.0	0.25	0.125	0.0	0.0001986486
は|1	0.0	0.625	0.25	0.0	2.20721e-05
新しい	0.0	0.5	0.25	0.0	0.0004193694

UID: 59029
Difficulty: 0.5902287947

最近 何か 面白い 本 読む
Have you read any interesting books lately?

WD	DIA	CMN	KNJ	KTK	NHK
最近	0.0	0.5	0.5	0.0	0.0002207207
何か	0.0	0.75	0.25	0.0	0
面白い{おもしろい}	0.0	0.5	0.375	0.0	0
本(ほん)[01]	0.0	0.25	0.125	0.0	0.0001986486
読む{読んだ}	0.0	0.5	0	0.0	8.82883e-05

UID: 110603
Difficulty: 0.8064078699

彼 は 地元紙 の 本 の 批評家 です
He is the book critic for the local newspaper.

WD	DIA	CMN	KNJ	KTK	NHK
彼(かれ)[01]	0.0	0.25	0.875	0.0	0
は|1	0.0	0.625	0.25	0.0	2.20721e-05
地元紙~	0.0	0.0	0.25	0.0	0
の	0.0	0.0	1.0	0.0	0
本(ほん)[01]	0.0	0.25	0.125	0.0	0.0001986486
の	0.0	0.0	1.0	0.0	0
批評家	0.0	0.0	0.75	0.0	0
です	0.0	0.0	0.375	0.0	0

UID: 33438
Difficulty: 0.8113931856

凄く 沢山 の 本 を 持つ です ね
What an awful lot of books you have!

WD	DIA	CMN	KNJ	KTK	NHK
凄く{すごく}	0.0	0.0	1.0	0.0	0
沢山{たくさん}	0.0	0.0	0.875	0.0	0
の	0.0	0.0	1.0	0.0	0
本(ほん)[01]	0.0	0.25	0.125	0.0	0.0001986486
を	0.0	0.125	0	0.0	0
持つ{お持ち}	0.0	0.5	0.375	0.0	0.0011392037
です	0.0	0.0	0.375	0.0	0
ね{ねぇ}	0.0	0.0	0	0.0	0

Figure 4: Example sentences for the word "book", with the easiest and hardest difficulty levels

Here we can see the internal details as to why the particles have been so difficult. They seem to be marked as the most difficult on the kanji scale. This is a bug, since the kanji system is supposed to filter out anything that is not a kanji. Unfortunately, while we spent quite a lot of time on trying to fix this, we can not figure out why it is acting as it does, and we are soon reaching the deadline of the project.

5.2 Conclusion

While the system performs well by our random samples, there are still some impurities to be researched further. There are also some bugs left to be fixed.

There are many other factors that we haven't explored yet which could be useful. For example, many sentences in the Tatoeba Corpus are already labeled with tags, some of which could indicate whether a sentence is difficult or not. We could also look to the automatic text difficulty classifier project for additional ideas on which factors to consider.

We also think more research is necessary to establish the correct weighting for different factors and which curves should be used. This requires examining which factors of a word are the most important for determining its level of difficulty. This is crucial for ensuring that the sorting system works correctly. Additionally, we need to investigate how to handle sentences with unfamiliar words to ensure they are sorted in a reasonable way.

References

Agency for Cultural Affairs, G.o.J., 2023. 常用漢字表の音訓索引 [Online]. Available from: https://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/kijun/naikaku/kanji/joyokanjisakuin/index.html [Accessed April 17, 2023].

Ahlstrom, K., 2023a. *Japanese$_t$ransliterators.rb* [Online]. Available from: https://github.com/Kimtaro/ve/blob/master/lib/providers/japanese_transliterators.rb [Accessed April 19, 2023].

Ahlstrom, K., 2023b. *Jisho.org* [Online]. Available from: https://jisho.org/about [Accessed April 20, 2023].

Breen, J., 2004. Jmdict: a japanese-multilingual dictionary [Online]. Available from: https://www.edrdg.org/jmdict/jmdictart.html.

consortium, W. wide web, 2023. *Prolog and document type declaration* [Online]. Available from: https://www.w3.org/TR/xml11/#sec-prolog-dtd [Accessed April 22, 2023].

Curto, P., Mamede, N., and Baptista, J., 2015. Automatic text difficulty classifier - assisting the selection of adequate reading materials for european portuguese teaching [Online], pp.36–44. Available from: https://doi.org/10.5220/0005428300360044.

h7x4, 2023. *Jisho study tool* [Online]. Available from: https://github.com/h7x4/Jisho-Study-Tool [Accessed April 15, 2023].

Jurafsky, D. and Martin, J.H., 2000. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. 1st. USA: Prentice Hall PTR. ISBN: 0130950696.

Komiya, K., Sasaki, Y., Morita, H., Sasaki, M., Shinnou, H., and Kotani, Y., 2015. Surrounding word sense model for japanese all-words word sense disambiguation. *Proceedings of the 29th pacific asia conference on language, information and computation* [Online], pp.35–43. Available from: https://cir.nii.ac.jp/crid/1050282677488198784.

McCann, P., 2020. Fugashi, a tool for tokenizing Japanese in python. *Proceedings of second workshop for nlp open source software (nlp-oss)* [Online]. Online: Association for Computational Linguistics, pp.44–51. Available from: https://www.aclweb.org/anthology/2020.nlposs-1.7.

Tanaka, Y., 2001. Compilation of a multilingual parallel corpus [Online]. Available from: https://www.edrdg.org/projects/tanaka/tanaka.pdf.