*Human Factors in Design, Engineering, and Computing,* Vol. 159, 2024, 158–167
https://doi.org/10.54941/ahfe1005578

# Construction of Japanese-Chinese Onomatopoeia Corpus Based on Events and Behaviors

**Wenjing Zhang[1], Amika Chino[1], Siyu Yan[1], and Takehiro Teraoka[2]**

[1] Graduate School of Engineering, Takushoku University, Japan
[2] Department of Computer Science, Faculty of Engineering, Takushoku University 815–1 Tatemachi, Hachioji, Tokyo 193–0985, Japan

## ABSTRACT

Many manga are translated into various languages each year, but the onomatopoeia in them often goes untranslated, affecting the reader's experience. In this study, we first translate Japanese onomatopoeia into Chinese, with the ultimate goal of multilingual translation. We constructed a Japanese-Chinese onomatopoeia corpus (JCO), including accurate translations of onomatopoeia and its related words. The most suitable words are determined by calculating cosine similarity and Levenshtein distance, and if this is not possible, the International Phonetic Alphabet (IPA) transcription is used. We developed a tool to help manga translators determine the most appropriate translation. In the experiment, five manga were randomly selected from Manga 109 and 100 unique onomatopoeic words were extracted and tested, with positive feedback.

**Keywords:** Japanese-Chinese corpus, Manga, Translation of onomatopoeia, Translation with related word

## INTRODUCTION

Japan is one of the leaders in the comic industry, and the word "manga" is used worldwide to specifically refer to Japanese comics. Numerous manga are translated into various languages every year, but the onomatopoeia often remains untranslated, which may take away from the readers' experience. One of the characteristics of the Japanese language is its rich variety of onomatopoeic words, which are used not only in manga but also in colloquial language. Having a clear understanding of the meaning of onomatopoeia is crucial for a deeper understanding of manga scenes. Since 2016, neural networks have been widely used in machine translation, and the introduction of neural machine translation (NMT) in particular gradually improved the accuracy and fluency of translations. Subsequently, the Transformers architecture led to more natural translations and handled long sentences and complex semantics more effectively. However, current machine translation methods are incapable of accurately translating onomatopoeia. Tables 1 and 2 show the results of translating Japanese onomatopoeia into Chinese using Google, Deepl, Baidu, and ChatGPT 4. As shown, even the latest version of ChatGPT cannot provide an accurate translation.

This research aims to address the issue of translating onomatopoeia. Starting with Chinese, we construct a Japanese-Chinese onomatopoeia (JCO) corpus and develop a tool to assist translators with selecting appropriate translations.

**Table 1.** Results of machine-translated onomatopoeia.

| | **ド (do) "Thud"** | **タ (ta) "Tap"** |
|---|---|---|
| Google | 做 (Zuò) "Do" | 稲田 (Dàotián) "Paddy" |
| DeepL | 做(Zuo) "Do" | 淘寶(Táobao) "Taobao" |
| Baidu | 文档(Wéndàng) "Document" | 數據(Shùjù) "Data" |
| ChatGPT 4 | lacks a direct translation | lacks a direct translation |

**Table 2.** Results of machine-translated onomatopoeia.

| | **ポンポン (ponpon) "Pom-pom"** | **ペコペコ (pekopeko) "Hungry"** | **ピリピリ (piripiri) "Tingling"** |
|---|---|---|---|
| Google | 絨球 (Róng qiú)<br>"Pompom" | 飢餓的 (Jı'è de)<br>"Hungry" | 刺痛 (Cì tòng)<br>"Stinging" |
| DeepL | 貝貝(Bèi bèi)<br>"Shell" | 餓了(Èle) "Hungry" | 灼舌(Zhuó shé)<br>"Scalding" |
| Baidu | 蓬蓬(Péng péng)<br>"Fluffy" | 佩科佩科(Pèi ke pèi ke)<br>"Peckish" | 火辣辣的(Huo là là de)<br>"Spicy" |
| ChatGPT 4 | 拍拍(pai pai)<br>"Patting" | 餓(è)<br>"Hungry" | 緊張(jın zhang)<br>"Tense" |

## RELATED WORK

### Problem of Onomatopoeia Translation

In everyday life, everyone may encounter vocabulary gaps, such as when searching for information in another country and not knowing the corresponding words in the target language. This often occurs with Japanese onomatopoeia, where the equivalent keywords in other languages are unknown. Chen et al. (2013) investigated whether foreign readers could understand manga as well as Japanese readers, focusing on onomatopoeia. They found that Chinese readers familiar with Japanese noticed onomatopoeia but struggled with comprehension due to the diversity and nuances of these expressions. Those unfamiliar with Japanese mostly did not notice onomatopoeia and felt no discomfort. However, when onomatopoeia is translated into Chinese, readers can better appreciate the emotions conveyed, suggesting that appropriate translation enhances manga enjoyment. Onomatopoeia includes both onomatopoeic and mimetic words

(Kawasaki, 2006; Muramoto, 2006). Onomatopoeic words mimic sounds, while mimetic words represent states of things, motion, human senses, and psychological states.

### Chinese Translations of Manga Onomatopoeia

Translations of Japanese onomatopoeia are often ambiguous due to cultural differences. Hou and Matsuo (2019) analyzed the accuracy of onomatopoeia translations in the manga series "Crayon Shin-chan" and "Doraemon" and found that onomatopoeia was often mistranslated or omitted entirely. The limitations of onomatopoeia translation can be attributed to the following three key aspects: 1. Differences in linguistic structure, 2. Differences in manga structure, 3. Unique onomatopoeia with no equivalents in other languages.

### Typology of Chinese Translations for Japanese Onomatopoeia

Zhang et al. (2017) proposed a method for translating Japanese onomatopoeia that does not have direct equivalents in Chinese by dividing Japanese onomatopoeia into mimetic words and phrases and then translating the mimetic phrases into Chinese. They determined that it is not necessary to find the exact words that match the original meaning. Instead, using the context and sentiment of the original text can also provide an accurate translation of mimetic words, that is, using alternative expressions for the same semantic meaning.

### Machine Translation From Japanese to Chinese Onomatopoeia

Yang et al. (2018) proposed a method of translating Japanese onomatopoeia to Chinese by transforming and classifying their Mel-frequency cepstral coefficients (MFCC). After obtaining the MFCC of the Chinese translation of Japanese onomatopoeia, a Chinese onomatopoeia database is used to output the Chinese onomatopoeia that most closely matches.

### Automated Selection of Onomatopoeia in Japanese-Chinese Texts

Japanese onomatopoeic words play an important role in daily communication due to their rich expressions and deep cultural connotations, but these words are often difficult to be accurately conveyed in translation, which is especially obvious when Chinese learners use machine translation tools. To address this issue, Harada et al. (2017) and team propose a new translation selection method, in which they first collect candidate translations from multiple translation tools, and then evaluate their accuracy by calculating the co-occurrence rate of onomatopoeic words with other words in each candidate translation. In addition, the study also considered the rationality of grammatical structures, and further screened the translation results by setting specific grammatical rules (e.g., the correct use of modifiers). Tested on real Chinese and Japanese corpora, Harada's team's method showed higher accuracy than conventional machine translation.

### Sound-Based Translation of Japanese Onomatopoeia

Tsuchiya et al. (2012) and Shimizu et al. (2012) noted that the meaning of onomatopoeia is inferred from Japanese sound semiotic features such as the pattern of morpheme sequences in onomatopoeia and the meaning of each morpheme. Han et al. (2019) proposed a method for transcribing Japanese onomatopoeia to Chinese by breaking down the input onomatopoeia into sound units called moras and using the International Phonetic Alphabet (IPA) to transcribe the moras into Chinese onomatopoeia. If there is Chinese onomatopoeia in the dataset that corresponds to the input onomatopoeia, the sound similarity between the Chinese characters is calculated and the closest result is output. If there is no corresponding Chinese onomatopoeia, the method calculates the similarity between the moras and the sounds of Chinese characters and outputs the Chinese characters that sound most similar. Finally, all of the Chinese characters are organized in the same order as the corresponding original Japanese onomatopoeia. For example, the process of transcribing the word "ゴリゴリ" is shown in Figure 1.

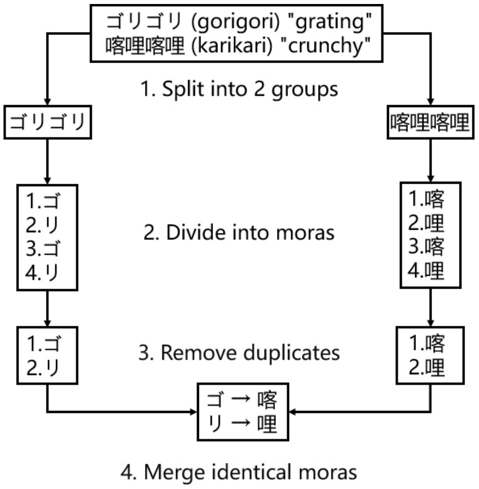

**Figure 1:** Process of transcribing "ゴリゴリ".

## CORPUS

### Japanese–Chinese Onomatopoeia (JCO) Corpus

We collected numerous onomatopoeic words extracted from Manga 109, supplemented by additional entries sourced from the internet, and their related words from dictionaries and online to expand bilingual dictionaries. This includes 1017 words that come from the "Grand Dictionary" (大辞典), online compilations of onomatopoeia, and comics by five different manga artists, or mangaka. In addition, we used Word2Vec to calculate the cosine similarity of related words. If the similarity is high, they were added as related words. For example, the onomatopoeia "ビリビリ" (biri biri) can be interpreted as "numbness" or "an electric shock." If the meaning is "numbness," words such as "麻痺" (paralysis), "麻醉" (anesthesia), and "麻

婆豆腐” (mapo tofu) are shown as related words. For the meaning “electric shock,” words such as “電撃” (electric shock), “雷電” (thunderbolt), and “ライトニング” (lightning) are included as related words.

**Table 3.** Japanese–Chinese onomatopoeia (JCO) corpus.

| Japanese Onomatopoeia | Chinese Translation | Related Word |
|---|---|---|
| ぺこぺこ (pekopeko) “Hungry” | 餓 (è) | 凹む(hekomu) “To dent” |
| ゲロゲロ (gerogero) “Vomit” | 嘔嘔 (ou ou) | 吐き気(hakike) “Nausea”<br>吐金く属 (haku) “To vomit” |
| ガリッ (gari) “Crunch” | 咬咬 (yao yao) | 金属 (kinzoku) “Metal”<br>磁器 (jiki) “Ceramic”<br>揺れる (yureru) “To shake” |

### Chinese Onomatopoeia Dictionary

If the output from the JCO corpus does not provide an accurate result, we also constructed a Chinese Onomatopoeia Dictionary to refine the answer. This dictionary includes 322 onomatopoeia and their descriptions from Chinese dictionaries and online websites. By extracting vocabulary from these descriptions, we created a dictionary of onomatopoeia-related words in Chinese, which enables users to find the most appropriate onomatopoeia by entering the related words.

## PROPOSED METHOD

### System Flow

Figure 2 illustrates the working flow of the system of the proposed method. First, the user inputs a Japanese onomatopoeic word and its related word, and the system checks if a Chinese translation exists in the JCO corpus. If a corresponding translation exists, it will be output as the result. If not, Word2Vec (Mikolov et al., 2013) is then used to calculate the cosine similarity between the input Japanese onomatopoeia and all of the related words in the corpus to find all related words with a similarity of over 0.7. Next, the Levenshtein distance is calculated for the input Japanese onomatopoeia and all results with a similarity over 0.7, and the Chinese onomatopoeia translation with the closest Levenshtein distance is output. If the system cannot find any related words with similarity over 0.7, the user needs to input words related to the Japanese onomatopoeia. Then the Chinese Onomatopoeia Dictionary is used to calculate the similarity between the related word input by the user and the descriptions in the Chinese Onomatopoeia Dictionary. The word with the highest similarity will be output as the result. If there is no word with a similarity over 0.7, then the IPA is used to transcribe the Japanese characters into their closest Chinese phonetic counterparts.

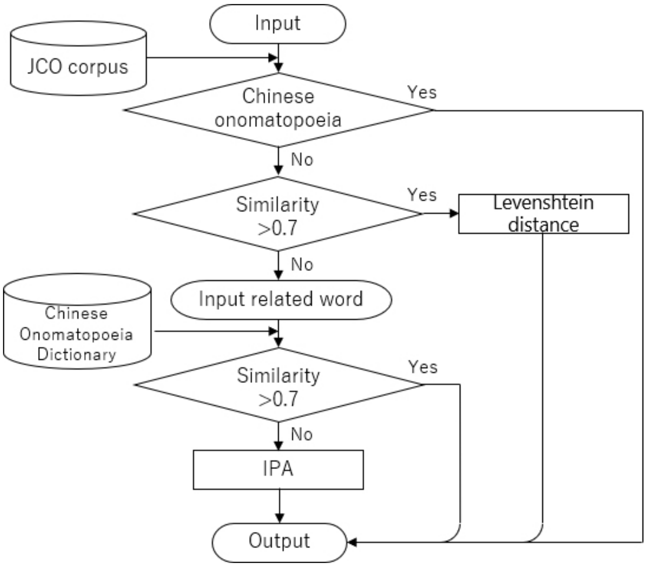

**Figure 2:** Working flow of the system of the proposed method.

### Word2Vec

Cosine similarity is commonly used to calculate the similarity between words or sentences. We trained Word2Vec with the data of the Japanese Wikipedia, which will then be used to determine the relation of words through cosine similarity. The Word2Vec model maps words into vectors in a high-dimensional space, and by calculating the cosine similarity, the semantic similarity between two words can be determined. The formula is as follows:

$$\text{Cosine Similarity} = \frac{v_1 v_2}{\|v_1\| \|v_2\|}$$

The cosine similarity ranges from –1 to 1. A cosine similarity closer to 1 indicates that the words are more semantically similar; if it is closer to –1, they are more dissimilar; if it is close to 0, there is no significant similarity.

### Levenshtein Distance

The Levenshtein distance is the minimum number of single-character edits required to change one word into another. For example, the Levenshtein distance between “ビリビリ” and “ビリビリ” is 0, and that between “ビリビリ” and “ピリピリ” is 2.

## EXPERIMENT

### Experiment 1

In our experiment, we chose 100 onomatopoeic words from five manga series in Manga109 dataset and translated them into Chinese with our proposed method. The manga series are YumeNoKayoiji, AosugiruHaru, MisutenaideDaisy, JangiriPonpon, and MoeruOnisan_vol19. After that, we conducted a questionnaire consisting of two questions: the first one asks whether the translation is appropriate (Table 3), and the second asks whether the translation fits the context of the manga (Table 4). For the second question, the participants were instructed to fill in blank manga panels with the corresponding onomatopoeia. All 100 translations were included in each question, for a total of 200 translations. The proposed method provides four ways to translate Japanese onomatopoeia. In our experiment, 63 translations were translated by the JCO corpus, 20 were translated by similarity calculations, six were translated by the Chinese Onomatopoeia Dictionary, and the last 11 were transcribed by IPA. The participants were 17 native Chinese speakers (NS) with advanced Japanese proficiency (JLPT N1) and two Japanese NS with high Chinese proficiency (HSK level 6). The translations were evaluated using a five-point scale.

**Table 4.** Results of meaning accuracy.

| | Chinese NS (17) | | | | | | Japanese NS (2) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | Average | 1 | 2 | 3 | 4 | 5 | Average |
| JCO corpus | 0.03 | 0.04 | 0.11 | 0.26 | 0.56 | 4.28 | 0.11 | 0.07 | 0.12 | 0.28 | 0.42 | 3.83 |
| Similarity calculation | 0.05 | 0.03 | 0.15 | 0.25 | 0.52 | 4.16 | 0.03 | 0.08 | 0.03 | 0.30 | 0.58 | 4.33 |
| Chinese dictionary | 0.05 | 0.09 | 0.16 | 0.25 | 0.45 | 3.97 | 0 | 0.08 | 0.17 | 0.25 | 0.50 | 4.17 |
| IPA | 0.05 | 0.03 | 0.17 | 0.33 | 0.43 | 4.06 | 0.32 | 0.05 | 0.18 | 0.27 | 0.18 | 2.95 |
| Overall | 0.04 | 0.04 | 0.13 | 0.26 | 0.53 | 4.21 | 0.11 | 0.07 | 0.11 | 0.28 | 0.43 | 3.85 |

**Table 5.** Results of nuance accuracy.

| | Chinese NS (17) | | | | | | Japanese NS (2) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | Average | 1 | 2 | 3 | 4 | 5 | Average |
| JCO corpus | 0.02 | 0.04 | 0.09 | 0.25 | 0.59 | 4.35 | 0.13 | 0.06 | 0.12 | 0.21 | 0.48 | 3.84 |
| Similarity calculation | 0.04 | 0.03 | 0.14 | 0.25 | 0.54 | 4.20 | 0.28 | 0.10 | 0.13 | 0.23 | 0.28 | 3.13 |
| Chinese dictionary | 0.06 | 0.07 | 0.15 | 0.26 | 0.46 | 4.00 | 0.42 | 0.17 | 0.17 | 0.17 | 0.08 | 2.33 |
| IPA | 0.03 | 0.03 | 0.12 | 0.30 | 0.30 | 4.26 | 0.05 | 0.05 | 0.09 | 0.14 | 0.68 | 4.36 |
| Overall | 0.03 | 0.04 | 0.11 | 0.26 | 0.56 | 4.29 | 0.17 | 0.07 | 0.12 | 0.21 | 0.44 | 3.67 |

Likewise, the results of Table 4 show that for Chinese NS all of the results of the proposed method are fairly accurate, but for Japanese NS, using the Chinese Onomatopoeia Dictionary did not produce convincing translations.

In Experiment 1, both native Chinese and Japanese speakers were able to effectively understand the translation results overall. This may be because the differences between the methods are small. For example, if simply using a bilingual dictionary without similarity calculation, words that do not exist in the dictionary cannot be translated. If the Japanese onomatopoeia is simply transcribed into similar-sounding Chinese characters using the IPA, many words in the output would be unnatural or completely unreadable due to cultural and linguistic differences. However, since the words in the bilingual dictionary are all manually translated and the translation accuracy is fairly high, we decided to use the bilingual dictionary first for our method. Because the phonetic translation method is likely to produce out-of-context vocabulary, it was the least prioritized. The direct phonetic translation method ensures output when the previous methods fail to produce appropriate output. Therefore, in terms of the overall results, the difference between the four methods is negligible.

### Experiment 2

We also evaluated our method for sound-based translation of Japanese onomatopoeia on a five-point scale. We used 30 words that were tested in previous research (Han et al., 2019). The participants were 10 Chinese NS with JLPT N1 proficiency. We randomized the questions so that the participant could not identify which method was used for the translation. As the results in Table 5 show, our method achieved higher scores than that of the previous research.

**Table 6.** Comparison of evaluation scores between our method and previous research.

| | Bad | Poor | Normal | Good | Excellent | Average |
|---|---|---|---|---|---|---|
| Sound-based Translation (Han et al., 2019) | 0.05 (15/300) | 0.01 (4/300) | 0.03 (8/300) | 0.29 (86/300) | 0.62 (187/300) | 4.42 |
| Proposed Method | 0.03 (8/300) | 0.01 (2/300) | 0.01 (3/300) | 0.22 (65/300) | 0.74 (222/300) | 4.64 |

Sound-based translation of Japanese onomatopoeia presents some significant limitations. For example, the Japanese onomatopoeia "チリンチリン" (chirinchirin) is used to imitate the sound of a bicycle bell or bells ringing. When the sound-based method is used to translate the word into Chinese, the result is "其鈴鈴" (qi ling ling), which is meaningless for Chinese NS. In contrast, the proposed method produces the translation "叮鈴鈴" (ding ling ling), which is the exact sound of a bicycle bell expressed in Chinese. Another example is "哗啦啦" (hua la la), which means rustling or flowing smoothly in Chinese. The sound-based method produces "ファララ" (fa ra ra), which is meaningless for Japanese NS, and the correct answer is "サラサラ" (sarasara).

## CONCLUSION

We have constructed a Japanese-Chinese onomatopoeia corpus and proposed a method to translate Japanese onomatopoeia into Chinese using related words that make the same sounds or indicate events or behaviors expressed by the onomatopoeia. The most appropriate words are determined by calculating the cosine similarity and Levenshtein distance. If the answer is not in our dataset, the IPA is used to transcribe Japanese characters into their closest Chinese phonetic counterparts. Our ultimate goal is to extend this method to multiple language models and translate onomatopoeia in manga in various languages.

The present study only used related words to help translate the onomatopoeia. In the future, we aim to introduce a wider range of contexts, such as character dialogues or descriptions of scenes in manga, to assist in the translation of onomatopoeia. Furthermore, the present method requires users to input words to use the system. We intend to make it easier to use by automating this input process.

## ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers JP22K00646.

## REFERENCES

Chen Yan, Nanae Shirozu, Mitsunori Matsushita. A Survey of Onomatopoeia in Japanese Comics Created for Chinese Speakers. Faculty of Informatics, Kansai University, Graduate School of Informatics, Kansai University. (2013) (In Japanese).

Chisato Harada, Kosuke Yamazaki, Ailin Meng, Wenyu Zhang, Shiho Nobesawa. Automatic Selection of Japanese-Chinese Translation for Onomatopoeia Phrases. (2017) (In Japanese).

Kai Yang, Tsuyoshi Nakamura, Masayoshi Kanoh, Koji Yamada. Proposal of Machine Translation from Japanese to Chinese Onomatopoeias. (2018) (In Japanese).

Kayo Kawasaki. Nihon no comics no Supeingoyaku niokeru Giongo Gitaigo no Hyogenbunseki. (In Japanese) Lingua, 105–123 (2006).

Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. Building a Manga Dataset "Manga109" with Annotations for Multimedia Applications. IEEE MultiMedia, 27(2), 8–18 (2020). doi: 10.1109/mmul.2020.2987895.

Mai Muramoto. Doitsugo no Onomatope nikansuru Ichikosatsu. (In Japanese) Graduate School of Humanities and Social Sciences Studies in Humanities and Cultures, No. 6, 151–170 (2006).

Ren-feng Hou, Miho Matsuo. A study on Chinese translation of onomatopoeia in manga. (2019) (In Japanese).

Seiji Tsuchiya, Motoyuki Suzuki, Fuji Ren, Hirokazu Watabe. A Novel Estimation Method of Onomatopoeic Word's Feeling based on Mora Sequence Patterns and Feeling Vectors. (In Japanese) Journal of Natural Language Processing, 19(5), 367–379 (2012).

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space." In: Proceedings of the International Conference on Learning Representations (2013). URL: https://api.semanticscholar.org/CorpusID:5959482.

Yuichiro Shimizu, Maki Sakamoto. Creation Support System for Japanese Onomatopoeia based on Sound Symbolism. (In Japanese) The Japanese Society for Artificial Intelligence (2012).

Xinli Zhang. Patterns of Chinese translations of Japanese onomatopoeia. (2010) (In Japanese).

Yihong Han, Yoko Nishihara, and Ryosuke Yamanishi. A Method for Converting Onomatopoeic Words in Japanese Comicsinto Chinese Based on International Phonetic Alphabet. Procedia Computer Science, 159, 850–859 (2019).

Yating Zhang, Adam Jatowt, and Katsumi Tanaka. Is Tofu the Cheese of Asia? Searching for Corresponding Objects across Geographical Areas. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 1033–1042 (2017).