Japanese Rule-based Grapheme-to-phoneme Conversion System and Multilingual Named Entity Dataset with International Phonetic Alphabet

Yuhi Matogawa[r], Yusuke Sakai[r], Taro Watanabe[r], Chihiro Taguchi[I]
[r]Nara Institute of Science and Technology, [I]University of Notre Dame
{matogawa.yuhi.na2, sakai.yusuke.sr9, taro}@is.naist.jp, ctaguchi@nd.edu

Abstract

In Japanese, loanwords are primarily written in Katakana, a syllabic writing system, based on their pronunciation. However, the transliterated loanwords often exhibit spelling variations, such as the word "Hepburn" being written as "ヘボン (hebon)", "ヘプバーン (hepubaan)", "ヘップバーン (heppubaan)". These orthographical variants pose a bottleneck in multilingual Named Entity Recognition (NER), because named entities (NEs) do not have one-to-one matches. In this study, we introduce a rule-based grapheme-to-phoneme (G2P) system for Japanese based on literature in linguistics and a large-scale multilingual NE dataset with annotations of the International Phonetic Alphabet (IPA), focusing on IPA to address the Katakana spelling variations in loanwords. These rules and dataset are expected to be beneficial for tasks such as NE aggregation, G2P system, construction of cross-lingual language models, and entity linking. We hope our work advances research on Japanese NER with multilingual loanwords by solving the spelling ambiguities[1].

1 Introduction

Japanese orthography consists of three unique writing systems: Hiragana, Katakana, and Kanji. Among these, Katakana is mainly used for transliteration of loanwords originating from languages outside the Sinosphere, in particular, for named entities (NEs) such as proper nouns. However, there are no clear rules for this transliteration process, and NEs are transliterated into the closest Katakana based on the pronunciation of the source language. While some loanwords close to the Japanese pronunciation are often transliterated into Katakana representing mostly similar sounds (e.g., "Obama" to "オバマ (obama)"), this is not the case for other loanwords, leading to ambiguities due to the lack of unified common transliteration (e.g., "Hepburn" to "ヘボン (hebon)", "ヘプバーン (hepubaan)", and "ヘップバーン (heppubaan)"). More details are described in Appendix A.

This ambiguity poses challenges in unifying Katakana-written loanword NEs within Japanese and identifying the original NEs. This issue stems from significant differences in phonological and writing systems between Japanese and other languages, especially in historical documents written when the reading of foreign words was not customary among general readers, resulting in a wide variety of transliterations based on the sound perception of individuals without established transliteration rules. Furthermore, the inconsistent transliteration of loanword NEs in Japanese can be an obstacle for Japanese learners.

Given these issues, we first developed a rule-based Japanese Grapheme-to-Phoneme (G2P) system. Although Katakana can be mapped into IPA by rule-based conversion, the development is challenging because constructing precise rules requires knowledge of linguistics and phonetics. Currently, only neural-based approaches support Japanese G2P. Our rule-based G2P system, founded on linguistic principles, accurately represents pronunciation even in cases where IPA is automatically extracted from a source whose phonetic accuracy is not guaranteed. We hope that our G2P system can serve as a useful learning aid for Japanese learners and is expected to be applicable to pronunciation-based Japanese text analysis methods, such as NER and entity linking.

Next, we constructed a large-scale multilingual NE dataset with IPA annotations to address the spelling variations such as Katakana spellings of loanwords in Japanese. This dataset contains over 69 million pairs of NE and IPA, and over 14 million

[1]The original code and dataset are available from https://github.com/lart-rt/Japanese_ipa_rule_and_NE_dataset. Moreover, our work is merged with Epitran (Mortensen et al., 2018), a widely-used G2P library: https://github.com/dmort27/epitran/pull/143. You can access the demo site on https://yusuke1997.com/Japanese_G2P.

Proceedings of the 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 77–86

IDs used to identify NEs from 68 languages. On average, each ID is associated with five different pairs of NE and IPA. Ours is the largest dataset among the multilingual NE datasets with phonetic annotation. We hope our work advances research dealing with the spelling variations in Japanese, including approaches of cross-lingual word alignments such as Knight and Graehl (1997); Ren (2023), which utilizes phoneme sequences (though non-IPA) as an intermediate to align English and Japanese words.

To summarize, our contributions are as follows:

- We developed a rule-based G2P conversion system from the Japanese linguistic literature.
- We constructed a large-scale multilingual NE dataset with IPA annotations considering the transliteration ambiguity.

2 Related Work

G2P Conversion Systems. G2P conversion systems are mainly classified into two types: rule-based (Pine et al., 2022; Sar and Tan, 2019; Kłosowski, 2022; Deri and Knight, 2016; Wang and Tsai, 2009; Alam et al., 2011; Narasimhan et al., 2004) and neural-based (Li et al., 2022; Yamasaki, 2022; Peters et al., 2017; Arora et al., 2020) / machine learning-based (Rama et al., 2009; Laurent et al., 2009; Kienappel and Kneser, 2001) approaches. The neural-based approaches achieve high performance for high-resource languages, but the performance is significantly degraded when the quality of phonemic representation in the training data is not ensured or when the dataset size is small (Clematide and Makarov, 2021). On the other hand, rule-based approaches can easily obtain accurate G2P results and are faster than neural-based approaches, when rules are built on correct pronunciation based on linguistic features.

Japanese G2P. There have been attempts to develop Japanese systems for G2P or grapheme-phoneme alignment using neural-based (Makarov and Clematide, 2020; Clematide and Makarov, 2021; ElSaadany and Suter, 2020; Vesik et al., 2020) and machine learning-based approaches (Waxmonsky and Reddy, 2012; Bhargava and Kondrak, 2011; Baldwin and Tanaka, 1999; Nagata, 2000). The systems are applied to tasks such as estimating pronunciation in Japanese (Hatori and Suzuki, 2011; Yencken and Baldwin, 2005) and transliterating named entities (Tsuji et al., 2012; Bilac and Tanaka, 2004; Yamashita et al., 2018; Ren, 2023). However, these systems require training data, which are often not from sources of ensured quality, bearing the possibility of predicting incorrect IPA[2]. They are not based on linguistic insights, and although they have achieved success in some tasks of natural language processing, such a problem remains that they do not reflect truly correct pronunciation from the perspective of linguistics and phonetics. For rule-based approaches (Bilac et al., 1999; Shiga and Kawai, 2012; Terada and Lee, 2017; Masuda and Umemura, 1997; Sagisaka and Sato, 1983), the rules for G2P or grapheme-phoneme alignment that reflect accurate pronunciation based on literature about Japanese phonetics and phonology have not been published in any academic paper or presentation yet. The difficulty of Japanese G2P can be attributed not only to the complexity of its writing systems, which employ the three different systems (i.e., Hiragana, Katakana, and Kanji), but also to the syllabic characteristics of Hiragana and Katakana. More details are described in Appendix B.

Datasets for IPA. There are several multilingual datasets that match NE sequences in their original language with their corresponding IPA representations. However, they are not necessarily suitable for downstream tasks such as NER assumed in this study. WikiPron, for example, includes IPA from non-NE entries and is not readily applicable for solving tasks related to NEs. Klumpp et al. (2022) adds annotation of IPA to speech in six languages, but it still does not suit our purpose because it is also not specific to NE.

3 Building the Japanese Rule-based G2P System

3.1 Creating the rules

We created the G2P rules that reflect the description of Japanese phonetics and phonology (NKG, 2005; Saito, 2006). Specifically, our system was developed based on the chapter on phonetics and phonology in (NKG, 2005), the encyclopedia of

[2]For example, WikiPron (Lee et al., 2020): https://github.com/CUNY-CL/wikipron, the data regarded as gold in Ashby et al. (2021), is the corpus which comprises pairs of graphemes and IPA automatically extracted from online dictionary "Wiktionary": https://www.wiktionary.org. However, the correctness and consistency of these IPA representations are not guaranteed because these IPA are manually annotated by Wiktionary users, including non-experts of unknown academic backgrounds.

Japanese language education. In addition, we referred to the phonetic description in (Saito, 2006) to take into account some peculiar phonetic realizations. Additionally, we also created a simpler version of G2P rules based on other references. Upon implementing the G2P rules, the format conforms to the notation of the existing multilingual G2P framework Epitran [3] (Mortensen et al., 2018). Our work is the first contribution to G2P conversion for a language with syllabary scripts in the framework of Epitran.

The mechanism of Epitran Epitran has three types of conversion rules: "map", which exists for all languages, and "pre" and "post", which are optional to some languages. The basic conversion is done by "map", which is a one-to-one correspondence between the letters of each language and the IPA symbols, but "pre" and/or "post" are applied before and/or after "map", respectively, as needed to handle phenomena that cannot be handled by the one-to-one correspondence, such as when the pronunciation changes depending on the environment of the preceding and/or following sounds. If necessary, "pre" and/or "post" are applied before and/or after "map" is applied, respectively. For example, German "ö" is basically converted to [ø] according to "map", though to [œ] when two or more consonants follow it due to "post".

Among the three types of rules in Epitran, we created "map" and "post" for Japanese because all the pronunciation mappings can be done by these two in the language. In other words, we created "map" for each combination of Katakana / Hiragana and an IPA symbol, and "post" to deal with phenomena that cannot be handled by "map". Table 1 shows the examples of both in the detailed version of the rules.

Detailed version of rules In our work, the criteria for IPA granularity prioritize phonetic accuracy over phonemic representation, without exceeding what is necessary. While phonetically accurate transcription is important, perfectly phonetic representation can be not only redundant but also impractical, since we do not have spoken data. For our purpose, the description in NKG (2005) meets these criteria. It was originally published for Japanese language education and includes linguistically precise descriptions of various aspects of the Japanese language including phonetics and phonology, which we mainly referred to in this study.

[3] https://github.com/dmort27/epitran

Type	Conversion rules	Conditions
map	ラ ⇒ ɾa	–
post	ɾ ⇒ ɖ / # _	if word-initial

Table 1: Examples of "map" and "post" in Japanese G2P rule.

We created "map" mostly based on basic mappings between Katakana and IPA described in NKG (2005). However, the pronunciation of characters can vary depending on their surrounding environment. For such cases, we incorporated such phonetic variations in "post", drawing from the description of phonetics and phonology of Japanese given in the literature. To cover the phonetic rules comprehensively, we also referred to the other work Saito (2006), which provide a more detailed description of Japanese phonetics than NKG (2005). For instance, the moraic nasal /N/ (Katakana: "ン") has different phonetic realizations [m], [n], or [ŋ] depending on the articulation of the following consonant. We describe this phenomenon using "post". Namely, we write the rules to update [ɴ] (the tentative IPA symbol for /N/ in "map") to either [m], [n], or [ŋ] conditioned by its succeeding sound.

After creating "map" and "post" for Katakana, we created the rules for Hiragana by converting Katakana to Hiragana. This is simply because Katakana and Hiragana have a complete one-to-one correspondence with each other.

Simplified rules In addition to the fine-grained rules for the mapping from Katakana to IPA, we prepared a set of simplified G2P rules based on the articles related to Japanese writing systems and phonology in the English Wikipedia.[4] Specifically, the number of the more accurate rules of "map" is 150 while for the simplified rules 112. Also for "post", the more accurate rules comprise 46 whereas the more simplified rules are 20.

The simplified rules are more phonemic and phonetically less fine-grained than the detailed rules. However, creating the simplified version allows users to have multiple choices; for example, a user

[4] "Katakana - Wikipedia" (https://en.wikipedia.org/w/index.php?title=Katakana&oldid=1103341275, viewed in 2022 August) and "Sokuon - Wikipedia" (https://en.wikipedia.org/w/index.php?title=Sokuon&oldid=1096454475, viewed in 2022 August. "sokuon" means the first part of a geminated consonant).

may only want a reduced system for IPA without phonetic details. The simplified version also includes the mapping rules for both Katakana and Hiragana.

3.2 Evaluation

We evaluate the G2P conversion with our rules in comparison to the one by WikiPron. WikiPron is constructed by automatic crawling from Wiktionary and thus the quality of IPA conversion is not ensured. Some of them seem linguistically incorrect and different from the actual pronunciation. On the other hand, our rules can reflect correct pronunciation even in these cases because our rules are fully based on the literature in linguistics and phonetics. We compare IPA in WikiPron to IPA converted by our rules and show our rules are more preferable when IPA in WikiPron is wrong.

We used 2,348 Katakana–IPA pairs in WikiPron and compared WikiPron's IPA to IPA converted by our rules from Katakana in the dataset. We show the patterns of differences between WikiPron and ours in Table 2. As shown in (a) and (b) in the table, WikiPron incorrectly represents pronunciation for more than a quarter of the words. For instance, word-initial /r/ is transcribed as [ɾ] in WikiPron, though it is inappropriate according to (Saito, 2006). In contrast, ours converts word-initial /r/ to [ɖ], as pointed out in the literature, representing the appropriate pronunciation. We cannot judge which is more correct between WikiPron and ours in pattern (c), while in (d) ours are wrong. However, note that the proportion of (a) and (b), the pattern where WikiPron's is wrong, is much higher than pattern (d) where ours are inappropriate. Moreover, the cases in (d) are only limited to either of the following, both of which are highly exceptional in Japanese:

- When the word includes a less frequent or seldom used mora. These moras are not native to Japanese phonology but originate from foreign languages (e.g. "ヴャ (vya)").
- When the word is written in the archaic orthography that is no longer used in modern Japanese. In addition to the example in Table 2, an interesting instance is "シャッター". The literal pronunciation is "shiyattaa", though it is actually pronounced as "shattaa". The word is now almost totally replaced with "シャッター", which Japanese speakers also read as "shattaa".

		Examples		
Pattern	# notations	Katakana	WikiPron	Ours
(a)	673 (28.7%)	ラム	[ramɯ]	[ɖamɯ]
(b)	5 (0.213%)	ユータナジー	[oitanaʑiː]	[juːtanaʑiː]
(c-1)	1679 (71.5%)	ディスコ	[dʲisɯko]	[disɯko]
(c-2)	528 (22.5%)	ベンチ	[bẽɲtɕi]	[bentɕi]
(d)	36 (1.53%)	ヰ	[i]	"ヰ"

Table 2: The comparison of WikiPron and our rule-based system. (a) means some sounds represented in WikiPron are wrong according to the literature. (b) indicates IPA in WikiPron actually represents different Katakana from what is aligned with the IPA in the dataset, where the Katakana is a variant of the same word (the example in this table belongs to this type) or the Katakana is a completely different word. Pattern (c) refers to the cases where it is not clear whether the sound in WikiPron is wrong based on the description given in the literature. Among (c), WikiPron and ours differ in supplemental symbols in (c-1) and in main symbols in (c-2). Pattern (d) includes cases where some characters in Katakana are not supported in our rules. Note that the sum is not 100% since one sample can include multiple error patterns. For instance, IPA of "リンチ (rinchi)" falls into both of patterns (a) and (c).

4 The multilingual dataset of pairs of NE and IPA

4.1 Constructing the dataset

In addressing the challenge of Japanese NE variants, we also constructed a multilingual NE dataset. This dataset comprises pairs of NEs and their respective IPA representations, derived from the NE dataset ParaNames[5](Sälevä and Lignos, 2022). We achieved this by converting NEs of each language using Epitran for each language and Japanese NEs using the Japanese rules we introduced in Section 3. There are also other multilingual datasets of NEs such as "TRANSLIT" (Benites et al., 2020), but we chose ParaNames because it has the largest size and the widest coverage of languages. We created 69 million pairs in 68 languages by leveraging G2P rules in Epitran for each language, in which approximately 671K pairs were Japanese.

ParaNames entirely derives from the structured knowledge base of entries in Wikipedia. Specifically, NEs registered in Wikidata as instances of either "human", "geographic region", or "organization" are extracted for each language supported in Wikipedia. One set of data consists of "wikidata_id"

[5]https://github.com/bltlab/paranames

	Key	Value		
ParaNames	wikidata_id	Q19618413		
	type	LOCATION		
	language	English	Chinese	Japanese
	label	Paris	巴黎	パリ
+ Ours	*ipa*	**pɛɹɪs**	**pali**	**paɾji**

Table 3: An Example of ParaNames and our contributions

for the ID given in Wikidata, "label" for the notation of the NE, "language" for the language tag associated with the notation in "label", and "type" for the type of the NE. Appendix C describes more details about ParaNames entries.

We converted notations in "label" columns to IPA by Epitran for 68 languages and added IPA to the original data as shown in Table 3. The rules for most languages consist of some or all of "map", "pre", and "post" except English and Chinese, for which it is difficult to implement the one-to-one G2P mapping. For this reason, we leverage external pronunciation dictionaries for these two exceptional languages: "flite"[6] for English and "CC-CEDICT"[7] for Chinese.

4.2 Statistics of the dataset

Table 4 presents the overview of the statistics of our dataset. The pairs of NE–IPA amounts to more than 69 million and the number of IDs associated with pairs is over 14 million. This results in 4.964 pairs of NE-IPA per ID on average. The average number of pairs of notations per language tag and per language is 732K and 1.023 million, respectively. This size is much larger than the existing dataset with graphemes and phonemes of WikiPron, which has only about 14K pairs per language. This suggests the effectiveness of this dataset when applied to linking or alignment between different notations of the same NE. Therefore, the dataset we constructed in this study is distinguished from the multilingual datasets with IPA in the previous studies in that not only it is specific to NEs but also the amount of data per language is much greater than that of existing datasets.

The type of NE with the highest frequency is PER (person). The average sequence lengths of the original NE and IPA are 15.085 and 15.894, respectively. The number of notations per language tag are provided in Appendix D.

Measurement	Value
Total number of notations	69,573,951
– PER	48,625,240
– LOC	13,905,603
– ORG	7,043,108
Total number of IDs	14,016,907
– PER	8,897,440
– LOC	3,464,982
– ORG	1,654,485
Number of language tags	95
Number of actual languages	68
Average character length of NE notations	15.085
Average character length of IPA sequences	15.894

Table 4: Key statistics of our dataset.

The number of writing systems of original NE is 20 in total[8]. 15 languages have more than one million pairs of NE-IPA, approximately over average per language, for each. All of them use Latin script except Russian and Chinese while all of these belong to Indo-European except Hungarian and Chinese.

5 Conclusion

In this paper, we introduced the new G2P rules for Japanese based on the literature in linguistics and phonetics, and constructed the largest multilingual dataset of NE with IPA using rule-based G2P including our own rules for Japanese. These resources will be beneficial for solving NE-related tasks such as NER and entity linking in Japanese. The actual application of our G2P tool and dataset to downstream tasks on Japanese NEs like transliteration, text-to-speech, and so on is left for future work.

6 Limitations

G2P conversion system While our research primarily focused on the development of G2P for Katakana, we have also made it compatible with Hiragana and Kanji as a prototype, allowing for the input of any Japanese text. However, the complexity of Kanji readings far exceeds our initial estimations, making full support a future challenge.

[6] "flite: A small fast portable speech synthesis system" (https://github.com/festvox/flite, viewed in 2023 January).

[7] "CC-CEDICT Home [CC-CEDICT WIKI]" (https://cc-cedict.org/wiki/#what_is_cc-cedict, viewed in 2023 January).

[8] Latin, Ge'ez, Arabic, Cyrillic, Bengali, Devanagari, Katakana, Hiragana, Khmer, Rao, Malayalam, Burmese, Oriya, Gurmukhi, Sinhalese, Tamil, Telugu, Thai, simplified Chinese, and traditional Chinese.

Nonetheless, we have confirmed that G2P conversion is possible with a general level of accuracy.

Dataset In this paper, our contributions are the development of the NE dataset with IPA annotations and do not include experiments in downstream tasks. However, applying NE datasets with IPA annotations to downstream tasks has been reported in recent studies (Hentona et al., 2022), and we intend to apply ours to downstream tasks in future work. Additionally, the development of our large-scale dataset, comprising over 69 million NEs with associated IPA representations, demanded a significant investment of computational resources and time. We used a total of 96 CPU cores of Intel(R) Xeon(R) Platinum 8160 CPU @ 2.10GHz and 384GB RAM, taking nearly two months to complete annotating IPA. The availability of the large-scale dataset enables rapid experimentation, rendering it a highly valuable resource for advancing future research.

7 Impact

7.1 Effectiveness of our Japanese G2P conversion system

Our rule-based G2P conversion system is based on linguistic literature on Japanese, as mentioned in Section 3, allowing it to perform accurate G2P transliteration. Furthermore, a rule-based conversion system enables faster transliteration than neural-based systems. Furthermore, as shown in Figure 1, we have launched a demonstration site that supports both PC and mobile environments to allow anyone to easily use our system. This effort is groundbreaking because it is not supported by existing G2P systems such as Epitran. Our demonstration site also supports most of Japanese characters, including Kanji, Katakana, and Hiragana. We hope that it will be utilized as a tool for learners of Japanese to accurately predict pronunciations. We plan the demonstration site to be supported long-term, with plans including OCR and mobile-native support, as well as expansion to other languages in the future.

7.2 Effectiveness of our dataset

When comparing the other datasets containing IPA or writing systems such as Katakana or Latin script, our dataset has mainly two benefits.

Phonetic accuracy. IPA in this dataset takes into account more accurate pronunciation than the one in WikiPron and simple romanization, since the rules were based on the specialized literature in phonetics. For example, "ラザフォード (razafoodo)" is transcribed as [ɖazafoːdo] in our system unlike inappropriate [razafoːdo] in WikiPron-like manner.

Figure 1: The screenshot of our Japanese G2P demonstration site. Index 0 in the results is written using Katakana, index 1 is written using Hiragana, and index 2 is written using Kanji characters. We input some Japanese words in the text input area, then output each IPA as G2P results. We mainly support Katakana, but any Japanese characters are accepted. We support both PC and mobile environments and continue to improve and ensure long-term support, so we hope our G2P site helps non-native or Japanese learners to know the pronunciation of Japanese words.

Smaller edit distance. IPA can contribute to identifying different notations of Katakana for one ID as the same entity with less cost than using the original Katakana. Given one entity with two different forms "ラザフォード" and "ラザホード (razahoodo)", the edit distance between them in IPA is only 1 ([ɖazafoːdo] for the former and [ɖazahoːdo] for the latter) while the edit distance for Katakana is 2.

This multilingual dataset includes not only NEs widely acknowledged in Japanese (e.g. "Paris", "Madrid", etc.) but also NEs hardly used in Japanese. Thus, our dataset can link or align NEs in Katakana even when they are rare words.

References

Firoj Alam, S.M. Murtoza Habib, and Mumit Khan. 2011. Bangla text to speech using festival. In *Proceedings of Conference on Human Language Technology for Development*, pages 154–161. Bibliotheca Alexandrina.

Aryaman Arora, Luke Gessler, and Nathan Schneider. 2020. Supervised grapheme-to-phoneme conversion of orthographic schwas in Hindi and Punjabi. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7791–7795, Online. Association for Computational Linguistics.

Lucas F.E. Ashby, Travis M. Bartley, Simon Clematide, Luca Del Signore, Cameron Gibson, Kyle Gorman, Yeonju Lee-Sikka, Peter Makarov, Aidan Malanoski, Sean Miller, Omar Ortiz, Reuben Raff, Arundhati Sengupta, Bora Seo, Yulia Spektor, and Winnie Yan. 2021. Results of the second SIGMORPHON shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 115–125, Online. Association for Computational Linguistics.

Timothy Baldwin and Hozumi Tanaka. 1999. The applications of unsupervised learning to Japanese grapheme-phoneme alignment. In *Unsupervised Learning in Natural Language Processing*.

Fernando Benites, Gilbert François Duivesteijn, Pius von Däniken, and Mark Cieliebak. 2020. TRANSLIT: A large-scale name transliteration resource. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3265–3271, Marseille, France. European Language Resources Association.

Aditya Bhargava and Grzegorz Kondrak. 2011. How do you pronounce your name? improving G2P with transliterations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 399–408, Portland, Oregon, USA. Association for Computational Linguistics.

Slaven Bilac, Timothy Baldwin, and Hozumi Tanaka. 1999. Incremental japanese grapheme-phoneme alignment (日本語における漸進型書記素・音素アラインメント). *The Special Interest Group Technical Reports of IPSJ. Technical Reports of NL, IPSJ Natural Language Processing*, 130:9–16.

Slaven Bilac and Hozumi Tanaka. 2004. A hybrid back-transliteration system for Japanese. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 597–603, Geneva, Switzerland. COLING.

Simon Clematide and Peter Makarov. 2021. CLUZH at SIGMORPHON 2021 shared task on multilingual grapheme-to-phoneme conversion: Variations on a baseline. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 148–153, Online. Association for Computational Linguistics.

Aliya Deri and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 399–408, Berlin, Germany. Association for Computational Linguistics.

Omnia ElSaadany and Benjamin Suter. 2020. Grapheme-to-phoneme conversion with a multilingual transformer model. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 85–89, Online. Association for Computational Linguistics.

Jun Hatori and Hisami Suzuki. 2011. Japanese pronunciation prediction as phrasal statistical machine translation. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 120–128, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Asahi Hentona, Takamichi Toda, Yuta Tomomatsu, Masakazu Sugiyama, Yuki Azuma, and Sho Shimoyama. 2022. Unsupervised entity linking based on word alignment considering similarity about word embeddings phoneme sequences (単語の分散表現および音素列の類似性を考慮した単語アラインメントに基づく教師なしEntity Linking). In *Proceedings of the 28th Annual Conference of the Association for Natural Language Processing*, pages 1568–1572. Association for Natural Language Processing. In Japanese.

Anne K. Kienappel and Reinhard Kneser. 2001. Designing very compact decision trees for grapheme-to-phoneme transcription. In *Proceedings of 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pages 1911–1914. International Speech Communication Association.

Philipp Klumpp, Tomas Arias, Paula Andrea Pérez-Toro, Elmar Noeth, and Juan Orozco-Arroyave. 2022. Common phone: A multilingual dataset for robust acoustic modelling. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 763–768, Marseille, France. European Language Resources Association.

Kevin Knight and Jonathan Graehl. 1997. Machine transliteration. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 128–135, Madrid, Spain. Association for Computational Linguistics.

Piotr Kłosowski. 2022. A rule-based grapheme-to-phoneme conversion system. *Applied Sciences*, 12(5).

Antoine Laurent, Paul Deléglise, and Sylvain Meignier. 2009. Grapheme to phoneme conversion using an

smt system. In *Proceedings of Interspeech 2009*, pages 708–711. International Speech Communication Association.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

Xinjian Li, Florian Metze, David Mortensen, Shinji Watanabe, and Alan Black. 2022. Zero-shot learning for grapheme to phoneme conversion with language ensemble. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2106–2115, Dublin, Ireland. Association for Computational Linguistics.

Peter Makarov and Simon Clematide. 2020. CLUZH at SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 171–176, Online. Association for Computational Linguistics.

Keiko Masuda and Kyoji Umemura. 1997. Extracting kana - alphabet rules from a non - japanese name reading table (人名辞書から名前読み付与規則を抽出する試み). *The Special Interest Group Technical Reports of IPSJ. Technical Reports of NL, IPSJ Natural Language Processing*, 69:97–102. In Japanese.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Masaaki Nagata. 2000. Synchronous morphological analysis of grapheme and phoneme for Japanese OCR. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 384–391, Hong Kong. Association for Computational Linguistics.

Bhuvana Narasimhan, Richard Sproat, and George Kiraz. 2004. Schwa-deletion in hindi text-to-speech synthesis. *International Journal of Speech Technology*, 7:319–333.

Nihongo Kyoiku Gakkai NKG. 2005. *Encyclopedia of Japanese Language Education (new edition)*. Taishukan Shoten. In Japanese.

Ben Peters, Jon Dehdari, and Josef van Genabith. 2017. Massively multilingual neural grapheme-to-phoneme conversion. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 19–26, Copenhagen, Denmark. Association for Computational Linguistics.

Aidan Pine, Patrick William Littell, Eric Joanis, David Huggins-Daines, Christopher Cox, Fineen Davis, Eddie Antonio Santos, Shankhalika Srikanth, Delasie Torkornoo, and Sabrina Yu. 2022. G_i2P_i rule-based, index-preserving grapheme-to-phoneme transformations. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 52–60, Dublin, Ireland. Association for Computational Linguistics.

Taraka Rama, Anil Kumar Singh, and Sudheer Kolachina. 2009. Modeling letter-to-phoneme conversion as a phrase based statistical machine translation problem with Minimum Error Rate training. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium*, pages 90–95, Boulder, Colorado. Association for Computational Linguistics.

Yuying Ren. 2023. Back-transliteration of English loanwords in Japanese. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, pages 43–49, Toronto, Canada. Association for Computational Linguistics.

Yoshinori Sagisaka and Hirokazu Sato. 1983. Accentuation rules for japanese word concatenation (日本語単語連鎖のアクセント規則). *The IEICE Transactions*, J66-D:849–856. In Japanese.

Yoshio Saito. 2006. *Introduction to Japanese Phonetics (revised)*. Sanseido. In Japanese.

Jonne Sälevä and Constantine Lignos. 2022. ParaNames: A massively multilingual entity name corpus. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 103–105, Seattle, Washington. Association for Computational Linguistics.

Vathnak Sar and Tien-Ping Tan. 2019. Applying linguistic g2p knowledge on a statistical grapheme-to-phoneme conversion in khmer. *Procedia Computer Science*, 161:415–423. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia.

Yoshinori Shiga and Hisashi Kawai. 2012. Multilingual speech synthesis system. *Journal of the National Institute of Information and Communications Technology*, 59:21–28.

Takuya Terada and Akinobu Lee. 2017. Automatic construction of a robust pronunciation dictionary for spoken language using statistical learning by g2p in japanese (日本語におけるG2Pによる統計的学習を用いた話し言葉に頑健な発音辞書の自動構築). *The Special Interest Group Technical Reports of IPSJ. Technical Reports of SLP, IPSJ Spoken Language Processing*, 11:1–6. In Japanese.

Rieko Tsuji, Yoshinori Nemoto, Wimvipa Luangpiensamut, Yuji Abe, Takeshi Kimura, Kanako Komiya, Koji Fujimoto, and Yoshiyuki Kotani. 2012. The transliteration from alphabet queries to Japanese product names. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 456–462, Bali, Indonesia. Faculty of Computer Science, Universitas Indonesia.

Kaili Vesik, Muhammad Abdul-Mageed, and Miikka Silfverberg. 2020. One model to pronounce them all: Multilingual grapheme-to-phoneme conversion with a transformer ensemble. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 146–152, Online. Association for Computational Linguistics.

Yu-Chun Wang and Richard Tzong-Han Tsai. 2009. Rule-based Korean grapheme to phoneme conversion using sound patterns. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2*, pages 843–850, Hong Kong. City University of Hong Kong.

Sonjia Waxmonsky and Sravana Reddy. 2012. G2P conversion of proper names using word origin information. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 367–371, Montréal, Canada. Association for Computational Linguistics.

Tomohiro Yamasaki. 2022. Grapheme-to-phoneme conversion for Thai using neural regression models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4251–4255, Seattle, United States. Association for Computational Linguistics.

Michiharu Yamashita, Hideki Awashima, and Hidekazu Oiwa. 2018. A comparison of entity matching methods between English and Japanese katakana. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 84–92, Brussels, Belgium. Association for Computational Linguistics.

Lars Yencken and Timothy Baldwin. 2005. Efficient grapheme-phoneme alignment for Japanese. In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 143–151, Sydney, Australia.

A Background of Loanword in Japanese and the other Non-Latin Alphabet Writing System Languages

In languages with non-Latin-alphabet writing systems like Chinese or Korean, loanwords are often written using a mixture of their native script and the Latin characters, known as code-switching. However, in Japanese, loanwords are almost always transliterated into Katakana based on their pronunciation, causing various transliterations for the same word caused by individual differences and preferences, especially for new or not yet standardized NEs.

B The difficulty of Japanese G2P

The difficulty of Japanese G2P can be attributed not only to the complexity of its writing system, which employs three different systems (Hiragana, Katakana, Kanji), but also to the syllabic characteristics of Hiragana and Katakana. Unlike alphabetical writing systems such as Latin and Cyrillic, the basic unit of Hiragana and Katakana is a syllable, not a phoneme. A syllable is a kind of phonological unit, usually composed of a core (nucleus) vowel and potentially consonants before and/or after the core. In Hiragana and Katakana, characters representing different syllables are completely different from each other, even when they share the same vowel or consonant. For instance, the Katakana character for /ka/ is “カ”, while /ki/, which shares the same consonant, is represented by a totally different form: “キ”.

C Details of ParaNames entries

One of “PER”, “LOC”, and “ORG” is assigned to “type”, corresponding to “human”, “geographic region”, and “organization”, respectively, in the type in the original Wikidata. Table 3 shows the example of the data in ParaNames. Note that the number of language tags stored in “language” does not agree with the actual number of languages in the dataset, since a language may have multiple tags reflecting differences in writing systems, regions, and so on.[9]

D The number of notations per language tag in our dataset

The number of writing systems of original NE is 20 in total[10]. 15 languages have more pairs of NE-IPA for each than the approximately average number of pairs per language, one million. All of them use a Latin-based script except Russian and Chinese and belong to the Indo-European language family except Hungarian and Chinese. Table 5 shows the number of notations per each language tag.

[9] For example, there are two tags for Uzbek: “uz-cyrl” for Uzbek written in Cyrillic and “uz-latn” for Uzbek in the Latin script.

[10] Latin, Ge’ez, Arabic, Cyrillic, Bengali, Devanagari, Katakana, Hiragana, Khmer, Lao, Malayalam, Burmese, Oriya, Gurmukhi, Sinhalese, Tamil, Telugu, Thai, Simplified Chinese, and Traditional Chinese.

language tag: language name	number of notation	language tag: language name	number of notation
aa: Afar	28,894	ny: Chewa	29,309
am: Amharic	4,478	om: Oromo	30,961
ar: Arabic	830,648	or: Oriya	15,045
av: Avar	1,155	pa: Punjabi	18,733
az: Azerbaijani	115,014	pl: Polish	1,526,678
bn: Bengali	437,838	pt: Portuguese	373,237
ca: Catalan	3,057,109	pt-br: Portuguese (Brazil)	1,897,670
cs: Czech	1,283,030	rn: Rundi	28,017
de: German	4,177,379	ro: Romanian	894,080
de-at: German (Austria)	295,193	ru: Russian	1,333,970
de-ch: German (Switzerland)	31,433	rw: Kinyarwanda	30,374
de-formal: German (formal)	34	sg: Sango	27,867
en: English	13,715,761	si: Sinhala	19,001
en-ca: English (Canada)	424,497	sn: Shona	29,662
en-gb: English (UK)	141,742	so: Somali	30,748
es: Spanish	6,071,612	sq: Albanian	2,855,562
es-419: Spanish (Latin America)	1,271	sv: Swedish	2,733,009
es-formal: Spanish (formal)	506	sw: Swahili	294,945
fa: Persian	630,954	ta: Tamil	89,757
ff: Fulah	30,456	te: Telugu	52,395
fr: French	5,003,611	tg: Tajik	76,492
ha: Hausa	46,317	tg-cyrl: Tajik (Cyrillic)	566
hi: Hindi	74,366	tg-latn: Tajik (Latin)	29,674
hr: Croatian	426,170	th: Thai	91,844
ht: Haitian	88,589	ti: Tigrinya	288
hu: Hungarian	1,056,422	tk: Turkmen	28,707
hu-formal: Hungarian (formal)	2	tl: Tagalog	172,360
id: Indonesian	774,023	tr: Turkish	586,329
it: Italian	3,096,623	ug: Uighur	3,125
ja: Japanese	671,429	ug-arab: Uighur (Arabic)	113
jv: Javanese	151,768	uk: Ukrainian	654,641
kk: Kazakh	58,089	ur: Urdu	149,481
kk-cyrl: Kazakh (Cyrillic)	47,204	uz: Uzbek	192
kk-kz: Kazakh (Kazakhstan)	761	uz-cyrl: Uzbek (Cyrillic)	1
kk-latn: Kazakh (Latin)	74,802	uz-latn: Uzbek (Latin)	7
kk-tr: Kazakh (Turkey)	25,389	vi: Vietnamese	568,179
km: Khmer	3,241	xh: Xhosa	32,628
ky: Kyrgyz	44,936	yo: Yoruba	274,206
lo: Lao	1,408	zh: Chinese	965,861
mi: Maori	51,825	zh-cn: Chinese (simplified)	16,178
ml: Malayalam	93,555	zh-hans: Chinese (simplified)	255,533
mn: Mongolian	12,020	zh-hant: Chinese (traditional)	9,508
mr: Marathi	41,352	zh-hk: Chinese (Hong Kong)	5,065
ms: Malay	545,598	zh-mo: Chinese (Macao)	305
mt: Maltese	68,109	zh-my: Chinese (Mandarin in Malaysia)	6
my: Burmese	9,331	zh-sg: Chinese (Mandarin in Singapore)	104
nl: Dutch	9,586,075	zh-tw: Chinese (Mandarin in Taiwan)	9,518
nl-informal: Dutch (informal)	5		

Table 5: The number of notations per each language tag.