ルールベース G2P による多言語固有表現の国際音声記号表記付きデータセットの構築

的川雄飛 [1]　坂井優介 [1]　平野颯 [1]　澤田悠治 [1]
大内啓樹 [1,2]　渡辺太郎 [1]
[1] 奈良先端科学技術大学院大学　[2] 理化学研究所
matogawa.yuhi.na2@is.naist.jp
{sakai.yusuke.sr9,hirano.hayate.hc2,sawada.yuya.sr7}@is.naist.jp
{hiroki.ouchi,taro}@is.naist.jp

概要

本研究では，多言語の固有表現（NE）をルールベースの grapheme-to-phoneme（G2P）によって国際音声記号（IPA）に変換した表記の大規模データセットを構築した．多様な文字・音韻体系からなる複数の言語の NE を IPA という一つの体系によって表記したデータセットは，NE を対象とした G2P や通言語の言語モデル構築，NE 表記のリンキングなどのタスクへの有益な応用を期待できる．本研究は，既存データセットの NE 表記をルールベース G2P のフレームワークを用いて IPA 列に変換し，IPA 表記付きのデータセットを構築した．また，これに伴い日本語から IPA への変換ルールを作成し，フレームワークに追加した．

1　はじめに

ある言語における固有表現（NE）を表すテキストとそれに対応する音声記号表記のデータは，テキストの発音に関する情報が求められる諸タスクに NE を対象として取り組む場合において必要となる．例えば，任意の文字体系で書かれた表記を国際音声記号（IPA）などの発音記号に変換する grapheme-to-phoneme（G2P）に，NE を対象としてニューラルベースの手法で取り組む場合，元の NE 表記と発音の関係を学習するためのデータが必要である．また，多様な言語・文字種の NE 表記を全て統一的な IPA に変換したデータが存在すれば，特定の言語に依らない非常に通言語性にすぐれた「IPA 言語モデル」の学習に利用できる．

これらに加え，特定の言語における NE 表記を正しい ID にリンキングさせるタスクにおいても，テキストと音声記号表記の対応データが必要な場合がある．例えば，他の言語の NE 表記と ID のペアが明示された情報が存在する場合，NE 表記を IPA 列に変換して，メインの研究対象である言語の NE 表記から変換された IPA 列と他言語の NE 表記から変換された IPA 列のあいだで音声的類似度を計算し[1)]，類似度が最も高い他言語の NE 表記と紐付いている ID をメインの研究対象である言語の NE 表記に紐付く ID とする，という手法が考えられる．この場合の音声的類似度の計算にも，やはり NE 表記と音声記号表記のマッピングのデータが必要となる．

元言語の表記と音声記号表記の対応データを公開した先行研究として，オンライン辞書 “Wiktionary” の記述に基づくデータセットを公開したものが挙げられる [2]．しかし，このデータセットは NE に特化したものではなく，上に挙げたような NE を対象とするタスク・研究における利用には適さない．

本研究では，既存の NE 表記データセットを用い，複数言語の NE 表記をルールベース G2P のフレームワークを用いて IPA 列に変換することによって，上記のような NE 関連のタスクに必要な IPA 付き NE 表記データセットを構築した．また，日本語については，このフレームワークにおけるサポートが存在しなかったため，言語学・音声学分野の文献を基に G2P ルールを新たに作成した．

この結果，68 言語について，NE 表記と対応する IPA 列の組数が 6,900 万以上にのぼるデータセットが完成した．このデータセットは，NE を対象とした研究における使用に適したものであるのみなら

1) 具体的な音声的類似度の計算方法としては，IPA 同士でナイーブな編集距離を計算する手法のほか，IPA の各シンボルを音声学的な素性の有無に基づくベクトルに変換してから類似度を計算する手法 [1] などが挙げられる．

表 1　ParaNames のデータの例

wikidata_id	label	language	type
Q19618413	Korshunikha	en	LOC

表 2　本研究において作成したデータの例

wikidata_id	label	language	type	**ipa**
Q19618413	Korshunikha	en	LOC	kɔɹʃunɪkhə

ず，1 言語あたりのデータの量についても先行研究の元言語表記-IPA 表記データセットを上回るものとなっている．

2　関連研究

既存研究で公開された特定の言語の表記と IPA 表記の対応データセットとして，“WikiPron”が挙げられる[2)] [2]．WikiPron は，オンラインの多言語辞書“Wiktionary”から，対象言語の表記とその IPA 表記の組を自動抽出して構築され，251 言語について 3,509,051 組が登録されている．しかし，その抽出元は辞書であるため，普通名詞・固有名詞や動詞，形容詞等に至るまで様々な種類の語が格納されており，1 章で述べたような特に NE をターゲットとした研究には適していない．

IPA への変換は行っていないが複数の言語の NE 表記と ID のペアを格納したデータセットとしては，“ParaNames”[3)] [3]，“TRANSLIT”[4)] [4] が存在する．両者は，言語のカバレッジや NE 表記数，ID 数のほか，データセットの作り方，データの形式などの面で違いがある．これらのうち，本研究は ParaNames を利用し，IPA 付き NE 表記の新たなデータセットを構築する．

468 の言語タグについて 124,343,696 個の NE 表記，14,017,168 個の ID を持つ ParaNames は，Wikipedia の項目を主に構造化した知識ベースである “Wikidata” から全てのデータが作られている．具体的には，Wikidata において “human”，“geographic region”，“organization”の各タイプのインスタンスとして登録されている NE 表記が，wikipedia でサポートされている各言語について抽出されている．1 組のデータは，Wikidata において付与されている ID を表す “wikidata_id”，NE の表記を表す “label”，“label”の表記に紐づく言語タグを表す “language”，NE のタイプを表す “type”によって構成されており，type には “PER”，“LOC”，“ORG”の値が，元の Wikidata におけるタイプ “human”，“geographic region”，“organization” それぞれに対応して割り当てられている（例：表 1）．

いっぽう，185 の言語タグについて 2,987,508 個の NE 表記，1,548,752 個の ID を持つ TRANSLIT は，独自に Wikipedia から収集した 1 種と既存の研究で公開されていた 4 種，計 5 つのデータセットが混合されて作られている．1 個の ID がキーとなり，それに対応する値として各言語の表記のリストが格納される形式となっている[5)]．

3　データセットの構築

本研究では，2 章に挙げた 2 つの NE 表記-ID のデータセットのうち ParaNames を用い，“label”に格納されている NE 表記をルールベース手法による G2P フレームワーク “Epitran”[6)] [5] によって IPA 列に変換して元の ParaNames のデータに加えることにより，IPA 付き NE 表記のデータセットを構築した（例：表 2）．研究開始時には Epitran でサポートされていなかった日本語については言語学・音声学分野の文献に基づいて自作した G2P ルールを，その他の言語については Epitran の既存の G2P ルールを用い，計 68 言語の NE 表記から IPA への変換作業を行った．

3.1　元データの選定

TRANSLIT ではなく ParaNames を用いてデータセットを作成した理由は，後者の方が言語タグ数，NE 表記数，ID 数いずれも前者よりも上回っているからのみならず，Wikidata という単一のソースに由来するデータセットを使用することでデータの質がより確保できる，と考えられるからである．複数のソースに由来する TRANSLIT では，データの質の面において，NE 表記と ID の紐付けが誤っていることがあるなどの問題が存在する．データの量の面で TRANSLIT を上回り，かつ一つのソースに由来する Paranames を基にデータを構築することで，量・質の両面でよりすぐれたデータセットを完成させることができると考えられる．

なお，ParaNames の “language”に格納されている言語タグの数は，実際の言語の数とは一致しない．1 つの言語につき，文字種などの違いや話される地

2）https://github.com/CUNY-CL/wikipron
3）https://github.com/bltlab/paranames
4）https://github.com/fbenites/TRANSLIT．ある文字体系を他の文字体系に変換するタスク「翻字（transliteration）」における利用のために作られたデータセットである．
5）簡易的な例を示す．　ID 001: [“en_name1”, “zh_name2”, ⋯]
6）https://github.com/dmort27/epitran

域の違いを反映して複数のタグが存在する場合があるためである[7)]．言語が同じだが言語タグが異なるデータであり，かつ文字種が同じデータについて，IPA 付きデータセットでは “wikidata_id”が重複しているデータを排除した．

さらに，ParaNames のうち，IPA に変換して新たなデータセットに組み込む対象は，95 個の言語タグのいずれかが付与されている計 68 言語のデータに絞った，これは，以下の 2 つの理由による．

1. 第一に，「Epitran がサポートする言語，あるいは日本語のいずれか」という制約を設けた．Epitran のサポート言語に絞った理由は，ルールベースの手法により容易に G2P を実行できる言語のデータを優先的に作成することが望ましいと考えられるからである．ただし，日本語のみ，Epitran では当初サポートされていなかったが本研究において新たにルールを作成し，そのルールを用いて IPA 変換を行った．これは，1 章で述べた応用先のうち，NE 表記を正しい ID にリンキングさせるタスクを例として想定し，さらにカタカナの NE 表記をリンキング実験の対象とすることを今後の研究の念頭に置いたためである．作成した日本語 G2P ルールについては，3.3 節で詳述する．
2. 第一の制約に加え，「TRANSLIT のデータ中で，少なくとも 1 つのカタカナ NE 表記と同じ ID に紐づいている表記が存在する言語」という制約を与えた．例えば，“ID 002: [“ja_{ カタカナ表記 }”, “en_…”, “fr_…”, …]”というデータが TRANSLIT に存在する場合，英語（en）とフランス語（fr）はいずれも本研究の対象に含まれる．この制約を課したのは，第一の制約の説明で述べたカタカナ NE 表記のリンキング実験の評価用データとして TRANSLIT を用いることを想定したためである．すなわち，任意のカタカナ NE 表記について，実際の TRANSLIT のデータ中で紐付いている ID を「正解 ID」とし，カタカナ NE 表記から変換された IPA 列と他言語の NE 表記から変換された IPA 列の音声的類似度を計算してリンキングを行う実験を想定している．

7） たとえば，ウズベク語については，キリル文字表記を表す “uz-cyrl”とラテン文字を表す “uz-latn”の 2 種類のタグが存在する．

表 3 日本語 G2P ルールの “map”, “post”の例

map	ラ，ɾa（「ラ」を [ɾa] に変換）
post	ɾ -> ɖ / # _（語頭の [ɾ] は [ɖ] に変換）

3.2 NE 表記から IPA 列への変換

“label”の NE 表記から IPA への変換は，すべて Epitran を用いて行った．Epitran は，各言語について用意された IPA への変換ルールに基づいて G2P を実現する，多言語対応のフレームワークであり，2023 年 1 月現在 97 言語についてサポートされている．変換ルールは，全言語において存在する “map”と，一部の言語にのみ存在する “pre”, “post”の 3 種類がある．基本的な変換は，各言語の文字（の組み合わせ）と IPA シンボル（の組み合わせ）が 1 対 1 に対応する map によって行うが，前後の音の環境により発音が変わる場合などの 1 対 1 対応では処理できない現象に対して，必要に応じ map の適用前に pre, 適用後に post がそれぞれ適用される．

また，表記と IPA の 1 対 1 対応によるマッピングが難しい言語については，外部の発音辞書が利用される．具体的には，英語については音声合成のためのソフトウェア “flite”[6] が，中国語については “CC-CEDICT”[7] が使用される．

3.3 日本語の G2P ルール作成

日本語からの G2P については，日本語の音声・音韻の概要を述べた複数の文献 [8, 9] を基に，第一著者が Epitran の記法に準拠してカタカナ・ひらがなからの G2P ルールを作成した．これは，研究を開始した時点で日本語が Epitran でサポートされておらず，また言語学的に正確な G2P ルールを発表した先行研究も存在しなかったためである．なお，漢字からの G2P は，文脈依存による発音の特定が必要であり対処が困難であるため，取り扱わなかった．

日本語について作成したルールは，3.2 節で挙げた 3 種のルールのうち，map と post である．すなわち，カタカナ・ひらがなの組み合わせ 1 通りと IPA シンボルの組み合わせ 1 通りの各対応を map として，それだけでは処理できない事象に対応するためのルールを post として作成した．両者の例を，表 3 に示す．

先に述べた日本語の音声・音韻に関する文献に基づくカタカナ・ひらがなからの G2P ルールのほか，英語版ウィキペディアにおける日本語の文字・音韻

~

表 4 作成したデータセットの統計概要

全表記数	69,573,951
PER	48,625,240
LOC	13,905,603
ORG	7,043,108
全 ID 数	14,016,907
PER	8,897,440
LOC	3,464,982
ORG	1,654,485
言語タグ数	95
実際の言語数	68
元言語の NE 表記の平均系列長	15.085
IPA 表記の平均系列長	15.894
IPA シンボルの種類数	393

関連の記事 [10, 11] を基に，簡易的な G2P ルール[8]も作成した．後者についてもカタカナ・ひらがな双方のルールを構築したため，前者と合わせ計 4 種類の G2P ルールを日本語の仮名について作成したこととなる．なお，IPA 付き NE 表記データセットの構築における日本語の G2P には，日本語の音声・音韻に関する文献に基づくルールの方を使用した．

本研究で作成した 4 種類の日本語 G2P ルールは，Epitran に統合された．カタカナ・ひらがなのような音節文字[9]を入力とするルールが Epitran に統合されたのは，初めてのことである．本研究で作成したルールは Epitran のフレームワーク内で現在利用可能であるほか，ルールを用いた IPA 変換のデモンストレーションも作成済みである[10]．

4 データセットの分析

データセットの統計的情報の概要を，表 4 に示す．NE 表記と IPA 表記の組数，および表記の ID の数はそれぞれ 6,900 万以上，1,400 万以上にのぼり，1 つの ID につき平均 4.964 組の NE 表記-IPA 表記の組が存在している．また，1 つの言語タグあたり，1 言語あたりの表記の組数はそれぞれ平均 732,357.379 組，1,023,146.338 組であり，1 言語につき平均して 100 万以上の NE 表記-IPA 表記の組が得られたことになる．1 言語あたりの表記数が 13,980.283 表記で 1 万あまりである WikiPron と比較するとこれは非常に規模の大きいデータである．したがって，本研究で作成したデータセットは，NE に特化しているという点で先行研究における多言語表記と IPA 表記の対応データセットと区別されると同時に，1 言語あたりのデータ量が既存のデータセットを大幅に上回っている，という点でも特徴づけられる．

最も多くのデータを得た NE のタイプは，PER であった．また，データセットに格納されている元言語の NE 表記と IPA 表記の平均系列長はそれぞれ 15.085，15.894 であり，ほとんど差がない．

NE 表記の文字種の数は，計 20 種である[11]．表記数が 100 万以上（およそ平均以上）にのぼる言語の数は 15 言語であり，ロシア語，中国語を除き全てラテン文字を使用する言語，またハンガリー語，中国語を除き全て印欧語族に属する言語である．言語タグごとの表記数の統計を，付録 A に付す．

5 おわりに

本研究では，NE の G2P や通言語的な言語モデルの構築，NE 表記のリンキングなどへの応用を見据え，複数の言語における NE 表記の既存データを基に，ルールベース G2P のフレームワークを用いて 68 言語の 6,900 万以上の IPA 表記付き NE からなる大規模データセットを構築した．また，これに伴って，日本語のカタカナ・ひらがなから IPA への言語学的に正確な G2P ルールを作成し，フレームワークに追加した．今後の課題として，これまで応用先として述べてきたタスク，すなわち言語モデルにおける事前学習や NE の G2P，NE 表記リンキングなどの実験を，本研究で作成したデータセットを用いて行い，データセットの効果を検証する必要がある．

参考文献

[1] David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**, pp. 3475–3484. The COLING 2016 Organizing Committee, December 2016.

[2] Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. Mc-

8） 具体的には，日本語の音声・音韻に関する文献に基づく，より正確な方の map のルール数が 150 であるのに対し，簡易的な方の map のルール数は 112 である．post のルール数についても，前者が 46 であるのに対し，後者は 20 である．

9） 「音節」は，音韻論における単位の一つ．通常は母音が「核」となり，その前後に存在しうる子音とともに構成されるが，母音のみ，子音のみで構成される場合もある．ラテン文字，キリル文字などが音素を表記の単位とする「音素文字」であるのに対し，カタカナ・ひらがな等は音節を表記の単位とする「音節文字」である．

10） https://yusuke1997.com/Japanese_G2P/

11） ラテン文字，ゲエズ文字，アラビア文字，キリル文字，ベンガル文字，デーヴァナーガリー文字，カタカナ，ひらがな，クメール文字，ラーオ文字，マラヤーラム文字，ビルマ文字，オリヤー文字，グルムキー文字，シンハラ文字，タミル文字，テルグ文字，タイ文字，漢字（簡体字），漢字（繁体字）．

Carthy, and Kyle Gorman. Massively multilingual pronunciation modeling with WikiPron. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 4223–4228. European Language Resources Association, May 2020.

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

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

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

[6] flite: A small fast portable speech synthesis system,（2023-01 閲覧）. https://github.com/festvox/flite.

[7] CC-CEDICT Home [CC-CEDICT WIKI],（2023-01 閲覧）. https://cc-cedict.org/wiki/#what_is_cc-cedict.

[8] 斎藤純男. 日本語音声学入門 改訂版. 三省堂, 2006.

[9] 日本語教育学会編. 新版 日本語教育事典. 大修館書店, 2005.

[10] Katakana - Wikipedia,（2022-08 閲覧）. https://en.wikipedia.org/w/index.php?title=Katakana&oldid=1103341275.

[11] Sokuon - Wikipedia,（2022-08 閲覧）. https://en.wikipedia.org/w/index.php?title=Sokuon&oldid=1096454475.

A　付録：作成したデータセットの言語タグ別表記数

言語タグ: 日本語名称	表記数	言語タグ: 日本語名称	表記数
aa: アファル語	28,894	ny: チェワ語	29,309
am: アムハラ語	4,478	om: オロモ語	30,961
ar: アラビア語	830,648	or: オリヤー語	15,045
av: アヴァル語	1,155	pa: パンジャーブ語	18,733
az: アゼルバイジャン語	115,014	pl: ポーランド語	1,526,678
bn: ベンガル語	437,838	pt: ポルトガル語	373,237
ca: カタルーニャ語	3,057,109	pt-br: ポルトガル語（ブラジル）	1,897,670
cs: チェコ語	1,283,030	rn: ルンディ語	28,017
de: ドイツ語	4,177,379	ro: ルーマニア語	894,080
de-at: ドイツ語（オーストリア）	295,193	ru: ロシア語	1,333,970
de-ch: ドイツ語（スイス）	31,433	rw: ルワンダ語	30,374
de-formal: ドイツ語（フォーマル）	34	sg: サンゴ語	27,867
en: 英語	13,715,761	si: シンハラ語	19,001
en-ca: 英語（カナダ）	424,497	sn: ショナ語	29,662
en-gb: 英語（イギリス）	141,742	so: ソマリ語	30,748
es: スペイン語	6,071,612	sq: アルバニア語	2,855,562
es-419: スペイン語（ラテンアメリカ）	1,271	sv: スウェーデン語	2,733,009
es-formal: スペイン語（フォーマル）	506	sw: スワヒリ語	294,945
fa: ペルシャ語	630,954	ta: タミル語	89,757
ff: フラニ語	30,456	te: テルグ語	52,395
fr: フランス語	5,003,611	tg: タジク語	76,492
ha: ハウサ語	46,317	tg-cyrl: タジク語（キリル文字）	566
hi: ヒンディー語	74,366	tg-latn: タジク語（ラテン文字）	29,674
hr: クロアチア語	426,170	th: タイ語	91,844
ht: ハイチ語	88,589	ti: ティグリニャ語	288
hu: ハンガリー語	1,056,422	tk: トルクメン語	28,707
hu-formal: ハンガリー語（フォーマル）	2	tl: タガログ語	172,360
id: インドネシア語	774,023	tr: トルコ語	586,329
it: イタリア語	3,096,623	ug: ウイグル語	3,125
ja: 日本語	671,429	ug-arab: ウイグル語（アラビア文字）	113
jv: ジャワ語	151,768	uk: ウクライナ語	654,641
kk: カザフ語	58,089	ur: ウルドゥー語	149,481
kk-cyrl: カザフ語（キリル文字）	47,204	uz: ウズベク語	192
kk-kz: カザフ語（カザフスタン）	761	uz-cyrl: ウズベク語（キリル文字）	1
kk-latn: カザフ語（ラテン文字）	74,802	uz-latn: ウズベク語（ラテン文字）	7
kk-tr: カザフ語（トルコ）	25,389	vi: ベトナム語	568,179
km:　クメール語	3,241	xh: コサ語	32,628
ky:　キルギス語	44,936	yo: ヨルバ語	274,206
lo: ラーオ語	1,408	zh: 中国語	965,861
mi: マオリ語	51,825	zh-cn: 標準中国語	16,178
ml: マラヤーラム語	93,555	zh-hans: 中国語（簡体字）	255,533
mn: モンゴル語	12,020	zh-hant: 中国語（繁体字）	9,508
mr: マラーティー語	41,352	zh-hk: 中国語（香港）	5,065
ms: マレー語	545,598	zh-mo: 中国語（マカオ）	305
mt: マルタ語	68,109	zh-my: 中国語（官話・マレーシア）	6
my: ビルマ語	9,331	zh-sg: 中国語（官話・シンガポール）	104
nl: オランダ語	9,586,075	zh-tw: 中国語（官話・台湾）	9,518
nl-informal: オランダ語（インフォーマル）	5		