漢字の音読みにおける象徴素のデータ駆動的探索の試み

吉田晶 [1]　松平茅隼 [1]　加藤大貴 [2,1]　平山高嗣 [3,1]
駒水孝裕 [1]　井手一郎 [1]
[1] 名古屋大学 [2] 名城大学 [3] 人間環境大学
{yoshidaa,matsuhirac,katoh}@cs.is.i.nagoya-u.ac.jp
t-hirayama@uhe.ac.jp
taka-coma@acm.org
ide@i.nagoya-u.ac.jp

概要

象徴素とは，形態素未満の語の構成単位で，特定の意味を想起させるものを指す．本研究では，意味的にまとまった漢字クラスタのうち，音読みにおいて特定の子音を有意に多く含むものを象徴素クラスタと称し，その抽出および意味のラベリングを試みた．前者では，漢字を分散表現に変換してクラスタリングした後，各クラスタに含まれる子音の比率と母比率を比較し，比率に偏りがある意味クラスタを抽出した．後者では，辞書の定義文を利用する手法と，WordNet を利用する手法の 2 つによるラベリングを比較した．常用漢字にこれらを適用した結果，計 133 の象徴素クラスタを抽出し，2 つのラベリング手法共に，それらの 8 割以上に対して意味ラベルの付与に成功した．

1　はじめに

象徴素とは，形態素未満の語の構成単位で，特定の意味を想起させるものを指す．これは，イギリスの言語学者 Firth が提唱した概念であり [1]，例えば，象徴素「gl」で始まる英単語には「gleam」や「glitter」など光に関係する語が多く含まれる．

象徴素は英語やスウェーデン語などで存在が確認されており，新規の象徴素の探索や既知の象徴素の分析が行なわれている．しかし日本語を対象とした研究は少ない．これは，一般に象徴素は 2，3 個の子音列で構成される [2] のに対し，日本語では子音が連続しないという障壁による．Hamano ら [3] は子音と母音を象徴素とみなして日本語における象徴素の分析を行なったが，対象がオノマトペに限られていたため，英語における象徴素ほど具体的な意味は特定されていない．また評価を人手で行なっているため，客観性や労力について留意する必要がある．

そこで本研究では，オノマトペよりも具体的な意味を表す漢字を用い，音読みの子音に着目することで，英語に対する分析に近い形で，日本語における象徴素の分析を自動的に行なうことを試みる．具体的には，意味的にまとまった漢字クラスタのうち，特定の子音を有意に多く含むものを「象徴素クラスタ」と称し，漢字の音読みに着目した象徴素クラスタの抽出と，その具体的な意味の特定を自動的に行なう手法を提案する．これは，英語の象徴素と同様に，日本語においても同じ音を有する語は似通った意味を表すと仮定したとき，白（ハク）・黒（コク）・赤（セキ）・緑（リョク）など色を表す頻出漢字は第 2 子音に k を，殺（サツ）・滅（メツ）・罰（バツ）・切（セツ）など攻撃性を感じる漢字は第 2 子音に t をもつなど，子音と意味の間に相関がある可能性が考えられるためである．音読みに着目した理由は，音読みが必ず 1 音節からなり，方言の影響が少なく，分析が容易であると考えたためである．

以降，まず 2 節で関連研究を紹介し，3 節で象徴素クラスタの抽出手法を，4 節で象徴素クラスタのラベリング手法を提案する．次に 5 節で提案手法の処理結果と考察を示したのち，最後に 6 節で本研究をまとめ，今後の課題と展望について述べる．

2　関連研究

Otis ら [4] は，分散表現を用いて英語の象徴素を分析する手法を提案した．具体的には，分散表現により英単語を意味空間へ写像し，単語間の距離を検定比較することで象徴素の存在を確認した．また，Abramova ら [5] は，英語の象徴素の意味を自動的に特定する手法を提案した．具体的には WordNet [6] における上位語と下位語の階層を利用し，象徴素を共有する単語集合へのラベリングを実現した．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

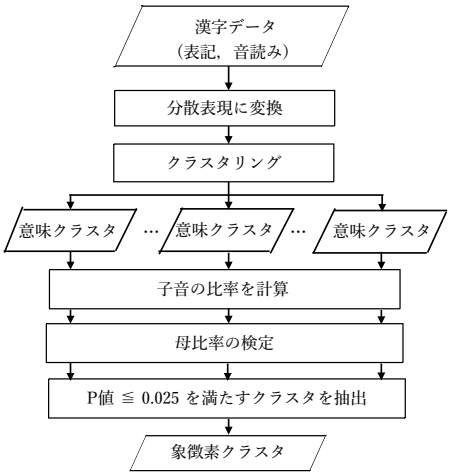

図 1　象徴素クラスタ抽出の処理手順

本研究では，これらの先行研究に基づき，漢字を分散表現に変換して扱う．また，ラベリング手法の 1 つとして日本語 WordNet [7] を用いた手法を提案する．先行研究では，英単語を既知の象徴素でグループ化した後に，単語間の距離に検定を施している．しかし，音読みにおいては既知の象徴素が存在しないため，本研究では，まず意味に基づいて漢字データをクラスタリングし，その後に母比率の検定により子音の偏りを分析することで，特定の意味と子音との間の共起性の抽出を試みる．

3　象徴素クラスタの抽出手法

本節では，漢字の分散表現と母比率の検定を用いた象徴素クラスタの抽出手法を提案する．提案手法の処理手順を図 1 に示す．

まず漢字データとして表記と音読みを入力し，次に各漢字の分散表現をクラスタリングし，最後に得られた各意味クラスタにおける子音の比率を検定することで象徴素クラスタを抽出する．

3.1　漢字データの準備

常用漢字表 [8] に収録されている漢字 2,136 字を対象とし，同表から各漢字の表記，音読み，用例を抽出した．この際，音読みが複数存在する場合はそれら全てを使用した．また，音読みが存在しない場合も，子音をもたない漢字として含めた．例えば漢字「亜」について，音読み「ア」と用例「亜流，亜麻，亜熱帯」を漢字データとして得た．

表 1　常用漢字データにおける各第 1 子音の比率

b	d	g	h	k	m	n
0.046	0.025	0.052	0.088	0.201	0.030	0.017
r	s	t	w	y	z	ϕ
0.053	0.217	0.108	0.002	0.031	0.063	0.068

表 2　常用漢字データにおける各第 2 子音の比率

k	N	t	ϕ
0.108	0.205	0.056	0.631

3.2　漢字分散表現への変換

漢字を分散表現に変換するために，文字単位の Bidirectional Encoder Representations from Transformers（BERT）の学習済みモデル [9] を利用した．本モデルは日本語の Wikipedia 記事で事前学習されている．2,136 件の漢字データのうち 2,135 件について分散表現への変換に成功した[1)]．

3.3　意味クラスタの抽出

英語において象徴素を共有する語は似た意味を表すことが知られているが，日本語においてもこの性質があると仮定する．そこで，3.2 項で得られた分散表現に対して，Ward 法で階層的クラスタリングを施し，その過程で出現した全てのクラスタを漢字の意味クラスタとして抽出し，それらを象徴素クラスタの候補とみなす．なお，包含関係にあるクラスタ同士も別の意味クラスタとする．

3.4　子音の比率の分析

得られた各意味クラスタに対し，第 1 子音および第 2 子音の比率の偏りを調べる．

まず，母集団となる常用漢字データにおける第 1，第 2 子音の比率をそれぞれ計算する．この際，昭和 29 年内閣訓令第 1 号で告示された「ローマ字のつづり方」[10] に基づいて子音を決定した．ここで，音読みをもたない漢字は子音がないもの「ϕ」として扱う．また，「ン」の子音は「N」，ナ行の子音は「n」と表して区別する．得られた第 1，第 2 子音の割合をそれぞれ表 1，表 2 に示す．次に，各意味クラスタに対しても同様に子音の比率を計算する．ここで，クラスタの要素数が少なすぎると，含まれる子音の比率が極端に大きくなってしまうため，要素数が 8 未満の意味クラスタは除外した．

最後に，各子音について，各クラスタにおける比

1)　変換に失敗したのは「憬」で，使用した BERT の語彙に含まれていなかったことが理由であった．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

率に対して母比率の検定を行なう．具体的には，いずれかの子音について P 値が有意水準を満たす意味クラスタを子音に偏りがあると判断し，象徴素クラスタとして抽出する．検定は片側検定とし，有意水準は経験的に 0.025 とした．この際，最小の P 値を示す子音をそのクラスタを代表する象徴素とする．

4 象徴素クラスタのラベリング手法

本節では，3 節で抽出した各象徴素クラスタに対して，その意味を自動的にラベリングする手法を提案する．ここでは辞書を用いた手法と WordNet を用いた手法の 2 つの手法を提案し，比較する．

4.1 辞書を用いたラベリング手法

まず，辞書における各漢字の定義文を利用したラベリング手法を提案する．本研究では Weblio 辞書中のデジタル大辞泉 [11] を使用した．

まず，対象とする象徴素クラスタに含まれる全ての漢字に対し，辞書からその定義文を得る．この際，辞書中に見出し語が見つからない場合には，3.1 項で準備した用例を代替の見出し語として定義文を抽出する．次に，定義文に対して MeCab [12] で形態素解析を施し，ストップワードを除いた動詞，形容詞，名詞，副詞を抽出する．その後，抽出した各語について，クラスタ内での出現回数を数える．なお，ある語が 1 つの定義文中に複数回現れた場合には，出現回数は 1 として扱う．また，「時間」と「年月」のような類義語は統合する．類義語の決定には Weblio 辞書中の日本語 WordNet [13] および Weblio 類語辞書 [14] を用いた．最後に，クラスタ内で最も多く出現した単語を候補ラベルとし，その中で出現率が閾値（以下の実験では経験的に 0.3 に設定）以上のものを象徴素クラスタのラベルとして採用する．ここで，出現率とは象徴素クラスタ内の全ての漢字のうち，その定義文に該当ラベルの語が 1 回以上出現する漢字の割合を意味する．

4.2 WordNet を用いたラベリング手法

次に，英単語を対象とした Abramova ら [5] の手法を拡張したラベリング手法を提案する．

まず，対象とする象徴素クラスタに含まれる全漢字に対し，日本語 WordNet における上位語を取得する．この際，入力漢字が WordNet の階層構造中に存在しない場合には，3.1 項で準備した漢字データの用例を代替の語として入力する．次に，クラスタ内の各漢字とその全ての上位語を統合し，候補ラベルとする．この際，WordNet の階層構造において根ノードからの最短距離が 3 以下の名詞は抽象的すぎると考えたため，除外した．次に，各候補ラベルに対して，以下の式で類似スコア S を計算する．

$$S(h, C) = \sum_{w \in C} \alpha(w, h) \tag{1}$$

$$\alpha(w, h) = \begin{cases} \frac{1}{\mathrm{dist}(w,h)^2} & \text{if } h \in H(w) \\ -0.1 & \text{if } h \notin H(w) \end{cases} \tag{2}$$

ここで C は対象とする象徴素クラスタを，w は当該クラスタ内の漢字を，h は候補ラベルを表す．また，$H(w)$ は w とその全ての上位語の集合を表す．$\mathrm{dist}(w, h)$ は WordNet の階層構造における w から h までの最短経路の長さを表すが，w と h が一致する場合は 1 を返す．

最後に，類似スコアが正の値かつ被覆率が閾値（以下の実験では経験的に 0.2 に設定）以上のものを象徴素クラスタのラベルとして採用する．ここで，被覆率とは象徴素クラスタ内の全ての漢字のうち，その漢字とその上位語の集合に当該の語が含まれる漢字の割合を意味する．

表 3 得られた象徴素クラスタの例

	クラスタ	子音	比率	Z 値
第 1 子音	価値果均献 貢勲功績	k	7/ 9	4.311
	講教学校究 研祉術療	k	5/ 9	2.649
	浄清静鎮聖 宮神仙天	s	5/ 9	2.460
第 2 子音	携連関係絡 結接雑密	t	4/ 9	5.083
	黒白青赤紫 緑紅黄	k	4/ 8	3.601
	紡蚕絹繭糸 繊桑藍麻綿	N	6/10	3.111

5 提案手法の処理結果と考察

本節では，象徴素クラスタの抽出とそのラベリング結果および考察を述べる．

5.1 象徴素クラスタの抽出

第 1，第 2 子音のそれぞれについて象徴素クラスタを抽出した結果，それぞれ 100，33 のクラスタが得られた．得られた象徴素クラスタの例を表 3 に示す．なお，表中の Z 値は，母比率の検定における検定統計量を示す．第 1，第 2 子音による結果を比較すると，17 のクラスタが共通し，4 のクラスタが包

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 4　象徴素クラスタに対するラベリング結果の例

クラスタ	辞書ベース		WordNet ベース		
	ラベル	出現率	ラベル	被覆率	類似スコア
黒白青赤紫緑紅黄	色	7/ 8	spectral color	5/ 8	4.70
	名	5/ 8	color	7/ 8	1.65
	草木	3/ 8	piece	2/ 8	1.40
割擦裂張貼塗削剥	切れる	4/ 8	–	–	–
滴晶泡豆穀麦粉菓粒	年間	4/ 9	grain	3/ 9	2.40
			sphere	3/ 9	1.51
			food product	4/ 9	1.25
短長少中小大半高低	低い	6/ 9	size	2/ 9	1.30
	高い	5/ 9	concept	2/ 9	0.55
	程度	4/ 1	–	–	–
置立持存産生出入成行発	起こる	7/11	human action	3/11	1.31
	出す	6/11			
	生きる	4/11			

含関係にあった．例えば「紡・蚕・絹・繭・糸・繊・桑・藍・麻・綿」は第 1 子音による分析では子音 m，第 2 子音による分析では子音 N において，母比率の検定で有意水準を満たした．しかし，第 1 子音が m かつ第 2 子音が N である漢字は「綿」のみであるため，子音の組 (m, N) に象徴素的な役割があるとは考えにくい．

5.2　象徴素クラスタのラベリング

辞書を用いたラベリング手法では，100 の第 1 子音による象徴素クラスタのうち 87，33 の第 2 子音による象徴素クラスタのうち 30 のクラスタで 1 つ以上の意味ラベルが得られた．一方，WordNet を用いた手法では，100 の第 1 子音による象徴素クラスタのうち 83，33 の第 2 子音の象徴素クラスタのうち 29 のクラスタで 1 つ以上の意味ラベルが得られた．両手法によるラベリング結果の例を表 4 に示す．まず，辞書を用いた手法の特徴として，WordNet を用いた手法と比べて，動作を表す漢字を多く含む象徴素クラスタに対して，より適切な意味ラベルの付与が確認された．例えば「割・擦・裂・張・貼・塗・削・剥」に対して，後者ではラベルが付与されなかったが，前者では「切れる」のラベルが付与された．

一方で，不適切なラベルが適切なラベルと同等の出現率を示してしまう事例も複数見られた．例えば「滴・晶・泡・豆・穀・麦・粉・菓・粒」に対して，「年」のラベルが出現率 4/9 で付与された．これは，「携・連・関・係・絡・結・接・雑・密」に対して付与されたラベル「つながり」の出現率と同じ値であるが，後者のラベルが前者のラベルと同程度に象徴素クラスタの意味を表しているとは考えられない．この結果は，各漢字の定義文が短かったためラベル候補の単語が少数だったこと，「一年麦」や「年の豆」など，「年」を含む用例が複数回あったことが原因である．この問題は，複数の辞典からの定義文の抽出や，ラベリングの適切さに対する出現率以外の尺度の導入により改善できると考えられる．

次に WordNet を用いたラベリング手法の特徴として，辞書を用いた手法と比べて，より包括的なラベルの付与が確認された．例えば「短・長・少・中・小・大・半・高・低」に対して，後者では「低い」や「高い」のラベルが同時に付与されたのに対して，前者ではクラスタ内の漢字全体に共通するラベル「size」が付与された．しかし包括的であるがゆえに，具体性に欠けるラベルも複数見られた．例えば「置・立・持・存・産・生・出・入・成・行・発」に対して，「human action」が第 1 候補のラベルとして付与された．本実験では，各候補ラベルに対して，被覆率が 0.2 以上という条件を設けたため，被覆率が 0.18 であった適切なラベル「beginning」が付与されなかった．このラベルは類似スコアでは「human action」に劣ったものの，根ノードからの距離は「human action」よりも離れており，より具体的である．この問題は，被覆率の制限を緩め，類似スコア，被覆率，根ノードからの距離それぞれの重み付き和を計算した尺度を導入することで改善できると考えられる．

6　おわりに

本研究では，音読みに着目した日本語漢字における象徴素クラスタの抽出と，その意味のラベリングを試みた．提案手法を常用漢字データに適用した結果，133 の象徴素クラスタと，そのうちの 8 割以上に対する意味ラベルが得られた．

各子音が複数の象徴素クラスタに属したため，結果的に英語の象徴素ほど一意に意味は定まらなかった．しかし，「清・静・聖・神・仙」といった清潔さや神聖さを表す漢字の音読みは子音 s で始まるなど，直感に沿う子音と意味の相関を確認できた．

抽出した象徴素クラスタの客観的評価，ラベリングの成功可否に影響を及ぼす要因の解明などが今後の課題である．また将来の展望として，漢字の訓読み，和語や熟語への拡張などが挙げられる．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究の一部は，JSPS 科研費 JP22H03612 の助成を受けたものである．

参考文献

[1] John Firth. **Speech**. Oxford University Press, 1930.

[2] Leanne Hinton, Johanna Nichols, and John J. Ohala. **Sound Symbolism**. Cambridge University Press, 1995.

[3] Shoko Hamano. **The Sound-Symbolic System of Japanese**. CSLI Publications, 1998.

[4] Katya Otis and Eyal Sagi. Phonaesthemes: A corpus-based analysis. In **Proceedings of the 30th Annual Meeting of the Cognitive Science Society**, 2008.

[5] Ekaterina Abramova, Raquel Fernández, and Federico Sangati. Automatic labeling of phonesthemic senses. In **Proceedings of the 35th Annual Meeting of the Cognitive Science Society**, pp. 1696–1701, 2013.

[6] Christiane Fellbaum. **WordNet: An Electronic Lexical Database**. MIT Press, 1998.

[7] Francis Bond, Hitoshi Isahara, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. Extending the Japanese WordNet. **言語処理学会第 15 回年次大会発表論文集**, pp. 80–83, 2009.

[8] **文化庁**. **常用漢字表（平成 22 年内閣告示第 2 号）**, 2010.

[9] **東北大学乾研究室**. BERT base Japanese (IPA dictionary),（2022-12-18 **閲覧**）. https://huggingface.co/cl-tohoku/bert-base-japanese-char.

[10] **文化庁**. **ローマ字のつづり方　第 1 表（昭和 29 年内閣訓令第 1 号）**, 1954.

[11] **（株）小学館**. **デジタル大辞泉**,（2022-12-18 **閲覧**）. https://www.weblio.jp/cat/dictionary/sgkdj.

[12] Taku Kudo. Mecab: Yet Another Part-of-speech and Morphological Analyzer,（2022-12-18 **閲覧**）2006. https://taku910.github.io/mecab/.

[13] **（国研）情報通信研究機構**. **日本語** WordNet,（2022-12-18 **閲覧**）. https://thesaurus.weblio.jp/category/nwnts.

[14] GRAS **グループ（株）**. Weblio **類語辞書**,（2022-12-18 **閲覧**）. https://thesaurus.weblio.jp/category/wrugj.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

付録

表5　抽出した象徴素クラスタとそのラベリング結果の例

	クラスタ	子音	比率	Z 値	辞書ベース		WordNet ベース		
					ラベル	出現率	ラベル	被覆率	類似スコア
第 1 子音	賄拾伺拭漂浮洗覆	w	1/ 8	6.529	–	–	drift	2/ 8	1.40
	台団落土所場地道路域区駅線境界点	d	4/16	5.789	場所	15/16	place	4/16	2.80
	世代武文治住政民	m	2/ 8	5.723	世の中	8/ 8	geologic time	2/ 8	1.40
	解釈説論談話証識知告報示表録記述	w	1/16	4.510	話す	11/16	–	–	–
	胎娠妊姻婚篤乳酪縫癒	n	2/10	4.427	宿す	3/10	physical condition	3/10	1.36
	夏冬秋春週日月年季旬	n	2/10	4.427	時間	9/10	time period	11/11	5.50
	事業務役職任商工農働労	n	2/11	4.180	仕事	8/11	duty	3/11	2.20
	准準添副従伴沿倣随逐	z	4/10	4.174	従う	4/10	attendant	2/10	1.20
	堂院寺塔洞坊亭楼	d	2/ 8	4.093	住まい	6/ 8	building	5/ 8	3.2
	緩快急厳激著極微労	g	3/ 8	4.092	激しい	3/ 8	intense	2/ 8	1.40
	壇廷典殿威誉儀宗礼	d	2/ 9	3.806	規範	3/ 9	activity	3/ 9	1.51
	幻夢魂霊怪妖呪魔	m	2/ 8	3.649	不思議	5/ 8	spirit	3/ 8	1.75
第 2 子音	一二三四五六七八九	t	3/ 9	3.630	数字	9/ 9	digit	9/ 9	9.00
	刷稿刊版筆書描閲読	t	3/ 9	3.630	書く	6/ 9	textual matter	2/ 9	1.30
	抜切掛込組選編予決定	t	3/10	3.367	取る	5/10	statement	2/10	1.20
	素実能格質材品物気風	t	3/10	3.367	本質	5/10	artifact	4/10	2.51
	昙孔窟穴栓槽膜筒圧液蓋胴	t	3/12	2.933	穴	6/12	opening	3/12	2.10
	衣服靴帽飲食酒醸浴粧濯	k	3/11	2.727	身	5/11	clothing	3/11	1.45
	角翼針標玉鈴環輪	k	3/ 8	2.429	形	2/ 8	toroid	2/ 9	1.40
	憶択答考思想郭構	k	3/ 8	2.429	考える	3/ 8	belief	3/ 8	1.75
	糖蜜酸窒硫塩炭油	t	2/ 8	2.395	化合	3/ 8	chemical compound	5/ 8	2.12
	引巻押追越超衝突	t	2/ 8	2.395	地点	3/ 8	force	2/ 8	1.40
	易難便通略緩快急厳激著極微	k	3/13	2.316	–	–	chemical compound	5/ 8	2.12
	進流信伝交情運展集達	N	5/10	2.315	伝わる	6/10	state	4/ 8	1.17

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).