日本語語彙正規化コーパス アノテーションガイドライン Version 0.2

2022 年 6 月 14 日

東山 翔平 †

1 概要

本稿では、Blog and Q&A Site Normalization Corpus（BQNC）[1](Higashiyama et al. 2021) を構築する過程で定めた、日本語のユーザ生成テキスト（User-Generated Text; UGT）に対して語彙正規化情報を含む所定の語情報を付与するためのアノテーション基準を説明する。BQNC は、国立国語研究所（国語研）による現代日本語書き言葉均衡コーパス（BCCWJ）[2]・非コアデータの特定目的サブコーパス「ブログ」「知恵袋」の一部に対し、語情報の修正と追加のアノテーションを施したコーパスである [3]。

本基準でアノテーション対象とする語情報は次のものである。BCCWJ のオリジナルデータと同様に、各文を国語研の短単位（以下、単に**語**と呼ぶ）に分割した上で、各語に出現形（語形出現形)、品詞、活用型、活用形、読み（発音形出現形)、語彙素、語彙素 ID の情報（これらをまとめて**語情報 B** と呼ぶ）を付与する。さらに、本基準で独自に付与する情報として、各語に正規表記 ID、語彙種別、異表記種別（これらをまとめて**語情報 N** と呼ぶ）を付与する。**正規表記**とは、UGT で頻繁に用いられる「崩れた」語の表記に対して、公用文や新聞記事などの公的なテキストで用いられる規範的・標準的な表記を指す。たとえば「でしょ」という出現形に対しては「でしょう」という正規表記が該当する。**語彙種別**は「オノマトペ」、「外国語」、「顔文字」などの語彙の意味的分類を指し、**異表記種別**は「代用表記」、「音変化」などの語の異表記の分類を指す。

例として、入力文「まあねっ、時間かかったでしょ」に対するアノテーション結果を表 1 に示す [4]。例では、「ねっ」と「でしょ」がともに「音変化」の異表記であり、その正規表記 ID がそれぞれ「28754: ね」と「25653: です_デショウ」となっている。正規表記 ID は、「(語彙素 ID):(語彙素)_(出現形発音)」の形式で定め、活用しない語や、終止形・連体形の活用語について

† 国立研究開発法人情報通信研究機構, NICT

[1] https://github.com/shigashiyama/jlexnorm

[2] https://pj.ninjal.ac.jp/corpus_center/bccwj/

[3] 本稿で述べるアノテーション基準には BQNC 構築後に修正した点があり、BQNC version 1.0 のアノテーション結果と一部一致しない点がある。BQNC の更新版データは今後公開予定である。

[4] 各語の品詞は、細分類まで含めると順に「副詞」「助詞-終助詞」「補助記号-読点」「名詞-普通名詞-助数詞可能」「動詞-非自立可能」「助動詞」「助動詞」である。表中では余白の都合で省略した。

表 1　入力文「まあねっ、時間かかったでしょ」に対する語情報の付与

出現形	品詞	活用型	活用形	読み	語彙素	語彙素 ID	語彙・異表記種別	正規表記 ID
まあ	副詞			マー	まあ	35378		
ねっ	助詞			ネッ	ね	28754	音変化	28754: ね
、	補助記号				、	24		
時間	名詞			ジカン	時間	17768		
かかっ	動詞	五段-ラ行	連用形-促音便	カカッ	掛かる	6016		
た	助動詞	助動詞-タ	終止形-一般	タ	た	21642		
でしょ	助動詞	助動詞-デス	意志推量形	デショ	です	25653	音変化	25653: です_デショウ

表 2　正規表記定義表の例

正規表記 ID	正規表記
8754: ね	ね
25653: です_デショウ	でしょう
26203: 通り	通り, とおり

は「(語彙素 ID):(語彙素)」の部分のみとする。本基準は、崩れた表記に対する正規表記を一つに限定せず、所定の基準を満たしたものを複数許容する。つまり、表 2 の**正規表記定義表**で示すように、各正規表記 ID に対して一つ以上の正規表記が対応する。

本基準に基づくアノテーション作業は、次の手順からなる。

(1) 自動形態素解析結果が付与された文の集合から、アノテーションを行う文を選択する。特に、BQNC の構築にあたっては、極力多様な表現を収録する目的から、UGT に特有な表現を含み、かつこれまでに選択した文に含まれていない表現を含む文を優先して、人手により選択した。

(2) 「『現代日本語書き言葉均衡コーパス』形態論情報規程集 第 4 版（上／下）」(小椋他 2011)（以下、**BCCWJ 規程集**と呼ぶ）に基づき、選択した文に元々付与されている語の区切りまたは語情報 B に誤りがあれば修正する。さらに、語情報 N で該当するものがあれば付与し、必要に応じて新たな正規表記 ID、正規表記集合を定義して正規表記定義表を更新する。

本アノテーションでは、形態素解析誤りに直結し得る、崩れた異表記の正規化のための情報の付与に焦点を当てる。したがって、あり得るあらゆる表記揺れの正規化、語彙的に異なる同義語の同定、略語・通称名の正式名称への変換、方言の共通語への変換、古語の現代語への変換、意味的・文法的な誤りの訂正などは考慮しない。

表 3　語彙種別、異表記種別の一覧

		例	正規表記
語彙種別	スラング	コピペ	
	固有名	ドラクエ	
	オノマトペ	キラキラ	
	感動詞	おお	
	外国語	ＥＡＳＹ	
	方言	ほんま	
	顔文字・ＡＡ	（＾－＾）	
異表記種別	異文字種	カワイイ	かわいい, 可愛い
	代用表記	大きぃ	大きい
	音変化	おいしーい	おいしい, 美味しい
	誤表記	つたい	つらい, 辛い

2　語彙種別と異表記種別の概要

本基準では、表 3 に示すように、7 種の語彙種別と 4 種の異表記種別を考慮する。これらの語彙、異表記に該当する語は、新聞記事などの既存のアノテーション付きコーパスで学習された形態素解析システムにとって、解析誤りの原因となりやすい [5]。このような種別を付与することで、形態素解析システムや語彙正規化システムの解析精度を現象の種類ごとに評価可能とすることを意図している。

7 種の語彙の特徴として、現在までに定着した語に加え、命名者や言語使用者により新たな語が際限なく生み出され得る点がある。現代日本語共通語の表現を含む語彙として「スラング」、「固有名」、「オノマトペ」、「感動詞」があり、その他の言語表現を含む語彙として「方言」、「外国語」があり、非言語表現として「顔文字・ＡＡ」（アスキーアート）がある。

4 種の異表記種別には、「異文字種」、「代用表記」、「音変化」、「誤表記」がある。この 4 種のいずれかに該当する語（正規化対象表記）には正規表記を付与する。いずれにも当たらない場合には正規表記を付与しない。

3　語彙種別の定義

§3.1–§3.7 において表 3 の 7 種の語彙種別の定義を示し、最後に §3.8 で全体に関する補足を述べる。

[5] 鍛治 他 (2015) は、Twitter テキストにおける形態素解析の分割誤りを分類し、表 3 の語彙・異表記とほぼ同等の項目について誤りが生じていることを報告している。

3.1 スラング

新しく社会に現れ、インターネットや放送メディアを通じて普及、定着し、集団的に使用されるようになった語のうち、固有名を除くものをスラングとみなし、語彙種別「スラング」を付与する。

なお、明解日本語学辞典 (森山，渋谷 2020) では、スラングを「集団語の一種。特定の社会集団・生活集団で使われる用語や表現。（中略）もっぱら集団内の情報伝達の効率化や仲間意識の保持のためにある。」と定義している。「集団語」は、スラングの他に職業語、専門語なども含むものとして、柴田 (1956) が提唱した概念である。松田 (2006) は集団語を「一般集団語」と「ネット集団語」に分け、後者を「インターネットを仲介して繋がった人々がネット上（中略）で交流するうちに発生した集団語」と定義した。本基準で扱うスラングは、多くの場合この「ネット集団語」に該当するが、インターネット上で発生したものに限定しない違いがある。また、類似の概念である「新語」、「俗語」、「若者語」との対象範囲の主な違いとして、「新語」と異なり固有名を含まない点、「俗語」と異なり卑俗なものに限らない点、「若者語」と異なり若者が使用するものに限らない点が挙げられる。

BQNC にてスラングを認定する際には、「新しく社会に現れた」という観点について、「概ね 2000 年以降に使用頻度が急増したと考えられる語」という基準を設け [6]、使用頻度の計測には BCCWJ 中納言版 [7] を使用した。具体的には、「概ね 1999 年以前の使用頻度が 5 件未満かつ 2000 年以降の使用頻度が 10 件以上」であるものをスラングとした。以下、例を挙げる。

- 次のものは、短単位検索の「書字形出現形」の条件で、1999 年以前の使用が見られないかごくわずかで、2000 以降に使用頻度が 10 件以上に増えており、スラングと認定した。
  - 「即買い」(1999 年以前頻度 →2000 年以降頻度：0→32)、「コメ返」(0→13)、「スク水」(0→21)、「プロフ」(0→52)、「コピペ」(0→204)、「ネタバレ」 (0→263)、「激混み」(1→18)、「ゲーセン」(1→107)、「プリクラ」 (3→79)、「バツイチ」 (2→96) など。
- 次のものは、短単位検索の「語彙素」の条件で、1999 年以前の使用が見られないかごくわずかで、2000 以降に使用頻度が 10 件以上に増えており、スラングと認定した。
  - 「コメ」（コメントの意味）(0→315)、「連ちゃん」（麻雀用語の「連荘」）(0→78)。
- 次のものは、文字列検索で、1999 年以前の使用が見られないかごくわずかで、2000 以降に使用頻度が 10 件以上に増えており、スラングと認定した。
  - 「リスカ」 (1→10) (“[^ア-ン] リスカ [^ア-ン]” の条件を設定)、「コメ返し」 (0→13) (“コメ返し [^てま]” の条件を設定)、「２ちゃんねらー」 (0→19) など。

[6] BQNC の原文のテキストが 2004〜2005 年に公開された「Yahoo!知恵袋」の投稿と 2008〜2009 年に公開された「Yahoo!ブログ」の記事であることから、2000 年という時期とした。

[7] `https://chunagon.ninjal.ac.jp/`

- 次のものは、BCCWJ 中納言版では 2000 年以降の 10 件以上の使用が確認できなかったが、インターネット上の辞書サイトに立項されているなど一定の認知・使用が認められるものは、スラングと認定した。
  - 「恋ばな」、「むずい」、「キャラ弁」(goo 国語辞書など)、「キュン死に」(Weblio 辞書)、「写メる」(Weblio 類語辞書)、「社割」(Weblio 英和辞典・和英辞典)、「イタ車」(Wikipedia 記事タイトル)、「デジサイ」(Google 検索によりデジタルサイネージを指すと考えられる異なる「デジサイ」の言及(会社名除く)を 10 件以上発見)。

一方、次のものは、1999 年以降にも 5 件以上の使用が確認されたことからスラングと認定しなかった。

- 「トラブる」(6→29)、「コンビニ」(11→1539)、「ばれる」(26→130)、「モテる」(32→182)。

3.2 固有名

1 語で固有名詞に該当するものに語彙種別「固有名」を付与し、細分類「一般」、「特殊」、「HN」(ハンドルネーム)も同時に付与する。なお、BCCWJ 規程集に基づく品詞「名詞-固有名詞」が付与される語と、語彙種別「固有名」が付与される語とは一致しない場合がある。

固有名-一般 BCCWJ 規程集下巻 p.48 の 1.1「名詞」(11)「名詞-固有名詞-地名-一般」および (12)「名詞-固有名詞-地名-国」に該当する国名・地名は「固有名-一般」と認定する。

規程集下巻 p.103「同語異語判別規程」細則 4「人名の扱い」において人名とみなすものは、後述する細則 4 の 1 (2) に該当する場合を除き、「固有名-一般」と認定する。細則 4 の 1 (2)「通称や仮名，一般人のペンネームやハンドルネームなどのうち，形式や語感から人名とみなし得るもの」に該当する場合は、著名人の通称名とみなしうる場合には「固有名-一般」とし(例:「ほりえもん」)、そうでない場合には後述する「固有名-HN」を付与する(例:「さっちゃん」)。欧米等の人名について、著名人でなく人名かどうか迷う場合には、和英辞書等で「人名」の項目が存在する場合には一般的な人名とみなし、「固有名-一般」を付与する(例:「カイム」、「モリー」)。BQNC のアノテーション作業では英辞郎 [8] を用いた。

BCCWJ 規程集下巻 p.106「同語異語判別規程」細則 5「固有名の扱い」に基づき、元号、生物相当の個体(ペット、キャラクター、神仏)の名、特定の集団の名、特定のプロダクトの名など「名詞-固有名詞-一般」に該当するものは「固有名-一般」と認定する(例:ポケットモンスターのキャラクター「ズバット」)。その他、Web サービスなどの商業サービス名についても「固有名-一般」と認定する(例:「Ｖｅｃｔｏｒ」)。

固有名の略称・通称については、認知度の高い呼称と考えられる場合には「固有名-一般」と

[8] https://eow.alc.co.jp/

認定する（例：「ドラクエ」、「ＦＦ」（ファイナルファンタジー）、「ＰＳ」（プレイステーション）、「ファミコン」、「パワプロ」）[9]。

BQNC のアノテーション作業では、認知度の判断根拠として Web 上の百科事典（リダイレクトページを含む Wikipedia、ピクシブ百科事典）に該当項目の記事が存在する場合に認知度が高いとみなした。

固有名-特殊　作品中の架空の概念（例：「ホイミ」）や架空の地名（例：「ガルバディア」）とみなせるものは「固有名-特殊」と認定する。

固有名-ＨＮ　前述のように、文脈上または形式や語感から、人名や生物個体名とみなし得る通称、仮名、ハンドルネーム等については「固有名-ＨＮ」とする（例：「亀 (チャン)」、「ぴー (さん)」、「いっちぃ」、「ペコ」；「()」内は前後に出現する文字列）。

3.3　オノマトペ

擬音語と擬態語を包括してオノマトペとし、オノマトペに該当する語に語彙種別「オノマトペ」を付与する。

オノマトペの語の認定基準　オノマトペの扱いは、基本的には BCCWJ 規程集の規則に準じ、下巻 p.115 細則 7「 擬音語・擬態語の扱い」に記載されている以下の 1 (1)〜(5) および 2 (5) の規則に従う（具体例は規程集を参照）。つまり、下記規則に該当するものは、オノマトペの 1 回的描写とみなし、1 短単位とした上で、語彙種別「オノマトペ」を付与する。

1 (1) a　動物などの鳴き声の描写
　【例】こけこっこー　にゃおーん　わん

1 (1) b ①　同一の 1 音（末尾に長音・促音が付加された場合を含む）
　【例】が　がっ　がー　ぱ　ぱっ　ぱあ

1 (1) b ②　同一の 1 音の連鎖（末尾に長音・促音が付加された場合を含む）
　【例】がが　ががっ　ががががー　ぱぱ　ぱぱぱっ　ぱぱぱー

1 (1) b ③　同一の 2 音の間に長音・促音が挿入されたもの
　【例】がっが　ぱっぱ　がーが

1 (1) c ①　2 音で構成されるもの
　【例】がく　ぐにゃ　ずば　がつ　ざぶ　どん

[9] 「ドラクエ」や「ファミコン」は UniDic において品詞「名詞-普通名詞-一般」と登録されているが、このような場合に品詞の修正は行わない。

1 (1) c ②　2 音に一つ以上の派生要素が付いたもの（語末の { っ, り, ん, ー } および語中の { っ, ん, ー } を派生要素とみなす）
【例】がくっ　がくり　がくん　がっく　がっくり　がっくん　がくー　がくーん

1 (1) c ③　上記 1 (1) c の ① および ② のうち同一の 1 音が連鎖したもの
【例】ずばばば　ずばばー　ずばばっ　ずばばばばばん　どどーん　どどどどん

1 (2)　2 音の形または 1 音に長音・促音が付加された形の 2 連続、3 連続。その派生的な形で、長音・促音が最大 1 つずつ追加されたもの、2 音のうち 1 音が変更されたもの。
【例】ぐるぐる　ぐるぐるぐる　ぐるぐる | ぐるぐる　からころ | からころ

1 (3)　1 (2) 以外の形の連続は、それぞれを 1 回的描写とする。
【例】ががっ | ががっ　さっさ | さっさ　ばばっ | ばばば　ずばばん | ずどどん

1 (5) c　擬音語・擬態語に語調を整える要素や語源未詳の要素が結合したもの
【例】ぽかすか　ほにゃらか

2 (5)　オノマトペに付く「と」を含めて副詞として 1 語化したとみなす以下のもの
うかと　うんと　おいそれと　おのずと　がっしと　かっと　きちりと　きちんと
きっと　きと　ぎょっと　ぐんと　けろっと　ごそっと　ごまんと　さっさと　さと
しかと　じっと　しゃんと　しゅんと　しらっと　しれっと　しんと　じんと
ずいと　すきっと　すっくと　ずっと　せっせと　そこはかと　そっと　ぞっと
そと　たんと　ちくと　ちゃんと　ちょうと　ちょこっと　ちんと　ついと　つと
てんと　とくと　どっかと　どっと　とっとと　とんと　ぬっくと　はたと
はっしと　はっと　びくと　ひしと　ひたと　ひょっと　ふっつと　ふと　ぼうっと
ぼけっと　ほっと　まんまと　むさと　むずと　むっくと　むっと　もそっと
もっと　よよと　りゅうと　りんと　わざと　わりと

3 (4)　泣き声・笑い声 [10]
【例】あはは　えへへ　えーん　わーん

さらに、規程集で明示されていない以下の場合もオノマトペとみなし、1 短単位とした上で、語彙種別「オノマトペ」を付与する。

2 (5)′　2 (5) の語中に長音が挿入されたもの

U　1 (1)〜(5)、2 (5) のいずれにも当てはまらないが、オノマトペにあたる語彙素の異語形として UniDic に登録されているもの

[10] BCCWJ 規程集では、笑い声は感動詞、鳴き声は副詞と品詞を認定している。

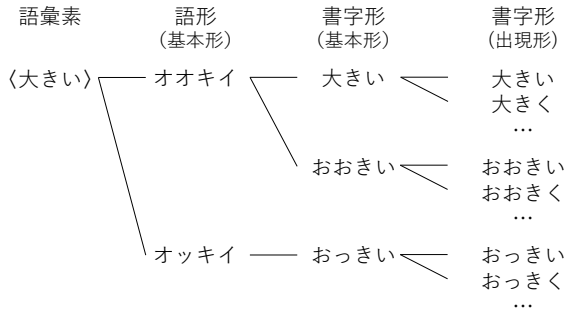

図 1　UniDic の階層的な見出し構造。語彙素は異なる語形を束ね、語形は異なる書字形（表記の違い）を束ね、書字形基本形は活用の異なる書字形出現形を束ねる。

D　オノマトペ辞典に立項されているもの。BQNC のアノテーションでは『擬音語・擬態語辞典』(山口 2002) を用いた。

なお、規程集の規則 1 (5) a および b に該当する擬音語・擬態語と一般語とが結合した語には、語彙種別「オノマトペ」を付与しない（例：「ざく切り」「ピリ辛」「ムカつく」）。

オノマトペの語彙素の認定基準　オノマトペの各出現形について、規程集下巻 p.119 の 4「同語異語判別」に基づき、同一語彙素か別語彙素かを判断する。同一語彙素と判断された異なる出現形は、異表記とみなし、後述する正規表記付与の対象とする。

4「同語異語判別」によると、同一語彙素（同一語形の場合も含む）とするのは次の場合である。なお、語彙素と語形の関係は、図 1 に示す通りである。

(1)　長音を示す母音、小書きの母音、長音符号の差異
　【例】ぎゅう／ぎゅー／ぎゅぅ ⟶ 語形 ギュウ

(2)　母音連鎖の「エイ」と「エエ」、「オウ」と「オオ」の差異
　【例】ぜいぜい／ぜえぜえ／ぜーぜー ⟶ 語形 ゼイゼイ

(3)　長音・促音とそれが連鎖している形との差異
　【例】ばたーん／ばたーーーん／ばたぁーん ⟶ 語形 バターン
　　ばったん／ばっったん ⟶ 語形 バッタン

(4)　単独の促音と、促音に長音を示す長音符号や小書きの母音が付いた形との差異
　【例】ごっ／ごっー ⟶ 語形 ゴッ

(5)　語末に促音が付加されたものと元の形との差異
　【例】ぎゅう／ぎゅっ ⟶ 語彙素 ⟨ ぎゅう ⟩

加えて、規程集で明示されていない次の場合についても、UniDic での定義に従い、同一語彙素とする。

(6)　漢字、ひらがな、カタカナ表記の差異

【例】ふわふわ／フワフワ ⟶ 語彙素⟨ふわふわ⟩
態と／わざと／ワザと ⟶ 語彙素⟨態と⟩

(7) その他、UniDic において同一語彙素の異なる出現形として登録されているもの
【例】ゆっくり／ゆつくり／ゆーっくり／ゆ〜っくり ⟶ 語彙素⟨ゆっくり⟩
ぐんと／ぐーんと／ぐうーんと ⟶ 語彙素⟨ぐんと⟩
きちんと／きちっと ⟶ 語彙素⟨きちんと⟩

一方、4「同語異語判別」によると、異なる語彙素とするのは次の場合である（上記 (7) と競合する場合は (7) を優先する）。

a. 語中の促音の付加
バタリ ⟷ バッタリ

b. 長音の付加
ギュ ⟷ ギュー　バタン ⟷ バターン

c. 撥音の付加
バタ ⟷ バタン　グニャリ ⟷ グンニャリ

d. 「リ」の付加
バタ ⟷ バタリ

e. 同一音の繰り返し
ズバ ⟷ ズババ ⟷ ズバババ　パン ⟷ パパン ⟷ パパパン
シュルン ⟷ シュルルン ⟷ シュルルルルン　ガ ⟷ ガガ ⟷ ガガガ ⟷ ガガガガ

f. 語基の繰り返し
トントン ⟷ トントントン

g. 清濁
カリカリ ⟷ ガリガリ

h. 濁と半濁
バタン ⟷ パタン

i. 直音と拗音
ビヨン ⟷ ビョン

オノマトペの正規表記の認定基準　以降、表 4 の表記を用いて説明する。各オノマトペの正規表記については、異なる語彙素にまたがらない範囲、つまり同一語彙素の異表記の中で正規表記を定める。この際、語形や語彙素の代表表記を必ずしも正規表記としない。具体的に、以下の基準を定める。

(1) ひらがな、カタカナによる表記（例：「がっかり」、「ガッカリ」）をともに正規表記に含める。語ごとの個別の傾向を考慮せず、オノマトペ全体について一括して両方の表記を

標準的な表記とみなす。ただし、語末が促音のものは、「全体がひらがな」、「全体がカタカナ」、「語末の促音のみひらがな」の場合のみ正規表記に含める（例：「がっ」、「ガッ」、「ガっ」）。また、語末が「と」のものは、「と」の直前の促音について前述の規則を満たした上で「語末の『と』がひらがな」の場合のみ正規表記に含める（例：「ぴたっと」、「ピタッと」、「ピタっと」）。

(2) 語末の促音の有無については、次に指定するものは、促音が付加されていないものと付加されたものの両方を正規表記に含める。

- 1 音のオノマトペ $AQ?$（例：「きゅ」と「きゅっ」）
- 1 音のオノマトペに長音が付加されたもの $ARQ?$（例：「ぎゅう」と「ぎゅうっ」）
- 1 音のオノマトペの連鎖型 $A^nQ?$（例「きゅきゅ」と「きゅきゅっ」）、
- 1 音のオノマトペに促音が付加されたものの連鎖型 $AQAQ?$（例：「きゅっきゅ」と「きゅっきゅっ」）、$AQAQAQ?$（例：「きゅっきゅっきゅ」と「きゅっきゅっきゅっ」）
- 2 音のオノマトペ $ABQ?$（例：「ずど」と「ずどっ」）
- 2 音のオノマトペの語末に長音が付加されたもの $ABRQ?$（例：「ずどー」と「ずどーっ」）
- 2 音のオノマトペの語末音連鎖型 $AB^nQ?$ とそれに長音が付加されたもの $AB^nRQ?$（例：「ずどど」と「ずどどっ」、「ずどどどー」と「ずどどどーっ」）

その他のものは、語末に促音が付加されていないもののみ正規表記に含める。

(3) 長音と母音の差異については、長音符号と母音で表記されたものをともに正規表記に含

表 4　オノマトペの一般形の記述用文字の定義

記号	外延的定義	説明
A		ひらがなまたはカタカナの直音または拗音 1 音
B		同上。ただし A とは異なる音とする。
L	{ ー }	長音符号
V		長音を表す母音（直前がア段の音の場合：{ あ, ア }）
v		長音を表す母音の小書き文字（直前がア段の音の場合：{ ぁ, ァ }）
R	$L \cup V \cup v$	長音（「〜」など長音符号の亜種も含める）
Q	{ っ, ッ }	ひらがなまたはカタカナの促音
N	{ ん, ン }	ひらがなまたはカタカナの撥音
λ	{ り, リ }	「り」または「リ」
T	{ と, ト }	「と」または「ト」
$X?$		X の 0 回または 1 回の出現
X^+		X の 1 回以上の繰り返し
X^n		X の n 回の繰り返し ($n \geq 2$)

める（例：「ぴいぴい」と「ぴーぴー」）。母音の小書き文字で表記されたものや、「〜」などの記号で表記されたものは正規表記に含めない。

(4) 長音の連鎖については、長音が二つ以上連続しないもののみ正規表記に含める（例：「すーっ」「すうすう」）。

オノマトペのアノテーション例　以上の基準に基づく、オノマトペの出現形に対するアノテーション例を付録 A の表 11〜19 に示す。

3.4　感動詞

感動詞については、BCCWJ 規程集下巻 p.121「同語異語判別規程」「細則８　感動詞の扱い」に基づいて品詞「感動詞」の認定を行った上で、挨拶表現かそうでないかに応じて語彙種別および異表記・正規表記の認定について異なる扱いとする。

挨拶の感動詞、つまり 1 語の特定の活用形（例：「ありがとう」）または 2 語以上からなる複合的表現（例：「こんにち+は」）が定型化した挨拶表現とみなされ、規程集により 1 最小単位と認定される語については、語彙種別「感動詞-挨拶」を付与し、異表記に該当する場合は正規表記も付与する。具体的には、品詞「感動詞-一般」として UniDic に登録されている「有り難う」、「おおきに」、「御早う」、「今日は」、「今晩は」、「さようなら」、「さらば」、「初めまして」の 8 語彙素を対象とする。

その他の感動詞については、語彙種別「感動詞-非挨拶」を付与するとともに、語彙素の特定を行い、正規表記の付与は行わない。以下に例を挙げる。

- 「はぁ」→〈はあ〉、「ふふっ」→〈ふふ〉、「フ〜」→〈ふう〉、「わあぁ」→〈わあ〉、「ひーひっひー」→〈ひひひ〉、「きゃっほい」→〈きゃっほう〉

UniDic 未登録語についても、感動詞とみなせるものがあれば適宜認定する。BQNC でのアノテーションでは、以下のような事例について、新たに品詞「感動詞-一般」、語彙種別「感動詞-非挨拶」として語の認定を行った。

- 「オホン」「ギャハハ」「んがっ」「あぼーん」「ばっちこーい」

3.5　外国語

日本語テキスト中で外国語に由来する表現が用いられる場合、「ｐｌａｓｔｉｃ」[11]のように原言語の文字体系で表記される場合と、「プラスチック」や「プラスティック」のようにカタカナなどの日本語の文字体系で表記される場合がある（「合羽」、「煙草」のように漢字表記され、外来語意識が希薄になっているものもある）。前者の原言語の文字で表記されたものを外国語とみなし、語彙種別「外国語」を付与する。後者のように日本語の文字で表記された場合は、日

[11] BCCWJ およびその一部を利用した BQNC では、ASCII 文字を全角文字として収録している。

本語の語彙に取り入れられた借用語ならびに外来語であると判断し、特別な種別は付与しない。

「ＳＮＳ」(Social Networking Service)、「ＬＶ」(level)、といった英大文字の略語については、「エスエヌエス」のように仮名表記することは稀で、日本語テキスト中でも通常英字（ラテン文字）で表記されるため、外国語と認定しないものとする。したがって、「GW」(ゴールデンウィーク)、「ＮＧ」(ノーグッド)、のような和製英語の略語、通常日本語内でしか使用されない略語も外国語とみなさない。ただし、「Ｂｌｏｇ」、「ｂｌｏｇ」(weblog) のように英語でも使用され、英小文字を交えて表記される略語については外国語と認定する。その他、キリル文字で表記されたロシア語の語「икра」(イクラ) なども外国語となる。

「ＶＥＳＰＡ」、「ＮＹ」(New York の意味) のような固有名詞や固有名詞の略語については、語彙種別「固有名」とみなし、外国語と認定しない。同様に日本国内の人物・組織が命名した固有名「ｇｏｏ」なども外国語とみなさない。

古代中国語に由来する語種を指す漢語は、外国語とみなさない。現代中国語の語や中国語由来の語について、表記の一部または全部に通常日本語で使用される漢字が含まれる場合、日本語の語として定着していれば外国語と認定せず（例：「麻婆」)、定着していなければ外国語と認定する（例：「ニイ好」)。

3.6　方言

現代日本語共通語で通常使用されず、限られた地域でのみ使用される方言語彙に該当する語に語彙種別「方言」を付与する。方言語彙に該当する場合、共通語において意味的に対応する語の異表記とはみなさい。たとえば、関西・大阪方言の「おもろい」に対して「面白い」や「おもしろい」を正規表記として付与することは行わない。しかし、方言における標準的な語形から変化した表記であると考えられる場合には異表記と認定し、元の標準的な語形を正規表記として付与する。たとえば、「おもろー」(終止形) には語彙種別「方言」とともに異表記種別「音変化」を付与し、正規表記として「おもろい」を付与する。

方言かどうかの判定のための方言辞書として、『都道府県別全国方言辞典』(佐藤 2009) を参

表 5　「方言」と認定した語の例（辞書見出し語を根拠とするもの）

出現形	文脈	地域　共通語表現	備考
あか | ん	あかんのに〜	京都　だめだ	
おもろー	むっちゃおもろー	大阪　面白い	方言辞書では「おもろい」で立項
ちゃう	こっちちゃうん	大阪　違う	
ほん | で	ほんで僕が	長野　それで, そして	
ほん | じゃ	ほんじゃ、また	長野　それじゃあ	方言辞書では「ほんじゃあ」で立項
ほんま	ほんまにあほやと	京都　本当, 真実	

表 6　「方言」と認定した語の例（辞書例文を根拠とするもの）

出現形	文脈	地域	共通語表現	辞書例文
えー	えーんやけど	大阪	よい, いい	ええかげんにしなはれや
や	えーんやけど	京都	だ	考えとるんや
せ | ん	勉強もせんでいいし	大阪	しない, できない	よーせんのか
だす	父の誕生日だす	三重	です	そーだすのかー
で	書いてあるで	奈良	ぞ, よ	おやっさんにどよされるで
でっ | せー	たのしいでっせー	大阪	ですよ	もうそろそろでっせ
とる	頭にきとる	香川	てる, ている	うまげな服きとるのー

照した。表 5 および表 6 に「方言」と認定した語を示す。出現形と同一またはそれに近い表現が辞書の見出し語として存在する場合、表 5 のように「方言」と認定した。辞書の見出し語にない場合には、出現形と同等の表現が何らかの見出し語の例文中に含まれていれば表 6 のように「方言」と認定した。表中の「あか | ん」などは短単位で複数語に分かれることを示しているが、複数語でまとまった表現をなす場合は、構成語すべてに「方言」を付与した。また、前述の「おもろー」に加え、「えー」、「(でっ) せー」は異表記「音変化」と認定し、それぞれ正規表記「ええ」、「せ」を付与した。

3.7　顔文字・ＡＡ

顔文字とアスキーアートは、感情や態度の表現のために用いられる、文字や特殊記号で構成された非言語表現と捉えられる (プタシンスキ 他 2017)。顔文字・アスキーアートに該当する表現に語彙種別「顔文字・ＡＡ」を付与する。

BCCWJ 規程集での顔文字とアスキーアートに関する記述としては、下巻 p.53 の 1.15「補助記号」において、品詞「補助記号-ＡＡ-一般」および「補助記号-ＡＡ-顔文字」と扱う事例が 3 件ずつ挙げられているに過ぎない。

(6) 補助記号-ＡＡ-一般
【例】　ｏｒｚ　　ミ田　　ε＝

(7) 補助記号-ＡＡ-顔文字
【例】　（＾ｏ＾）　　ｍ（．＿.）ｍ　　（＝゜ω゜）ノ

一方、BCCWJ コアデータでは次のような単位で語の認定を行っている。以下、語境界を「|」で表し、各語の品詞については「補助記号-ＡＡ-顔文字」を「F」、「補助記号-ＡＡ-一般」を「A」、「補助記号-一般」を「補」、「記号-文字」を「C」と表示する。

(a)	＾＾ｂ	[F]
(b)	＾＾；	[F]
(c)	ｍ｜（＿＿）｜ｍ	[A, F, A]
(d)	（ｏ゜◇゜）｜ノ	[F, A]
(e)	（＊＾＾）|v	[F, A]
(f)	（＾▽＾）｜／｜☆｜＊｜：	[F, A, 補, 補, 補]
(g)	σ｜（σ＿σ＊）｜ｚ｜ｚ｜ｚ	[A, F, C, C, C]
(h)	―|―|（´・ω・｀）|―|―	[補, 補,F, 補, 補]
(i)	☆彡	[A]
(j)	（；´－ω－）|～ 3	[F, A]

上記 (a)、(b) のように輪郭を表す括弧がない顔文字については、手「ｂ」や、汗または頬「；」を表す部分を含めて 1 語としている。(c)〜(g) のように輪郭を表す括弧を持つ顔文字については、輪郭の左右に隣接する手・腕を別語として切り離して「補助記号-ＡＡ-一般」を付与しており、規程集の例と異なる扱いをしている。(f)〜(h) に含まれる輪郭の外側の装飾的表現については、英字など文字にあたるものは「記号-文字」とし、その他の記号類は「補助記号-一般」としている。また、(i)、(j) の「流れ星」、「ため息」といった顔以外の概念を表現したアスキーアートについては、顔文字に隣接しているか否かによらず「補助記号-ＡＡ-一般」としている。

以上のように、規程集とコアデータで一部相違点が見られるが、実際のアノテーション例であるコアデータから推測される付与基準を踏襲し、本基準では顔文字、アスキーアートの認定を次のように行う。

(1) **アスキーアート**：顔以外の体の一部や顔を含む体全体、その他の概念を表すアスキーアートは、該当する範囲を 1 語とし、品詞「補助記号-ＡＡ-顔文字」、語彙種別「顔文字・ＡＡ」を付与する。
- 「ｏｒｚ」「凹○」「ノシ」

(2) **輪郭のある顔文字**：顔の輪郭を表す括弧がある場合、左右輪郭に該当する括弧までを 1 語とし、品詞「補助記号-ＡＡ-顔文字」、語彙種別「顔文字・ＡＡ」を付与する。壁などで顔の一部が隠れている表現についても、壁にあたる記号を輪郭と同様に扱う。
- 「（ＴｏＴ）」「(-__-｜｜｜)」「｜ω・)」

(3) **輪郭が複数ある顔文字**：顔の輪郭を表す括弧が複数個連続している場合、連続する括弧全体を含めて 1 語とし、品詞「補助記号-ＡＡ-顔文字」、語彙種別「顔文字・ＡＡ」を付与する。
- 「((((　；゜Д゜)))」「((´∀｀)))」

(4) **輪郭外側の目・耳・頬**：輪郭の外側に目、耳、頬（汗）など顔の部位を表す文字・記号がある場合、それらは別語とし、品詞「補助記号-ＡＡ-一般」とする。顔および目・耳・頬にそれぞれ語彙種別「顔文字・ＡＡ」を付与する。

- 「・|゜|(∀　)|゜|・」「⊂|(・(ェ)・)|⊃」「(＋ｏ＋)|；」

(5) **輪郭外側の腕・脚**：輪郭の外側に腕や脚を表す文字・記号がある場合、それらは別語とし、品詞「補助記号-ＡＡ-一般」とする。顔および腕・脚にそれぞれ語彙種別「顔文字・ＡＡ」を付与する。

- 「ｖ|(゜０＾＊)|ｖ」「ι|(´Д｀υ)」「＼|(◎ｏ◎)|／」
  「ｄ|(・・。)」「＼|(≧▽≦)|ﾉ」「(≧▽≦)|ゞ」
  「ヽ|(#゜Д゜)|ノ|┌┘」、「ヽ|(゜д。ヽ≡ノ゜д。)|ノ」

(6) **輪郭外側の動き表現**：輪郭の外側に動きや残像等を表す文字・記号がある場合、それらは別語とし、適切な品詞（「記号-文字」や「補助記号-一般」など）を付与し、特に語彙種別を付与しない。ただし、アスキーアートと認定できる箇所には語彙種別「顔文字・ＡＡ」を付与する。

- 「(＾__－)|－|☆」「(Ｔ__Ｔ)|／|￣|￣|￣」
  「(〃＾∇＾)|ｏ|__|彡☆」「☆|⌒|(　＊＾-゜)|ｂ|ネッ|！」

(7) **輪郭外側の感動詞・オノマトペ**：輪郭の外側に感動詞などの台詞やオノマトペなどの言語表現がある場合、それらは別語とし、適切な品詞を付与した上で、場合には語彙種別「感動詞」や「オノマトペ」を付与する。

- 「(＾ー＾＊　)|フフ|♪」「(ｏ__　__)|ｏ|バタッ」

(8) **2 個体に分離できる顔文字**：同一個体または別個体を表す顔文字が連続し、2 個体に分離できる場合、語として分割した上で、それぞれ品詞「補助記号-ＡＡ-一般」、語彙種別「顔文字・ＡＡ」を付与する。

- 「ｏ|(＾－＾ｏ)|(ｏ＾－＾)|ｏ

(9) **輪郭のない顔文字**：左右両方または一方について顔の輪郭を表す括弧が存在しない場合、目、鼻、口、頬、耳など顔の部位を表す範囲全体を 1 語とし、品詞「補助記号-ＡＡ-顔文字」、語彙種別「顔文字・ＡＡ」を付与する。腕、脚など顔の部位以外を表す文字・記号は別語とし、品詞「補助記号-ＡＡ-一般」とした上で語彙種別「顔文字・ＡＡ」を付与する。

- 「＾＾」「￣▽￣；」「@@；」「＞__＜」「＾＾；；」「(＾＾；；」「Ａ|＾＾；；」
  「＾＾|ｖ」」「＞__＜|ノ゛」「ｍ|－－；|ｍ」

ただし、UniDic 登録語は、それにしたがう。

- 「(＾＾ゞ」「ｆ＾__＾；」

(10) **その他**：括弧文字については、顔文字やアスキーアートではなく言語表現と判断し、括弧と括弧内部の文字に分けた上で、内部の文字には「名詞-普通名詞-一般」など適切な品

詞を付与する。

- 「(| 笑 |)」「(| 爆 |)」「(| 泣 |)」「(| 汗 |)」

3.8 補足

BQNC のアノテーションにおいて、§3.1–§3.7 で挙げた 7 種の語彙種別の他に、古語・歴史的仮名遣い（「ませう」、「思ふ」）や、キャラ語尾 (定延 2007) のような特殊な文末表現（「(問題) にゃ (のかな)」、「(どうなり) ましゅ (か)」）の事例が少数存在した。これらは、UGT に特有かつ解析上問題となる語のクラスを構成し得るため、今後、出現状況に応じて語彙種別として定義することも視野に入れる。

オノマトペについては、BCCWJ 規程集に基づいて同一語彙素を超えない範囲で正規表記を定めたが、語彙素をまたがって表記・意味・用法が近い語を正規化したい場面があり得、またそのように語彙素を超える正規化を行ったシステム出力を、誤った正規化と判定してしまうことも問題となり得る。オノマトペに関して類似の語彙素を束ねる上位クラスの概念を導入し、応用上より適切な評価を可能にするなどの対応が考えられる。

顔文字については、BCCWJ および UniDic の実態に沿うよう、一つの顔文字・アスキーアートを複数語に分ける基準を定めた。この基準は、類似する要素を持つ顔文字を検索する際などには有用と考えられるものの、連続する顔文字・アスキーアートのまとまりを一つの単位として抽出しづらくなる問題や、補助記号が細かく分割されていることで直感よりも高い単語分割精度が計算されてしまう問題などが起きうる。別途、顔文字・アスキーアートをチャンキングする方法を用いたり、一般の語と顔文字・アスキーアートに関する語を分けて評価するなどの対応が考えられる。

4 異表記種別の定義と正規表記・正規化対象表記の認定基準

表記揺れの概観 UGT に特有の表記揺れに注目する前に、新聞・雑誌などの出版物に見られる表記揺れとして知られているもの (国立国語研究所 1983) を挙げる。

【同一文字体系内の揺れ】

(i) 異なる漢字の対立（「特徴」と「特長」）
(ii) 送り仮名の対立（「行った」と「行なった」）
(iii) 表音文字列における異なる文字の対立
  a 仮名遣いの対立（「いなずま」と「いなづま」）
  b 外来語のカタカナ表記の対立（「ドライブ」と「ドライヴ」）
  c ローマ字表記の対立（「ＨＵＪＩ」と「ＦＵＪＩ」）

【異なる文字体系内の揺れ】

(iv) 漢字とひらがなの対立（「工夫」と「くふう」）
(v) 漢字とカタカナの対立（「硫黄」と「イオウ」）
(vi) ひらがなとカタカナの対立（「うなぎ」と「ウナギ」）
(vii) カタカナとローマ字の対立（「ワット」と「W」）
(viii) 漢数字と算用数字の対立（「百メートル」と「一〇〇メートル」と「１００メートル」）
(ix) 文字と繰り返し記号の対立（「人人」と「人々」）
(x) 文字とその他の記号類の対立（「パーセント」と「%」）

上記のうち、標準的でない表記が用いられると、他の語と部分文字列が一致することで解析誤りの要因となりやすい項目として、(iii)-a 仮名遣いの対立、(iii)-b 外来語のカタカナ表記の対立、(iv) 漢字とひらがなの対立、(v) 漢字とカタカナの対立、(vi) ひらがなとカタカナの対立、が挙げられる。本基準では、これらの項目を注目すべき異表記と捉え、(iii)-a、(iii)-b を異表記種別「**代用表記**」、(iv)、(v)、(vi) を異表記種別「**異文字種**」として扱う。

また、UGT に特徴的な表記の揺れとして、「たのしー」（たのしい）、「ぁなた」（あなた）、「あやぃい」（あやしい）のように通常とは異なる文字を用いた表記や、「ありがと」（ありがとう）、「ほっっっとんど」（ほとんど）、「あちい」（あつい）のように元の表記に対して発音の変化を伴う表記が使用される (鍛治 他 2015; 大崎 他 2017)。そこで、前者を異表記種別「**代用表記**」、後者を異表記種別「**音変化**」として扱う。その他に、入力誤りや変換誤りなどに起因する誤った表記と考えられるものを異表記種別「**誤表記**」として扱う。

本基準における異表記の定義　同一の語彙素（例：⟨ 大きい ⟩）に対応付けられる各出現形（例：「大きい」、「おっきい」、「おおきい」）のうち、異なる書字形基本形に対応付けられる出現形同士（例：「大きい」と「おっきい」と「おおきい」）を異表記とみなす[12]。異表記の中で、後述する基準により標準的な表記と認定されたものを**正規表記**と呼び、また非標準的な表記と認定され、正規化処理の対象となったものを**正規化対象表記**と呼ぶ。つまり、同一語彙素の異表記には、正規表記と、正規化対象表記と、どちらでもないその他の表記が含まれる。

異表記に関するアノテーションの基本方針　次の方針とする。

- 同一語彙素に対応付けられる出現形のうち、BCCWJ における出現頻度の割合（出現割合）が概ね $\theta_{\mathrm{v}} = 0.1$ 以上である出現形をすべて**正規表記**とする。たとえば、語彙素 ⟨ 面白い ⟩ の BCCWJ コアデータでの 136 件の出現中、「面白い」が 67%、「おもしろい」が 30%で、その他の出現形は合計 3%以下であるため、「面白い」と「おもしろい」が正規表記となる。

[12] 同一の書字形基本形の異なる活用形にあたるもの（例：「おおきい」と「おおきく」）は異表記と考えない。

表 7　異表記種別・正規表記の付与例

	出現形	語彙素	語彙・異表記種別	正規表記 ID	正規表記
(a1)	面白い	面白い			
(a2)	おもしろい	面白い			
(a3)	面白ーい	面白い	音変化	5261: 面白い	面白い, おもしろい
(b)	やきゅー	野球	異文字種; 代用表記	38139: 野球	野球
(c)	パニクル	パニクる	スラング; 異文字種	71627: パニクる	パニクる

  - 出現割合を算出するデータとして、コアデータでの頻度が 100 件以上ある語彙素はコアデータでの出現割合を採用し、そうでない場合、非コアデータでの頻度が 100 件以上ある語彙素は非コアデータでの出現割合を採用し、いずれの頻度も 100 件未満の語彙素は、「中納言」での文字列検索なども併用しながら個別に判断する。
  - 新たに正規表記と認定された表記について、正規表記 ID と正規表記全体の対応を正規表記定義表のファイルに追記する（§1 参照）。
- 同一語彙素に対応付けられる出現形のうち、§4.1-§4.4 で後述する基準に基づき、**正規化対象表記**と認定されたものについてのみ、異表記種別と正規表記 ID の二つの情報を付与する。たとえば、語彙素 ⟨ 面白い ⟩ の出現形である表 7 の (a1)〜(a3) のうち、異表記種別「音変化」に該当する正規化対象表記である (a3) についてのみ、二つの情報を付与する。
  - 正規化対象表記の異表記種別は、一つの語に対して複数該当する場合がある。たとえば、表 7 の (b)「やきゅー」は複数の異表記種別「異文字種」と「代用表記」に該当し、(c)「パニクル」は語彙種別「スラング」かつ異表記種別「代用表記」に該当する。このように語彙種別・異表記種別あわせて複数の種別に該当する場合には、半角セミコロン「;」で並記する。

正規表記と正規化対象表記の認定基準　異表記種別全体に対し、正規表記と正規化対象表記をどのように認定するかという基準を示す。異表記種別ごとの詳細な認定方法は §4.1-§4.4 で後述する。

(1) **基本規則**：語彙素 l の出現形全体の中で、出現割合が概ね $\theta_{\mathrm{v}} = 0.1$ 以上である出現形 w を正規表記とする。

(2) **異文字種の規則**：語彙素 l において、§4.1 の規則 (1)〜(3) に該当する表記上の特徴を持つ同一発音 p の出現形全体の中で、出現割合が概ね $\theta_{\mathrm{ct}} = 0.05$ 以下である出現形を異表記種別「異文字種」の正規化対象表記とする。

(3) **代用表記の規則**：語彙素 l において、§4.2 の規則 (1)、(3)、(4) に該当する表記上の特徴を持つ出現形と、(2) で示すように外来語のカタカナ表記の出現形全体の中で出現割合が概ね $\theta_{\mathrm{lw}} = 0.05$ 以下である出現形を、異表記種別「代用表記」の正規化対象表記とする。

(4) **音変化の規則**：§4.3 の規則 (0) に基づいて追加の正規表記を認定し、§4.3 の規則 (1)〜(7) に該当する発音・表記上の特徴を持つ出現形を異表記種別「音変化」の正規化対象表記とする。両者の認定にあたっては、規則 (0)〜(7) に記載の通り、出現割合の閾値 $\theta_{\mathrm{ct}} = 0.15$ を考慮する。

(5) **誤表記の規則** §4.4 の規則 (1)〜(5) に該当する出現形を異表記種別「誤表記」の正規化対象表記とし、適切と考えられる語の正規表記 ID と対応付ける。

正規表記 ID の定義方法　アノテーション作業中、正規化対象表記と認定した出現形に付与すべき正規表記 ID、正規表記集合がない場合、それらを定義する必要がある。正規表記 ID の定義と正規表記の関連付けは次の方針で行う。

(a) 通常は、正規表記と認定した出現形からなる正規表記集合を定め、「(語彙素 ID):(語彙素)」の形式の正規表記 ID と対応付ける（例：表 8 (i)）。

(b) 活用語の終止形・連体形に関しては、終止形・連体形の出現形の中で正規表記集合を定めるとともに、「(語彙素 ID):(語彙素)」の形式の正規表記 ID と対応付ける（例：表 8 (ii)）。

(c) 活用語の終止形・連体形以外の活用形に関しては、同一活用形の出現形の中で正規表記集合を定めるとともに、「(語彙素 ID):(語彙素)_(読み)」の形式の正規表記 ID を設定する（例：表 8 (iii)）。

(d) §4.3 音変化の規則 (0)-c に基づき、「音変化」の正規化対象表記ではないと認定された発音 p（$\neq$ 語彙素読み p_0）の出現形に関しては、発音 p の出現形の中で正規表記集合を定め、「(語彙素 ID):(語彙素)_(読み)」の形式の正規表記 ID と対応付ける（例：表 8 (iv)）。

(e) §4.3 音変化の規則 (7) に基づき、融合後の語形（例：「ちゃう」）が融合前の語形（例：「て | しまう」）と異なる語彙素として立項されている場合、融合後の語彙素の出現形の中で正規表記集合を定めるとともに、(a)、(b) または (c) に応じた形式の正規表記 ID を設定する（例：表 8 (v)）。

(f) §4.3 音変化の規則 (7) に基づき、融合後の語形（例：「ありゃ」）が融合前の語形（例：「あれ」）と同一の語彙素として扱われている場合、融合前・後の出現形が属する語彙素の中

表 8　正規表記定義の例

	正規化対象表記	品詞	正規表記 ID	正規表記集合
(i)	みーんな	名詞	36678: 皆	皆, みんな, みな
(ii)	ゆー	動詞	1571: 言う	いう, 言う
(iii)	ゆっ	動詞	1571: 言う_イッ	いっ, 言っ
(iv)	ン	助詞	28989: の_ン	ん
(v)	ゃい	助動詞	23635: ちゃう_チャイ	ちゃい
(vi)	ありゃぁ	代名詞	1270: 彼れ_ありゃ	あれ

で正規表記集合を定めるが、「(語彙素 ID):(語彙素)_(読み)」の形式の正規表記 ID と対応付ける（例：表 8 (vi)）。

4.1 異文字種

使用文字種の差異によって異表記となる出現形が「異文字種」の候補である。ただし、これら異表記のうち同一語彙素中で出現割合が大きいものは、形態素解析器の学習用コーパスに多くの事例が含まる、形態素解析用辞書に登録されているといった理由で、解析誤りを引き起こす可能性は低いと考えられる。そこで、使用文字種の違いで異表記となった同一語彙素かつ同一発音の出現形全体のうち、出現割合が概ね $\theta_{\mathrm{ct}} = 0.05$ 以下の出現形を「異文字種」の正規化対象表記と認定する。一方、出現割合が θ_{ct} を超える出現形については特に情報を付与せず、正規化の対象としない。

具体的に異文字種として認定するのは以下の三つのケースである。

(1) **和語・漢語**では、漢字、ひらがな、カタカナ表記の違いによって生じた異表記を対象とする。たとえば、語彙素「素直」については、（出現割合の小さい）「すなお」、「スナオ」が正規化対象表記となる。

表 9 漢字、カタカナ、ひらがな表記の出現割合の分布

ひらかな・カタカナ表記				漢字表記			
ある	0.997			**有る**	0.001		
ない	0.994			**無い**	0.006		
たくさん	0.90			沢山	0.08		
くる	0.79			来る	0.21		
いい	0.57	よい	0.22	良い	0.09		
だいたい	0.47			大体	0.53		
みる	0.42			見る	0.56	観る	0.01
がんばる	0.41			頑張る	0.58		
つくる	0.39			作る	0.55	**造る**	0.04
いっしょ	0.25			一緒	0.75		
なん	0.23	なに	0.13	何	0.62		
ぜんぜん	0.18			全然	0.80		
ちがう	0.17			違う	0.82		
へん	0.13	ヘン	0.01	変	0.76		
すなお	0.06	スナオ	0.004	素直	0.94		
すき	0.02	スキ	0.003	好き	0.97		
ちから	0.004	チカラ	0.001	力	0.994		
すすめる	0.001			進める	0.999		

- ただし、漢字を含む異表記について、漢字・音訓が常用漢字表[13]に記載されているなど、社会生活において広く使用されていると考えられる場合、正規化対象表記とみなさない。たとえば、「有る（ある）」、「無い（ない）」、「造る（つくる）」はこの条件に該当するため、正規化対象表記としない（表 9 のゴシック体箇所）。

(2) **外来語**では、カタカナ、ひらがな表記の違いによって生じた異表記を対象とする。たとえば、語彙素「カード」（card）については「かーど」や「かあど」、語彙素「ザッツ」（That's）については「ざっつ」などが正規化対象表記となる。なお、英字表記の出現形（例：「Ｌｏｃｋ」）は語彙種別「外国語」と認定し、カタカナ表記（例：「ロック」）と対応付けることは行わない。

(3) **英字による短縮語**では、英大文字、英小文字表記の違いによって生じた異表記を対象とする。たとえば、語彙素「ＵＲＬ」については「ｕｒｌ」などが正規化対象表記となる。

ただし例外として、

- 「ニッポン | 放送」の「ニッポン」、「あっ | ！ | と | おどろく | 放送 | 局」の「おどろく」など固有表現を構成する語の出現形は、出現割合によらず正規化対象表記とみなさない。
- 「オオ | ルリ」、「さとう | きび」、「オ | ススメ」、「カッコ | イイ」など複合語を構成する語の出現形は、複合語単位での出現割合が十分大きければ正規化対象表記とみなさない。

参考に、和語・漢語十数語について、表記の出現割合の分布を表 9 に示す。異表記のうち正規化対象表記に該当する出現形を下線で示した。

4.2 代用表記

語の一部が特殊な文字等で代替された出現形を異表記種別「代用表記」の正規化対象表記と認定する。具体的には、以下に該当する現象により生じた出現形を正規化対象表記とする。

(1) **仮名遣いによる表記の違い**：現代仮名遣いによる一般的な表記と異なる仮名遣い（歴史的仮名遣い含む）により生じた異表記。

- 「思う」→「思ふ」、「ましょう」→「ませう」、「まじ」→「まぢ」、助詞「は」→「わ」など。
- UGT でよく見られるものとして、四つ仮名（「じ」、「ぢ」、「ず」、「づ」）を入れ替えたり、助詞「は」、「を」、「へ」を「わ」、「お」、「え」と表音的に表すような表記があるが、これらを誤用かどうか判断するのは自明でないため、後述する異表記種別「誤表記」ではなく「代用表記」とする。

(2) **外来語のカタカナ表記の違い**：外来語のカタカナ表記に関する、慣用的な表記を用いるか（例：「アイデア」）、原音に近い表記を用いるか（例：「アイディア」）などの違いによ

[13] https://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/kijun/naikaku/kanji/index.html

る異表記。§4.1 の異文字種の場合と同様に、カタカナ表記の出現形全体の中で、出現割合が概ね $\theta_{\mathrm{lw}} = 0.05$ 以下の出現形を正規化対象表記とする。

- たとえば、語彙素「オーケストラ」(orchestra) の出現形は「オーケストラ」が 99% 以上を占めており、「オオケストラ」は正規化対象表記となる [14]。
- 「ベッド」(bed) を「ベット」、「スタッドレス」(studless) を「スタットレス」と表記するような清濁の違いについては、後述する「音変化」ではなく「代用表記」の候補とし、出現割合を基に正規化対象表記を判定する。

(3) **仮名に関するその他の表記の違い**

a **長音の表記の交替**：長音を表す文字（母音、小書きの母音、長音符号、それに準じる記号）が入れ替わった異表記。

- 「おいしい」→「おいしー」、「チャレンジャー」→「チャレンジャ〜」など。
- エ段音に続く「い」と「え」の交替（「ねえ」(姉) →「ねい」）、オ段音に続く「う」と「お」の交替（「そう」→「そお」、「とおり」→「とうり」）も対象とする。

b **仮名と小書き文字の交替**により生じた異表記。

- 「わたし」→「ゎたし」、「ちゃん」→「ちやん」など。

c **英字・記号等による音韻的・視覚的な代用**により生じた異表記。

- 「いい」(良い) に対する「E」、「(気分が) 下がる」に対する「↓る」、「こんにちは」に対する「こωにちは」など。
- 「言う」や「いう」を「ゆう」とする異表記 [15]、「アップ」(up) を「うｐ」とする異表記も本区分に含める。

(4) **発音の変化を伴わない記号等の挿入**：特に発音しない記号や空白文字が挿入された異表記。

- 「ファイト」に対する「ファ・イ・ト」、「こども」に対する「こ␣ど␣も」（全角空白を「␣」で表示）など。
- すでに長音や促音がある語にさらに長音や促音が追加された異表記も対象とする。「しいん」に対する「しいーーん」、「やっと」に対する「やっっっと」など。

4.3 音変化

語の一部が異なる文字に変更される、削除される、新たな文字が挿入されるといった要因により、発音の変化が生じた出現形を異表記種別「音変化」の正規化対象表記とする。具体的に

[14] 外来語の表記の変化に発音の変化が伴うかについては、使用者や状況に依存する面があり、本ケースに該当する異表記は後述する「音変化」には含めないこととした。たとえば、「コンピュータ」という表記を「コンピューター」のように読み上げたり、「ヴォイス」という表記を [vo] ではなく [bo] と読み上げるという状況は十分考えられる。

[15] 「言う」の発音は通常「ユー」であるとされることから、発音上の変化は生じておらず「音変化」ではないと判断した。(https://www.nhk.or.jp/bunken/research/kotoba/20160801_2.html)

は、以下に該当する現象により生じた出現形を正規化対象表記とする。

(1) **促音化**：語の一部が促音「っ」に置き換えられて生じた出現形。
- 「どこ (か)」→「どっ」、「です」→「っす」など。

(2) **撥音化**：語の一部が撥音「ん」に置き換えられて生じた出現形。
- 「これ (だけ)」→「こん (だけ)」、「すみ (ません)」→「すん」など。

(3) **母音の交替**：語を構成する母音（二重母音、母音＋長音、半母音 [y] および [w] を含む[16]）が他の母音に交替することで生じた出現形。
- 「もったい (ない)」→「もったえ」、「すごい」[-oi]→「すげえ 」[-ee]、
「わたし」[wa-]→「あたし」[a-]、「かわいい」[-ii]→「かわゆい」[-yui] など。

(4) **子音の交替**：語を構成する子音が他の子音に交替することで生じた出現形。
- 「ぴったり」[-ri]→「ぴったし」[-shi]、「くさい」[-sa-]→「くちゃい」[-cha-] など。

(5) **脱落**：語を構成する V（母音）や CV（子音+母音）のモーラ、特殊モーラ（長音、促音、撥音）に相当する文字が脱落して生じた出現形。
- 「でしょう」→「でしょ」、「ところ」→「とこ」、「ちょっと」→「ちょと」）など。

(6) **長音・促音・撥音の挿入**：語頭、語中または語末に長音、促音、撥音などが挿入されて生じた出現形。
- 「ファイト」→「ファイトォー」、「なぜ」→「なぁぜ」、「きつい」→「きっつい」、「すごく」→「すんごく」など。

(7) **融合**：連語などが融合して 1 語として扱われる語形・出現形は、融合前の複数語に復元することは考えず、融合後の出現形の中で正規表記、正規化対象表記を定める。
- 融合形助動詞「じゃ」（「で | は」）、「てる」（「て | いる」）、「てく」（「て | いく」）、「とく」（「て | おく」）、「ちゃ」（「て | は」）、「ちゃう」（「て | しまう」）、「つう」（「と | いう」）、「てらっしゃる」（「て | いらっしゃる」）、「なきゃ」、「なけりゃ」（「なけれ | ば」）などは、UniDic で融合前と異なる語彙素として立項されている扱いに準じ、融合前の語形からの「音変化」とみなさない。
- 「痩せれ (た)」（痩せ | られ）などのら抜き言葉も、UniDic の扱いに準じ、融合前の語形からの「音変化」とみなさない。
- 「こりゃ」、「こりゃあ」（「これ | は」）などは、UniDic での扱いと同じく語彙素「此れ」の異表記「音変化」とみなし、正規表記「これ」を付与する。同様に、「にゃあ」（「に | は」）、「のぁ」（「の | は」）も語彙素「に」、「の」の異表記「音変化」とみなす。

ただし、元の表記から発音の変化が生じた異表記であっても、広く使用され、特に解析誤り

[16] 注目する発音の一部をローマ字表記で「[]」内に示す。

の要因とならないものもある。このような場合、同種の音変化が生じた出現形全体の出現割合が高ければ「音変化」とみなさないものとする。具体的には、以下の規則 (0) を設ける。

(0) 語彙素 l の読みを p_0 とし、語彙素 l の出現形全体の中での各出現形 w の出現割合を r とし、互いに同一の発音 p である出現形全体の集合を W_p とする。

a 4 節の基本規則に基づき、出現割合が概ね $\theta_{\mathrm{v}} = 0.1$ 以上の出現形を正規表記集合に追加する。

b 発音 p_0 について、各出現形 $w \in W_{p_0}$ のいずれも正規表記集合に含まれない場合、W_{p_0} 中で出現割合が最大の出現形を正規表記集合に追加する。

c 発音 $p \neq p_0$ について、各出現形 $w \in W_p$ の出現割合 r の和 $R_p = \sum_{w \in W_p} r$ が概ね $\theta_{\mathrm{sc}} = 0.15$ 以上である場合、各 w を「音変化」の正規化対象表記とみなさない。

- さらに、出現割合 r が概ね $\theta_{\mathrm{sc}} = 0.15$ 以上の出現形 $w \in W_p$ を正規表記集合に追加する。
- 他の異文字種の規則に該当する出現形は正規化対象表記となる。

d 発音 $p \neq p_0$ かつ出現割合の和 R_p が概ね $\theta_{\mathrm{sc}} = 0.15$ 以下である W_p において、音変化の規則 (1)〜(7) に該当する発音・表記上の特徴を持つ出現形 $w \in W_p$ を異表記種別「音変化」の正規化対象表記とする。

以下、「()」内に BCCWJ コアデータまたは非コアデータでの出現割合を示しつつ、正規表記および異表記種別「音変化」の正規化対象表記とした具体例を挙げる。

- 準体助詞「の」は、「の」(0.83)、「ん」(0.18) とも正規表記とする。
- 終助詞「もの」は、「もの」(0.65)、「もん」(0.35) とも正規表記とする。
- 接続助詞「けれど」は、「けれど」(0.52)、「けど」(0.48) とも正規表記とする。
- 接続詞「けれど」は、「けれど」(0.83)、「けど」(0.16) とも正規表記とする。
- 接続詞「で」は、「で」(0.43)、「じゃあ」(0.29)、「じゃ」(0.22) を正規表記、「んで」(0.01)、「じゃー」(0.001) などを正規化対象表記とする。
  - 「じゃあ」および「じゃ」は、「で | は」が融合した発音・表記と考えられるため、正規表記 ID を「25516: で_ジャア」と発音を付加した形式として正規表記定義表に登録し、正規化対象表記「じゃー」には同 ID を対応付ける。
- 代名詞「私」（わたし） は、「わたし」(0.80)、「あたし」(0.19) を正規表記とする。
  - 発音「アタシ」について $R_p = 0.19$ であることから、これらの発音に該当する出現形を「音変化」の正規化対象表記としないが、「アタシ」(0.004) は異文字種の出現割合の基準 5%を下回ることから「異文字種」の正規化対象表記とする。
- 代名詞「貴方」は、「あなた」(0.80)、「あんた」(0.17) を正規表記、「あーた」(<0.001) を正規化対象表記とする。
- 副詞「余り」は、「あまり」(0.82) を正規表記、「あんまり」(0.01) を正規化対象表記と

する。

- 副詞「迚も」は、「とても」(0.92) を正規表記、「とっても」(0.01) を正規化対象表記とする。
- 副詞「矢張り」は、「やはり」(0.70) と「やっぱり」(0.28) を正規表記、「やっぱ」(0.01) を正規化対象表記とする。
- 助詞「ばかり」は、「ばかり」(0.95) を正規表記、「ばっかり」(0.04) や「ばっか」(0.006) を正規化対象表記とする。
- 形容詞「ごつい」は、「ごつい」(0.48)、「ごっつい」(0.38) とも正規表記とする。
  - 発音「ゴツイ」について $R_p = 0.55$、発音「ゴッツイ」について $R_p = 0.41$ であることから、これらの発音に該当する出現形を「音変化」の正規化対象表記としないが、「ゴッツい」(0.02) は異文字種の出現割合の基準 5%を下回ることから「異文字種」の正規化対象表記とする。
- 名詞「格好」は、「格好」(0.49)、「カッコ」(0.15)、「恰好」(0.13)、「かっこ」(0.12)、「かっこう」(0.1) を正規表記とし、「かっちょ」、「カッチョ」(合わせて 0.005) を正規化対象表記とする。
  - 発音「カッコウ」について $R_p = 0.73$、発音「カッコ」について $R_p = 0.27$ であることから、これらの発音に該当する出現形を「音変化」の正規化対象表記としないが、「カッコウ」(0.009) は異文字種の出現割合の基準 5%を下回ることから「異文字種」の正規化対象表記とする。
- 名詞「思い切り」は、「思い切り」(0.13) に加え、「思いっきり」(0.69) を正規表記とする。
- 動詞「突き込む」は、「突っ込む」(0.57)、「つっこむ」(0.21)、「突っこむ」(0.163) を正規表記とする。規則 (0)-b に基づき、語彙素読みと同一の発音である出現形の中で出現割合が最大である「突き込む」(0.01) も正規表記とする。

4.4 誤表記

入力誤り、仮名漢字変換誤り、テキスト編集誤り、使用者の知識の誤りなどにより、誤字、脱字、衍字などが生じた出現形を異表記種別「誤表記」の正規化対象表記とする。

ただし、次のものは誤表記としない。

- 規範的でないとされる表記であっても、集団的な使用が認められるもの。たとえば、ら抜き言葉は誤表記としない。
- 使用者が意図的に用いたと考えられる代用表記・音変化。たとえば、「わたし」に対する「ゎたし」は代用表記、「です」に対する「っす」は音変化とする。
- 語の単位を超える文法的な誤り・用法。たとえば、「すごい素敵」の「すごい」は誤表記としない。

以下、誤表記と認定する例を挙げる。

(1) **不要な語**：「どれくれくらい」の「くれ」。このように 2 語の間に挿入された不要な文字列は、1 語と認定し、品詞「ｗｅｂ誤脱」、正規表記「<EMPTY>」を付与する。

(2) **誤入力や使用者の知識の誤り**と考えられるもの：「だだ」（「だだ(･･･) だと少し違ってきます」、正規表記は「ただ」）、「とりあえうず」（正規表記は「とりあえず」）「べ」（「どうすれべ」、正規表記は「ば」）、「ちまみ」（「ちまみに」、正規表記は「ちなみ」）、「わらら」（「わららないこと」、正規表記は「わから」）、「つたい」（「プールが 3 日連ちゃんであるつたいね」、正規表記は「つらい」）、「うら覚え」（正規表記は「うろ覚え」）、「チョ」（「チョミスが多く」、正規表記は「チョイ」）など。

(3) **誤変換**：「盛」（可能盛大、正規表記は「性」）

(4) **入力途中のローマ字を含むもの**：「すｇ」（「かっこよすｇ」、正規表記は「すぎる/過ぎる」）、「そｒ」（「なにそｒ」、正規表記は「それ」）、「ｎ」（「すみませｎ」、正規表記は「ん」）など。

(5) **通常の揺れの範囲を超える外来語のカタカナ表記**：「スパー」（正規表記は「スーパー」）、「コミュニュケーション」（正規表記は「コミュニケーション」）など。

4.5 複合的な異表記

前述の通り、一つの出現形が複数の異表記種別に該当する場合、該当する異表記種別をすべて認定する。表 10 に例を示す。一つ目のブロックには、異文字種と代用表記、ならびに異文字種と音変化の複合的な異表記の例を示している。たとえば、出現形「やきゅー」では、正規表記「野球」が異文字種の異表記「やきゅう」を経て代用表記の異表記「やきゅー」に至ったと判断できる。二つ目のブロックでは、音変化の異表記の例を示している。これらの例は代用表記的な要素（「ヤッヴァい」における「ヴァ」など）を含むが、代用表記的な要素のみあるいは純粋な音変化のみが生じた中間的な状態を仮定すると複雑になるため、音変化とのみ認定した。つまり、**音変化かつ代用表記の可能性がある出現形には、異表記種別「音変化」のみを付与した**。三つ目のブロックには、何らかの語彙種別に該当し、かつ異表記であるものを示した。

参考文献

Higashiyama, S., Utiyama, M., Watanabe, T., and Sumita, E. (2021). “User-Generated Text Corpus for Evaluating Japanese Morphological Analysis and Lexical Normalization.” In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5532–5541, Online. Association for Computational Linguistics.

プタシンスキミハウ，奥村紀之，ジェプカラファウ (2017). 顔文字の現象および研究の概観―記号の遊びが科学されるようになった道―. 人工知能, **32** (3).
松田謙次郎 (2006). ネット社会と集団語. 日本語学, **25** (10).
国立国語研究所 (1983). 現代表記の揺れ. 国立国語研究所報告 75.
大崎彩葉，北川善彬，小町守 (2017). 日本語 Twitter 文書を対象とした系列ラベリングによる表記正規化. 情報処理学会研究報告, **2017-NL-231/2017-SLP-116** (12), pp. 1–6.
小椋秀樹，小磯花絵，冨士池優美，宮内佐夜香，小西光，原裕 (2011). 『現代日本語書き言葉均衡コーパス』形態論情報規程集 第 4 版（上／下）. 国立国語研究所内部報告書.
鍜治伸裕，森信介，高橋文彦，笹田鉄朗，斉藤いつみ，服部圭悟，村脇有吾，内海慶 (2015). 形態素解析のエラー分析. 言語処理学会第 21 回年次大会ワークショップ「自然言語処理におけるエラー分析」.
森山拓郎，渋谷勝己 編 (2020). 明解日本語学辞典. 三省堂.

表 10　複合的な異表記

出現形	中間状態	正規表記	語彙・異表記種別
やきゅー	やきゅう	野球	異文字種; 代用表記
ぶちょぅ	ぶちょう	部長	異文字種; 代用表記
マァマァ	マアマア	まあまあ	異文字種; 代用表記
こおひい	こーひー	コーヒー	異文字種; 代用表記
けこーん	けっこん	結婚	異文字種; 音変化
イパーイ	イッパイ	一杯, いっぱい	異文字種; 音変化
べんと	べんとう	弁当	異文字種; 音変化
ホント	ホントウ	本当	異文字種; 音変化
まぁ〜ッス	まス	ます	異文字種; 音変化
ヨォ	ヨ	よ	異文字種; 音変化
ムリイイイイイー	ムリ	無理	異文字種; 音変化
ス	デス	です	異文字種; 音変化
ぴーぷー	ぴーぷる	ピープル	異文字種; 音変化
ヤッヴァい	~~ヤッバい/ヤヴァい~~	ヤバい	音変化
こんばんわ〜	~~こんばんは〜/こんばんわ~~	こんばんは	音変化
悪ぅーぃ	~~悪ぅーい/悪ぃ~~	悪い	音変化
Ｌａｎｇｕｅｇｅ		Ｌａｎｇｕａｇｅ	外国語; 誤表記
ニホン		日本	固有名; 異文字種
でぃずにー		ディズニー	固有名; 異文字種
ピカチュー		ピカチュウ	固有名; 代用表記
谷ッ		谷	固有名; 音変化
ｉｐｏｔ		ｉＰｏｄ	固有名; 誤表記
おもろー		おもろい	方言; 音変化

山口仲美 編 (2002). 擬音語・擬態語辞典. 講談社.
柴田武 (1956). 集団生活が生むことば. 石黒修也 編『ことばの講座』. 東京創元社.
定延利之 (2007). キャラ助詞が現れる環境. 金水敏 編『役割語研究の地平』. くろしお出版.
佐藤亮一 編 (2009). 都道府県別全国方言辞典. 三省堂.

付録

A　オノマトペのアノテーション例

3.3 節の基準に基づく、オノマトペの出現形に対するアノテーション例を表 11〜19 に示す。各出現形に対し、「オノマトペの語の認定基準」のうちどれを適用してどのように出現形を「導出」したかを示し、語彙素（UniDic に登録されていないものは「〈〉」で囲んだ）、正規表記と、異表記種別（正規化対象表記である場合のみ）を記載した。「導出」では、基本となる 1 音または 2 音の部分に付加された要素を「()」で示した。

表 11　1 音のオノマトペとその派生形

認定基準	出現形	導出	語彙素	正規表記	異表記種別
	AQ?			A,AQ	
1-1-b-1	きゅっ	きゅ (っ)	〈きゅっ〉	きゅ, きゅっ, キュ, キュッ, キュっ	
	$AR^{+}Q$?			AV, AVQ, AL, ALQ	
1-1-b-1	ちゅ〜っ	ちゅう (っ)	〈ちゅう〉	ちゅう, ちゅうっ, ちゅー, ちゅーっ, チュウ, チュウッ, チュウっ, チュー, チューッ, チューっ	代用表記
1-1-b-1	すーーっ	すう (ーっ)	〈すう〉	すう, すうっ, すー, すーっ, スウ, スウッ, スー, スーッ, スーっ	

表 12　1 音連鎖型のオノマトペとその派生形

認定基準	出現形 A^nQ?	導出	語彙素	正規表記 A^n, A^nQ	異表記種別
1-1-b-2	チャチャッ	ちゃちゃ (っ)	⟨ ちゃちゃ ⟩	ちゃちゃ, ちゃちゃっ, チャチャ, チャチャッ, チャチャっ	
1-1-b-2	ドドドドッ	どどどど (っ)	⟨ どどどど ⟩	どどどど, どどどどっ, ドドドド, ドドドドッ, ドドドドっ	
	$ARARQ$?			$AVAV, ALAL$	
1-1-b-3	ピーピー	ぴ (ー) ぴ (ー)	⟨ ぴいぴい ⟩	ぴいぴい, ぴーぴー, ピーピー, ピイピイ	
	$AQAQAQ$?			$AQAQA, AQAQAQ$	
1-2	きゅっきゅっ きゅっ	きゅ (っ) きゅ (っ) きゅ (っ)	⟨⟨ きゅっきゅっ きゅ ⟩⟩	きゅっきゅっきゅ, きゅっきゅっきゅっ, キュッキュッキュ, キュッキュッキュッ	

表 13　2 音のオノマトペとその派生形

認定基準	出現形	導出	語彙素	正規表記	異表記種別
	$ABQ?$			AB, ABQ	
1-1-c-1	ギョロ	ぎょろ	⟨ ぎょろ ⟩	ぎょろ, ぎょろっ, ギョロ, ギョロッ, ギョロっ	
	$ABR^{+}Q?$			$ABR, ABRQ$	
1-1-c-2	ムカ〜〜っ	むか (ー$^{+}$っ)	⟨⟨ むかー ⟩⟩	むかー, むかーっ, ムカー, ムカーッ, ムカーっ	代用表記
	$AR^{+}BQ?$			AVB, ALB	
1-1-c-2	ガーーーン	が (ー$^{+}$) ん	⟨ があん ⟩	があん, がーん, ガアン, ガーン	代用表記
	$ABNQ?$			ABN	
1-1-c-2	キュイン	きゅい (ん)	⟨⟨ きゅいん ⟩⟩	きゅいん, キュイン	
1-1-c-2	ガコンっ	がこ (んっ)	⟨⟨ がこん ⟩⟩	がこん, ガコン	音変化
	$ABQNQ?$			$ABQN$	
1-1-c-2	ぽっちん	ぽ (っ) ち (ん)	⟨⟨ ぽっちん ⟩⟩	ぽっちん, ポッチン	
	$ABR^{+}NQ?$			$ABVN, ABLN$	
1-1-c-2	ずどーん	ずど (ーん)	⟨⟨ ずどーん ⟩⟩		
1-1-c-2	ベロ〜ン	べろ (ーん)	⟨⟨ べろーん ⟩⟩	べろーん, べろおん, ベローン, ベロオン	代用表記
1-1-c-2	チュィ——ン	ちゅい (ー$^{+}$ん)	⟨⟨ ちゅいーん ⟩⟩	ちゅいーん, ちゅいいん, チュイーン, チュイイン	代用表記
	$AQB\lambda Q?$			$AQB\lambda$	
1-1-c-2	しっかり	し (っ) か (り)	⟨ しっかり ⟩	しっかり, シッカリ	
1-1-c-2	にょっこり	にょ (っ) こ (り)	⟨⟨ にょっこり ⟩⟩	にょっこり, ニョッコリ	
	$AB\lambda NQ?$			$AB\lambda N$	
1-1-c-2	とろりん	とろ (りん)	⟨⟨ とろりん ⟩⟩	とろりん, トロリン	

表 14　2 音の末尾連鎖型のオノマトペとその派生形

認定基準	出現形	導出	語彙素	正規表記	異表記種別
	$AB^{n}Q?$			$AB^{n}, AB^{n}Q$	
1-1-c-3	ズドドドド	ずどどどど	⟨⟨ ずどどどど ⟩⟩	ずどどどど, ズドドドド, ずどどどどっ, ズドドドドッ, ズドドドドっ	
	$AB^{n}R^{+}NQ?$			$AB^{n}VN, AB^{n}LN$	
1-1-c-3	ぼよよ〜〜〜ん	ぼよよ (ー$^{+}$ん)	⟨⟨ ぼよよーん ⟩⟩	ぼよよーん, ぼよよおん, ボヨヨーン, ボヨヨオン	代用表記

表 15　2 音反復型のオノマトペとその派生形

認定基準	出現形	導出	語彙素	正規表記	異表記種別
	$ABABQ$?			$ABAB$	
1-2	ツヤツヤ	つやつや	⟨ つやつや ⟩	つやつや, ツヤツヤ	
1-2	とてとて	とてとて	⟨⟨ とてとて ⟩⟩	とてとて, トテトテ	
	$ABABR^{+}Q$?			$ABABV$, $ABABL$	
1-2	ふにゃふにゃー	ふにゃふにゃ (ー)	⟨⟨ ふにゃふにゃー ⟩⟩	ふにゃふにゃあ, ふにゃふにゃー, フニャフニャア, フニャフニャー	
	$AQBAB$			$AQBAB$	
1-2	ぴっかぴか	ぴ (っ) かぴか	⟨⟨ ぴっかぴか ⟩⟩	ぴっかぴか, ピッカピカ	
	$ABQAB$			$ABQAB$	
1-2	ダラッダラ	だら (っ) だら	⟨⟨ だらっだら ⟩⟩	だらっだら, ダラッダラ	
	$ABABABQ$?			$ABABAB$	
1-2	パタパタパタ	ぱたぱたぱた	⟨⟨ ぱたぱたぱた ⟩⟩	ぱたぱたぱた, パタパタパタ	

表 16　複雑な反復型のオノマトペ

認定基準	出現形	導出	語彙素	正規表記	異表記種別
	AQB			AQB	
1-3	ぽっか | ぽっか ↓ ぽっか	ぽ (っ) か	⟨⟨ ぽっか ⟩⟩	ぽっか, ポッカ	
	$AQBR^{+}$			$AQBV$, $AQBL$	
1-3	ワッサー | ワッサー ↓ ワッサー	わ (っ) さ (ー)	⟨⟨ わっさー ⟩⟩	わっさあ, わっさー, ワッサア, ワッサー	

表 17　2 音に語調を整える要素が付いたオノマトペ

認定基準	出現形	導出	語彙素	正規表記	異表記種別
1-5-c	ぱかすか	ぱか (すか)	⟨⟨ ぱかすか ⟩⟩	ぱかすか, パカスカ	

表 18　1 音または 2 音に「と」が付いたオノマトペ

認定基準	出現形	導出	語彙素	正規表記	異表記種別
	AQT			AQ と	
2-5	ムっと	むっと	⟨ むっと ⟩	むっと, ムッと, ムっと	
	$ABNT$			ABN と	
2-5	きちんと	きちんと	⟨ きちんと ⟩	きちんと, キチンと	
	$AR^{+}QT$			AVQ と, ALQ と	
2-5	ボーッと	ぼうっと	⟨ ぼうっと ⟩	ぼうっと, ぼーっと, ボウッと, ボウっとボーッと, ボーっと	
	$AR^{+}?QT$			AQ と	
2-5-x	ず〜っと	ず (ー) っと	⟨ ずっと ⟩	ずっと, ズッと, ズっと	
	$AR^{+}NT$			AN と, AVN と, ALN と	
2-5-x	ぐーんと	ぐ (ー) んと	⟨ ぐんと ⟩	ぐんと, ぐうんと, ぐーんと, グンと, グウンと, グーンと	
2-5-x	ジーンと	じ (ー) んと	⟨ じんと ⟩	じんと, じいんと, じーんと, ジンと, ジインと, ジーンと	
2-5-x	し〜〜〜んと	し (ー$^{+}$) んと	⟨ しんと ⟩	しんと, しいんと, しーんと, シンと, シインと, シーンと	代用表記
	$ABR^{+}QT$			ABQ と	
2-5-x	しらーーっと	しら (ー$^{+}$) っと	⟨ しらっと ⟩	しらっと, シラッと, シラっと	

表 19　その他の特殊なオノマトペ

認定基準	出現形	導出	語彙素	正規表記	異表記種別
	$AR^{+}?QB\lambda$			$AQB\lambda$	
U	ゆ〜っくり	ゆ (ーっ) く (り)	⟨ ゆっくり ⟩	ゆっくり, ユックリ	音変化
U	タ〜ップリ	た (ーっ) ぷ (り)	⟨ たっぷり ⟩	たっぷり, タップリ	音変化
	$A'QB$			AB, AQB, $A'QB$	
U	むっちゃ	め (っ) ちゃ	⟨ 目茶 ⟩	めちゃ, めっちゃ, むっちゃ, メチャ, メッチャ, ムッチャ	