# オノマトペパタンと音韻規則を適用した未知語処理法の提案

馬場　睦也† 　楠　 和馬† 　波多野賢治††

† 同志社大学大学院文化情報学研究科　〒 610–0394 京都府京田辺市多々羅都谷 1–3
†† 同志社大学文化情報学部　〒 610–0394 京都府京田辺市多々羅都谷 1–3
E-mail: †{baba,kusu}@ilab.doshisha.ac.jp, ††khatano@mail.doshisha.ac.jp

**あらまし**　本研究では，音韻論でいわれている音韻規則とオノマトペパタンを用いた未知語処理法を提案する．音韻規則を適用して崩れた表現を元の表現に修正し，オノマトペパタンを適用して新たなオノマトペに対応することで未知語処理の精度向上を試みる．評価実験では，形態素解析器が未知語を既知語であると判定した数と，既知語と判定された未知語のうち正しく判定されている語の数で未知語処理法を提案している先行研究との比較を行う．

**キーワード**　未知語，音韻論，形態論

## 1 はじめに

自然言語処理の研究分野では，日本語のような空白で区切られていない言語を扱う際には，形態素解析を行い分かち書きされた状態にする必要がある．形態素解析とは，文を言語で意味を持つ最小単位である形態素に分割し，各形態素の品詞や活用形などを判定する処理である．形態素の品詞推定や分割箇所の誤りを取り除こうと試みているが完全に無くすことは困難である．この原因の一つとして，形態素解析器が用いる辞書に登録されていない語（以降，未知語）の存在がある．特にソーシャルネットワークサービス (以降，SNS) 上ではネットスラングや砕けた表現が多く見られるため，SNS 上の文書を扱う研究では，大きな問題である．

未知語による形態素解析結果の誤りを取り除くため，これまでに形態素解析器の辞書に未知語を登録する方法 [1] や形態素解析を行う際に未知語を形態素解析器の辞書に登録されている語（以降，既知語）として推定する方法 [2] などが提案されている．しかし，上記の方法では対応不可能な未知語もあるため，笹野らのように，人手で形態素解析結果の誤りを確認し，その傾向から未知語処理の規則（以降，未知語処理規則）を設定する方法 [3] も提案されている．ただし既存手法は，人手で未知語処理規則を形態素解析の誤りから抽出するため，未知語処理規則の抽出漏れを引き起こす可能性が大いに考えられる．

このため，本研究では音韻論と形態論で定義されているオノマトペパタンと音韻規則に基づく未知語処理法を提案をする．ただし，本研究で扱う未知語は，形態素解析器の辞書に登録されている語（以降，既知語）との関係を持たない語であるオノマトペと既知語から派生した語である．音韻論と形態論は，現在に至るまで多くの言語学者によって語の変換規則について研究されてきていることから未知語処理規則に網羅性があるといえる．このため，既存手法と提案手法を比較した際，より効果的な手法になる可能性がある．このことから，オノマトペパタンと音韻規則を用いた未知語処理法を提案してきた [4, 5] が，この手法にはオノマトペパタンにより形態素解析の誤りが引き起こされる問題が残されている．

そこで本研究では，オノマトペパタンにより形態素解析の誤りを分析し得られた誤りの特徴に基づいたコスト調整を行う．また，これまでは音韻規則を入力文にまとめて適用していたため，音韻規則を適用する順によって得られる結果に違いが生じるという問題点があった．これに対しては，各音韻規則を適用した形態素解析結果を統合することで音韻規則を適応する順番に関わらない方法を採る．

## 2 先行研究

未知語処理に関する研究では，人手で文の形態素ごとの分割や品詞情報の付与を行った大規模コーパスから形態素解析器の辞書に登録されていない語を見つけ出し，辞書へ登録する方法が採られることが多い．また形態素解析器の辞書に登録されていない語の切れ目や品詞，語義を統計的に推定し既知語かどうかを判定する方法もしばしば採られている．統計的方法としては，ネットスラングや砕けた表現のような特定分野に特化し，未知語処理の規則を抽出する研究が見られる [6]．このような規則は，Project Next NLP[1]で行われているような形態素解析結果の誤りに基づいて構築されることが多く，笹野らも分析結果をもとに未知語処理規則を構築している [3].

### 2.1 形態素解析結果の誤りに基づく未知語処理法

笹野らは，WEB の文書を JUMAN [7] で形態素解析した際に見られた形態素解析結果の誤りを分析し，「連濁」，「小文字化」，「長音化」，「オノマトペパタン」といった規則を見つけている．これらの規則を適用した未知語処理法（以降，既存手法）を提案しているが，池田らにより，ツイートから得られた形態素解析結果の誤りを分析したところ，長音化や小文字化による未知語よりも表記ゆれや形態素解析器の辞書に未登録の固有名詞が多く見られたことが報告されている [8]．この原因は，形態素解析結果の誤りを分析するデータセットによって構築され

1：Project Next NLP: https://sites.google.com/site/projectnextnlp/（2019 年 2 月 12 日閲覧 ）

表 1　1 モーラのオノマトペの生成パタン

| パタン ID | パタン | 例 |
|---|---|---|
| 1a | CV | ふ |
| 1b | CV と | ふと |
| 1c | CVQ | ふっ |
| 1d | CVN | ばん |
| 1e | CVV | がー |
| 1f | CVVQ | ばーっ |
| 1g | CVVN | ばーん |
| 1h | CVQCVQ | くっくっ |
| 1i | CVNCVN | ばんばん |
| 1j | CVVCVV | がーがー |

る未知語処理規則に差が生まれるためである．このことから，データセットの依存しない未知語処理規則の構築が求められる．

### 2.2　音韻論と形態論に基づいた未知語処理法

既存手法の問題を解決するため，これまでに音韻論と形態論で定義されている音韻規則とオノマトペパタンを用いた未知語処理法を提案している [4,5]．音韻論と形態論は言語学者によって語の構造や変化に関する規則を分析する学問である．このため，既知語から未知語へと変化してしまった語に対する未知語処理規則に網羅性が担保されている．音韻論とは，音韻と語の意味を変化させる働きを持つ音の最小単位で記述できるようにし，音の機能・構造・パタンを分析する理論である．形態論とは，形態素を最小単位とし，語の構造を分析する理論である．このことから音韻論と形態論には，既知語が未知語化する規則に網羅性があり，未知語を効果的に処理できる可能性がある．また，時代とともに派生語や造語が多く発生するといわれており [9]，オノマトペに関しても母音や子音，撥音，促音，長音などの音素の組合せと語の形態に注目した規則化もなされている [10,11]．

これまでの手法にて扱っているオノマトペパタンを表 1，表 2 に示す．日本語において，オノマトペは，モーラの観点からみると 1 もしくは 2 モーラの基本形にまとめることが可能だとされている [10]．モーラとは音韻論でいわれている音の長さを表す概念である [12]．例えば,「モーラ」 は 「モ」,「ー」,「ラ」に分けられるため，3 モーラからなる単語である．1 モーラのオノマトペの生成パタンを表 1 に，2 モーラのオノマトペの生成パタンを表 2 に示す．表 2 中の記号 V は母音，C は子音，Q は促音「っ」，N は撥音「ん」を意味している．$p_1$ と $p_2$ はそれぞれ丸括弧内のパタンが別のパタンであることを意味する．

音韻規則とは，語の変化規則である．このことから，既知語から派生した未知語を自動的に変換することを可能にすると考えられる．音韻規則は多くの種類が見つけられているが，それらの規則の中には曖昧性を含む規則がある．しかしながら，音韻規則の中には曖昧性を含む規則もある．このため，提案している手法では音韻規則の中でも，規則が一般化され，処理の際に曖昧性が含まれていない規則のみを採用する．次に挙げる四つの音韻規則が，用いた規則である．

表 2　2 モーラのオノマトペの生成パタン

| パタン ID | パタン | 例 |
|---|---|---|
| 2a | CVCV | がば |
| 2b | CVCVQ | ばた |
| 2c | CVCVri | ばたり |
| 2d | CVCVN | ばたん |
| 2e | CVQCV | どっか |
| 2f | CVNCV | むんず |
| 2g | CVQCVri | ばっさり |
| 2h | CVNCVri | ぼんやり |
| 2i | CVCVCVCV | ばさばさ |
| 2j | $p_1$(CVCV)$p_2$(CVCV) | どたばた |
| 2k | CVCVriCVCVri | ばたりばたり |
| 2l | CVCVNCVCVN | ばたんばたん |
| 2m | $p_1$(CVCVri)$p_2$(CVCVri) | のらりくらり |
| 2n | $p_1$(CVCVN)$p_2$(CVCVN) | がたんごとん |

表 3　母音融合の規則

| 変化前 | 変化後 | 変化前 | 変化後 |
|---|---|---|---|
| a + i → | e | a + u → | o |
| a + e → | e | a + o → | o |
| i + u → | u | i + e → | i |
| i + o → | u | u + i → | i |
| u + e → | i | u + o → | u |
| e + a → | a | e + i → | e |
| e + u → | o | e + o → | o |
| o + a → | a | o + i → | e |
| o + u → | o | o + e → | e |

表 4　母音融合の例

| 変形前 | 変形後 |
|---|---|
| すごい （sug**oi**） | すげ （sug**e**） |
| さむい （sam**ui**） | さみ （sam**i**） |
| ふるい （hur**ui**） | ふり （hur**i**） |

表 5　重音脱落の例

| 変形前 | 変形後 |
|---|---|
| ナガアメ（長雨） | ナガメ |
| ミチノオク（陸奥） | ミチノク |
| マツウラ（松浦） | マツラ |

**母音融合**

母音融合とは，母音が連続してしまうことを避けるために連続する母音が変化する規則である．唇や舌などの調音器官の状態や音の特性など言語音の性質を表す素性である弁別素性を用いて語の表示を行うことにより，母音融合の一般化が可能であるといわれている [13]．具体的な弁別素性の組合せに基づいた母音融合の変化規則を表 3 に，母音融合の例を表 4 に示す．

**重音脱落**

重音脱落とは，同音の音節が隣接している場合，後方が脱落する規則である [12]．重音脱落の具体例を，表 5 に示す．

**長音の短音化**

長音の単音化とは，語末にある長音が短音化する規則である [12]．長音の短音化の具体例を，表 6 に示す．

**連続同音母音の長音化**

連続同音母音の長音化とは，同音の母音が隣接している場合，先頭の母音以降の母音が長音になる規則である [12]．連続同音母音の長音化の具体例を，表 7 に示す．

表 6　長音の短音化の例

| 変形前 | 変形後 |
|---|---|
| オニンギョウ（お人形） | オニンギョ |
| シンコウ（新香） | シンコ |
| ビンボウ（貧乏） | ビンボ |

表 7　連続同音母音の長音化の例

| 変形前 | 変形後 |
|---|---|
| オカアサン（お母さん） | オカーサン |
| オニイサン（お兄さん） | オニーサン |
| オネエサン（お姉さん） | オネーサン |

しかしながら，オノマトペパタンと音韻規則をただ適用しただけでは，不十分であることが分かっている [4,5]．これは，形態素解析時に用いられる品詞ごとに付与されているコスト（以降，形態素コスト）と隣接する形態素間のコスト（以降，連接コスト）が正しく設定されていないためである．このため，本研究ではオノマトペパタンと音韻規則に適切なコストを調節した未知語処理法を提案する．

## 3　提 案 手 法

本節では，各音韻規則を適用した文の形態素解析結果の統合方法，コストの調整方法について説明した後，提案手法の処理の流れを述べる．

### 3.1　音韻規則を適用した文の形態素解析結果の統合

これまでは，音韻規則を一つずつ順に適用していたため，未知語や未知語を含む文を変換する音韻規則の順番により，結果が異なってしまうという問題点がある．例えば,「ビイル」の正しい表記を「ビール」としたとき，連続同音母音の長音化を適用することで「ビイル」を「ビール」へと元に戻すことができる．しかし先に重音脱落を適用した場合,「ビイル」は「ビル」へと変換され別の語になってしまい，元の「ビール」に戻すことができなくなる．

この問題を解決するため，形態素解析した際に構築されるラティスを統合する方法を採る．ラティスとは図 1 のような形態素解析を行った際に構築される有向グラフである．図中の丸はノードであり一つの形態素を表現している．また，矢印はエッジであり連接する形態素の繋がりを表現している．また各ノードには形態素コスト，各エッジには連接コストが付与されており，これらのコストが最小になる経路が形態素解析結果として求められる．ラティス統合は次の手順で行う．

（1） 入力文を形態素解析し，ラティスを構築する
（2） 入力文に音韻規則を適用し，文を変換する
（3） 変換後の文を形態素解析し，ラティスを構築する
（4） (1)〜(3) のラティスを統合する
（5） (4) で得たラティスの最小コストの経路を求める

(2) では，母音融合を適用した入力文，重音脱落を適用した入力文，長音の短音化を適用した入力文，連続同音母音の長音化を適用した入力文を作成する．(4) では，図 2 のように (3) で得たラティスの統合を行う．

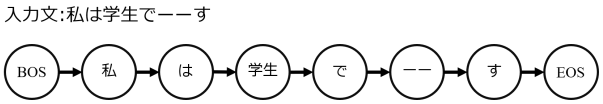


図 1　ラティス構築

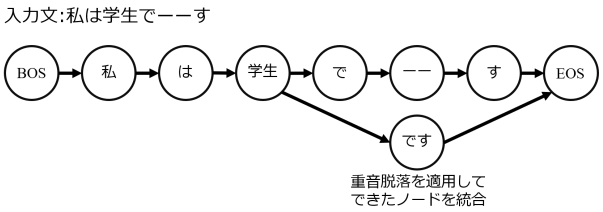


図 2　ラティス統合

### 3.2　形態素解析結果の誤り分析に基づくコスト調整

音韻規則やオノマトペパタンをそのまま適用すると，形態素解析結果の誤判定が生じてしまうことが分かっている [5]．このため，形態素解析結果の誤りを分析し，適切なコストを求める．具体的には，次のようにしてコスト調整を行う．

（1） 未知語 400 個を含むデータセットを作成する
（2） 音韻規則とオノマトペパタンを実装した手法にて，データセットの形態素解析を行う
（3） 処理結果を分析し，誤判定の種類と原因を特定する
（4） コストを調整する

(1) で対象にするテキストデータには，Twitter 社の SNS[2] から得られるツイートをデータとして利用する．これは，Twitter では新語や派生語，砕けた表現が高頻度で使用されており，ツイートとして投稿されることが多いと述べられているためである [14]．ツイートデータの取得には，リアルタイムに投稿されているツイートからランダムにツイートを得ることが可能な，Twitter 社が提供する Application Programming Interface（以降，Twitter API）である Sample realtime Tweets[3] を用いる．

未知語数を 400 件と設定しているがこれは，無作為抽出における母比率と標本比率の許容標準誤差の計算式を変形した下記の式 (1) から算出した．

$$n \geqq Z_\alpha^2 \frac{\hat{p}(1-\hat{p})}{\sigma_p^2} = 384.16 \fallingdotseq 400 \qquad (1)$$

式 (1) 中の $Z_\alpha$ は有意水準 $\alpha$ % における標準正規分布に従う統計量 $Z$，$\sigma_p$ は比率の標準誤差，$p$ は母比率，$\hat{p}$ は標本比率，$n$ は標本サイズを表している．本実験では標本比率 $\hat{p}$ が明らかでないため，標本サイズを最大値にとるように標本比率 $\hat{p} = 0.5$ に設定した．また，有意水準 $\alpha$ は一般的に使用される 5 %に，比率の標準誤差は 5 % 以内に定めた．以上より，標本サイズ

2 : Twitter: https://twitter.com/（2019 年 2 月 12 日閲覧）
3 : Sample realtime Tweets: https://developer.twitter.com/en/docs/tweets/sample-realtime/overview.html（2019 年 2 月 12 日閲覧）

$n$ は式 (1) のように求まった．作成した標本は，ツイート母集団の特徴を抽出した縮図のようなデータ集合であるため，母集団が有するデータの特徴を標本も有すると保証することができる [15]．

ツイート文で使用されている形態素が未知語かどうかの判断を行う形態素解析器には，既存手法が実装されている JUMAN Ver. 7.01 を用いる．この理由は，既存手法が実装されており，また既存手法により処理できなかった未知語を提案手法により既知語であると判定することが出来れば，提案手法の有用性を示すことが可能なためである．

データセットの形態素解析結果の誤りを分析した結果，誤りを 214 件であった．また，誤りの種類をみられた誤り内で割合の高い順に並べたところ，図 3 のようなロングテールであることが分かった．このため，本研究では，上位 8 割に含まれる誤りの種類のコスト調整を行う [16]．

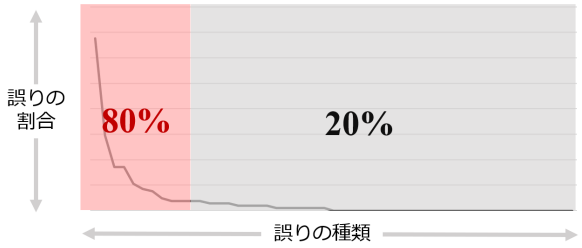


図 3　誤りの種類別割合

上位 8 割に含まれる誤りの種類は以下のようである．

- パタン 2j のオノマトペは見られない
- パタン 1a のオノマトペは見られない
- 母音融合，重音脱落，長音化を適用した形態素解析結果に分かち書き・品詞推定の誤りがみられた
- オノマトペパタンの文字列が隣接している
- パタン 2a は，名詞や形容詞を誤ってオノマトペと推定している

本研究ではこれらの特徴に対して，以下のようにしてコスト調整を行う．

- オノマトペパタン同士の連接コストを最大に設定し，オノマトペパタンが連接する経路を通らないように設定する
- パタン 2j の文字列のコストを未知語よりも大きく設定する
- パタン 2a の文字列のコストを名詞・形容詞よりも大きく設定する
- パタン 1a の文字列のコストを未知語よりも大きく設定する
- 母音融合，重音脱落，長音化された入力文のラティスに適切なコストを設定する

### 3.3　ラティス統合法を反映した未知語処理法の流れ

これまでの未知語処理法にラティス統合法を反映させる．具体的には，以下のようにしてラティス統合を行う．

（1） オノマトペパタン [10] に基づいて文字列を生成する
（2） (1) の文字列を形態素解析器の辞書に登録する
（3） 入力文を形態素解析器で分析し，ラティスを得る
（4） 母音融合の変換規則を適用した入力文に，形態素解析器で分析し，ラティスを得る
（5） 入力文に重音脱落の変換規則を適用した後，形態素解析器で分析し，ラティスを得る
（6） 入力文に長音の短音化の変換規則を適用した後，形態素解析器で分析し，ラティスを得る
（7） 入力文に連続同音母音の長音化の変換規則を適用した後，形態素解析器で分析し，ラティスを得る
（8） 図 2 のように手順 (3)〜(7) で得られたラティスを統合し，ノードとエッジのコストの合計値が最小になる経路を求める

## 4　評価実験

本節では，本研究で提案した音韻論と形態論に基づく方法と文献 [3] で提案された方法による未知語処理性能の比較実験を行う．具体的には，未知語処理の規則がどれだけ未知語を既知語へと変換できているのか（以降，影響度）を各手法の未知語から既知語へと変換された語の数から，未知語処理の規則が未知語から既知語へと変換できた語のうちどれだけ正しく変換できているのか（以降，精度）を未知語から既知語へと変換された語のうち正しく変換されている語の割合から確認する．

### 4.1　実験内容

次の手順で未知語処理の実験を，既存手法と提案手法のそれぞれで行う．

（1） Twitter API を介して未知語が 400 件になるようにツイートを取得し，それを 1 データセットとする
（2） 既存手法にて各データセットの各ツイートの処理を行い，影響度・精度を確認する
（3） 提案手法にて各データセットの各ツイートの処理を行い，影響度・精度を確認する
（4） 既存手法と提案手法それぞれの影響度・精度を比較する

### 4.2　実験結果

データセットの確認をしたところ，表 8 に示す結果が得られた．既知語だと判定された語は，表 8 より同じ件数であることから未知語に対して既存手法と提案手法は，同程度の影響度であることが分かった．また精度は式 (2) で求める．

$$A = \frac{n}{N} \tag{2}$$

式 (2) の $A$ は精度（Accuracy）を，$n$ は正しく既知語へと変換されていた未知語数を，$N$ は既知語だと判定された未知語数を意味する．

表 8 より，既存手法と比較して提案手法の精度は高くないことが分かった．

提案手法で正しく扱えなかった未知語の原因を確かめたとこ

表 8　精度の比較

| 手法 | 既知語だと判定された未知語数 ($N$) | 正しく既知語へと変換されていた未知語数 ($n$) | 精度 ($A$) |
|---|---|---|---|
| 提案 | 84 | 47 | 0.560 |
| 既存 | 84 | 32 | 0.381 |

ろ，未知語をパタン 1e のオノマトペであると誤って判断している場合が 3 割を占めていることが分かった．このことから，長音を扱う音韻規則のコスト調整が正しくできていないことが考えられる．

## 5 おわりに

本研究では，音韻論と形態論に基づく規則を適用した未知語処理法を提案した．これまでにオノマトペパタンと音韻規則を適用した未知語処理法をコスト未調整で行った際に結果の誤りを分析したところ，コスト調整を行わなければ，形態素解析の精度が下がることが分かっている．この問題を解決するためには，形態素解析結果の分析を通して誤りの種類を特定し，コスト調整を行うことで精度を上げる必要がある．コスト調整を行った結果，既存手法と提案手法の影響度は同程度であり，提案手法の精度は既存手法より低かった．提案手法の精度が低い原因として，誤りの種類に取りこぼしがあることが考えられる．このため，各規則が少なくとも 1 件は含まれており，正誤が付与されているデータセットを無作為に抽出することで作成し，勾配降下法を用いてコストの最適化を行う．これは，全規則のコスト調整を行うことで扱う誤りの種類の取りこぼしを防ぐためである．また勾配降下法を用いる理由としては，コスト最適化の学習に使用するデータセット内のデータ数が定まっており，全件に使用できるためである．

## 謝　辞

本研究は，北見工業大学工学部地域未来デザイン工学科テキスト情報処理とインフォマティクス研究室との共同研究によるものである．また本研究の一部は文部科学省私立大学戦略的研究基盤形成支援事業，JSPS 科研費 JP18H03342 の助成および同志社大学大学院文化情報学研究科の研究推進補助金を受けて遂行された．ここに記して謝意を表す．

## 文　献

[1] Shinsuke Mori and Makoto Nagao. Word Extraction from Corpora and Its Part-of-speech Estimation Using Distributional Analysis. In *Proceedings of the 16th Conference on Computational Linguistics*, pp. 1119–1122. ACL, 1996.

[2] Masaaki Nagata. A Part of Speech Estimation Method for Japanese Unknown Words Using a Statistical Model of Morphology and Context. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 277–284. ACL, 1999.

[3] Ryohei Sasano, Sadao Kurohashi, and Manabu Okumura. A simple approach to unknown word processing in japanese morphological analysis. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 162–170. AFNLP, 2013.

[4] 馬場睦也, 楠和馬, 波多野賢治. 日本語の音韻変化規則に基づく未知語処理法の提案. 第 17 回情報科学技術フォーラム (FIT 2018) 講演論文集, pp. 21–24, 2018.

[5] Tokiya Baba, Kazuma Kusu, and Kenji Hatano. An approach for unknown word processing based on japanese phonological rules. In *Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services*, pp. 348–352. ACM, 2018.

[6] Itsumi Saito, Kyosuke Nishida, Kugatsu Sadamitsu, Kuniko Saito, and Junji Tomita. Automatically extracting variant-normalization pairs for japanese text normalization. pp. 937–946. AFNLP, 2017.

[7] Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. Improvements of japanese morphological analyzer juman. In *Proceedings of The International Workshop on Sharable Natural Language Resources*, pp. 22–28. NAIST, 1994.

[8] Taishi Ikeda, Hiroyuki Shindo, and Yuji Matsumoto. Japanese text normalization with encoder-decoder model. In *Proceedings of the 2nd Workshop on Noisy User-generated Text*, pp. 129–137. ACL, 2016.

[9] Manabu Okumura, Atsushi Okumura, and Suguru Saito. Automatic construction of a japanese onomatopoeic dictionary using text data on the www. In *International Conference on Application of Natural Language to Information Systems*, pp. 209–215. Springer, 2006.

[10] 田守育啓, ローレンススコウラップ. オノマトペ —形態と意味—, 第 6 巻. くろしお出版, 1999.

[11] 角岡賢一. 日本語オノマトペ語彙における形態的・音韻的体系性について. くろしお出版, 2007.

[12] 安部清哉, 加藤大鶴, 吉田雅子. 日本語の音. 朝倉書店, 2017.

[13] 窪園晴男. 日本語の音声. 岩波書店, 1999.

[14] Suman Maity, Anshit Chaudhary, Shraman Kumar, Animesh Mukherjee, Chaitanya Sarda, Abhijeet Patil, and Akash Mondal. Wassup? lol: Characterizing out-of-vocabulary words in twitter. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, pp. 341–344. ACM, 2016.

[15] William G. Cochran. *Sampling Techniques*. John Wiley, 3rd edition, 1977.

[16] Vilfredo. Pareto. *Cours d'économie politique professé à l'Université de Lausanne*. Cours d'économie politique professé à l'Université de Lausanne. F. Rouge, 1896.