# オンラインコミュニティにおける単語頻度の通時的変化を利用した新語リストの獲得

阿部 香央莉 [1]　松田 耕史 [2,1]　吉川 将司 [3,2]　乾 健太郎 [1,2]
[1] 東北大学大学院情報科学研究科　[2] 理化学研究所　[3] 東北大学 TCPAI 研究センター
{abe-k,matsuda,inui}@ecei.tohoku.ac.jp　yoshikawa@tohoku.ac.jp

## 1　はじめに

近年，機械翻訳などの言語処理技術の発展は目覚ましいが，それらのシステムに対して未知の単語（未知語）が入力として与えられると，その性能は顕著に落ちることが知られている [1, 2, 3].

未知語の例としては，漫画のタイトルなどの固有名詞（「呪術廻戦」，「あつ森」など）や，特定のコミュニティ内で生まれ幅広く広まった用語（「草」，「エモい」など），医学や法学などの専門用語などがあげられる．これらの語は日々生み出され続けており，大規模データを学習した BERT[4] のようなモデルにおいても，現状の学習済モデルが静的な状態で運用されている以上，これらの新語全てに対応するのは不可能である．したがって，今後の自然言語処理技術の発展に向け，未知語（特に新語）に対応する術を考えることの重要性が増している．

しかし，日本語において新語を対象にした分析を行うための基盤は，我々の知る限り十分に整えられていない．そのため，我々は新語分析基盤を整えるための最初のステップとして，具体的にどのような新語が存在するのかを把握するべく，分析対象となる新語のリストを収集することを試みる．新語リストの収集後には，リスト中の語の体系的分析や，自然言語処理システムの新語に対する頑健性を評価するデータセットの構築等を行っていく．本研究では，日本語における新語のリストをデータ駆動的に獲得することを目的として，時系列の記録が紐づいたテキストデータから (1) 頻度 0 から生起した語の取得 (2) 年々右肩上がりに頻度上昇した語の時系列クラスタリングによる取得の 2 つを試す．結果として，前者で満足に取得できなかった新語を後者で取得できた[1)]ものの，実際に新語リストとして活用するには目視による確認が必要であり，今後の新語研究基盤を整えるためさらに精査する心算である．

1)　詳細なクラスタリング結果は https://chanabe-k.github.io/time_clustering_novel_words を参照.

## 2　関連研究

近年の言語処理分野においては，新語（novel words）や造語（neologism）に着目した研究が広がりつつある．英語では，オンラインコミュニティの Reddit データ [5, 6] や新聞データ [7] を利用した新語についての研究が行われている．Cole ら [5] は，新語の生み出される数が年々増加するという過去の報告通りの現象が Reddit においても生じることを確認している．これは，日々生まれる新語に対処する方法論を考える必要性を示唆する結果である．また，Pinter ら [7] は，新聞データにおいて出現した新語に対し，細かい粒度（混成語，方言，ドメイン特化の専門用語等）での分類を行い，新語データセットの作成を行っている．この研究からもわかる通り，一言で新語と言ってもその中身は多様であり，これら全てに対応するための方法論は自明ではない．既存語の組み合わせや変形で生じた語に関しては，Hofmann ら [6] がある特定の語を語幹として語形変化によって生み出された新語（例：trump → trumpy, trumpish）のまとまりを形態論的家族（morphological family）と呼び，その広がりを推測するという新しいタスクを提案している．また，Schick らは BERT の埋め込み表現と語形を考慮したベクトルを組み合わせて（新語などの）希少語に対応する手法を提案している [8]．さらに，英語だけでなくヘブライ語においても，ヘブライ語の語形成の特徴を考慮した造語自動生成の研究などが行われている [9]．このように，新語に関する研究は言語を跨って行われており，その多くは新語問題に取り組む手始めとして，新語研究を行うための基盤（データセット・タスク）を整えている．しかし，日本語において同様に新語についての分析を行いたいと考えた時に，このようなデータセットなどの資源は現状整備されて

いない．そのため，本研究では新語分析を行うためのデータセット構築に向け，まず分析対象とする新語のリストを獲得することを試みる．

日本語において新語の研究を行いたいと考えた時，日本語ではひらがな・カタカナ・漢字・アルファベットの 4 種類もの文字セットを常用していることにより，英語等の既存研究の枠組みを適用するだけでは新語のモデリングが困難と考えられる．藤井ら [10] の研究では，現状の機械翻訳モデルは異表記（本来漢字等で表すべきところを，ひらがなやカタカナ等で表現している）を含む文に対処できないという指摘がある．このように，一つの概念に対する表現手法が一意に定まらない点が日本語における新語研究を複雑化させていると考えられる．

## 3　新語リストの獲得

新語を対象とした分析を始めるにあたり，まずどのような新語を対象とすべきか，対象とする新語をどう同定するかという点が第一の問題となる．特に日本語では，分かち書きがされていないことにより，文中のどの範囲を新語として取り出すかという課題も生じる．この議論は，Unidic[11] や MeCab の NEologd[12] など，単語分割用辞書の自動獲得の文脈にも近い．本研究では，新語の分割自体は NEologd に頼り，この辞書中で特に幅広く話題となっている新語をどう獲得するか，という点に焦点を当てる．単語分割方法を定めた上で，以下の 3 種類の見方で新語を取得することを考える．

1. 予め新語とわかっている単語
2. ある基準（日時等）を境に，頻度 0 から生起した単語
3. 頻度が年々右肩上がりに上昇した単語

1 は言語学の知見や人手による観察を元にリストを得る方法で，2 および 3 は大規模コーパスを利用してデータ駆動的に得る手法である．しかし，1 の手法では選定する新語が人間の感覚や観察に大きく依存するため，対象としたい新語を大規模に獲得することができない．本研究では，先行研究 [5, 7] でも用いられている 2 と，本研究で新たに試みた 3 の結果を報告する．

### 3.1　データセット

本研究では，通時的な単語頻度の変化を集計・分析することで，新語の隆盛の傾向を捉えることを試みる．このような分析を行うためには，なるべく大規模かつそのテキストがいつ記述されたものかという情報が付加されている，各年代別に分かれたデータセットを分析対象とする必要がある．今回は，この条件に該当するデータとして，Twitter をクローリングして得られたデータ（Twitter データ）および国立情報学研究所のダウンロードサービスにより株式会社ドワンゴから提供を受けた『ニコニコ動画コメント等データ [13]』（ニコニコデータ）を用いた．Twitter データに関しては，2013 年から 2018 年までクローリングした結果から，各年別に一部を抽出した[2]データを使用した．Twitter データ及びニコニコデータの各年ごとのデータ規模を表 1 に示す．

単語頻度の傾向分析をする際，データの規模によって正規化を行った．以降の実験結果におけるグラフの縦軸は，「対象とする単語の頻度 / 総単語数」で算出した割合となっている．

### 3.2　頻度 0 を基準とした新語獲得

ここでは，Twitter およびニコニコデータセットにおいて，ある年を境に出現した新語を抽出して新語リストを獲得することを試みた．具体的な手順としては，表 1 に示している各年コーパスにおいて，コーパス中に出現する全単語を集めた語彙集合（$V_y, y = 2013, ..., 2018$）を作成し，この語彙集合間の差分を取った．例えば，2015 年データセットの語彙集合 $V_{2015}$ から，それまでの（2013〜2014 年）データセットの語彙集合の和 $V_{2013} \cup V_{2014}$ を引いたものを 2015 年に出現した新語とみなした．なるべく広範に使用されている語を獲得したいため，初めて出現した年（上記の例の場合，2015 年時点の頻度）において頻度が 100 より大きいものを抽出した．

表 2 にて，実際に獲得された新語のリストの一部を例として示す．表 2 より，Twitter・ニコニコデータそれぞれにおいて獲得された新語は，数・内容共に大きく異なることがわかった．取得される数は元データの規模に大きく依存していた．また内容に関しては，Twitter データの方がより時事的な話題やテレビの番組等の固有名詞が多いのに対して，ニコニコデータではゲーム・アニメに関する固有名詞が多い傾向がみられた．また，それぞれのデータから抽出された時点では，ノイズとみなされるような単語[3]も多く抽出されていたが，Twitter・ニコニコ

2）　一部抽出する際，http から始まる URL を含むツイートを除外し，ツイート内からユーザ名を除く前処理を行った．
3）　特にニコニコデータでは「弾幕」あるいは「荒らし」と見な

| | | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 |
|---|---|---|---|---|---|---|---|
| Twitter | 行数 | 16,606,755 | 31,621,890 | 25,580,766 | 27,021,802 | 16,102,743 | 5,018,650 |
| | 総単語数 | 220,483,276 | 365,082,739 | 284,907,703 | 331,827,896 | 220,344,924 | 72,563,153 |
| ニコニコ | 行数 | 284,142,829 | 261,880,468 | 273,056,009 | 224,814,667 | 185,160,409 | 164,703,008 |
| | 総単語数 | 1,353,147,623 | 1,350,161,846 | 1,501,833,943 | 1,232,911,450 | 1,037,159,431 | 923,427,432 |

**表 1**　各年ごとの Twitter およびニコニコデータの規模（行数，総単語数）．ニコニコデータは 1 コメント 1 行として数える．総単語数は単語の異なり数ではなく，データ中の全単語数である（実験節におけるグラフの正規化のために用いる）．

| Twitter データから得られた新語リスト | | |
|---|---|---|
| | 総数 | 例 |
| $V_{2014} \setminus V_{2013}$ | 491 | N のために, 水曜日のダウンタウン, 痛バ, ジェラトーニ, ... |
| $V_{2015} \setminus V_{2013-2014}$ | 366 | Alexandros, ゴレライ, 格安スマホ, ウデマエ, バズリズム, BiSH, ... |
| $V_{2016} \setminus V_{2013-2015}$ | 317 | レムりん, 生前退位, オタユリ, トランプ大統領, 魔剤ンゴ, 共感性羞恥 ... |
| $V_{2017} \setminus V_{2013-2016}$ | 170 | 加計学園問題, FE ヒーローズ, 希望の党, MeToo, iPhoneX, バチェラージャパン, PUBG, ... |
| $V_{2018} \setminus V_{2013-2017}$ | 36 | ニンテンドーラボ, 西日本豪雨, にじさんじ, 米朝首脳会談, ルパパト, 万引き家族, ... |
| **ニコニコデータから得られた新語リスト** | | |
| | 総数 | 例 |
| $V_{2014} \setminus V_{2013}$ | 1705 | レットイットゴー, 親ロシア派, 思い出のマーニー, 花子とアン, ゲラゲラポー, ... |
| $V_{2015} \setminus V_{2013-2014}$ | 1364 | がっこうぐらし!, 馬場豊, 平和安全法制, ツクッター, スーパーマリオメーカー ... |
| $V_{2016} \setminus V_{2013-2015}$ | 1389 | ポケストップ, ヨルシカ, ガオガエン, 嘗若, (「°　∀°　)」, 君の名は。, ... |
| $V_{2017} \setminus V_{2013-2016}$ | 1229 | ブルゾンちえみ, ミニスーファミ, ミライアカリ, てるみくらぶ, ... |
| $V_{2018} \setminus V_{2013-2017}$ | 2798 | キュアアムール, ニンテンドーラボ, ゾンビランドサガ, バ美肉, セネガル戦, ... |
| **両データでともに新語と判定された語の例** | | |
| | 総数 | 例 |
| $V_{2014} \setminus V_{2013}$ | 64 | STAP 細胞, SHIROBAKO, 危険ドラッグ, マイルドヤンキー, ダメよ～ダメダメ ... |
| $V_{2015} \setminus V_{2013-2014}$ | 49 | 安保法案, ねこあつめ, 刀剣男士, レモンジーナ, デレステ, 心が叫びたがってるんだ, ... |
| $V_{2016} \setminus V_{2013-2015}$ | 51 | 安倍マリオ, ニンテンドースイッチ, メンタルリセット, 無人在来線爆弾, ... |
| $V_{2017} \setminus V_{2013-2016}$ | 23 | にゃんこスター, ウルトラマンジード, フレンズなんだね, PUBG, ハンドスピナー ... |
| $V_{2018} \setminus V_{2013-2017}$ | 5 | 西日本豪雨, ボブネミミッミ, ヴァルハザク, ニンテンドーラボ |

**表 2**　Twitter およびニコニコデータにおける各年コーパスの語彙集合の差分から得られた新語リスト

データ双方で新語と判定された語（2 つのリストの積集合）を取ることで，より妥当な新語リストを抽出することができた．しかし，積集合を取ることで新語リスト内の単語の総数は大きく減少するため，積集合を取って質を高めることと獲得される新語の数はトレードオフの関係にあると言える．一方で，良く知られた新語である「エモい」「まじ卍」などの表現はこの手法ではうまく獲得できなかった．原因として，「エモい」は 2013 年時点から Twitter データで 198 件，ニコニコデータで 452 件出現しており，当時から既に幅広く使用されていた語彙であること，「まじ卍」に関しては 2013 年時点で頻度 0 だったものの，3～5 年のスパンをかけて徐々にその勢力を拡大しており，頻度の下限で足切りされたことがあげられる．

されるものにより，一般的でない表現（「。・。★・。」など）が新語とみなされ獲得されてしまうという現象が起きていた．

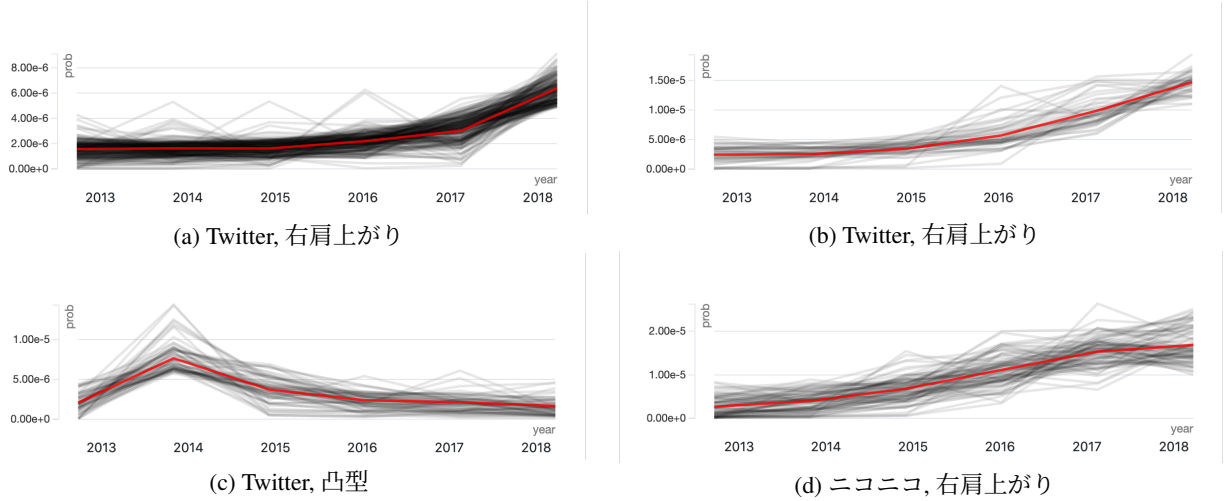
## 3.3　時系列クラスタリングによる新語獲得

頻度 0 から急激に使用頻度が増えた単語を抽出するだけでは，「エモい」等の既知の新語をデータ駆動で獲得することができなかった．そこで，このような単語も抽出する方法として，各年コーパスを比較した際に顕著に頻度が増加した単語を自動抽出する手法を考えた．具体的には，各単語の 2013～2018 年の頻度の推移を時系列データとみなして時系列クラスタリングを行い，単語頻度が上昇傾向にある単語を収集することを試みた．

### 3.3.1　実験設定

Python の `tslearn` ライブラリ[4]における TimeSeriesKMeans（時系列に対応した K-Means 手法）を用いてクラスタリングを行った．このとき，時系

4）https://tslearn.readthedocs.io/en/latest/index.html




(a) Twitter, 右肩上がり　(b) Twitter, 右肩上がり

(c) Twitter, 凸型　(d) ニコニコ, 右肩上がり

**図 1**　Twitter およびニコニコデータの単語に対する時系列クラスタリングの結果．赤線はクラスタの中心線を表す．

列データ間の距離尺度にはユークリッド距離を用いた[5]．クラスタ数は Twitter・ニコニコデータそれぞれにおいて 100，80 とした．実行時間を考慮し，Twitter データでは頻度 2,000 以上かつ 20,000 未満の 23,504 語，ニコニコデータでは頻度 2,000 以上かつ 200,000 未満の 14,586 語に対してクラスタリングを実行した．

### 3.3.2　実験結果

図 1 に，時系列クラスタリングの結果の例を示す．図 1 (a) のクラスタでは，2013 年当時の出現頻度がほぼ 0 に近い単語が，2018 年には 3 倍ほどの頻度に達している様子が読み取れる．このクラスタには 3.2 節で得られなかった「エモい」が含まれているだけでなく，「Bluetooth」「LGBT」など，次第に世間に広まった新語が含まれていることを確認できた．また，同じく右肩上がりの傾向が見られたクラスタ（図 1 (b) ）では，「すこ」「メルカリ」のような単語が見られた．一方で，この右肩上がりのクラスタ内には「アスリート」などの新語ではない単語も見受けられた．同様に，ニコニコデータでも右肩上がりのクラスタから流行した新語（「SSR」「野田内閣」など）のリストを得ることができた（図 1 (d) ）．しかし，Twitter と比べるとニコニコデータでは右肩上がりとみなせるクラスタが少なかった[6]．

時系列クラスタリングの副産物として，一時的に流行した単語リスト（凸型の傾向）や，以前は流行していたが年々衰退している単語のリスト（右肩下がりの傾向）の収集もできた．図 1(c) の凸型グラフのクラスタからは，「集団的自衛権」「アナ雪」などの 2014 年頃に話題となった単語が取得できた．

総じて，異なる 2 種類のデータセットから新語を含む語集合を得ることができたものの，中には一部新語に分類されないものが含まれており，新語分析用のデータセット等を作る上ではここからさらに目視による確認ステップを挟むなどして，その質を高める必要があることがわかった．また，3.2 節と同様に，データセットによって獲得できる新語の種類は異なっていた．これは各データセットの元となったコミュニティの文化やそこで取り上げられる主要な話題に依拠しているものであり，実世界にある新語を網羅的に獲得するには複数のデータセットで横断的に新語を獲得する必要があると考えられる．

## 4　おわりに

日本語において日々生まれ拡散されていく新語について，時系列情報がテキストと紐付いたデータを用い，その通時的な変化を元にデータ駆動で新語リストを獲得することを試みた．今後は，引き続き利用可能な時系列情報が付加されたデータセット（新聞データ等）を利用して新語リストを拡充し，この新語リストを利用した日本語における新語データセットの作成に取り組みたい．またデータセット構築後の次のステップとして，現状の BERT 等のモデルにおいて今回収集したような日本語の新語を扱えているかどうかの分析を行うことも考えている．

**謝辞**　本研究は JSPS 科研費 JP20J21694 の助成を受けたものです．

5）　距離尺度を Dynamic Time Warping (DTW) や Soft DTW とした実験も行ったが，結果に大きな差異は見られなかった．

6）　目視で確認した結果，Twitter データでは 24 / 100，ニコニコデータでは 13 / 80 のクラスタが右肩上がりの傾向を示していた．

# 参考文献


[1] Shoetsu Sato, Jin Sakuma, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. Vocabulary adaptation for domain adaptation in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4269–4279. Association for Computational Linguistics, 2020.

[2] Zorik Gekhman, Roee Aharoni, Genady Beryozkin, Markus Freitag, and Wolfgang Macherey. KoBE: Knowledge-based machine translation evaluation. In *Proceedings of EMNLP (Findings)*, pp. 3200–3207. Association for Computational Linguistics, 2020.

[3] Timo Schick and Hinrich Schütze. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. *Proceedings of AAAI*, Vol. 34, No. 5, pp. 8766–8774, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186. Association for Computational Linguistics, 2019.

[5] J. Cole, M. Ghafurian, and D. Reitter. Word adoption in online communities. *IEEE Transactions on Computational Social Systems*, Vol. 6, No. 1, pp. 178–188, 2019.

[6] Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. Predicting the growth of morphological families from social and linguistic factors. In *Proceedings of ACL*, pp. 7273–7283. Association for Computational Linguistics, 2020.

[7] Yuval Pinter, Cassandra L. Jacobs, and Max Bittker. NYTWIT: A dataset of novel words in the New York Times. In *Proceedings of COLING*, pp. 6509–6515. International Committee on Computational Linguistics, 2020.

[8] Timo Schick and Hinrich Schütze. BERTRAM: Improved word embeddings have big impact on contextualized model performance. In *Proceedings of ACL*, pp. 3996–4007. Association for Computational Linguistics, 2020.

[9] Moran Mizrahi, Stav Yardeni Seelig, and Dafna Shahaf. Coming to Terms: Automatic Formation of Neologisms in Hebrew. In *Proceedings of EMNLP (Findings)*, pp. 4918–4929. Association for Computational Linguistics, 2020.

[10] Ryo Fujii, Masato Mita, Kaori Abe, Kazuaki Hanawa, Makoto Morishita, Jun Suzuki, and Kentaro Inui. PheMT: A phenomenon-wise dataset for machine translation robustness on user-generated contents. In *Proceedings of COLING*, pp. 5929–5943. International Committee on Computational Linguistics, 2020.

[11] 岡照晃. 『国語研日本語ウェブコーパス』からの新規語彙素獲得の試み. 言語資源活用ワークショップ発表論文集, No. 3, pp. 586–592, 2018.

[12] 佐藤敏紀, 橋本泰一, 奥村学. 単語分かち書き辞書 mecab-ipadic-neologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会 (NLP2017), pp. 875–878. 言語処理学会, 2017.

[13] 株式会社ドワンゴ. ニコニコ動画コメント等データ. text(json), 2018. 国立情報学研究所情報学研究データリポジトリ.