# 頻出語ではなく使用者が多い語が自然な日本語である

荒牧英治 * ** 増川佐知子* 宮部真衣 * 森田瑞樹 * *** 保田祥 ****

* 東京大学 知の構造化センター　　** 科学技術振興機構 さきがけ
*** 独立行政法人 医薬基盤研究所　　**** 国立国語研究所
eiji.aramaki@gmail.com

## 1. はじめに

日本語とはどういった語の集合なのであろうか？ 例えば，雑誌や新聞に頻出する語，例えば「大学」や「東京」は常識的に日本語であることに異議を唱える人は少ないであろう．一方，ネットやテレビなどのメディアに頻出する「腐女子」「地味変」といった語については，これを日本語と扱うことについて疑問を呈する人もいるかもしれない．さらに，「DVD」など日本語か英語か曖昧な語もある．このように，日本語を構成する語の集合（以降は，<日本語>と表記する）は不明確である．

このような境界が曖昧な<日本語>であるが，時として明確な区切りをもって自然な<日本語>が必要となる時がある．例えば，辞書や用語集を編纂する際には，収載する語を選ばなくてはならない．実際に，岩波国語辞典[1]の見出し語には，62,000語が収載されており，日本語教育の教材「みんなの日本語（初級）」 [2] には1,060語が収載されている．これらの語は，どのような基準で選ばれたのであろうか？

本研究では，国語辞典[1]に収載されている語を自然な<日本語>とみなし，これを統計的な客観指標で選別することを目指す．これを実現するための素朴な方法は，使用頻度の高い語を<日本語>とみなすことである．しかし，この方式は，直感にそぐわない語を含んでしまう場合がある．例えば，ソーシャルメディア上には「ww」といった表現が高い頻度で出現しており，この基準に照らせば，これらも収載されればならない．これらの語を日本語とみなすことになぜ違和感を覚えるのであろうか？　おそらく，次のような理由が考えられる．

- **使用者の偏り**：一部の使用者（若者）だけが使っており，使用していない人口が多い．
- **使用期間の短さ**：最近使われだしたので，定着したかどうかが分からない．

本研究では，前者の「使用者の偏り」という点に注目する．現在の大量のウェブ・テキストを用いれば，だれがどのような語を使っているのか大規模に調査することができる[3]．そこで，本研究では，約10万人の語の使用統計をもとに，いくつかの指標を導入し，自然な<日本語>の選別を試みる．

実験では，Wikipediaの見出し語を材料に，岩波国語辞典に収載されているかどうかを判定した．実験の結果，高い精度（最大の適合率 0.980）で分類することでき，頻度による手法の精度（適合率 0.890）を大幅に上回った．本研究により，これまで内省によるところの大きかった自然な日本語について，その要因の一部が明らかになったと考える．

## 2. 関連研究

本研究は，日本語を構成する語の集合を統計的に得ることを目的としているが，これを直接議論している先行研究は少ない．ただし，教育のための語彙調査など関連する研究は多い．

日本語使用者がどれくらいの語彙を持っているのか調査する試験は語彙数推定テストとよばれ，これまでに多くの調査が試みられてきた（小学校卒業時[4]，中学校卒業時[5]）．ただし，調査コストが高く少数のサンプルしか得られないという問題がある．

一方，被験者を用いず，雑誌やテレビを材料に語彙統計を得る調査も多い．英語では，コーパスにおける単語出現頻度の高いものから抽出した結果，2,000 語にて，話し言葉の90%をカバーし[6]，6,000語にて書き言葉の80%をカバーしたとの報告がある[7]．

また，ティーン向け小説を読むのに3,000〜5,000語[8]，経済5,000語，アカデミック12,000語[9]が必要だとの研究がある．

日本語においても，10,000 語が現代雑誌 [10]，17,000 語がテレビ[11]に出現する語をほぼカバーした

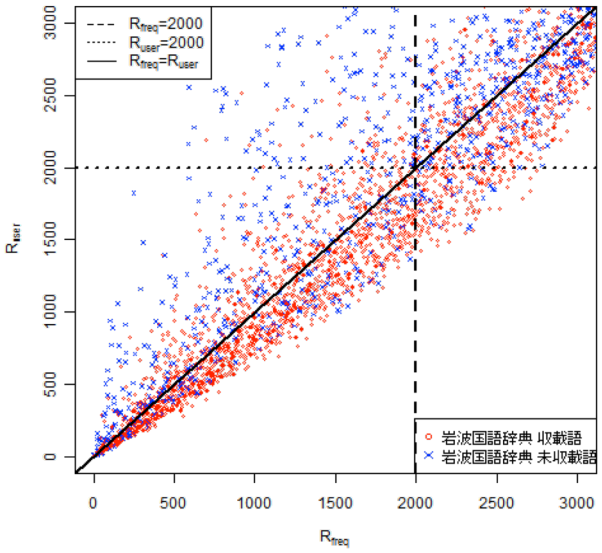


図 1: 語の使用者数と出現頻度の順位.

Y 軸は語の使用者数の順位，X 軸は語の出現頻度の順位を示す．実線は Y=X を示す．右下は頻度の割に使用者が多い，左上は頻度の割に使用者が少ない語となる．

と報告されている．

これらのいずれの値も，ある種の日本語の境界と考えることができるが，これらはすべて頻度に頼っている．本研究は，使用者数という新たな統計を導入する点が新しい．

## 3. 材料

使用者ごとの語の統計を得るために，オンライン・コミュニケーションツールである Twitter 上のテキストから，約 10 万人の継続的な発言を得た．クローラの限界のため，各個人の発言について網羅性はなく，取得できていない発言もあるが，発言の取得にバイアスはないため，問題はないと考える．データの抽出基準と統計は以下である．

- **データ期間**:2009/11/3 から 2010/3/25 の 143 日間（約 5 カ月間）
- **全ツイート数：**約 2.5 億ツイート（253,482,784 ツイート）
- **ユーザ数：**約 10 万人（99,964 人）
- **ユーザ抽出条件**
  - 毎月 5 ツイート以上の投稿
  - 総発言語数が 5000 以上.

形態素解析には juman7.0 [12]を使用した．本研究では，この解析器が出力した形態素の単位を語とみなす．

表 1: 使用者順位／頻度順位の比ごとの語.

(a) 使用者数順位／出現頻度順位 Ruser(w)/Rfreq(w) 小

| w | freq(w) | R freq(w) | user(w) | R user(w) | R user(w) / R freq(w) |
|---|---|---|---|---|---|
| 週間 | 379183 | 554 | 77943 | 282 | 0.5 |
| 復活 | 293265 | 697 | 70103 | 392 | 0.56 |
| 予定 | 917124 | 243 | 88721 | 146 | 0.6 |
| 気分 | 601588 | 351 | 82794 | 211 | 0.6 |
| 昨日 | 1519917 | 160 | 93673 | 97 | 0.6 |
| 原因 | 212165 | 958 | 60124 | 619 | 0.64 |
| 決定 | 320819 | 642 | 68865 | 417 | 0.64 |
| 時間 | 3933947 | 69 | 97927 | 45 | 0.65 |

(b) 使用者数順位／出現頻度順位 Ruser(w)/Rfreq(w) 中

| w | freq(w) | R freq(w) | user(w) | R user(w) | R user(w) / R freq(w) |
|---|---|---|---|---|---|
| 文化 | 204188 | 991 | 48142 | 991 | 1.0 |
| 綺麗 | 319449 | 647 | 58856 | 648 | 1.0 |
| 更新 | 653784 | 328 | 74442 | 330 | 1.0 |
| 宇宙 | 220376 | 923 | 50081 | 929 | 1.0 |
| ログ | 92127 | 2028 | 29652 | 2036 | 1.0 |
| 撮影 | 220362 | 924 | 49896 | 933 | 1.0 |
| ゆず | 59392 | 2771 | 19283 | 2779 | 1.0 |
| 神社 | 98206 | 1915 | 31113 | 1919 | 1.0 |

(c) 使用者数順位／出現頻度順位 Ruser(w)/Rfreq(w) 大

| w | freq(w) | R freq(w) | user(w) | R user(w) | R user(w) / R freq(w) |
|---|---|---|---|---|---|
| 旦那 | 210886 | 966 | 27914 | 2157 | 2.23 |
| てら | 315380 | 656 | 36352 | 1562 | 2.38 |
| 爆発 | 581952 | 359 | 51831 | 867 | 2.41 |
| 原稿 | 328386 | 634 | 34422 | 1680 | 2.64 |
| たん | 792173 | 270 | 55067 | 747 | 2.76 |
| ボク | 256087 | 792 | 24774 | 2396 | 3.02 |
| おつ | 485454 | 431 | 39277 | 1398 | 3.24 |
| ノシ | 352862 | 592 | 22786 | 2559 | 4.32 |

## 4. 手法

**【尺度】**ある語 w を辞書に載せるかどうかにおいて以下の 4 つ指標を用いる．

- ***freq(w)***: 語 *w* の出現頻度
- $R_{freq}(w)$: 語 w の出現頻度の順位
- ***user(w)***: 語 *w* の使用者数
- $R_{user}(w)$: 語 *w* の使用者数の順位

***freq(w)***と $R_{freq}(w)$はこれまでも使われてきた尺度である．一方，***user(w)***と $R_{user}(w)$は本研究で提案する尺度である．

**【素朴な方法（ベースライン）考え方】**ベースラインとして，出現頻度が多いものが辞書に収載されるとみなす．この場合，$R_{freq}(w)$が閾値順位内の語が収載されることになる．

**【提案する考え方】**提案手法は使用者数が多いものが辞書に収載されるとみなす．この場合，$R_{user}(w)$が閾値順位内語が収載されることになる．

**【もう一つの考え方】**もう一つの考え方として，出現頻度と使用者数のバランスを考慮する方法もある．仮に，語の使用に個人の偏りがない場合，出現頻度 *n* 位の語は使用者数も *n* 位であるはずであり，以下の式を満たす:

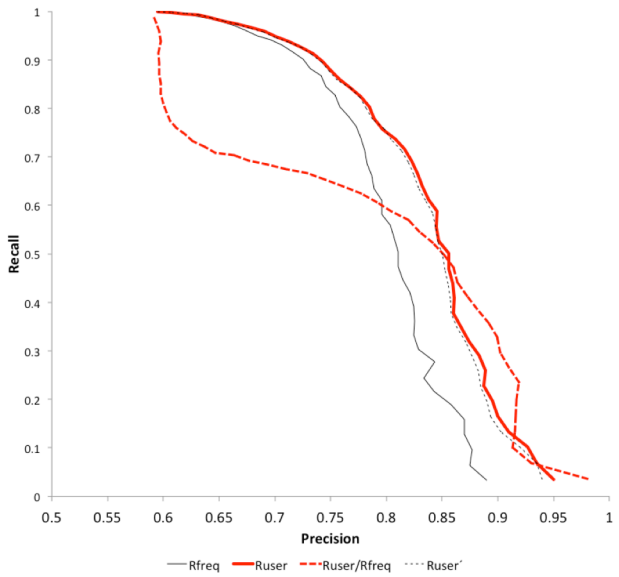


図 2：各手法の精度.
X 軸は適合率，Y 軸は再現率を示す

表 2：各手法の精度（最大の F 値と適合率）

| | 最大の F 値 | | | 最大の適合率 | | |
|---|---|---|---|---|---|---|
| | F | P | R | F | P | R |
| Rfreq | 0.804 | 0.716 | 0.916 | 0.062 | 0.890 | 0.003 |
| Ruser | 0.813 | 0.734 | 0.912 | 0.066 | 0.950 | 0.066 |
| Ruser' | 0.812 | 0.733 | 0.911 | 0.066 | 0.940 | 0.034 |
| Ruser/Rfreq | 0.741 | 0.590 | 0.993 | 0.069 | 0.980 | 0.035 |

表 3：辞書に収載されておらず，かつ Ruser/Rfreq が低い語の例.

| w | freq(w) | R freq(w) | user(w) | R user(w) | R user(w) / R freq(w) |
|---|---|---|---|---|---|
| お気に入り | 133888 | 1479 | 46152 | 1071 | 0.72 |
| パターン | 132207 | 1497 | 45351 | 1101 | 0.73 |
| コンビニ | 286394 | 715 | 63647 | 531 | 0.74 |
| 花粉症 | 86712 | 2116 | 33987 | 1709 | 0.8 |
| 日本語 | 429239 | 488 | 70023 | 394 | 0.8 |
| メイン | 192997 | 1036 | 52759 | 832 | 0.8 |
| ダウンロード | 130200 | 1517 | 40373 | 1329 | 0.87 |
| 平成 | 91887 | 2033 | 32635 | 1808 | 0.88 |
| 焼肉 | 99661 | 1890 | 34311 | 1690 | 0.89 |
| バレンタイン | 183282 | 1089 | 48703 | 971 | 0.89 |
| 飲み会 | 170438 | 1175 | 46734 | 1058 | 0.9 |
| 店長 | 78652 | 2314 | 28397 | 2126 | 0.91 |
| 最終回 | 98431 | 1911 | 33378 | 1753 | 0.91 |
| ガキ | 78574 | 2316 | 28465 | 2121 | 0.91 |
| 忘年会 | 171181 | 1167 | 46468 | 1065 | 0.91 |

$$R_{user}(w) = R_{freq}(w)$$

ここで，もし，若者の流行語のように，限られた使用者のみが複数回使っている語があると，その語は出現頻度に相対して使用者数が少ないため，以下の式を満たすはずである:

$$R_{user}(w) > R_{freq}(w)$$

逆に，使用者に偏りなく使われる語は以下の式を満たす:

$$R_{user}(w) < R_{freq}(w)$$

このような出現頻度と使用者のバランスを考慮すると，重み定数 $\alpha$ を用いて，以下の式によって辞書収載を判定することも考えられる:

$$R_{user}(w) < \alpha \cdot R_{freq}(w)$$

## 5. 実験

**【比較手法】**以下の 4 つの方法を比べる：

- **頻度ベース: Rfreq**
  $R_{freq}(w) < \alpha_1$
- **使用者数ベース: Ruser**
  $R_{user}(w) < \alpha_2$
- **頻度重み付け使用者数ベース: Ruser'**
  $-\log(freq(w)) \cdot user(w) < \alpha_3$
- **使用者数と頻度の比ベース: Ruser/Rfreq**
  $R_{user}(w) / R_{freq}(w) < \alpha_4$

いずれもパラメータ $\alpha$ を動かし（$\alpha_{1..4}$=0〜無限大），分類精度を評価した.

**【材料】**対象となる語は Wikipedia の日本語エントリーとなっている語から Twitter 上で一定頻度出現していることを条件に無作為抽出した 4,000 語を対象とした．これらのうち 2,598 語は岩波国語辞典に収載されており，残り 1,402 語は収載されていない．語の集計に関しては，形態素境界を考慮し，語ごとに独立に集計した．すなわち，「東京｜大学」が出現すると「東京｜大学」「東京」「大学」という 3 つのエントリーの頻度となる．ただし，形態素境界が一致しない「京大」の頻度には数えない.

**【評価】**分類精度は以下の尺度で測る.

- **適合率 (P):** 適合した語の数／システムが出力した語の数
- **再現率 (R):** 適合した語の数／収載語の数
- **F 値 (F)：**上記の調和平均

**【結果】**語のプロットを図 1 に示す．Y 軸は語の使用者数の順位，X 軸は語の出現頻度の順位を示す．実線は Y=X を示す.

図の多くの点は Y=X 付近に散らばっている．このことから，使用者数と出現頻度はおおむね対応していることがわかる.

その一方，時としてグラフ上方向へのはずれ値（Y＞＞X）となる語がある．これは出現頻度に相対して使用者が極端に少ない語があることを示している．そのような語の例を表 1(C)に示す．「てら」「ノシ」といった若者言葉や「原稿」といった職業／専門用語に相当する語が含まれている.

逆にグラフ下方向へのはずれ値 (Y<<X) は少ない．これは使用者が均等な語が少ないことを示している.

ここで，Y<X の領域には辞書収載語の多くが含まれている．また，Y>X の領域には非収載語が多い.

このことから，使用者に偏りのない語を辞書が選好していることがわかる．

頻度ベース，使用者数ベース，提案手法による辞書収載語の分離精度を図 2 に示す．**Ruser** は一部について **Ruser／Rfreq** に劣るものの全体的に精度が高く，常に **Rfreq** よりも高精度である．このことから，辞書収載の基準を説明するのに使用者数の重要性を示唆している．

各手法の最大の F 値と適合率を表 2 に示す．F 値の観点からは，**Ruser／Rfreq** が低く，それ以外の各手法に大きな精度差はない．一方，適合率の観点からは **Rfreq** が低く，**Ruser／Rfreq** が高い精度を示す．どのような場合においても，**Ruser** は安定して高い精度を保っており，使用者数の重要性をここでも確認できる．また，適合率が最大となる（最大の適合率 0.980）のは **Ruser／Rfreq** であり，高い確信度が必要な場合には本尺度が有効である．

## 6. 考察

提案手法の考え方を推し進めると，現在の辞書が収載していない語でも，今後収載される可能性のあるものを抽出できる可能性がある．辞書に収載されていないが，使用者数が多い語のリストを表 3 に示す．提案手法は，このような語を自動的に大規模に収集でき，従来内省によるところが大きかった辞書編纂に客観的な指標を提供できる可能性がある．

**【調査の限界】**本研究には以下の限界がある：

- **使用者バイアス:** ソーシャルメディアに参加しているユーザは日本語話者の一部であり，偏った集団から語彙を採取している可能性がある．実際に，本研究で扱った Twitter は，30%近くのユーザが東京に集中し，かつ，20 代のユーザが多いとされている[12].
- **環境のバイアス:** ソーシャルメディアという環境が，語彙の使用のバイアスとなる．例えば，キーボード／スマートフォンでの入力には，IME の語の選好が影響している可能性がある．

上記をはじめとし，他にも期間，様々なバイアスがあり，今後，どのようにバイアスを低減するかが課題である．

**【応用可能性】**本研究結果は辞書の収載以外に様々な応用が可能である．例えば，特定のコミュニティで使われている語彙（ビジネス英会話，医療通訳など）から学習することで効率的な語彙習得が可能となる．また，ある使用者と近い語彙を持つ使用者の推薦なども可能となる．

## 7. おわりに

本研究では使用者数という新しい統計を用いて，直観に沿う日本語の語を選別する手法を提案した．実験の結果，使用者数順位（**Ruser**）が安定して高精度を，一部については，使用者順位／頻度順位比（**Ruser／Rfreq**）が高い精度を示した．これらのいずれの場合も使用者数を用いることから，使用者数は辞書収載に関して重要な要因であることが分かった．また，分類に失敗した語も単なる誤りとみなせず，一部については今後，辞書収載が検討されてもよいものが含まれていると考えられる．

本研究は，従来，内省によるところが大きかった辞書編纂に使用者数という統計を導入し，その有効性を示した．このように従来，単なる頻度を用いてきたが，実際は使用者数がより適切である指標や課題は他にもあると思われる．今後はより幅広く可能性を検討したい．

**データの公開:** 本研究で使用した語はウェブ・サイトにて公開している[1].

**謝辞:** 本研究は，JST 戦略的創造研究推進事業（さきがけタイプ）「情報環境と人」及び，科研費補助金(若手研究 A)による．

## 参考文献

1. 西尾実, 岩淵悦太郎, and 水谷静夫, *岩波国語辞典第五版* 1994: 岩波書店.
2. スリーエーネットワーク, *みんなの日本語 初級* 1998: スリーエーネットワーク.
3. 荒牧英治, et al., *日本人のオンライン・コミュニケーション上での平均使用語彙数は 8,000 語である*, in *情報処理学会 第 208 回 自然言語処理研究会 (SIG-NL)*2012.
4. 阪本一郎, *教育基本語彙* 1958: 牧書房.
5. 森岡健二, *義務教育終了者に対する語彙調査の試み*. 国立国語研究所年報, 1951. **2**: p. 95-107.
6. West, M., *A General Service List of English Words*1953: Longman.
7. Francis, W.N., *Frequency Analysis of English Usage*1982: Houghton Mifflin Company.
8. Hirsh, D. and P. Nation. What vocabulary size is needed to read un-simplified texts for pleasure? Reading in a Foreign Language 8(2): 689-696 1992.
9. Sutarsyah, C., I.S.P. Nation and G. Kennedy , *How useful is EAP vocabulary for ESP? A corpus-based case study. RELC Journal* 25, 2: 34-50, 1994.
10. 国立国語研究所, *現代雑誌 90 種の用語用字* 1984: 秀英出版.
11. 国立国語研究所, *高頻度語彙から見たテレビ放送語彙の特徴* 1999: 大日本図書.
12. Kurohashi, S., et al. *Improvements of Japanese Morphological Analyzer JUMAN*. in *The International Workshop on Sharable Natural Language Resources*. 1994.
13. アスキー総合研究所, *Twitter 利用実態調査*. 2010.

[1] http://mednlp.jp/resources/