Vol.2012-NL-208 No.9
2012/9/3

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

荒牧英治[†,††] 増川佐知子[†] 森田瑞樹[†] 保田祥[†]

これまで言語学で高い関心を集めている問題の 1 つに人間の語彙数がある．数々の調査がなされてきたが，その多くは，理解できる語彙（理解語彙）の調査にとどまり，実際に使用する語彙（使用語彙）についてはどのくらいのものか，いっこうにわからないとされてきた．本研究では，ウェブ上の発言データを利用し，10 万人という大規模な人数で使用語彙調査を行った．調査の結果，使用語彙は平均 8,000 語であることが明らかになった．さらに，同データを用いて，語のユーザ数の調査を行った．この結果，ユーザに偏りがある語や偏りがない語のリストが得られた．このようなユーザ数にもとづいたリストは本研究で初めて得られたものである．

Average Japanese Vocabulary for Online Communication is 8,000 words

Eiji ARAMAKI[†,††], Sachiko MASKAWA[†], Mizuki MORITA[†] and Sachi YASUDA[†]

The active vocabulary size – the number of words that we can use– is one of the biggest issues in linguistics. Although various studies had been challenged this issue, the precise size of our active vocabulary is still unknown. To solve this issue, this study utilized an online communication text produce by 100,000 people. The result revealed that the average vocabulary consists of 8,000 words. Furthermore, this study also presents a method to estimate the number of users for each word. By using this method, this paper presents a user size based word list.

1. はじめに

語彙数は，言語学分野にてもっとも熱心な関心が払われていた問題の 1 つであり，これに答えるため，数々の語彙調査がなされてきた．例えば，辞書の見出しを用いた調査[1]やマスメディア（雑誌やテレビ）に出現する語の調査[2, 3]が行われてきた．しかし，これらの調査の多くは，読者が理解できる語彙（理解語彙）の調査であり，使用された語彙（使用語彙）については行われていない．使用語彙の調査が困難であるのは様々な理由があるが，第一に膨大な調査コストがかかることが大きい．例えば，鶴岡調査[4]では，被験者に対して 24 時間録音または速記を行ったが，このような実験を遂行するのに必要なコストは膨大なものになり，僅か 3 人の被験者のみの調査にとどまる．また，被験者不足の問題とは別に，サンプリングするという行為そのものが，平常時の発話とは異なる環境を実験参加者に強いる恐れがある．このため，「日本人の平均使用語彙量がどのくらいのものか，いっこうにわからない」[5]と言われてきた．

そこで，本研究ではウェブ上のオンライン・コミュニケーションのデータに注目する．オンライン・コミュニケーションのデータを用いれば，個人に紐付いたテキスト・データを大量に入手可能である．また，調査を後ろ向きに行う（過去のデータを用いる）ことにより，調査バイアスのないデータを得ることができる．このデータを用いて調査が可能な課題は数多くあるが，本研究では，このデータを用いて語彙数調査を試みる．

本研究ではウェブ上の発言を利用して，10 万人という大規模な人数で，対象者が実際に使用した語彙（使用語彙）を調査する．この結果，平均 8,000 語の語彙がオンライン・コミュニケーションで用いられてきたことが分かった．

また，誰がどのような語を使っているかが分かるということは，逆にいうと，ある語が誰によって使われているか分かるということでもある．これを調査すると，全数としての使用頻度が高くても使用人口が少ない言葉や，逆に使用頻度がそれほど高くなくても高い使用人口を持つ言葉を抽出できる．本調査では，使用人口の多少による語のリスト化を行った．このような語の使用率の集計は世界で初めて得られたものである．

† 東京大学知の構造化センター
Center for Knowledge Structuring, University of Tokyo.
†† 科学技術振興機構さきがけ
JST PRESTO
††† 独立行政法人医薬基盤研究所
The National Institute of Biomedical Innovation

Vol.2012-NL-208 No.9
2012/9/3

2. 関連研究

本研究は，ある特定の集団を追いかけるという意味で，言語コホート研究と近い調査方法である．そこで，これまでの言語コホート研究を概観する（2.1 節）．さらに，過去に実施された語彙数についての調査を俯瞰する（2.2 節）．

2.1 コホート研究

コホート調査は通常は疫学調査の研究デザインの一つであり，次の手続きを踏む．まず，多数の健康人の集団を対象として，疾病の原因となる可能性のある要因（喫煙・食生活・血液データなど）を調査する．次に，この集団を追跡調査して，疾病にかかる者を確認する．その上で，最初に調査した要因と，その後の疾病の発生との関連を分析する．例えば，喫煙者と非喫煙者で，その後の肺がんの発生率を比較する．喫煙と肺がんの関連など，今日では常識的な知見も，この研究方法で明らかにされた成果である[6].

これを言語調査に応用したのが，言語コホート調査であり，これまで，以下の 3 つのコホート研究が実施されている．

- **鶴岡調査** [4]（1953, 1974, 1994, 2007）
- **岡崎敬語調査** [7]（1953, 1972, 2008）
- **Seattle Longitudinal Study (SLS)** [8]：（1956～）

これらは，**前向きコホート**と言われるタイプの研究で，観察対象となる集団を決定して，定期的にこれを追うスタイルの研究である．一方，記録がとられている場合は後ろ向きにこれを行うことができ，**後ろ向きコホート**と呼ばれる．本調査も記録を遡って行うため後ろ向きコホート研究の一種ともいえる．本研究の特徴を表 1 にまとめる．本研究は他のコホート研究と比較して以下のような利点を持つ．

- **【大規模】**従来にない大規模な調査が可能となる．鶴岡調査では 24 時間調査として，発話すべてを速記と録音によって残した．岡崎調査では知識・意見・内省について，調査員が個別面談を実施した．SLS の言語関連調査は被験者に筆記テストを課した．これらの方法では被験者は少人数に限定されてしまうが，本研究では 10 万人といった大規模な集団を扱う．
- **【低バイアス／低コスト】**従来型の前向き調査では調査する行為自体が使用言語を変化させてしまう恐れがある．本研究は後ろ向き調査であるため，調査（観察）バイアスがない．また，調査コストも小さい．

	調査方向	網羅性	調査間隔	調査期間
SLS	前向き	6,000 人 (現在まで 26 人継続)	7 年間隔	49 年～
岡崎敬語調査	前向き	300～400 人 (現在まで 20 人継続)	20 年･36 年間隔	55 年～
鶴岡調査	前向き	500 人 (現在まで 53 人継続) うち 24 時間調査 3 人	20 年間隔	50 年～
本研究	後ろ向き	100,000 人	1 日間隔	5 ヶ月

表 1: これまでの言語コホート研究と本研究との比較．

- **【細密度】**細かい時間間隔で調査が可能となる．鶴岡調査では 20 年おきにサンプリングを行なっている．SLS は 7 年おきである（言語関連のテストがすべての年度で実施されているのではない）．本調査は期間こそ半年と短いものの 1 日単位でのサンプリングが可能である．

前述の利点があるものの，コホート調査として以下の欠点も存在する．

- **【音声会話でない】**オンライン・コミュニケーション上での発言である．これは口語体の書き言葉とみなせるものの，実際の発話とは異なる．
- **【短期間】**半年という言語変化の観察においては短い期間である．また，今後，継続的に調査しようにも，オンライン・サービスの寿命に依存する．岡崎調査や鶴岡調査のような半世紀を越える長さに渡って単一のサービスが存在するとは考えられないため，長期間の調査は絶望的である．

これらの欠点を考慮すると，ウェブ・データは，長期間での言語変化といった従来のコホート調査の課題（通時的研究）には不向きであると言える．しかし，誰がどのような語彙を利用したかといった調査（共時的研究）は大規模性が重要であり，本データと親和性が高い．

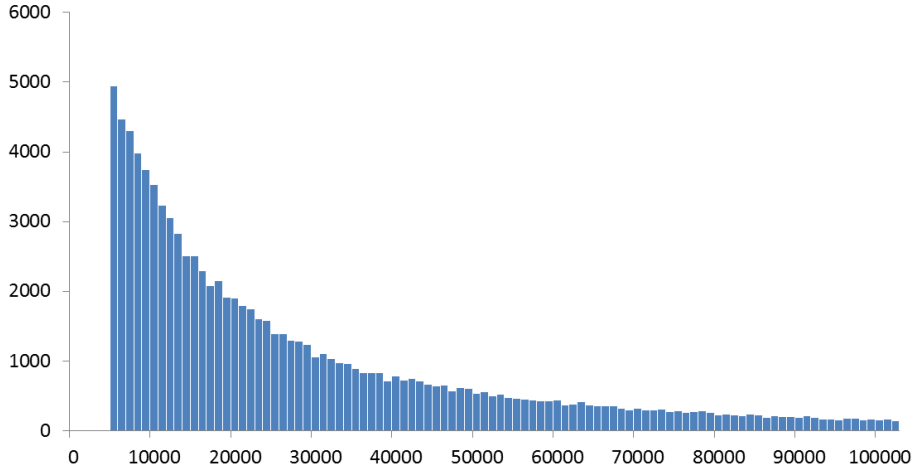

図 2: 発言語数のヒストグラム. X 軸は発言語数，Y 軸は発言者数とする．

		対象	語彙の種類	語彙数
West[15]	英語	コーパスでの頻度	理解語彙	2,000 語
Bates[20]	英語	16-30 ヶ月児	一部使用語彙	50-600 語
Hirsh [14]	英語	中高生向け小説	理解語彙*	3,000-5,000 語
Goulden [12]	英語	大学終了者	理解語彙	20,000 語
D'Anna[13]	英語	大学生	理解語彙	17,000 語
天野[19]	日本語	2-5 歳児	一部使用語彙	300-2000 語
玉村[2]	日本語	雑誌	理解語彙*	12,000 語
国研[3]	日本語	テレビ	理解語彙	17,000 語
阪本[9]	日本語	中学生	理解語彙	20,000 語
甲斐[18]	日本語	小学生用辞書	理解語彙	25,000 語
森岡[1]	日本語	義務教育終了者	理解語彙	30,000 語
林[5]	日本語	過去の調査	理解語彙	40,000 語
本研究	日本語	Twitter ユーザ	使用語彙	8,000 語

表 2: これまでの語彙調査と本研究の比較．

* 玉村[2]で議論される語彙数は正確には日本語教育のために必要な語彙数であるが，簡単のため理解語彙と分類した．

2.2 語彙調査に関する研究

- **【語彙数推定テストによる調査】**語彙数を調査する試験は**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．もっとも基本的な語彙数推定テストは，辞書を用いた調査と推計である．阪本[9]は，保有語彙量（理解語彙・vocabulary size）は，小学校卒業時で約 20,000 語，中学校卒業時で 36,000 語と推定した．森岡[1]は全辞書見出し（約 37,000 語）を用いて義務教育終了者 15 名の理解語彙を調査し，平均 30,000 語（max 36,000 語; min 23,000 語）とした．林[5]は，上記の結果を受けて，「*日本人の成人の理解語彙量は大体 40,000 語程度であろう*」と述べている．ただし，「*使用語彙は調査法がむずかしく，日本人の平均使用語彙量がどのくらいのものか，いっこうにわからないが，理解語彙量よりずっと少ないことは確かだろう*」とも述べている．
- **【第二言語習得に関する調査】**語彙数の把握は第二言語習得においても重要な調査であり，英語を中心に多くの報告がある[10, 11]．英語圏の調査では，大学卒業程度の理解語彙は 17,000 語から 20,000 語と言われている[12][13]. ただし，これらすべてが学習者に必要なわけでなく，Hirsh 等[14]は中高生向けの小説をカバーする語彙量から英語を第 2 言語とするために必要な語彙数は 3,000 語～5,000 語と報告している．
- **【出版物／メディアを用いた類推】**被験者を用いず出版物やメディア（雑誌やテレビ）を調査することで，間接的に語彙数を推定する調査も行われている．英語では，コーパスにおける単語の出現頻度の高いものから 2,000 語を抽出し，これを General Service List として英語教育に用いている．この GSL は話し言葉の 90%以上をカバーしている[15]．Brown コーパスを用いた調査から，書き言葉は上位 6,000 語で約 90%をカバー可能とされている[16]．日本語においても，現代雑誌の調査では見出し語の上位 10,000 語までが．全体の 91.7%をカバーしており[17]，日本語を運用するためには 12,000 語が必要と推定されている[2]．同様に，テレビについても，上位 17,000 語を知っていれば現れた語をほぼ 100%カバーするという結果が出ている[3]．また，小学生用の各国語辞典は平均して約 25,000 語の見出しを持っている[18].
- **【言語獲得に関する調査】**理解語彙ではなく，使用語彙を調査したものに幼児期における言語獲得の研究がある．一般的には，2 歳で 300 語，3 歳で 1000 語，4 歳で 1500 語, 5 歳で 2000 語を獲得しているとされている[19, 20]. このように，使用語彙が少ない幼児期に限っては，使用語彙の調査は可能であるが，その場合においても実データの収録だけでなく，質問紙による調査（MacArthur-Bates Communicative Development Inventories）[21]が併用されている．

以上のように，語彙数に関する研究はさまざまな分野で実施されている．これらをま

とめると，成人の理解語彙は 17,000 語～40,000 語であるものの，実際に必要な語は 12,000 語（雑誌）～17,000 語（テレビ）と予想される（表２）．ただし，いずれも理解語彙調査であり，使用語彙調査は幼児期を除いて難しい．本研究は，成人に対して使用語彙の調査を行う点が新しい．

3. 材料

使用語彙調査を行うためには，誰がどんな発言したか，発言者とその使用語彙の紐付けが必要である．しかし，これが保証されたウェブ上のリソースは少ない．例えば，ブログは複数の執筆者により記述されることがある．また，大規模な引用が起こりうる．そこで，本研究では代表的なオンライン・コミュニケーションツールである Twitter に注目した．Twitter はユーザ数も 1,400 万人（2011 年 1 月）と多く，また，文字数制限のため大規模な引用がない．

本研究では以下の基準で約 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，発言の取得に語彙のバイアスはなく，また，発言数から使用語彙を推定するため，語彙調査にあたっての問題はない．統計を以下に示す．

- **データ期間**：2009/11/3 から 2010/3/25 の 143 日間（約 5 カ月間）
- **ユーザ数**：約 10 万人（99,964 人）
- **ユーザ抽出条件**
  - 毎月 5 ツイート以上投稿している．(継続的な発言)
  - 総発言語数が 5000 以上.
  - 最初の 100 ツイート中に「の」が含まれている．(日本語使用者に限定)これは非日本語使用者を除くため行った．
- **全ツイート数**：約 2.5 億ツイート（253,482,784 ツイート）
- **全形態素数**：約 43 億語（4,258,707,255 語）.

図 2 にユーザの期間中の全発言語数分布を示す．なお，形態素解析には juman7.0[22]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．

4. 方法

ある人がどれくらいの語彙をもっているかは，十分な長期間観察を行い，どのような語を使ったかを調べればよい．しかし，数万と言われる語彙すべてが使われるためには，気の遠くなるような長期間の観察が必要となってしまう．さらに，どれだけ観察しても，対象者のすべての語彙を観察したという保証を得ることはできず，調査を終了するタイミングが分からない．

そこで，本研究では，一定期間にユーザが発言した語数から，潜在的な語彙数 N を推測する．この推測はユーザがジップ則[23, 24]に従って語を発生していると仮定することで可能になる．例えば，あるユーザが 10,000 語の語彙を持っていると仮定する．個人の発言がジップ則に従っているならば，1 位の語は全発言の 10.2%を占めるはずであり，2 位の語は 5.1%を占めるはずである．この場合，この対象者の発言を延べ 1,000 語集めた段階（以降，延べ語数を**トークン**と呼ぶ）で，期待される語の異なり（以降，**タイプ**と呼ぶ）はおよそ 509 語である．逆に言えば，1,000 トークン集めて 509 タイプを得たならば，その人の語彙数は 10,000 語であると推測できる．

実際の様子を図 3 に示す．X トークンごとに Y タイプが観測されるはずという期待値を潜在語彙数（N=5000…30000）ごとにプロットしている．以降，この曲線を**タイプ・トークン曲線**と呼ぶ．タイプ・トークン曲線は語彙数 N が大きくなれば，傾きが急になる．これは，巨大な語彙を持っているならば，いくら観測しても，次から次へと新しいタイプが観測できるからである．逆に，語彙が少ないならば観測早々にして（すなわち，小さいトークンで）あらゆるタイプが出尽くし，曲線は飽和してしまう．

ここで，コーパスから抽出したユーザのタイプ・トークン曲線の例を図中の点線で示す．@xxx については，実験の 5 ヶ月間で 18,000 トークンが観測され（X 軸の値），4,000 タイプが得られている（Y 軸の値）． このユーザのタイプ・トークン曲線ともっとも近い語彙数の曲線は N=10,000 であり，ここからこのユーザが潜在的に語彙数 10,000 と推測することができる．

実際には，N=1000 から 50000 まで 1000 刻みの語彙数の曲線 50 本を用意した．10 万人すべてのユーザに対して，各ユーザのタイプ・トークン曲線が 50 本の中のどの線に近いかを算出し，語彙数の推定を行った．

5. 結果と考察

推定された使用語彙数のヒストグラムを図 4 に示す．語彙数の最頻値は 7,000 語，平均は 8,000 語であった．この 8,000 語という値は先行研究で推定されてきた理解語彙 40,000 語と比較すると大きな格差がある．これは，普段，理解できる語彙のおよそ 1/5 のみしか使用していないことになり，なぜ，このようなギャップが生じるか，今後のさらなる研究が必要である．

また，使用語彙の分布は冪分布となっており，非常に大きな語彙をもつユーザもわずかながら存在することを示している．例えば，通常の語彙の 6 倍近い 5 万といった語彙を使用しているユーザも 0.08%存在する．このような語彙数の分布は本研究によって初めて明らかになった．

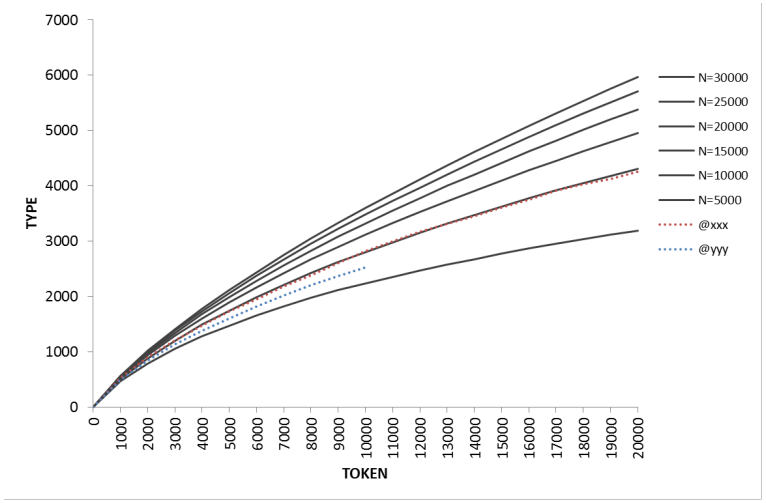

図 3: 語彙数 N（N=5000..30000）のタイプ・トークン曲線．
X 軸はトークン数を示す．Y 軸はタイプ数を示す．実線は理論曲線で，語彙数（N）で区別されている．どの曲線も，トークンが多くなるにつれ，タイプの増加が鈍くなる．同じトークンに対しては，より大きな語彙数をもっていた場合ほど，タイプ数が多いと推定される．逆に，少ない語彙数では，比較的少量のトークンでタイプが飽和してしまう． 破線は実際のユーザの例を示している．@xxx は N=10000,@yyy は N=7000 と推定された．

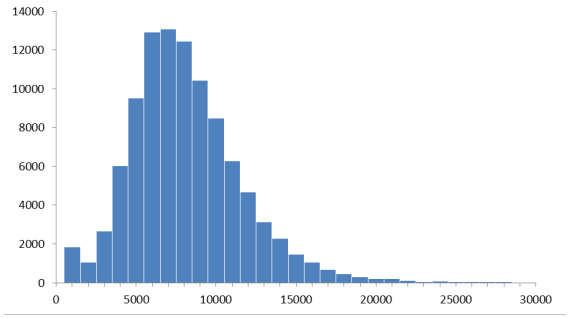

図 4: ユーザ数と推定使用語彙数．X 軸は語彙数，Y 軸はその語彙数を持つユーザ数を示す．

5.1 語のユーザ数調査

語彙数調査は，あるユーザがどれくらいの語彙を使っているかの統計であった．これと対照をなす調査として，ある語についてどれくらいのユーザが使っているのかを調べることができる．同じ出現頻度であっても，多くのユーザが用いる語はより一般的な語と考えられるため，この調査により語の一般性を定量化できる．

語のユーザ数のプロットを図 6 に示す．Y 軸は語のユーザ数を示す．ユーザ数は，調査対象の全員（約 10 万人）が使った場合を 1.0（100%）として表示している．X 軸は語を使用頻度順に並べた際の順位である．

1 位から 10 位までの語はほぼ全ユーザが使用しており，Y=1 付近にプロットされている．しかし，出現頻度が 100 位くらいから，ユーザ数が 0.5 を切っている語が存在しはじめ，2000〜3000 位の語ではユーザ数 0.5 以下の語が大半となる．

図の実線は，全ユーザが各語をジップ則に従った確率で使用すると仮定した場合のユーザ数の期待値である．図示されるように，実測値は期待値よりも低いユーザ数を持っていることがわかる．これは，ユーザが均一でなく，ユーザごとに使用する語彙が異なるからだと考えられる．これを詳細にみるため，各語について，ジップ則から推定される語のユーザ数と実測値の比を以下のように計算した：

$$逸脱率 = \frac{ジップ則から推定される語のユーザ数の期待値}{ユーザ数の実測値}$$

語の逸脱率のリストを付録に掲載する．逸脱率が低い語はユーザに偏りがない語である．例えば，もっとも逸脱率の低い語は「あける」であるが，これは年始の挨拶「あけましておめでとうございます」に起因しており，直感的にも多くの人間が一様に使うことが理解できる．逆に，逸脱率が高い語は，「オレ」:「ぼく」などの選択可能な人称や，「ww」や「〜だん」といった若者言葉／スラング等であり，ユーザに偏りがある語といえる．このように，使用頻度とは異なるユーザ数という概念があり，これらの 2 軸の比から算出される逸脱率によって，語彙は異なった様相をみせる．

5.2 仮定の検証

この語彙数調査は語の出現頻度がジップ則に沿うと仮定していた．この仮定の検証を図 5 に示す．図示されるように，語の発言頻度は上位数語を除いてジップ則にそっており，決定係数 R^2 も高いため，妥当な仮定であると言える．

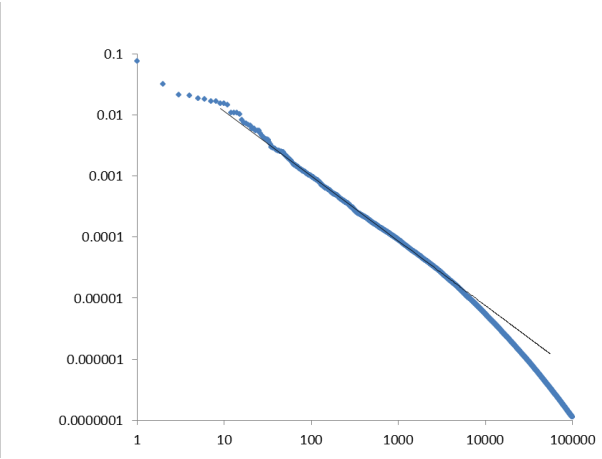

図 5：語の出現順位と出現数．X 軸は語の順位，Y 軸は語の相対頻度を示す．両軸とも対数で表示してあるので，ジップ則は直線として確認される．

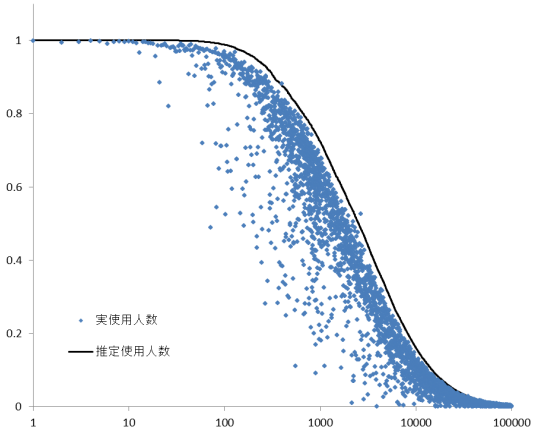

図 6: 語のユーザ数のプロット．X 軸は語の使用頻度順位，Y 軸は語のユーザ数（人口比）を示す．ユーザ数は，調査対象の全員（10 万人）が使った場合を 1.0（100%）として表示した．点は実際のユーザ数，実線は全ユーザが各語をジップ則に従った確率で使用すると仮定した場合のユーザ数の期待値である．

5.3 調査の限界

本研究は語彙数や語のユーザ数といった集計が困難な統計を得ることができるものの，以下の限界がある：

- **【形態素単位での集計のバイアス】**本研究の調査はすべて形態素の単位での集計であり，複合名詞は，ばらして集計されている．例えば，「サンシャイン牧場」は「サンシャシン」と「牧場」に分けて集計される．このため，先行研究の語彙数とずれる可能性がある．

- **【ユーザのバイアス】**オンライン・コミュニケーション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，本研究で扱ったサービス「Twitter」では，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[25]．このバイアスが語彙を実際よりも小さくしている可能性がある．
- **【環境のバイアス】**オンライン・コミュニケーションという環境自体が，語彙を変化させている可能性がある．例えば，キーボード／スマートフォンでの入力が語彙に影響している可能性がある．

上記をはじめ，材料／集計方法により様々なバイアスがあるが，先行研究の方法においてもバイアスは存在した．例えば，辞書の見出し語の調査では，選定した辞書の影響を大きく受け，少人数の被験者を用いる場合も，日本人の均一なサンプルである保証はない．したがって，本調査が特に信頼できない調査という根拠にはならない．むしろ，膨大な人数の調査により，統計的には正確である可能性がある．今後，本調査を様々な材料に広げ，これまでの調査と組み合わせることで，より緻密な語彙数推定が望まれる．

5.4 応用可能性

本研究結果は様々に応用することができる．例えば，日本語会話辞書や旅行会話辞書など実用的な言語リソースに含まれる語彙は，実際に使用しない語彙を多く含んでいる事になり，大幅なコンパクト化が可能である．また，ターゲットとなるユーザ数の多い語彙から学習することで効率的な語彙習得なども可能となる．

Vol.2012-NL-208 No.9
2012/9/3

6. おわりに

これまで，日本人の平均使用語彙量については，どのくらいのものか，いっこうにわからないとされてきた．本研究では，オンライン・コミュニケーションのデータを用い，10 万人という大規模な人数で，対象者が実際に使用した語彙（使用語彙）を調査した．結果，平均語彙数 8,000 語と推定され，我々の理解語彙の多くは実際には使用されていないことが判明した．また，語のユーザ数を調査し，使用率によって，一般的な語や逆に非一般的な語をリスト化した．このリストは，本研究によって初めて可能となったものである．

倫理的配慮

多くの人を対象として追跡する研究では，個人の特定が可能な情報を使用せざるをえないことなど，倫理的な配慮が必要である．本研究では，ウェブ上で公開されているデータ（Twitter のタイムラインデータ）を用いた．また，形態素単位で集計することにより，個人が特定できない統計情報としてデータを公開した．

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) (挑戦的萌芽)による．本論文を書くにあたって有益な議論をいただいた東京大学医学部附属病院篠原恵美子氏に謹んで感謝の意を表する．

参考文献

1. 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.
2. 玉村文郎, *NAFL Institute 日本語教師養成通信講座 8 日本語の語彙・意味* 2002: アルク.
3. 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.
4. 国立国語研究所. *鶴岡調査*. 2012.
5. 林四郎, *語彙調査と基本語彙*. 国立国語研究所報告, 1971. **39**.
6. 平成 17 年度祖父江班報告書, 2006.
7. 国立国語研究所. *岡崎敬語調査*. 2008.
8. Willis, K.W.S.a.S.L. *Seattle Longitudinal Study*. 2011 [cited 2012/05/27.
9. 阪本一郎, *教育基本語彙* 1958: 牧書房.
10. Laufer, B., *The development of L2 lexis in the expression of the advanced language learner*. Modern Language Journal, 1991. **75**(4): p. 440-448.
11. Laufer, B. and P. Nation, *Vocabulary size and use:Lexical Richness in L2 Written Production*. Linguistics, 1995. **16**(3): p. 307-322.
12. Goulden, R., P. Nation, and J. Read, *How Large Can a Receptive Vocabulary Be?* Applied Linguistics, 1990. **11**(4).
13. D'Anna, C.A., E.B. Zechmeister, and J.W. Hall, *Toward a meaningful definition of vocabulary size*. Journal of Reading Behavior, 1991. **23**: p. 109-122.
14. Hirsh, D. and P. Nation, *Nation. What vocabulary size is needed to read unsimplified texts for pleasure?* Reading in a Foreign Language, 1992. **8**(2): p. 689-696.
15. West, M., *A General Service List of English Words*1953: Longman.
16. Francis, W.N., H. Kučera, and A.W. Mackie, *Frequency analysis of English usage: lexicon and grammar*1982: Houghton Mifflin.
17. 国立国語研究所, *現代雑誌 90 種の用語用字* 1984: 秀英出版.
18. 甲斐睦朗, *語彙指導の方法（語彙表編）*1986: 光村図書.
19. 天野清, *言語心理学*. 現代双書（第三巻）1976: 新読書社.
20. Bates, E., et al., *Developmental and stylistic variation in the composition of early vocabulary*. J Child Lang, 1994. **21**(1): p. 85-123.
21. Fenson, L., et al., *MacArthur communicative development inventories; User's guide and technical manual*1993: Singular Publishin.
22. Kurohashi, S., et al. *Improvements of Japanese Morphological Analyzer JUMAN*. in *The International Workshop on Sharable Natural Language Resources*. 1994.
23. Zipf, G.K., *The Psychobiology of Language*1935: Houghton-Mifflin.
24. Zipf, G.K., *Human Behavior and the Principle of Least Effort*1949: Addison-Wesley.
25. リンクシェア・ジャパン株式会社. *Twitter 利用実態調査*. 2010.

付録　上位 1 万語の使用率別リスト

低逸脱率リスト

基本形	1/逸脱率	基本形	1/逸脱率	基本形	1/逸脱率
あける	1.09	を	1.00	から	0.99
象徴	1.03	と	1.00	中	0.99
今年	1.01	だ	1.00	ぬ	0.99
ぶり	1.00	で	0.99	れる	0.99
の	1.00	も	0.99	なる	0.99
は	1.00	か	0.99	ある	0.99
明ける	1.00	ない	0.99	のだ	0.99
に	1.00	の	0.99	る	0.99
する	1.00	ます	0.99	よ	0.99
が	1.00	お	0.99	の	0.98

中逸脱率リスト

基本形	1/逸脱率	基本形	1/逸脱率	基本形	1/逸脱率
うむ	0.50	ハック	0.50	日テレ	0.50
てら	0.50	桂	0.50	合同	0.50
隊長	0.50	宮崎	0.50	批評	0.50
封	0.50	手当	0.50	だん	0.50
駒	0.50	ピカチュウ	0.50	イラレ	0.50
べろ	0.50	橋本	0.50	神保	0.50
離脱	0.50	ワンピ	0.50	メシ	0.50
双	0.50	プレー	0.50	子ども	0.50
ウチ	0.50	怪談	0.50	ぜん	0.50
株	0.50	照れる	0.50	乱	0.50

高逸脱率リスト

基本形	1/逸脱率	基本形	1/逸脱率	基本形	1/逸脱率
理沙	0.19	春香	0.16	けいたい	0.10
例大祭	0.19	金貨	0.16	にゃも	0.10
車載	0.19	是	0.16	パーソン	0.09
快	0.18	イナイレ	0.15	入室	0.07
なえる	0.18	しゅっしゃ	0.15	曇	0.07
候	0.18	ルルーシュ	0.15	ルーン	0.07
了	0.18	奈落	0.15	デリヘル	0.07
高専	0.17	値下がり	0.14	ポッチャマ	0.05
千早	0.17	ロイター	0.13	看る	0.04
賽銭	0.17	アイドルマスター	0.11	野幌	0.01