オンライン・コミュニケーション上での平均使用語彙数に関する研究

荒牧 英治（東京大学 知の構造化センター／JST さきがけ）†
増川 佐知子 （東京大学 知の構造化センター）
森田 瑞樹 （東京大学 知の構造化センター／医薬基盤研究所）
保田 祥（東京大学 知の構造化センター）

Study on Average Japanese Vocabulary for Online Communication

Eiji Aramaki (University of Tokyo)

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(森岡健二 1951)やマスメディア（雑誌やテレビ）に出現する語の調査(国立国語研究所 1999; 玉村文郎 2002)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）は行われていない．使用語彙を調査するの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(国立国語研究所 2012)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」(林四郎 1971)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．

本研究ではネット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 7,000 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．

２．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 Web 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．

eiji.aramaki@gmail.com

- **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）.

なお，形態素解析には juman7.0 (Kurohashi, Nakamura et al. 1994)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.

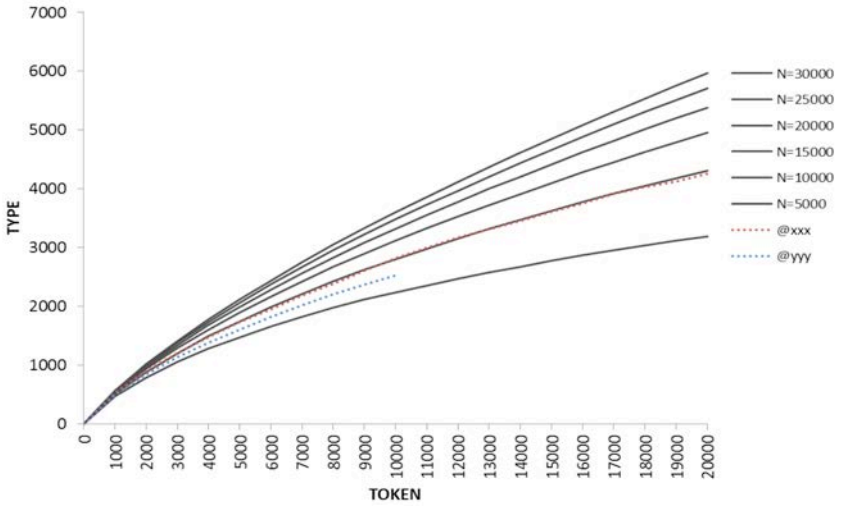

図 1: 語彙数 N（N=5000..30000）のタイプ・トークン曲線

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.

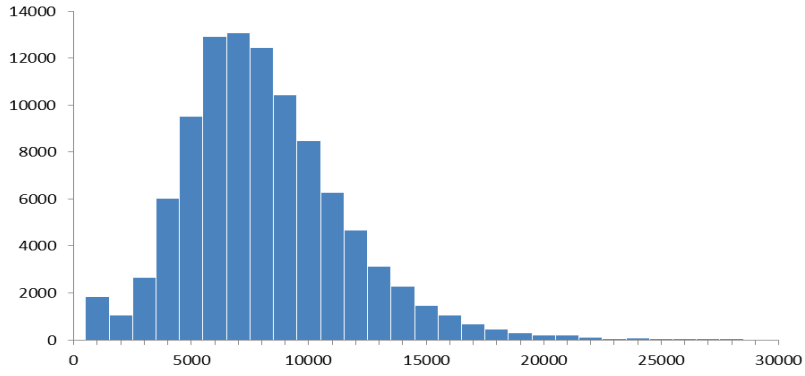

図 2: ユーザ数と語彙数．X 軸は語彙数，Y 軸はその語彙数を持つユーザ数を示す

３．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話した語数から，潜在的な語彙数を推測する．この推測はユーザがジップ則(Zipf 1935; Zipf 1949)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 1 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=5,000…30,000）ごとにプロットしている．語彙数（N）は 1,000 きざみに 50 本を用意している．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，実際にコーパスから抽出したユーザのタイプ・トークン曲線を図中の点線で示す．ユーザ@xxx については実験の 5 ヶ月で 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 を持つと推測できる．このようにして，12 万人すべての語彙数を推定することができる．

４．結果

推定された語彙数のヒストグラムを図 2 に示す．最も多い語彙数（最頻値）は 7,000 語であった．この 7,000 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(林四郎 1971)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 1/6 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．

４．１ 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．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%近くのユーザ東京に集中し，かつ，20 代のユーザを多いとされている(リンクシェア・ジャパン株式会社 2010)．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事態が，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るであろうし，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．

４．２ 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る．よって，今後，旅行会話辞書の大幅なコンパクト化や，ユーザ数の多い語彙を優先して学習することで効率的な語彙習得なども可能となる．

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．

文　献

Kurohashi, S., T. Nakamura, et al. (1994). Improvements of Japanese Morphological Analyzer JUMAN. The International Workshop on Sharable Natural Language Resources.

Zipf, G. K. (1935). The Psychobiology of Language, Houghton-Mifflin.

Zipf, G. K. (1949). Human Behavior and the Principle of Least Effort, Addison-Wesley.

リンクシェア・ジャパン株式会社. (2010). "Twitter 利用実態調査."

国立国語研究所（1999). 高頻度語彙から見たテレビ放送語彙の特徴, 大日本図書.

国立国語研究所. (2012). "鶴岡調査."

林四郎（1971). "語彙調査と基本語彙." 国立国語研究所報告 **39**.

森岡健二（1951). "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 **2**: pp.95-107.

玉村文郎（2002). NAFL Institute 日本語教師養成通信講座８ 日本語の語彙・意味, アルク.