小・中・高校生の語彙数調査および単語親密度との関係分析

藤田 早苗† 小林 哲生† 山田 武士† 菅原 真悟‡ 新井 庭子‡ 新井 紀子‡
† NTT コミュニケーション科学基礎研究所 ‡ 教育のための科学研究所

{sanae.fujita.zc, tessei.kobayashi.ga, takeshi.yamada.bc}@hco.ntt.co.jp

1 はじめに

学習者の語彙数を測定することは教育支援につなげるためにも重要である [1, 2]．語彙数の調査はこれまでも多くなされてきた．しかしその多くは，大学生 [1, 2, 3] や成人 [4, 5]，あるいは幼児 [6] を対象としており，小学生～高校生を対象とする大規模な語彙数調査はこれまであまり報告がない．阪本 [7] は，6～20 歳まで計約 4 万人という大規模な調査を行っているが，調査実施は 1937 年であり，調査結果がそのまま現代に当てはまるとは考えにくい．

語彙数調査の方法も様々な方法が提案されている．荻原 [3] は，辞書の全見出し語について知っているか調べる全数調査を行っているが，全数調査は時間と労力がかかるため，大規模な調査の実施は現実的ではない．また，正解を問う問題を用いて語彙数を調べるテスト [8, 9] も考案されているが，例えば佐藤ら [8] のテストでは調査時間が 125 問で 40 分かかるなど，調査に時間がかかるうえ，多肢選択式の問題を人手で個々に作成する必要があるため拡張や変更は容易ではない．

一方で，天野ら [4, 5, 10] は，単語親密度 [11] に基づいて調査対象語を選ぶことで，少数の語の調査だけで語彙数を推定できるテストを考案した．天野らは 50 語だけで語彙数を推定するテストを過去に Web で公開しており，心理学調査などで用いられきた．単語親密度に基づく語彙数推定テストは実施が簡単で拡張性が高く，大規模調査に向いている．ただし，単語親密度は約 20 年前に大学生以上を対象とした評定実験によって得られたデータであり，これに基づく語彙数調査も大学生以上を対象としたものがほとんどである [2, 4]．

そこで本研究では，小学生～高校生約 2,500 人を対象に単語親密度に基づく語彙数調査を実施する．本研究では，現代の小学生～高校生の語彙数を調査するだけでなく，単語親密度との関係を分析することで，語彙数テストを小学生～高校生に適用する場合の有効性を検証する．なお，本調査は，語彙数と読解力の関係分析にも利用することを目的に，読解力を測るテストであるリーディングスキルテスト (以下，RST)[1][12] と同時に実施している．

本研究の貢献は次の通りである．

(1) RST との関係分析にも利用できる単語親密度に基づく語彙数推定テストを新たに作成 (2 章)．

(2) 小中高校生 2,469 名に対し，語彙数推定テストを実施 (3.1 節)．

(3) 学年と語彙数増加の関係を分析．語彙数は学年とともに増加する傾向があり，小 6 と高 2 で語彙数が約 2 倍になることを示す (3.2 節)．

(4) 調査結果と単語親密度の関係を分析．両者に強い相関があることを示す (3.3 節)．

2 語彙数推定テストの作成方法

単語親密度 [11] は，大学生以上を対象とした評定実験で語のなじみ深さを 1-7 の値で数値化したものであり，数値が高いほどなじみのある語であることを示している．天野らは，文字単語親密度，文字音声単語親密度，音声単語親密度の 3 種類の親密度を取得しているが，本調査は，文字での提示になるため，文字単語親密度 (以下，親密度) に基づき，テスト用の語を選定する[2]．

基本的な語彙数推定方法は次の通りである．まず親密度が付与された辞書の語を親密度順に並べ，一定の間隔で試験用の語を選択する．ここで，親密度ごとに語数は異なるため，親密度の間隔は一定にはならない[3]．次に，被験者には，語を「知っている」かどうかを回答してもらい，「知っている」と回答された語の中で，最も親密度の低い語より親密度の高い語は全て知っていると仮定し，語彙数を推定する．ただし，知っている語と知らない語の親密度境界付近では，ば

[1] https://www.yozemi.ac.jp/rst/

[2] 天野らは，2002 年に，第 1 巻に含まれなかった約 3 万語の文字単語親密度を追加調査しているが，本稿では，Web 公開版と同じく第 1 巻のデータのみを利用する

[3] 例えば，親密度が 5.5 の語は 749 語あるが，2 の語は 338 語であるため，親密度 2 より 5.5 の周辺の語の方を多く用いる．

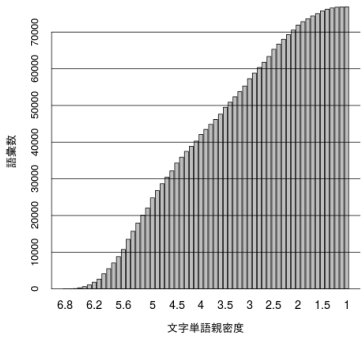

図 1: 文字単語親密度ごとの累積語数

らつきがあると考えられるため，Web 公開版では，親密度順に知らない語が 2 つ以上連続する語の親密度と，知っている語が 2 つ以上連続する語の親密度との中間点を親密度の境界としている．本稿でも同様に語彙数の推定を行う．

図 1 に親密度ごとの累積語数を示す．図 1 からもわかるように，例えば，親密度が 5 までの語を知っている場合は語彙数 24,801 語， 4 までの語を知っている場合は語彙数 42,090 語と推定される．

単語親密度に基づく語彙数推定方法の利点としては，(1) テストに利用する語は同じ親密度の語であればどの語を利用してもよい，(2) 実施が簡単である，(3) 短時間で回答できる，ということがあげられる．ただし，自己申告に基づくため，うそや誤解に基づく誤りが含まれる可能性はある．

本調査で用いるテスト用の語の選択方法は上記の手順に準ずるが，次の制限をかけた．
(a) RST の問題文に出現する語であること．これは RST の成績との関係分析にも利用するためである．
(b) 表記の妥当性[4]が 4 以上の語であること．これは，通常使われない表記を出題することによる受験者の混乱を避けるためである．

これにより，小学生用として 50 語，中高生用として 54 語を選択した．両者に共通する語は 32 語である．なお，両者のテストが同一でないのは，RST では小学生と中学生以上では問題が異なっており，問題文からテスト用の語を選ぶという制約を課すと，完全に同一にできなかったためである．

[4]日本語の語彙特性 [11] 第 2 巻参照．同じエントリに対して複数の表記がありえる時の妥当性を評定しており，最大値は 5．例えば,「食い違う」 は 4.70 で,「食違う」 は 3.55 など．

表 1: 参加学校数と回答者数

小中高	学年	回答者数 男	女	その他	計
小学校 (8 校)	6	182	163	68	413
中学校 (1 校)	1	60	45	2	107
	2	71	48	3	122
高校 (5 校)	1	450	346	175	971
	2	464	349	43	856
合計 (14 校)		1,227	951	291	2,469

3 語彙数推定テストの結果

3.1 調査の時期・規模

RST の受験と語彙数推定テストを同時に実施した．テストはタブレットで実施しており，提示した語について,「知っている」「わからない」「知らない」の 3 択で回答してもらった．ただし，以降の分析では「わからない」は「知らない」と同等に扱う．

調査の実施時期は，2019 年 1 月 13 日から 2019 年 2 月 15 日である．表 1 に調査の参加学校数と回答者数を示す．なお，参加校は全て公立学校である．

3.2 語彙数推定結果

Web 公開版と同様，語彙数推定を実施した．図 2 に，教育段階別の推定語彙数を示す．グラフには，各学年での推定語彙数の第一四分位数，中央値，第三四分位数も記載している．

図 2 からもわかるように，学年があがるにつれ，おおむね語彙数は上昇している．小 6 では，中央値で 19,267 語だが，高 2 では， 38,395 語となっており，ほぼ倍増している[5]．

さらに，男女別に分けた結果を図 3, 4 に示す．性別に関する質問に無回答だった人は含まれていない．男性のみの結果 (図 3) と，女性のみの結果 (図 4) を比較すると，すべての学年で男性の推定語彙数の方が高くなっている．ただし， 2 章でも述べたように，推定語彙数は自己申告に基づいているため，この結果を鵜呑みにはできない．特に男性では，全ての学年で推定語彙数が最大値，すなわち出題した語をすべて知っていると回答した人が相当数いる．一方で，女性では出題した語をすべて知っていると回答した人は圧倒的に少ない．女性の方が，正直に回答した人が多いとすると，小 6 の実際の推定語彙数は中央値で 17,895 語，

[5]なお，荻原 [3] によると，大学 4 年生の理解語彙は，中央値で 43,473 語だった

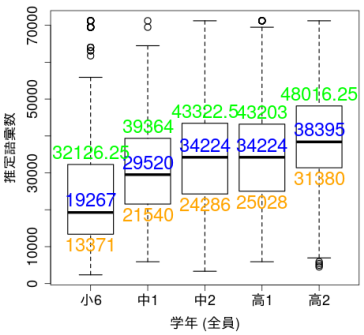

図 2: 教育段階別推定語彙数

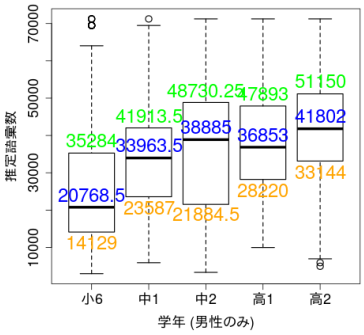

図 3: 教育段階別 (男性)

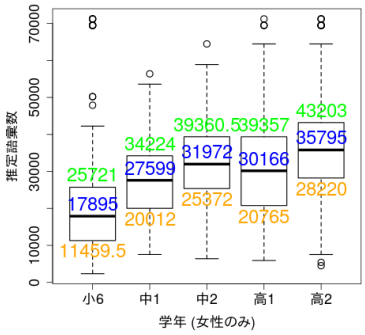

図 4: 教育段階別 (女性)

高 2 で 35,795 語となる．いずれにしろ，小 6 から高 2 になる間に語彙数はほぼ倍になっていると言える．

なお，全語を知っていると回答しても外れ値とならない学年もあり，いい加減な回答の発見・対処方法は今後の課題としたい．

3.3 各語を「知っている」割合

3.3.1 単語親密度との関係

本節では，単語親密度と各語を知っていると回答した人の割合の関係を分析する．図 5 に小学生，図 6 に中高生の散布図を示す．図 5，6 に示したように，単語親密度と各語を知っている人の割合には強い相関がある．単語親密度の調査が行われたのは 20 年以上前であるが，現在でも多くの語で通用する信頼性の高い値であると言える．特に中高生の場合，$r = 0.868$ と非常に強い相関がある．単語親密度の調査が成人で行われているため，より大人に近い中高生の相関の方がより強くなったと考えられる．

小学生でも $r = 0.72$ と強い相関があるが，親密度が 5 から 6 あたりの語では大きくばらついている．親密度が 6 以上の語でも「銀行」($psy = 6.406$) で 99.3%，「経済」($psy = 6.281$) で 73.8%，「大部分」($psy = 6.188$) で 48.6%と差がある．親密度 5.3 の周辺の語では，「取り込む」($psy = 5.375$) は 84.8%と高いが，「総称」($psy = 5.344$) は 27.7%，「成就」($psy = 5.219$) は 27.2%と低い．

親密度 5 以上の語は成人の 95%が知っていると言われるが，小学生はあまり知らない語も多く，小学生以降早期に覚えていく語が多く含まれると考えられる．

また上の例から，近い親密度であっても，抽象度の高い語や文語的な語のほうが知っている人が少ない傾向が見える．今後は抽象度なども反映し，特に小学生の語彙数推定方法を改良したい．

3.3.2 教育段階別の割合変化

小学生と中高生で重複している 32 語について，各語を知っていると回答した人の教育段階ごとの割合変化を分析した．親密度順に 3 グループに分け，図 7-9 に割合変化を示す．

多くの語は学年が上がるにつれ，知っていると回答した人の割合も高くなっている．また，親密度の高い語 (図 7) は，小 6 から知っている人の割合が高い語が多いが，親密度が低くなるに従い，小 6 で語を知っている人の割合が低い語が多くなる．ただし，小 6 で知っている人の割合が低い語であっても，「弦」($psy = 4.844$) までの語は，「成就」を除き，高 2 ではおおむね 80%の人が知っていると回答するようになる[6]．一方，「首長」($psy = 4.781$) 以下の親密度の語は，高 2 でも知っている人の割合は 60%を下回っている．

また，多くの語では知っている人の割合はなだらかに上昇しているが，図 7 の「比喩」は中 1 で急激に知っている人の割合が高くなる．これは「喩」が中 1 で習う漢字であるためだと考えられる．このように，学校教育によって強く影響されている語も見られた．

4 まとめと今後の課題

本稿では，単語親密度に基づく語彙数推定テストを作成し，小中高校生 2,469 名に対する実施結果を分析した．その結果，語彙数は学年とともに増加する傾向があり，小 6 と高 2 で語彙数が約 2 倍になることを示した．また，本稿で用いた単語親密度は 20 年以上前の 18〜29 歳での調査[7] に基づいた値だが，本調査で各語を知っていると回答した人の割合との間に強い相

[6]「成就」は，高 2 でも知っていると回答したのは 57.6%のみだった

[7]日本語の語彙特性第 1 巻の親密度の評定実験の時期は 1995 年 9 月から 1997 年 7 月

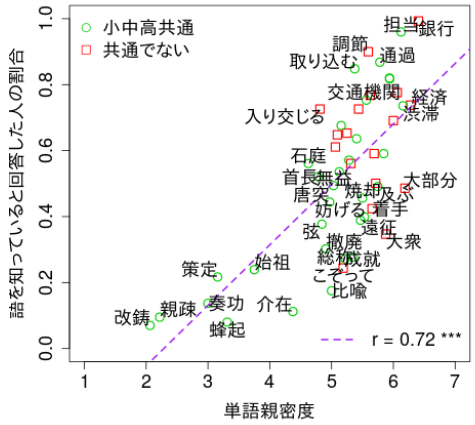

図 5: 単語親密度と語を知っている人の割合 (小学生)

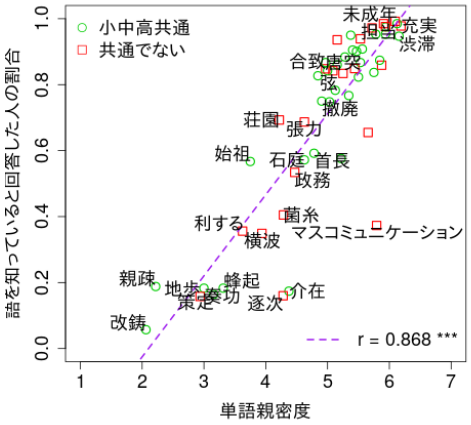

図 6: 単語親密度と語を知っている人の割合 (中高生)

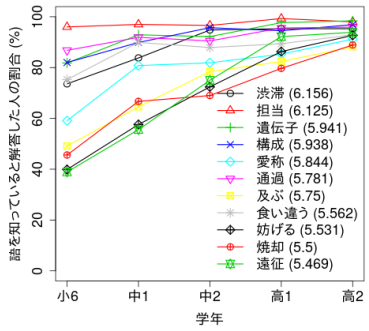

図 7: 教育段階と語を知っている人の割合 (1)

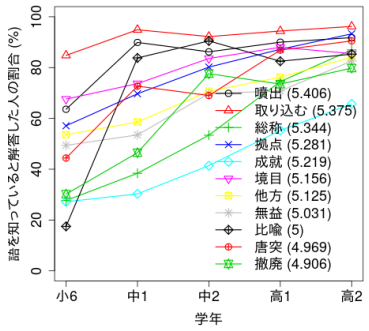

図 8: 教育段階と語を知っている人の割合 (2)

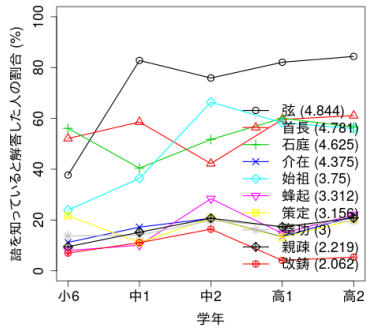

図 9: 教育段階と語を知っている人の割合 (3)

関があることを示した．NTT では単語親密度の再調査も行っており [13][8]，その結果を反映した分析や語彙数推定テストの再作成にも取り組みたい．

一方で，特に小学生の場合，親密度が同程度の語でも知っている人の割合が大きく変わる語が見られた．抽象度や文語的な語かといった要素の影響が考えられ，今後の語彙数調査への反映方法を検討したい．

最後に，本調査では中学生の回答者数が他学年に比べて少なかったため，今後，中学生の追加調査と，大人での調査も計画している．RST との関係分析の結果についても稿を改めて報告したい．

参考文献

[1] 田島 ますみ, 佐藤 尚子, 橋本 美香, 松下 達彦, 笹尾 洋介. 日本人大学生の日本語語彙測定の試み. **中央学院大学人間・自然論叢**, (41):3–20, 2016.

[2] 松浦 年男. 大学初年次の学生に対する日本語語彙力調査の試行. **北星学園大学文学部北星論集**, 52(2):53–61, 2015.

[3] 荻原 廣. 日本人の語彙量 (理解語彙、使用語彙) 調査を行うにあたっての基礎的研究 (日本語学特集). **京都語文**, (21):1–30, 2014.

[4] 天野 成昭, 近藤 公久, 片岡 良治. 単語親密度を利用した語彙数推定 -インターネットによる大規模調査-. **日本認知科学会第 *22* 会大会**, pp. 58–59, 2005.

[5] 小林 哲生, 天野 成明, 正高 信男. **モバイル社会の現状と行方**. NTT 出版, 2007.

[6] 小林 哲生, 奥村 優子, 南 泰浩. 語彙チェックリストアプリによる幼児語彙発達データ収集 の試み. **電子情報通信学会技術研究報告**, 115(418):1–6, 2016. (HCS2015-59).

[7] 阪本 一郎. **読みと作文の心理**. 牧書店, 1955.

[8] 佐藤 尚子, 田島 ますみ, 橋本 美香, 松下 達彦, 笹尾 洋介. 使用頻度に基づく日本語語彙サイズテストの開発：50,000 語レベルまでの測定の試み. **千葉大学国際教養学研究**, pp. 15–25, 2017.

[9] 松下 達彦. 「日本語を読むための語彙サイズテスト」の開発. *2012* **年日本語教育国際研究大会第 1 分冊**, pp. 310, 2012.

[10] Shigeaki Amano and Tadahisa Kondo. Estimation of mental lexicon size with word familiarity database. In *Proc. of ICSLP-98*, vol. 5, pp. 2119–2122, 1998.

[11] 天野 成昭, 近藤 公久. **日本語の語彙特性**. 三省堂, 東京, 1999.

[12] Noriko Arai, Naoya Todo, Teiko Arai, Kyosuke Bunji, Shingo Sugawara, Miwa Inuzuka, Takuya Matsuzaki, and Koken Ozaki. Reading skill test to diagnose basic language skills in comparison to machines. In *Proc. of CogSci-2017*, pp. 1556–1561. 2017.

[13] 藤田 早苗, 小林 哲生. 単語親密度の再調査と過去のデータとの比較. **言語処理学会第 *26* 回年次大会** *(NLP-2020)*.

[8] 2018 年 9 月から 2019 年 12 月にかけて実施