コーパス統計量と読解能力値に基づいた単語の既知率の予測

相原理子[1]　石川香[1]　藤田早苗[2]　新井紀子[3]　松崎拓也[1]
[1]東京理科大学 理学部第一部 応用数学科
[2]NTT コミュニケーション科学基礎研究所, [3]教育のための科学研究所
{1417001, 1417009}@ed.tus.ac.jp, sanae.fujita.zc@hco.ntt.co.jp, arai@s4e.jp, matuzaki@rs.tus.ac.jp

1 はじめに

小学校から大学までどの学習段階においても教科書は最も主要な知識源である．よって児童・生徒・学生が教科書をどの程度理解できるか，あるいは逆に，教科書がその対象読者に理解可能な形で書かれているかを知ることは重要である．一方，読解に関するこれまでの研究から，あるテキストを理解するためには，そこに含まれる単語の大部分が既知であることが必要だと知られている．そこで，様々な学年の児童生徒にとって，教科書中の単語がどれくらい既知であるかを正確に知りたい．

この目的の下，本研究ではコーパス統計量や児童生徒の読解能力スコアを用いて，各単語の既知率を正確に予測することを試みた．ここで，ある単語の既知率とはある集団（例えば同学年の児童）における，その単語を知っている人の割合を意味する．語彙量に従ってテキストの理解度も向上するという結果は様々な実験から知られており，例えば、Schmitt ら [1] は，テキストの内容を 60%以上理解するためには，テキスト中の単語の約 98%を把握している必要があるとしている．

特に教科書中の語彙が，その対象である児童生徒にとってどの程度まで既知であるかは重要な基礎情報だが，これを調査した例は我々の知る限り存在しない．その主な理由は，教科書中の単語についての全数調査はもちろんのこと，十分な量のサンプリング調査でも児童生徒への負担が極めて大きく，現実的でないことであろう．そこで本研究では，コーパス統計量などを用いて，教科書中の各単語の既知率を正確に予測することが可能かどうかを調査する．

2 関連研究

2.1 語彙知識予測問題

語彙知識予測問題とは，ある人がある単語を知っているかどうかを予測する問題である．例えば，江原 [2] は単語テストに用いる単語を選ぶ際に，学習者の能力を正確に把握するのに適した項目を選ぶことによって，語彙知識予測の性能を向上させる手法を提案した．この研究では，項目の難易度を正確に把握することが重要であるが，単語の難易度は，学習者によって異なるため，項目反応理論の Rasch モデルを発展させ，学習者にとっての単語の難易度を計算するモデルを提案している．

2.2 単語親密度による語彙数調査

単語親密度とは語のなじみ深さを 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す．天野ら [3] は 18～29 歳の評定者，藤田と小林 [4] は 18~35 歳の評定者からこの数値を得ている．藤田ら [5] はこの単語親密度を用いて小学生から高校生を対象に語彙量調査を行っている．この調査により，学年が上がるにつれて，おおむね語彙量が増大していること，また単語親密度と各語を知っている人の割合（既知率）には強い相関があることが分かっている．ここで注意すべき点は，学年ごとに単語親密度と既知率の関係は異なること，さらに，特に小学生では同程度の親密度を持つ語の間でも既知率に大きな差がある場合があることである．よって個人の語彙量の推計ではなく，具体的な教科書中の語彙のカバー率（既知の単語の率）を推計する際には，単語親密度以外の情報も併用して個々の語の既知率を正確に推計することが望まれる．

3 背景

3.1 現代日本語書き言葉均衡コーパス

現代日本語書き言葉均衡コーパス(BCCWJ)は,短単位で数えて約 1 億語を含む大規模均衡コーパスである．BCCWJ は計 13 のレジスターから構成されている．表 1 にレジスターごとの延べ語数（短単位）を示す（数値はマニュアル [6] に基づく）．

表 1 各レジスターの語数

レジスター	語数(万)	レジスター	語数(万)
書籍(PB)	2866	ベストセラー(OB)	375
雑誌(PM)	450	Yahoo!知恵袋(OC)	1030
新聞(PN)	138	Yahoo!ブログ(OY)	1028
書籍(LB)	3044	韻文(OV)	23
白書(OW)	494	法律(OL)	108
教科書(OT)	93	国会会議録(OM)	510
広報紙(OP)	383	合計	10542

3.2 教育基本語彙

教育基本語彙[7]とは学習者に教育する基本的な言葉であり，国立国語研究所が 7 種の教育基本語彙をデータベース化したものである．本研究では，教育基本語彙に含まれるかどうかとその単語の既知率との関係を調べた．

3.3 リーディングスキルテスト

リーディングスキルテスト（RST） [8] は主として教科書から採った 1〜2 文からなる短いテキストを用いた読解能力テストである．これまで中高生を中心に，小学生および成人も含め約 20 万人が受検している．RST の問題は以下の 6 タイプからなる：係り受け解析（DEP），照応解決（ANA），含意関係の認識（INF），定義に適合する具体例の選択（INST），画像の理解（REP），同義文か否かの判定（PARA）．テスト結果に従って，各受検者に対し問題タイプごとに項目反応理論に基づく能力値が推定される．本研究では受検者がある単語を知っているか否かを能力値をもとに推測することを試みる．

4 調査対象とするデータ

4.1 藤田らによる語彙調査結果

藤田ら [5] は，親密度がおよそ 2.0〜6.5 の単語およそ 50 語について，小学校 6 年生から高校 2 年生までの約 2500 名を対象に，各単語を知っているか否かを調査した．本研究ではこのデータを調査対象の一つとした．以下，これを「データ A」と呼ぶ．

4.2 語彙テストと RST 能力値データ

小学 6 年生から中学 3 年生までの約 4300 人を対象に語彙テストを行った．テストした単語は 40 単語で，各被験者はそのうち 10 単語について，「知っている/たぶん知っている/たぶん知らない/知らない」のうち 1 つを選んだ上で，語義を問う 4 択のテストに答えた．「知っている」または「たぶん知っている」を選び，かつ 4 択のテストに正解した語を「既知」とみなす．語彙テストに加え DEP，INF，INST，REP の 4 つの RST 問題タイプに対する各被験者の能力値を測定した．以下，これを「データ B」と呼ぶ．

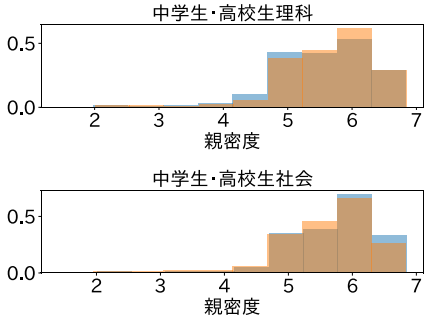

図 1 中学高校教科書における単語親密度の分布

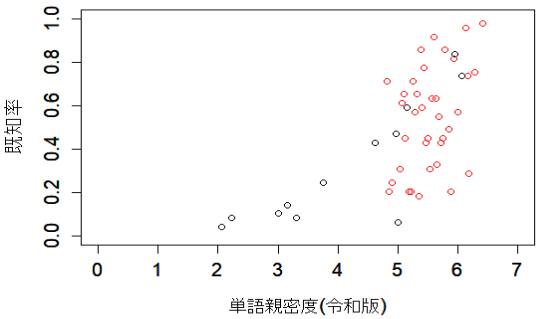

図 2 親密度（令和版）と既知率の関係

5 分析結果

5.1 中学校・高校の教科書テキスト比較

図 1 は，中学校・高校の理科・社会の教科書中の単語の親密度の分布である．オレンジのヒストグラムが中学校，青のヒストグラムが高校の教科書における分布を表す．具体的な科目としては，中学校理科は理科 3 年 [9]，中学校社会は公民 [10]，高校理科は物理 [11]，高校社会は現代社会 [12] の教科書を調査した．図から，理科の科目どうし，社会の科目どうしの親密度の分布は比較的似ていることが分かる．言い換えれば，類似する科目の教科書であれば，中学・高校の教科書中の単語の親密度の分布には大きな差がない．このことから，もし中学生・高校生の間で使用する教科書中の単語カバー率に差が

表 2 学年ごとの各レジスターの線形回帰係数と自由度調整済み決定係数

	小6	中1	中2	中3	高1	高2
書籍(LB)	-11.7	5.95	5.52	2.44	4.64	-0.33
Yahoo!知恵袋	0.79	‡8.59	†6.46	2.21	1.17	1.29
国会会議録	4.05	0.57	1.66	3.72	2.51	0.32
白書	-5.66	-1.07	-1.34	-3.77	-3.68	-1.88
書籍(PB)	4.44	-8.06	-9.03	-2.42	-7.27	1.15
新聞	4.2	-2.76	-4.32	-3.85	-3.9	-3.3
ベストセラー	1.27	3.98	3.41	5.62	4.84	3.16
法律	-4.31	-0.59	-0.75	-1.47	-1.42	-1.03
広報紙	1.32	-0.21	-0.47	0.34	0.41	0.91
韻文	-0.13	2.5	3.76	2.83	2.58	1.53
Yahoo!ブログ	5.34	-1.16	1.98	3.72	7.31	‡6.30
雑誌	†8.28	4.22	3.73	0.13	2.53	-0.45
教科書	2.46	0.03	1.46	2.27	3.73	1.39
R^2	0.35	0.44	0.54	0.41	0.47	0.39
R^2(単回帰)	-1.12	0.23	0.5	0.52	0.55	0.63

あるとすれば，それは中学教科書が高校教科書に比べ「易しい」単語を多く含むためではなく，主として，同程度の親密度の単語に対する中学生・高校生の既知率の違いに起因すると予想される．

5.2 教育基本語彙・単語親密度と既知率

図 2 はデータ A のうち小学校 6 年生を対象とする結果について，教育基本語彙に含まれる単語(赤点)と含まれない単語（黒点）に分けて，横軸に単語親密度(令和版)，縦軸に既知率を取った散布図である．教育基本語彙は教育上重要と考えられている語を集めているため，特に低い学年では，同程度の親密度の単語でも教育基本語彙に含まれるものの方が既知率が高い可能性がある．しかし，図 2 から分かるようにそのような傾向は見られなかった．

5.3 単語出現頻度を用いた既知率の予測

図 3 は BCCWJ の 13 個のレジスターごとに，含まれる単語の親密度（平成版）の分布を示す（単語延べ出現数に対する分布)．各レジスターにおける親密度の分布には多少の違いが見られる．

データ A の各テスト語について BCCWJ の各レジスターにおける出現頻度を調べ，説明変数を各レジスターでの相対頻度の対数，目的変数を小学校 6 年生～高校 2 年生におけるその単語の既知率として，学年ごとに線形回帰を行った．ただし，頻度が 0 となるレジスターに対する説明変数の値は，その他のレジスターでの相対頻度の平均値の対数とした．

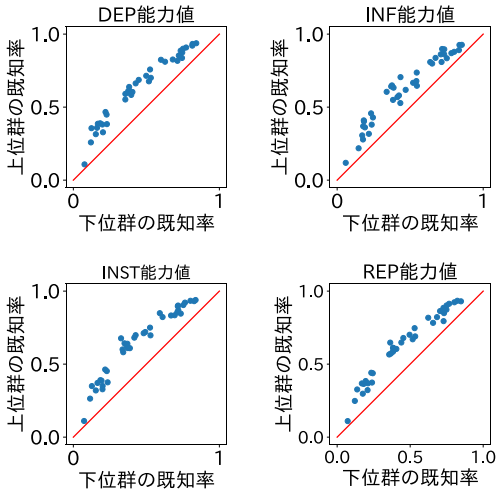

図 4 読解能力値の違いによる既知率の差

表 2 に，各レジスターに対応する回帰係数及び自由度調整済み決定係数(R^2)を学年ごとに示す．表中で，係数が非ゼロであることが 5%有意であるものには‡，10%有意であるものには†を付した．非ゼロであることが有意となる係数がごく少数であり，R^2の値も低いことからモデルのあてはまりは良くない．また，予測値が実測値を上回る単語，下回る単語が全ての学年でほぼ共通していた．このことからモデルの線形性の仮定が妥当でない可能性がある．

また，表 2 の最終行に BCCWJ 全体における相対頻度の対数を説明変数とする場合(すなわち単回帰の場合)の自由度調整済みR^2を示す．表の最後の 2 行の比較から，低い学年ではレジスターを分けた方が回帰の当てはまりが良く，高い学年ではレジスターを分けない方が良いことが分かる．

5.4 読解能力値を用いた既知語予測

各児童生徒の読解能力値を入力値とすることで，ある単語がその児童生徒にとって既知であるかどうかを予測したい．これが高い精度で可能ならば，個々の児童生徒にとっての教科書テキストの単語カバー率が推計できる．そこで，まず読解能力値と語彙テスト結果の関係について確認しておく．

図 4 はデータ B について 4 つの能力値ごとに縦軸に能力値が高い上位 1/4 の受検者における各テスト語の既知率，横軸に能力値が低い下位 1/4 の受検者における既知率をとった散布図である．図 4 から，

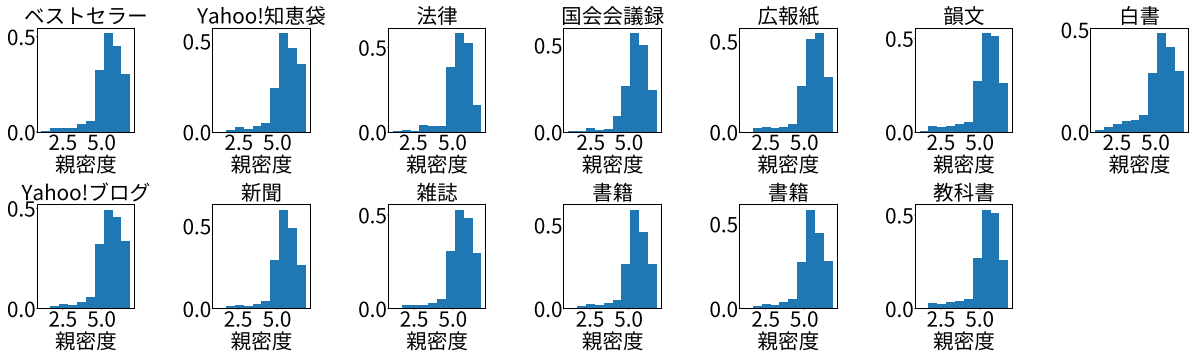

図 3　BCCWJ のレジスターごとの分布

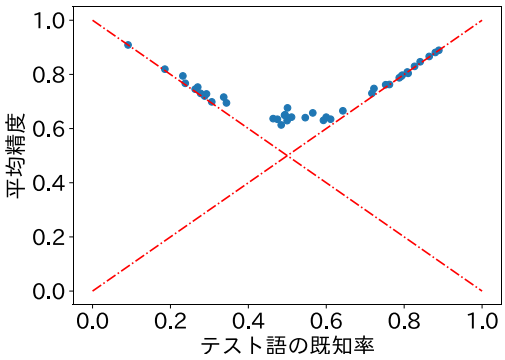

図 5 既知率と既知語予測の平均精度

どの能力値，どのテスト語についても能力値上位 1/4 群における既知率が下位 1/4 群におけるそれを上回っており，読解能力値と語彙量には関係があることが分かる．

次に，説明変数を受検者の 4 つの RST 能力値，目的変数を各受検者にとってテスト語が既知であった(1)か否か(0)として，個々の単語ごとにロジスティック回帰を行った．受検者を 10 分割した交差検証を行ったときの平均精度を図 5 に示す．赤い直線は $y = x$，$y = 1 - x$ のグラフで，全て既知，あるいは全て未知と予測した場合のベースライン精度を表す．赤線との比較から，既知率 50%付近の単語については予測精度がベースラインを上回るが，それ以外の単語についてはベースラインと同等の精度でしか予測できていないことが分かる．すなわち，4 つの RST 能力値を総合しても，個々の受検者にとって単語が既知かどうかを予測することは難しいことが分かる．

さらに，受検者の 4 つの RST 能力値を説明変数，各受検者にとってテスト語が既知であったか否かを目的変数とする上記のロジスティック回帰モデルに，説明変数として単語固有のパラメータ(項目反応理における困難度に相当する)を加えたもの（モデル 1）と，単語親密度を加えたもの（モデル 2）について，全テスト語に対するデータをまとめて訓練データとし 10 分割交差検証を行った．結果を表 3 に示す．単語親密度を説明変数として加えたモデル 2 の平均精度は 0.64 であり，個々の単語の既知率の推定に用いるには不十分であることが分かった．

表 3 困難度あるいは親密度を加えた回帰の結果

	正則化	RST 能力値 DEP	INF	INST	REP	親密度	平均精度
モデル 1	なし	0.256	0.211	0.226	0.403	—	0.44
	あり	0.147	0.104	0.116	0.159	—	0.60
モデル 2	なし	0.225	0.340	0.170	0.173	0.574	0.64
	あり	0.230	0.340	0.170	0.173	0.574	0.64

5 おわりに

教育基本語彙，BCCWJ での単語出現頻度，および RST 能力値を用いて，ある集団内での単語の既知率の予測および各受検者にとって単語が既知か否かの予測を試みた．いずれの予測も難しく，受検者ごとに単語が既知か否かを予測する精度が 70%を超えることはなかった．予測精度を上げるためには，受検者に関しさらに情報を加える必要があると考えられる．例えば，受検者の読書傾向や読書量，また家庭環境などの情報は有用であろう．しかし，全児童生徒にとっての教科書の単語カバー率を推定するという当初目的を考えると，そのような情報を全児童生徒について調査できるかという問題がある．

謝辞

教科書データを提供いただいた東京書籍株式会社に感謝します．本研究の一部は JST，さきがけ，JPMJPR175A の支援を受けたものである．

参考文献

[1] Norbers Schmitt, Xiangying Jiang, and Willian Grabe. The Percentage of Words Known in a Text and Reading Comprehension. *Modern Language Journal*, 95(1), pp.26-43, 2011.

[2] Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. Mining words in the minds of second language learners: learner-specific word difficulty. In *Proceedings of COLING*, 2012.

[3] 天野成昭, 近藤公久. 日本語の語彙特性. 三省堂, 1999.

[4] 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, 2020.

[5] 藤田早苗, 小林哲生, 山田武士, 菅原真悟, 新井庭子, 新井紀子. 小・中・高校生の語彙調査および単語親密度との関係分析. 言語処理学会第 26 回年次大会, 2020.

[6] 国立国語研究所. 『現代日本語書き言葉均衡コーパス』利用の手引. 2015.

[7] 国立国語研究所. 教育基本語彙の基本的研究：増補改訂版. 明治書院, 2009.

[8] Noriko H. 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 *Proceedings of the 39th Annual Cognitive Science Society Meeting*, pp. 1556-1561, 2017.

［9］岡村定矩ほか. 新編 新しい科学３. 東京書籍, 2016.

［10］坂上康俊ほか. 新編 新しい社会 公民. 東京書籍, 2016.

［11］三浦登ほか. 改訂 新編物理基礎. 東京書籍, 2017.

［12］間宮陽介ほか. 現代社会. 東京書籍, 2017.