技術・研究報告

小学生向け教育番組の音声に用いられる語彙の予備調査

淺井優介[a], 北村達也[a], 川村よし子[b]

[a] 甲南大学 知能情報学部 知能情報学科
神戸市東灘区岡本８－９－１, 658-8501
[b] 東京国際大学 言語コミュニケーション学部 英語コミュニケーション学科
川越市的場北１－１３－１, 350-1197

（受理日 2020 年 5 月 12 日）

概 要

日本語教育が必要な児童向けの教材作成の基礎データを提供するために，NHK Eテレの小学生向け教育番組の音声を書き起こし，語彙表を試作した．低学年向け 17 番組計 510 分，高学年向け 19 番組計 570 分を音声認識を利用して書き起こし，その中に現れた単語を有用度指標にもとづいて降順に並べ，語彙表を作成した．得られた語彙表は，先行研究にて作成された書き言葉コーパスに基づく語彙表よりも易しい単語が抽出されており，本研究の方法論の有効性が示された．

キーワード： 日本語教育，語彙表，出現頻度，単語親密度，NHK Eテレ

1 はじめに

近年，日本では出入国管理及び難民認定法の改正などにより在留外国人数が増加し続けている．その数は 2018 年 6 月時点で 2,829,416 人と日本の総人口の約 2 パーセントに達している [1]．このような中，日本語指導が必要な生徒数の増加が問題視されている．日本語指導が必要な生徒とは，日本語で日常会話が十分にできない，または日常会話ができても学年相当の学習語彙が不足しているため学習活動への参加に支障が生じている生徒を指す．特に小学生が多く，2018 年で 33,685 人と 10 年前から 45.8 %増加している [2]．親の就労などによって日本で生活する児童が増える中，経済的事情からインターナショナルスクールや日本語教室などに通わず，日本人が通う一般の学校に通うケースが増え，十分な日本語指導を受けられていないためであると考えられる．

その対策として，日本政府は 2014 年に学校教育法施行規則の一部の改正を行った．これによって，在籍する学級でなく，別教室にて個人の日本語能力に応じた指導を行うことが推奨され，児童の日本語能力の向上を目指した環境作りが行われている．しかし，教員が児童ひとりひとりに指導計画を立てて実行する必要があるため，日々の授業に加え新たに日本語指導も行わなければならず，教員の負担は重い．その上，小学校の教員には日本語指導の経験がないため，教材やノウハウがほとんど存在しないことが特別の教育課程を導入する妨げとなっている [3].

このような状況を改善するため，本研究は日本語指導が必要な生徒向けの教材作成の基礎となる語彙表を作成することを目的とする．語彙表とは指導対象が学ぶべき単語をリストアップしたものである．本研究では「小学生向け教育番組ではその対象学年の一般の日本人児童が理解できる単語や表現で構成されている」と仮定し，NHK Eテレの教育番組の音声に基づいて語彙表を試作する．その理由は，(1) 番組に対象学年が設定されていること，(2) ナレーション，人形劇，多人数による掛け合いなど多彩な場面で用いられる単語を抽出できること，である．

以下では，低学年向け番組と高学年向け番組に分けて語彙表を作成し，比較することによって，学年間の語彙力の違いを調査する．さらに，現代日本語書き言葉均衡コーパス (BCCWJ) [4] に基づいて作成された日本語教育用語彙表 [5] と比較し，抽出された単語に差異が見られるか調査する．そして，得られた語彙表が小学生向けの易しいものであるかを単語親密度 [6] や日本語能力試験の出題基準に基づいて評価する．

2 語彙表の作成方法

2.1 対象

NHK for School [7] にて公開されている小学生向けの教育番組を対象とした．NHK for School では，NHK Eテレにて放送された種々の教科の番組を web ブラウザ上で視聴することができる．教科は，理科，社会，国語，算数，生活，実技 (音楽，体育，図工)，道徳，総合，英語，特別活動 (特活)，特別支援教育 (特支)，その他の 12 種である．それぞれの番組には対象学年が設定されており，本研究ではこの情報を利用して低学年と高学年で使われている語彙を比較する．1 話の長さは 5 分，10 分，または 15 分である．本研究では以下の条件にて調査対象を選択した．

1. 低学年向け番組は対象が 1 年生から 3 年生，または 1 年生から 6 年生のものとし，高学年向け番組は対象が 4 年生から 6 年生のものとする．
2. 抽出する番組数は，低学年，高学年それぞれ 1 教科から最大 3 番組とする．
3. 1 番組から 30 分間の動画を対象とする (5 分番組は 6 話，10 分番組は 3 話，15 分番組は 2 話)．

低学年向け番組と高学年向け番組の各教科ごとの番組数と合計時間をそれぞれ表 1，表 2 に示す．

2.2 音声の書き起こし

Google Chrome 上で利用できる音声認識アプリケーション Speechnotes を用いて上記の番組の音声をテキストに変換した．その後，番組の動画を見ながらテキストを修正した．ナレーション部は比較的高精度に認識できたが，会話部の認識精度は十分ではなく，番組によってはほぼ全ての音声を手作業で書き起こす場合もあった．

表 1: 低学年向け番組

教科	番組数	合計時間 [分]
理科	3	90
社会	1	30
国語	2	60
算数	1	30
道徳	2	60
生活	1	30
特活	2	60
特支	3	90
その他	2	60
合計	17	510

表 2: 高学年向け番組

教科	番組数	合計時間 [分]
理科	3	90
社会	3	90
国語	1	30
算数	2	60
道徳	2	60
実技	3	90
総合	3	90
特活	2	60
合計	19	570

2.3 単語の抽出

UniDic [8] の短単位で形態素解析を行い，単語 (語彙素) を抽出した．UniDic とはコーパス日本語学への応用を指向した形態素解析用電子辞書で，BCCWJ の解析を目的として開発されたものである．短単位は基準が明確で表記ゆれが少なく，意味を持つ最小の単位に分割するため，元の文で使われている語彙の形から離れすぎないので，用例収集や基本語彙の選定に適している [9].

本研究では，Web 茶豆 [10] を利用して形態素解析を行い，単語に分割した．このシステムは，web ブラウザ上で MeCab エンジンにより形態素解析を行うシステムであり,「現代語」,「現代話し言葉」,「上代 (万葉集)」など複数の辞書を切り替えて使用できる．本研究では「現代話し言葉」を使用した．

抽出した語彙から記号 (句点，句読点など)，助詞，助動詞，接頭辞，接尾辞，固有名詞，数詞，指示代名詞，接尾辞的形状詞，時に関わる語彙を除外した．これは本田 [5] と同じ条件である．その他，話し言葉に多く現れるフィラーも除外した．

2.4 語彙の有用度指標

本田 [5] は BCCWJ を対象にして語彙表を作成した．その際，それぞれの語彙の出現頻度と Deviation of Proportion (DP) [11] を乗じた有用度指標を用いて基本語彙を選定している．しかし，BCCWJ のサブコーパスに対して，本研究の調査対象のサブコーパス (番組) 数が少ないため，DP を計算する際にサブコーパスにおける語彙の期待値が大きくなり，単語間に大きな差が現れない．

そこで，本研究では多数の番組に出現する語彙の有用度が高くなるように次の指標を用いた．ある単語 w_i の出現頻度を tf_i，単語 w_i が出現する番組数を $\{d : d \in w_i\}$，総番組数を D とすると，その単語の有用度指標を以下の式で求める．

$$\log tf_i \times \frac{\{d : d \in w_i\}}{D} \tag{1}$$

本研究にて得られた語彙の出現頻度は最大で約 1000，最小で 1 と差が大きいため，出現頻度の影響を押さえるために対数化した．さらに，多数の番組に出現する語彙の有用度を高くするため，単語 w_i が出現する番組の割合を乗じた．この有用度指標に関して単語を降順に並べかえたものを本研究の語彙表とした．

3 語彙表の評価

得られた語彙表を以下の 3 種の指標を用いて評価した．

3.1 本田 [5] の語彙表

本研究で得られた語彙表と書き言葉コーパスから得られた語彙表との差異を調査するため，本田 [5] の語彙表と比較する．

3.2 旧日本語能力試験の出題基準

リーディング・チュウ太の語彙チェッカー [12] は，旧日本語能力試験の出題基準に基づいて入力文章中の単語の難易度を判定できる[1]．このシステムを用いて語彙表に含まれる単語の旧日本語能力試験における級 (1 級，2 級，3 級，4 級，級外) を調査し，易しい単語が抽出されているかを検討する．3 級，4 級の語を易しい単語と見なすことにした．

3.3 単語親密度

単語親密度とは，ある単語がどの程度なじみがあると感じられるかを数値で表した指標である．天野と近藤 [6] は大学生以上を対象にして単語親密度を調査した．単語親密度には，親密度を文字だけ

[1] 現在の日本語能力試験は出題基準が公開されていないため，語彙チェッカーでは公開されている旧日本語能力試験の出題基準が用いられている．

表 3: 低学年向け番組，高学年向け番組から抽出された語彙に関するデータ

	番組数	総語数	総異なり語数	1 文あたりの平均語彙数	標準偏差
低学年	17	19,820	2,927	6.80	2.17
高学年	19	27,293	3,770	8.14	1.59

で評価した文字単語親密度，音声だけで評価した音声単語親密度，文字と音声で評価した文字音声単語親密度の 3 種が存在する．山口ら [13] は小学 5 年生の児童を対象にして単語親密度を調査し，音声単語親密度が 5 以上の単語は児童にとてもなじみ深いと報告している．そこで，小学生にとって易しい単語が抽出されたかを検討するために，音声単語親密度 5 以上の語が含まれる割合を調査する．

4 結果と考察

4.1 低学年向け番組，高学年向け番組から得られたデータ

低学年向け番組，高学年向け番組から抽出された語彙に関するデータを表 3 に示す．また，得られた語彙表の上位 20 語を付録に掲載する．総語数は約 2 万語から約 3 万語，総異なり語数は約 3 千語から約 4 千語であった．BCCWJ を対象にした本田 [5] の調査では，総語数 9,463 万語，総異なり語数 3 万 6 千語であるので，本研究の規模は総語数で 10^{-4} のオーダー，総異なり語数で 10^{-1} のオーダーであり，大幅に小さいことがわかる．

低学年，高学年を比較すると，高学年の総語数，総異なり語数が低学年に対してそれぞれ 37.7%，28.8%多い．林 [14] は，6 歳の理解語彙が約 6,000 語，11 歳で約 20,000 語と報告している．このような発達による理解語彙への配慮が，上記の総語数や総異なり語数の違いに現れていると考えられる．また，1 文あたりの平均語数は低学年 6.80 語，高学年 8.14 語であり，これらの差異は有意傾向であった ($t(34) = 2.032,\ p = 0.0505$)．高学年の方が修飾語などの係り受けが多く，複雑な構造を持つ長い文を用いていることがわかる．

異なり語数 100 語ごとのテキストカバー率を図 1 に示す．低学年と高学年のグラフは同様の推移を示しており，上位 200 語で 50%，上位 800 語で 70%近くを占めていることがわかる．

4.2 本田 [5] の語彙表との比較結果

本田 [5] と本研究で得られた語彙表の上位 800 語における一致率を表 4 に示す．一致率は低学年で 50.1 %，高学年で 57.6 %となり，低い結果となった．

本田 [5] の語彙表には含まれず，低学年，高学年の語彙表に含まれる単語の例をそれぞれ表 5，表 6 に示す．低学年における名詞では児童にとって身近な単語といえる「友達」,「父」,「母」,「先生」，代名詞では「おまえ」という単語が見られた．動詞では「遊ぶ」,「見せる」,「集まる」,「開ける」,「探す」,「会う」など，話者本人の行動を表す単語が多く,「ござる」,「おいら」などのキャラ語も見られ

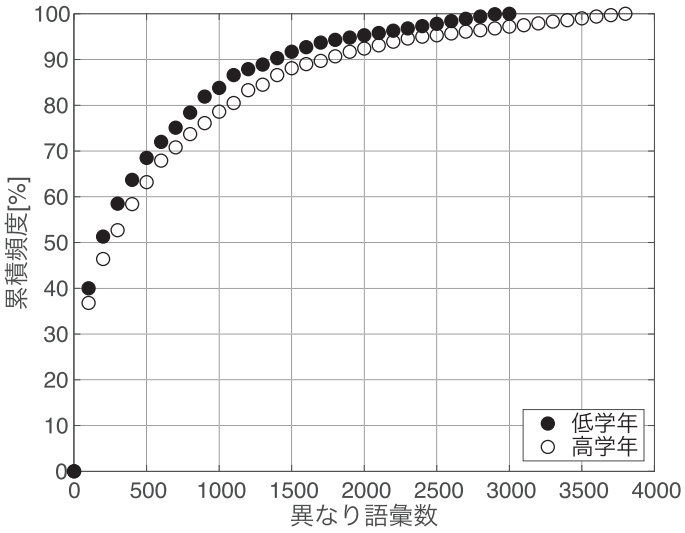

図 1: テキストカバー率．●：低学年，○：高学年．

た．高学年における名詞では「様子」,「答え」,「温度」,「量」などの観察に用いるような単語，動詞では「落ちる」,「調べる」,「起こる」など状態を表す単語が見られた．

4.3 旧日本語能力試験の出題基準，単語親密度との比較結果

低学年，高学年の語彙表および本田 [5] の語彙表の上位 800 語において，旧日本語能力試験の出題基準 3 級，4 級の単語が含まれる割合，音声単語親密度 5 以上の単語が含まれる割合を表 7 に示す．前者は，低学年では 64.8%，高学年では 60.4%，本田 [5] では 56.0%であり，低学年のものが最も高いという結果であった．また，低学年，高学年の上位 800 語には「パワー」,「葉っぱ」,「しゃべる」,「じゃんけん」,「ケンカ」,「気付く」,「クイズ」などの小学生にとって不可欠な単語と考えられる語が含まれていた．これらは，日本語能力試験では，級外の単語，つまり 1 級から 4 級のいずれにも含まれていないものである．したがって，本研究の方法によって，より易しい単語が抽出できるだけでなく，小学校生活を送る児童にとって必要な基本的な単語が抽出され，上位にランクされるといえる．一方で，音声単語親密度 5 以上の単語の割合は，3 つの語彙表の間で大きな差異はなかった．これは，単語親密度が大学生を対象に調査した結果をもとにしていることも影響しているものと考えられる．

以上のように本研究の有用性は明らかになったものの，小学生にとって必要不可欠だと考えられる単語であるにもかかわらず，出現頻度，番組数が共に 1 のため有用度が低くなり，下位にランクされてしまう場合があった．例えば，低学年での「ボール」,「マーク」,「まあまあ」,「わいわい」,「安心」,「協力」,「学年」,「怪しい,「丸める」などである．逆に，番組数が 1 であるにもかかわらず出現頻度が 2 であったため,「入水」,「鞭毛」,「微塵」などの専門的な単語が「ボール」などよりも上位にランクされてしまっていた．このような問題については，調査対象を拡大することによって解消されるものと予想される．

表 4: 上位 800 語における本田 [5] の語彙表との一致率

	一致率 [%]
低学年	50.1
高学年	57.6

表 5: 低学年の語彙表にのみ見られる単語

品詞	単語
名詞	友達，父，母，先生
代名詞	おまえ
動詞	遊ぶ，見せる，集まる，開ける，探す，会う
キャラ語	おいら，ござる

表 6: 高学年の語彙表にのみ見られる単語

品詞	単語
名詞	様子，答え，量，温度，水，国，風
代名詞	あなた
動詞	集める，落ちる，調べる，起こる

表 7: 上位 800 語における割合

	能力試験 3 級，4 級の割合 [%]	単語親密度 5 以上の割合 [%]
低学年	64.8	91.0
高学年	60.4	92.0
本田 [5]	56.0	96.0

5 おわりに

本研究では，日本語教育が必要な児童向けの教材作成の基礎データを提供することを目的として，NHK Eテレの小学生向け教育番組の音声を書き起こし，語彙表を作成した．低学年向け番組，高学年向け番組に現れる単語を調査したところ，総語数，総異なり語数ともに高学年の方が多く確認された．得られた単語を本研究にて提案した有用度指標にて降順に並べ，語彙表を試作した．今後は収集データを増やし，語彙表の完成度の向上を図りたい．

謝辞

本研究の一部は日教弘本部奨励金の支援により行われた．書き起こし作業にご協力いただいた皆様に感謝します．

参考文献

[1] 法務省入国管理国, “平成 30 年末現在における在留外国人人数について,” http://www.moj.go.jp/nyuukokukanri/kouhou/nyuukokukanri04_00081.html (2020 年 1 月 20 日閲覧).

[2] 文部科学省総合教育政策局, “日本語指導が必要な児童生徒の受け入れ状況に関する調査（平成 30 年度）の結果について,” https://www.mext.go.jp/content/1421569_002.pdf (2020 年 1 月 20 日閲覧).

[3] 文部科学省初等中等教育局国際教育課, “外国人児童生徒等教育の現状と課題,” https://www.bunka.go.jp/seisaku/kokugo_nihongo/kyoiku/todofuken_kenshu/h30_hokoku/pdf/r1408310_04.pdf (2020 年 1 月 20 日閲覧).

[4] 丸山岳彦, 山崎 誠, 柏野和佳子, 佐野大樹, 秋元祐哉, 稲益佐知子, 田中弥生, 大矢内夢子, “「現代日本語書き言葉均衡コーパス」におけるサンプリングの原理と運用,” 特定領域研究「日本語コーパス」平成 22 年度研究成果報告書, 2011.

[5] 本田ゆかり, 大規模コーパスに基づく日本語教育語彙表の作成. 東京外国語大学大学院総合国際学研究科博士 (学術) 論文, 2015.

[6] 天野成昭, 近藤公久, NTT データベースシリーズ 日本語の語彙特性 (第 1 期). 三省堂, 1999.

[7] NHK for School, https://www.nhk.or.jp/school/ (2020 年 1 月 20 日閲覧).

[8] 伝 康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵, “コーパス日本語学のための言語資源: 形態素解析用電子辞書の開発とその応用,” 日本語科学, vol. 22, pp. 101–123, 2007.

[9] 小椋秀樹, “「現代日本語書き言葉均衡コーパス」における短単位の概要,” 特定領域研究「日本語コーパス」平成 18 年度公開ワークショップ（研究成果報告会）予稿集, pp. 101–108, 2007.

[10] 堤 智昭, 小木曽智信, “歴史的資料を対象とした複数の Unidic 辞書による形態素解析支援ツール「Web 茶豆」,” じんもんこん 2015 論文集, pp. 179–184, 2015.

[11] Stefan Th. Gries, “Dispersions and adjusted frequencies in corpora,” *International Journal of Corpus Linguistics*, vol. 13, issue 4, pp. 403–437, 2008.

[12] 川村よし子, “語彙チェッカーを用いた読解テキストの分析,” 講座日本語教育, vol. 34, pp. 1–22, 1998.

[13] 山口俊光, 渡辺哲也, 大杉成喜, “教育基本語彙と成人の単語親密度との関係,” 情報処理学会研究報告音声言語情報処理 (SLP), vol. 2000, no. 12 (2006-SLP-060), pp. 31–35, 2006.

[14] 林 四郎, “語彙調査と基本語彙,” 電子計算機による国語研究, vol. 3, pp. 1–35, 1971.

付録 語彙表の上位 30 語

低学年向け番組から作成した語彙表

No.	単語	有用度指標
1	する	2.846
2	いる	2.516
3	何 (ナニ)	2.473
4	見る	2.452
5	ない	2.384
6	なる	2.371
7	ある	2.336
8	行く	2.258
9	こと	2.197
10	言う	2.193
11	来る	2.155
12	良い	2.045
13	やる	2.023
14	もう	2.013
15	皆	1.957
16	わかる	1.894
17	できる	1.820
18	くれる	1.770
19	うん	1.691
20	物 (モノ)	1.679
21	僕	1.627
22	あっ(感動詞)	1.609
23	時 (トキ)	1.594
24	ちょっと	1.567
25	で (接続詞)	1.563
26	本当	1.556
27	思う	1.539
28	出る	1.499
29	えっ(感動詞)	1.493
30	食べる	1.487

高学年向け番組から作成した語彙表

No.	単語	有用度指標
1	する	3.029
2	いる	2.798
3	こと	2.591
4	ある	2.543
5	なる	2.529
6	言う	2.525
7	見る	2.519
8	来る	2.377
9	良い	2.338
10	何 (ナニ)	2.320
11	ない	2.270
12	行く	2.255
13	できる	2.193
14	やる	2.083
15	皆	2.036
16	ちょっと	1.941
17	物 (モノ)	1.930
18	もう	1.907
19	わかる	1.840
20	時 (トキ)	1.833
21	出る	1.823
22	所 (トコロ)	1.820
23	人 (ヒト)	1.774
24	中 (ナカ)	1.766
25	前 (マエ)	1.711
26	まず	1.671
27	あっ(感動詞)	1.650
28	うん (感動詞)	1.643
29	ため	1.640
30	私	1.637