論　文

小学校低学年向け教育番組の音声における単語出現頻度の調査

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 2022 年 5 月 10 日）

概 要

日本語教育が必要な児童，特に低学年の教材作成に関する基礎データを提供するため，テレビの教育番組の話し言葉にて用いられる語とその出現頻度を調査した．NHK for School にて公開されている小学校低学年向け教育番組 53 時間の音声を書き起こし，形態素解析により語を抽出した．フィラー，記号，数詞，固有名詞以外の異なり語数が 12,398 語，延べ語数が 244,848 語のデータを得て，各番組の放送時間の差異を考慮した出現頻度に基づいてランキングを行った．

キーワード: 日本語教育，小学校低学年，NHK E テレ，教育番組，話し言葉，語彙調査

1　はじめに

近年，日本の学校では日本語指導が必要な児童生徒が増加している．2018 年度には公立学校における日本語指導が必要な児童生徒は 5 万人を超え，それまでの 10 年間で 1.5 倍の増加となった．一方で，これらの児童生徒の 2 割以上が日本語に関する特別な指導を受けることができていない．このような中，2014 年度に学校教育法施行規則が一部改正され，日本語指導が必要な児童生徒の日本語能力向上を目指した環境作りが進められている [1]．本研究は，日本語指導が必要な児童生徒に対する日本語教材作成に寄与するため，小学校低学年の教室活動の話し言葉で用いられうる語彙に関する調査を行う．しかし，教室活動の音声の収集には個人情報保護の障壁がある上，実際の作業も膨大となる．そこで，本研究では「教育番組はその対象学年の日本語母語話者が理解できる語や表現で構成されている」と仮定し，小学校低学年向け教育番組を対象に語彙の調査を実施するとともに，教育番組をこのような児童生徒を含む教育に利用する際に必要となる配慮についても考察する．

日本語教育分野ではかねてより学習者が優先的に学ぶべきいわゆる基本語に関する研究が行われてきた [2], [3]．コンピュータ能力の向上に伴い，近年は実際のデータに基づく手法が一般的となり，コーパスや教科書等の使用語彙調査が行われるようになった [4]-[10]．しかし，これらの研究は書き言葉を対象としたものがほとんどで，著者らが知る限り話し言葉を対象にした研究は獅々見 [11] などわ

ずかであり，特に小学校低学年の教室活動における話し言葉に着目した調査は見当たらない．そこで，本研究ではNHK Eテレの小学校低学年向けの教育番組を対象として語彙調査を実施した．NHK Eテレの教育番組は生徒の学習支援および学校教育の支援を目的とし，そのWebページ (NHK for School) には教材や教師向けの資料が提供されている．そのため，その教育番組にて用いられる語彙や表現を分析することによって，日本語を母語としない児童の学習・教育支援につながると考えた．本研究では，番組内の音声で使用されている語およびそれらの出現頻度のリストを作成し，使用語彙の傾向を分析する．

2 方法

2.1 分析対象

本研究では，NHKの学校教育に関するWebページであるNHK for School[1]にて公開されている教育番組を対象にした．NHK for Schoolでは，NHK Eテレにて過去に放送された教育番組をWebブラウザ上で視聴することができる．これらのうち，以下の条件を満たす全ての番組の全ての放送回を分析対象にした．

1. 2020年度上半期の作業時点でこのWebページにて公開されていること．
2. 小学3年生以下を対象とすること．対象に小学4年生以上を含む番組，例えば，対象が小学1年生〜6年生となっている番組は含まない．

本研究にて対象にした番組の一覧を表1に示す.「おばけの学校たんけんだん」など一部の番組の放送回数に半端な数が含まれるのは，作業時点で公開されている放送回数がそこまでだったためである．また，本研究同様にEテレの教育番組を対象にした淺井ら [12] では科目間で番組数のバランスを取っていたが，本研究ではできる限り多くのデータを収集するため，彼らのような調整は行っていない．

本研究の対象となった番組は，国語，算数，理科，社会，特活，生活，道徳，その他の8科目15番組である．科目間のバランス調整を行わなかった結果，対象番組のうち，小学3年を対象にした理科が3番組を占めることとなった．このことは，他科目に比べて理科の番組が多いことによるものだが，これが本研究の結果に何らかの偏りを与えている可能性があることを付記する．

表1に示すように番組によって放送回数や1話あたりの放送時間が異なる．そのため，1話あたりの放送時間に放送回数を乗じた値を d [分] とすると，この値が最大となるのは「銀河銭湯パンタくん」で380分，最小となるのは「えいごでがんこちゃん」と「すたあと」で100分であり，約4倍の開きがある．そこで，分析の際には，後述するように番組ごとの放送時間の差異の補正を行った．また，全番組の d の総和，すなわち総放送時間は3,190分 (53時間10分) であり，淺井ら [12] が対象にした低学年向け番組の総時間の6倍を超える量となった．

[1] https://www.nhk.or.jp/shcool/

表 1: 本研究にて対象にした番組とその放送回数，1 話あたりの放送時間

番組名	対象学年	科目	放送回数	時間 (分)
えいごでがんこちゃん	小学 1〜2 年	特活	20	5
おはなしのくに	幼保・小学 1〜3 年	国語	17	10
おばけの学校たんけんだん	幼保・小学 1〜2 年	生活	16	10
銀河銭湯パンタくん	小学 1〜2 年	道徳	38	10
ことばドリル	小学 1〜2 年	国語	20	10
こどもにんぎょう劇場セレクション	幼保・小学 1〜2 年	その他	10	15
コノマチ☆リサーチ	小学 3 年	社会	20	10
さんすう犬ワン	小学 1〜3 年	算数	19	10
しぜんとあそぼセレクション	幼保・小学 1 年	その他	25	15
新・ざわざわ森のがんこちゃん	幼保・小学 1〜2 年	道徳	28	10
すたあと	幼保・小学 1 年	生活	20	5
で〜きた	幼保・小学 1 年	特活	20	10
ふしぎエンドレス理科 3 年	小学 3 年	理科	20	10
ふしぎがいっぱい 3 年	小学 3 年	理科	20	10
理科 3 年ふしぎだいすき	小学 3 年	理科	19	15

2.2 書き起こし作業

音声の書き起こし作業は，複数名の日本語を母語とする大学生が行った．各放送回にはあらすじが提供されており，作業者はこのデータを参照しつつ動画を視聴しながら，PC への入力作業 (以下「書き起こし」) を行った．当初，音声認識の利用も検討したが，誤認識や漢字の誤変換を修正する作業の手間が多大であり，むしろ全て人手で入力した方が正確さを担保できるため，本研究ではこの方法を選択した．

書き起こしの際には，その後に行う予定の形態素解析における誤解析を減らすため，作業者にできる限り漢字を用いた表記を使うよう指示した．しかし，作業者による表記のばらつきや誤入力は避けられなかったため，書き起こし作業の終了後に著者の 1 人が全テキストデータを点検し，可能な限り書き起こしの誤りおよび表記ゆれを訂正した．

2.3 単語出現頻度の計測

上記の作業によって得られたテキストデータを形態素解析システム MeCab[2]で自動的に形態素に分割し，基本形や品詞の情報を得た．形態素解析用の辞書としては IPA 辞書を用いた．これは，現代話し言葉 UniDic [13] と IPA 辞書を比較した結果，後者ではユーザー辞書への単語登録が容易であり，

[2]https://taku910.github.io/mecab/

今回対象とした書き起こしデータに多数含まれる固有名詞や音声変化を伴う感動詞や役割語 (キャラ語) の解析に有利であったためである．さらに，固有名詞 (例：えみり，パンキチ，銀河富士)，話し言葉特有の表現 (例：そっかー，やったー，ぎょえー) やその他 IPA 辞書に含まれない語 (例：他人事，一発芸，特上) 計 1,670 語についてユーザー辞書を作成し，IPA 辞書に追加した．ユーザー辞書に追加した語の半数以上は感動詞であった．しかし，このようなユーザー辞書を作成しても，特に今回のように話し言葉を対象にした場合には，形態素解析において誤解析が生じやすい．そのため，本研究では解析結果を目視で確認し，できるだけその影響を避けるよう配慮した．

形態素解析により得られる品詞情報は 2 階層になっている．例えば,「挨拶」は「名詞：サ変接続」,「走る」は「動詞：自立」である．本研究では，これらの例の「名詞」,「動詞」にあたる部分を品詞 1,「サ変接続」,「自立」にあたる部分を品詞 2 と呼ぶ．本研究の品詞 1 の分類は，IPA 辞書に従い，名詞，動詞，形容詞，副詞，連体詞，接頭詞，感動詞，助詞，助動詞，フィラー，記号の 11 種である．

本研究の目的は小学生にとって必要な語彙の抽出にあるため，形態素解析の結果，フィラー，記号，名詞のうち数詞または固有名詞と判定された語は分析対象から除外した．さらに，分析対象となった単語に対して，以下 (1) から (5) の処理を施し，番組ごとの各語の出現頻度を表の形に整理した．

(1) 単一の番組にしか出現しない語の除外

(2) 表記の統一

- 動植物の名称は原則カタカナ表記
- 外来語語源はカタカナ表記
- 常用漢字外や当て字はひらがな表記
- 副詞は原則ひらがな表記 (ただし，漢字の意味が強く，一般に漢字で書かれるものについては一部漢字表記)
- 形容詞は原則ひらがな表記 (漢語表現から来たものは一部漢字表記)
- 動詞は原則漢字かな交じり表記 (補助的用法が多いものや多義で判別が難しいものはひらがな表記)
- 動詞の複合語において表記に複数の可能性があるときや第 2 項の意味が本義から離れているときはその部分をひらがな表記
- 感動詞のうち，外来語語源，音や動物の鳴き声はカタカナ表記，それら以外は原則ひらがな表記

(3) 表記上の変化や音声変化は「変化形」として扱い，対応する見出し語に集約

(4) 誤解析の可能性が高い語の確認作業および修正

(5) ひらがな表記の多義語をマーク

処理 (1) を行ったのは，単一の番組にしか現れない語は特殊なものである可能性があり，汎用性が高い語は複数の番組に現れると考えたためである．処理 (2) は，ルールに沿って表記を統一する作業である．処理 (3) は,「アイディア」と「アイデア」のような表記上のバリエーションや,「言う」と「ゆ

う」のような話し言葉特有のバリエーション,「じゃあ」,「じゃぁ」,「じゃー」のような作業者による表記上のバリエーションをまとめてカウントするためのものである．また，この処理では,「あはは」,「あははは」,「あはははは」のような繰り返しもバリエーションと見なしている．これらのバリエーションのうち著者らが代表的と判断したものを見出し語として選定し，それ以外を単語リストの「変化形」欄に記載した．出現頻度は見出し語および変化形の出現頻度の和とした．処理 (4) では，誤解析が疑われる単語について元の文に戻って確認し，必要に応じて出現頻度，品詞 1，品詞 2 を修正した．その結果，別の見出し語として分類されていた複数の語が 1 つの見出し語に集約されたケースもあった．最後の処理 (5) では，単語出現頻度リストに多義語の項目を設け,「おる」,「かける」などのひらがな表記の多義語にマークを付けた．本来であれば，これらの語も文脈上の意味によって分類し，それぞれの出現頻度を求めるべきであるが，この処理を自動的に行うことは困難であるため，本研究では多義語であることを利用者に示すにとどめた．

2.4 番組ごとの放送時間の差異の補正

上述の通り，番組ごとに d (=1 話あたりの放送時間 $\times$ 放送回数) の値が異なるため，この値が大きい番組，つまり，総放送時間が長い番組に現れやすい語の出現頻度は大きくなりやすい．この影響を抑えるため，前節にて求めた単語出現頻度にその番組の d と総放送時間 (=3,190 分) の比を乗じ，時間正規化単語出現頻度を求めた．

正確を期すために以下に数式を用いた説明を加える．番組 i ($i = 1, \ldots, 15$) における語 j の出現頻度を f_{ij} とし，番組 i の 1 話あたりの放送時間に放送回数を乗じた値を d_i，総放送時間を D とすると，番組 i における語 j の時間正規化単語出現頻度 f'_{ij} は以下の式で求められる．

$$f'_{ij} \quad = \quad f_{ij} \times \frac{d_i}{D} \tag{1}$$

このようにして得られた f'_{ij} の 15 番組分の和を求め，その値が大きい順に見出し語を並べかえ，各語の出現頻度を分析した．

3 結果

3.1 単語出現頻度の分析

書き起こしデータを形態素解析し，フィラー，記号，数詞，固有名詞と判定された単語を除外した異なり語数は 12,398 語，延べ語数は 244,848 語であった．さらに，上記の処理 (1)，(3)，(4) により，見出し語の異なり語数は 3,463 語，延べ語数は 228,012 語に減少した．これは特に処理 (1) の影響が大きいのだが，見出し語数は処理前より 70 %以上削減されたのに対して，延べ語数は 7%しか減少していない．このことは，1 番組にしか出現しない語の数は多いものの，そのほとんどは出現頻度が小さく，これらを除外しても出現頻度の分析に及ぼす影響は限られることを意味している．

最終的に得られた語彙について品詞 (品詞 1) ごとの異なり語数と延べ語数を表 2 に示す．3,463 語の見出し語のうち，最も数が多い品詞は名詞であった．見出し語全体の半数以上の 1,913 語 (55%) が

表 2: 品詞 1 の異なり語数と延べ語数 [語]．複数の品詞に判定された語を除く．

品詞 1	異なり語数	延べ語数
名詞	1,913	57,478
動詞	752	35,532
副詞	287	8,964
感動詞	198	12,940
形容詞	136	6219
助詞	79	71,922
接続詞	48	1,819
助動詞	19	28,062
連体詞	18	2,827
接頭詞	18	1,374

名詞であり，その延べ語数は 57,478 語 (25%) であった．次いで多かったのは動詞で 752 語 (22%)，延べ語数は 35,532 語 (16%) であった．

時間正規化単語出現頻度の和の上位 50 語を表 3 に示す．上位 50 単語の約半数，24 語を助詞または助動詞が占めている．また，上位の語およびその変化形には話し言葉で多用されるものが多い．例えば，終助詞の「よ (10 位)」,「ね (13 位)」,「な (21 位)」，感動詞の「あ (36 位)」,「はい (46 位)」,「うん (48 位)」などが挙げられる．変化形としては，例えば「だ (3 位)」の変化形が目立つが,「じゃろ」,「じゃん」,「だぁ」,「だー」,「だぁー」,「だーい」,「だい」,「だろぉ」など，話し言葉特有の変化である．

さらに，表 3 に示した 50 語が出現した番組数を調べたところ，50 語中 47 語は 15 番組全てにおいて用いられており，残りの 3 語も 14 番組で用いられていた．すなわち，上位に現れた語は，特定の番組で多用されたものではなく，汎用性の高いものであった．

時間正規化単語出現頻度のカバー率を図 1 に示す．カバー率は，上位 360 語で 80%に達し，上位 840 語で 90%になった．つまり，上位の 840 語とその変化形を聞き取ることができれば，低学年向けの教育番組の 9 割の語を把握できることが期待できる．言うまでもなく，文を構成する語の 9 割を知っていてもその文全体の意味を理解できるとは限らず，残りの 1 割の単語こそが学習内容の理解に不可欠なキーワードであるケースも多いことも考えられるが，1 つの目安となる数字といえよう．

時間正規化単語出現頻度は，単語出現頻度に各番組の放送時間の比がかけられているため，当該の語が実際に何回使われたかを把握しにくいという欠点がある．そこで，以下にいくつかの語の出現頻度を示す．1 位の「の」は 8,534 回，10 位の「よ」は 3,292 回，100 位の「お母さん」は 241 回，360 位 (カバー率 80%) の「春」は 67 回，840 位 (カバー率 90%) の「向かう」は 37 回，1000 位の「背中」は 29 回，2000 位の「手入れ」は 7 回，3000 位の「決定」は 2 回であった．1 位の「の」は延べ語数全体の 3.7%を占めており，約 30 語に 1 語の割合という高頻度で出現していたことになる．

本単語リストを旧日本語能力試験出題基準 (以下,「出題基準」) における最も易しいレベルである

表 3: 時間正規化単語出現頻度の上位 50 語 (次のページに続く)

No.	見出し語	変化形	品詞 1	品詞 2	多義語	出現頻度
1	の		助詞	格助詞/終助詞/連体化		638.473
2	て	てーっ	助詞	格助詞/接続助詞		622.911
3	だ	じゃろ/じゃん/だぁ だー/だぁー/だーい だい/だろぉ/だぁー だーい/だい/だろぉ	助動詞			573.879
4	た	たぁ/たー/たーっ/たり	助動詞			551.149
5	に		助詞	格助詞/副詞化		498.028
6	は		助詞	係助詞		496.908
7	が		助詞	格助詞/接続助詞		441.583
8	を		助詞	格助詞		385.870
9	ない	なーい/せん/ないっ なきゃ	助動詞 助動詞			265.519 265.519
10	よ	よぉ/よぉー/よーん	助詞	終助詞		261.958
11	する		動詞	自立		243.801
12	と		助詞	格助詞/接続助詞 副詞化/並立助詞		223.834
13	ね	ねぇ/ねえ/ねー/ねえー ねん	助詞 助詞	終助詞 終助詞		223.235 223.235
14	か		助詞	副助詞/並立助詞/終助詞		214.412
15	ます	まーす	助動詞			211.635
16	ん		名詞	非自立		209.591
17	も		助詞	係助詞		184.514
18	で		助詞	終助詞/接続助詞		182.903
19	いる		動詞	自立	○	180.533
20	です	でーす	助動詞			156.544
21	な		助詞	終助詞		153.212
22	う		助動詞			122.343
23	から	からぁ	助詞	終助詞		118.270
24	なる		動詞	自立/非自立	○	103.197
25	何	なあに/なーに/なーん	名詞	代名詞		101.080
26	てる		助動詞			98.100
27	事		名詞	一般/接尾/非自立		92.074
28	これ	これっ	名詞	代名詞		88.113
29	お		接頭詞	名詞接続		82.558
30	できる		動詞	自立/非自立		79.870

No.	見出し語	変化形	品詞 1	品詞 2	多義語	出現頻度
31	この		連体詞			77.328
32	来る	きたぁ/こーい	動詞	自立/非自立		75.737
33	って	ってな	助詞	格助詞		71.803
34	そう	そ/そっ	副詞	一般		71.760
35	有る		動詞	自立		69.528
36	あ	あっ	感動詞			68.254
37	行く	いっく/ゆく	動詞	自立/非自立		66.914
38	どう		副詞	助詞類接続		66.321
39	みる		動詞	自立/非自立		64.053
40	いい	ええ	形容詞	自立		63.434
41	さん	さーん	名詞	接尾		61.942
42	じゃ		助動詞/助詞	接続助詞/副助詞		60.641
43	みんな	みんなー	名詞	代名詞		59.107
44	見る	見ん	動詞	自立		59.041
45	言う	ゆう	動詞	自立		55.384
46	はい	はぁい/はーい/はいっ	感動詞			53.525
47	やる		動詞	自立/非自立		53.506
48	うん	うんー	感動詞			51.605
49	ちゃ		助詞	接続助詞		48.787
50	もう		副詞	一般		48.039

4 級の単語リストと比較したところ，そのほとんどが本単語リストに含まれていた．しかし,「月曜」,「おとい」,「階段」,「午前」,「午後」,「曇り」など，基本的な語であるにもかかわらず本単語リストに含まれないものが存在し (ここに例示した語以外にも本単語リストに含まれない語は存在する)，その中には本研究で対象にした 15 番組に 1 度も出現していない語もあった．

その一方で，本単語リストには出題基準に含まれない語や上級とされている語も多く含まれていた．語の難易度レベルの判定には「リーディング・チュウ太」の語彙チェッカー[3]を用いた．上位 1,000 語のレベル判定を行った結果，級外あるいは N1 (上級) と判定された語を表 4 に示す．これらの中には,「幼虫」,「バッタ」,「ドングリ」,「砂鉄」などの理科で用いられる名詞も多いが,「葉っぱ」,「明かり」,「給食」,「作戦」,「ロボット」,「ホームページ」,「和尚」,「小僧」など日本人の小学生なら知っているはずの語が並んでいる．また，名詞以外では,「やってくる (動詞)」,「格好いい (形容詞)」,「仲良い (形容詞)」などがある．さらに，副詞の「なんだか」,「もしかして」,「本当は」など話し言葉で用いられることが多い表現や「もぐもぐ」,「ごしごし」,「わくわく」のような擬声語・擬態語なども含まれている，以上のような形で小学校生活において頻繁に用いられている語が抽出できたことは本研究の成果といえよう．

[3] https://chuta.cegloc.tsukuba.ac.jp/

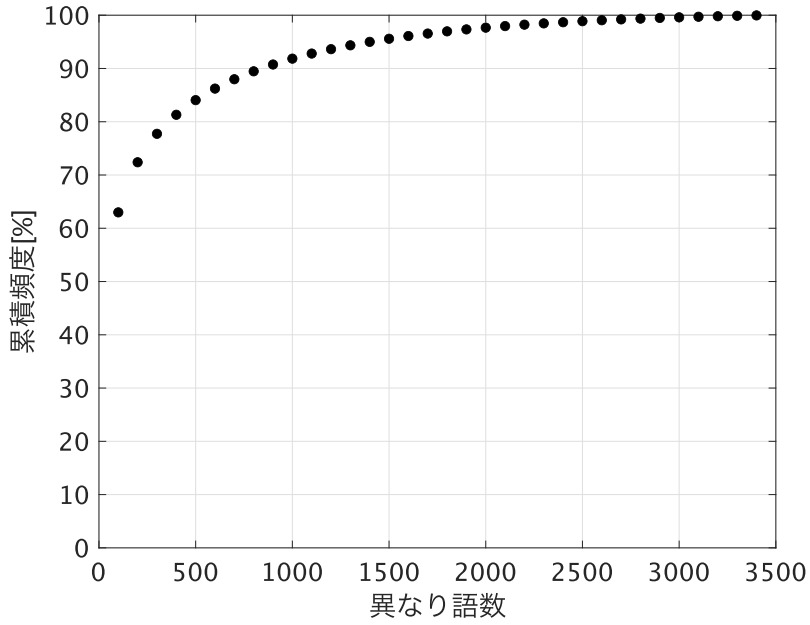

図 1: 時間正規化単語出現頻度のカバー率

3.2 単語リストから除外された語

2.3 節にて述べたように，本研究では単一の番組にしか出現しない語は除外した．しかし，このような語の中にも出現頻度の多いものが存在した．それらの大半は，科目特有の語や役割語 (キャラ語) であった．役割語とは，ステレオタイプな人物像やキャラクターに関係した言葉遣いであり，アニメの中の博士の一人称が「わし」であったり，語尾が「○○じゃ」であったりするのが代表的な例である [14]．以下では，単一の番組にしか出現しない語も含めた単語出現頻度ランキング 1 万 2 千語中の上位 1,000 位に現れた語について議論する．

第 1 の科目特有の単語としては，社会科の番組に現れた「工場」,「市」，算数の番組に現れた「分の」(「2 分の 1」などの「分の」) の 3 語があった．いずれもそれぞれの科目で学習に欠くことのできない基本的な語である．しかし，表 1 に示すように本研究では社会と算数はそれぞれ 1 番組しか対象にできなかったため，残念ながらこれらの語がリストから除外されてしまった．これはより多くのデータが必要なことを示している．

第 2 の役割語は一部の番組で頻出していた．特定のキャラクターによる「○○ギャオ」,「○○ッピ」,「○○ねーん」などの語尾や,「ギャイ」,「おっけ (OK)」,「ホーレイ」などの感動詞であり，しかも多用されていた．また，付言すれば，これらの役割語が頻出していた番組は，自然な会話というよりも子供たちの興味を引き付けるためのキャラクターによる会話で構成されていた．こうした特徴は日本語を母語としない生徒にとっては理解の妨げになる可能性があり，視聴させる番組の選択には十分な配慮が必要であることを示唆している．

表 4: 時間正規化単語出現頻度上位 1,000 語の中で旧日本語能力試験出題基準の級外または 1 級 (N1) と判定された語

品詞	単語
名詞	葉っぱ，お前，おいら，あたし，幼虫，ドリル，産む，俺，雌，雄，人達，バッタ，蝶，明かり，モンシロチョウ，奴，ホウセンカ，お宝，河童，ウサギ，お家，本日，オクラ，僕ら，あんた，カマキリ，とこ，お湯，お化け，お客，ラッパ，変身，アリ，おら，カブトムシ，わし，ヒマワリ，子象，挑戦，昆虫，正解，尻尾，茎，銀河，トノサマバッタ，さなぎ，美味い，タイム，木の実，ホームページ，モール，カエル，和尚，どんぐり，クリップ，キャベツ，リス，鶴，作戦，勝手，タヌキ，折り紙，同士，草むら，セミ，砂鉄，ロボット，給食，ナイス，仕業，通り道，粘土，隊員
動詞	やってくる，知れる，気が付く
副詞	なんだか，もぐもぐ，もしかして，ごしごし，わくわく，本当は
接続詞	まずは
連体詞	そういう
形容詞	格好いい，でかい，気持ちいい
感動詞	こりゃ，オッケー，やあ，ふーん，ガーン，ようこそ
助詞	によって

4 考察

本研究では，小学校低学年の教育現場における話し言葉に現れる語の特性の一端を明らかにすることを目的とし，NHK E テレの低学年向け教育番組の音声にて用いられている語の出現頻度を調査した．8 科目 15 番組，53 時間 10 分の動画を対象にして分析を行い，2 つ以上の番組に出現した 3,463 語の出現頻度を求めた．

3.1 節に示したように，本研究で得られた単語リスト全体には話し言葉の特徴が現れている．まず，品詞全体に占める感動詞の比率が極めて高い．先行研究において，小学 1 年生から 4 年生の児童の作文では感動詞の延べ語数の比率が 0.5% [15] であるのに対して，本研究の結果得られた感動詞については，延べ語数の比率が 5.7%と 10 倍以上である．延べ語数の 5.7%という比率は，本研究のデータでは，およそ 18 語に 1 語の割合で感動詞が出現していたことを意味している．加えて，前節に示したように，感動詞以外でも話し言葉特有の表現や音声変化が多数現れている．これらの点は，既存の書き言葉コーパスに基づく単語リストと大きく異なる特徴である．

また，本単語リストには出題基準の 4 級の語の大半が含まれる一方で，頻度上位の語であっても級外となる語があった．これらの語には理科の用語が多く，日本語を母語としない生徒の指導において，この点に配慮する必要があることを示唆している．さらに，話し言葉特有の表現や擬音語・擬態語等も多く含まれていたことから，本単語リストを用いることによって，出題基準とは異なる視点で小学

校低学年の生徒に必要な語を把握できるものと考えられる．

一方，本研究の限界としては，以下の点が挙げられる．

(1) 分析対象のデータの量が不十分であること．

(2) 形態素解析における誤りを一部含むこと．

(1) の問題は対象とする番組を増やすことによって解決でき，これによって 3.2 節に示した特定の科目に出現する語が取りこぼされてしまうという問題も解消される．その上で課題となるのは，(2) の形態素解析の精度である．本研究では，形態素解析結果の確認と誤解析の修正に相当な時間を要した．話し言葉も高精度に解析できる形態素解析技術の登場に期待したい．

5 おわりに

本研究では，日本語を母語としない児童の学習・教育支援への応用を目的として，NHK E テレの小学校低学年向けの教育番組における使用語彙を調査した．15 番組の全 312 回，53 時間 10 分の音声を書き起こし，複数の番組で用いられていた語を出現頻度順にリスト化した．得られた単語リストには小学校生活で不可欠な語も多く含まれており，感動詞や音変化が多い等，話し言葉特有の特徴が現れていた．さらに，単語リストから除外された語の分析から，一部の番組で役割語が多用されていることも明らかとなった．テレビの教育番組は，映像を伴っているため実物を見せることができない場合の補助具としての役割は大きいものの，番組の選択にあたっては日本語を母語としない児童への配慮が必要であることも明らかになった．

本研究の結果得られた単語出現頻度および時間正規化単語出現頻度のデータは第 1 著者の Web ページにて公開する予定である．ただし，著作権の問題があるため書き起こしたテキストデータそのものの公開は行わない．今後はこのデータをどのような形で実際の学習支援につなげていくのかを考えていく必要がある [16], [17].

謝辞

本研究の一部は 2020 年度 公益財団法人 日教弘本部奨励金の支援によるものである．

参考文献

[1] 文部科学省総合教育政策局国際教育課, “外国人児童生徒等教育の現状,” 文部科学省, https://www.mext.go.jp/content/20210526-mxt_kyokoku-000015284_03.pdf, 2021-05. (参照 2021-12-15).

[2] 工藤真由美, 児童生徒に対する日本語教育のための基本語彙調査. ひつじ書房, 1995.

[3] 国立国語研究所, 教育基本語彙の基礎的研究 増補改定版. 明治書院, 2009.

[4] 仁科喜久子, 楊接期, 小島聡, 赤堀侃司, “オンライン科学技術日本語学習システム構築のためのテキスト解析 (2): 論文の形態と語彙の特徴を中心に,” 日本語教育方法研究会誌, vol. 5, no. 1, pp. 34–35, 1998.

[5] 野村愛, 川村よし子, “介護福祉士候補者の自立学習支援のための語彙リスト作成,” 日本語教育方法研究会誌, vol. 18, no. 1, pp. 14–15, 2011.

[6] 安藤句美子, “小・中学校教科書の語彙分析: 連語の観点から,” 日本語教育方法研究会誌, vol. 23, no. 1, pp. 68–69, 2016.

[7] 本田ゆかり, “コーパスに基づく「読解基本語 1 万語」の選定,” 日本語教育, vol. 172, pp. 118–133, 2019.

[8] 本田ゆかり, “「初級日本語教科書共通語彙リスト」の開発,” 日本語教育方法研究会誌, vol. 25, no. 2, pp. 130–131, 2019.

[9] 李在鎬, “BCCWJ に含まれる学校教科書コーパスの計量的分析: 日本語教育のためのリーダビリティと語彙レベルの分布を中心に,” 計量国語学, vol. 32, no. 3, pp. 147–162, 2019.

[10] 田中祐輔, “COSMOS: 帰国・外国人児童のための JSL 国語教科書語彙シラバスデータベース,” 計量国語学, vol. 32, no. 5, pp. 277–287, 2020.

[11] 獅々見真由香, “日本語の会話におけるオノマトペの基本語彙選定: 「BTS による多言語話し言葉コーパス」と「BTSJ による日本語話し言葉コーパス」,” 日本語教育, vol. 165, pp. 73–88, 2016.

[12] 淺井優介, 北村達也, 川村よし子, “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,” 甲南大学紀要知能情報学編, vol. 13, no. 1, pp. 67–75, 2019.

[13] 岡 照晃, “言語研究のための電子化辞書,” in コーパスと辞書, pp. 1–28. 朝倉書店, 2019.

[14] 金水 敏, ヴァーチャル日本語 役割語の謎. 岩波書店, 2003.

[15] 茂呂雄二, 村石昭三, “児童の作文使用語彙 (4): 小学校中学年児童の使用語彙,” 日本教育心理学会第 27 回総会発表論文集, pp. 258–259, 1985.

[16] 中川健司, “専門日本語の語彙研究を学習支援につなげていくためには何が必要か: 介護用語学習ウェブサイト開発の事例を基に,” 専門日本語教育研究, vol. 19, pp. 11–18, 2017.

[17] 松下達彦, “語彙リストの利用法: コーパス分析に基づく語彙研究は何を目指すべきか,” 専門日本語教育研究, vol. 19, pp. 19–24, 2017.