絵本の言語分析：
ブックスタート・セカンドブック事業の対象書籍の比較を通して

近藤 可奈子[1] 内田 諭[2]
[1]九州大学共創学部 [2]九州大学大学院言語文化研究院
kondo.kanako.545@s.kyushu-u.ac.jp, uchida@flc.kyushu-u.ac.jp

概要

本研究の目的は，幼児のいる家庭での読書習慣の形成を目的としたブックスタート事業とセカンドブック事業で配布対象となっている絵本における出現語彙の違いを明らかにすることである．ひらがなテキストの形態素解析として，精度比較の結果，GiNZA (ver. 5.1.0)を用いることで，スペースありの絵本の原文テキストについてカタカナ語の前処理をすることで, 単語区切りで 96.1%, 品詞付与で 94.0%の精度が見込まれるため，これを利用することとした．語彙分析の結果，ブックスタート事業の対象絵本では，身体的な表現や具体性の高い表現が多いことが明らかとなり，セカンドブック事業の対象絵本では，思考に関わる表現や時間把握などに関わるより認知的に高度な表現が多く出現することが明らかとなった．

1 はじめに

OECD（経済協力開発機構）が 2018 年に実施した PISA（国際学習到達度調査）において，日本の読解力は 15 位となり, 前調査年度の 8 位から大幅に下落した．日本児童の読解力低下は国内でも問題となっており，その原因の一つは読書量の減少だとする主張もあり，2018 年度の調査では，読書習慣のある生徒の方が読解力を測る問題における平均点が高いことも明らかにされた（『産経新聞』2019 年 12 月 3 日の記事より）．子どもの読書離れは以前から問題視されており，その懸念を背景に 2001 年「子どもの読書活動の推進に関する法律」が制定され，基本理念や国・地方公共団体の責務が明示された．

このような状況を背景に，幼児のいる家庭での読書習慣を促進することを目的として絵本を対象者に配布する「ブックスタート事業」や「セカンドブック事業」が各自治体で開始された．これらの事業は地方自治体の単位で行われ，選定される絵本は地域によって異なる．そのため，どのような絵本がどのような基準で選ばれているかは明らかではない．そこで本研究では，これらの事業の配布対象となっている書籍のリストを収集し，語彙分析をすることでその特徴を示す．これにより，特にマスターリストが存在しないセカンドブック事業（後述）における絵本の選定に貢献することを目指す．

自然言語処理分野における絵本の分析は，ひらがなの形態素解析精度の向上の研究[1, 2]や，自然言語処理技術を応用した絵本推奨システムの開発[3]，対象年齢の推定[4]などが行われている．本稿では特にひらがなの形態素解析について最新の日本語 NLP ライブラリである GiNZA[cf. 5]と，MeCab ベースの Python ライブラリである janome を用いて精度比較を実施する．

2 絵本普及のための事業の概要

2.1 ブックスタート事業

子どもの読書習慣向上を目指した活動は数多く実施されているが，主に各自治体や公共図書館が主体となって実施する「ブックスタート事業」もその一つである．ブックスタートとは，各自治体において乳幼児とその保護者を対象に絵本や読み聞かせ体験などをプレゼントする事業で, 1992 年に Share books with your baby!というキャッチフレーズのもと, イギリス・バーミンガム市で始まった．日本では「子ども読書年」である 2000 年にその活動が紹介されたのをきっかけに，2001 年に本格的な実施へと踏み出した．現在は NPO ブックスタートの支援により，全国 1,087（2021 年 12 月 31 日現在）の市区町村自治体の公的事業として実施されている[i]．森ら（2011）は，

[i] https://www.bookstart.or.jp/coverage/

7 年間ブックスタート事業に取り組んでいる自治体と未だ取り組んでいない自治体において，小学 1 年生の子どもを持つ保護者を対象にアンケート調査を行った．当調査により，「子どもの読書習慣が高まる，子どものさまざまな本への興味が高まる，（中略）保護者の図書館利用頻度が高まる，（中略）保護者による読み聞かせが多くなる」等の結果が得られ，ブックスタート事業は子どもの読書習慣や保護者の行動へ影響を与えることが示された[6].

2.2 セカンドブック事業

ブックスタートのフォローアップ事業として行われているのが「セカンドブック事業」である．対象年齢が 0 歳児とされるブックスタートに対し，セカンドブックは大多数の自治体で 3 歳前後もしくは小学 1 年生とされる．前者は先述の通り 1,000 以上の自治体が実施している一方で，後者は 176 の自治体でしか行われていないのが現状である（NPO ブックスタートの資料より）．第一の要因は，セカンドブックの実施にかかる絵本の費用や事業運営を行うためのスタッフの人件費等であると考えられる. また，ブックスタートに関しては，NPO ブックスタートが三年毎に開く「絵本選考会議」において選ばれた 30 冊の絵本が対象リストとして公開されるが，セカンドブックに同様のものは存在せず，事業の普及の障壁になっていると考えられる．

3 ひらがな文章の形態素解析

3.1 絵本の形態素解析に関する先行研究

絵本のテキストの大部分はひらがなで構成され，また単語や文節区切りで空白を含むという点で特殊である．藤田ら(2014)は絵本と一般コーパスを比較した場合，前者では 74.5%がひらがな，7.8%が空白であったのに対し，一般コーパスである BCCWJ ではそれぞれ 49.2%，1.3%であることを示している[2]. これまで言語モデルを用いた形態素解析の手法[1]や空白区切りや句読点，ひらがな化したデータ，さらに絵本そのもののデータを学習に使う手法などが提案され[2]，精度の向上を報告している．さらに，絵本の分析の課題として，表記の揺れ[3]，固有名詞の扱い[2]などの問題が指摘され，数量的な分析の基礎となる形態素解析での課題が多い．

本稿は，形態素解析の精度向上を目的としたものではなく，ブックスタート事業およびセカンドブック事業で用いられている絵本の言語的特徴を明らかにすることを目的としているため，これらの点には深く立ち入らず，既存の形態素解析プログラムの精度を比較した上で，最もパフォーマンスが良いものを利用して分析を進める．

3.2 分析精度の比較

既存の形態素解析プログラムで精度が高いものを利用するため，Python の NLP ライブラリである GiNZA (ver. 5.1.0)と janome(ver. 0.4.1)を用いて精度比較を実施する．形態素解析の実行環境は Google Colaboratory を利用する．GiNZA のフレームワークとして spaCy (ver. 3.2.1)を利用し，辞書として ja-ginza-electra (GiNZA ver. 5.1.0)を指定した．また，janome は mecab-ipadic-2.7.0 を内包している．

3.2.1 評価データ・評価方法

評価データは後述する自作の絵本コーパスからランダムに 10 冊（コーパス全体の約 10%にあたる 8,515 文字分のデータ）を選定したものを利用し，絵本のテキストに含まれる空白の有無（例：「11 ぴきの　のらねこが　いました」vs「11 ぴきののらねこがいました」）の影響についても調査した．「単語区切り」と「品詞付与」について，GiNZA の分析結果を基本として人手で確認した（判断に迷う部分は筆者らで合議して決定した）．ただし，形態素解析器のポリシーによる違いによって生じる差異については吸収できる形でアノテーションした．例えば，「おさとう」は GiNZA では「お（接頭辞）＋さとう（名詞）」，janome では「お（接頭詞）＋さとう（名詞）」となり「お」の品詞が異なるが，どちらも正答として扱った．また，「とびかかろう」は GiNZA では「とびかかろう（動詞）」，janome では「とびかかろ（動詞）＋う（助動詞）」となり，単語区切りが異なるが，これらはそれぞれのプログラムで一貫して用いられる区切りであるため，どちらの場合も正答として扱った．なお，GiNZA の出力には pos と tag があるが，品詞の照合に際してはより詳細な情報が得られる tag の最初の区分を利用した．例えば，「名詞-普通名詞-一般」が解析結果であった場合，最初の区分である「名詞」を照合対象とした．

3.2.2 評価結果

スペースあり（絵本の原文のまま）とスペースなし（原文データからスペースを一括削除）のそれぞれについて，GiNZA, janome の精度評価結果を表 1 に示す.

表 1 形態素解析の精度比較

	スペースあり	スペースなし
GiNZA_単語区切り	**95.0**	94.5
janome_単語区切り	85.6	80.6
GiNZA_品詞付与	**93.3**	92.8
janome_品詞付与	82.1	76.8

検証の結果，単語区切り・品詞付与どちらの場合も GiNZA を用いて絵本の原文のままスペースありで形態素解析を実施したものが最も精度が高い結果となった[ii]．GiNZA の場合，スペースなしの場合でもスペースありの場合と遜色のない結果となったことは特筆に値するだろう．なお，spaCy の辞書で ja-core-news-sm (3.2.0)を指定した場合，複合動詞（例：かけ＋こむ）や複合語（例：お＋かあさん）の区切りが細かくなるという点を除いて結果は同一であったが，語彙分析に際しては複合形式を単一単位として扱うほうが望ましいため，本稿の分析では辞書として ja-ginza-electra を利用する.

3.3 ひらがな解析の精度向上の手法提案

絵本のテキストの分析精度を上げる手法として，藤田ら(2014）は固有名詞（例：「ぐり」，「ぐら」など）を辞書に登録することで精度の向上が図れると提案している．これに加えて，本稿では「リュックサック」，「ボール」などのカタカナ語を，事前に辞書を用いて，ひらがなからカタカナに変換する前処理を提案する．評価データを観察すると，このようなカタカナ語の誤分析が多い．例えば，「べっどのうえでかすてらをたべた」という文は，「べっ（名詞） どの（連体詞） うえ（名詞） でかす（動詞） てら（名詞） を（助詞） たべ（動詞） た（助動詞）」のように誤分析されるが，「ベッドのうえでカステラをたべた」とカタカナにすると「ベッド（名詞） の（助詞） うえ（名詞） で（助詞） カステラ（名詞） を（助詞） たべ（動詞） た（助動詞）」と正しく分析される．本稿のデータではカタカナ語の変換を実施した場合，GiNZA を用いると単語区切りで 96.1%，品詞付与で 94.0%まで精度が向上した.

[ii] 藤田ら(2014)では MeCab のデフォルト設定での品詞推定精度として 83.2%であったことを報告しており，janome の結果(82.1%)はこれとほぼ一致する[2].

4 特徴語の分析

4.1 データセット

ブックスタート事業，セカンドブック事業の対象絵本の特徴を明らかにするために，全国で配布されている絵本のリストをインターネットおよび図書館等への問い合わせによって入手し，採用件数を集計した．最終的に，ブックスタート事業を実施している 24 自治体，セカンドブック事業を実施している 30 自治体のリストを集約し，いずれも 3 地域以上で掲載されているものを分析の対象とした. その結果，「ブックスタート絵本コーパス」として 25 冊（4,345 文字），「セカンドブック絵本コーパス」として 29 冊（80,884 文字）が対象となった．これらの絵本のテキスト情報を手入力で入力し，テキストファイルとして保存した．3 節での実験結果を受けて，絵本のスペースありデータ（原文のまま）を GiNZA で解析し，品詞ごとに出現語彙を集計した.

4.2 高頻度語・特徴語の抽出

平ら(2012)は絵本の高頻度語を分析し，ランダムに売れ筋の絵本を選択した場合のカバー率（20 冊程度で上位 1,000 語，150 冊程度で上位 2,000 語をカバー）や，ウェブ日誌法のテキストデータと比較して幼児語（まんま，ねんねなど）の出現率が低いことなどを示している[7]．一方，対象年齢別の語彙の違いやその特徴までは論じていない．以下では，0 歳児向けの絵本が多く含まれるブックスタート絵本コーパスと 3 歳～6 歳児向けの絵本が多く含まれるセカンドブック絵本コーパスの高頻度語および特徴語を抽出し，その違いについて考察する．高頻度語はブックスタート絵本コーパスでは頻度 5 以上（形容詞は 3 以上），セカンドブック絵本コーパスでは頻度 30 以上（形容詞は頻度 20 以上）のものを抜き出した（ただし，誤解析と思われるものは省く）．また，特徴語の抽出にあたっては Python の SciPy (ver.

1.4.1)を利用してχ二乗検定を実施し，p<=0.05 のものを特徴語として認定した（*で記す）．

4.3 ブックスタート絵本コーパスの特徴語

ブックスタート絵本コーパスで出現する名詞の高頻度語および特徴語は以下の通りである．

【名詞】おかあさん(24)*，あかちゃん(21)*，こども(11)*，おく(11)*，くつ(10)，かお(9)*，ぱん(9)*，こんど(8)*，なか(7)，おとうさん(7)，ねこ(6)，おなか(6)*，みち(6)*，たまご(6)*，かくれんぼ(6)，いぬ(5)，みんな(5)，うさぎ(5)，ぞう(5)，ふわふわ(5)，てんてん(5)，ぱっぱ(5)

「おかあさん」「あかちゃん」「おとうさん」など，乳児が日常生活において頻繁に語りかけられると考えられる名詞や「かお」「おなか」など身体部を指す名詞が上位にある．小椋・綿巻（1999）は，日本の幼児の早期表出語彙 50 語において，普通名詞では身体部分を示す語が最多であるという結果を示しており，絵本での高頻度語はこの観察と符合する[8]．

【動詞】いる(28)*，くる(17)*，でる(7)*，はこぶ(7)*，こぼす(7)*，よぶ(6)*，くっつく(6)*，にげる(6)*，のせる(5)*，あける(5)*

また，動詞では「くる」「でる」「はこぶ」など移動を表すものや具体的な動作を示すものが多い．

【形容詞】いい(17)*, おいしい(8)*, いろいろ(4)*, おおきい(3), きれい(3)*

形容詞については，「いい」「おいしい」「きれい」など肯定的な語が上位に複数含まれる．

【副詞】きゅっ(12)*，よく(10)*，どうぞ(10)*，ぎゅう(8)*，ぴょーん(8)*，わん(6)*，また(5)，びり(5)*

副詞として分析された語をみると，「きゅっ」「ぎゅう」「ぴょーん」「わん」など擬音語が頻繁に見られる．

4.4 セカンドブック絵本コーパスの特徴語

次にセカンドブック絵本コーパスの高頻度語および特徴語について品詞別に考察する．

【名詞】こと(155)*，ライオン(81)*，かえる(79)*，ところ(66)*，うち(63)*，もの(62)*，さま(55)，とき(53)，中(52)，わに(51)，こえ(50)，おうち(48)，ねこ(48)，川(48)，オオカミ(45)，しっぽ(42)，はなし(42)，め(40)，き(39)，どうぶつ(39)，なか(39)，ひと(39)，まえ(39)，りゅう(39)，ベッド(39)，つぎ(38)，ホネ(38)，とら(37)，もも(35)，こども(31)，しま(31)

名詞では「オオカミ」「りゅう」「とら」など，一般的な日常生活においては身近な存在でない動物を表す名詞が上位に多く含まれる．また，「もの」「とき」など抽象的な名詞も上位にある．絵本の対象年齢が上がるため，子どもたちが想像できる範囲がより拡大されることが理由であると考えられる．また，「中」，「川」など簡単な漢字の単語も含まれている．

【動詞】いう(401)*, いる(363), くる(257), する(256), なる(230), ある(180), いく(140), みる(131), やる(107), しまう(99), おもう(73), たべる(66), くれる(56), きく(53), わかる(52), つく(47), あるく(40), だす(40), かける(38), でる(38), はじめる(38), みえる（34），できる(33)，まつ(32)，もつ(32)

「いる」「なる」「ある」など抽象的な状態動詞や「みる」「おもう」などの感覚・思考動詞が頻繁に見られることがわかる．O'Grady(2005)は，子どもの発話において最も頻繁に出現するのは「走る」「遊ぶ」「乗る」などの行為に関する動詞で，「思う」「信じる」などの感覚的な動詞は子どもにとって習得が難しいものとして位置づけており[9]，これらの絵本は子どもの認知的な成長につながるよいインプットとなることが示唆される．

【形容詞】いい(80)，ない(67)，ちいさい(66)，おおきな(50)，おおきい(29), 大きい (25), くろい(24), ながい(23), うまい(22), 大きな(20)，すごい(19)，はやい(18)

形容詞は特徴語と認定されるものはなかったが，ブックスタート絵本コーパスと比較して，次元や色彩，味覚，速度など様々な属性を表す語が観察された．

【副詞】もう(82)*，そう(59)*，どう(50)，まだ(36)，とても(34)，みんな(33)，また(30)，すこし(29)，こう(25)，ずっと(24)

副詞では「とても」「すこし」など程度を表す語や「もう」「ずっと」など時の流れを表す語など，より認知的に高度な副詞が多く見られた．

5 まとめ

本論文では自作したブックスタート絵本コーパスとセカンドブック絵本コーパスについて GiNZA を用いて語彙分析を実施し，それぞれの特徴を明らかにした．前者には身体的・具体的な表現が多い一方，後者には抽象的でより認知的に高度な表現が多いことが明らかとなった．比較的小規模なデータでの検証ではあるが，対象年齢別の特徴が明確に現れており，セカンドブック事業のための絵本リスト作成に示唆を与えるものであると考える．特に，心理・思考の描写のある絵本や時間把握能力が必要となる絵本が適していることが示唆される．

謝辞

本研究の成果の一部はJSPS科研費JP 18H00693の助成を受けたものです.

参考文献

1. 工藤拓, 市川宙, Talbot, D., 賀沢秀人(2012)「Web上のひらがな交じり文に頑健な形態素解析」『言語処理学会第18回年次大会発表論文集』1272-1275.
2. 藤田早苗, 平博順, 小林哲生, 田中貴秋(2014)「絵本のテキストを対象とした形態素解析」『自然言語処理』21(3), 515-539.
3. 藤田早苗, 服部正嗣, 小林哲生, 奥村優子, 青山一生(2017)「絵本検索システム「ぴたりえ」～子どもにぴったりの絵本を見つけます～」『自然言語処理』24(1), 49-73.
4. 藤田早苗, 小林哲生, 平博順, 南泰浩, 田中貴秋(2014)「絵本を基にした対象年齢推定方法の検討」『人工知能学会全国大会論文集第 28 回全国大会』3D4-4.
5. 松田寛(2020)「GiNZA-Universal Dependencies による実用的日本語解析」『自然言語処理』27(3), 695-701.
6. 森俊之, 谷出千代子, 乙部貴幸, 竹内惠子, 高谷理恵子, 中井昭夫(2011)「ブックスタート経験の有無が子どもの生活習慣や読書環境等に及ぼす影響」『仁愛大学研究紀要 人間学部篇』10, 61-67.
7. 平博順, 藤田早苗, 小林哲生 (2012)「絵本テキストにおける高頻度語彙の分析」『情報処理学会関西支部支部大会』F-103.
8. 小椋たみ子, 綿巻徹(1999)「早期表出・理解語彙の日米比較」『日本教育心理学会第 41 回総会発表論文集』141.
9. O'Grady, W. (2005) How children learn language. Cambridge University Press.（内田聖二監訳(2008)『子どもとことばの出会い:言語獲得入門』研究社）