英語絵本コーパスの構築と教育利用を目指した検討

藤田 早苗　服部 正嗣　小林 哲生　奥村 優子　永田 昌明
NTT コミュニケーション科学基礎研究所
{sanae.fujita.zc,takashi.hattori.cp,tessei.kobayashi.ga,yuuko.okumura.sf,masaaki.nagata.et}@hco.ntt.co.jp

1　はじめに

近年，日本の英語教育改革が進められてきている．2020 年度からは，小学校 3 年生から「外国語活動（英語）」が必須化，小学校 5 年生からは教科となることが決定しており，ますます低年齢から英語や外国文化に触れ，英語力を伸ばすことが要求されている．

絵本は，少なくとも母国語を学ぶ時には言語発達に貢献すると考えられるが [5, 8, 13]，その理由として，日常の会話ではほとんど出現しない語やフレーズが絵本には多数含まれていることや [6, 11, 16]，現在の物やイベントに縛られないこと [10] など，多くの点が挙げられている．

無論，これらは母国語として学ぶ時の効果を示した研究だが，特に小学校という低年齢から外国語を学ぶ場合，共通する効果がある可能性がある．一方で，外国語教育では，母国語の場合とは異なる要素も求められるだろう．特に，学校教育に沿った利用を考えると，子どもの英語力にあった絵本を選べること，教育課程にあった絵本を選べることが重要だろう．そこで，英語絵本の特徴を分析し，教育利用への可能性や活用方法を検討する．

本稿ではまず，NTT で構築している英語絵本のコーパスを紹介する (2 章)．また，英語の難易度推定方法の先行研究と英語絵本コーパスへの適用結果を述べる (3 章)．次に，教育利用を目的として開発された語彙表を紹介し，英語絵本コーパスとの対応付けと分析を行う (4 章)．最後にまとめと今後の課題について述べる (5 章)．

2　NTT 絵本・児童書コーパス

NTT ではこれまで，日本語の絵本のコーパスを構築してきており (NTT 絵本・児童書コーパス)，図書館や幼稚園などでの絵本の推薦 [14] や，語彙発達に遅れの見られる子どもの訓練用絵本の推薦 [18]，言語発達との関係調査 [7, 17] などに利用してきている．

英語の絵本コーパスも，日本語と同様に構築中である．取得している情報は日本語とほぼ共通だが，英語版では出版国 (US, UK など) の情報を追加している．

英語絵本の選定規準としては，次の 3 点があげられる．(1) 多くの幼児に読まれていると考えられる (ニューヨーク・タイムズベストセラーや，出版社ごとのベストセラー) や，著名な賞の受賞作（アンデルセン賞，コルデコット賞，ニューベリー賞，チルドレンズ・ブック賞，ボローニャラガッツィ賞など）．(2) 同じ内容の絵本が日本語でも読める (翻訳本が出版されている）．(3) テキストの難易度が明確に示されている (I Can Read! シリーズ[1], *Harper Collins*. 以下，ICR)．

英語の絵本コーパスのサイズを，表 1 に示す[2]．すでに約 145 万語からなり，英語絵本のコーパスとしては，著者の知る限り最大である．

ここで，選定理由 (3) について補足しておくと，出版社によって一貫した難易度や対象年齢が付与されている絵本は多くはない．出版社によって対象年齢が付与されている場合でも，書き方はバラエティに富んでいる．表 1 の絵本の内，出版社によって対象年齢かレベルが付与されている絵本は，ICR 以外に 795 冊あったが，その表現には 100 以上のバリエーションがあった（例えば，“Ages 4 to 8”,“Ages 4 up”, “3+”, “Grade 1-3”, “lower age group” など)．そこで，難易度推定の評価への利用を検討するため，難易度が明示された ICR を選定した．ICR のコーパス化が進めば難易度推定の定量評価に利用する予定である．

3　難易度推定に関する既存研究

英語力にあった絵本を選ぶためには，絵本のテキストの難易度を推定する必要がある．本章では，既存の難易度推定方法を簡単に紹介する．

[1] www.icanread.com
[2] 2018 年 12 月現在のサイズ．含まれるタイトルは http://www.kecl.ntt.co.jp/icl/lirg/members/sanae/ehon/ehon-list-english.1717.html= で閲覧可能．

表 1: NTT 絵本・児童書コーパスのサイズ

分類		冊数	話数	ページ数		行数		語数 (Type)		のべ語数 (Token)	
				平均	計	平均	計	平均	計	平均	計
日英	E2J	209	228	42	8,778	64.63	14,736	215.63	10,031	732.64	167,043
翻訳	J2E	7	7	34.43	241	67.29	471	252.71	1,052	684	4,788
あり	X2JE	5	5	32.2	161	68.8	344	277.4	875	836.6	4,183
ICR	0	111	112	29.86	3,315	46.63	5,223	90.46	2,391	278.52	31,194
	1	155	168	37.61	5,829	86.73	14,571	212.66	5,695	702.92	118,090
	2	69	92	55.83	3,852	138.05	12,701	267.32	4,867	1223.03	112,519
	3	6	9	65.67	394	154.33	1,389	283.33	1,293	1275.56	11,480
	4	2	2	50	100	233	466	453	714	2447	4,894
翻訳なし・ICR 以外		1,166	1,424	38.25	44,597	65.13	92,739	224.08	36,392	701.85	999,431
全体		1,717	2,047	38.88	67,267	69.68	142,640	217.55	41,176	710.12	1,453,622

※「翻訳あり」とは，NTT 絵本・児童書コーパスに日英両方があることを示す．実際には，より多くの翻訳された絵本が存在する可能性がある．また，I Can Read! シリーズ (ICR) で，翻訳もある絵本が 13 冊存在するため，両者で重複がある．

英語を対象とした難易度推定方法の研究は非常に多くなされてきており，難易度を求める公式は 200 以上提案されている [2]．

難易度推定問題は，回帰や分類，ランキング問題として捉えられる．例えば，難易度の値を求める回帰式としては，Flesch-Kincaid score [3, 4] や New Dale-Chall Readability Formula [1] などが提案されている．前者は語の音節数の平均値と文の平均単語数を変数として利用し，後者はよく知られた 3000 語のリストを用意し，そこに含まれない語の出現割合や，1 文中の平均語数を変数として利用している．これらの方法は，簡単で使い易いため広く利用されている．

難易度クラスに分類する手法としては，統語的特徴量と言語モデルから得られる特徴量の両方を利用して SVM で分類する手法 [9] などが提案されている．

難易度推定をランキング (ソート) 問題として捉える研究では，田中ら [12] が，全対象テキストで一対比較を繰り返して難易度順にソートする手法を提案している．新しいテキストに対しても，難易度順にソートするためにはソート済テキストとの一対比較を行う．しかし，全テキストのソートには時間がかかることと，同じ特徴量で統一的に比較できないという問題点がある．

このように非常に多くの難易度推定方法が提案されてきているが，絵本を対象とした研究は多くない．藤田ら [15] は，日本語の絵本を対象として，言語モデルや平均文長などの特徴量を用いたランキング学習による手法を提案している．絵本のように語彙数の少ないテキストに対してもロバストな難易度推定ができるが，規準として学習に利用するテキストが必要である．

そこで，規準とする学習データを必要としない既存手法をいくつか英語絵本コーパスに適用し，有効性を検証した．本稿では Flesch-Kincaid と Dale-Chall の結果を紹介する．

Flesch-Kincaid の場合，推定結果のスコアが大きい方がより簡単であり，スコア $90-100$ 程度で 5 年生程度とされているが，絵本の 72.8%は 100 以上のスコアとなり，5 年生より簡単だと推定された．一方で，中高生や大学生以上レベルと判定された絵本は 2.5%だった．

Dale-Chall の場合，スコアが小さい方がより簡単であり，4.9 以下で 4 年生以下とされる．しかし，4.9 以下と推定された絵本は 2.7%のみであり，7.0-7.9 (9-10 年生．中高生レベル) が 36.7%と最も多かった．

このように，全体的に Flesch-Kincaid より Dale-Chall の方が難しいと推定される傾向が見られる．Dale-Chall の場合は，3000 語のリストに含まれない語が多ければより難しいと推定されるが，絵本の場合，リストには含まれないが難しくはない語も多いためだと考えられる．一方の Flesch-Kincaid は，単語リストは用いず，音節数や単語数を用いるため，絵本に対しては Dale-Chall よりも直感にあった結果となっていると考えられる．

ICR のコーパス化が進めば，ICR を用いた定量評価も行いたい．特に，日本語の絵本での難易度推定方法 [15] が英語絵本でも有効かどうかを検証したい．

4 学校教育での利用に向けた検討

CEFR 学校教育で用いることを考えると，教育過程に合った難易度推定や推薦ができることが重要である．文部科学省による「グローバル化に対応した英語

表 2: CEFR-J Wordlist (CWJ) のサイズと絵本コーパスとの対応結果

レベル	見出し語数	項目数（品詞別）	絵本に出現 No.	(%)	例	絵本に出現せず No.	(%)	例
A1	1,068	1,165	1,155	99.1	a (determiner)	10	0.9	CD (noun)
A2	1,359	1,416	1,331	94.0	ability (noun)	85	6.0	acceptable (adjective)
B1	2,358	2,451	2,059	84.0	abandon (verb)	392	16.0	abnormal (adjective)
B2	2,696	2,782	1,802	64.8	abandoned (adjective)	980	35.2	abnormally (adverb)
合計	7,481	7,814	6,347	81.2		1,467	18.8	

教育改革実施計画」[3] では，CEFR (外国語の学習，教授，評価のためのヨーロッパ共通参照枠）に基づく学習や評価を行うことが示されている．

CEFR では，言語能力を A1，A2（基礎段階の言語使用者），B1，B2（自立した言語使用者），C1，C2（熟達した言語使用者）の 6 段階に分ける．また「読むこと」,「聞くこと」,「やりとり」,「表現」,「書くこと」の 5 つの能力カテゴリーに分けて言語活動の内容を表している.「実施計画」では，高校卒業までに B2 レベルに達することを目標としている．CEFR では「英語を用いて〜することができる」という形式による目標設定（CAN-DO リスト）が示されており，日本の英語教育での利用を目的に構築された CEFR-J も公開されている[4]．そこで当初，CEFR-J の CAN-DO リストと絵本との対応付けを検討したが困難だったため[5]，まずは，基礎語彙力の指標として A1〜B2 の各レベルごとに作成された語彙表『CEFR-J Wordlist Version 1.3』[6](以下，CWL) との自動対応付けと分析を行う．

CEFR-J Wordlist と絵本の比較　まず，絵本に出てくる語は CWL の語をどの程度カバーしているのかを調査する．

CWL は，同じ見出し語でも品詞が異なれば別項目となり，レベルも異なる場合がある．例えば，“round” の場合，“adverb” は A2，“adjective/noun” は B1，“preposition/verb” は B2 が付与されている．そのため，コーパスと対応させるときには，品詞と見出し語の両方が一致する必要がある．ただし，一意に決まらない場合や，解析誤りの可能性があるため，本稿では見出し語のみが一致する場合も許した．また，“out of” の様に，複数の形態素からなる場合，そのままの語順で全形態素が一致する場合のみ対応づけた[7]．

例えば，文 (1)[8]の場合，自動対応付けの結果，A1 レベルの語が 4 語，B2 レベルの語が 1 語となった．ただし，文 (1) の “Biscuit” は固有名詞（NNP) なので，CWL にある “biscuit” とは本来は別項目だが，見出し語が一致するため対応づけられた．

(1) Biscuit likes his rag doll .
NNP VBZ PRP NN NN .
A1 A1 A1 B2 A1 .

表 2 に CWL のレベルごとの語数と絵本との対応結果を示す．表 2 から，A1, A2 レベルの語のほとんどが絵本に出現することがわかる．レベルが高くなるに従って絵本に出現する語は減るが，B2 レベルでも 64.8 % の語は絵本に出現しており，絵本を通してかなりの基本語彙に触れられると考えられる．

次に，絵本には出現するが，CWL にない語の数を表 3 に示す．表 3 から，絵本コーパスの語 (Token) の 12.6% が CWL にない語だった．ただし，そのうち 41.5%を固有名詞が占めており，数字や解析誤り，省略形での出現による不一致も多い．こうした語を除けば，一致率はさらに向上するだろう．

表 3: 絵本に出現するが CWL と一致しなかった語

品詞	Type No.	(%)	Token No.	(%)	例
NNP*	11,952	34.4	59,727	41.5	George
NN*	11,604	33.4	43,245	30.0	dragon
VB*	3,307	9.5	11,294	7.8	’ve, fell
JJ*	5,577	16.0	10,986	7.6	nest, hic
RB*	782	2.2	7,791	5.4	n’t, faster
CD	1,127	3.2	4,682	3.3	1, 2
MD	17	0.0	3,418	2.4	’ll, ’d
UH	64	0.2	1,490	1.0	Oh, Oops
その他	383	1.1	1,360	0.9	de, ozoni
Total	34,791	100	143,993	100	

* は変化形をまとめたことを示す．

[3] http://www.mext.go.jp/b_menu/houdou/25/12/1342458.htm

[4] http://cefr-j.org/cefrj.html

[5] 絵本の多くは，A1 レベルの「簡単な語を用いて書かれた，挿絵のある短い物語を理解することができる.」に該当するなど

[6] 東京外国語大学投野由紀夫研究室．http://cefr-j.org/index.html より，2018 年 8 月 30 日ダウンロード

[7] Token 化と品詞推定は，nltk ライブラリの word_tokenize と pos_tag を利用

[8] Biscuit (Alyssa Satin Capucilli, Pat Schories, 2008, *Harper Collins*) より

例えば，“The Very Hungry Caterpillar”[9] では，CWL と一致した語がのべ 206 語 (A1 が 180 語 (87 %)，A2 が 17，B1 が 5，B2 が 4)，一致しなかった語が 18 語，句読点などの記号が 39 個だった．一致しなかった理由としては，CWL に語が存在しない (“caterpillar”, “Swiss” など (12 語))，“-” の有無 (“ice-cream” (1))，省略形での登録がない (“n’t” (2))，解析誤り (“ate” は “eat” の過去形だが名詞として解析されて原型推定に失敗，など (3)) などだった．

3000 語のリストを用いた Dale-Chall では，絵本が比較的難しく推定される傾向があったが，CWL は 7000 語以上からなり，絵本に出現する語彙のカバー率が高い上，レベル分けもされている．今後，CWL との対応結果を用いた難易度推定や，教育過程に合った絵本の推薦を検討したい．

5　まとめと今後の課題

NTT では，教育利用などを目的として，英語の絵本コーパスを構築中である．

2 章では，NTT で構築している英語絵本のコーパスを紹介した．コーパスサイズは約 145 万語である（2018 年 12 月現在）．3 章では，既存の英語の難易度推定方法を英語絵本コーパスに適用した．その結果，3000 語のリストを用いる Dale-Chall は，音素や平均単語数を使う Flesch-Kincaid に比べて，絵本では難しく判定される傾向があることがわかった．今後は，テキストの難易度が 5 段階で明示されている I Can Read! シリーズのコーパス化を進め，難易度推定の評価・提案を行いたい．

4 章では，基礎語彙力の指標として開発された CEFR-J Wordlist と，英語絵本コーパスの対応と分析をおこなった．WordList で定義されている語の多くは絵本に出現しており，特に，A1 レベルの語では 98.9 %，A2 レベルの語では，93.8 %が出現していた．解析誤りを減らせば，さらに一致率を向上できるだろう．今後，CWL も用いた難易度推定にも取り組みたい．

一方で，CEFR の各レベルは，例えば A1 レベルで，小学校〜中学 2 年程度と幅が広い．今後は，教育課程での利用を念頭に，より細かい教科書単元などへのマッピングに取り組みたい．さらには，教育現場のニーズに合わせた検索インタフェースを実現し，実際の英語教育に役立つシステム開発を目指したい．

[9]Eric Carle, 1969, *Philomel Books*

参考文献

[1] Jeanne S. Chall and Edgar Dale. *Readability Revisited: The New Dale-Chall Readability Formula*. Cambridge, 1995.

[2] William H. DuBay. *The Principles of Readability*. Impact Information, 2004. http://www.impact-information.com/impactinfo/readability02.pdf.

[3] R. F. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32:221–223, 1948.

[4] J.P. Kincaid, R.P. Fishburne, and R.L. Rodgers. *Derivation of new Readability Formulas for Navy Enlisted Personnel*. Research Branch Report 8–75, 1975.

[5] Suzanne E. Mol, Adriana G. Bus, Maria T. de Jong, and Daisy J. H. Smeets. Added value of dialogic parent-child book readings: A meta-analysis. *Early Education and Development*, 19(1):7–26, 2008.

[6] Jessica L. Montag, Michael N. Jones, and Linda B. Smith. The words children hear: Picture books and the statistics for language learning. *Psychological science*, 26, 2015.

[7] Yuko Okumura, Tsssei Kobayashi, Sanat Fujita, and Takashi Hattori. Why is shared book reading effective for children’s theory of mind development?: Frequency analysis of cognitive mental state terms in japanese picture books. In *International conference on Language Acquisition*, 2016.

[8] Elaine Reese and Adell Cox. Quality of adult book reading affects children’s emergent literacy. *Developmental Psychology*, 35(1):20–28, 1999.

[9] Sarah Schwarm and Mari Ostendorf. Reading level assessment using support vector machines and statistical language models. In *Proceedings of ACL-2005*, pp. 523–530, 2005.

[10] Catherine Snow and Anat Ninio. *Emergent literacy: Writing and reading*, chapter The contracts of literacy: What children learn from learning to read books, pp. 116–138. 1986.

[11] Elizabeth Sulzby. Children’s emergent reading of favorite storybooks: A developmental study. *Reading research quarterly*, 20(4):458–481, 1985.

[12] Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada. Sorting Texts by Readability. *Association for Computational Linguistics*, 36(2):203–227, 2010.

[13] G. J. Whitehurst, F. L. Falco, C. J. Lonigan, J. E. Fischel, B. D. DeBaryshe, M. C. Valdez-Menchaca, and M. Caulfield. Accelerating language development through picture book reading. *Developmental Psychology*, 24(4):552–559, 1988.

[14] 藤田 早苗, 服部 正嗣, 小林 哲生, 奥村 優子, 青山 一生. 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, 2 2017.

[15] 藤田 早苗, 小林 哲生, 南 泰浩, 杉山 弘晃. 幼児を対象としたテキストの対象年齢推定方法. **認知科学**, 22(4):604–620, 2015.

[16] 藤田 早苗, 奥村 優子, 小林 哲生, 服部 正嗣. 絵本と幼児向けの発話に出現する語の多様性比較. **言語処理学会第 *23* 回年次大会** *(NLP-2017)*, pp. 1264–1267, 2018.

[17] 樋口 大樹, 奥村 優子, 藤田 早苗, 服部 正嗣, 小林 哲生. 幼児のカタカナ読み書き習得に及ぼす文字特性の影響. **心理学会**, 2018.

[18] 阿久津 由紀子, 小林 哲生, 服部 正嗣, 奥村 優子, 藤田 早苗, 塚田 徹, 鈴木 啓章, 渡辺 佐和. 病院の言語室から絵本を貸し出せるようになるとどうなるか？－絵本を用いた言語発達支援の新しい試み－. **言語聴覚研究**, 15(240), 2018.