絵本と幼児向けの発話に出現する語の多様性比較

藤田 早苗　奥村 優子　小林 哲生　服部 正嗣
NTT コミュニケーション科学基礎研究所
{fujita.sanae,okumura.yuko,kobayashi.tessei,hattori.takashi}@lab.ntt.co.jp

1 はじめに

絵本の読み聞かせは幼児の言語発達を促す重要な情報の 1つと考えられる [6, 10, 14]．例えば，読み聞かせを開始する月齢が早いほど， 2才や 4才の時点での言語理解や発話の能力が高くなること [1, 9]，そして 8ヶ月時点での絵本の読み聞かせが多い方が， 12，および，16 ヶ月時点での語彙が発達していること [4] などが示されている．

このように絵本が言語発達に貢献する要因としては，聞き手 (幼児) と読み手のインタラクションや [13] 対話が促進されること [14]，同一の対象に注意を向ける共同注意が数多く起こること [5]，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 [7, 11] など，数多く挙げられている．このうち，Montag ら [7] は，日常会話における養育者などから幼児に向けた発話 (child-directed speech. 以下，CDS) と絵本とで，出現する**語の多様性 (lexical diversity)** を統計的に比較している．具体的には，のべ語数 (Token) や異なり語数 (Type)，Token に対して Type がどのくらい出てくるか (以下，Type-Token 比) を調査し，絵本の方が CDS より Type-Token 比が高い，つまり，同じ語数で比較すると絵本の方がより多様な語が出現していることを示した．幼児の言語発達の調査では，幼児に向けた発話における語の多様性は言語発達に影響すると報告されており [3]，絵本の語の多様性も言語発達に貢献している可能性を示した．

しかし，Montag ら [7] が用いた絵本は英語で書かれた 100 冊のみであり，他の言語での報告はない．そこで本稿では，より多くの絵本を用いて日本語で，絵本と幼児に向けた発話を比較する．さらに，幼児の年齢による変化や両者に出てくる語の違いも調査する．

2 コーパス

NTT 絵本・児童書コーパス　絵本に出現する語を調査するために，構築中の NTT 絵本・児童書コーパス [1][17] を利用する．NTT 絵本・児童書コーパスは，子供の興味や発達に応じた絵本の推薦 [17] や，発達心理学における絵本に出てくる心的状態語の頻度の調査 [8] を目的として構築しており，多くの幼児に読まれていると考えられる本 (ベストセラー)，名作として専門家によって推薦されている本，長年に渡り出版されつづけている本 (ロングセラー) を含むよう選定し，本文を書き起こしたものである．本稿では，児童書をのぞいた 4,246 冊，5,911 話 [2] を調査に用いる．ここで，冊数と話数が一致しないのは一冊に複数話含まれるお話集を含むためである．これらの絵本の形態素数は 400 万以上であり，言語が異なるので一概には比較できないが，Montag ら [7] の調査 (約 6.8 万語) の約 59 倍である．

CHILDES 日本語コーパス　CDS に出現する語を調査するため，公開されている日本語の CHILDES コーパス [12] を用いる．CHILDES は幼児の言語発達に関する調査を目的として，幼児の発話や周囲の発話を共通のフォーマットで書き起こしたコーパスであり，調査対象である幼児 (Target_Child) の年齢や性別も記録されている．

また，CHILDES では，発話者毎に発話内容が記述されている．例 (1) の場合，母親 (MOT) の発話と，幼児 (CHI) の発話がローマ字で記述されており，CHI は，例を抽出したファイルでの Target_Child である．また，“%ort” から始まる行には，発話内容が日本語で記述されているが，コーパスによっては存在しない．

(1)　*MOT: okaachan no oshigoto yo .
%ort: おかあちゃん の おしごと よ .
*CHI: soo ka [/][3] soo ka .
%ort: ソウカ [/] ソウカ .
(浜崎コーパス [2] (ファイル 030203.cha) より)

本稿では，CHILDES の各コーパスを，Target_Child の発話 (以下，CHI) とそれ以外の発話，つまり幼児に向けた発話 (CDS) に分ける．例 (1) の場合，MOT の発話は CDS に含まれる．また CDS には，兄弟など，Target_Child 以外の幼児の発話も含まれる．表 1 に

[1] 現在も拡張中．過去の論文では「絵本データベース」と紹介．

[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.4246.html で閲覧可能．

[3] [/] は繰り返しを示す記号

表 1: 日本語 CHILDES コーパスの概要とデータサイズ

コーパス [a]	子ども	ファイル数	年齢 (年; 月齢)	平均月齢	取得時期	データ内容	%ort 有無	発話数 [b] CHI[c]	CDS[d]
野地	Sumihare	141	0;0 - 6;11	35.4	1948 - 1955	日記データ	有	39,876	22,732
横山	Kiichan	93	1;0 - 3;0	24.5	1974 - 1976	日記データ	有	31,976	8,775
石井	Jun	100	0;6 - 3;8	27.3	1977 - 1981	親子会話 (父親)	無	44,454	47,097
宮田	Ryo	81	1;3 - 3;0	26.9	1986 - 1988	親子会話	有	12,142	9,588
	Aki	56	1;5 - 3;0	28.9	1989 - 1990	親子会話	有	23,069	26,780
	Tai	77	1;5 - 3;1	27.2	1994 - 1995	親子会話	有	35,167	55,008
浜崎	Taro	32	2;2 - 3;4	33.3	1999 - 2000	親子会話	有	14,864	29,736
MiiPro	ArikaM	54	3;0 - 5;1	46.9		親子会話 (母親)	有	48,694	44,046
	Asato	63	3;0 - 5;0	30.6		親子会話	有	34,218	62,564
	Nanami	57	1;2 - 5;0	34.0		親子会話	有	31,320	74,200
	Tomito	19	2;11 - 5;1	45.9		親子会話	有	11,581	23,678
岡山	130 名	130	2;2 - 4;11	42.3		親子会話 (各 1 日分)	有	27,352	37,629
合計		903[e]	0;0 - 6;11	32.9				354,713	441,833

(a): いずれも https://childes.talkbank.org/access/Japanese/ から取得．(b): 解析対象外 (CHILDES では，xxx が付与されている) を除いてカウント．(c): CHI は対象の子ども (Target_Child) の発話．(d): CDS は CHI 以外の全発話．(e): このうち CHI 以外の発話が収集出来たのは 882 ファイル．

表 2: 絵本と CHILDES(CDS): 子どもの年齢ごとの形態素数

	絵本					CHILDES (CDS)				
		年齢ごと		累積			年齢ごと		累積	
年齢	話数	異なり	のべ	異なり	のべ	ファイル数	異なり	のべ	異なり	のべ
0	285	3,682	31,528	3,682	31,528	16	360	3,454	360	3,454
1	287	6,219	89,295	7,558	120,823	210	5,380	256,420	5,398	259,874
2	336	6,526	104,935	9,976	225,758	383	8,274	660,646	9,569	920,520
3	1,124	10,642	417,407	14,180	643,165	173	8,796	548,170	12,395	1,468,690
4	2,139	17,684	1,252,851	21,341	1,896,016	80	6,880	236,041	13,702	1,704,731
5	1,467	22,069	1,440,300	28,854	3,336,316	17	2,972	41,090	13,964	1,745,821
6	273	21,558	705,589	34,849	4,041,905	12	1,135	8,293	14,015	1,754,114

本調査で用いた CHILDES コーパスの概要とデータサイズを示す．

3 比較方法

形態素解析とカウント方法 出現する語を調査するため両コーパスを形態素解析する．形態素解析器は MeCab[4](ver.0.994) を利用し，品詞体系は UniDic[5][16] 短単位に準拠する．ただし，辞書はひらがなの解析精度を向上させるために再学習したものを利用する [17, 19].

CHILDES では，“%ort” が付与されている場合には “%ort” の日本語を形態素解析し，付与されていない場合，ローマ字による記述を日本語に自動的に変換してから形態素解析を行う．例 (1) の MOT の発話を形態素解析した結果を，例 (2) に示す．

(2) お/接頭辞/オ/御，かあ/名詞-普通名詞-一般/カア/母，ちゃん/接尾辞-名詞的-一般/チャン/ちゃん，の/助詞, 格助詞/ノ/の，お/接頭辞/オ/御，しごと/名詞-普通名詞-サ変可能/シゴト/仕事，よ/助詞-終助詞/ヨ/よ，．/補助記号-句点//.
(解析結果の “出現形/品詞/語彙素の読み/語彙素” を記載)

[4] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[5] http://pj.ninjal.ac.jp/corpus_center/unidic/

本稿では，記号や空白を除いて Type, Token を数える．そのため，例 (2) の場合，Token は 7，Type は “お/接頭辞/オ/御” が 2 回出現するため 6 となる．

年齢ごとの比較 絵本や CDS は，幼児の年齢によって使われる語や多様性も変わる可能性がある．そこで本稿では，年齢ごとの Type と Token の変化も調査する．CDS は，各ファイルに記述された Traget_Child の年齢によって 0 から 6 歳までの各年齢に分けた．また，絵本は，幼児向けテキスト (絵本) を対象とした難易度推定方法 [18] に基づいて一話ごとに対象年齢を推定し，推定された各対象年齢に分けた．表 2 に，絵本と CDS の各年齢ごとの形態素数を示す．

サンプリング方法 Type-Token 比は Token 数によって変化する．そこで，Motag ら [7] と同様，両コーパスからランダムに 100 語ずつ選んで追加していき，各 Token 数での Type 数を調査する．各 Token 数で 100 回試行を繰り返し，CDS の最大 Token 数に合わせ，1,754,100 語まで調査する．ここで，“出現形/品詞/語彙素の読み/語彙素” が一致する語は同じ語だとみなす．また，個々の絵本や対話における傾向を調査するため，絵本の一話ごと，および，CHILDES のファイルごとでも Type-Token 比を調査する．

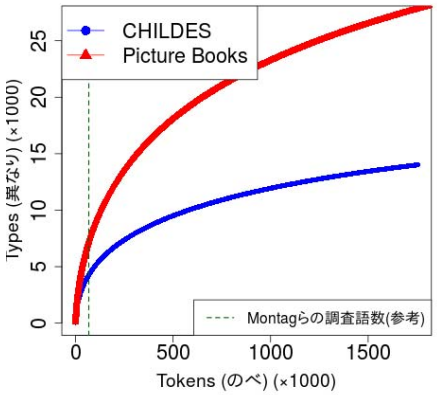

図 1: 絵本と CHILDES(CDS) の比較

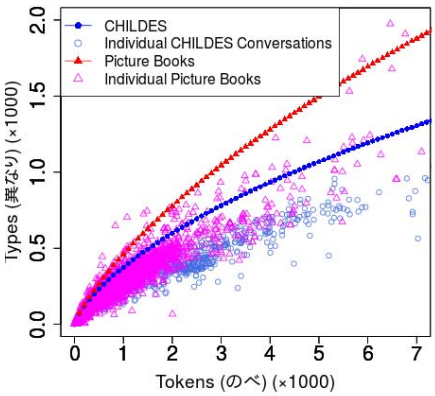

図 2: 個々のファイルのプロットを含む

表 3: 絵本と CDS の一方にのみ出現する語の品詞割合

	絵本のみに出現			CDS のみに出現		
品詞	異なり語数	(%)	例	異なり語数	(%)	例
名詞	12,998	56.6	若者, 人々, 大声	1,112	51.9	ナイナイ, おんも
固有名詞	4,418	19.2	ジョージ, ロッタ	629	29.4	ナナミ, ホノ, シマジロウ
動詞	2,422	10.5	見上げる, 見詰める, 過ごす	194	9.1	放す, 食べさす, 積む
副詞	1,760	7.7	のんびり, せっせと, うっとり	108	5.0	かんかーん, ぱこぱこ
その他	1,378	6.0	おぎゃあ, 貧しい, 氏	99	4.6	あちち, きしょい, わりゃあ
合計	22,976	100		2,142	100	

4 結果と分析

絵本と発話:全体の比較 絵本および CDS の全語から，ランダムに 100 語ずつ抽出した場合の Type と Token の変化を図 1 に示す．さらに，図 2 には，図 1 の一部を拡大し，かつ，絵本の一話ごと，および，CHILDES のファイルごとの CDS の値をプロットした．図 2 ではプロット数が多いために重なってしまっているが，個々のファイルごとの Type-Token 比 [6] の平均は，絵本で 0.390, CDS で 0.246 であり，平均値でも絵本の方が高く，より多様な語が含まれる傾向があることが分かる．

これらの結果から，(1) 同じ語数をランダムに選んだ場合，絵本の方が含まれる語が多様であること，(2) 個々の絵本と CHILDES のファイルを比較しても，個々の絵本の方が Type-Token 比が高い傾向があること，(3) 絵本でも CDS でも，コーパス全体からランダムに語を選択した方が，個々の絵本やファイルより Type-Token 比が高い傾向があること，が分かる．これらは，Montag らの報告 [7] と同傾向の結果である．つまり，日本語でも英語と同様，CDS よりも絵本に含まれる語の方が，より多様であるといえる．

Montag ら [7] は，英語絵本 100 冊 (68,103 語) で比較した結果，同じ語数の CDS の 1.72 倍の異なり語が含まれていたと報告している．ここで，調査に用いた絵本の平均語数は 684 語/話だった．つまり，100 冊で 68,400 語程度となり，Montag らの報告と大きくは変わらない．また，68,400 語を 100 回ランダムに選んだ場合の平均異なり語数を，絵本と CDS で比較すると，絵本は CDS の 1.71 倍の異なり語が含まれており，この比率も英語での報告と非常に近い．

Young ら [15] によると，0-3 歳までの子どもの親の 39%が毎日一回以上絵本を読み聞かせている一方で，16%の親は絵本を全く読んでいない．本調査結果から分かるように，絵本を多く読んであげる方が日常会話だけより多様な語に接する機会が多くなるということであり，言語発達に貢献する可能性もある．

なお，各コーパスにのみ出現する語を抽出し，品詞割合を表 3 に示した．絵本のみ，CDS のみに出現する語は，どちらでも半分以上が名詞であり，固有名詞を合わせると 8 割前後を占め，動詞，副詞，と続いている．

年齢ごとの変化 幼児の年齢による Type-Token 比の変化を調査した (図 3, 4)．この結果，少なくとも絵本の場合，対象年齢が上がると Type-Token 比もより高くなっている．また，CDS では 0, 5, 6 歳のデータは少ないためわかりにくいが，少なくとも 1-4 歳までは，年齢が上がるにつれて Type-Token 比も高くなっている．つまり，CDS でも絵本でも，幼児の成長に伴い，より多様な語が用いられている．さらに，0 歳を除いたすべての年齢で，絵本の方が Type-Token 比が高い，つまり，より多様な語が出現している．

[6] Token を Type で割ったもの．0-1 の間の値を取る．

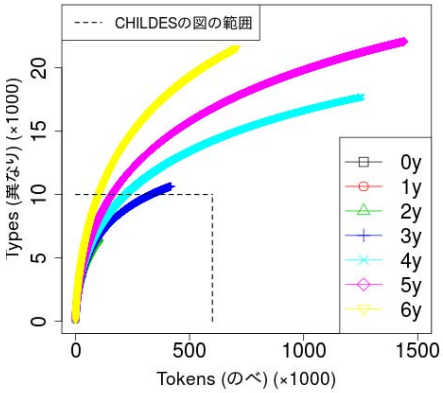

図 3: 絵本の対象年齢ごとの変化

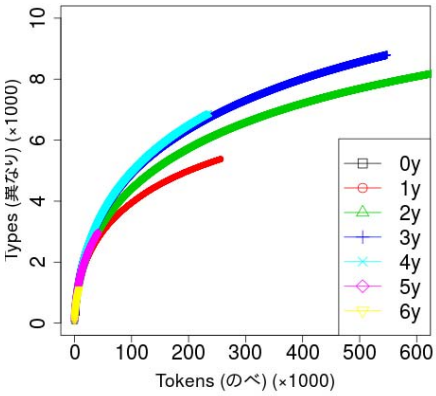

図 4: CHILDES(CDS) の幼児の年齢ごとの変化

5 まとめと今後の課題

本稿では，これまで英語でしか調査されていなかった日常会話における幼児に向けた発話 (CDS) と絵本中の語の多様性の統計的な比較を大規模に行った．言語が異なるので単純に比較はできないが，英語での調査 [7] の約 59 倍の語数で調査した．その結果，英語と同様，絵本の方が多様な語を含み，絵本 100 冊程度の語数 (68, 400 語) で比較すると，同じ語数の CDS の約 1.71 倍の異なり語が含まれることがわかった．

絵本は幼児の言語発達に貢献すると言われており，その要因は数多く挙げられているが，語の多様性もその一つである [7, 11]．本調査結果から，日本語の絵本でも日常会話に比べて絵本に出現する語は多様であり，語の多様性という面からも幼児の言語発達に寄与している可能性を示すことができた．今後は，絵本と日常会話との比較をより詳細にすすめ，教育支援に活かしたい．

参考文献

[1] Barbara D Debaryshe. Joint picture-book reading correlates of early oral language skill. *Journal of child language*, 20(02):455–461, 1993.

[2] Naomi Hamasaki. *Japanese - Hamasaki Corpus*. Pittsburgh, PA: TalkBank, 2004.

[3] Janellen Huttenlocher, Heidi Waterfall, Marina Vasilyeva, Jack Vevea, and Larry V. Hedges. Sources of variability in children 's language growth. *Cognitive Psychology*, 61:343–365, 2010.

[4] Jan Karrass and Julia M Braungart-Rieker. Effects of shared parent–infant book reading on early language acquisition. *Journal of Applied Developmental Psychology*, 26(2):133–148, 2005.

[5] Jan Karrass, Meghan C VanDeventer, and Julia M Braungart-Rieker. Predicting shared parent-child book reading in infancy. *Journal of Family Psychology*, 17(1):134, 2003.

[6] 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.

[7] 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.

[8] 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.

[9] Adam C Payne, Grover J Whitehurst, and Andrea L Angell. The role of home literacy environment in the development of language ability in preschool children from low-income families. *Early Childhood Research Quarterly*, 9(3):427–440, 1994.

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

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

[12] 宮田 Susanne, 森川 尋美, and 村木 恭子, editors. **今日から使える発話データベース－初心者のための** *CHILDES* **入門**. ひつじ書房, 2004.

[13] Michael Tomasello and Michael Jeffrey Farrar. Joint attention and early language. *Child development*, 57(6):1454–1463, 1986.

[14] 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.

[15] Kathryn Taaffe Young, Karen Davis, Cathy Schoen, and Steven Parker. Listening to parents: A national survey of parents with young children. *Archives of Pediatrics & Adolescent Medicine*, 152(3):255–262, 1998.

[16] 伝 康晴, 山田 篤, 小椋 秀樹, 小磯 花絵, and 小木曽 智信. *UniDic version 1.3.9* **ユーザーズマニュアル**, 2008.

[17] 藤田 早苗, 服部 正嗣, 小林 哲生, 奥村 優子, and 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.

[18] 藤田 早苗, 小林 哲生, 南 泰浩, and 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.

[19] 藤田 早苗, 平 博順, 小林 哲生, and 田中 貴秋. 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, 2014.