幼児の語彙獲得と絵本コーパスの関係を探る

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

1　はじめに

本稿では，幼児が語を獲得する時期と絵本コーパス中の頻度の分析を通して，幼児にとっての語の獲得の難しさの分析に絵本コーパスが有効であることを示す．

発達心理学の分野では，幼児に「ネケ」や「ヘク」などの実験用の無意味語を学習させる実験を通して，幼児にとってどういった語の学習が難しいかを調べる研究がされている．例えば，今井ら [5] によると，基礎カテゴリである物や動物の名前は，教えられるとすぐに覚えることのできる「即時マッピング」が可能だが，より抽象的な概念である上位カテゴリは基礎カテゴリに比べて学習が難しい．こうした研究は，幼児にとって学習が難しい語のタイプやカテゴリを明らかにできるが，無意味語を用いるため，実際の語では調査できないという問題がある．

一方，小林ら [6] は，幼児がいつどんな語を習得するかを調査した幼児語彙発達データベース (CVD) を構築した．CVD からは,「イヌ」という語を 796 日齢で 50%の子どもが発話できるようになり,「動物」なら 964 日齢,「サイ」なら 1100 日齢であるといったことがわかる．つまり上位カテゴリの語である「動物」は，基礎カテゴリの「イヌ」より獲得時期は遅いが,「サイ」よりは早い．とはいえ,「サイ」の方が「動物」より幼児にとって学習が難しいとは限らない．単に「動物」の方が,「サイ」よりインプットされる機会や回数が多いため，難しい語の割に早く獲得されるだけかもしれない．つまり，実際の語で学習の難しさを調べるためには，獲得時期だけでなく，インプットの量と組み合わせる必要がある．

幼児へのインプットとしては，会話やメディア，絵本の読み聞かせなどが考えられる．幼児への全てのインプットを記録することは現実的には不可能だが，本稿では，重要なインプットの一つである絵本 [9] に着目する．絵本コーパス中の頻度を幼児に対するインプット量だと見なすと，インプット量に比べて獲得時期が早いか遅いかという分析が可能となる．分析の結果，実際の語のデータから，無意味語を用いた今井らの研究結果を支持する結果が得られることを示す．

2　先行研究

今井ら [5] は実験を通して，(a) 幼児にとっては名詞より動詞や形容詞の学習が難しく，(b) 同じ名詞でも基礎カテゴリより抽象度の高い上位カテゴリの学習が難しく，(c) 子どもに対しては大人に対する場合より擬態語が使われる頻度が高いこと，等を示している．本稿の分析では，これらの知見と比較する．

次に，CVD[6] を紹介する．CVD は，0 − 4 歳頃までの乳幼児の保護者約 1300 名に，調査時点で自分の子が該当の語を理解できるか，発話できるかどうかをチェックリスト形式で解答してもらい，その情報をデータベース化したものであり，2700 語が収録されている．さらに南ら [7] は，ある語を獲得する子どもの割合は，ロジスティック回帰曲線に一致することを示し，任意の割合の子どもが各語を理解・発話できるようになる日齢を算出可能にした．本稿では，50 %の子どもが発話できるようになる日齢 (以下，獲得日齢) を調査に用いる．ここで，理解できる日齢ではなく，発話できる日齢を用いるのは，発話されたかどうかの方が保護者にわかりやすく，信頼度が高いからである．

獲得日齢とコーパス頻度の関係を調査した先行研究は存在しないが，単語親密度とコーパス頻度との相関関係を調査した先行研究を紹介する．単語親密度とは，大人を対象とした心理実験によって語のなじみ深さを 1-7 の間で数値化したものである [1]．寺田ら [8] は，単語親密度と各種コーパス (毎日新聞，Wikipedia，Web からのクローリングデータ，青空文庫，日本語話し言葉コーパス (CSJ)) 中の頻度との相関係数[1]を算出し，コーパスのサイズと種類の影響を調べた．具体的には，サイズは 10 億語程度まで対数線形的に相関係数が高くなり，種類は話し言葉である Web データや CSJ の方が，書き言葉よりも相関係数が高い傾向が見られることを示した．

3　コーパスと頻度の数え方

構築中の絵本データベース [2] 中の本文データを絵本コーパスとして利用する．絵本データベースは，発達心理学における研究や，子供の興味や発達に応じた

[1] ピアソンのモーメント相関係数とスピアマンの順位相関係数の両方を算出している．

表 1: コーパスサイズ

	形態素数 のべ	異なり	話数/ファイル数
絵本	81,986	28,024	3,455
BCCWJ	19,560,391	116,981	172,675

絵本は 2,661 冊で，1 冊に複数話収録されたお話集も含む．

絵本の推薦を目的として構築しており，ベストセラーやロングセラーを含むよう選定している．また，絵本コーパスでの結果と比較するため，現代日本語書き言葉均衡コーパス[2](以下，BCCWJ) でも調査する．表 1 にこれらのコーパスのサイズを示す．

CVD に含まれる語は必ずしも 1 形態素で構成されるとは限らない．例えば「お姉ちゃん」は， UniDic では「お」「姉」「ちゃん」の 3 形態素となる．また，「お姉ちゃん」「姉」「ねーね」のように，同じ対象について複数の表現が存在する場合もあり，CVD では異なる語として獲得日齢を算出している．

本稿では，CVD と同様にまとめるため，形態素の最長一致によって頻度を数えている．また,「ちゃん」ではなく「さん」が使われている場合や,「お」が省略されている場合，表記ゆれとみなせる場合も同様にまとめた．つまり,「お姉ちゃん」「姉ちゃん」「お姉さん」「姉さん」は「お姉ちゃん」[3],「姉」「あね」は「姉」,「ねえね」「ねぇね」は「ねーね」として数える．

また，頻度としては，出現頻度 (FREQ) と，いくつの話に出てきたか (DF)[4]の対数値を用いた．

4 結果と分析

本章では，まず，幼児の語の獲得日齢と絵本コーパス中の頻度の相関を調べ，高い相関があることと，カテゴリごとの分析の重要性を示す (4.1 節)．口語表現や家庭環境が影響する事例も紹介しつつ (4.2 節)，両者を利用するとインプットの量と獲得の難しさの関係をよくモデル化できることを示す (4.3, 4.4 節)．

4.1 順位相関係数

表 2 に，獲得日齢と絵本，BCCWJ を用いて算出したスピアマンの順位相関係数 (ρ) を示す．表 2 には，CVD 中の全語をまとめて用いた場合と，CVD に付与された 28 カテゴリに分けた場合の結果を示した．なお，相関係数の算出にはコーパス中に 1 度以上出現した語のみ用いた．

表 2 の結果から分かることを (1)-(3) に示す．
(1) ほぼ全カテゴリで，BCCWJ より絵本の方が相関が高い．全語をまとめた場合，絵本とは弱い相関，BCCWJ とは相関があるとは言えない程度だった．

(2) 絵本はコーパスサイズの割に相関が非常に高い．寺田ら [8] は，単語親密度との相関係数は，10 億語程度までは対数線形的に高くなるとしている[5]．BCCWJ の形態素数は絵本の 239 倍あるにも関わらず，カテゴリによっては絵本の方が相関がはるかに高く，獲得日齢と絵本は相関が高いと言える．
(3) カテゴリによって相関の程度が大きく異なる．
(3-1) 相関が高いカテゴリは,「動物」「かたち・色」「時間・数量」「乗り物」.
(3-2) 相関がほぼ見られないカテゴリは,「擬音語・擬態語」「キャラクター」「助詞・助動詞」など．これらのカテゴリの語の品詞は，感動詞，固有名詞，助詞，助動詞などが多い．

4.2 口語表現や家庭環境の影響

「家族」カテゴリには,「お姉ちゃん」と「ねーね」のように，同じ対象に対する異なる表現が含まれる場合がある．図 1 に,「家族」カテゴリの散布図と回帰直線を示す．図 1 から,「ママ」と「パパ」のように対になる語は散布図でも近くにプロットされており，獲得日齢も出現頻度も近いことが分かる．また,「ねーね」「にーに」のような口語表現は「お姉ちゃん」「お兄ちゃん」のような表現に比べ，獲得日齢は早いが出現頻度は低く，両者の傾向は大きく異なることがわかる．

そこで,「お姉ちゃん」と「ねーね」のように，同じ対象に対して複数の表現が含まれる場合，より口語的な表現を除いた 10 語だけを用いて，順位相関係数を算出しなおした (表 2，最下段)．その結果，絵本の DF で $\rho = -0.93$ という高い値が得られた[6].「ねーね」「にーに」などの言葉を早期に覚えるかどうかは，家族構成によるところが大きいためコーパスとの相関は低いのだと考えられる．

このように，口語表現や家庭環境に依存する語を除くと，絵本コーパスとの相関はより高くなる．

4.3 基礎カテゴリ

「動物」「乗り物」カテゴリの語は，ほとんどが基礎カテゴリにあたる．本節では，順位相関係数が最も高かった「動物」カテゴリを取り上げる．

「動物」カテゴリでは，FREQ よりも DF を用いる方が相関が高かったため，DF を用いた場合の散布図と回帰直線を図 2 に示す．図 2 からも，ほとんどの語が回帰直線に沿ってプロットされており，出現頻度 (インプット量) が高い方が獲得日齢が早いという関係をよく表わしている．

[2]http://www.ninjal.ac.jp/kotonoha/
[3]「おねえちゃん」などの平仮名表記も含む．
[4]DF では，1 話の中に複数回出現しても頻度 1 と数える．
[5]日本語の場合，単語親密度と最も順位相関係数が高かったのは，毎日新聞の 8 千万語で，$\rho = 0.5330$ だった．
[6]表 2 で,「家族」カテゴリでは BCCWJ の方が順位相関係数が高かったが,「家族」カテゴリの語は BCCWJ に 12 語しか出現せず，これらのみで算出したためである．

表 2: Spearman's rank correlation (ρ)

カテゴリ	全語数	例	絵本 ρ log(DF)		絵本 ρ log(FREQ)		絵本 一致語数	BCCWJ ρ log(DF)		BCCWJ ρ log(FREQ)		BCCWJ 一致語数
全語	2,512		-0.39	***	**-0.41**	***	1,905	-0.21	***	-0.20	***	1,910
動物	100	虫	**-0.83**	***	-0.78	***	100	-0.61	***	-0.50	***	99
かたち・色	26	丸	**-0.78**	***	-0.77	***	22	-0.62	***	-0.61	**	26
時間・数量 ▲	92	朝	**-0.73**	***	-0.71	***	91	-0.44	***	-0.37	***	91
乗り物 ▽	41	車	-0.72	***	**-0.74**	***	41	-0.60	***	-0.60	***	37
人称代名詞 ▲	18	私	**-0.66**	**	-0.63	**	17	-0.61	*	-0.56	*	17
空間 ▲	31	あいだ	-0.63	***	**-0.65**	***	30	-0.47	**	-0.47	**	30
衣類	75	服	-0.63	***	**-0.65**	***	71	-0.35	**	-0.29	*	73
からだ	59	身体	-0.63	***	**-0.65**	***	58	-0.55	***	-0.49	***	57
自然 △	54	火	-0.63	***	**-0.66**	***	54	-0.45	***	-0.52	***	54
ひと △	52	人間	-0.60	***	**-0.65**	***	51	-0.06		-0.02		52
動作・事象 ▲	244	会う	-0.58	***	**-0.60**	***	237	-0.45	***	-0.44	***	237
その他 △	76	声	-0.57	***	**-0.58**	***	71	-0.27	*	-0.28	*	74
食べ物・飲み物 ▼	170	飴	**-0.54**	***	-0.53	***	163	-0.53	***	-0.52	***	170
部屋と設備 △	69	家	**-0.53**	***	-0.52	***	64	-0.53	***	-0.47	***	65
こころ ▲	59	こわい	-0.51	***	**-0.52**	***	55	-0.39	**	-0.39	**	56
屋外の物・場所	101	お外	**-0.50**	***	-0.47	***	88	-0.20	*	-0.18	.	99
おもちゃ・道具 ▽	116	ボール	-0.48	***	**-0.53**	***	107	-0.39	***	-0.42	***	110
ようす △	72	明るい	**-0.47**	***	-0.47	***	71	-0.35	**	-0.42	***	71
家庭用品 ▽	120	鍵	**-0.44**	***	-0.42	***	117	-0.41	***	-0.34	***	119
あいさつ ▽	27	どうも	-0.43	*	**-0.44**	*	27	-0.07		-0.08		26
家族	16	ママ	-0.41		-0.46	.	16	**-0.74**	**	**-0.74**	**	12
助詞・助動詞 ▲	45	が	**-0.39**	*	-0.35	*	38	-0.37	*	-0.33	*	38
やりとり・かけ声 ▽	62	どうぞ	**-0.30**	*	-0.29	*	60	-0.08		-0.11		56
代名詞・疑問詞 ▲	35	あそこ	-0.28		**-0.31**	.	34	-0.24		-0.23		34
日課 ▼	52	ごはん	-0.13		-0.20		46	0.22		**0.22**		47
擬音語・擬態語 ▼	110	がおー	-0.09		**-0.12**		100	-0.04		0.04		79
キャラクター ▽	72	神様	-0.01		**0.13**		44	-0.11		-0.12		49
遊び・活動	36	お歌	0.15		0.18		32	0.27		**0.32**	.	32
家族 (10 語) ▲	10	お母さん	**-0.93**	***	-0.83	**	10	-0.70	*	-0.70	*	10

▲△▽▼は 4.4 節参照. ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, . : $p < 0.1$. 一致語数はコーパス中に 1 度以上出現した語の数.

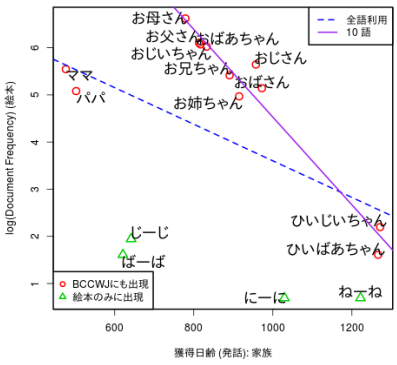

図 1: 絵本と獲得日齢 (家族)

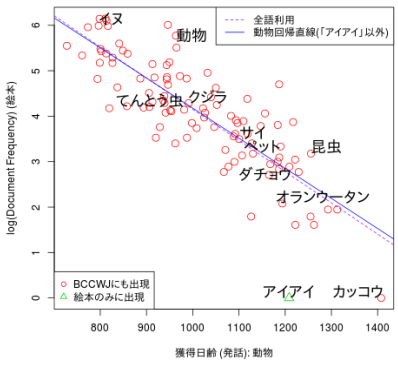

図 2: 絵本と獲得日齢 (動物)

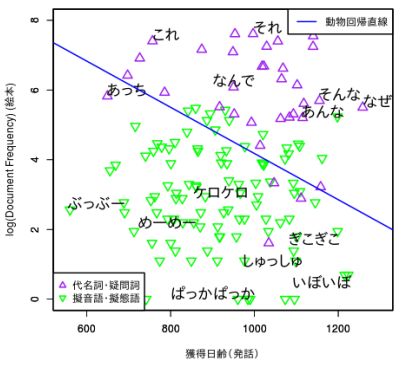

図 3: 「動物」カテゴリとの比較

さらに,「動物」カテゴリに含まれる語でも,「動物」「ペット」「昆虫」は上位カテゴリの語である. これらの語は, 回帰直線より上, つまり, 出現頻度に比べて獲得日齢が遅めであり, 上位カテゴリの方が学習が難しいという知見を支持する結果となった.

ここで, 回帰直線からの残差が最も大きい「アイアイ」を外れ値として除き[7]絵本の DF を用いた回帰式を求め直すと, 回帰式は $age = 1450.7 - 107.1 * log(DF)$ となった. 以下, この回帰式で描かれる回帰直線を動物回帰直線と呼ぶ.

[7]「アイアイ」は絵本での出現頻度が低い割に獲得日齢は早い. これは, 童謡の影響だと思われる.

CVD に収録されていない語でも，この回帰式を用いれば獲得日齢の推定ができる．例えば CVD の「動物」カテゴリ中の語は，日本語語彙大系 [4] では，意味クラス <536:動物（個体）> 配下の語が該当する．これらの意味クラスの語は，絵本データベースに 460 語出現している．CVD に収録されている「動物」カテゴリの語は 100 語であり，残りの 360 語は収録されていない．高頻度語ほど収録されている語は多いが，比較的頻度の高い語でも収録されていない語も存在する．こうした収録外の語でも，本稿で得られた回帰式を用いることで，獲得日齢を推定することが可能である．例えば，CVD に収録されていない「タイ」や「タカ」でも，「タイ」は「クジラ」と近い日齢に，「タカ」は「ダチョウ」と近い日齢で覚えられると推定できる．

CVD で収録語を増やすためには，大規模なデータ収集が必要なため，簡単に増やすことはできない．そのため，収録対象外の語でも，獲得日齢の推定ができることは有意義である．

4.4 「動物」カテゴリとの比較

前述の通り，「動物」カテゴリの語のほとんどは基礎カテゴリである．動物回帰直線が，基礎カテゴリにおける絵本の DF と学習の難しさの関係をモデル化できたとすると，動物回帰直線と比較すれば，基礎カテゴリと比べて学習が難しいかどうか判別する手がかりになる可能性がある．

そこで，各カテゴリで，動物回帰直線より上にプロットされる語の割合を調べた．上にプロットされる語は，出現頻度と比べて獲得が比較的遅く，下にプロットされる語は，出現頻度と比べて獲得が比較的早いと考えられる．調査の結果，カテゴリによって，動物回帰直線のどちら側に分布するかが偏る場合が見られた．偏りが見られる場合，表 2 のカテゴリ名の右肩に「▲△▽▼」のマークを付与した．「▲」は各カテゴリの語の 80% 以上が動物回帰直線の上にプロットされる場合，「△」は 60 %以上，「▽」は 40 %以下，「▼」は 20 %以下であることを示している．

CVD のカテゴリのうち，動詞や形容詞が多く含まれるカテゴリは，「動作・事象」「こころ」「ようす」だが，これらはいずれも，動物回帰直線の上に偏っている (2 節, (a) に対応)．また，名詞ばかりのカテゴリでも「人称代名詞」「空間」等の抽象度の高いカテゴリは上に偏っている ((b) に対応)．逆に，「日課」や「擬音語・擬態語」は下に偏っている．こうした，頻度に比して獲得が早いカテゴリは，基礎カテゴリ以上に獲得が容易であるという可能性と，身近な物や日常的によく使う表現なため，会話等によるインプットが多いという可能性が考えられる．これは，少なくとも擬態語は幼児に対して高頻度で用いられるという知見とも一致する ((c) に対応)．

例として「代名詞・疑問詞」と「擬音語・擬態語」の散布図と動物回帰直線を図 3 に示す．図 3 からも，抽象度の高い「代名詞・疑問詞」は動物回帰直線より上に，「擬音語・擬態語」は下に偏ってプロットされることが見て取れる．これらのカテゴリは，順位相関係数だけを調べると相関はほぼない程度である．しかし，動物回帰直線と比較することで，これまで実験的に示されてきた知見を支持する結果を得ることができた．

5 まとめ

本稿では，幼児の語の獲得日齢と絵本コーパス中の頻度の相関関係を調査し，特に基礎カテゴリでは相関が非常に高いことを示した．

また，絵本中の頻度が幼児へのインプット量に対応すると仮定し，インプット量の多さと獲得日齢の対応関係を分析，獲得日齢だけではわからない傾向を明らかにした．つまり，上位カテゴリ中の獲得の早い語と基礎カテゴリ中の獲得の遅い語は，獲得日齢だけで比較すると，上位カテゴリの語の方が獲得が容易なように見えるが，インプット量と比べて獲得が早いか遅いかという評価を行えるようになり，上位カテゴリの語の方が頻度に比べて獲得が遅めであることを示せた．こうした結果は，発達心理学の分野で無意味語を用いて実験的に得られた知見とよく一致している．本稿では，これらの知見を支持する結果を実際の語のデータから得ることができた．

今後は，動詞や形容詞等の活用する語について，頻度だけでなく，活用のバリエーションや項構造，構文の難しさを調査し，学習の難しさとの関係をより詳細に調べたい．また，単語親密度でも，本稿同様，語の属するカテゴリによって相関の傾向が変わるかどうかを調査したい．

参考文献

[1] 天野 成昭, 近藤 公久. **日本語の語彙特性**. 三省堂, 東京, 1999.

[2] 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), 2 2017.

[3] Hans Stadthagen-Gonzalez and Colin J. Davis. The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, 38(4):598–605, 2006.

[4] 池原 悟, 宮崎 雅弘, 白井 諭, 横尾 昭男, 中岩 浩巳, 小倉 健太郎, 大山 芳史, 林 良 彦. **日本語語彙大系**. 岩波書店, 1997.

[5] 今井 むつみ, 針生 悦子. **レキシコンの構築 -子どもはどのように語と概念を学んでいくのか-**. 岩波書店, 2007.

[6] 小林 哲生, 奥村 優子, 南 泰浩. 語彙チェックリストアプリによる幼児語彙発達データ収集 の試み. **電子情報通信学会技術研究報告**, 115(418):1–6, 2016. (HCS2015-59).

[7] 南 泰浩, 小林 哲生. 幼児の発達に応じた語彙検索システム. **電子情報通信学会和文誌** *D*, J96-D(10):2612–2624, 2013.

[8] 寺田 博視, 田中 久美子. 単語親密度と頻度に関する考察. In **言語処理学会第** *14* **回年次大会**, pages 713–716, 2008.

[9] 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.