

# ウェブ上で収集した幼児語彙発達データの信頼性検証

小林 哲生・永田 昌明

NTT コミュニケーション科学基礎研究所 協創情報研究部

〒619-0237 京都府相楽郡精華町光台 2-4

**発表要旨** 本研究では、ポータルサイトの育児情報コーナーに、わが子が覚えた語彙(わんわん, まんま等)を継続記録できるウェブ日誌ツールを用意し、全国のウェブユーザから無報酬で収集した幼児語彙発達データがどの程度信頼性があるかを検証するために、研究者の指示下で親が1年間記録した語彙発達データとの比較を試みた。その結果、語彙の獲得速度(初語, 25 語, 50 語到達日齢), 及び意味カテゴリ分布(NV 比)の点において両データ間に有意な差がなかった。これは、ユーザが自発的に無報酬でウェブへ入力したデータであっても、研究者の指示下で収集したデータと同等の信頼性を持ち、ユーザの多くが正直かつ勤勉にデータを入力してくれたことを示唆している。

## 1. 背景と目的

幼児がどのように言語を獲得するかを解明するには、彼らがいつ、どんな単語を覚える傾向にあるかを特定し、その特徴から語彙獲得メカニズムとその機序を明らかにしていく必要がある。

従来は、こどもが獲得した単語を親が質問紙で回答する語彙チェックリスト法を用いて多くの研究が行われてきたが(Caselli ら, 1999; Fenson ら, 1994; 小椋・綿巻, 2008), データの多くが複数の月齢群(例えば 12, 18, 24 ヶ月齢)の横断データ(cross-sectional data)からなり、群間の比較しかできなかつたため、個人差を考慮できず、詳細な発達過程を十分に検討できないという問題があった。

より有効な方法は、こどもが覚えた単語を随時日誌に記録していく日誌法や、親子の会話場面を一定期間毎に撮影して書き起こす自然観察法により、特定の個人を追跡調査し、縦断データ(longitudinal data)を取得する方法である。しかし、こうした縦断データは、その取得に多大な時間と労力を要するため、これまで少人数のデータしか取得できなかった(天野ら, 2008; MacWhinney, 2000; Miyata ら, 2004)。そのため、見出された結果の一般化や理論化が十分に行えないという問題があった。

こうした問題を解決する方法のひとつが、ウェブを利用して縦断データを多数収集するウェブ日誌法

(web diary method)である。この手法を用いて、小林・永田(2009)は、ポータルサイトの育児情報コーナーに、わが子が覚えた語彙(まんま, わんわん等)を日齢単位で記録・管理できるウェブ日誌ツールを公開し、ウェブユーザから初期語彙の縦断データを大規模に収集できるかどうかを検討した。その結果、約2年間に日本全国の約800名のウェブユーザから約12,000語のデータを収集するのに成功した。また取得したデータを用いて各単語の平均獲得月齢(AoA: age of acquisition)を算出することにより早期出現50語(first 50 words)を特定し、日本語学習児の初期語彙発達の特徴を明らかにしてきた(小林・永田, 2008)。

このウェブ日誌法は、初期語彙発達の縦断データを効率的に多数取得できるという点で有効な手法と言えるが、その一方でウェブユーザから収集されたデータがどの程度信頼できるかについては検討の余地がある。語彙チェックリスト法や自然観察法などの従来研究では、データ取得に研究者が介在するため、データの信頼性が問題となることはなかった。しかし、ウェブ日誌法は、無報酬で投稿された匿名ユーザからのデータであり、また研究者がデータ記入の指導や指示を逐一できないため、ユーザが正直にかつ勤勉にデータを投稿してくれたかを確認できない。したがって、ウェブ日誌法で収集したデータは何らかの手段でその信頼性を検証することが必要となる。

こうしたウェブデータの信頼性検証法のひとつとしてこれまでに行ってきたのは、ウェブ日誌法のデータから算出した各単語の平均獲得月齢(AoA)と、語彙チェックリスト法の横断データから算出した各単語の「50%到達月齢」(50%の子どもに獲得された月齢時点; 小椋・綿貫, 2008)との一致を見ることであった。早期出現 50 語における平均獲得月齢と 50%到達月齢の級内相関係数を算出した結果, 有意に高い一致性を示した( $ICC[3,2] = .753, p < .001$ ; 図 1)。この結果から、ウェブ日誌法から得られた各単語の平均獲得月齢が、既存データと概ね一致することがわかった。

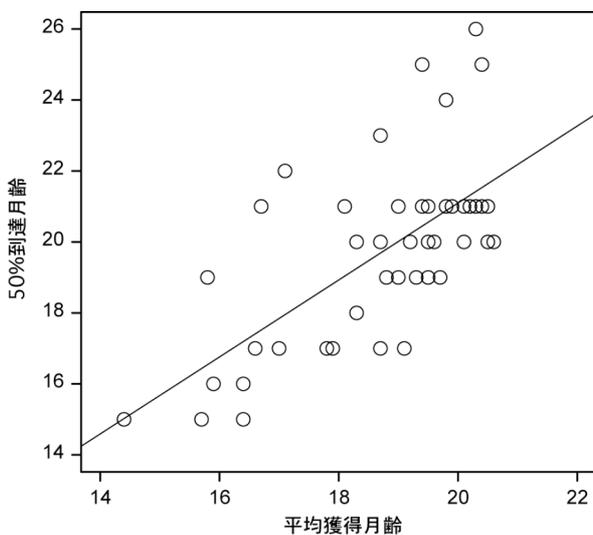


図 1. 早期出現50語(first 50 words)の平均獲得月齢(AoA)と50%到達月齢の関係 (Kobayashi & Nagata, under review).

しかし、上記の検証法は、各単語の平均獲得月齢という全体傾向からデータの信頼性を検討したものであり、各ユーザがどの程度データを正直にかつ勤勉に報告しているかは検証できていない。そのため、ウェブ日誌法により収集した縦断データが有効なものかどうかは未解決のままである。こうした問題を解決するひとつの方法は、通常の方法で取得したデータとウェブデータを、語彙の獲得速度や意味カテゴリ分布の点から直接比較することである。しかし、縦断研究の不足により、比較可能なデータは現在、存在していない。

そこで本研究では、研究者の指導・指示の下で約1年間親にわが子の語彙発達を記録してもらい、その収集した縦断データを、全国のウェブユーザからのデータと突き合わせることにより、ウェブ日誌法で取得した縦断データの信頼性を検証する試みを行った。この結果から、ウェブ日誌法によるデータ取得の有効性を検討する。

## 2. 方法

### 2.1. 参加者

**京阪奈データ:**京阪奈地区に住む母親(9-11ヶ月齢児がいる方)に広告媒体を通じて調査協力を呼びかけ、応募のあった18名( $M = 33.8$ 歳,  $SD 3.9$ )に、研究者の指示下で約1年間の語彙発達データの記録をお願いした。調査は2007年9月から翌年9月迄の約1年間実施した。最終的に、24ヶ月齢に達するまで継続記録いただいた15名分のデータを分析対象とした(男児8名, 女児7名)。

**ウェブデータ:**京阪奈地区で実施したデータ収集と同時期に、約1年間ウェブより継続投稿のあったウェブユーザ20名分( $M = 33.6$ 歳,  $SD 3.9$ ; 全員母親)を分析対象とした(男児11名, 女児9名)。

### 2.2. 手続き

ウェブデータは、ポータルサイト goo にある育児情報サイト「goo ベビー」(<http://baby.goo.ne.jp/>)の会員を対象に収集した。本サイトでは、こどもがいつ、どんな単語を覚えたかを日齢単位で記録・管理できる日誌ツールが無料で提供され、会員は誰でも、それを自由に利用することができた(小林・永田, 2009)。

データを継続投稿すると、「語彙成長グラフ」と「月齢別獲得語彙リスト」が自動表示された。また日記や写真、身長・体重も同時に記録できるため、成長記録ツールとして利用価値が高くなるよう工夫した(図2を参照)。さらに、携帯電話からのデータ投稿もできるようにし、育児に忙しい母親でも手軽に記録できるよう工夫した。



図2. gooベビー「赤ちゃん成長ダイアリー」のデータ入力画面(左)と成長グラフ画面(右)。入力の際は、こどもがはじめて覚えた単語を、発話(しゃ)・意味(電車)・日付(2009.1.7)の点から記入。URLは、<http://baby.goo.ne.jp/>。

一方、京阪奈の参加者にも同様に、ウェブ日誌ツールを利用して、こどもがいつ、どんな単語を言えるようになったかを日齢単位で記録してもらった。事前説明会で、調査の目的・概要及びウェブ日誌ツールの利用法などを丹念に説明した。参加者は全員、1-2ヶ月に1度、研究所に招集され、データの記録状況などを担当研究者に報告した。なお、データ記録について不明な点がある場合は、メールと電話にて研究担当者に連絡するように指示をした。なお、参加者へは、調査協力の謝金を支払った。

### 2.3. データ分析

ウェブデータ及び京阪奈で取得したデータのいずれにおいても2名のスタッフにより登録データのチェックを行った後、語彙の獲得速度および意味カテゴリ分布の点から両データの比較検定を行った。

## 3. 結果

### 3.1. データ間比較1(語彙の獲得速度)

ウェブユーザが正直にかつ勤勉にデータを投稿しているのであれば、語彙の獲得速度の点において京阪奈で取得したデータと差が見られないはずだと予測した。この点を検証するために、データの種類(被験者間要因: ウェブ vs. 京阪奈)と、各累積語彙数時点の到達日齢(被験者内要因: 初語 vs. 25語 vs. 50語)における2要因混合計画の分散分析を行った。

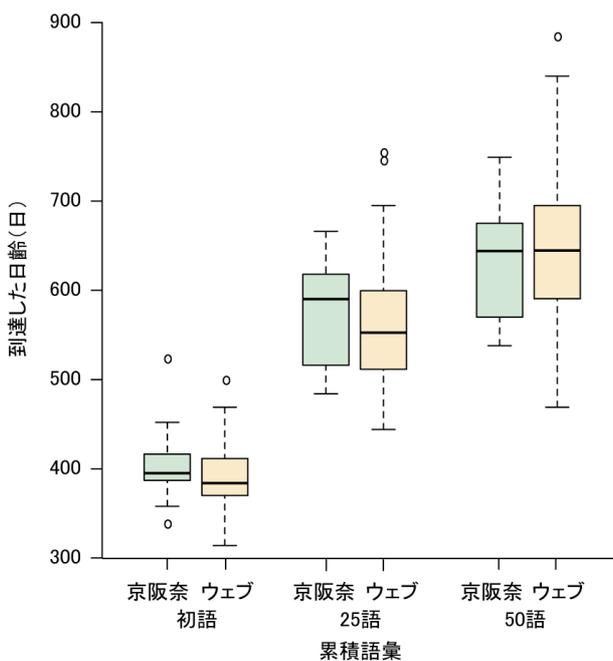


図3. ウェブデータと京阪奈で取得したデータにおける、各累積語彙数時点(初語、25語、50語)に到達した平均日齢の比較。三時点ともに、各データ間に有意差はなかった。

その結果、データの種類に関する主効果 ( $F[1, 31] = .001, p = .970$ )及び交互作用 ( $F[2, 30] = 1.476, p = .245$ )に有意な差は見られなかった。したがって、初語開始時(ウェブ:  $M = 391.5$ 日 [ $SD 44.1$ ]), 京阪奈:  $M = 405.2$ 日 [ $SD 43.9$ ]), 25語到達時(ウェブ:  $M = 564.8$ 日 [ $SD 86.0$ ]), 京阪奈:  $M = 571.7$ 日 [ $SD 61.4$ ]), 50語到達時(ウェブ:  $M = 649.4$ 日 [ $SD 104.7$ ]), 京阪奈:  $M = 630.3$ 日 [ $SD 56.2$ ])のいずれの時点においても、両データ間の到達日齢に差がなく、語彙の獲得速度に違いがないことを示唆している(図3)。

### 3.2. データ間比較2(意味カテゴリ分布)

データを不正確に投稿するユーザが増えると、初期語彙に含まれる意味カテゴリ分布に偏りが生じる可能性が高いため、この点からもウェブデータの信頼性を検証できると考えた。各個人の早期出現50語を対象に、幼児言語発達の分野で使用される意味カテゴリ(社会語、普通名詞、動作語、形容詞、人々、幼児語、機能語; Fenson ら, 1994)を付与し、データごとに、早期出現50語に占める各意味カテゴリの平均語数を算出し、表1にまとめた。

表1. 各データ(京阪奈とウェブ)における早期出現50語に占める意味カテゴリごとの平均語数(括弧内は標準偏差)

	意味カテゴリ						
	社会語	名詞	動詞	形容詞	人々	幼児語	機能語
京阪奈データ (N=15)	13.7 (4.7)	12.3 (5.4)	4.9 (1.6)	5.1 (2.9)	6.2 (2.7)	4.6 (2.3)	3.2 (1.5)
ウェブデータ (N=20)	10.3 (5.0)	16.4 (6.9)	4.4 (1.5)	5.2 (2.6)	5.7 (2.1)	5.4 (3.2)	2.8 (1.3)
計	12.0 (5.1)	14.4 (6.6)	4.7 (1.5)	5.1 (2.7)	5.9 (2.3)	5.0 (2.8)	3.0 (1.4)

次に、初期語彙発達の特徴を知る際に使用されることの多い「NV比(noun-verb ratio)」を指標として、各データ間の意味カテゴリ分布の違いを検証した。このNV比は、語彙の中心的な役割を担う名詞と動詞に注目し、「名詞数÷(名詞数+動詞数)」の式で算出される。NV比が0.55以上の場合は名詞優位型、0.45-0.54の場合はバランス型、0.44以下は動詞優位型と分類され、個人差や言語間の特徴を知る際に用いられることが多い(Miyata ら, 2004)。こうしたNV比に対して両データ間でt検定を行った結果、有意な差は見られず、両データともに名詞優位型の値を示した(京阪奈:  $M = .69$  [ $SD .13$ ], ウェブ:  $M = .75$  [ $SD .13$ ])。

[ $SD .15$ ];  $t [33] = 1.332, p = .192$ , two-tailed)。この NV 比の結果から、ウェブデータは、意味カテゴリ分布の点においても、研究者の指示下で収集したデータと相違ないことが示唆される。

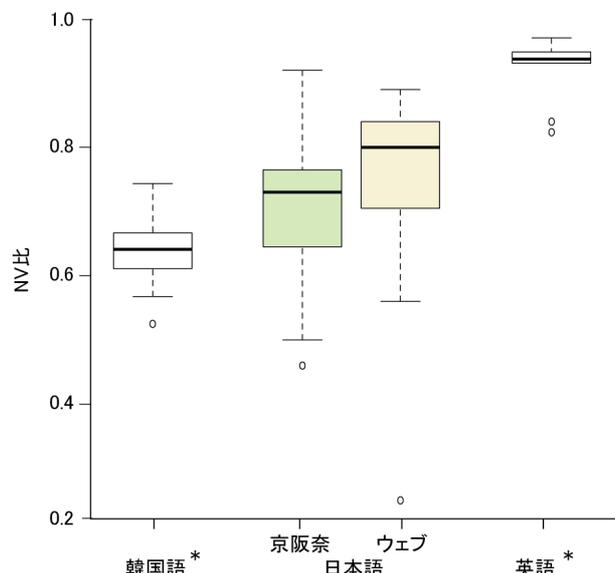


図4. ウェブデータと京阪奈のデータにおけるNV比の比較(両者に有意差なし)。なお、Kimら(2000)の韓国語と英語のNV比を参考のため追記。

#### 4. 考察

本研究では、ウェブユーザから収集した幼児語彙発達データがどの程度信頼性があるかを検証するために、研究者の指示下で親が1年間記録した語彙発達データとの比較を試みた。その結果、語彙の獲得速度と意味カテゴリ分布の点において両データ間に有意な差がないことがわかった。これは、ウェブユーザが自発的に無報酬でウェブへ入力したデータであっても、研究者の指示下で収集したデータと同等の信頼性を持ち、ユーザの多くが正直かつ勤勉にデータを入力してくれたことを示唆している。

一般に、ウェブユーザの投稿・評価したデータがどの程度信頼できるかについては重要な問題でありながら、実際に検証を行った事例は少ない(Callison-Burch, 2009)。本研究は、幼児語彙発達という特定領域のデータではあるけれども、ウェブデータの信頼性検証に成功したという点で重要な試みだったと考えている。今回、ウェブデータの信頼性が高かったのは、わが子の成長記録という課題自体が重要な要因だったのではないかと考えている。というのも、わが子の記録を不正確に行ってもユーザには何の意味もないため、ユーザ個人にとって重要な事項であればあるほど、そのデータの信頼性は自ずと高まるのかもしれない。

本研究でウェブ日誌法によるデータの信頼性を検証できたことにより、今後の展開において2つの重要な意義をもたらすことができたと考えている。1つは、ウェブから得たこの縦断データを、幼児言語発達の基礎データとして使用可能であることを実証した点にある。今後、ウェブ日誌法でさらにデータを収集できれば、その大規模縦断データから語彙獲得メカニズムの新しい発見が可能となるかもしれない。もう1つの意義は、無報酬であっても語彙発達データを正直にかつ勤勉に入力してくれるユーザが数多く存在することから、こうした枠組みを、こどもの発達を見守るウェブシステム等に適用できる可能性があるという点である。言語能力は、幼児における発達全般の有効な指標でもあるので、ウェブ上での語彙発達記録から発達の遅れや障がい等を早期発見・療育することが今後可能となるかもしれない。こうした基礎及び応用研究の視点を考慮しつつ、今後も、幼児の言語発達研究を進めていきたい。

#### 文献

- 天野成昭・近藤公久・加藤和美 (2008). NTT 乳幼児音声データベース (INFANT). 国立情報学研究所 音声資源コンソーシアム.
- Caselli, M.C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weir, J. (1995). A cross-linguistic study of early lexical development. *Cognitive Development*, 10, 159-199.
- Challison-Burch, C. (2009). Fast, cheap and creative: evaluating translation quality using Amazon's Mechanical Turk. *Proc. of the 2009 Conference on EMNLP*, 286-295.
- Fenson, L., Dale, P.S., Reznick, J.S., Bates, E., Thal, D., & Pethick, S. (1994). Variability in early communicative development. *Monographs of SRCD*, 59(5), No.242.
- Kim, M., McGregor, K.K., & Thompson, C.K. (2000). Early lexical development in English- and Korean-speaking children: language-general and language-specific patterns. *Journal of Child Language*, 27, 225-254.
- 小林哲生・永田昌明(2008).日本語を母語とする幼児の初期語彙発達:ウェブ日誌法による早期出現語彙の特定. 第72回日本心理学会発表論文集, 1122.
- 小林哲生・永田昌明(2009).ウェブを用いた幼児言語発達研究:大規模縦断データの収集. 言語処理学会第15回年次大会発表論文集, 534-537.
- Kobayashi, T., & Nagata, M. (under review). Early lexical development in Japanese children: Age of acquisition of first 50 words specified using web diary method.
- MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Miyata, S., Oshima-Takane, Y., & Nishisawa, H. (2004). Noun bias in early Japanese vocabulary and characteristics of maternal input. *Studies in Language Sciences*, 4, 87-101.
- 小椋たみ子・綿巻徹(2008).日本の子どもの語彙発達の規準研究:日本語マッカーサー乳幼児言語発達質問紙から. 発達・療育研究, 24, 3-42.

(Tessei Kobayashi & Maasaki Nagata)