語彙獲得の月齢分布に基づく語彙学習機構の推定
Estimation of Learning Process based on Distribution of Age of Acquisition of Words

日高　昇平[†]
Shohei Hidaka

[†]Indiana University
shhidaka@indiana.edu

Abstract

What determines vocabulary growth patterns? The research presented here examines the growth pattern of words listed in the McArthur-Bates Communicative Development Inventory using a computational model. Our model characterizes vocabulary growth curves based on the sampling of learning relevant events and a threshold (the number of such events needed) for acquisition of the word. Using this general class of models, fits of vocabulary growth curves suggests a transition from one in which acquisition is primarily limited by the threshold for acquisition to one in which acquisition is primarily limited by sampling speed. Further analyses suggest that these parameters of the learning model link to meaningful psychological factors: specifically the acquisition of threshold limited (and earlier learned) words are correlated with frequency whereas sampling-speed-limited words are correlated with imageability of the word in the input.

Keywords — Vocabulary growth, Age of Acquisition, Learning based on accumulation and acceleration.

1. はじめに

言語発達の研究において、語彙獲得の月齢を予測することは一つの目標となっている．しかし、獲得月齢の大きな個人差のため[1]、それは容易ではない．乳幼児の獲得単語を調べる主な方法として、質問紙を用いた養護者による報告が用いられる．MCDI [5]は30ヶ月児が獲得(1語以上の発話として定義)する典型的な単語を網羅した標準的なリストである．MCDIを用いた研究によれば、最初の50語を、最も早く獲得する幼児集団(上位10%)と最も遅く獲得する集団(下位10%)には12ヶ月以上もの獲得月齢の違いがある[6]．この大きな個人差にも関わらず、語彙発達に関する多くの先行研究では、主に集団分布の代表点を扱ってきた(e.g., 50%以上の幼児が獲得した最初の月齢)．これに対し本研究では、代表点のみならず乳幼児集団の獲得月齢分布から、語彙学習機構の推定を試みる．もし全ての幼児が近似的に同じ学習機構を経由し、ある単語を獲得するならば、語彙獲得の月齢分布はその学習機構の一側面を反映するはずである．具体的には、次に述べる学習モデルを提案し、乳幼児の語彙獲得分布に対するモデルの予測性を検討した．

2. 学習モデルから獲得月齢の分布へ

本研究のモデルでは、幼児がある１単語に関連する「事象X」を観察し、その観察の累積回数が一定値に達したときに、幼児は単語を獲得(発語)すると仮定する(詳細は付録を参照)．また事象Xは月齢によらず一定頻度で確率的に観測されるか、月齢の関数として観察頻度が増加すると仮定する．つまり、このモデルは事象Xの累積数N・観測頻度の増加率Dで記述できる．以下ではこれを累積・加速モデルと呼ぶ．このモデルは、事象Xの内容に拠らず、強い一般性・抽象性を持つ一方で、月齢分布の形状に関して定量的な予測ができる．事例Xの特定に関しては、分析結果に基づき後に議論する．累積・加速モデルは2つの特徴的な場合を含む．1つは、事象Xの必要累積数がN(＞１)で、観測頻度が一定(D=1)である場合である(図1a)．この確率過程の結果、獲得月齢はガンマ分布に従う(累積モデル)．もう一つは、事象Xの必要累積数が1 (N=1)で、観測頻度が増加する場合(D>1)である(図1b)．この確率過程の結果、獲得月齢はワイブル分布に従う(加速モデル)．累積・加速モ

デル(族)は、この二つを下位に含み、事例Xの必要累積数Nでかつ観察頻度が増加する場合の獲得月齢分布のモデルにあたる(ワイブル・ガンマ分布).

従来の語彙成長曲線の研究の多くではロジスティックモデルが用いられてきた[10]. 提案モデルとロジスティックモデルとの違いを視覚的に示すために、ガンマ分布、ワイブル分布、ワイブル・ガンマ分布および、ロジスティック分布の累積確率分布とハザード関数を図２に示した. ハザード関数とは、この場合、ある時点で、まだある単語を学習していない幼児のうち、その次の瞬間にその単語を獲得する幼児の割合を示したものである. ４つの累積確率分布の概形は、どれも“S字”であり、一見すると良く類似しているが、そのハザード関数に本質的な違いが表れている. ガンマ分布では上に凸、ワイブル分布では下に凸、ロジスティック分布は変曲点を持つ単調増加関数であり、またワイブル・ガンマ分布はある条件では単峰性の関数となる. 従って分布の形状によって、新たにある単語を学習する幼児の割合が、初期に増加(ガンマ分布)、後期に増加(ワイブル)、ある時点を境に増加(ロジスティック分布)、増加の後に減少(ワイブル・ガンマ分布)という学習特性を捉える事ができる.

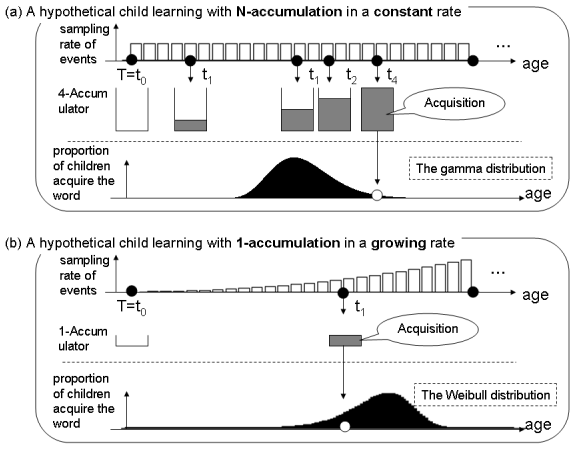

図１：(a)事象Xがある一定頻度で確率的に観察され(上段)、それが累積４回に達したときに単語を獲得する場合(中段、N=4). 獲得月齢はガンマ分布に従う(下段). (b)事象 X の観測頻度が月齢を追って高くなり(上段, D>1)、一度の観察で単語が獲得される場合(中段). 獲得月齢はワイブル分布に従う(下段).

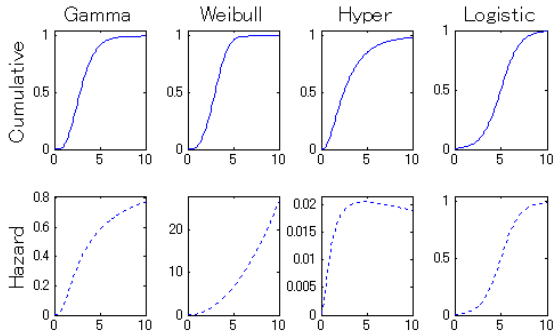

図 2: ワイブル・ガンマモデルおよびロジスティックモデルの累積確率分布とハザード関数. 全て同一の平均と分散を持つ.

3. 分析方法

MCDI(Lex2005)には、654 語に関して、16 から 30 ヶ月児の 1 ヶ月ごと 15 点の獲得割合が公開されている. MCDI は名詞・動詞・形容詞・代名詞・関係代名詞・前置詞・その下位分類など 21 の語彙カテゴリを含む. 各単語、16 から 30 ヶ月児の月齢分布に対し、上記 3 つの累積・加速モデル族(ガンマ・ワイブル・累積加速モデル)および、先行研究の多くで語彙獲得曲線の記述に用いられるロジスティックモデル[10]の適合性(BIC[9])を分析した.

4. 結果・考察

モデル選択の結果、654語のうち88%の単語に対して、3つの累積・加速モデル族のいずれかが適合し、ロジスティックモデルはわずか12%の単語に適合した. 従って、ロジスティックモデルを以下の分析から除外した. 累積加速(上位)モデルには最多45%の単語が適合した. これは多くの単語の獲得が累積・加速両方の特性を持つ事を意味するが、単語の学習がより累積的または加速的かを調べるため、両極であるガンマ・ワイブル分布に適合する単語の割合と、平均の獲得割合に注目し、21の語彙カテゴリを分析した(図2). 名詞(vehicle, animals, etc.)や動詞(action words)は獲得が早く、その多くの単語はよりガンマ分布に適合した. 一方、question words などの機能語は獲得が遅く、その多くの単語はより

ワイブル分布に適合した．つまり、語彙カテゴリの平均的な獲得時期に相関して、分布形状が系統的に変化する事を意味している(R=-0.67)．モデルの仮定に沿って考察すると、この結果は、初期の語彙獲得がある種の経験の累積によって記述でき、また後期には学習速度の増加によって記述できる事を示唆する．

次に、モデルの仮定である「事象 X」を特定するため、複数のデータベース([2], [3],[7],[8])に含まれる単語の心理的要因(frequency, familiarity, imageability, etc.)を用いて、月齢分布から推定したモデルのパラメタとの相関を調べた．その結果、ガンマモデルの観察頻度と単語の出現頻度[7]とが有意な相関を示し、一方、ワイブルモデルの観察頻度の増加率と単語の心象性(imageability)獲得月齢間の統計的従属性とが有意な相関を示した．つまり、初期の学習において事象 X は発話単語の観察であり、その累積が単語獲得に関連する、後期の学習において事象 X は独立した単語の観察ではなく、複数単語の関係性・文脈である事が示唆される．

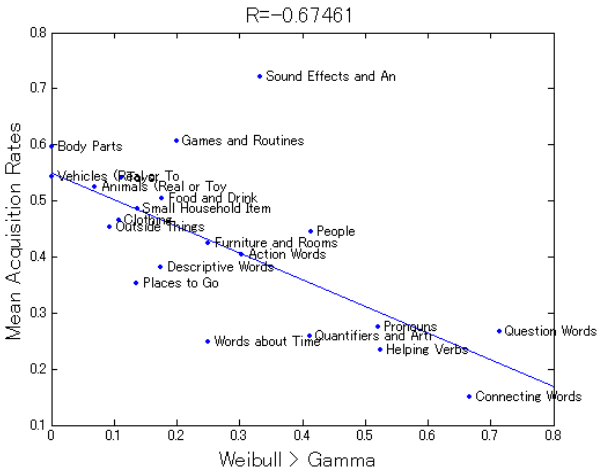

図2: 語彙カテゴリ内のワイブルモデルに適合した単語の割合(X軸)と平均の16-30ヶ月児の獲得割合(Y軸). 平均獲得割合が高いほど単語獲得が早い事を意味する．

謝辞

This study was supported by grants from NIH MH60200.

参考文献

[1] Bates, E., Dale, P., & Thal, D. (1995). Individual differences and their implications for theories of language development. In P. Fletcher & B. MacWhinney (Eds.), *Handbook of child language* (p. 96-151). Oxford: Basil Blackwell.

[2]Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33A*, 497-505.

[3]Cortese, M. J., & Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behavior Research Methods, Instruments, & Computers*, *36* (3), 384-387.

[4] Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, *28* (1), 125-127.

[5] Fenson, L., Dale, P., Reznick, J. S., Bate, E., Hartung, J., Pethick, S., et al. (1993). *Macarthur communicative development inventories*. San Diego: CA: Singular Publishing.

[6] Jahn-Samilo, J., Goodman, J., Bates, E. & Sweet, M.(2000) Vocabulary Learning in Children from 8 to 30 Months of age: A Comparison of Parental Reports and Laboratory Measures. *Technical Report CND-0006*, *Project in Cognitive and Neural Development Center for Research in Language University of California*, San Diego.

[7] MacWhinney, B., & Snow, C. (1990). The child language data exchange system: An update. *Journal of Child Language*, *17* (2), 457-472.

[8] Miller, G. A. (1995). WordNet: An on-line lexical database for English. *Communications of the ACM*, *38* (11), 39-41.

[9] Schwarz, G. (1978). Estimating the dimension of a

model. *The Annals of Statistics*, *6* (2), 461-464.
[10] van Geert, P. (1998). A dynamic systems model of basic developmental mechanisms: Piaget, Vygotsky, and beyond. *Psychological Review*, *105* (4), 634-677.

付録：ワイブル・ガンマ分布の導出

この付録では、ワイブル・ガンマ分布の導出を示す．語彙の獲得がある一定数 N の事象 X の経験によって起こり、全ての事象の経験数 M のうち、事象 X が確率 f で発生すると仮定する．このとき、語彙を獲得する確率は、事象 X の経験数が N 以上である場合であり、これは以下のように累積二項分布で表される．

$$P(k \geq N) = \sum_{k=N}^{M} \Gamma(M+1)\Gamma(k+1)^{-1}\Gamma(M-k+1)^{-1} f^{k}(1-f)^{M-k}$$
$$= \Gamma(M+1)\Gamma(N)^{-1}\Gamma(M-N+1)^{-1}\int_{0}^{f} t^{N-1}(1-t)^{M-N}\,\mathrm{dt}$$

ただし、$\Gamma(N)$はガンマ関数であり、第二行目は、累積二項分布に等価な不完全ベータ分布への変換である．事象の経験数 M が十分に大きい場合、ベータ分布はガンマ分布によって近似できる．また、事例の発生確率 f を時間の多項式的な関数 $f(M)=(\delta^{-1}T)^{d}$ とみなすことで、以下のワイブル・ガンマ累積確率分布を得る．

$$P(T;\delta,N,d) = \Gamma(N)^{-1}\int_{0}^{x} t^{N-1}\exp(-t)\,\mathrm{dt}$$

ただし、$x=(\delta^{-1}T)^{d}$ また $\delta, N, d>0$. δは時間に不変な事象の基本頻度の逆数、d は頻度の時間に伴う増加・減少を表す指数、Nは獲得までに要する事象の累積数である．