偏った頻度分布はどこに宿るか?
―表層パターンの分布分析に基づく統語発達に関する一考察―

吉川 正人（慶應義塾大学）

1. はじめに

任意のテキストやコーパス内の単語の頻度分布がその出現順位に逆相関するという性質は，Zipf 則 (Zipf 1935, 1949) として広く知られている．この性質は任意の構文 (e.g, 二重目的語構文) における動詞 (e.g., *give*) の頻度分布にも見られ，それが言語習得に重要な役割を果たしているという指摘もある (e.g, Goldberg, Casenhiser, & Sethuraman 2004)．しかしながらこれらの研究で対象となっているのは [動詞 目的語 $_1$ 目的語 $_2$] (VOO) といった抽象的な構文パターン(所謂項構造構文: Goldberg 1995) であり，このような抽象度の高い構造が本当に言語習得の過程で幼児の学習に活用されているかどうかは疑問の余地がある．分布の知覚には生起環境の認識が不可欠であり，生起環境の認識には構文の学習が前提になることから，論理的にも矛盾を指摘できる．

本稿では，具体的な単語に根差したパターン(e.g, [give me a …]) がまず学習されそれらが徐々にまとめあげられることで抽象的な構文が獲得されるとする Tomasello (2003) の議論を受け，このような具体的なパターンにおける特定の生起位置に現れる単語の分布こそが重要であり，それが Zipf 則に従うようなものになっていることを示す．

2. Zipf 則と言語習得

2.1 Zipf 則とは

Zipf 則とは，提唱者の George Kingsly Zipf にちなんで名づけられた自然言語の統計的性質で，任意のテキストにおいて文字や単語の出現頻度がその出現頻度順位に対数反比例する性質である．数学的には，以下の式で表現される:

$$P_n \sim \frac{1}{n^a}$$

(ただし a は定数, P_n は n 位の単語の頻度; Zipf 1935, 1949)．基本的には $a=1$ であることが知られているため，頻度が [1/順位] に反比例するということになる．つまり，出現頻度 2 位の要素の頻度は 1 位の頻度の 1/2, 3 位であれば 1/3, …となる．

この時，頻度・順位の両変数に対数を取り散布図を描くと，分布は線形になる．Figure 1 はアメリカ英語の中規模均衡コーパスである Brown コーパスの単語出現頻度を，縦軸を頻度，横軸を頻度順位としてプロットした散布図である．頻度トップ 10 の単語と頻度は Table 1 の通りである．

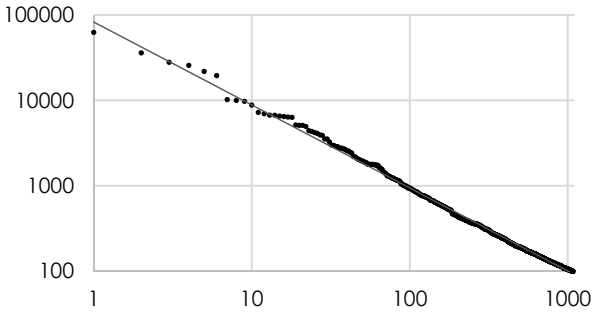

Figure 1. Brown コーパス単語出現頻度 (ただし頻度 ≥ 100) と順位の相関 (対数, 直線は回帰直線)

Table 1. Brown コーパス単語出現頻度トップ 10

順位	単語	頻度
1	the	62713
2	of	36080
3	and	27915
4	to	25732
5	a	21881
6	in	19536
7	that	10237
8	is	10011
9	was	9777
10	for	8841

2.2 Zipf 分布と言語習得の関係

Zipf 則に従う分布 (以降 Zipf 分布と呼ぶ) を示すのはテキスト内の要素の頻度分布だけではなく，任意の構文における動詞の頻度分布にも同様の性質がみられ，そのような分布をなしていることが構文の習得にとって重要な役割果たしているという指摘もある．

例えば Goldberg et al. (2004) では，Bates コーパス (Bates, Bretherton, & Snyder 1988) における幼児に向けられた母親の発話を対象に，1) 自動詞移動構文 (VL [(1a)])，2) 使役移動構文 (VOL [(1b)])，3) 二重目的語構文 (VOO [(1c)]) の実例を収集し，それぞれどのような動詞が何回使用されたのかカウントを行った結果，Table 2 に示すような分布の偏りが確認されたことが報告されている．最上位の動詞が他の追従を許さないほど極めて頻度が高く，総数にすると 20%以上を占めている．

(1) a. Verb Oblique [VL] (e.g, *I went to the store.*)
b. Verb Object Oblique [VOL] (e.g, *Marty put the milk in the fridge.*)

c. Verb Object Object$_2$ [VOO] (e.g., *Pat gave Chris a book.*)

Table 2. Bates コーパスにおける当該構文で母親が使用した動詞 (Goldberg et al 2004: 298, Table 5 より一部改編)

構文	最高頻度の動詞	相対頻度 (絶対頻度/総数)	使用された動詞の異なり数
VL	go	39% (136/353)	39
VOL	put	38% (99/256)	43
VOO	give	20% (11/54)	13

また Casenhiser & Goldberg (2005) では, 構文の習得には「偏った分布 (skewed input)」が重要であるということが新奇構文の習得実験で示されている．実験では，英語話話者の幼児 (5-7 歳) を対象に，[Subject Object Verb] という標準的な英語には存在しない文法構造に対し「ある対象 (Subject) がある場所 (Object) に出現する」という意味を対応付けた新奇な構文を作成し，その習得に「偏った分布」が有益か否か検証された．具体的には，新奇な動詞 (e.g., anoopo) を用いたこの構文の実例 (e.g., *The rabbit the hat anoopoed.*) を，「出現」場面 (e.g.,「ウサギが帽子の上に現れる」) を示した動画と共に音声で提示し，5 種類の新奇動詞 (anoope, vako, suto, keebo, fego) の頻度分布が比較的均等である条件 (「均等頻度条件」, それぞれ頻度が 1, 1, 2, 2, 2 [順不同]) と一つの動詞だけが極端に頻度が高い条件 (「偏り条件」, それぞれ頻度が 4, 1, 1, 1, 1 [順不同]) とで，この構文，即ち [Subject Object Verb] という構造と「出現」という意味とのペアの学習成績に有意な差が生じるかどうかを検証した．結果，後者の「偏り条件」で優位に学習成績が高いことが確認され，Zipf 分布に見られるような一つの要素が極端に高い頻度を示す頻度分布が文法的な構文のような抽象的なパターンの学習にとっても有益である可能性が示された．

また母語だけでなく第二言語習得においても同様の議論があり (e.g, Ellis & Ferreira-Junior 2009)，言語に対するヒトの学習全般に対しての有効性が見て取れる．

2.3 量知覚の対数的性質

Zipf 分布，或いはより一般的に，偏った分布がヒトの言語習得にとって有意義であるとすれば，そこにはヒトの認知に関する何らかの数量的な原理が存在すると考えられる．その最も有力な候補は，所謂 Weber-Fechner 則として知られる，ヒトの量知覚における対数的性質である．

ヒトはほぼありとあらゆる量の知覚において，物理量をそのまま知覚するわけではないことが知られている．例えば，物体の重量を二倍にしても，その重さの知覚は二倍にならないし，液体の塩分濃度を二倍にしても，塩辛さの知覚は二倍にはならない (Vershney & Sun 2013: 28)．また，数の知覚に関しても，1 と 2 の差は大きく感じられるが，同じ 1 の差であっても，10 と 11 の差はあまり大きく感じられないだろう．

一般に物理量と知覚量の関係は以下の式で表現されるような関係にあるとされている．これが Weber-Fechner 則である (Vershney & Sun 2013: 30):

$$P = K \log \frac{S}{S_0}$$

(ただし P は知覚量，S は物理量，S_0 は知覚可能な最小の知覚量，K は定数)．従って，知覚量は物理量に対して対数的であるということになる．

単語等言語要素の頻度も「何回その要素に出くわしたか」という数量の知覚であり，Weber-Fechner 則が当てはまると考えられるため，頻度の知覚は頻度の対数で近似が可能であり，頻度の差を知覚するには，その差は対数的でなくてはならないと考えられる．従って，Zipf 分布は，任意の環境 (e.g, 同一テキスト，構文の特定の生起位置) における出現要素の頻度の差を知覚するのに極めて有益な性質であると言える．何故だか分からないが，言語要素の頻度分布は，ヒトの知覚にとって都合のいいようにできているのである．

2.4 先行研究の問題点とその解決案

2.2 節で提示した知見と 2.3 節で提示したヒトの量知覚の対数的性質と合わせて考えると，確かに言語習得において Zipf 分布をなす要素の頻度分布が重要であることは伺える．しかしながら，2.2 節で提示した先行研究の議論には少なからず疑問の余地がある．

当該研究では，構文 C (e.g, 二重目的語構文) の習得にはその構文に生起する動詞群 V (= {*give*, *tell*, *ask*, *send*, …}) の「偏った頻度分布」が有効であるとしているが，そもそも，V は「C に生起する動詞」であって，V の頻度分布を知覚するためには C の認識が前提になるため，習得の効率化に習得の対象そのものを利用してしまうことになる．例えて言うなら,「インターネットのつなぎ方をインターネットで調べる」ような状況となってしまっている．

この問題を解決するには，習得の対象となる抽象的な構造 (e.g, [VOO]) ではなく，それを体現すると思われるより具体的で知覚可能 (と思われる) 生起環境を想定し，その環境下において動詞の分布が「偏った」ものであることを示す必要がある．

そこで本稿では，具体的な単語に根差したパターン(e.g, [give me a …]) がまず学習されそれらが徐々にまとめあげられることで抽象的な構文が獲得されるとする Tomasello (2003) の議論を受け，このような具体的なパターンにおける特定の生起位置に現れる単語の分布こそが重要であり，それが Zipf 則に従うようなものになっていることを示す．以下で，調査の概要と結果を提示する．

3. 調査

3.1 データ

データには，上述の先行研究 (Goldberg et al. 2004) で使用されたものと同様，幼児と周囲の大人の対話データベースである Childes (MacWhinney 2000) に含まれている Bates コーパス (Bates et al. 1988) を使用した．このコーパスは 20 ヶ月児 27 名，28 ヶ月児 89 名分のデータが含まれる，月齢を統一した横断データである．

3.2 方法

調査方法は以下の通りである: Goldberg et al. (2004) で対象となっていた 3 つの構文 ((2) に再掲) を体現すると思われる 3 つのパターン (3) を対象に，正規表現検索を用いて動詞生起位置に出現する要素の一覧を取得し，そこから頻度分布表を作成した．[1]

(2) a. Verb Oblique [VL] (e.g., *I went to the store.*)
b. Verb Object Oblique [VOL] (e.g., *Marty put the milk in the fridge.*)
c. Verb Object Object$_2$ [VOO] (e.g., *Pat gave Chris a book.*)

(3) a. [X {into, to} {a, an, the}]
b. [X {a, an, the} … {in, into, on}]
c. [X {me, us, him, her, them} {a, an, the}]

つまり，“X” を動詞出現位置として，a. 直後に前置詞 into, to が生起しその後ろに冠詞が生起するパターン，b. 直後に冠詞と何らかの要素が生起しその後ろに前置詞 in, into, on が生起するパターン，c. 直後に代名詞 me, us, him, her, them が生起しその後ろに冠詞が生起するパターン，ということである．尚，これらのパターンは，半ば恣意的ではあるが，事前に判明している当該構文の性質と，予備調査を通して探索的に得られた知見とに基づいて選定している．

幼児が実際に得ているインプットが調査の対象となるため，コーパスに含まれる幼児自身の発話を取り除き，周囲の大人の発話のみを分析に用いた．また，同様の理由から，頻度分布の測定には，コーパスに現れている表層形をそのまま用い，レンマ化を行わなかった．

3.3 結果

調査の結果を以下にまとめる．まず [X {into, to} {a, an, the}] パターンであるが，観察された X の値は Table 3 の通りである (頻度 2 以上)．予想通り最上の go が極めて高頻度で生起しており，次点より二倍以上の生起頻度となっているため，Zipf 分布をなしているものと考えられる．2 位の goodnight や 3 位の out など，動詞ではないものも混在しているが，分布に影響を与えるものではないためここでは問題としない．観察された事例を以下に提示する:

(4) a. does the man want to go to the store ?
b. then put the dollies in the car so that we can drive to the picnic .
c. do you want to go to the airport # yeah ?
d. then she ran into the snow .

Table 3. [X {into, to} {a, an, the}] の X に生起した要素

順位	X	頻度
1	go	11
2	goodnight	5
3	out	4
3	went	4
5	ran	3
5	say	3
5	going	3
8	back	2
8	said	2
8	up	2

続いて [X {a, an, the} … {in, into, on}] パターンであるが，観察された X の値は Table 4 の通りである (頻度 2 以上)．こちらも予想通り最上位の put が 2 位以降を大きく引き離す高頻度の分布となっており，極めて偏った分布であることが確認された．観察された事例を以下に提示する:

(5) a. put the people in the car .
b. let's bring the doggie in the house .
c. oh # you don't put the doggies on the table !
d. did Miffy find the bird in the snow ?

(6d) に関しては使役移動構文の事例ではないが，このような事例は稀であったため，結果に影響はないものと見做す．

Table 4. [X {a, an, the} … {in, into, on}] の X に生起した要素

順位	X	頻度
1	put	35
2	for	5
2	with	5
4	down	3
4	bring	3
6	find	2
6	#	2
6	got	2

最後に [X {me, us, him, her, them} {a, an, the}] パターンであるが，観察された X の値は Table 5 の通りである (頻度 2 以上)．このパターンに関しては最上位の tell が次点の give と拮抗する頻度となっており，またそれに次ぐ build ともほとんど頻度差がなく，Zipf 分布になっているとは言い難い頻度分布となった．

この結果は，Goldberg et al (2004: 298, Table 5 注) でも言及されているが，コーパスに含まれるデータの性質に起因するものと考えられる．というのも，Bates コーパスには自由

[1] 正規表現検索・頻度集計にはスクリプト言語 Python (Ver. 2.6.5) を用いて作成した自作のスクリプトを使用した．

会話と読み聞かせデータ，タスク中の会話とが混在しており，例外的な結果に影響を与えていたのは全て読み聞かせデータに含まれる発話であることが確認された．読み聞かせデータを取り除くと，結果はTable 6 のようになり，確かにX に生起する要素の頻度分布はZipf 則に従うことが確認できた．

Table 5. [X {me, us, him, her, them} {a, an, the}] のX に生起した要素

順位	X	頻度
1	tell	12
2	give	10
3	build	9
4	read	4
5	show	3
5	get	3
7	is	2
7	make	2
7	turn	2
7	gave	2

Table 6. [X {me, us, him, her, them} {a, an, the}] のX に生起した要素 (読み聞かせデータ以外)

順位	X	頻度
1	give	9
2	build	3
2	show	3
4	is	2

最後に，観察された事例を以下に示す (読み聞かせデータも含む):

(6) a. okay # can you give them a ride ?
b. get some of those and build me a tower .
c. can you show me a green one ?
d. can you read me the book ?

4. 考察と課題

前節の結果から，先行研究で検討された抽象的な構文を典型的に体現すると思われるパターンにおいて，動詞の頻度分布がZipf 則か，それに準じる「偏った分布」を示すことが確認された．

このことは，構文の習得には今回調査対象としたような語彙の指定された比較的具体的なパターンが活用され，少なくとも初期段階においてはそこに際立って高い頻度で生起する動詞を含む事例をそのパターンが持つ典型的な意味として習得される，という習得の実態を示唆している．

ただ，今回調査対象としたパターンは対象とする構文の全貌を明らかにする上では極めて限定的であるということには注意が必要である．特にVL 構文では，Table 2 にあるようにGoldberg et al. (2004) の調査ではgo が136 回生起している一方，本稿の調査では11 回と極めて少ない．レンマ化の影響を考えても，過半数が調査対象外となってしまっていることは確かで，方法論の改善が必要と言える．

5. 結語に変えて: Zipf 分布の社会的意義

頻度分布がZipf 分布をなすということは，単に知覚上の優位性を持つだけにとどまらない利点を持つと考えられる．Varshney & Sun (2013) でも指摘されているが，ヒトの量知覚の対数性は「エラーを最小限にする」という利点がある．つまり，数量が大きくなればなるほど「異なる」と感じるための絶対量が拡大するため，裏を返せば，よほどの差がない限り「異ならない」と感じられるようになり，少々の差であれば「誤差」で済むようになる．

このことは，個人のレベルにとどまらず，社会的にも重要な意義を持つように思われる．例えばある人物A がある表現X に100 回遭遇し，別の人物B は110 回遭遇したとする．頻度の絶対量の差は10 であるが，対数をとれば両者の差はほんの僅かである．従って，量知覚の対数的性質は「個人差の最小化」に寄与すると言える．ヒトの言語習得が経験に基づく統計学習的な性質を持つとすると常に問題となる「頻度経験の個人差」の問題が，この性質によって大いに回収される可能性があるということである．

この意味するところは大きく，ヒトの知覚・認知と社会，ならびに言語の統計的性質とを結びつけるような議論に展開することが可能であり，今後もっと注目されるていくべき特徴づけであると考えられる．

参考文献

Bates, E., Bretherton, I., & Snyder, L. S. 1988. *From first words to grammar: Individual differences and dissociable mechanisms*. Cambridge: Cambridge University Press.

Casenhiser, D., & Goldberg, A. E. 2005. Fast mapping between a phrasal form and meaning. *Developmental Science, 8*, 500–508.

Ellis, N. C. & Ferreira-Junior, F. 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics, 7,* 187–220.

Goldberg, A. E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago; London: University of Chicago Press.

Goldberg, A. E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. 2004. Learning argument structure generalizations. *Cognitive Linguistics, 15*, 289–316.

MacWhinney, B. 2000. *The CHILDES project: Tools for analyzing talk*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Tomasello, M. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA.: Harvard University Press.

Varshney, L. R. & Sun, J. Z. 2013. Why do we perceive logarithmically? *Significance*, *10*, 28–31.

Zipf, G. K. 1935. *The psycho-biology of language: An introduction to dynamic philology*. Boston: Houghton Mi□in.

Zipf, G. K. 1949. *Human behavior and the principle of least e□ort*. Boston: Addison-Wesley.