

幼児の統語発達モデル:
日本語, 英語, 中国語の言語構造を反映した統語範疇の獲得過程
A model for syntactic development of children:
Acquisition processes of syntactic categories reflecting structures of
Japanese, English, and Chinese languages

河合 祐司[†], 大嶋 悠司^{†*}, 笹本 勇輝[†], 長井 志江[†], 浅田 稔[†]
Yuji Kawai, Yuji Oshima, Yuki Sasamoto, Yukie Nagai, Minoru Asada

[†] 大阪大学大学院工学研究科
Graduate School of Engineering, Osaka University
yuji.kawai@ams.eng.osaka-u.ac.jp

Abstract

Children acquire syntactic categories reflecting structure of their native language. Five-year-old children can estimate a syntactic category of a novel word in a sentence, and map it to a perceptual target more readily than three-year-old children. This study proposes a computational model to explain the children's syntactic development and its language-dependency in the inference of a target indicated by a novel word. Our model estimates a syntactic category (a hidden state) of a novel word by a Bayesian hidden Markov model, and selects a target based on the category. Here, an increase of the number of the hidden states represents syntactic development: A small number of the hidden states results in an unclear estimation of the syntactic categories. The model with sufficient number of the hidden states almost correctly acquires differentiated categories. The model learned Japanese, English, or Chinese and reproduced the results of the target inferred by children from three to five years old. Our analysis revealed that the acquired syntactic categories were dependent on the learned language.

Keywords — Syntactic Development, Syntactic Category, Bayesian Hidden Markov Model

1. はじめに

幼児の統語発達の理解は, 人の言語処理メカニズムの解明へ向けた重要なアプローチとなるため, これまでに多くの観察実験やモデルが報告されてきた。人は生後およそ12~24ヶ月から多語文を産出し始め, その初期の多語文にも統語規則があることが報告されている [1]。その一方で, 文中での語の役割 (動詞など) を表すカテゴリである統語範疇は未発達であるとされる [2, 3]。そして, 36ヶ月ごろから徐々に精緻な統語範疇が獲得され

るようになる [3]。このように, この時期の幼児に統語範疇の発達があるとされているが, その発達メカニズムは未だ明らかでない。

幼児の統語発達を調査する一つの方法として, 統語的な手がかりにより新奇語の統語範疇を推定し, その語を正しく対象 (動作や物体など) に結びつけられるかどうかを試す, 名詞・動詞般用課題がある [3, 4]。まず, 女性が新奇な物体を用いて新奇な動作をしている標準刺激が幼児に提示される。この映像刺激と同時に, 日本語を母語とする幼児に対しては, 新奇語「ネケ」を

- 名詞条件: 「ネケがある」
- 動詞 (項省略) 条件: 「ネケってる」
- 動詞 (項明示) 条件: 「お姉さんが何かをネケってる」

のいずれかの文で与える。次に, 物体のみが標準刺激と異なる刺激 (物体変化刺激) と, 動作のみが異なる刺激 (動作変化刺激) が同時に幼児に提示される。そして, 新奇語「ネケ」がどちらの刺激に対応するか幼児は尋ねられ, どちらかを選択する。名詞条件の場合は動作変化刺激, 動詞条件の場合は物体変化刺激が正しい選択となる。このとき, 幼児は統語的な手がかりをもとに新奇語の指示対象を選択しなければならない。例えば, 幼児は新奇語の前後の語列から統語範疇 (名詞もしくは動詞) を推定し, 名詞は物体, 動詞は動作と対応するという知識を利用することで指示対象を正しく選択できる。

Imai et al. [3] は日本語, 英語, または中国語を母語とする幼児 (三歳, 五歳) に対し, この課題を実施した。その結果, 三歳児は母語に関わらず名詞般用可能だが, 動詞般用に失敗することがわかった。そして, 五歳になると母語に依存した動詞般用を示すことが明らかとなった。日本語は接尾辞で動詞と名詞の区別が可能であるため, 日本語児は項を省略しても動詞を般用できる。その一方で, 英語もしくは中国語を母語とする幼児は項

*2014年よりNTT

省略条件で動詞般用ができない。英語は項の省略が少ないこと、中国語では項を省略すると名詞と動詞の区別がつかないことが原因とされている。したがって、三歳から五歳にかけて母語の言語構造を反映した語の統語範疇が獲得されていくと考えられる。しかし、観察実験からその統語発達背後にある統語範疇構造の詳細を記述することは困難である。

そこで、内部構造の解析が可能な計算論モデルによる研究が期待される。単純なりカレントニューラルネットワーク(RNN)による単語予測学習によってその内部に統語範疇が表現されることが知られており、人の言語発達との対応が議論されている[5, 6]。Toyomura & Omori [7]はこのRNNの統語表現能力を用いて新奇語の指示対象を推定するモデルを提案した。三語文程度の簡単な英語コーパスがRNNに与えられ、その内部に統語範疇が表現される。その統語範疇が指示対象と連合されることによって、未知語の指示対象を推定できる。しかし、日本語のように語の省略や語順の変化のある複雑な言語構造を単純なRNNが学習することは困難である。それゆえ、このモデルはImai et al. [3]の示した母語に依存した統語発達を再現するには至っていない。

そこで本研究では、Imai et al. [3]が報告した三歳から五歳の統語理解の発達的变化とその言語構造依存性を再現可能なモデルを提案し、そのときのモデルの内部構造を解析することで、それらの現象を引き起こした統語範疇構造を明らかにすることを目的とする。ここで、ベイジアン隠れマルコフモデル(BHMM) [8]を語の統語範疇推定器として用い、このモデルの隠れ状態数を増加させることで統語発達を表現する。また、このモデルは統語を確率的に表現するため、語の省略や語順の入れ替えのように入力データの規則が複雑であっても学習できる。

2. 統語発達モデル

2.1 モデル概要

提案モデルのグラフィカルモデルを図1に示す。図下部のBHMMは語列 w を入力され、統語情報をもとに各語の隠れ状態 s を統語範疇として推定する。そして、 s は事物範疇 c (物体や動作といった抽象的な離散感覚空間)と、さらに c は指示対象 o と確率的に対応づけられている。ある語 w_i が既知の場合、 $P(o|w_i)$ の関係から o を直接推定できる(図1中のピンクの矢印)。しかし、 w_i が未知語の場合、それに対応する統語範疇 s_i を介した o の間接推定 $P(o|c)P(c|s_i)P(s_i|w_i)$ が必要である(図1中の青い

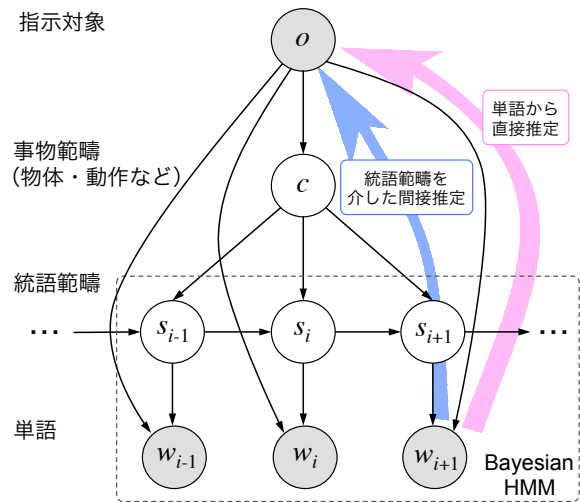


図1 視覚刺激から単語列を生成するグラフィカルモデル。今回の課題では、その生成過程を逆転させ、語からそれが指す対象を推定する。未知語の場合、指示対象の直接推定(ピンクの矢印)ができないため、統語範疇を介した指示対象の推定(青い矢印)が必要となる。

矢印)。したがって、後者の推定には s_i を正確に推定する必要がある。もし、 s の状態数 S が小さければ、一つの隠れ状態に複数の品詞が対応することとなり、例えば図2(a)のように名詞と動詞、助詞と接頭辞が混在するような統語範疇表現となる。その結果、新奇語 X から c と o を正確に推定することが困難となる。これに対し、 S が大きいと、ある隠れ状態が特定の品詞と対応することになり、 c と o の正確な推定が可能となる(図2(b))。この推定の特性を利用し、小さな S を持つモデルを三歳児、十分に大きい S を持つモデルを五歳児とみなして、統語発達を表現する。

2.2 指示対象の推定

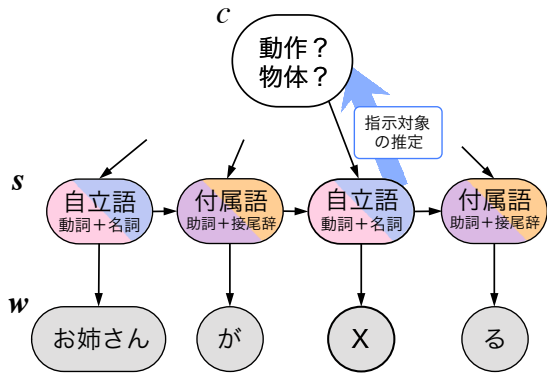
文中の単語 w_i から推定される指示対象 o は、次式で得られる。

$$P(o|w, s_{-i}) = \sum_{s_i} \sum_c P(o|c)P(c|s_i)P(s_i|w, s_{-i})P(o|w_i) \quad (1)$$

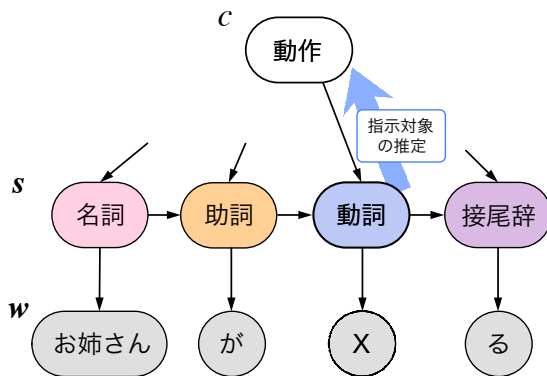
上式はベイズの定理より、

$$P(o|w, s_{-i}) \propto \sum_{s_i} \sum_c P(c|o)P(s_i|c)P(c)P(s_i|w, s_{-i})P(w_i|o)P(o) \quad (2)$$

で表される。ここで、右辺乗算の第一項は指示対象と事物範疇の対応であり、与えられるものとす



(a) 未分化な統語範疇による指示対象の推定



(b) 分化した統語範疇による指示対象の推定

図2 統語範疇の精緻化による統語発達。(a) 隠れ状態 s の状態数 S が小さいと未分化な統語範疇が獲得され、曖昧な指示対象の推定となる。(b) S が大きいと精緻な統語範疇が獲得され、明確な指示対象の推定が可能となる。

る。第二項は統語範疇と事物範疇の対応関係であり、次式で計算される。

$$P(s_i | c = c_j) = \frac{n(s_i, c_j)}{n(c_j)} \quad (3)$$

ここで、全ての入力中で s_i と c_j が同時に出現する回数を $n(s_i, c_j)$ 、また、 c_j が出現する回数を $n(c_j)$ とする。 c と o の事前確率を意味する式(2)の右辺第三項と第六項はそのとき提示された物体や動作に対して一様分布とする。式(2)の第四項は統語範疇の推定に相当し、BHMM [8] により求められる。また、式(2)の第五項は語と指示対象の対応、すなわち語彙であり、次式で与えられる。

$$P(w_i | o = o_t) = \frac{n(w_i, o_t)}{n(o_t)} \quad (4)$$

上式の $n(w_i, o_t)$ は学習データセット中で w_i と o_t を同時に観測する回数、 $n(o_t)$ は o_t を観察する回数である。ただし、新奇語の場合、この確率分布は一様とする。

まず、モデルはBHMMによって学習コーパス w の各語の統語範疇 s を推定、すなわち、式(2)の第四項の確率分布を計算する。この際、 s の状態数 S は設計者が与え、ギブスサンプリングによりこの確率分布を近似的に求める [8]。そして、統語範疇と事物範疇の対応(式(3))と、語と指示対象の対応の確率(式(4))を計算する。評価の際は、未知語 X を含む文をモデルに与え、式(2)によって X の指示対象を推定する。

3. 実験

3.1 実験設定

学習文として、日本語、英語、および中国語の各言語構造を有する人工コーパスを生成する。それぞれ名詞、動詞、形容詞などで構成され、接尾辞は分離されている(例、動詞:-てる, -ing, 形容詞:-い)。それぞれの言語のコーパスに含まれる文法構造とその割合は二歳から五歳の幼児に対する大人の発話コーパス [9, 10, 11] やコーパス解析研究 [12] をもとに決めた。その結果、日本語コーパスでは主語や目的語が省略されることが多く、英語コーパスと中国語コーパスではその省略が少なくなった。特に英語の「動詞-ing」単体文の割合は0.2%と非常に少ない。

また、指示対象として、文中の名詞、動詞、形容詞のそれぞれに対応するラベルがモデルに入力される。例えば、「お姉さんが赤い本を読んでいる」という文であれば、「お姉さん」「赤い」「本」「読んで」に対応する指示対象が与えられる。今回、各言語コーパスとその指示対象の10,000セットを用いてモデルを学習させた。

モデルが上述の各コーパスを学習した後に、Imai et al. [3, 4] の実験に倣って、コーパスに含まれない新奇語 X を以下の文でモデルに与える。

日本語

- 名詞条件「 X がある。」
- 動詞(項省略)条件「 X ってる。」
- 動詞(項省略)条件「お姉さんが何かを X ってる。」

英語

- 名詞条件「There is a X .」
- 動詞(項省略)条件「 X ing .」
- 動詞(項省略)条件「She is X ing it .」

中国語

- 名詞条件「有 X .」
- 動詞(項省略)条件「 X .」
- 動詞(項省略)条件「彼女(t) 在 X 某物 .」

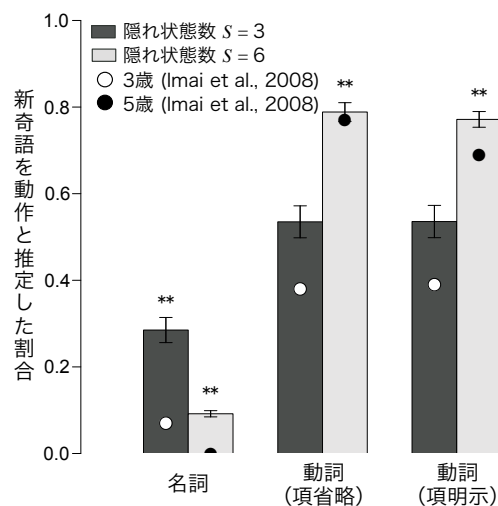
この文と同時に女性と形容詞(いずれも既知語)、新奇物体、および新奇動作に対応する四つの指示

対象をモデルに与える．そして，モデルは式(2)を用いてXから指示対象 o を推定する．なお，女性や形容詞に対応する対象が o として推定された場合，モデルは新奇物体と新奇動作を等確率で選択する．これは，名詞・動詞般用実験 [3, 4]で幼児が新奇語を物体や動作以外と推定していたとしても，物体・動作変化刺激のどちらかを選択することに対応する．各条件でXに対する o を推定する実験をBHMMの初期値を変えて20回実施した．

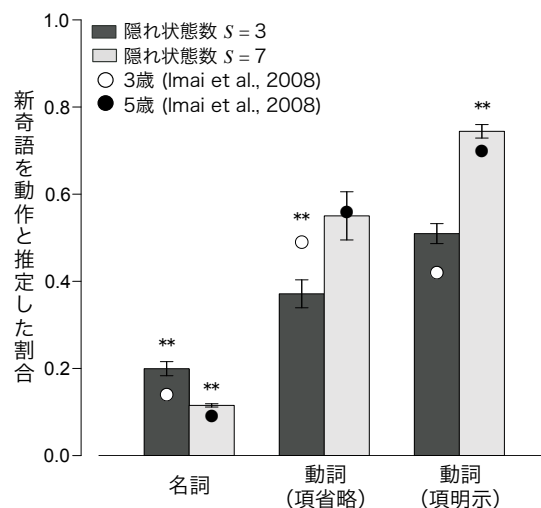
Xを新奇動作と推定する割合の平均値でモデルを評価した．一標本t検定により，その評価値がチャンスレベル(0.5)を有意に上回れば，モデルはXを新奇動作に対応付けたと結論付けられる．また，評価値が有意に下回れば，モデルは新奇物体に対応付けたとする．さらに，獲得された隠れ状態表現(統語範疇)の持つ品詞構造を解析する．学習コーパスの各語に品詞を割り当て，品詞ごとに式(2)の右辺第四項の和を計算する．そして，各隠れ状態についてその品詞の占める割合を求める．Imai et al. [3]の実験結果を再現できるように，隠れ状態数 S を三歳児モデルでは3, 五歳児モデルでは5~7に設定した．

3.2 実験結果

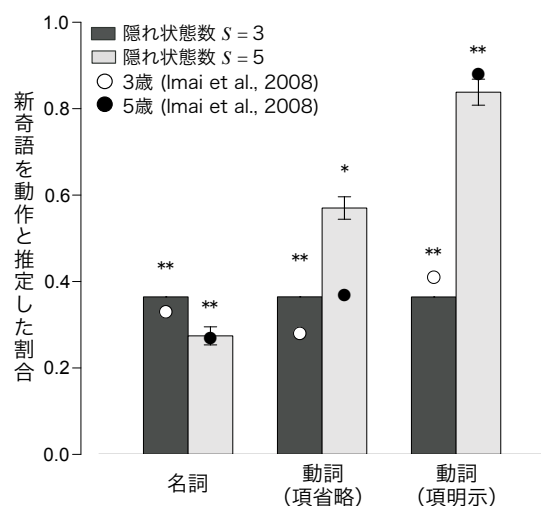
図3にそれぞれの言語で，新奇語を動作と推定した割合を棒グラフで示す．名詞条件ではチャンスレベル0.5に対して値が有意に小さく，動詞条件では値が有意に大きいと正しい般用といえる．また，参考のためImai et al. [3]の実験結果を点で描き加えた．これらの図より，提案モデルはImai et al. [3]の結果をよく再現していることがわかる．全ての条件で名詞般用に成功し(全て $p_s < .01$)， S が大きい場合には動詞般用に成功する条件があり，特に，項省略条件での動詞般用成績には言語間で差異がみられる．日本語入力の場合(図3(a))で $S = 6$ のとき，接尾辞で名詞と動詞を区別できるため，項明示・省略条件の両方で動詞般用可能である(どちらも $p_s < .01$)．その一方で，英語入力(図3(b))で $S = 7$ のときは項明示条件で動詞般用に成功している($p < .01$)が，項省略条件では失敗している($p > .05$)．これは学習コーパスに「動詞ing」単体文が少ないため，正しく統語範疇を推定できなかったためであると考えられる．中国語入力(図3(c))も $S = 5$ の項省略動詞条件で有意にチャンスより大きくなっている($p < .05$)ものの，項明示条件よりも小さな値となっており，幼児の実験結果の傾向を再現しているといえる．この傾向はそもそも中国語では項を省略すると名詞と動詞の区別がつかないためであると考えられる．な



(a) 日本語入力



(b) 英語入力



(c) 中国語入力

図3 未知語の指示対象の推定実験結果．エラーバーは標準誤差であり，記号はモデル推定値とチャンスレベル(0.5)とのt検定結果を示す(*: $p < .05$, **: $p < .01$)．点はImai et al. [3]の幼児の実験結果を示す．

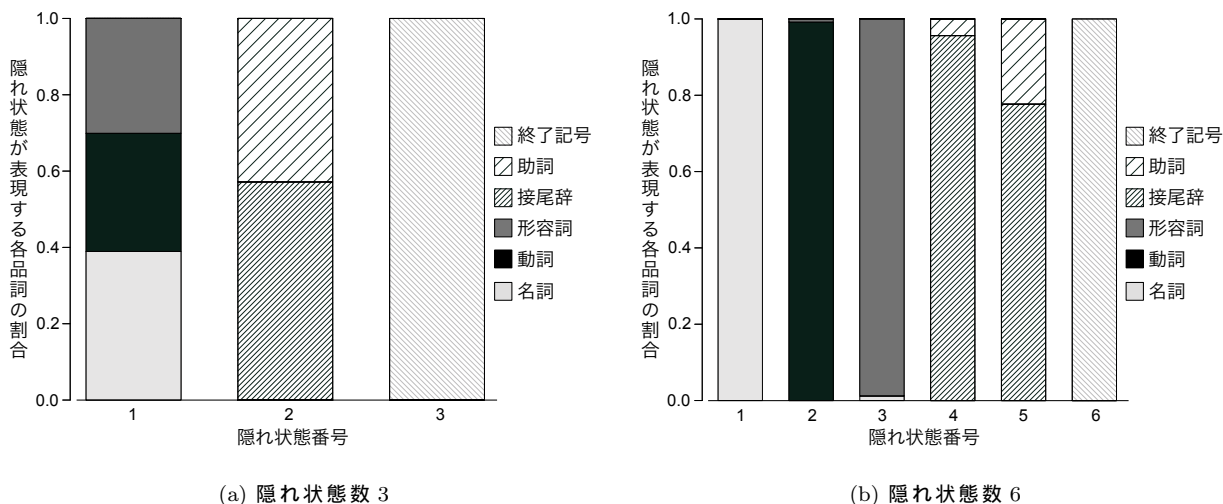


図 4 日本語入力に対する統語範疇の典型表現

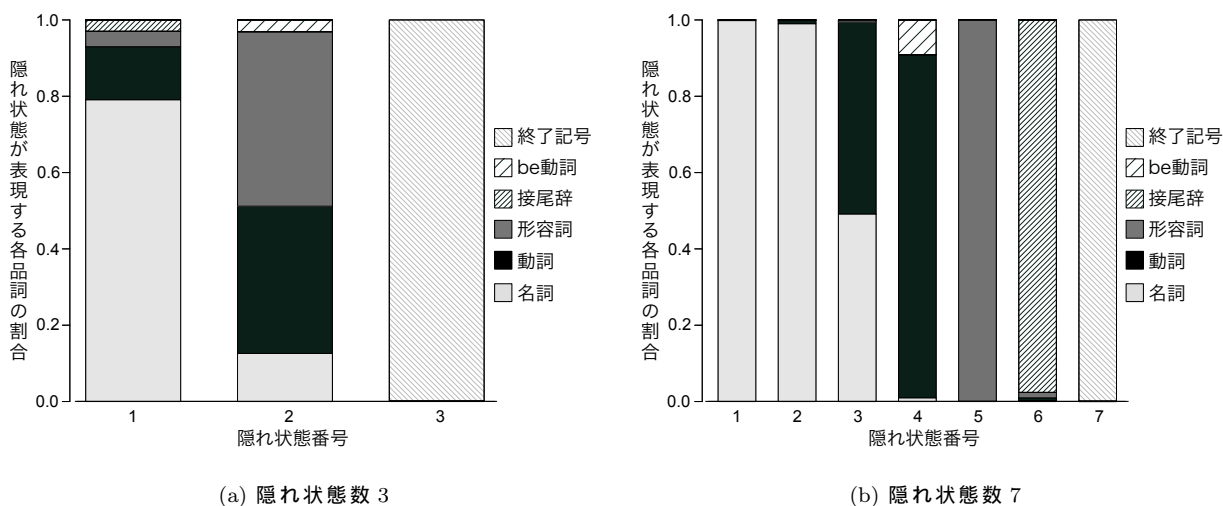


図 5 英語入力に対する統語範疇の典型表現

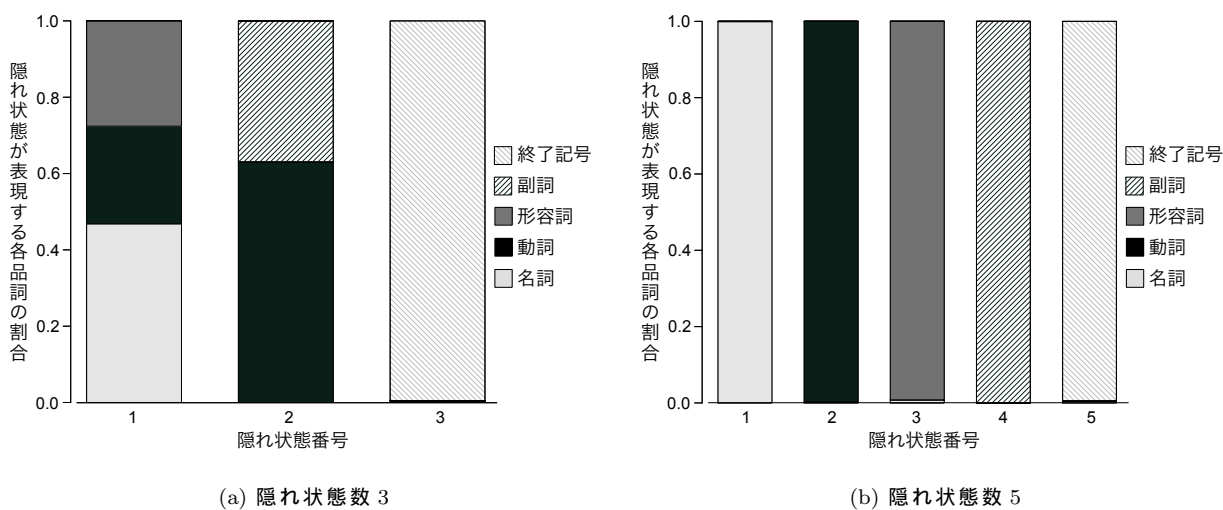


図 6 中国語入力に対する統語範疇表現の典型例

お、今回は五歳児の名詞・動詞般用課題成績を最もよく再現するSを選択し、言語間で異なるSとなったが、三つの言語のS = 7のモデルの結果でも上記と同様の傾向がある。

このときの各統語範疇が表現する品詞を図4-6に示す。まず、日本語の場合(図4), S = 3では名詞, 動詞, 形容詞が混同され, その結果, 動詞般用に失敗する。ただし, 名詞表現の割合が比較的大きく, また, 初期値によっては名詞が分離する場合があるため, 名詞般用は可能である。S = 6になると, 各品詞の表現が分離され, 動詞般用も成功するようになる。特に, 接尾辞(4, 5番)をさらに解析した結果, 助動詞と形容詞語尾が分離して表現されていることがわかり, これにより項が省略された場合にも動詞の推定が可能であったと考えられる。次に図5に示す英語の場合, S = 3であっても名詞が独立し(1番), それ以外の品詞が混同される(2番)。このような範疇の獲得により, 名詞般用には成功するが, 動詞般用に失敗する結果となったといえる。そして, S = 7になると, 各品詞の分離表現が得られる。名詞の範疇(1, 2番)のさらなる解析から, これらは主語の範疇とそれ以外の名詞の範疇に分離しており, 文頭に配置されやすい語といった語順に依存した統語範疇構造となっていることがわかった。その結果, 動詞単体で文が与えられると, 文頭の語は名詞という推定になり, 項省略動詞条件での動詞般用に失敗する。中国語入力の場合(図6)も同様に, Sが増加することによって, 各品詞表現が分離していく。しかし, 中国語は日本語のような名詞と動詞を区別する接尾辞表現を持たないため, 項が省略されると, 動詞般用ができない。

図7にSを2から6ないし7まで増加させたときに獲得した隠れ状態の表現を示す。提案モデルでは各Sの結果間の連続性は仮定されないが, それらの表現の包含関係から統語範疇表現間が線で結ばれている。図7(a)より, 日本語の場合はまず付属語と自立語が分離し, 付属語から助詞や助動詞が分化していく。したがって, 日本語は助詞や接尾辞に基づいた統語範疇の獲得過程となっていることがわかる。次に, 英語では名詞が主語(文頭の語)とそれ以外に分化するといった, 語順に依存した統語範疇獲得となっている(図7(b))。そして, 中国語ではまず副詞(在・正在)が分離する(図7(c))。同一の単語が動詞にも名詞にもなりえる中国語にとって, 副詞が重要な言語構造であるといえる。このように, 本モデルによって, 言語構造を反映した統語範疇の獲得過程を表現することができた。

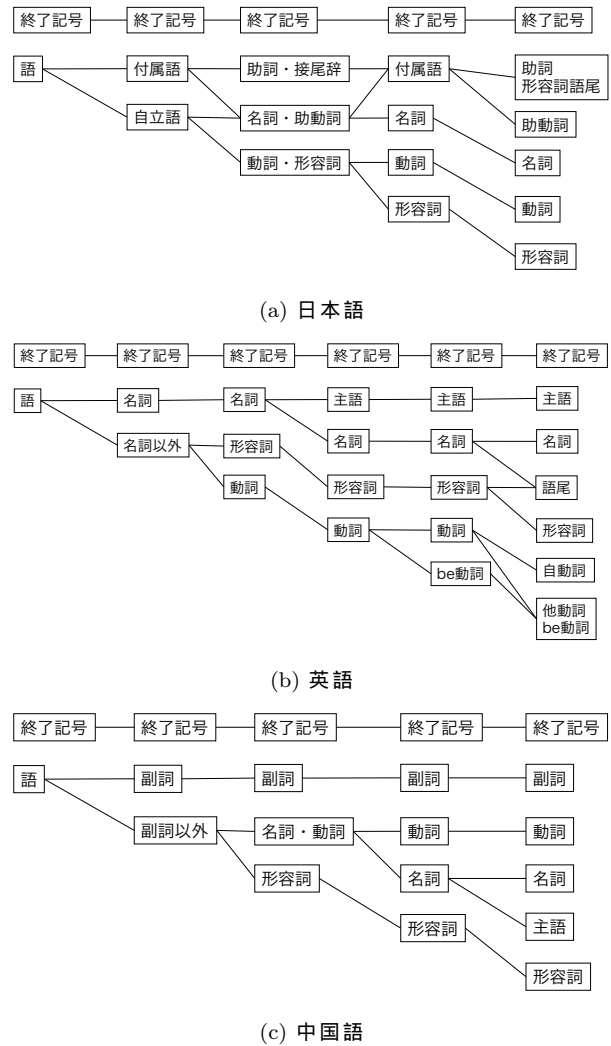


図7 統語範疇の分化過程

4. 議論

本論では、BHMMを用いて統語的手がかりから未知語の隠れ状態（統語範疇）を判断し、その指示対象を推定するモデルを提案した。そして、BHMMの隠れ状態数 S を増加させることによって、幼児の統語発達を表現した。図3に示したシミュレーション結果より、Imai et al. [3]が報告した名詞・動詞般用課題成績の発達的变化と一致する結果を得た。特に、五歳児で観察された、項を省略したときの動詞般用の母語依存性を一つのモデルで再現することができた。さらに、隠れ状態が表現する品詞を解析することによって、名詞・動詞般用における現象の背後にある統語範疇構造を明らかにした。日本語は助詞や接尾辞、英語は語順、中国語は副詞といった、それぞれの言語構造を反映した範疇構造となっており、その結果、名詞・動詞般用課題での成績に母語間の差異が現れたと考えられる。

以上の結果は、三歳から五歳の統語理解能力がBHMMのような単純な系列規則の統計学習によって実現される可能性を示唆する。また、この能力の発達的变化は隠れ状態の表現能力が向上し、統語範疇表現が精緻化されることによって引き起こされると考えられる。しかし、本モデルでは隠れ状態数 S はあらかじめ設定され、 S の増加メカニズムの問題が残る。動詞構文の統語構造が幼児とその養育者で類似する[13, 14]ことから、言語経験が幼児の統語発達のトリガの一つであると考えられるが、どのような言語入力か幼児の統語発達を導くかは不明な点が多い。今後、BHMMの最適な S を自動決定する手法[15, 16]を用いて S の増加に寄与する言語入力の特徴を調査したい。

提案モデルは幼児の名詞・動詞般用の発達的变化を統語範疇の精緻化という統語機構によって説明した。一方で、Imai et al. [3]は言語に依らず三歳児が名詞般用に成功し、動詞般用に失敗することについて、物体よりも動作の知覚的な切り出しが困難であること[4, 17]が原因と議論している。しかし、この般用課題でのその知覚発達の影響は自明でない。Waxman et al. [18]は同様の名詞・動詞般用において、目的語を代名詞でなく既知語として与えると、三歳児でも動詞般用が可能であることを示した。したがって、三歳での動詞般用の失敗は、動作の知覚的切り出しの困難というよりは動詞の統語的な理解の困難であると結論づけられている[18]。今回、知覚発達を考慮せずとも幼児の名詞・動詞般用課題の発達を再現できたことから、この課題の発達的变化は幼児の文法的側面の発達に依拠するというWaxman et al. [18]の立場を支持する。しかしながら、知覚的な切り出し

の能力は、特に語彙獲得において重要な要因である。提案モデルでは指示対象や事物範疇の空間は離散的で不変であったが、これらの空間の精緻化が統語発達や語の般用に及ぼす影響も興味深い。

5. 結論

本研究ではImai et al. [3]が報告した三歳から五歳でみられる名詞・動詞般用の発達的变化とその母語依存性を再現可能なモデルを提案した。そして、その発達の現象の背後には幼児の統語範疇の精緻化が存在し、入力言語構造を反映した統語範疇を獲得することで、動詞般用成績に言語間の差が生じる可能性を示した。さらに、獲得された統語範疇の解析により、日本語は助詞や接尾辞、英語は語順、中国語は副詞を手がかりに統語範疇を獲得することを明らかにした。このように、一つのモデルで複数の言語に対する統語範疇構造の発達を詳細に記述することができた。これらの結果から導かれる統語発達過程の仮説を検証するための、さらなる発達心理学的研究が期待される。

謝辞

慶應義塾大学環境情報学部 今井むつみ教授からご助言をいただいたことに対し、ここに記して謝意を表す。また、本研究は日本学術振興会科学研究費補助金 特別推進研究(24000012) および特別研究員奨励費(13J00756)の補助を受けた。

参考文献

- [1] Braine, M. D. S. and M. Bowerman (1976) "Children's first word combinations," *Monographs of the society for research in child development*, Vol. 41, pp. 1-104.
- [2] Olguin, R. and M. Tomasello (1993) "Twenty-five-month-old children do not have a grammatical category of verb," *Cognitive Development*, Vol. 8, pp. 245-272.
- [3] Imai, M., L. Li, E. Haryu, H. Okada, K. Hirsh-Pasek, R. M. Golinkoff, and J. Shigematsu (2008) "Novel Noun and Verb Learning in Chinese-, English-, and Japanese-Speaking Children," *Child Development*, Vol. 79, pp. 979-1000.
- [4] Imai, M., E. Haryu, and H. Okada (2005) "Mapping novel nouns and verbs onto dynamic action events: Are verb meanings easier to learn than noun meanings for Japanese children?" *Child Development*, Vol. 76, pp. 340-355.
- [5] Elman, J.L. (1990) "Finding structure in time," *Cognitive science*, Vol. 14, pp. 179-211.
- [6] Morifuji, M. and T. Inui (2004) "A connectionist model of vocabulary acquisition based on development of grammatical knowledge," in *Conference on Neuro-Computing and Evolving Intelligence*.

- [7] Toyomura, A. and T. Omori (2005) “A computational model for taxonomy-based word learning inspired by infant developmental word acquisition,” *IEICE transactions on information and systems*, Vol. 88, pp. 2389–2398.
- [8] Goldwater, S. and T. Griffiths (2007) “A fully Bayesian approach to unsupervised part-of-speech tagging,” in *Annual meeting-association for computational linguistics*, Vol. 45, p. 744.
- [9] Miyata, S. (2013) *MiiPro Nanami Corpus*: PA: TalkBank, ISBN 1-59642-473-7.
- [10] Henry, A. (2004) *English Belfast Corpus*: PA: TalkBank, ISBN 1-59642-037-5.
- [11] Tardif, T. (2007) *Chinese Beijing2 Corpus*: PA: TalkBank, ISBN 1-59642-287-4.
- [12] Cameron-Faulkner, T., E. Lieven, and M. Tomasello (2003) “A construction based analysis of child directed speech,” *Cognitive Science*, Vol. 27, pp. 843–873.
- [13] Choi, S. (1999) “Early development of verb structures and caregiver input in Korean: Two case studies,” *International Journal of Bilingualism*, Vol. 3, pp. 241–265.
- [14] Theakston, A. L., E. V. M. Lieven, J. M. Pine, and C. F. Rowland (2001) “The role of performance limitations in the acquisition of verb-argument structure: An alternative account,” *Journal of child language*, Vol. 28, pp. 127–152.
- [15] Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006) “Hierarchical dirichlet processes,” *Journal of the American Statistical Association*, Vol. 101, pp. 1566–1581.
- [16] Fox, E. B., E. B. Sudderth, M. I. Jordan, and A. S. Willsky (2007) “The sticky HDP-HMM: Bayesian nonparametric hidden Markov models with persistent states,” Technical Report 2777, MIT Laboratory for Information and Decision System.
- [17] Gentner, D. (1982) *Why nouns are learned before verbs: Linguistic relativity versus natural partitioning*: Erlbaum, Hillsdale, pp.301–334.
- [18] Waxman, S. R., J. L. Lidz, I. E. Braun, and T. Lavin (2009) “Twenty four-month-old infants’ interpretations of novel verbs and nouns in dynamic scenes,” *Cognitive Psychology*, Vol. 59, pp. 67–95.