繰り返し学習モデルにおける高速化のための文字省略プロセスの提案
Fast Process for Iterated Learning Model with Omitting String

須藤 洸基†, 的場 隆一†, 萩原 信吾†, 中村 誠‡
Hiroki Sudo, Ryuichi Matoba, Shingo Hagiwara, Makoto Nakamura

† 富山高等専門学校, ‡ 名古屋大学
National Institute of Technology, Toyama Colledge, Nagoya University
sstudyhiroki@gmail.com

Abstract

An infant learns his/her language by communication with others. For simulating 1st language acquisition, Iterated Learning Model (ILM) is used in related studies. Expanding ILM to change a learning environment where an infant agent receives utterances from more than one parent agents, length of uttered string from the parent agents increases rapidly. This phenomenon is clearly distant. The purpose of this sturdy is to prevent increase of uttered string length by clipping utterance method. Applying this method, average of uttered string length decreases without potent influence in intergenerational language propagation.

Keywords — Language acquisition, Iterated learning model

1. はじめに

幼児は成長と共に母語を獲得して，言葉によって多種多様な表現をできるようになる．従来，この幼児の言語獲得は，生得説と環境説によって議論されてきた．生得説と環境説とは，次のような説である[?].

生得説 幼児の言語発達には一定の法則と順序が存在し，その基礎にある認知発達段階に基づき特別に障害のない限り，幼児に共通する過程として言語獲得を強調する

環境説 特別な訓練や教育を受けなくとも．言語を自然に習得するという概念に基づくため，出生後の環境が強く影響し，環境（社会，文化，人間）の模倣により言語が修得されるという模倣説を強調する

この生得説において，幼児に共通する言語獲得のための能力を生得的能力という．藤井[?]によると，近年の生得説に関する論議では，生得的能力があるかないかという単純命題ではなく，何が生得的能力か，それがどのようなメカニズム・プロセスで言語獲得に寄与するのかという内容に変化してきたと述べている．この幼児の生得的能力に関して，コンピュータシミュレーションを用いて追求する研究がある[?, ?, ?]．これらの研究では，幼児の言語獲得に関するモデル（語彙獲得のモデルや構文獲得のモデル）を構築して，仮定した生得的能力の有用性や実在性を評価している．

我々は親と子の垂直伝達と世代交代をモデル化した繰り返し学習モデル（Iterated Learning Model; ILM）[?]を用いて，生得的要因である認知バイアスの言語獲得における効果を検証してきた[?, ?]．しかし，このモデルでは，幼児の言語獲得は親の影響のみを受けるようにモデル化されている．現実では，幼児は親以外の人間からの影響を少なからず受けて言語を獲得していく．したがって，これをモデル化して，改めて認知バイアスの言語獲得における効果を再検討する必要がある．このような集団における言語獲得のモデルをILMの汎化学習機構[?]を利用して構築すると，集団の用いる言語の発話文字列長が際限なく増加することが経験的に知られている[?]．これは，ある意味を表現する発話が時間と共に際限なく増加していくことを意味している．我々が扱う言語は，省略語の浸透により，言葉を表現する文字列が短くなることはあるものの，時間と共に文字列が長くなっていくことはないと考える．したがって，既存のILMは現実の言語変化とは，文字列長の変化においてかけ離れたモデルであると言える．さらに，この文字列の増加現象は，充分な計算資源や膨大な計算時間を必要とするといったシミュレーションを実行する上での問題が生じる．

我々はこの問題の解決と，ILMを現実的なモデルとするために，人間が行う省略語の生成ルール[?, ?]を基に発話文字列省略プロセスの提案を行う．ただし，この発話文字列省略プロセスは文字列長の際限ない増加とシミュレーションを実行する上での問題を解決するものであり，実験結果に影響を与えないことを理想とする．これをILMに

組み込み，言語進化シミュレーションの経過を観察した．

本章に続いて，第2章では本研究の実験モデルであるILMについて詳述する．第3章では我々が提案する発話文字列省略プロセスについて説明する．第4章では本研究のILMにおけるシミュレーションの経過を示す言語変化の定義を行う．第5章では発話文字列省略プロセスを適用した場合とそうでない場合の文字列長と言語変化を比較した結果を示し，第6章に結論を置く．

2. 繰り返し学習モデル

本章では本研究の実験モデルであるILMについて詳述する．ここで説明するILMはKirbyの実験[?]で用いられた設定に基づいている．

ILMは幼児の言語獲得と人間の世代交代を単純化したモデル（図??）である．親エージェントは，意味空間からランダムに選ばれた意味に対して，自らの言語知識を用いて発話を生成する．意味空間とは，5種類の動詞（述語）と5種類の名詞（主語，目的語）から成る100種類の2項述語形式の意味の集合である（図??）．親エージェントが意味に対する発話を生成できない時は，できるだけ自らの言語知識を利用して発話できるような新しい発話規則を1つ作成する．これをinventionと呼ぶ．inventionの際に作られる発話規則は，3文字以下のランダムな文字列を用いて作られる．inventionを使用した場合は，inventionで作成された発話規則を親エージェントの知識に加える．このような親エージェントによる発話生成を意味空間の意味の数の半分である50回行う．親エージェントの発話生成後，子エージェントは親エージェントから意味とその発話のペアを受け取り記憶する．この時，子エージェントは最大50種類の発話を親エージェントから受け取る．よって，子エージェントは親エージェントから意味空間の全ての意味（100種類）を表現する発話を受け取ることはできない．子エージェントは記憶した内容から汎化学習（図??）を行って文法を解釈する．子エージェントの持つ汎化学習機構はchunk, merge, replaceと3つの汎化学習で構成されている[?]．これにより，子エージェントの言語知識が変数で表現され，より多くの意味を表現できるようになる．この学習を自分の知識に変化がなくなるまで繰り返し行う．汎化学習が終わった時点で，子エージェントは親エージェントとなり，1世代のプロセスが終了する．このプロセスを繰り返すモデルがILMである．

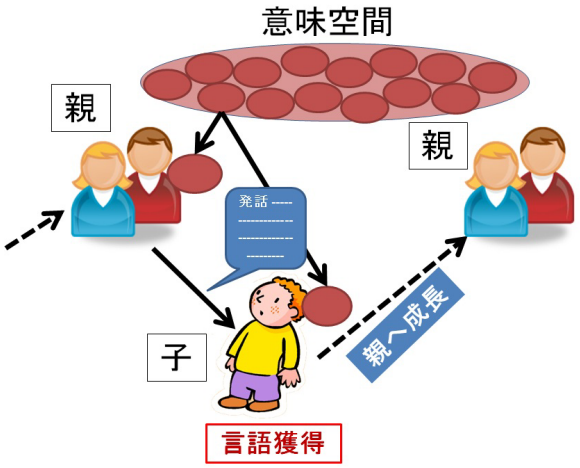

図 1 Kirby's ILMの図示

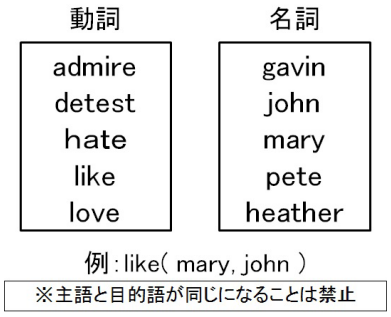

図 2 ILMで扱う意味の表現

$$\begin{cases} S/\mathrm{detest}(\mathrm{mary},\mathrm{john}) \rightarrow marydetestsjohn \\ S/love(mary,john) \rightarrow marylovesjohn \end{cases}$$
$$\downarrow chunk$$
$$\begin{cases} S/X_1(mary,john) \rightarrow mary\ N_0/X_1sjohn \\ N_0/detest \rightarrow detest \\ N_0/love \rightarrow love \end{cases}$$

$$\begin{cases} N_1/gain \rightarrow gain \\ N_2/gain \rightarrow gain \end{cases}$$
$$\downarrow merge$$
$$\{N_2/gain \rightarrow gain$$

$$\begin{cases} S/admire(john,pete) \rightarrow johnadmirespete \\ N_3/admire \rightarrow admire \end{cases}$$
$$\downarrow replace$$
$$\begin{cases} S/X_1(john,pete) \rightarrow john\ N_3/X_1spete \\ N_3/admire \rightarrow admire \end{cases}$$

図 3 汎化学習：chunk, merge, replace

3. 発話文字列省略プロセスの提案

本章では人間が行う省略語の生成ルールを基にした発話文字列省略プロセスについて詳述する．また，文字列を不必要に削らないように，文字列を省略するか否かの判定基準を定義する．

本研究では，人間が行う省略語の生成ルールを基にしている．この人間が行う省略語の生成ルールについて説明を行う．人間は省略語の生成を単純語または複合語を構成する単純語単位で行う[?]．そして，単純語単位において，前から削る，後ろから削る，または両方から削る，といういずれかの手法で省略を行っている．どの手法で削るか，どこまで削るかについては，音韻を解析することである程度の形式化ができるという研究がされている[?, ?]．しかし，人間の省略語には形式化できない場合があり，省略の仕組みについては完全に明らかとなっているわけではない．

我々が提案する発話文字列省略プロセスは，単語単位で文字列の省略を行う．ここで用いる単語単位というのは，発話規則において変数を含まない文字列単位という意味である．ILMでは音韻という概念を用いていないため，文字列から形式的にどの部分を削るか，どこまで削るかを音韻から決定することは困難である．そのため，本研究では削る場所を文字列の一番後ろとし，既存知識に省略対象の文字列がなく省略対象の文字列が1文字でない場合に削るという判定基準を定義した．後ろから削るという手法は，人間が行う文字列省略において頻出するパターンであるため，この手法を採用した．そして，誤読される可能性がないならば削り，最低1文字は残すという方針で省略の判定基準を決定した．具体的な発話文字列省略プロセスの例を図??に示す．この図は子エージェントが$like(john, mary)$に対する発話を生成する過程での発話文字列省略を示している．まず，子エージェントは$like(john, mary)$を生成するための発話規則を言語知識から選び出す．そして，$cba, ed, abef$といった単語単位で省略を行う．例として，cbaの文字列省略を順を追って説明する．まず，cbaが省略可能か判断する．cbaという文字列は他の発話規則に存在しないので省略可能である．そして，cbaを後ろの文字から省略して，cbとなる．cbは他の発話規則にない文字列なので，もう一文字削ってcとなる．cは他の発話規則にある文字列であるため，誤読される可能性がある．したがって，cbを省略語文字列とする．

4. 繰り返し学習モデルにおける言語変化

本章ではILMにおけるシミュレーションの経過を数値的に観察するために，言語変化について定義を行う．発話文字列省略プロセスを入れた場合と入れない場合での言語変化を比較することで，発話文字列省略プロセスが実験の結果に影響を与えているかを判断する．

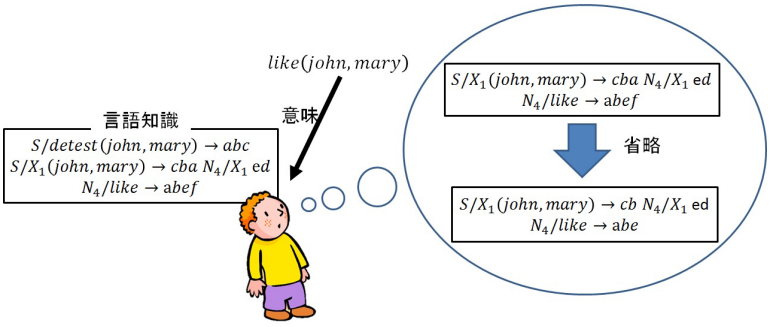

図 4 発話文字列省略プロセスの例

本研究における実験の結果とは，ILMの各世代の言語知識がどのように変化していくかということを指す．我々は言語知識を表面的な変化と内面的な変化に分けて捉えることとした．表面的な変化とは，汎化学習後の子エージェントが持つ発話規則数の変化と表現できる意味の数の変化である．ILMでは，汎化学習と世代交代により，より少ない発話規則数でより多くの意味を表現できるように変化していく[?]．内面的な変化とは，子エージェントが親エージェントの文法を正しく理解したかを示す言語間距離[?]の変化である．先に述べたように，ILMでは世代を経ることでより少ない発話規則数 でより多くの意味を表現できるようになる．そのため，子エージェントは親エージェントの文法を推測しやすくなる．したがって，世代交代を重ねることで，子エージェントと親エージェントの言語間距離は小さくなっていく．

ここで，本研究で用いた言語間距離の計算方法を詳述する．言語間距離は子エージェントが親エージェントの発話から発話規則をどれだけ理解できたかを示す尺度として設計されている． 以下に言語間距離の測定手法を述べる．

1. 親エージェント，子エージェントそれぞれの言語知識で生成可能な意味とその発話のペアをすべて生成する．
2. 子から生成した意味と発話のペアを 1 つ選ぶ．
3. 子から選択したペアの意味と最も近い意味を持つペアを親のペアから選ぶ．意味の類似度はハミング距離で測る．
4. 同ペア間の発話の距離をレーベンシュタイン距離で測る．ただし，レーベシュタイン距離は，0から1に正規化した値を用いる．

5. 子から生成したペア全てにおいて2から4の操作を行い，4で求められる距離の総和を求める．
6. 言語間距離は5で求めた総和を子エージェントのペアの数で割った値となる．これは，4で求めたレーベンシュタイン距離の平均となる．

図??は言語間距離のアルゴリズムのイメージを示している．図??の点（·）は言語知識から生成された意味と発話のペアを表現している．この図からわかるように親エージェントの発話には，言語間距離に影響しない発話が出現することがある．したがって，言語間距離は親エージェントと子エージェントの言語知識を入れ替えて計算すると，入れ替える前と同じ数値が出てくるとは限らない．しかし，これは親エージェントが子エージェントに伝えていない文法で生成された発話を計算から除外するためである．子エージェントは親エージェントが発話しなかった文法については，その文法を正しく理解できるはずがない．したがって，このような文法から生成された親エージェントの発話は計算から除外するべきノイズである．

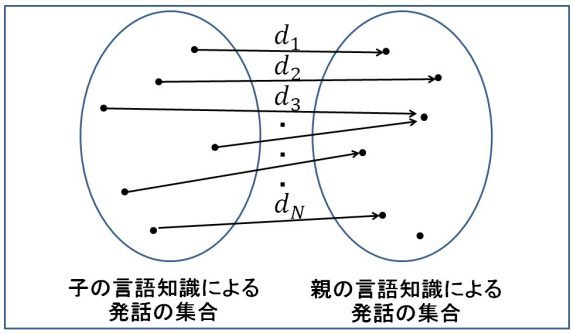

図 5 言語間距離のイメージ

5. 発話文字列省略プロセスの適用

本章では発話文字列省略プロセスを垂直伝達環境のILMに適用して，適用しなかった場合との差異を明らかにする．第4章で示した表面的な変化，内面的な変化と平均発話文字列長で発話文字列省略プロセスを適用した場合と適用しなかった場合での比較を行った．

本研究では，実験モデルとしてinventionで親エージェントの言語知識が増えないILMを構築した．inventionで親エージェントの言語知識を増やさないようにすると，発話文字列長が際限なく増えていくことがわかっている．実験は，5世代毎の発話規則数と表現できる意味の数，および，言語間距離で評価し，300世代まで行った．実験結果は100試行の平均で行い，発話文字列省略プロセスを適用した場合と適用しなかった場合で比較した．

5.1 表面的な変化

本節では発話文字列省略プロセスを適用した場合と適用しなかった場合について表面的な変化で比較する．図??, ??は表現できる意味の数の変化と発話規則数の変化のグラフである．どちらも横軸は世代数で，図??の縦軸は表現できる意味の数，図??の縦軸は発話規則数を意味している．各図のWithout ClippingとWith Clippingは，発話文字列省略プロセス適用無しと適用有りのグラフである．

図??を見ると，表現できる意味の数は300世代目で適用無しが約88.5種類，適用有りが約90.8種類であり，各世代での差を平均すると，適用有りの方が約2.58種類多くなっている．図??を見ると，発話規則数は300世代目で適用無しが約44.6個，適用有りが約44.3個であり，各世代での差を平均すると，適用有りの方が約1.47個少なくなっている．

表面的な変化における発話文字省略プロセスの適用は，より少ない発話規則でより多くの意味を表現できるようになるように働くことがわかった．しかし，表現できる意味の数でも発話規則数でも全体からみると小さい値であり，発話文字列省略プロセスの影響は少ないと考える．

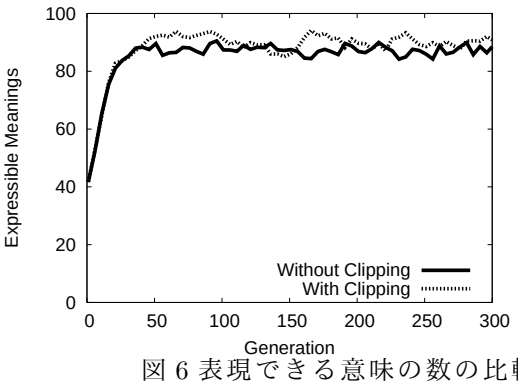

図 6 表現できる意味の数の比較

5.2 内面的な変化

本節では発話文字列省略プロセスを適用した場合と適用しなかった場合について内面的な変化で比較する．図??は親エージェントと子エージェントの言語間距離の変化を示したグラフである．横軸は世代数で縦軸は言語間距離を表している．図??のWithout ClippingとWith Clippingは，発話文

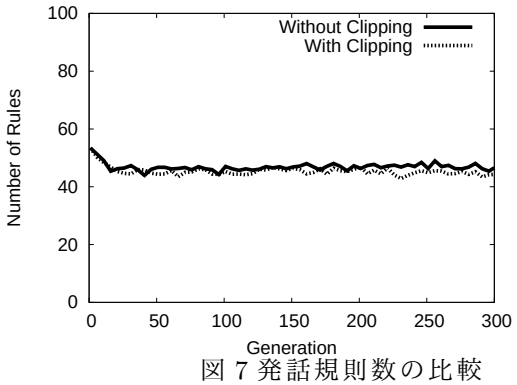

図 7 発話規則数の比較

字列省略プロセス適用無しと適用有りのグラフである．

図**??**を見ると，300世代での言語間距離は発話文字省略プロセスの適用無しで約0.196，適用有りで約0.227であった．このように値の振動により，多少の差はでるが，各世代での差を平均すると，適用有りの方が約0.00325大きくなり，全世代を通して値に差がでなかったので文字省略プロセスの影響はほとんどないと考える．

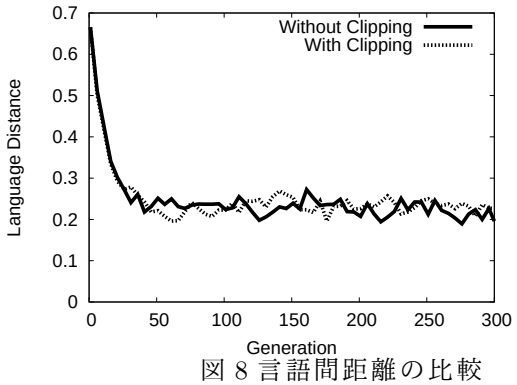

図 8 言語間距離の比較

5.3 発話文字列長の変化

本節では発話文字列省略プロセスを適用した場合と適用しなかった場合について発話文字列長の変化で比較する．図**??**は発話文字列省略プロセス適用無しの平均発話文字列長のグラフである．そして，図**??**は発話文字列省略プロセスを適用した平均発話文字列長のグラフである．図**??**, **??**のどちらとも横軸は世代数で，縦軸は平均発話文字列長を表している．そして，Without ClippingとWith Clippingは，発話文字列省略プロセス適用無しと適用有りを意味する．

図**??**を見ると，発話文字列省略プロセス適用無しの平均発話文字列長が300世代で約574.1文字であるのに対して，図**??**の適用有りの平均発話文字列長は300世代においても約5.3文字と大幅に抑えることができている．さらに，図**??**では平均発話文字列長が収束していることがわかる．

このように発話文字列省略プロセスの適用により，発話文字列長を抑えることができ，また一定の値で収束させることができた．発話文字列省略プロセスは発話文字列長を小さくするだけでなく，発散性をも無くすことができることがわかった．

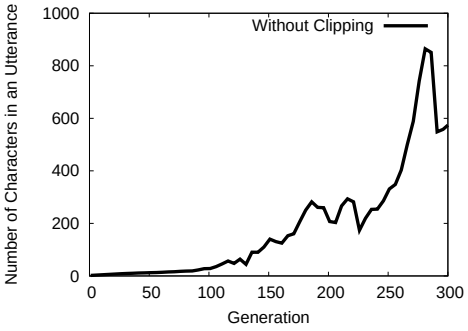

図 9 平均発話文字列長の比較

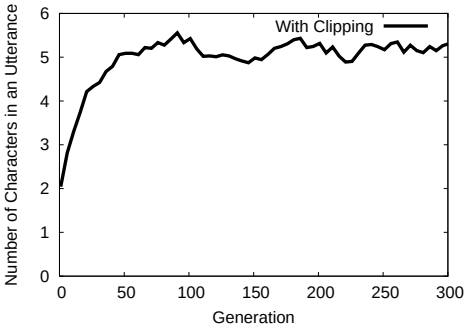

図 10 発話文字列省略プロセスを適用した場合の平均発話文字列長の変化

6. おわりに

ILMでは親以外から発話を受け取ると，世代を経るごとに，発話文字列長が増加していく現象が発生する．このような現象は現実では発生しない．この現象を解消するために，発話文字列を省略するプロセスを提案した．この省略プロセスは，人間が行っている省略語生成ルールを基にして考案

したものである．この発話文字列省略プロセスの有効性を検証するために，発話文字列長が増加するモデルを構築し実験を行った．実験では表面的な変化と内面的な変化，および発話文字列長で発話文字列省略プロセスの適用有りと適用無しを比較した．その結果，表面的な変化への影響は小さく，内面的な変化への影響は見られなかった．発話文字列長では単純に文字が減っただけではなく，ある一定の値に収束させることができた．

実験では，発話文字列長が大きく変化しているため，表面的な変化への影響があることは妥当である．なぜならば，発話文字列長が減ることは相応の情報量が減ることであり，子エージェントの言語獲得難易度が上がることを意味するからである．しかし，実験結果の発話文字列長は大きく省略されているが，表面的な変化と内面的な変化への影響は小さい．これは，人間が行う省略語の生成ルールを基に発話文字列省略プロセスを定義したためであると考えられる．しかし，発話文字列をどこまで削るかの判定基準は，多様性を持たないならば削るという方針で定義しており，その点は人間が行う省略を模倣している訳ではない．それにもかかわらず，内面的な変化には影響がなく，表面的な変化に対しては小さい影響で発話文字列を省略できた．よって，この判定基準は，現実でも無意識に使用されている可能性がある．

今後の課題として，発話文字列長省略プロセスを適用した集団における言語獲得のモデルで，認知バイアスの言語獲得における効果を検証する事が挙げられる．また，発話文字列省略プロセスにおいてどこまで文字列を削るか判定する基準が，人間が行う判定基準と同じであるかを再度検討する必要がある．

参考文献

[1] 橋本敬，中塚雅也（2007） “文法化の構成的モデル化 - 進化言語学からの考察-”，認知言語学会論文集，7, 33-43.
[2] 藤井聖子,（2001）“構文理論と言語習得（特集　構文理論の現在）” 英語青年，研究社出版，Vol. 147, No. 9, pp. 536-540,549.
[3] Jamet, D.（2009）“A morpho-phonological approach of clipping in English. Can the study of clipping be formalized ?”, Lexis - E-Journal in English Lexicology.
[4] 今井康晴,（2011）“幼児の言語獲得に関する一考察 : ブルーナーの言語獲得論を中心に” 学習開発学研究，No. 4, pp. 21-27.
[5] Kenny, S., and Hurford, James R. (2003) “Language Evolution in Populations: Extending the Iterated Learning Model”, Advances in Artificial Life, pp. 507-516.
[6] Kirby, S. (2002) “Learning, bottlenecks and the evolution of recursive syntax”, Linguistic Evolution through Language Aquisition, Cambridge University Press.
[7] 窪薗晴夫,（2010）“語形成と音韻構造-短縮語形成のメカニズム-” 国語研プロジェクトレビュー，Vol. 1, No. 3, pp. 17-34.
[8] 的場隆一，中村誠，東条敏（2008）“構文獲得における対称性バイアスの有効性”, 認知科学，Vol. 15, No. 3, pp. 457-469.
[9] Matoba, R., Sudo, H., Tojo, S., Hagiwara, S. (2013) “Evaluation of the Symmetry Biass in Grammar Acquisition”, AROB, 18th.
[10] 篠原修二，田口亮，橋本敬，桂田浩一，新田恒夫（2007）“語彙学習エージェントにおけるバイアスの自律調整について”，人工知能学会論文誌，Vol. 22, pp.103-114
[11] 須藤洸基，的場隆一，萩原信吾，中村誠，東条敏（2013）“第一言語獲得における認知バイアスに基づいた言語知識修正”，認知科学会第30回大会