# 確率的言語モデルと第2言語習得理論：学習者コーパスの理論化に関する一考察

# A Probabilistic Model of Language and SLA Theory

投野 由紀夫

There is a growing awareness that many aspects of linguistic structures, variations and use are probabilistic. This probabilistic view will help us restructure the model of second language acquisition. I will propose a new framework of integrating learner corpus findings into an L2 acquisition theory by employing the Bayesian analysis as the foundation of probability theory and the DOP model as the basis of a probabilistic model of language.

## 1. はじめに

学習者コーパス（learner corpus）は第2言語として当該言語を学ぶ学習者の作文・発話などの産出データを大量に収集電子化し，コーパス言語学の手法を用いて分析する分野である。この分野は1990年代初めから徐々に研究が進展し，ヨーロッパではInternational Corpus of Learner English（ICLE）（Granger 1998）のように10数カ国の母語の異なる大学生のデータを収集・比較するような試みや，Longman Learner's CorpusやCambridge Learner Corpusのように出版社が辞書編纂や言語テスト作成の目的で大規模学習者コーパスを収集してきた。日本においても，アルクの開発したStandard Speaking Test（SST）の英語インタビュー・データを1200人以上電子テキスト化したNICT JLE Corpus（和泉 他2001）や，日本人中高生1万人以上の英作文をコーパス化したJEFLL Corpus（投野2007）などが知られている。

これらのコーパスから得られる情報は，外国語学習者が使用するいろいろな側面の言語特徴であり，それらは一般に過剰使用（overuse），過少使用（underuse），エラー（error）という3つのタイプに分類され，母語話者との比較やレベルの異なる学習者同士での比較が行われる。日本においてもこれらの研究成果による英語学習者の語彙文法の発達過程の記述研究はかなりの数に上ってきている。学習者によるこれら語彙・文法の使用または誤用の頻度やその推移がわかることは，中間言語プロセスの記述研究としてはそれなりに価値がある。しかし，現在の学習者コーパス研究は記述研究が

中心であり，第2言語習得研究への理論面での貢献が弱い。一方で，コーパス言語学は大量に収集したテキストを研究対象にする分野であり，20世紀後半の数理統計学(数式重視で，推定量や検定統計量の性質を導出する）からこの15年ほどで計算機統計学（大量のデータを計算機を徹底活用して確率情報を中心に処理する）へ移行しているように，データ科学（data science）の最先端の手法を駆使できる分野である。

こういった流れの中で本論では，今まで記述研究のみであった学習者コーパス研究を補完するために，学習者コーパスから取得できる大量の言語統計データを利用して，どのように第2言語習得理論に対して提案していくべきか，その新しい理論的可能性について試案を提示したい。

## 2. ベイズ理論を応用した第2言語習得理論の構築

まず第2言語習得理論を考える前に一般的な科学の方法から出発する。通常，我々が問題に取り組む際に，演繹法（deductive logic）がよく用いられる。これは図示すれば(1)のようになる：

(1)　原因（cause）● ⇒ ○ ○ ○　結果（effects）

これはさまざまな有益な結果が有限個の明確に定義された公理（axiom）の論理的な帰結として生じるという考え方である。Chomskyの考える生成文法などの理論構築も，有限個の文法規則で無限の文を生成するという考え方ではこれと同じである。

しかし，科学では上記とは反対の状況に置かれる場合がよくある。すなわち，ある一連の結果が観察される場合に，それらの根底にある原因は何か，という問いである。この場合は帰納法（inductive logic）またはplausible reasoningと呼ばれる「最も確からしい推論」を行うことになる。これを図示したものが(2)である。図を見ればわかるが，この場合は原因が不明であるので，与えられたデータからの推論の方法を明確にしておく必要がある。

(2)　原因 ● ● ● ⇔ ○ ○ ○　結果

この推論の方法論的な根拠が確率論（probabilistic theory）である。これを従来の古典的統計学（classical statistics）では，分析対象となる母集団（population）から繰り返し一定量の標本（sample）をランダム抽出し，標本を使って未知の母集団特性（パラメーター parameter）を推測するといったことを行っていた。しかし，近年，人工知能や自然言語処理を含むデータサイエンス全体で注目されているのは古典的統計学とは違う発想のデータ分析方法である。それにはいくつかの流派があるが，ここではベイズ統計学（Bayesian statistics）を取り上げる。ベイズ統計学をコーパス分析に応用するのは，コンピューター分析による計算機統計学（computational statistics）の分野において，大量データから有用な情報を計算機処理によって獲得し複雑なシステムの背後にある真のモデルをイメージする，といった分析の基礎に，このベイズ統計学が利用されてきており，大量データを扱うコーパス分析への応用が期待されるからである。

ベイズの理論（Bayes' theorem）は，計算式では(3)のように表せる：

$$(3) \qquad P(X \mid Y,I) = \frac{P(Y \mid X,I)P(X \mid I)}{P(Y \mid I)}$$

$P(X \mid Y,I)$ は | の右側の事象Yおよび I が生起したことを条件として | の左側の事象 X が生起する確率，つまり条件付確率（conditional probability）を表す。(3)は確率論の基礎である加法定理(4)および乗法定理(5)から導かれる：

$$(4) \qquad P(X \mid I) + P(\overline{X} \mid I) = 1$$

$$(5) \qquad P(X,Y \mid I) = P(X \mid Y,I)P(Y \mid I)$$

(3)を具体的に見てみると，ベイズの理論で言わんとしているのは，I（関連する背景的な情報）やY（手元にあるデータ）が与えられた時，X（仮説）の確かさを確率的に産出しよう，ということである。$P(X \mid I)$ は事前確率（prior probability）という。これは手元にあるデータを分析する前に一般的に我々が知りえる情報をもとに仮説に対してどのくらい確信（あるいは不確定な見方）があるかを指す。これに $P(Y \mid X,I)$ を掛け合わせているが，これは事前確率を尤度関数（likelihood function）によって補正しているわけで，簡単に言うと事前確率（過去の経験などからわかっている仮説に対する確かさ）を仮説を知識として与えられた時の実際のデータの様子を見て修正

している，ということである。そして最終的に左辺の $P(X \mid Y,I)$ が事後確率(posterior probability) と呼ばれる。これをもう少し具体的に言葉で書くと (6) のようになる

(6)　　事後確率 = (データによる補正) × (事前確率)

これは「今までこうだったから多分将来…なるだろう」というような予測の仕方を確率的に定式化したものである。人間の認知的な判断や心理学的なさまざまな状況判断は，この過去の経験を基にした瞬時の状況判断によっており，これを日々刻々変わる新しいデータをもとに補正しつつ適用している，と考えるわけである。

近年，世界におけるさまざまな現象を確率論的に捉える視点が注目されている。自然現象（エルニーニョ現象など）だけでなく人間の行動（マーケティングから遺伝子制御まで）にもこれが当てはまる。人は問題解決の際に，過去の経験から最適な方法を選択している。現実的な問題は常に不確実性を含んでおり，この不確実性をどう扱うかが重要な鍵となる。その大きな潮流が，ベイジアン・モデリングなのである。

## 3. 確率的言語モデルの言語習得データへの応用

言語習得も基本的な学習プロセスは人間の認知システムの中での事後確率の変化として捉えることができる。たとえば，Levelt (1989) に基づく一連の言語処理の流れを (7) に示す：

(7) CONCEPTUALIZER:　伝達内容生成 → 言語化前の伝達内容

↓

FORMULATOR:　語彙選択・文法コード化 → 音韻コード化

↓

ARTICULATOR:　音声プラン → 実際の発声された発話

この場合，言語の知識として Levelt のモデルでは概念化装置 (conceptualizer) で生成された言語化される前の伝達内容が，メンタル・レキシコンの語彙部門(lemma) において伝達内容に合致した単語が活性化 (activate) され，その語彙に対応する文法構造が連動して活性化して選択される。この部分のメカニズムは多分に確率情報に基づいて制御されていると考えられる。たとえば，概念化装置で「X (私) が Y (所有物) を Z (誰か) に与える」というような伝達内容を意図として生成したとする。この際に，「与える」という意図に合致した語彙 give がレキシコンで活性化する。この際には，語彙項目内に概念にネットワークされた複数語彙が候補として考えられる

が，give, grant, impart, provide, vest, yield など文脈によってみな「与える」という意味になるものの中から，その話者の学習経験に応じて最も適当なものが確率的に高い値をもっており，閾値を超えれば選択される。学習は新しい発話や状況で既存知識による事前確率が補正されて新しい予測や選択が出来るようになることを指す。これも先のベイズの定理に基づいて，(8) のように書くことができる：

(8)（選択される動詞 | 意図，文脈）
=（新しいデータによる補正値）×（選択される動詞 | 既存知識）

母語習得の場合には (8) の（既存知識）部分が第 2 言語習得と大きく異なり，実際に母親とのやりとりの中で意味と文法の獲得を行っていくので，確率情報もより多様な分布を身につけている。それに対して，教室環境での第 2 言語習得の場合には，（既存知識）にあたる部分は教科書からの文法・語彙指導であることが多いので，「与える」= give といった対応関係を初学者は 100% の確率で対応させており，徐々にそこから習得段階が上がるにつれて確率分布の仮説を修正していく（つまり学習していく）と考えられる。

同様のことは give を選択した後の項構造の選択にもいえる。項構造の表示の仕方はいろいろあるが，give の場合には概略 (9) のような情報が格納されており，3 項動詞であることが明示されている：

(9) f(x,y,z) = give<x,y,z> ただし x = actor, y = possession, z = recipient

そして Levelt（1989）のモデルに従えば，FORMULATOR の内部で give の選択と同時に文法コード化（grammatical encoding）が起こる。この場合にも give の下位範疇化情報に関する確率分布が関与する。たとえば British National Corpus を見ると，give の後の項構造の頻度は概略 (10) のようになる：

| | | |
|---|---|---|
| (10) | give + NP + NP | .53 |
| | give + NP + PP | .16 |
| | be given | .13 |
| | give + PAR | .07 |
| | give + PP | .04 |
| | give + others | .07 |
| | | 1.00 |

母語話者はこのような確率情報を辞書項目内に有しており，これらを元に最適な構

造選択を行う。しかし，これだけが選択の要因ではなくManning（2003）が指摘しているような最適化理論（optimality theory）に準じるような制約の束が構造選択の判断を確率的に左右している，と考えたほうがよい。

このあたりの精密なモデル化は紙数が足りないので別の機会に譲るが，これらの確率情報をインプットから学習していくということが言語習得の主要なプロセスになっていることは大いに予想できるだろう。

## 4. Data Oriented Parsing (DOP) Model の学習者データへの拡張

確率的言語モデルを第2言語習得データに援用していく際に，大きな枠組みとして生成文法的なモジュール理論をとるか，コネクショニスト的な考え方をとるかは理論的な立場としては大きな分かれ目になる。現在，理論言語学の分野でもこれに関しては多様な考え方がある。しかし，衆目の一致するところでは，確率的に言語モデルを捉えようとする研究者は，チョムスキーの考える生成文法の枠組みはカテゴリー主義（categorical：すべての言語要素を分類された枠内に押し込もうとする考え方）が強く，いろいろな言語の比較分析の結果ある言語ではカテゴリーとして定義したほうがよくても，別の言語ではそうではない，という場合も報告されるなど（cf. Bresnan, Dingare and Manning 2001），カテゴリー的，固定的な対立概念の設定よりも連続体の上での確率的（probabilistic）な分布という捉え方の方が実態に則しているという考え方が主流になってきている（Bod et al. 2003）。しかし，これは即コネクショニスト的な立場をとるということではなく，今までの生成文法で構築されてきた言語理論を発展させていく方向性も模索されている。

先のベイズの理論と自分が有する大量の学習者コーパスを基にした時に，理論的に極めて重要かつ魅力的な考え方であると思われるのが，Data Oriented Parsing (DOP) model（Bod, Scha, and Sima'an 2003）である。これは確率的言語モデル（stochastic language model）の1つであるが，最もよく知られている確率文脈自由文法（probabilistic context-free grammar）などに比べると，句構造規則の派生関係から説明できない単語間の依存関係を比較的柔軟に対応できるなど利点が多い（Bod 2003: 26）。

DOPモデルでは，言語を構文解析済み情報が付帯したサブツリー（subtree）の集合体と考える。たとえば，John likes Mary. Peter hates Susan. という2文をもとにすると，(S(NP *John*)(VP (V *likes*)(NP *Mary*))), (S(NP *Peter*)(VP (V *hates*)(NP *Susan*))) から始めて，(S(NP)(VP (V *likes*)(NP *Mary*))), (S(NP *John*)(VP (V)(NP

*Mary*))), (S(NP *John*)(VP (V *likes*)(NP))) と次々とサブツリーを生成していくと全部で 34 個のサブツリーの集合になる。さらにこれらのサブツリーの集合を元にノード代入操作 (node substitution operation) によって，サブツリー同士を結合して新しい文を生成する。たとえば (11) がその例である：

(11) (S(NP)(VP (V)(NP Susan))) ○ (NP Mary) ○ (V likes)
= (S(NP Mary)(VP (V likes)(NP Susan))),

ここで○は結合操作を表し，「T○U」の場合には「U を T の一番左の非終端ノードに接続せよ」という意味になる。

このモデルがなぜ確率文法なのかというと，このサブツリーの結合によって文が生成される確率が計算できるからである。(11) の例でいえば，(12) のように各サブツリー選択の確率計算を行う：

(12) 事象
(a) サブツリーのルートが S であるすべてのサブツリーから (S(NP)(VP (V)(NP *Susan*))) を選択する確率 → 1/20
(b) サブツリーのルートが NP であるすべてのサブツリーから (NP *Mary*) を選ぶ確率 → 1/4
(c) サブツリーのルートが V であるすべてのサブツリーから (V *likes*) を選ぶ確率 → 1/2

これによって，この文を生成する確率は 1/20 × 1/4 × 1/2 = 1/160 となる。このような計算方法を用いることで，文の生成に関して確率的により可能性の高い構造とそうでないものをランク付けすることが可能になる。人間はこのようなサブツリーの断片を大量に記憶している，というのが Bod ら DOP モデルの提唱者の考え方である。文法を定義する際に，単語がすべて個別に辞書に格納されていて，それを Levelt のモデルに代表されるように，プロセス・モジュールを 1 つ 1 つ順番にリニアに経ながら発話に至る，というものよりも，DOP モデルは辞書格納部分のアイデアが極めて柔軟である。このモデルでは，単語とその構造情報が一体化しており，かつそこに確率分布が付与されているので，並列分散処理をして，瞬時に文を生成するようなモデルと親和性が高いと思われる。かつ，一般の確率文脈自由文法が説明できない離れた語彙間の依存性をサブツリーの概念で見事に解決している。この発想は Sinclair などが昔から行っている idiom principle の考えに非常によく似ている。人間の脳内の言

葉の格納の仕組みに対して，特にメンタル・レキシコンの状態に関して，興味深いモデルを提案している。

DOP モデルはまだ提案されて日が浅く，LFG への実装などいくつかの異なる提案が検討されている段階であるが，今後，学習者コーパスからの確率情報を説明する際に非常に面白い可能性を提供してくれるものであると筆者は考えている。

## 5. まとめ：問題点と今後の課題

本論では，学習者コーパスからスタートして，大量の学習者データから第 2 言語習得理論に結び付けていくための理論的枠組みに関して，ベイズ理論と確率的言語モデルを用いた枠組みを試案として提案した。その言語的な実装は DOP モデルによって行い，かつ学習者が日々接するインプットの質や量，また教授方法などによる言語知識の変容の状態，環境的なさまざまな要因を，ベイズ分析では柔軟にパラメーターとして取り込み，文生成の判断基準として，あるいは文生成の際の制約としてそれらを理論の中に組み込んでいくことが可能である。そして，学習者の中間言語の処理も DOP モデルで例示したような確率情報に基づく無自覚の自動処理のような言語操作部分と，より顕在的な意識の中で行う意思決定に似た言語操作部分とがあるはずである。

ベイズ理論が強力なのは，これらの学習者データとその環境的な要因を模式化（モデル化）し，どんどんパラメーターを追加したり，パラメーターを制御する超パラメーターというような概念を導入したりしても，基本のベイズの定理に従いまったく同様に処理可能な点である。特に第 2 言語習得のように，あまりに多様な要因が複合的に絡み合う現象には，不確実性を含む対象を計算機上に確率変数の集合とその間の条件付確率として表現して，計算を行わせることで，モデルの改善をしていくような手法が大きな可能性を秘めている。

今後の課題をいくつか挙げておこう。1 つはベイズ理論のより詳しい検討である。現在，ベイジアンモデリング（樋口他 2007）など，日本国内でも他分野でベイズ分析は用いられている。第 2 言語習得での利用はまだ筆者の知る限りないが，自然言語処理ではすでに御馴染みの手法となりつつある。それらの応用例から多くを学ぶことが可能であろう。

次に，DOP モデルをもとに学習者データの構文解析データからの確率分布を実験的にとってみることが必要である。これについては比較的小さなデータセットから開始し，母語話者の使用パターンの確率分布とのずれが著しいような項目を特定できる

とよい。そのためには，DOPモデルのような構造木を書けるような構文解析ツールの比較検討が必要である。これにはCharniakのparserが最も精度が高いといわれているので，これを中心に処理を検討する。

最後に，ベイジアンネットワークの構築である。複合的なモデル化をDOPモデルのような言語面と，学習者のインプットからの情報（例：教科書の頻度分布，授業観察でのインプットの頻度確率など），母語話者の言語使用頻度，といった複数要因をからめて，すでに知っている経験的な判断と既知データから言語習得モデルを予測する，といった試みを行いたい。これに関しては，Tono（2002）で動詞下位範疇化情報の獲得に関して，対数線形分析を用いて複合的なモデルの重み付けの可能性を示したが，同様の手法をDOPモデルによる確率分布とベイズ分析の適用で行うことができよう。

本論では，学習者コーパスの次の一手ということで，データと理論の融合に関して筆者なりの提案をしてみた。まだ筆者の理解が不十分な点もあるかもしれないし，議論が精密化していない部分も多々あるが，これらの論点に関して実証的なデータを添えながら具体的な提案をしていくことで，学習者コーパス研究と第2言語語彙習得研究，メンタル・レキシコン研究など，自分が過去に行ってきた諸分野の研究成果を統合する総合的な理論的枠組みを提供できる可能性に期待したい。

## 引用文献


Bod, R. (2003) Introduction to elementary probability theory and formal stochastic language theory. In Bod et al. (2003), pp. 11-37.

Bod, R., J. Hay, and S. Jannedy (eds.) (2003). *Probabilistic Linguistics*. Cambridge, Mass: MIT Press.

Bresnan, Joan, Dingare, S., and Manning, C.D. (2001). Soft constraints mirror hard constraints: Voice and person in English and Lummi. In M. Butt and T. Holloway King, (eds.) *Proceedings of the LFG 01 Conference*. Stanford, Calif.: CSLI Publication, pp. 13-32.

Granger, S. (ed.) (1998). *Learner English on Computer*. London: Addyson Wesley Longman.

Levelt, Willem J.M. (1989) *Speaking: From Intention to Articulation*. Cambridge, Mass: MIT Press.

Manning, Christopher D. (2003). Probabilistic syntax. In Bod et al. (2003), pp.289-342.

Tono, Y. (2002). *The Role of Learner Corpora in Second Language Acquisition Research and Foreign Language Learning: The Multiple Comparison Approach*. Unpublished Ph.D. thesis. Lancaster University.

和泉絵美・内元清貴・井佐原均（編著）(2004).『日本人 1200 人の英語スピーキング・コーパス』東京：アルク .
投野由紀夫（編著）(2007).『日本人中高生１万人の英語コーパス JEFLL Corpus — 中高生が書く英文の実態とその分析』東京：小学館 .
樋口知之（監修・著）(2007).『統計数理は隠れた未来をあらわにする：ベイジアンモデリングによる実世界イノベーション』東京電機大学出版局 .