用法基盤モデルが想定する言語知識のありようについて考える

大谷直輝
（東京外国語大学）

1. はじめに

本発表では，Joan Bybee や Ronald W. Langacker などの考えに端を発し（Bybee 1985; Langacker 1988），その後，認知言語学，構文文法，相互行為言語学，談話機能言語学などの分野で採用されている用法基盤モデル（the usage-based model）[1]の考え方を紹介する（Barlow and Kemmer 2000; Tomasello 2003; Goldberg 2006, 2019）。用法基盤モデルは，1980年代に生じたアプローチであり，「刺激の貧困」や「生得的な文法能力」のような，当時の言語研究の前提となっていた考え方に異を唱え，文法や語彙などの言語知識は，言語使用の場で繰り返し用いられる言語表現から生じるという考え方を提示した。用法基盤モデルに基づく研究は，コーパスの普及とともに広がりを見せており，現在では，上記の分野以外にも，言語獲得や，言語の変化・変種・変異等を扱う様々な分野において，言語使用の中から生じる言語知識の研究が進んでいる。

用法基盤モデルは，おもに英語を用いた研究を行う研究者から出てきた考え方であるが，発話の場で繰り返し用いられるパターンが，話者に認識され，抽象化や分割されることで言語知識が創発するという考え方は，英語という個別言語の分析に留まらず，自然言語に共通して適用できる原理だと考えられている。実際，用法基盤モデルを用いた言語分析は，初期の頃から現在まで，英語，ドイツ語，フランス語，等の印欧語族の言語に加え，日本語や中国語をはじめとした非印欧語族の研究においても広く見られる。

本発表の目的はおもに以下の三点である。第一に，用法基盤モデルの基本的な考え方を概観する。用法基盤モデルは，文法を分析する際に作例ではなく実例を使うことを奨励するだけの方法論ではなく，実際の言語使用から立ち現れる文法の在りようを捉えようとするモデルである。第二に，用法基盤モデルという一種の舶来物である理論を日本語の研究に適用をする際の注意点について考える。第三に，近年目覚ましい発展を遂げた大規模言語モデル（LLM）を用いた用法基盤モデルの研究の可能性を検討する。

本発表では，2 節で，文法に対する一般的な考え方を概観する。3 節では，用法基盤モデルが想定する文法が持つと考えられる特徴を見ていく。4 節では，英語の研究から生じた用法基盤モデルを日本語の分析に適用する際の注意点を検討する。5 節では，大規模言語モデル（LLM）と用法基盤モデルの類似点を挙げながら，LLM を用いた用法基盤モデルの研究の可能性を検討する。

2. 伝統的な文法の特徴

文法は言語学における最も重要な研究対象の一つであるが，用法基盤モデルが想定す

[1] the usage-based model には，使用基盤モデル，使用依拠モデル等の訳語もある。

る文法の姿は，一般的な文法とはかなり異なる。一般的には，文法という名称によっても示されるように，文法には，文（sentence）の存在が前提となる。文は，句点によって区切られ，文法規則が適用されるのは，文の中だけとされる。また，文は主部と述部からなり，日本語の場合，主語や目的語のような文法関係は助詞が標示するとされる。このような前提があるため，文法について論じる際には(1)のような例文が用いられる。

1. (a) 象は鼻が長い。
   (b) カキ料理は広島が本場だ。
   (c) 私は赤ん坊に泣かれた。

一方で，私たちが日常のコミュニケーションで話す言語は，一般的な言語分析で用いられる，完全な文とはおよそかけ離れたものである。

2. A：あのさー、ほら、結婚式のときにさー、あの、チャペルでさー。彼ってさー、返事早いじゃん、すごい。人にかぶせるじゃーん。
   B：そうだった？（名大コーパス）

一般的な文法から見ると，(2)のようなやりとりは，言い淀みや言い差しを含み，主語，述語，助詞などが欠けている不完全な文であるため，考察する対象としては不適切となる。しかし，（音声）言語は一義的には音声からなり，6000 から 8000 あるとも言われる世界の言語（Crystal 2013）の中で，固有の文字体系を持つのは，多く見積もっても 1 割以下という事実を考慮すると，(1)のような例文のみに基づいて作られた文法は「書き言葉バイアス」（written language bias）を内在すると言える。

以下，一般的な文法に含まれる特徴をまとめる。（cf. 中山・大谷 2020）

3. 1) 書き言葉に基づいて作られている。
   2) 有限の語を文法規則で組み合わせることで，無限の文が生成される。
   3) 構成性の原理に基づき，全体の意味は部分の要素に還元できる。
   4) 語彙と文法は排他的な部門を構成し，その例外がイディオムとなる。
   5) 論理的にどこまで言えてどこまで言えないかの境界を示すための規則である。

3. 用法基盤モデルが考える文法の姿

用法基盤モデルは，(3)とは対照的な言語観を持ち，文法も語彙も，抽象性や複雑性が異なるものの，形式と意味の対である構文からなると考える。以下に，用法基盤モデルが想定する文法の特徴を見る。

第一に，文法規則と呼ばれるものは，発話の場からボトムアップで出現すると考える。二重目的語構文を作るための規則を例にすると，幼児は，Gimmethat（それとって），Gimmethejuice（ジュース取って）というような，コミュニケーションの中で繰り返し用

いられるまとまり（holophrase）を聞く中で，Gimme, that, thejuice などの，内部構造を認識し，まとまり間の共通のパターン（Gimme O）を捉える。そこから，徐々に抽象化が進んでいき，Give me O（O を私にちょうだい）というパターンが，さらに，Give me O, Give her O, Give him O などのパターンから S Give OO （誰かが誰かに何かをあげる）というパターンが，最後に S give OO, S tell OO, S show OO のようなパターンから二重目的語構文（SVOO）が認識され，習得されると考える。そのため，用法基盤で「文法」という場合，抽象的な文法規則だけでなく，表現の一部がオープンスロットになった S give OO や Gimme O などの具体性の高いパターンも含む（Stefanowitsch and Gries 2003）。

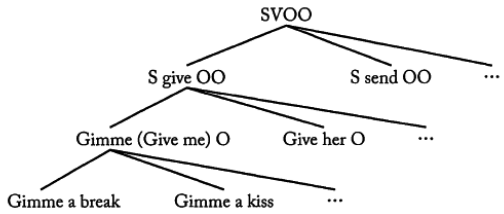


図１　ネットワーク構造としての言語知識

図 1 ボトムアップで生じるネットワーク構造としての言語知識（野村 2013）

第二に，用法基盤モデルは，言語知識は，文法と語彙に大別されるという伝統的な考え方とは対照的に，文法も語彙も言語使用において繰り返されるパターンから発現すると考える。すなわち，繰り返し用いられるパターンの内部構造が認識された後で，パターン間の類似性に基づき抽象化が進んでいくことで抽象的な文法規則が，パターンの内部が分節されていくことで語彙が習得されると考える。このような考え方をすると，文法であっても語彙であっても，言語使用の場で用いられるパターンから創発するため，両者はともに形式と意味の対からなる構文とみなせる。用法基盤モデルでは，抽象度や複雑度が異なる様々な構文が体系的に結びつき言語知識（Construct-i-con）を構成すると考える（図 1 参照）。

第三に, 発話における実際のやり取りの中から言語知識が創発すると考えることから，意味の中に多くの要素が含まれることになる。例えば，文は単独ではなく，談話を構成する要素として発話されることから, 内容的な意味だけでなく, 話者の捉え方（construal）, 談話内での機能，文脈的な意味なども意味として扱う。例えば，How has your day been? と Of what quality has your day been?は意味論的には「今日一日はどうであったか」を問う疑問文であるため類義的と言えるが，両者には，決定的な違いがある。前者は構文として話者や聞き手に共有された表現であり，一般的な挨拶や，短い会話を開始するという機能が表現全体に構文の意味として定着している。一方，後者については，構文として話者の記憶に蓄えられていることは考えにくく, 挨拶等の談話的な機能を持っていない。

第四に，意味がどこに存在するかという点でも，伝統的な考え方と異なる。現代の言語学では，話者が意味を言葉に詰め（encode）それを言語（code）という媒体を通して聞

き手に伝え，聞き手が頭の中で言葉から意味を取り出す（decode）ことでコミュニケーションが成立するという考え方が優勢である。一方，用法基盤モデルでは，百科事典的に基づいた言語観を採用する。すなわち，意味は言葉の中にあるのではなく，言語的なインプットを受け，百科事典的な知識の一部が活性化した結果，（意味と呼ばれるような）可能な事態が推測できるようになるため，意味は頭の中で構築されるものと考える[2]。

第五に，表現が違えば意味が違うという「no synonymy principle（同義語は存在しない原理）」の考え方を採用する。コミュニケーションでは，一般的に，同義語は一方が淘汰されるか，意味や機能が分化することが観察されるため，完全な同義語は存在しないと考える。これは抽象的な構文でも同様であり，例えば，二重目的語構文が持つ意味として，行為の結果所有権が移動するというものがある。これは，for 与格構文には見られない特徴である。例えば，He bought her a flower と He bought a flower for her.では，前者では，彼女が花を受け取った事態までを表すが，後者は，受け取ったところまでは含意しない。このような意味の違いは部分の意味に還元することができず，二重目的語構文と for 与格構文という抽象的な構文が持つ特徴とみなされる。

第六に，文法は，論理的に可能な文とそうでない文を分けるための規則ではなく，実際に発話されるか文を理解したり生成したりするために必要な知識とみなされる。そのため，言語学者しか使わないような「花子は太郎が誰が書いたかどうか誰に教えた本を読んだの？」というような文を生成するための規則も，言語能力の一部と認めつつも，そのような抽象的な文法規則より，実際のコミュニケーションで用いられる言語の産出や理解に必要と思われる知識に注目する。特に，会話では，瞬時的なやり取りが必要となるため，抽象的な文法規則を用いて語を組み合わせて発話するのではなく，先行する会話に出てきてすでに活性化されている表現など，具体的な表現をつなぎ合わせることで，発話がなされると考える（Du Bois 2014）。

4 用法基盤モデルを日本語の分析に用いる

現在，日本語の研究において，現代日本語書き言葉均衡コーパス（BCCWJ）や日本語話し言葉コーパス（CSJ）などのコーパスの整備に伴い，コーパスを用いたデータの収集や，コーパス内に現れる語の頻度や共起関係を用いた定量的な分析が盛んに行われている。そのような研究の中には，用法基盤モデル，認知言語学，構文文法などを用いた日本語の語彙や文法の研究も多く見られる。そこで，ここでは，英語の研究者が考案した用法基盤モデルの考え方を，日本語の分析に用いる際の注意点について二点記す。

一つ目は，用法基盤モデルは，実例を用いて言語を分析する方法論ではない点である。用法基盤モデルでは，作例よりも，実例を用いた分析を行う傾向があるが，実例を用い

[2] 例えば，「お年玉をもらった後で，神社に向かって出発したが，小銭を忘れてしまった」という文を聞いた後で，「いつ起こった出来事か」「何をしにお寺に行くのか」「小銭を何に使うのか」といった質問をされると，日本文化に関する知識がある場合，難なく回答ができるが，日本語がどれほど堪能であっても，日本文化に関する知識がない学習者にとって，回答は困難であるだろう。

た分析を行ったり，コーパス内で何らかのパターンを発見したりすることが，用法基盤モデルの研究になるわけではない。用法基盤モデルの研究では，3 節で示したような言語観に基づいて，言語使用から創発する言語知識のありようの解明を目指す必要がある。

二つ目は，舶来物の指標である用法基盤モデルをそのまま日本語の分析に用いる危険性である[3]。語順が比較的安定しており分析的な言語である英語に基づいた理論が，語順が比較的自由である日本語という膠着語の分析に，そのままの形で適用できるかは定かではない。一例を挙げると，英語を対象とした構文文法の研究では，項構造構文の分析をはじめとして，形式面において，語順によって伝わる意味の違いが重要視される傾向がある。しかし，語順が比較的柔軟な日本語で，英語と同様に，語順が意味の伝達において最重要であるかどうかは自明ではない。自然言語のコミュニケーションにおけるパターンの一つに過ぎない語順（つまり，語や句の並び）だけでなく，Du Bois (2014)で提示されている，進行中の談話の中に言語表現を産出するための重要な基盤があるという「対話統語論（dialogic syntax）」の考え方のように，意味・談話・パラ言語などの側面に見られる様々なタイプのパターンを考慮することが重要であるように思われる。

二点目で挙げた，理論に潜む英語のバイアスは，日本語だけでなく個別言語の研究者が直面する問題である。すなわち，現代の言語学のほとんどの概念は英語（あるいは，欧米の言語）を対象とする研究者によって考案，あるいは精緻化されたものである。そのため，前提として，英語の現象を分析するうえで適した理論や方法論となっている。英語と日本語を比べると，屈折型／膠着型，語順が厳格／比較的自由，主語が義務的／義務的でないという点を含め，本質的な違いが多く見られるため，理論に潜む英語バイアスを意識する必要があると思われる。一方で，用法基盤モデルをはじめとした英語発祥の理論が普遍的な言語分析の道具になる過程で，日本語のような英語とはかなり特性が異なる言語の研究が，理論の射程範囲や精緻化に貢献することは，大いに考えられる。

5 用法基盤モデルから見た大規模言語モデル

最後に，用法基盤モデルの観点から，生成 AI の基盤となる大規模言語モデル（BERT, GPT-4 など）を用いた言語研究の可能性について述べる。多くの言語学者にとって，人間と同じようなアウトプットを行っているように見える生成 AI は興味の対象になると考えられるが，現時点では，生成 AI の内部がブラックボックスとなっていることもあり，生成 AI を用いた言語学の研究は極めて少ない状況にある。一方，ChatGPT の衝撃以降，言語関係の学会では LLM の有用性が認識され，LLM をどのように活用すれば，どの程度，言語研究に役に立つのかという議論が始まりつつある。このような議論の中で出てきた話題の一つに，現在の深層学習に基づく LLM は初期の自然言語処理に比べて格段に用法基盤モデルとの相性が良くなっているという主張がある。

自然言語処理は，与えられた入力に対して何らかの処理を施して所望の出力を得るアプローチである。所望の出力をどのような処理により得るかという観点からパラダイム

[3] 本件に関して，大谷・中川・野元・長屋（近刊）で論じる予定である。

変遷を追っていくと，初期と現代のアプローチには大きな違いが見られる。2000 年以前の自然言語処理では，人が作成した解析の規則を用いて処理が行われていた。例えば，機械翻訳の場合，品詞，統語，意味などを解析する規則を機械に与え翻訳が行われていた。そのような規則は，言語学の知見や，内省（reflection）のような言語学的な方法論を用いて作成することが主流であった。一方，2010 年代に入り深層学習が導入されると，品詞解析，構文解析などの途中の解析を介さず，入力から直接所望の出力を得るアプローチが主流となった。例えば，上述の機械翻訳の例では，翻訳対象文を入力すると訳文のみ出力される。ここでは，品詞や統語情報などを機械に前もって与えなくても，機械が，用例からそれらを学び，アウトプットを行っている。この段階になると，品詞や統語というものをあらかじめ仮定する必要はなくなり，解析結果に必要な規則を文脈から抽出ができるようになったと言える。

このような，現代の深層学習の特徴の一部は以下のようにまとめられる。

1. 発話された文が重要なインプットになる。
2. 規則や言語知識は前もって与えられず，使用の中から学習することができる。
3. 言語使用から繰り返し用いられるパターンが抽出されて言語知識となる。
4. 文脈を均等ではなく，重要な部分とそうでない部分を区別しながら認識する。

(1)から(4)の特徴は，用法基盤モデルが想定する言語習得や言語知識と一致している。さらに，深層学習に基づく自然言語処理のアプローチを，用法基盤モデルの観点から見ると，深層学習を以下のように解釈することも可能となる。

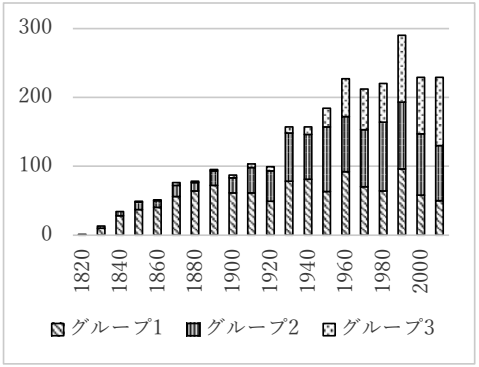
1. （大量の言語データが必要になるものの）品詞，文法関係，統語規則等を事前にデータに付与しなくても，生の言語データからそれらを学習できる点で，言語使用の中から文法を学べると主張する用法基盤モデルの考え方に似ている。
2. 文法も語彙もイディオムも生の言語データから同様のニューラルネットを用いて学習するということは，文法も語彙も言語使用の中から一般的な認知能力を用いて習得していくと考える用法基盤モデルと類似している。（文法は生得的に獲得し，語彙は学習で習得をするという異なる習得経路の違いは仮定されない。）
3. 生の入力データから言語を解析する規則を作り上げる際に，文脈の要素を均等に扱うわけではなく，文脈の中で重要な部分とそうでない部分を区別し，重要な部分に特に注目をするような設計がなされている（Vaswani et al 2017）点は，発話の中で重要な部分とそうでない部分を分けながら認識する人間の認知機能に近い。

このような類似性に注目をすると，大規模言語モデルを人間の言語処理や言語知識に見立てた研究の可能性について，検討を始める段階に達していると思われる。

最後に，LLM を用いた学際的な言語研究の一例として，大谷（2022）で提示した言語変化に関する仮説を，LLM を用いて検証することを試みた永田・大谷・高村・川崎（2022）を紹介する。現代英語には，*I'd be better off {dead/left alone/going home/without you}*のよう

に，*better off* に主語の仮想的な状態を表す補語句が後続し，文全体として「(補語句が表す仮想的な事態の方が）ましだ」という意味を表す構文が見られる。*better off* は語源的には，「裕福な」を意味する *well-off* の比較級であり，19 世紀の *better off* には，構文的な用法は見られなかった。そこで，大谷（2022）では通時的なコーパスを用いた *better off* の調査から得られた結果を認知言語学の観点から考察することで，字義的な意味を表していた *better off* に，(i) 意味の一般化，(ii) 仮想状況（＝セッティング）の前景化，(iii) セッティングの補語句化，が起こることで，他に類例がない，主語の仮想的な状態を表す補語句を持つ *better off* 構文が誕生したと考える仮説を提示した。

永田・大谷・高村・川崎（2022）では，この仮説に対して，最初に，人手による *better off* の分類を行い仮説の妥当性を検証した後で，深層学習に基づく言語処理的アプローチを用いることで検証結果を補強した。人手による分類では，「グループ 1: 字義的な用法」，「グループ 2: 仮想的な状況が前景化した用法」「グループ 3: *better off* が補語句を取る確実に構文とみなせる用法」に分けて，用法の経年変化を表すグラフの作成を行った。一方，大規模言語モデルでは，1) *better off* の用例に対して BERT を用いてベクトルに変換，2) 得られたベクトルを *k*-means クラスタリングでグルーピング，3) 年代ごとにグループ内の用例数を求めグラフ化，という手順でグラフの作成を行った。図 1 と図 2 は人手によるコーディングと，BERT によるクラスタリングに基づくグラフである。



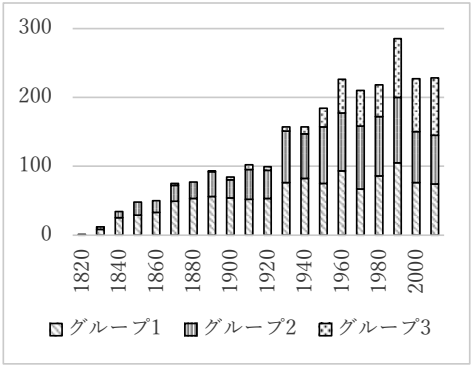


図 1 人手の分類を用いて得た用法変化　図 2 自動グループ化を用いて得た用法変化

図 1 と図 2 をみると，人手による分類と機械によるグループ化が酷似していることが分かる。BERT では文脈情報を考慮してベクトル化した各用例をグループ化しており，各グループは文脈的な類似度により求められている。その 3 つのグループが 1) 字義的なもの，2) 仮想的な状況における状態を表すもの，3) *better off* 構文と確実に分類できるものに対応をしており，また，各グループは 1 から 3 へと年代を経るにしたがって増えていった。この BERT に基づいた *better off* の用例のグループ化と，各グループの変遷は，人手による分類の結果ともかなりの程度一致したことから，BERT に基づいた *better off* の用例のグループ化の結果が，人手による分析を補強していると言える。

ただし，LLM の中身がブラックボックス化しており，内部で何が起こっているかよく分かっていない以上，現時点で，LLM は言語理論を直接的に検証する道具ではなく，言

語分析を間接的に補強する道具に留まっていると言える。しかし，LLM の性能が向上し，LLM の内部構造の解明が進み，用法基盤モデルとの類似点の精査が進んでいくと，さらに一歩進み，理論の検証に LLM が直接用いられる可能性も考えられる。特に，「意味は使用である（meaning is use）」という用法基盤モデルの考え方に基づくと，LLM が提供する文脈的な類似性に基づくクラスタリングは，文脈の特徴をより正確に捉えるという点で，人手による分類よりも量の面でも質の面でも上回っているように思われる。言語学の仮説の検証道具として LLM が使われる時代がそこまで迫っているかもしれない。

（参考文献）

Barlow, Michael and Suzanne Kemmer (2000) *Usage based models of language*. Chicago: The University of Chicago Press.

Bybee, Joan L. (1985) *Morphology: A study of the relation between meaning and form*. Amsterdam and Philadelphia: John Benjamins.

Crystal David (2011) *A little book of language*. New Heaven and London: Yale University Press.

Du Bois, John W. (2014) Towards a dialogic syntax. *Cognitive Linguistics* 25 (3): 359-410.

Fillmore, Charles J., Paul Kay and Mary C. O'Connor (1988) Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64: 501-538.

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

Goldberg, Adele E. (2019) *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton: Princeton University Press.

Langacker, Ronald W. (1988) A usage-based model. In Brygida Rudzka-Ostyn (ed.) *Topics in cognitive linguistics*, 127-161. Amsterdam and Philadelphia: John Benjamins.

永田亮・大谷直輝・高村大也・川崎義史 (2022) 「言語処理的アプローチによる *better off* 構文の定着過程の説明」言語処理学会第 28 回年次大会口頭発表.

中山俊秀・大谷直輝（編）（2020）『認知言語学と談話機能言語学の有機的接点−用法基盤モデルに基づく新展開』東京：ひつじ書房.

野村益寛（2014）『ファンダメンタル認知言語学』東京：ひつじ書房.

大谷直輝・中川裕・野元裕樹・長屋尚典（編）（近刊）『言語研究に潜む英語のバイアス』ひつじ書房.

大谷直輝 (2022) 「*better off* 構文の定着過程に関する認知言語学的考察」言語処理学会第 28 回年次大会口頭発表.

Stefanowitsch, Anatol and Stefan Th. Gries (2003) Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8: 209-243.

Tomasello, Michael (2003) *Constructing a language: A usage-based theory of language acquisition*. Cambridge and Massachusetts: Harvard University Press.

Vaswani, Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems* 30 (NIPS 2017).