語を構成単位としない統語論にむけて:
パターン束モデルを用いた文構造記述の理論と応用

吉川 正人
慶應義塾大学大学院/日本学術振興会
mail@yoshikawacademia.com

1　はじめに

近年、自然言語処理・言語教育・言語学など、言語に関わる様々な分野で、「連語 (Multiword Expressions: e.g., Sag, Baldwin, Bond, Copestake, & Flickinger 2002)」「定型表現 (Formulaic sequences: e.g., Wray 2002)」「構文 (Constructions: e.g., Goldberg 1995; Croft 2001)」といった、部分を構成する語や構成規則には意味・構造を還元できない、語よりも大きな構造体に関する議論が盛んに行われている。

このような大きな単位の構造体は、理論的にも技術的にも扱いが困難なため、長らく「例外」として無視されるか、アドホックな対処によって処理されてきたと言える。しかしながら、上に述べたような議論の高まりを前に、いつまでも「例外」扱いを続けるわけにもいかないのが現状である。

最も深刻なのは、何が「連語」「構文」で何がそうでない単なる複合的な表現なのかという判断が困難である点である。というより、より正確に言うならば、そもそも「連語性」「構文性」というのは、$\{0,1\}$ の二値で決まるものではなく、連続値 $[0,1]$ で表現されて然るべきものであると考えられる。

そうなると、そもそも基本単位を「語」と考え、その合成によって表現不可能な非構成的な単位を例外的なものとして扱うこと自体の正当性が疑わしくなってくる。むしろ、黒田 (2009b) などが指摘するように、事態その逆であり、常に言語単位は全体優先で (Cf. Sinclair 1991)、語レベルに分解可能 (= 構成的) であるのは極めて例外的であると考えた方が、言語事実にはよほど見合っているかもしれない。

本稿では、このような事態を鑑みて、従来の「『語』を構成単位とする統語論」を根本から解体し、語より大きな単位、即ち、「超語彙 (superlexical) 単位」を構成単位とした新たな統語論の一つの形を提案し、それが統語構造の記述に有益であり、また、言語処理技術への応用も可能であることを示す。

具体的には、「パターン束モデル (Pattern Lattice Model, PLM: e.g., 黒田・長谷部 2009; Kuroda 2009a; 吉川 2010)」の定義する「パターン」を文事例の「超語彙索引 (superlexical indices)」として利用し、データからボトムアップに構築された継承関係を持つパターンの集合によって文の構造もしくは類型を指定する記述モデルを提示する。

2　超語彙単位の重要性

本節では、1) 「語より大きな単位」= 超語彙単位の構造体に関するいくつかの先行研究を簡単に俯瞰し、2) その上で超語彙単位を構成単位とする統語論の必要性を確認する。

2.1　多義語のパラドクス

Taylor (2003) は、多義語の曖昧性解消の問題を取り上げ、以下の相反する二つの事実を対比した:

(1) a. n 個の語からなる文 $s = w_1 w_2 \cdots w_n$ の曖昧性は理論的には語 w_i の持つ語義数の総積となり、語の曖昧性が多ければ多いほど組み合わせ爆発的に増大する
b. ヒトは複数の多義語を含む文に対しても困難なく文意を解することができる

この事実が示唆する一つの可能性は、実は「語の多義性」というのは理論的な仮定でしかなく、ヒトが実際に処理しているのは、語よりも大きな単位であって、語は結局雑多な用法の寄せ集めに過ぎない、ということである (Taylor 2003: 653)。

2.2　構文文法

Goldberg (1995) は抽象的な「項構造構文 (Argument Structure Construction: e.g., 二重目的語構文)」には語に還元できない構文固有の意味 (Constructional meaning) が存在するとし、構文も語と同様な「形式と意味の対」としての記号体であるとした。これは、規模や抽象度の違いはあれ、「構文文法 (Construction Grammar)」に共有された想定である。

このような想定は、以下のような動詞の新奇な用法 (2a) や無意味動詞 (2b) の解釈を説明するのに有効である (Goldberg 1995: 29, 35):

(2) a. Pat sneezed the napkin off the table.
b. She topamased him something.

上のような文にも一定の解釈が与えられるという事実は、動詞等の構成語彙に文意の源泉を還元することが困難であることを物語っている。

2.3　連語

言語処理分野では記念碑的研究である Sag et al. (2002) を端緒に、「連語 (Multiword Expressions)」の研究が

盛んになっている (日本語の研究としては 首藤・田辺 2010)。連語は一般に信じられているよりずっとその数は多く、連語の対処なくして高精度の処理技術の達成はあり得ない。だが、連語は単なる「スペースを含む単語 (words with spaces: Sag et al. 2002: 2)」ではなく、定型性・逸脱性・生産性など様々な尺度において多様であり、その対処は困難を極める。

Sag et al. (2002) は連語を「語彙化された句 (lexicalized phrases)」と「慣用表現 (institutionalized phrases)」に、前者をさらに「固定表現 (fixed expressions)」「半固定表現 (semi-fixed expressions)」「統語上柔軟な表現 (syntactically-flexible expressions)」に分類し、分類毎に異なる対処法を与えている。

2.4 問題

以上のように、様々な論者が様々な観点から超語彙単位の構造体を言語の主要な構成要素として論じており、その重要性は明らかである。しかしながら、現状では以下 3 点の未解決問題が指摘できる:

(3) a. 超語彙単位を予めリストし尽くすことは困難であり、動的にその一覧を取得することが求められるが、データから超語彙単位を獲得するアルゴリズムは確立されていない
b. 超語彙単位には多くの場合「変異 (variation)」が存在するが、その対処は首藤・田辺 (2010) のようなフラットな記述が主で、階層的な性質は扱いきれていない
c. 超語彙単位は依然語とは別種の存在であると想定されており、両者を統合する可能性は考えられていない

本稿では、次節で紹介する「パターン束モデル (Pattern Lattice Model: e.g., 黒田・長谷部 2009; Kuroda 2009a)」が上記の問題、特に (3b, 3c) を解決し、(3a) の解決に一つの可能性を提供すると考える。

3 パターン束モデル (PLM)

3.1 概要

パターン束モデル (以下 PLM) とは黒田・長谷部 (2009) で提案されたヒトの言語知識とその構造化のモデルである。[1] PLM では、言語知識の構成要素は事例 (*exemplar*) e の集合 E と事例の索引であるパターン (*patterns*) p の集合 P とされる。またパターン集合 P は以下のアルゴリズムによって事例 e から得られる:

(4) a. 任意の分節モデル T (e.g., 単語分節) による e の分節化の結果を $T(e)$ とする
b. $T(e)$ の n 個の分節を 0〜n 個網羅的かつ再帰的に変項 X で置換する

事例 e に対するパターン集合を $P(e)$ とする。例えば、e = *John hit Mary*、T を単語分節とすると、

(5) a. $T(e)$ =[John, hit, Mary]
b. $P(e)$ = {(John, hit, Mary), (_, hit, Mary), (John, _, Mary), (John, hit, _), (_, _, Mary), (_, hit, _), (John, _, _), (_, _, _)}

となる ("_" は変項を表す)。

$P(e)$ は上位パターンが下位パターンに継承 (inherit) される形で `is-a` 関係の階層を持つ。継承関係の定義されたパターンの半順序集合を「パターン束 (*Pattern Lattice*)」と呼び、事例 e から得られたパターン束を $L(e)$ と表記する。

(5) の例における $L(e)$ をハッセ図に表すと Fig. 1 のようになる。尚 Fig. 1 では、簡略化のため連続する変項を単一の変項に縮約している (e.g., (John, _, _) → (John, _))

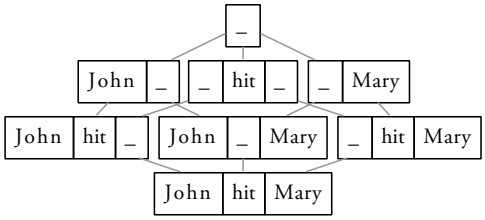

Fig. 1 e = *John hit Mary* の $L(e)$
(Pattern Lattice Builder (黒田・長谷部 2009) で作成)

複数の事例の集合 E から得られる $L(E)$ は個々の事例 e_i のパターン束 $L(e_i)$ を結合したものとるなり、従って $L(E)$ は膨大な反順序集合を形成する。例示のため、Fig. 2 に E ={*John hit Mary, John loves Mary*} の場合の $L(E)$ を提示する:

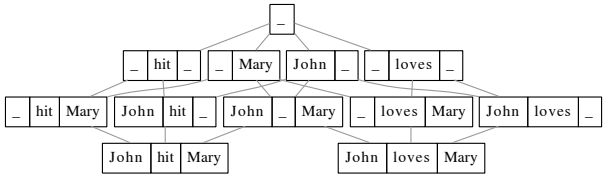

Fig. 2 E = {*John hit Mary*,*John loves Mary*} の $L(E)$
(Pattern Lattice Builder (黒田・長谷部 2009) で作成)

パターン束上のパターン p は (文) 事例 e の索引として機能する。PLM を想定した言語処理モデルでは、新奇入力の解析は事例ベースで行われ、再利用される既知の事例の探索にパターン束が活用されると考える (黒田・長谷部 2009; Kuroda 2009a; 黒田 2009b)。

3.2 下位概念

以下に PLM における重要な概念のいくつかを簡単に提示する:

(6) a. ランク **(rank)**
パターン $p = (s_1, s_2, \cdots, s_n)$ における分節 s_i のうち、変項 ("_") ではない分節の数; 即ち、p のランクを $r(p)$ とすると、
$$r(p) = |\{s_i \in p \mid s_i \neq \_\}|$$
b. 頻度 **(frequency)**
パターン p の頻度 $f(p)$ は、p を実現する (realize) 事例 e の延べ数で定義される; 即ち、

[1] 前提には「ヒトは一度見聞きした表現は全て覚えている」という完全記憶の仮説 (黒田 2010) があるが、本稿では紙面の都合上この問題には触れない。

$$f(p) = |\{e \in E \mid e \texttt{ realizes } p\}|$$

c. 有用性 **(utility)**
パターン p の有用性 $u(p)$ は、何らかの評価尺度 (関数) u によって計測される p の「再利用可能性 (recyclability)」(≈ 生産性) である。[2)]

d. 重ね合わせ **(superposition)**
ランク k のパターン p^k はランク $k-1$ のパターン群 P^{k-1} の部分集合 $Q := \{q \in P^{k-1} \mid p^k \texttt{is-a } q\}$ の「重ね合わせ (superposition: Kuroda 2009a)」として定義される (ただし $k \neq 1$)

e. (超) 語彙パターン **((super)lexical patterns)**
ランクが 1 のパターンを「語彙パターン」ランクが 2 以上のパターンを「超語彙パターン」と呼ぶ (黒田・長谷部 2009; Kuroda 2009a)

(5) の例において最上位のパターン (_) のランクは 0 である。一般に、分節数 n の事例からは、ランクが 0〜n のパターンが生成される。

パターン (John, _, Mary) はパターン {(John, _), (_, Mary)} の重ね合わせであり、(John, hit, Mary) は { (John, hit, _), (John, _, Mary), (_, hit, Mary) } の重ね合わせである。また事例 *John hit Mary* の構造は Fig. 1, Fig. 2 で図示した階層で与えられる。

このような性質から、PLM は上述の (3b, 3c) に対し明快な解決を与える。

3.3 PLM の利点と問題点

PLM のアルゴリズムは、分節モデルさえ与えられれば、任意の事例に対して網羅的に可能な超語彙単位 (= パターン) の全集合を与えることができる。従って、PLM の利点は 1) 事前知識の寡少性; 2) 生成可能なパターンの網羅性の二点だと言える。

ただ、上述のアルゴリズムだけではパターン p の有用性 $u(p)$ を与えることはできず、有用なパターンとそうでないパターンを区別できないため、有意義な記述を与えることは困難である。従って、効果的な有用性尺度 u を定義し、パターンに重みづけを行う必要があるが、現時点でこれは未達成の課題である。次節でその可能な候補の一つを提示する。

4 PLM を用いた構造記述の理論と応用

4.1 統語論

パターン束は (可能な) パターンの集合とその継承関係を定義するのみであって、パターン (を介した事例) の合成規則や制約を明示的に与えるものではない。従って、「統語論」を「部分から全体をくみ上げる操作」であるとするならば、パターン束は統語論の「入力」となるだけであり、統語論は独立に定義される必要がある。

ただ、統語論を単に部分と全体の関係を体系的に指定するものであると考えるのであれば、PLM は文事例 e の統語構造に対して以下のように言うことができる:

(7) (6d) より: 文事例 e の構造は PLM の定義するパターンの「重ね合わせの階層 (the hierarchy of superpositions)」として表現される

これに有用性尺度 (関数)u の定義によりパターンの重み付けが加えられれば有意義な構造記述が可能となるはずである。従って、PLM を利用した文事例 e の統語構造の指定は、以下の形をとることになると言える:

(8) パターンの重み付き重ね合わせ階層 (The weighted hierarchy of superpositions of patterns)

もちろんこのような構造の特徴づけは「重み」が何であるか定義して初めて内実を持つものである。しかしながら、(8) を構造記述の基礎原理と捉えれば、後は任意の重み付けの尺度を定義し適用することで様々な観点から多様な構造を規定できることになる。従って、現段階で統語構造記述に対し PLM が貢献可能なのは、「記述単位 (= 超語彙単位) の体系的な規定」と「構造記述のひな型の提供」の 2 点である。

4.2 有用性尺度 (関数) u の候補

ここで、イディオム原則 (Idiom Principle, Sinclair 1991) に示されるような、自然言語の「全体優先」の性質を鑑みて、Kuroda (2009a), 黒田 (2009b) の提案する PL 上の意味解釈モデルである Simulated Parallel Error Correction (SPEC) を参考に、パターンの階層に事例探索における [下から上] の優先順位を想定する。つまり、パターンのランクと優先順位が比例するものと考える。紙面の都合上 SPEC の詳細には触れられないが、簡単にその概要を述べると、新規事例 e の解釈は e が k 個に分節化されるとして、

(9) a. e から得たパターン $P(e)$ のうちランク k のパターンを PL 上で探索し、得られた事例の解釈を e の解釈に転用する
b. a が失敗した場合、ランク $k-1$ のパターン群を PL 上で探索し、得られた事例群の論理和を e の解釈とする
c. 以降、事例集合が得られなかった場合、ランクを 1 ずつ減らしていき、探索範囲を広げ同様の探索を繰り返す

これはより高ランクのパターンで事例 (の集合) をうまく収集できる場合、低ランクのパターンにへ伝播が起こらないということを意味し、全体優先の性質を自然に体現する。

以上から、パターン p の有用性 $u(p)$ を事例探索に利用された回数として規定できる可能性が浮上する。これは、事例 e_i が実現するパターン集合 $P(e_i)$ のうち、e_i 以外の事例も実現しているパターン $= P(e_i) \cap P(e_j) (i \neq j)$ におけるランクの最大値を $rankmax(e_i)$ とすると、以下のように定式化できる:[3)]

(10)
$$u(p) = |\{e_i \in E \mid e_i \texttt{ realizes } p \text{ かつ}$$

[2)] 現時点で u は未定義である。尚、Pattern Lattice Builder (黒田・長谷部 2009: http://www.kotonoba.net/rubyfca/) では、u に (ランク毎の) 頻度の正規化を採用し、$u(p)$ を正規化頻度、即ち頻度の z-スコアとしている。

[3)] このような有用性の定義は、傳康晴氏 (千葉大学) から頂いたコメントを元に考案された。この場を借りて謝意を表したい

$$r(p) = rankmax(e_i)\}|$$

これは文字通り p の有用性 (utility) となっている。

4.3 問題

しかし、PLM を用いて生データを解析し構造記述を得るにあたって、解決しなければならない問題が以下の 3 点ある。1 つは理論的なもので、1 つは理論的かつ技術的、そしてもう一つは純粋に技術的な問題である:

(11) a. パターン p にラベルが付与されない $= p$ が「何であるか」の解釈が困難である
b. 事例の分節数 n に応じて 2^n 個のパターンが生成されるため長い (= 分節数が多い) パターンに対しては組み合わせ爆発的にパターン数が増大し処理が困難である
c. 大規模コーパスの全ての事例からパターンを生成すると膨大な処理コストとデータ量となるたえ事実上解析不可能である

(11a) に関しては解決は難しいが、そもそも有限個のラベルをパターンにあてがうことは困難であり、得られたパターン集合に対する分析の段階で後付け的にラベル付けを行った方が有意義である可能性が高い。その前段階として、分布類似度などを利用したパターンのグループ化などを行うことは効果的かもしれない。

(11b) に関しては、おおよそ分節数が 7 を超えると処理コストが格段に高くなることが報告されている (黒田・長谷部 2009)。この対処としては、分節数がある閾値 l (e.g., 7) を超える事例に対しては n-gram に分割するなどして断片化し、そこからパターンを生成する、というような前処理を行うのが望ましい (Cf. 吉川 2010)。

このような対処には理論的な含意もある。即ち:

(12) 統語構造を指定するのに十分な単位は、l 語の範囲に収まるような局所的なユニットである。

ということである。しかしこう想定することの問題としては、長距離依存 (long-distance dependency) など大域的な現象を扱えないということがあげられる。この種の問題は重ね合わせの制約をうまく定義することで解決できる可能性もある。

(11c) に関しては完全に技術的な問題である。この解決には、1) 生データの効率的な解析アルゴリズムの開発; 2) 解析済みデータの効率的かつ省スペースな貯蔵法の開発、が不可欠である。

4.4 PLM パーサーの実装計画

上記の (11c) さえ解決されれば、十分なサイズの事例集合 E を解析し膨大なパターン束 $L(E)$ を獲得することができる。これを元に、PLM ベースの解析器、PLM パーサーを実装することが可能である。

PLM パーサーは、任意の入力 e_{new} に対し $L(e_{new})$ を構築し、$p^* \in L(e)$ を $L(E)$ 上から探索し、マッチするパターン p があった場合それを実現する事例の和集合と、$f(p)$ および $u(p)$ を取得し、取得した情報で $L(e_{new})$ を更新する。マッチするパターンの存在しなかったパターンは破棄されるか非活性化される。最終的に解析器は得られた事例集合と更新された $L(e_{new})$ を返す。

(11a) の問題があるため得られた結果が何を意味しているのかには解釈が必要となるが、パーサーが実装され解析結果が蓄積されれば、ある種の結果に一定の解釈を与えられるようになる可能性は見込める。また、そのような解釈の不要な機械翻訳などの形式ベースの処理であれば、効力を発揮する可能性は高い。

5 結語

本稿では、1) 近年高まりを見せる超語彙単位の構造体に関する議論を概観し、2) 従来の「語」を構成単位とした統語論に代わる超語彙単位を構成単位とした統語論の必要性を訴え、3) その実現のための道具立てとしてパターン束モデル (PLM) が有効であることを論じた。ただ、本稿では実際のデータの解析例を示し従来の手法よりも PLM の記述が有益であることを示すことはできなかったため、(11) に述べたような問題を解決し、実データの解析を進めていくことが当面の課題である。

参考文献

Croft, W. 2001. *Radical construction grammar: Syntactic theory in typological perspective.* Oxford; New York: Oxford University Press.

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

Kuroda, K. 2009a. Pattern lattice as a model for linguistic knowledge and performance. In *Proceedings of the 23rd pacific asia conference on language, information and computation* (pp. 278–287).

黒田航. 2009b. パターンのラティス下での疑似並列エラー修復に基づく文意の構築. 日本認知科学会第 26 回大会発表論文集 (pp. 236–237).

黒田航. 2010. 超常記憶症候群の理論的含意. 日本認知科学会第 27 回大会発表論文集 (pp. 789–792).

黒田航・長谷部陽一郎. 2009. Pattern Lattice を使った (ヒトの) 言語知識と処理のモデル化. 言語処理学会第 15 回大会発表論文集 (pp. 670–673).

Sag, I., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the third international conference on computational linguistics and intelligent text processing* (pp. 1–15).

首藤公昭・田辺利文. 2010. 日本語の複単語表現辞書: JDMWE. 自然言語処理, *17*(5), 51.

Sinclair, J. 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Taylor, J. 2003. Polysemy's paradoxes. *Language Sciences*, *25*(6), 637–655.

Wray, A. 2002. *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.

吉川正人. 2010. 「語」を超えた単位に基づくコーパス分析に向けて: パターンラティスモデル (PLM) とその有用性. 藝文研究, *98*, 221-207.