用法基盤モデルに基づいた複合語形成の生産的パターンの抽出

浅尾仁彦 (京都大学大学院)
asaokitan@ling.bun.kyoto-u.ac.jp

1　はじめに

本研究[1]の目的は，構文文法 [7, 8] および用法基盤モデル [12, 13, 14, 11, 17] の考え方を簡単なシミュレーションモデルとして実装することで，日本語複合語 (複合動詞および動詞由来複合語) の生産的パターンをコーパスから自動抽出することである。

「構文」は構文文法の記述の基礎単位でありながら[2]，その認定は多くの場合直観に依存してきた。本研究はコーパスベースで「構文」を自動的に認定するひとつの方法となる。と同時に，人間が構文を経験からどのように学習しているかという問題にも示唆を与える。

2　複合語の生産的パターン

本節で，複合動詞と動詞由来複合語に見られる生産的パターンについての先行研究を概観する。

複合動詞は，統語的複合動詞と語彙的複合動詞とに二分する分析がよく知られている [10]。

(1) a. **統語的複合動詞**
　　食べ始める，増え続ける，…
　　b. **語彙的複合動詞**
　　飛び込む，消え去る，…

統語的複合動詞を形成する「-始める」「-続ける」などの動詞は，基本的に任意の動詞と結びつきうるのに対して，語彙的複合動詞にはそのような生産性はない。

また，「名詞＋動詞→名詞」の複合語 (動詞由来複合語) は，(直接) 目的語の複合した内項複合語 (2a) と，それ以外の付加詞複合語 (2b) に分けられることが知られている [9]。

(2) a. **内項複合語**
　　魚釣り，皿回し，ゴミ拾い，…
　　b. **付加詞複合語**
　　手作り，日焼け，機械編み，…

内項複合語は，意味的に整合するかぎり任意の名詞と動詞を組み合わせうるのに対し[3]，付加詞複合語には限定的な生産性しかない。

ここで留意すべきは，「統語的複合動詞」の生産性と「内項複合語」の生産性とは質が異なり，前者は特定の少数の後項に関する性質であるのに対し，後者は前項・後項ともに語彙的な制限はないことである。また，付加詞複合語であっても「-生まれ」のように，特定の後項が高い生産性をもつ場合がある [16]。

このような事実を簡潔に表現するうえで，言語知識を，さまざまな抽象度をもった「構文」の集合とみる立場が有効である。複合動詞，動詞由来複合語に関する話者の知識は，それぞれ (3) (4) のような構文の集合として記述できる [4]。

(3) a. **統語的複合動詞**
- [[　]$_V$-始める / BEGIN TO [　]$_V$]$_V$
- [[　]$_V$-得る / BE ABLE TO [　]$_V$]$_V$

[1] 本研究を進めるにあたって中川奈津子氏 (京大), 黒田航氏 (NiCT) をはじめとする多くの方の助言，協力を頂いた。この場を借りて感謝申し上げる。本研究の誤りは全て著者の責任である。

[2] 構文を背後にある抽象的な文法規則によって生じる副次的現象とみなす Chomsky 理論と異なり，構文文法では構文は話者がもつ言語知識の記憶の単位となる。構文は形式と意味とのペアからなり (スラッシュで区切って示す。ただし本論のシミュレーションでは意味の側面は扱わない)，(i) のようにさまざまな抽象度をもちうる。

(i) a. [[　]$_{NP1}$ [　]$_V$ [　]$_{NP2}$ [　]$_{NP3}$ / [　]$_{NP1}$ CAUSE [　]$_{NP2}$ TO RECEIVE [　]$_{NP3}$ BY [　]$_V$ING]
　　b. [[　]$_V$-ed out / WORN OUT FROM TOO MUCH [　]$_V$ING]
　　c. [cat / CAT]

構文の定着度には程度差があり，言語理解・産出において頻繁に用いられるものは高い定着度をもつと考える。本稿で用いる「生産的パターン」というのは，定着度の高い構文とほぼ等価である ((ic) のような完全に語彙的なものは頻度が高くても通常は「生産的」とは言わないが，ここでは同列のものとして捉えている)。

[3] ただし，内項複合語であっても「?*服着 (ふくき)」など容認しがたい例もあり，内項複合語の可否には音韻的条件なども関わっていると思われる (杉岡洋子, personal communication)。

- etc.

b. **語彙的複合動詞**
   - [見破る / DETECT]$_V$
   - [切り倒す / CUT DOWN]$_V$
   - etc.

(4) a. **生産的な内項複合語**
   - [[ ]$_N$-[ ]$_V$/ [ ]$_V$ING OF [ ]$_N$]$_N$

b. **生産的な付加詞複合語**
   - [[ ]$_N$-生まれ / BORN IN [ ]$_N$]$_N$
   - [[ ]$_N$-沿い / ALONG [ ]$_N$]$_N$
   - etc.

c. **語彙的複合語**
   - [歯止め / RATCHET]$_N$
   - [秒読み / COUNTDOWN]$_N$
   - etc.

しかしながら，これらの記述は直観に基づいて書かれたもので，おのおのの構文を認定するための客観的手続きを欠いている。そこで，コーパスを利用して生産的パターンを認定することを考える。

これまでに，コーパスに基づいた生産性の指標である Baayen の $\mathcal{P}$ [5] を用いることで，統語的複合動詞と語彙的複合動詞を識別することが可能であることが示されている [2, 3, 15] [4]。しかしながらこの指標は個々の接辞の生産性を計算するためのものであり，例えば [[ ]$_N$-[ ]$_V$] のようなパターンが一般に生産的かどうかを記述する目的には適さない。

次節で，任意のパターンについてその生産性を記述するためには，用法基盤モデルに基づいたシミュレーションが有用であることを論じ，実際のコーパスデータに基づいてシミュレーションを行う。

[4] Baayen の生産性 $\mathcal{P}$ は，接辞の総トークン数 (その接辞をもつ語ののべ語数) を N，そのうちただ一度だけ出現した語 (hapax legomena) の数を n_1 として，次の式で求められる。

$$\mathcal{P} = \frac{n_1}{N} \tag{1}$$

3 シミュレーション

3.1 概要

用法基盤モデルでは，言語知識は経験からの共通性の抽出によって成立すると考える。例えば，[[ ]-始める] という構文は，「食べ始める」「読み始める」のような実例に接することによって成立し，定着度を増していく。

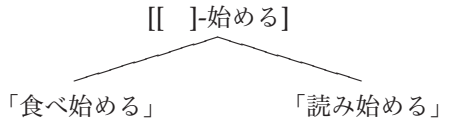

図 1

このとき重要なのは，構文はできるかぎり具体的なものが優先されることである [13, 11]。例えば，用例が全て「-始める」という形をしているかぎり，定着度を増すのは [[ ]-始める] という構文であり，より抽象的な [[ ]-[ ]] という構文ではない[5]。

これを踏まえたシミュレーションの基本的な仕組みは以下のようになる。

(5) a. コーパスでの実際の使用頻度に基づいて，モデルに複合語を 1 つずつ与えていく。
   b. 与えられた複合語どうしに共通性が見つかった場合，その共通性を抜き出した構文が成立する。
   c. 与えられた複合語が，既に成立している構文の実現例である場合，その構文の定着度が増加する。
   d. 上記の b. および c. において，どの構文が選ばれるか複数の可能性がある場合は，最も具体的な構文が選択される。

この方法で得られた構文の集合が，話者の言語知識を表現しているものとみなすことができる。

3.2 手順

コーパスとして『CD-毎日新聞 '95 データ集』を用いた。コーパスの規模は 29,547,574 語 (句読点など含む) である。まず，複合語の頻度情報を得るため，

[5] これは，最小一般化 (minimal generalization) [1] と呼ばれているアルゴリズムと同等である。

形態素解析ソフト MeCab 0.97，形態素解析用辞書 UniDic 1.3.9 [6] を用いてコーパスから複合動詞および動詞由来複合語を抽出したうえ，人手でのチェックを行った[6]。

抽出された複合動詞はタイプ数 5,820 トークン数 190,525，動詞由来複合語はタイプ数 10,120 トークン数 88,996 である。漢字表記の揺れは，UniDic の機能を用いて吸収してある。

次に，抽出された複合動詞からランダムな順序で 50,000 トークンを抜き出しシミュレーションモデルに通す試行を 5 回行った。同様の手順を動詞由来複合語に対しても繰り返した。

定着度の具体的な数値については，構文の成立時の定着度は 1，また構文の新たな実現例が現れるごとに定着度が 1 増加するとした。なお，構文を選択する際に，同じ抽象度の候補が複数ある場合は，定着度の高いほうを選択することとした (定着度が等しい場合はランダムに選択した)。

3.3 結果

高い定着度を得た順に 10 位までを表 1 に示した (5 回の試行の平均値)。

定着度の高いものはほとんど全て語彙的な複合語が占めているが，動詞由来複合語の 10 位に抽象的な [[　]-[　]] が現れている。語彙的なものを除き，空所をもつ構文のみを定着度の高い順に並べると，表 2 のようになった[7]。

複合動詞

順位	構文	定着度
1	[出-来る]	8 135.4
2	[仕-舞う]	2 181.8
3	[取り-組む]	694.6
4	[呼び-掛ける]	589.6
5	[受け-入れる]	566.6
6	[繰り-返す]	544.8
7	[盛り-込む]	476.6
8	[打ち-出す]	378.8
9	[受け-取る]	363.4
10	[見-直す]	318.6

動詞由来複合語

順位	構文	定着度
1	[気-持ち]	2 038.6
2	[先-駆け]	1 404.4
3	[役-割]	1 218.6
4	[手-続き]	1 013.4
5	[年-寄り]	703.6
6	[間-違い]	592.2
7	[枠-組み]	480.2
8	[町-作り]	473.8
9	[値-上げ]	462.0
10	[[　]-[　]]	412.8

表 1　定着度上位の構文

3.4 分析

表 2 の結果と，(3)(4) で示した記述とを比較すると，以下の点が指摘できる。

(6) a. 複合動詞に関しては，[[　]-始める]，[[　]-続ける] などの統語的複合動詞が高い定着度をもつことが正しく検出できている。ただし，語彙的複合動詞である [[　]-込む] と [[　]-上げる] も上位に現れており，統語的複合動詞と語彙的複合動詞を完全に分離することはできていない[8]。

b. [[　]-[　]] という完全に抽象化された構文は動詞由来複合語で高い定着度を示しており，(3)(4) の記述を正しく反映している。ただし，シミュレーション結果における複合動詞の [[　]-[　]] の定着度は高すぎるかもしれない。

c. 動詞由来複合語では，[[　]-[　]] のほかに [[　]-作り] [[　]-振り] などが上位に来ている。これらは (4b) に該当するパターンである [3] [9]。

[6] 動詞由来複合語の抽出にあたっては，ノイズを減らすために，格助詞，係助詞，またはコピュラが後続するものにデータを限定している。

[7] 表 2 に現れたそれぞれ 10 個の構文について，5 回の試行の各ペア (${}_5C_2 = 10$ 通り) について Pearson の積率相関係数を求めたところ，複合動詞で 0.983 以上，動詞由来複合語で 0.994 以上であった。従って表 2 の結果はきわめて安定した結果と言える。

[8] 「-込む」と「-上げる」は語彙的複合動詞としてはタイプ頻度が高い [3]。Baayen の生産性 $\mathcal{P}$ は「-込む」「-上げる」のタイプ頻度が高いにもかかわらず生産性は低いことを正しく反映できるが，今回のシミュレーションではその点を反映できていないことがわかる。

[9] [[　]-作り] の用例の多くは「町作り」「法案作り」のように直接目的語を取っているので，(4a) の生産的パターンの事例であるように見える。しかしながら，「-を作る」と「-作り」では選択制限に違いがあり (黒田航, personal communication)，また内項複合語では通常生じない連濁

複合動詞

順位	構文	定着度
1	[[　]-始める]	239.8
2	[[　]-続ける]	210.8
3	[[　]-出す]	181.2
4	[[　]-込む]	150.6
5	[[　]-[　]]	134.4
6	[[　]-合う]	125.2
7	[[　]-過ぎる]	110.2
8	[[　]-上げる]	81.8
9	[[　]-切れる]	81.2
10	[[　]-切る]	79.8

動詞由来複合語

順位	構文	定着度
1	[[　]-[　]]	412.8
2	[[　]-作り]	401.0
3	[[　]-振り]	287.2
4	[[　]-通り]	209.4
5	[[　]-付き]	201.0
6	[[　]-入り]	198.8
7	[[　]-向け]	196.8
8	[[　]-行き]	118.0
9	[[　]-好き]	111.8
10	[[　]-生まれ]	95.6

表 2　定着度上位の構文 (空所をもつもの限定)

d. 後項のみ抽象化した構文 (例えば [打ち-[ ]] のようなもの) の定着度が高くないことを正しく反映できている。

4　おわりに

本発表ではコーパスから構文を抽出するための簡単なシミュレーションを示した。今回表 1, 2 で得られた構文は (3)(4) と比較すると品詞や意味の情報を欠いているが，それらを素性の束として与えることで同様のシミュレーションを行うことができるので，今後の検討課題としたい。また，シミュレーション結果が直観をどの程度的確に反映しているか評価する手続きも必要となってくる。

参考文献

[1] Adam Albright and Bruce Hayes. Modeling English past tense intuitions with minimal generalization. In *Proceedings of the ACL-02 workshop on morphological and phonological learning*, Vol. 6, pp. 58–69. Association for Computational Linguistics, 2002.

[2] 淺尾仁彦. 分析性からみた複合動詞. 形態論・レキシコンフォーラム 2006 発表資料, 2006.

[3] 淺尾仁彦. 複合語の生産性と文法的性質. 日本言語学会 第 134 回大会 予稿集, pp. 416–421. 日本言語学会, 2007.

[4] 淺尾仁彦. 構文形態論による日本語動詞複合語の記述. 形態論・レキシコンフォーラム 2008 発表資料, 2008.

[5] R. Harald Baayen. Quantitative aspects of morphological productivity. In Geert Booij and Jaap van Marle, editors, *Yearbook of Morphology 1991*, pp. 109–149. Kluwer, Dordrecht, 1992.

[6] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源: 形態素解析用電子化辞書の開発とその応用. 日本語科学, Vol. 22, pp. 101–122, 2007.

[7] Adele E. Goldberg. *Constructions: a Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago, 1995.

[8] Adele E. Goldberg. *Constructions at Work*. Oxford, New York, 2006.

[9] 伊藤たかね, 杉岡洋子. 語の仕組みと語形成, 英語学モノグラフシリーズ, 第 16 巻. 研究社, 2002.

[10] 影山太郎. 文法と語形成. ひつじ書房, 1993.

[11] 熊代文子. 認知音韻論. 吉村公宏（編）, 認知音韻・形態論, pp. 3–78. 大修館書店, 2003.

[12] Ronald W. Langacker. *Foundations of Cognitive Grammar. Volume 1: Theoretical Prerequisites*. Stanford University Press, Stanford, 1987.

[13] Ronald W. Langacker. *Concept, Image, and Symbol: the Cognitive Basis of Grammar*. Mouton de Gruyter, Berlin and New York, 1990.

[14] Ronald W. Langacker. A dynamic usage-based model. In M. Barlow and S. Kemmer, editors, *Usage-based models of language*, pp. 1–63. CSLI, Stanford, 2000.

[15] 松村一登. 複合語の生産性といわゆる統語的／語彙的の区別—コーパスに基づく考察. 日本言語学会第 134 回大会予稿集, pp. 378–383. 日本言語学会, 2007.

[16] Yoko Sugioka. Regularity in inflection vs. derivation: rule vs. analogy in deverbal compound formation. *Acta Linguistica*, Vol. 45, pp. 231–253, 1996.

[17] 上原聡, 熊代文子. 音韻・形態のメカニズム —認知音韻・形態論のアプローチ—, 講座認知言語学のフロンティア, 第 1 巻. 研究社, 2007.

が生じる [16, p.250] という点からも，[[　]-作り] という構文は普通の内項複合語でなく，[[　]-[　]] と別個に個別に記憶されているとする十分な根拠がある。