日本語の動詞性複単語表現辞書

A Dictionary of Japanese Verbal Multiword Expressions

首藤　公昭†　田辺　利文†
†福岡大学大学院　工学研究科　電子情報工学専攻
{ shudo, tanabe }@tl.fukuoka-u.ac.jp

abstract

日常の自然言語文には意味・構文上の構成性(compositionality)に問題の有る相当数の複単語表現（MWE; Multi-Word Expression）が使われており、自然言語処理（NLP）におけるネックとなっている。 また、強い共起性で結ばれた単語からなる複単語(定型)表現も構文解析の効率や曖昧さ低減の観点から重要であるが十分には整備されていない。筆者らは以上の認識から日本語 MWE 辞書の構築を行ってきているが、本稿ではそのうち、「出席を取る」、「手を染める」、「聞くともなく聞く」、「浴びる程飲む」、「旗色を鮮明にする」、「負うた子に教えられる」、「犬も歩けば棒に当たる」など、動詞相当の MWE に絞って辞書の概要を報告する。これらの収録見出し数は約 50,000 である。

key words：

複単語表現(MWE)，軽動詞構文(LVC)，支援動詞構文(SVC)，機能動詞結合，動詞句，慣用句，決まり文句，連語，コロケーション，成句，語結合，派生語，複合語，格言，諺，フレーズ翻訳，予測変換，構文解析，意味解析，単語 n-グラム，日本語音声認識

1. はじめに

(Sag et al., 2002)がきっかけとなって、自然言語処理（NLP）における複単語表現（MWE; Multi-Word Expression）の重要性が近年、改めて認識されるようになった。ACLは2003年以降、MWEのworkshopをほぼ毎年開催しており、非構成的(non-compositional)なMWE辞書の構築を目指す研究を中心に活発な議論が行われている。しかし、現状では充分な網羅性を備えた成果は得られていない。筆者らは、近年主流となっている機械学習等、統計的手法の有効性は限定的であり、日常の自然言語を対象とする「深い」NLP の実現には人の内省による辞書構築が不可避であると考えており、1970 年前後から日本語を対象とした MWE の収集・整理を行ってきた。(首藤他 2010; Shudo et al., 2011) 本稿では、特に、動詞相当 MWE に特化した辞書の概要を報告する。

2. 関連研究

日本語 MWE に関する研究としては、古くから国語学の領域で人の利用を目的とした慣用句辞典類の編纂が種々行われてきた。(尾上(監修), 1993; 三省堂編修所(編), 1999; 白石, 1992; 田島, 2002; 米川他, 2005 など) しかし、これらの研究では、各表現の異表記、内部構造、どこまで変化が許されるかという柔軟性などは殆ど考慮されていない。これに対し、本研究では各表現に異表記、内部構造、内部修飾可能性の情報を体系的に記載している点に大きな特徴がある。

これまで、NLP の立場で日本語の動詞性 MWE に焦点を当てた研究には、名詞と動詞の共起頻度や名詞の意味の特異性を考慮してコーパスからの自動抽出を試みた(新納他, 1995a; 1995b)、支援動詞構文の言い換え問題を論じた(Fujita et al., 2004)、VV 型複合動詞の意味の曖昧さ解消について考察した(Uchiyama et al., 2003)

などがある。しかしながら、十分な網羅性と精密さを備え、一般的なNLPに耐える言語資源は未だ構築されていない。海外での類似の研究には英語の動詞・不変化詞構文(VPC)辞書を考察した(Villavicencio, 2003)、英語の動詞・名詞構文(VNC)の意味の多義性を考察した(Cook et al., 2008)、13,000個のエストニア語の動詞性複単語表現(MWV)データベースを構築した(Kaalep et al., 2008)などがある。

A	B	C	D	E	F
がらがらとおとをたててくずれる	がらがら-と-おと-を-たてて-て.くずれる	ガラガラ-と-音-を-立て-て-崩れる	Verb_VteV	[[[Oto][[Nwo]V23]]te]*V30	
あくにみをそめる	あく-に-み-を.そめる	悪-に-身-を-染める	Verb_(Np)^{2}V	[Nni][[Nwo]*V30]	
あくまにたましいをうる	あくま-に-たましい-を.うる	悪魔-に-魂-を-売る	Verb_(Np)^{2}V	[Nni][[Nwo]*V30]	
あすにきぼうをつなぐ	あす-に-きぼう-を.つなぐ	明日-に-希望-を-繋ぐ	Verb_(Np)^{2}V	[Nni][[Nwo]*V30]	
あたまにしもをおく	あたま-に-しも-を-おく	頭-に-霜-を-置く	Verb_(Np)^{2}V	[Nni][[Nwo]V30]	
あとにおをひく	あと-に-お-を.ひく	(後/アト)-に-尾-を-引く	Verb_(Np)^{2}V	[Nni][[Nwo]*V30]	
うえにあぐらをかく	うえ-に-あぐら-を.かく	上-に-胡座-を-かく	Verb_(Np)^{2}V	[Nni][[Nwo]*V30]	「の」連体修飾

図1　辞書の一部

3. 収録表現

筆者らは新聞記事、小説、雑誌記事、各種解説記事などの生データから次の二つの基準で動詞性 MWE を収集し、既存の事典類も参考にしながら確認・補強を行った。

1) イディオム性

要素単語から全体の意味を規則で導くことが難しいと思われる(non-compositionalな)表現、例えば、「油を売る」、「手を焼く」、「長い-(目/眼)-で-見る」、「手-の-内-を-読む」、「(爪/ツメ)-の-(垢/アカ)-を-煎じ-て-(飲/呑)む」、「(無/亡)き-者-に-する」、「一-皮-剥ける」、「枯(れ)-木-に-花-が-咲く」、「頭-の-上-の-(蝿/ハエ)-を-(追/遂)う」等々である。 通常の慣用句辞典類に収録されている表現はこの基準を満たす典型例と考えられる。

2) 高い単語間共起確率

単語間共起確率が高い表現。例えば、「信用-する-に-足る」、「(餌/エサ)-を-啄む」、「エスプリ-を-利かす」、「常識-に-外れる」、「シクシク-泣く」、「(渋(渋/々)/シブシブ)-折れる」、「逃げる-様-に-去る」、「論ずる-に-足る」等々である。[1]

本辞書は、動詞性のいわゆる慣用句、動詞性の決まり文句、動詞性のことわざ、動詞性の格言、動詞性の故事成句、動詞性の慣用的比喩表現(換喩、張喩等を含む)、支援動詞構文(SVC)、軽動詞構文(LVC)、非構成的複合動詞を収録している。

4. 統計的性質

Web 上の日本語 200億文コーパス中の N-gram 頻度データ LDC2009T08 (Kudo et al., 2009)を用いて、本辞書が収録する[名詞＋格助詞＋動詞]型表現の統計的性質を調べた結果、各[名詞＋格助詞]部に対して出現頻度の相対的に高い[動詞]が選ばれていること、また、[動詞]部のエントロピーの相対的に小さい[名詞＋格助詞]部が選ばれていることが確認されており、前節 3.の基準2)に関連する本辞書の有効性が推定される。(Shudo et al., 2011; 田辺他, 2012) また、収録表現の新聞記事における延べ出現頻度も10文中 7 か所程度と、かなり高い事が分っている。(首藤他, 2010)

5. 記載情報

[1] 以上の基準 1)、2)は排他的ではなく、双方を同時に満たす表現は数多い。

本辞書は約 50,000 行、6 欄（A 欄～F 欄）の MS-Excel 形式に作成されている。 図1に辞書の一部を示す。

5.1 平仮名ベタ見出し（A 欄）

音に基づいている。例えば，「九-死-に-一-生-を-得る」に対して「きゅうしにいっしょうをえる」、「きゅうしにいっしょうをうる」を別見出しとして立てている。

5.2 構成単語間の境界（B 欄）

ハイフン-あるいはドット.で単語(接辞)間境界を示す。ドットはこの位置に別の単語列（例えば動詞）が介在する可能性がある事を示す。活用語尾は切り離していない。

5.3 漢字、片仮名などの異表記（C 欄）

字種と表記の揺れ情報を与える。例えば、「振(る)っ-てる」のカッコ( )は文字の任意性、「(已/止)む-無き-に-(到/至)る」のカッコ( )と斜線/は文字の選択肢を与える。B 欄、C 欄を合わせれば、多数の異表記に対応でき、異表記も数に入れると、本辞書は 300,000 表現程度をカバーしていると推定される。

5.4 文法的な機能と種別（D 欄）

表現全体の文法的な機能が動詞性である事をコード Verb で表わし、大まかな下位分類をアンダースコア_を介してその後ろに記す。具体的には現在、以下のような数十種が区別されている。

Verb_(Np)V： ex.「賃-を-払う」、「名声-が-鳴(り)-響く」、「(腹/ハラ)-に-据え-兼ねる」、「血-眼-で-(捜/探)す」、「麻酔-から-覚める」

Verb_(Np)^{2}V： ex.「頼み-の-綱-が-切れる」、「(鳶/トンビ)-に-油-揚げ-を-攫わ-れる」、「体面-を-気-に-する」、「手-に-職-が-(有/在)る」、「血-が-頭-に-(上ぼ/昇)る」、「不安-が-頭-を-過(ぎ)る」

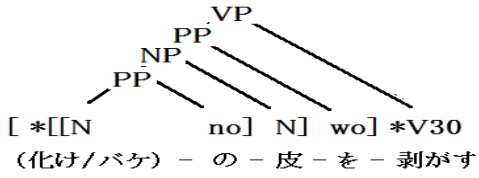
Verb_(Np)^{3}V： ex.「上手-の-手-から-水-が-(漏/洩)る」、「一-年-の-計-は-元旦-に-(有/在)り」

Verb_V^2： ex.「通じ-合える」、「掴み-掛(か)る」、「(付け/ツケ)-回す」

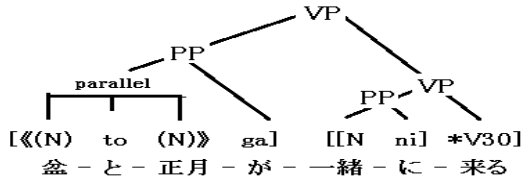
・・・・・・・・

5.5 表現の構文構造（E 欄）

表現内の依存（係り受け）構造を2項括弧表現[ ]で与える。ただし、概念語は品詞記号で、機能語は綴り英字列で表わす。[2] 例えば、「(化け/バケ)-の-皮-を-剥がす」には以下の構造記述 [*[[V22no]N]wo]*V30 が与えられている。[3]

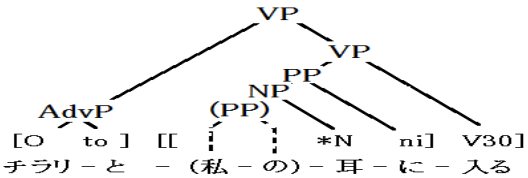


表現が並列構造を有する場合は括弧《 》あるいは< >で並列句がマークされている。例えば、「盆-と-正月-が--一緒-に-来る」では [《(N)to(N)》ga][[Nni]V30] によって、以下の構造が表わされる。



5.6 表現の柔軟性-内部修飾

収集した表現の一体性(rigidness)には無数の段階が有り、一律に扱う事は出来ない。表現の柔軟性を個別に担保するため、E 欄の構造記述中に内部被修飾可能性をアスタリスク*でマークした。例えば、「チラリ-と-耳-に-入る」という表現では、E 欄表示 [Oto][[*Nni]V30] 中のアスタリスク*によって、下図の「チラリと私の耳に入る」など、柔軟な派生表現に対応することができる [4]。



[2] 文節内の語の接続も便宜上、依存と同じ括弧表現で表示している。

[3] ただし、N は名詞、V22、V30 はそれぞれ動詞の連用形、終止形を表わす。品詞、活用型、活用形などの記号体系については(首藤他, 2012)を参照のこと。

[4] ただし、O はオノマトペを表わす。

5.7 文頭側条件(F 欄)

表現が存立するための条件として文頭側コンテクストを規定する。例えば、「形-を-する」は単独では用いられず、「妙-な-形-を-する」などのように、文頭側に連体修飾語が必要であることを＜連体修飾＞と記す、などである。

6. むすび

筆者らは、広大な言語現象に対処するためには、単語レキシコンに依存した NLP から句レキシコンに基づく NLP への脱皮がいずれ必要になると考えている。このとき、句を単語的にカプセル化するのではなく、「柔軟な句」として取り扱うことが必要である。本稿で述べた辞書は、この主張に沿って編纂された大規模表現辞書;JDMWE の最大のサブセットをなす。JDMWE は、日本語発話者が持つであろう N-gram 言語モデルの上澄み部分を(1≦N≦18 の範囲で) 特異性、慣用性の尺度によって掬い取り、かつ、各 N-gram に構造を付与した一種の構造付き tree bank である。

更なる許容変化形情報、および、しっかりした意味論に基づく意味関連情報の記載が今後の重要課題である。

謝辞

データの収集に協力頂いた方々、貴重な助言、励ましを頂いた島津明氏、荻野孝野氏に深甚の謝意を表します。

参考文献

[1] Cook, P., Fazly, A., Stevenson, S. 2008. The VNC-Tokens Dataset, Proceedings of the MWE workshop, ACL.

[2] Fujita, A., Furihata, K., Inui, K., Matsumoto, Y., Takeuti, K. 2004. Paraphrasing of Japanese Light-verb Constructions Based on Lexical Conceptual Structure, Proceedings of the MWE workshop, ACL.

[3] Kaalep, H., Muischnek, K. 2008. Multi-Word Verbs of Estonian: a Database and a Corpus, Proceedings of the MWE workshop, ACL.

[4] Kudo, T., Kazawa, H. 2009. Japanese Web N-gram Version 1, Linguistic Data Consortium, Philadelphia.

[5] 尾上兼英(監修). 1993. 成語林-故事ことわざ慣用句, 旺文社.

[6] Sag, I., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. 2002. Multiword Expressions; A Pain in the Neck for NLP, Proceedings of the 3rd CICLING.

[7] 三省堂編修所(編). 1999. 故事ことわざ慣用句辞典, 三省堂.

[8] 首藤公昭, 田辺利文. 2010. 日本語複単語表現辞書 JDMWE, 自然言語処理, 17-5.

[9] Shudo, K., Kurahone, A., Tanabe, T. 2011. A Comprehensive Dictionary of Multiword Expressions, Proceedings of the 49th Annual Meeting of the ACL.

[10] 首藤公昭, 高橋雅仁, 田辺利文. 2012. 日本語慣用句機械辞書, 情報処理学会研究報告, NL-205.

[11] 新納浩幸, 井佐原均. 1995a. 片方向の共起性による述語型定型表現の自動抽出, 自然言語処理, 2-3.

[12] 新納浩幸, 井佐原均. 1995b. 語義の特異性を利用した慣用表現の自動抽出, 情報処理学会論文誌 36-8.

[13] 白石大二(編). 1992. 擬声語擬態語慣用句辞典, 東京堂出版.

[14] 田辺利文, 高橋雅仁, 首藤公昭. 2012. 日本語表現辞書 JDMWE の統計的性質, 情報処理学会研究報告, NL-205.

[15] 田島諸介. 2002. ことわざ故事・成語慣用句辞典, 梧桐書院.

[16] Uchiyama, K., Ishizuka, S. 2003. A Disambiguation of Compound Verbs, Proceedings of the MWE workshop, ACL.

[17] Villavicencio, A. 2003. Verb-Particle Constructions and Lexical Resources, Proceedings of the MWE workshop, ACL.

[18] 米川明彦, 大谷伊都子(編). 2005. 日本語慣用句辞典, 東京堂出版.