日本語の複単語表現データベース(JMWE/C)の構築

Development of Japanese Multi-Word Expressions Database, JMWE/C

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構成性(compositionality)に問題の有る相当数の慣用句あるいは慣用句的な複単語表現(MWE; Multi-Word Expression)が使われており、構文・意味解析の大きなネックとなっている。 筆者らは、主としてこの問題に対処するため 1960 年代末から日本語 MWE のデータベース構築を行ってきたが、最近、その概要が定まったので、概念(自立)語相当表現データに絞って報告する。

key words：

慣用句(イディオム)、決まり文句、連語、コロケーション、成句、語結合、機能動詞結合、支援動詞構文、派生語、複合語、クランベリー語、四字熟語、格言、諺、擬態・擬音・擬声語(オノマトペ)、呼びかけ表現、応答表現、様態表現、フレーズベース翻訳、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

自然言語処理(NLP)における複単語表現(MWE; Multi-Word Expression)の重要性は、英語処理で問題となる MWE を概観し、その重要性を指摘した[1]がきっかけとなって、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 を毎年開催しており、非構成的(non-compositional)な MWE を統計的に自動評価・抽出する方法などが活発に議論されている。しかし、最近の研究でも Multiword Verb、Multiword Noun、Verb Particle Construction、Verb Noun Construction といった特定の構文構造の表現のみを対象とする研究が多く、未だ、表現の多様性を総括的に捉えた研究は見られないようである。筆者らは日常の自然言語を対象とする将来の NLP のためには、人の内省によって問題のある MWE 候補を出来るだけ網羅的に資源化しておくことが不可欠であると考え、1960 年代末から日本語を対象とした MWE の収集・整理を行ってきたが、今回、全体像がほぼ定まったので報告する。

2. 関連研究

日本語 MWE に関する研究としては、古くから国語学領域で人の利用のための辞典の編纂が数多く行われてきた[10]-[20]。しかし、これら個々の研究は表現、表記の多様性の不備や構造、用法の体系的記述の不備から、そのままでは NLP 向きとは言い難い。NLP の立場での日本語 MWE の研究は、機能(付属)語性 MWE を収集・整理し、単語的に組み込んだ“拡張文節モデル”を提案した[2]-[3]の研究が比較的古いほうだと思われる。近年の機能語性 MWE の研究には日本語文末 MWE の意味体系を提案した[4]や、表現を階層的に収集・整理する方法を提案した[5]などがある。 他方、NLP における概念(自立)語性 MWE の研究には、名詞、格助詞、動詞からなる述語性慣用句の日英翻訳を考察した[6]や、約 20,000 の NLP 用日本語 MWE を収集・整理し、公開した[7]がある。また、最近では市販の数種の慣用句辞典から約 3,400 個の慣用句を収集し、考察した[8]がある。しかし、これまでの NLP における MWE 研究は、

A	B	C	D	E	F	G	H	I
いまだかって	いまだ-かって	未だ-(嘗/曾)(っ)て	D		DD			否定
いまだかつて	いまだ-かつて	未だ-(嘗/曾)(つ)て	D		DD			否定
いまだかってない	いまだ-かって-ない	未だ-(嘗/曾)(っ)て-無い	Ya	aeb	nai			
いまだかつてない	いまだ-かつて-ない	未だ-(嘗/曾)(つ)て-無い	Ya	aeb	nai			
いまだしのかん	いまだし-の-かん	未だし-の-感	Mk		KnoM	No-De		
いまだしのかんあり	いまだし-の-かん-あり	未だし-の-感-(有/在)り	Yk	vb20	V'	X-De		
いまだしのかんがある	いまだし-の-かん-が.ある	未だし-の-感-が-(有/在)る	Yv	vb2	aru			
いまだしのかんのある	いまだし-の-かん-の.ある	未だし-の-感-の-(有/在)る	Tv	vb25	aru			
いまだに	いまだ-に	未だ-に	D		Dni			否定
いまだもって	いまだ-もって	未だ-以て	D		DD			否定

図1　データの一部

表現、表記の多様性や機能、構造記述等の点で未だ不十分と言わざるを得ない。本研究は、これらの問題を軽減すべく[7]を大幅に修正・拡張したものである。

3. 収録表現

筆者らは次の基準で概念語性 MWE を収集・整理した。

1) 慣用句(イディオム)性の表現

要素単語から全体の意味が規則で導くことが難しいと思われる表現(non-compositional な表現)、例えば、「赤-の-他人」、「耳-を-貸さ-ない」、「手-を-抜く」、「足-が-出る」、「首-が-回ら-ない」、「顔-を-売る」、「気-を-取(り)-直し-て」、「気-を-利か-せる」などである。 また、通常、慣用句とは呼ばれないが、やはり構成性(compositionality)に問題のある表現も出来るだけ網羅した。この意味で支援動詞構文(SVC)、一部の複合語、派生語が含まれる。 例えば、「一-票-を-投じる」、「批判-を-加える」、「(磨/研)き-を-(掛/懸)ける」、「伝票-を-切る」、「計画-を-立てる」、「辞書-を-(引/曳/牽)く」、「(バカ/馬鹿/莫迦)-を-(言/云)う」、「右-肩-上(が)り-に」、「練り-歩く」、「打(ち)-拉が-れる」、「積(み)-立てる」、「顔-を-する」、「ウロウロ-する」、「大学-を-出る」、「要求-を-(飲/呑)む」、「シドロモドロ-に-成る」などである。

2) 決まり文句的な表現

一体性の強い表現。例えば、「風前-の-灯」、「付きっ-切り」、「矢-継(ぎ)-早」、「禍-転じ-て-福-と-なす」、「雲-一つ-無い」、「時-は-金-なり」、「其れ-は-然う-と」、「オット-ドッコイ」、「程度-の-差-こそ-有れ」、「(眼/目)-に-も-(止/留)まら-ぬ-早-(技/業)」、「(腰/コシ)-を-抜かす-程」、などである。

3) 単語間共起確率の高い表現

例えば、「警鐘-を-鳴らす」、「相性-の-良い」、「訓戒-を-垂れる」、「対決-色-を-強める」、「喧騒-を-離れる」、「手-を-こまぬく」、「腰-を-抜かす-程-驚く」、また、例えば、「故郷-を」は「(思/想)う」、「出る」、「偲ぶ」、「恋う」、「離れる」などの動詞と共起しやすいことなども収録している。さらに、オノマトペとその派生形についても動詞との共起を出来るだけ網羅的にデータ化した。例えば、「ユルユル-と-動く」、「グラグラ-揺れる」、「グッスリ-眠る」、「クルクル-回る」、「ポッカリ-と-空く」などである。以上の3種の基準を兼備する表現も非常に多く、表現セットが3種に直和分割される訳ではない。

4. 記載した情報

全体は約95,000行見出し、9情報欄 (A欄～I欄) の表形式である。 図1にデータの一部で形式を示す。

4.1 平仮名ベタ見出し(A 欄)

音に基づいている。例えば、「良い」は「よい」と「いい」に、「得る」は「える」、「うる」に読み分けて見出しとする。

4.2 構成単語間の境界(B 欄)

「-ハイフン」あるいは「.ドット」で単語間境界を示す。ドットはこの位置に別の単語列(例えば副詞)が挿入される可能性を示す。活用語尾は原則として切り離さない。

4.3 漢字、片仮名などの異字種表記(C 欄)

字種と表記の揺れ情報を同時に与える。例えば、「組

(み)-付ける」などのカッコは文字の任意性、「(良/好/善)い」など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、殆ど全ての異表記に対応できる。例えば、B 欄「き-の-いい-やつ」、C 欄「気-の-(良/好/善)い-(奴/ヤツ)」から、次の様な 24 種の表記がカバーされる。

「きのいいやつ」、「きのいい奴」、「きのいいヤツ」、・・・・・、「気の良い奴」、「気の好い奴」、「気の好いヤツ」、「気の善いやつ」、「気の善い奴」、「気の善いヤツ」

4.4 文法的な機能と種別(D 欄)

表現全体の文法的な機能として、C: 接続詞性表現、D: 副詞性(連用修飾)表現、T: 連体詞性(連体修飾)表現、M: 名詞性表現、Ms: サ変名詞性表現、Md: サ変以外の動的名詞性表現、Mk: 形容動詞的名詞性表現、Yv: 動詞性表現、Ya: 形容詞性表現、Yk: 形容動詞、準形容動詞性表現、Yo: 擬態・擬音・擬声表現、の別、用途の種別として、-P: 格言、諺、-Self: 自問、独り言表現、-Call: 呼びかけ表現、-Grt: 挨拶表現、-Res: 応答表現、などを記載する。文解析には不可欠の情報である。

4.5 述語の表層格構造(E 欄)

表現に述語が含まれる場合、その表層格パタン約 80 種を以下のようにコード化して与える。

名詞+「を」+動詞 : va1
名詞+「が」+動詞 : va2
名詞+「に」+動詞 : va3
・・・・・・・・
名詞+「も」+名詞+「も」+形容詞 : ad1

例えば、「花-も-実-も-(有/在)る」には vd1 と記載される。これは、文の大局処理とのリンクを意識しているためである。

4.6 末尾の構造(F 欄)

例えば、表現の末尾が「思わ-れる」、「押(し)-掛ける」の場合、末尾に助動詞「れる」が使われていること、二動詞の複合であることを、それぞれ、「Vreru」、「VV」と記す。末尾表現の活用を表現全体の活用とみなすことができる。

4.7 派生形(G 欄)

形容動詞性(様態)表現(D、Mk、P、Yk、Yo)の場合、

＜連体修飾形＞{-＜連用修飾形＞{-＜動詞形＞}}

の形式で派生形を与える。

例えば、「(我/吾)-関せ-ず」という表現では、「(我/吾)-関せ-ず-**の**」、「(我/吾)-関せ-ず-**と-(言/云/謂)う**」、「(我/吾)-関せ-ず-**と-(言/云/謂)っ-た**」で連体修飾、「(我/吾)-関せ-ず-**と**」、「(我/吾)-関せ-ず-**で**」と連用修飾句が派生することを NoToiuToitta-ToDe と記す。同様に、擬態語「フラフラ」には、「フラフラ-**の**」、「フラフラ-**し-た**」、「フラフラ-**と-し-た**」で連体修飾、「フラフラ」、「フラフラ-**と**」、「フラフラ-**し-て**」、「フラフラ-**と-し-て**」で連用修飾、「フラフラ-**する**」、「フラフラ-**と-する**」と動詞化することを NoSitaTosita-EToSiteTosite-SuruTosuru と記す。また、同じ擬態語でも「グングン」では連用句としての「グングン」、「グングン-**と**」以外の派生は不自然なので、G 欄は X-ToE と記す。(E は空列、X は派生ナシの意。)このように、これらの派生パタンは多彩で、約 300 種にのぼる。X-ToE などのコードは様態表現の細密化した品詞の表記と考えることができる。

4.8 文頭側条件(H 欄)

表現が存立するための必要条件として文頭側コンテクストを与える。例えば、「割れ-に-なる」は単独では用いられず、「**元本**-割れ-に-なる」などのように、文頭側に名詞による連体修飾語が必要であることを＜名詞連体＞と記す、などである。

4.9 文末側条件(I 欄)

H 欄と同様、文末側コンテクストを与える。例えば、「如何-と-も」は文末側に「～難い」などの困難性を表す表現を必要とすることなどである。

5. むすび

見出し数は現在、約95,000であるが、G欄の派生形を加えれば、約 110,000 表現、さらに、B、C 欄の表記揺れを考慮すれば、500,000 表現程度をカバーしていると推定される。機械翻訳などの応用タスクを除く一般的意味で

の課題として次の点が残る。

ⅰ. 表現のカバレジの詳細な検証。(凡そは検証済み)

ⅱ. 意味上の多義性を持つかどうかの情報付与。例えば、「手-を-切る」はイディオム的な意味と文字通りの意味の両方で使い得ることなど。

ⅲ. 「です」、「ます」調表現、丁寧、尊敬表現等の追加。

ⅳ. 標準的な表現への言い換え情報付与。例えば、「油-を-売る」に「話などをして時間を過ごす」と(言い換え表現の文法構造と共に)与える。

ⅴ. 詳細な変化形情報付与。例えば、「手-が-回る」の変化形「手-も-回る」とか「回る-手」、「手-は-回る」は熟語の意味として許されるや否やなど。

ⅵ. 出現頻度(n-グラム)情報、条件付き確率付与。

ⅶ. 異表記の優先度情報付与。

ⅷ. 古語、現代語表現の区別情報付与。

以上の一部については既に整理を開始している。例えばⅴについては [9]を参照のこと。

謝辞

長期に亘りデータの収集に協力頂いた、江崎斗志子氏、竹内美津乃氏、高丘満佐子氏をはじめとする多くの方々、貴重な助言、励ましを頂いた元九州芸術工科大学長故吉田将氏、元言語処理学会会長、JAIST 教授島津明氏に深甚の謝意を表します。

参考文献

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

[2] 首藤公昭, 楢原斗志子, 吉田将, 日本語の機械処理のための文節構造モデル, 電子通信学会論文誌, 62-D-12, 1979.

[3] 首藤公昭, 文節構造モデルによる日本語の機械処理に関する研究, 福岡大学研究所報, 45, 1980.

[4] K. Shudo, T. Tanabe, M. Takahashi, K. Yoshimura, MWEs as Non-propositional Content Indicators, Proc. of the 2nd ACL Workshop on MWE, 2004.

[5] 松吉俊, 佐藤理史, 宇津呂武仁, 日本語機能表現辞書の編纂, 自然言語処理, 14-5, 2007.

[6] 奥雅博, 日本語慣用表現の分析と日英翻訳への適用, 情報処理学会研究報告, 87-NL-62, 1987.

[7] 首藤公昭, 日本語における固定的複合表現, 昭和63 年度文部省科学研究費特定研究(Ⅰ)「情報ドクメンテーションのための言語の研究」報告書, 1989.

[8] 佐藤理史, 基本慣用句五種対照表の作成, 情報処理学会研究報告, 07-NL-178, 2007.

[9] 安武満佐子, 小山泰男, 吉村賢治, 首藤公昭, 固定的共起表現とその変化形, 言語処理学会第 3 回年次大会発表論文集, 1997.

[10] 新村出編, 広辞苑 第 6 版, 岩波書店, 2008.

[11] 松村明監修, 大辞泉, 小学館, 1998.

[12] 松村明編, 大辞林 第 3 版, 三省堂, 2006.

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

[14] 三省堂編修所編, 故事ことわざ慣用句辞典, 三省堂, 1999.

[15] 白石大二編, 擬声語擬態語慣用句辞典, 東京堂出版, 1992.

[16] 竹田晃, 四字熟語・成句辞典, 講談社, 1990.

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

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

[19] 藤田保幸, 山崎誠編, 複合辞研究の現在, 和泉書院, 2006.

[20] グループ・ジャマシイ編, 日本語文型辞典, くろしお出版, 2007.