日本語の副詞性複単語表現辞書

A Dictionary of Japanese Adverbial 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表現も構文解析の効率や曖昧さ低減の点で重要であるが十分に整備されてはいない。筆者らは以上の認識から日本語特異表現辞書の構築を行ってきたが、本稿では、そのうち「嬉々として」、「言うに事欠いて」、「驚いて」、「小骨一本抜かずに」など連用修飾MWEに絞って辞書の概要を報告する。収録数は約13,000表現である。

key words：

複単語表現(MWE), 連用修飾, 副詞, オノマトペ, 慣用句, 決まり文句, 連語, コロケーション, 成句, 語結合, 様態表現, 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

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(Sag et al., 2002)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が活発に行われている。しかし、現状では充分な適合性、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などの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) 本稿ではその一部である連用修飾MWE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、古くから国語学領域で人の利用を目的とした慣用句辞典類の編纂が数多く行われてきた。(尾上(監修), 1993; 三省堂編修所(編), 1999; 白石, 1992; 田島, 2002; 米川他(編), 2005など) しかし、これらの研究では、多くの場合、表現の異表記や内部構造、副詞性であるなどの文法機能・用法についてあまり注意が払われていない。

これまで、NLPの立場で日本語の副詞性MWEに焦点を当てた研究は少なく、ほとんどがモダリティー副詞、程度副詞といった特定カテゴリーの副詞や個別副詞の意味に関する国語学領域の研究が多い。また、一部、単語、複単語の混在した議論もみられる。日本語副詞全体を俯瞰した意味に関する研究には(仁田, 2002)がある。外国では、フランス語の副詞性MWE (Multiword Adverbs) 6,800種を収録したNLP用辞書が(Laporte et al., 2008)で報告されている。その他、ドイツ語では(Seelbach, 1990)、スペイン語では(Català et al., 2007)、

ポルトガル語では(Baptista, 2003)、韓国語では(Jung, 2005)などの副詞性 MWE に関する NLP 領域での研究報告がある。

しかし、何れにおいても十分な実用性が認められている訳ではない様である。

A	B	C	D	E	F	G
いまだかつて	いまだ-かつて	未だ-(嘗/曾)(つ)て	Adv	DD		否定
いまだに	いまだ-に	未だ-に	Adv_D_ni	Dni		否定
いまだもって	いまだ-もっ-て	未だ-以-て	Adv_Verb_te	D[V23'te]		否定
いまでこそ	.いま-で-こそ	今-で-こそ	Adv_N_dekoso	[*Nde]koso		～が
いまでは	.いま-で-は	今-で-は	Adv_N_deha	[*Nde]ha		
いまでも	いま-でも	今-でも	Adv_N_demo	*Ndemo		
いまとなっては	.いま-と-なっ-て-は	今-と-なっ-て-は	Adv_Verb_teha	[[[*Nto]nat]te]ha		

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基準で副詞性 MWE を収集し、既存の事典類を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、「口-程-に-も-な-く」、「ケン-も-ホロローに」、「手-を-抜-い-て」、「気-を-取(り)-直し-て」、「取る-物-も-取り-あえ-ず」、「(已/止)む-に-(已/止)ま-れ-ず」等々である。 これらには、通常の慣用句辞典類には収録されていない表現が相当数含まれる。

2) 高い単語間共起確率

単語間共起確率が高く、一体性の強い表現。例えば、「付きっ-切り-で」、「矢-継(ぎ)-早-に」、「差(し)-向かい-で」、「力-尽く-で」、「其れ-は-然う-と」、「取り-敢え-ず」、「程度-の-差-こそ-有れ」、「(眼/目)-に-も-(止/留)まら-ぬ-早-(技/業)-で」等々である。[1,2,3]

JDMWE における表現採録の網羅性については（首藤他, 2010)で検証されている。

4. 記載情報

本辞書は約 13,000 行、7 欄（A 欄～G 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。

4.1 平仮名ベタ見出し（A 欄）

音に基づいている。例えば、「(～を)良い事に」は「よいことに」と「いいことに」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。活用語尾は切り離さない。

4.3 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、「差(し)-向かい-で」のカッコ( )は文字の任意性、「(已/止)む-に-(已/止)ま-れ-ず」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。異表記を数え上げると、本辞書は 60,000 表現程度をカバーしていると考えられる。

---

1 以上の基準 1)、2)は排他的ではなく、双方を満たす表現は数多い。

2 数量表現、単独のオノマトペは対象外とする。

3 この基準に照らした収録表現の一般的な妥当性検証については(Shudo et al., 2011; 田辺他, 2012)を参照されたい。

4.4 文法的な機能と種別(D 欄)

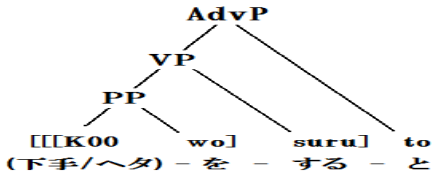
表現全体の文法的な機能として副詞性である事をコード Adv で表わし、末尾構造の大まかな種類をアンダースコア_を介してその後ろに記す。具体的には以下のような 200 種類程度が区別されている。

Adv_Adj:末尾が形容詞(連用形)　ex.「否-応-無く」
Adv_Adj_to:末尾が形容詞+「と」 ex.「今-や-遅し-と」
Adv_AdjVerb_daga:末尾が形容動詞+「だが」
ex.「生憎-だ-が」
Adv_AdjVerb_ni:末尾が形容動詞+「に」
ex.「速(や)か-に」
Adv_AdjVerb_to:末尾が形容動詞+「と」 ex.「猛然-と」
Adv_D_mo:末尾が副詞+「も」　ex.「(猶/尚)-も」
Adv_N:末尾が名詞　ex.「苦心-の-末」
Adv_N_ni:末尾が名詞+「に」　ex.「迂闊-な-事-に」
Adv_O_to:オノマトペ+「と」　ex.「ゼイゼイ-と」
Adv_Verb:末尾が動詞(連用形)　ex.「追っ-付け」
Adv_Verb_domo:末尾が動詞+「ども」
ex.「行け-ども-行け-ども」
Adv_Verb_gotoku:末尾が動詞+「ごとく」
ex.「先-に-述べ-た-如く」
Adv_Verb_ni:末尾が動詞+「に」　ex.「考える-に」
Adv_Verb_to:末尾が動詞+「と」
ex.「今-から-思う-と」
・・・・・・・・

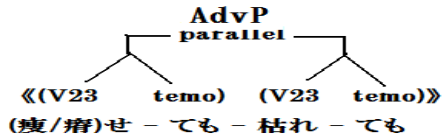
4.5 表現の構文構造(E 欄)

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[4]。

例えば、「(下手/ヘタ)-を-する-と」には以下の構造を表わす記述 [[K00wo]suru]to が与えられている[5]。



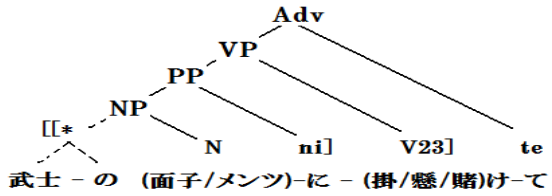
表現が並列構造を有する場合は括弧《 》あるいは< > で並列句がマークされている。例えば、「(痩/瘠)せ-ても -枯れ-ても」では 《(V23temo)(V23temo)》 によって、以下の構造が表わされる。



4.6 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、一律に単語として扱う事は出来ない。そこで、表現の柔軟性を留保するため、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、「(面子/メンツ)-に-(掛/懸/賭)け-て」という表現では、E 欄表示[[*Nni]V23]te のアスタリスク*によって、「武士の面子に懸けて」、「医師としての面子に懸けて」などの関連表現に対応可能としている。



4.7 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与える。例えば、「奇貨-と-し-て」は単独では用いられず、「震災-を-奇貨-と-し-て」などのように、文頭側に「を」格の修飾語が必要であることを<「を」連用修飾>と記す。

---

4 文節内の語の接続も便宜上、依存と同じ括弧表現で表示する。
5 ただし、K00 は形容動詞語幹を表わす。品詞、活用型、活用形などの記号系については(首藤他, 2012)を参照されたい。

4.8 文末側条件(G 欄)

H 欄と同様、文末側条件を与える。例えば、「如何-と-も」は文末側に「難い」など、困難性を表す表現を必要とすることなどである。古くから単語副詞と後続語との呼応が調べられているが(栗原他, 1968)、ここではそれらの MWE への一般化が行われている。対象が副詞相当表現であるから<非文末>がデフォールト条件である。

5. むすび

筆者らは、広範な自然言語現象に対応するためには、単語レキシコンに依存した NLP から(単語を包含した)句レキシコンに基づく NLP への転換が必要であると考えている。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肝要である。本稿で述べた辞書はこの主張に沿って編纂されている。表現の一般的な網羅性については(首藤他, 2010)を参照されたい。しっかりした意味モデルに基づく意味情報の記載が今後の主要な課題である。

謝辞

データの収集に協力頂いた多くの方々、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] Baptista, J. 2003. Some Families of Compound Temporal Adverbs in Portuguese, Proceedings of the Workshop on Finite-State Methods for Natural Language Processing, EACL.

[2] Blanco, X., Català, D. 1998. Quelques remarques sur un dictionnaire électronique d'adverbes composés en espagnol, Lingvisticae Investigationes 22.

[3] Català, D., Baptista, J. 2007. Spanish Adverbial Frozen Expressions, Proceedings of the MWE Workshop, ACL.

[4] Jung, E. J. 2005. Grammaire des adverbes de durée et de date en coréen, Thèse de doctorat en Informatique Linguistique, Université Paris-Est Marne-la-Vallée.

[5] 栗原俊彦, 吉田将, 鶴丸弘昭. 1968. 日本語文の分析-副詞の処理について(1)-, 九州大学工学集報, 41-5.

[6] Laporte, É., Voyatzi, S. 2008. An Electronic Dictionary of French Multiword Adverbs. Proceedings of the LREC Workshop towards a Shared Task for Multiword Expressions.

[7] 仁田義雄. 2002. 副詞的表現の諸相, くろしお出版.

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

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

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

[11] Seelbach, D. 1990. Zur Entwicklung von bilingualen Mehrwortlexica Französisch-Deutsch-Stützverbkon struktionen und adverbiale Ausdrücke, Lexicon und Lexikographie 11.

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

[13] 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.

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

[15] 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.

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

[17] 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.

[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.