文の部分構造を与える日本語フレーズレキシコン

首藤　公昭[†]　田辺　利文[††]　高橋　雅仁[†††]

[†]福岡大学名誉教授 koshoshudo@gmail.com
[††]福岡大学工学部電子情報工学科 tanabe@fukuoka-u.ac.jp
[†††]久留米工業大学工学部情報ネットワーク工学科 taka@cc.kurume-it.ac.jp

1. はじめに

日常の広範な文を対象とする自然言語処理(NLP)では単語と接辞の辞書だけでは語彙資源として十分ではなく、必要に応じてフレーズを処理単位と捉えることが不可欠である事はよく知られており、近年、フレーズベース統計翻訳などの研究が盛んに行われている(Galley et al., 2010 など)。しかし、統計的手法においては、相当大規模なコーパスでも言語表現の生起に希薄性が残るロングテール問題や考慮すべきフレーズの種類、特徴をどのように捉えるかなどの難しい問題があり、十分な成果はまだ見られない。特に、ギャップを含む不連続フレーズを扱うとなれば、統計処理の困難さは更に増大する。筆者らはフレーズベース機械翻訳のための日本語フレーズレキシコン(JDMWE: Japanese Dictionary of Multiword Expressions)を 1960 年代から開発してきており、現在、収録フレーズ基本形 11 万件を超え、一般日本語処理用辞書としても有効なレベルに達している。JDMWE は、基本的に2種の特異性に着目してフレーズを採録している。一つは、例えば、「赤の他人」の意味が「赤」、「の」、「他人」という要素語の通常の意味から導くことが難しいという性質(非構成性=イディオム性)、他の一つは、例えば、「こまねく」という動詞は「手をこまねく」以外には殆ど使われないというような性質(要素間強共起性=決まり文句性)を言う。この種のフレーズ=複単語表現(Multiword Expression: MWE)の重要性は、2002 年に I. A. Sag 氏らによって指摘されてから世界で再認識されるようになった(Sag et al., 2002)。しかし、まだ総括的なレキシコンは世界で得られていない。JDMWE は 2011 年の米国計算言語学会(ACL)年次大会における K. Church 氏の提題:"How Many Multiword Expressions do People Know?"に対する日本語における一つの解答である(Church, 2011)。本稿では「言語処理の観点から興味深い言語現象の収集・分析」の一例として JDMWE を紹介する。

2. 編成・規模

JDMWE は次の様な部分辞書から構成されている。

1. 日本語 MWE 辞書_慣用句編:「油を売る」、「真っ赤なウソ」などの慣用句約 4,400 表現

2. 日本語 MWE 辞書_動詞性表現(Ⅰ類)編:「手を結ぶ」、「思いが募る」、「沽券に関わる」のような『名詞+を+動詞』、『名詞+が+動詞』、『名詞+に+動詞』の形式の約 35,000 表現

3. 日本語 MWE 辞書_動詞性表現(Ⅱ類)編:「骨の髄までしゃぶる」、「猿も木から落ちる」、「ゼロからやり直す」など動詞が支配するⅠ類、Ⅲ類以外の約 13,000 表現

4. 日本語 MWE 辞書_動詞性表現(Ⅲ類)編:「放り出す」、「飲んだくれる」など、単語性や構成性に疑問が残る、複合動詞を中心とした比較的単純な構造の約 3,500 表現

5. 日本語 MWE 辞書_形容詞性表現編:「頭が痛い」、「傷跡が生々しい」など、形容詞が支配する約 4,800 表現

6. 日本語 MWE 辞書_形容動詞性(様態)表現編:「痘痕も笑窪」、「願ったり叶ったり」、「驚く程」など、広義の形容動詞性ともいえる様態表現約 10,000

7. 日本語 MWE 辞書_副詞性表現編:「思いがけず」、「心を鬼にして」など、連用修飾句としてよく用いられる約 16,000 表現

8. 日本語 MWE 辞書_連体詞性表現編:「気のきいた」、「世に云う所の」など、連体修飾句としてよく使われる約 15,000 表現

9. 日本語 MWE 辞書_名詞性表現編:「天下の一大事」、「無二の親友」など、名詞性の約 18,000 表現

10. 日本語 MWE 辞書_連結詞・文副詞性表現編:「そうは言うものの」、「このような状況で」など、文頭で用いられる連結詞・文副詞性表現、1,100 表現

11. 日本語 MWE 辞書_オノマトペ表現編:「グラグラ」、「シュ

ルシュルと」、「ガッツリ食う」など、擬声・擬音・擬態語およびそれらを用いた典型表現約 12,000

12. 日本語 MWE 辞書_格言・諺・成句・決まり文句編：「急がば回れ」、「義を見てせざるは勇無きなり」などの約 3,900 表現

13. 日本語 MWE 辞書_四字熟語編：「四面楚歌」、「切磋琢磨」などの約 500 表現

14. 日本語 MWE 辞書_挨拶・呼びかけ・応答・独言表現編：「何という事だ」、「御苦労さま」など良く使われる挨拶・呼びかけ・応答・独言表現約 450

15. 日本語 MWE 辞書_クランベリー型表現編：「しがみつく」、「後ろめたい」などのクランベリー型表現候補約 40

16. 日本語 MWE 辞書_機能語(助詞、助動詞)性表現編：「を理由に」、「において」、「における」、「ほうがいい」、「てください」など、助詞、助動詞相当の機能を持つ表現約 4,000

3. 特徴・統計的性質

本辞書の主な特徴は、

1. 本辞書が対象外とする各種専門用語、各種固有表現、時間空間表現を除けば、上記の如く収録表現の種類が網羅的であること

2. 表現の構文機能だけでなく、内部係り受け構造をギャップ可能性を含めて記載していること

3. 記載情報が多様かつ詳細であること

4. 機械処理への即応性が高いこと

である。

また、本辞書の統計的性質を調べる 2 種の調査を行った結果は、概略、次の通りであった。

(1) ランダムに選んだ日本経済新聞の 1 ページと最終ページの記事 1,459 文中、1,928 か所に JDMWE の収録表現が何らかの形態(活用変化形など)で使われていた。10 文中に 13 か所程度という比率である。(このうち、6 か所は上記辞書 16.の助詞・助動詞性表現。) このように少くとも新聞記事に対しては JDMWE は相当高いトークン・カバー率を持つ。

(2) Google 社が公開した Web 上の 200 億日本語文に現れる n-単語連鎖の出現頻度データ LDC2009T08(Kudo et al., 2009)と上記 **2.**の辞書 2、動詞性表現(Ⅰ類)とを比較した結果、その『名詞＋助詞(が、を、に)』から『動詞』に移る遷移確率が高いものほど JDMWE に多く採録されているという傾向が確認され、表現採録の妥当性が推定された。また、同調査から JDMWE 収録表現の 10%程度は Google のデータに表れていないことが分り、コーパスにおける MWE のロングテール分布が覗われた。さらに、Web 上に生起する動詞性表現(Ⅰ類)タイプの約 2.5%にすぎない JDMWE 収録の動詞性表現(Ⅰ類)が、トークンレベルでは Web 上に生起する同形式表現の約 14%をカバーしていることが分った。このように JDMWE にはかなり高頻度で生起する表現が選ばれている(首藤ほか, 2010)。

4. 記載情報

上記の各辞書には、原則として以下の **a.**～**f.**の情報が記載されている。

a. 区切り、表記情報： 辞書見出しは「ばかをみる」のように平仮名ベタ表記で与え、「ばか/を/みる」と要素単語に区切られること、そのうち、「ばか」は「バカ」、「馬鹿」、「莫迦」、「みる」は「見る」、「観る」と表記可能であることが記載されている。従って、この場合、4×3=12 種の異表記が与えられていることになる。

b. 文法機能情報： 例えば、「墓穴を掘る」は全体として動詞句(VP)、「命の洗濯」は名詞句(NP)、「年がら年中」は副詞句(AdvP)の働きをする、など、表現全体の相当文法カテゴリーが記載されている。

c. 文法構造情報： 例えば、「目が点になる」は、目→が→なる、点→に→なる という依存構造を持つことを自立語は品詞記号、付属語はローマ字綴りを使い、カッコ[ ]で [[Nga][[Nni]V_{30}]] と 2 項の句表示をしている。(N は名詞、V_{30} は動詞終止形の品詞記号。) また、並列構造はカッコ< >、あるいは《 》で、並列要素はカッコ( )で表示している。例えば、「泣く子と地頭」の構造は <([V_{40}N])to(N)>である。(V_{40}は動詞連体形の品詞記号。)

d. 内部修飾可否情報： 例えば、慣用句「油を売る」は「油をいつもの店で売る」のように内部修飾句をとり、ギャップが生じることがある。このことを **c.**の構造記述中にアスタリスクで [[Nwo]*V_{30}] のように記載する。この情報は、慣用句などをいつも単語化して扱う弊害を避け、より柔軟なギャップ付きフレーズ(不連続フレーム)として扱うための枠組みである。

e. 文脈情報： 例えば、軽動詞構文「顔をする」は「困った顔をする」の様に文頭側に連体修飾句を要求する。また、副詞的表現

「一つたりとも」は後方に「与えない」の様な否定句を要求する。この種の必須(あるいは選好)文脈が記載されている。

f. 連体、連用、動詞化情報： 本辞書は「独りよがり」、「針で突いた程」、「子供だまし」のような、物事の様態を表わす形容動詞的と言える表現を含んでおり、これらが連体、連用修飾句として用いられたり、動詞化して用いられたりする際の後続要素を記載している。例えば、擬態語「フラフラ」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、これに対して「グングン」は、「グングン」、「グングンと」の連用修飾形のみ存在することなどが記載されている。

5. 応用について

5.1 構文解析

意味的な纏まりをもつ表現ブロックを処理単位とする手法は構文・意味解析、機械翻訳などにおいて威力を発揮する。例えば、JDMWE には「手に付かず」、「散歩に出る」、「事にする」がそれぞれ、副詞性表現(AdvP)、動詞性表現(VP)、助動詞性表現(Aux)として登録され、それぞれに前記 c.の構造 [[Nni][V_{12}zu]]、[[Nni]V_{30}]、[[Nni]suru]が記載されている。また、「手に付かず」には前記 **e.**の文頭側必須文脈条件として「が」格の後置詞句が指定されている。これらの情報によって入力文：「彼は仕事が手に付かず、散歩に出る事にした」の構文解析を行った場合の解析例を図 1 に示す。(細部は省く。) この入力文は 8 文節からなっているが、JDMWE を優先的に採用すれば、3 文節からなる文であるかの如く取り扱うことができ、処理が簡素化される。同時に、相当数の解析不正解(曖昧さ)が排除できていることは明らかである。

また、前述の統計的性質から、JDMWE は入力単語を予測しながら読み進み、解析木を効率よく生成していく**予測的構文解析**を実現するためにも有効な資源となる。むろん、誤って MWE を認定する可能性もあるが、巧く設計すれば、妥当な解析に早期に導かれる確率を高くすることができると思われる。

5.2 機械翻訳

また、この例の場合、JDMWE の各表現に英訳情報、例えば、“as *SUB* is unable to get down to doing *SUB's N*”、“go out for a walk”、“decide to”を与えたと仮定すれば、図 2 のように上記の解析にほぼ並行した形で自然な(意味に忠実な)訳出を行うことが出来る可能性が在る。(細部は省く。)

5.3 仮名漢字変換

収録表現数 68,000 程度であった JDMWE の旧バージョンを当時市販されていたワープロソフト WXG v1.25 に組み込み、見出し、および、前記 **a.**の単語間区切り情報と異表記情報を用いることにより、仮名漢字変換の初回正変換率を 7 ポイント程度向上させたことが報告されているが(小山ほか, 1998)、現状の JDMWE では収録表現、記載情報がより充実しているため、仮名漢字変換のさらなる精度向上にも寄与できると思われる。これらの情報は、漢字部に対するいわゆる「ルビ振り」にも直接的に応用可能である。

5.4 音声認識

JDMWE には要素単語間に確率的な縛りの強いものが網羅されており、特に、次発生単語のパープレキシティーの小さい表現ほど優先的に採録されている(Shudo et al., 2011; Tanabe et al., 2014)。また、見出しは表音の仮名表記となっており、「える」と「うる」、「よい」と「いい」など、同一表現であっても発音が異なれば別見出しとしているなど、音声処理への入口に配慮した構成となっている。

5.5 その他

JDMWE のそれぞれの部分辞書は、新聞、雑誌などの実際の生データから長期間にわたって収集された表現を出来るだけ漏れなく収録したもので、記載情報と併せて国語学の各研究領域に何らかの有効な情報を提供するのではないかと思われる。

6. むすび

人間がこの種のフレーズを一纏まりに認知・生成しているという側面は否定できないと思われ、今後の NLP システムが単語レキシコンに加えて本格的 MWE レキシコンを装備・活用することは寧ろ自然であろう。なお、講演時に聴講者からの希望により、JDMWE における個別フレーズ収録の有無を確認・検証する時間を設けたい。

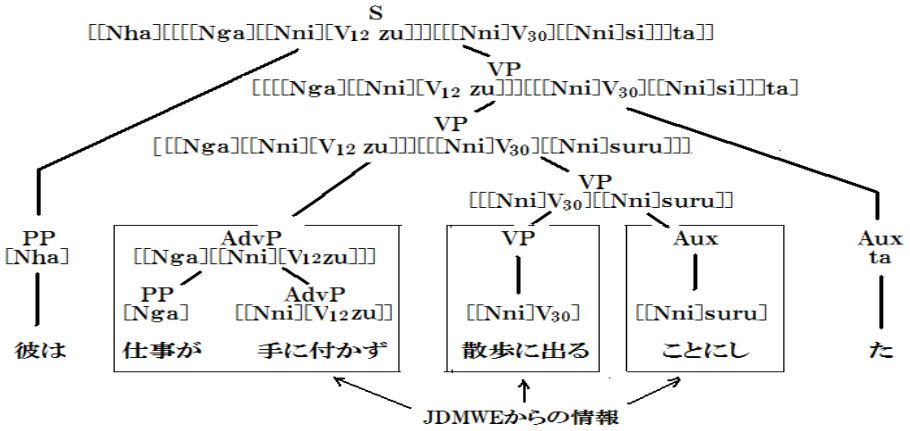

図 1　JDMWE による構文解析例

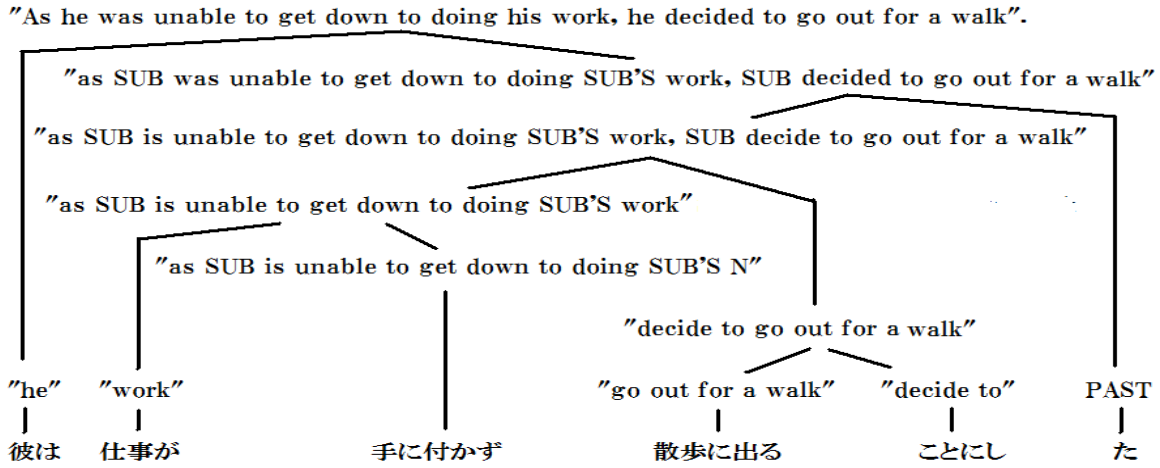

図 2　JDMWE による日英機械翻訳例

参考文献

[1] Church, K. 2011. How Many Multiword Expressions do People Know? , Proceedings of the MWE workshop, ACL.

[2] Galley, M., Manning, C. D. 2010. Accurate Non-Hierarchical Phrase-Based Translation, Proceedings of NAACL-HLT.

[3] 小山泰男, 安武満佐子, 吉村賢治, 首藤公昭. 1998. 連語データを利用した仮名漢字変換. 情報処理学会論文誌, 39-11.

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

[5] 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.

[6] 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.

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

[8] Tanabe, T., Takahashi, M., Shudo, K. 2014. A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing, Journal of Computer, Speech, and Language, Elsevier. (in press)