日本語複単語表現レキシコン（JMWEL）の概要と現状
― 動詞性複単語表現を中心として ―

髙橋雅仁* 田辺利文** 首藤公昭***
*久留米工業大学 **福岡大学工学部 ***福岡大学名誉教授
* taka@kurume-it.ac.jp ** tanabe@fukuoka-u.ac.jp *** viggo_ksf@jcom.home.ne.jp

1. はじめに

21 世紀に入り，コロケーション，慣用句，決まり文句などの単語の境界を越えた長単位表現の取扱いが重視されるようになり，自然言語処理（NLP）分野で「複単語表現 Multiword Expression; MWE[1]」，言語学では「定型言語 Formulaic Language[2]」,「単語連鎖 Lexical Bundles[3]」などの名称で種々の観点から研究が行われている．これらの表現が日常言語でかなり多種類，高頻度で使われていることは多くの研究者の指摘しているところであり[1]，また，近年，人が行う文意の認知機構の点からも興味を持たれている[2]．

本論文では，筆者の一人が 1960 年代から意味ベースの日本語処理研究の一環として編纂を進めてきた，見出し数，10 万件を超える規模の日本語複単語表現レキシコン（JMWEL ; Japanese MWE Lexicon, 以下 JMWEL と記す）[7][8][9]について，その現状を報告する．JMWEL は，構文・意味解析をはじめとする広範な日本語処理のための言語資源のプロトタイプを目指すものであり，広く研究用に公開されている．本稿では，JMWEL を構成する辞書のうち，特に 2017 年 11 月に言語資源協会（GSK）より公開された

（1）日本語動詞性複単語表現（1 類）レキシコン：JMWEL_verbal（class1）v3.2 ― ガ格，ヲ格，ニ格を介した動詞と名詞のコロケーション集（慣用句等を含む）― （GSK2017-C），

（2）日本語動詞性複単語表現（2 類）レキシコン：JMWEL_verbal（class2）v3.2 ― 述語動詞と種々の語とのコロケーション集 ― （GSK2017-D）

を中心に述べる．

2. 日本語複単語表現レキシコン JMWEL

2.1. 特徴

JMWEL の主な特徴は，以下の通りである．

（1）表現の種類と異表記形が比較的充実している．

（2）表現の構文構造を与えている．

（3）内部修飾によるギャップの可能性を記載している．（例えば，「手を広げる」に対する「手を/外国にまで/広げる」の可能性）

（4）呼応現象をデータ化している．（例えば，「たった一つも・・・ない」）

（5）不完全慣用表現を収録している．（例えば，「猫に小判」，「ピンからキリまで」）

2.2. 採録表現

JMWEL では，新聞記事，雑誌記事，小説，随筆，事典・辞書類などの広範な文書から，主として編者の内省により，非構成(イディオム)性，および，要素語間の強い共起性のうち少なくとも一方の性質を持つ単語列を MWE として抽出・収集した．JMWEL の見出し 2,000 個程度をランダムに抽出して調べたところ，約 38%が非構成性を，約 92%が強い単語間共起性を持ち，両方を併せ持つ MWE は 30%程度であった．なお，MWE も単語同様，自立語性 MWE と機能語性 MWE（複合辞に相当）に大別される．

2.2.1. 非構成性

要素単語の標準的な機能から表現全体の意味を規則で導くことが難しい表現を非構成性 MWE として収録した．形式的には，単語列 $w_1w_2\cdots w_n$ がまとまった構文・意味・談話上の機能をもち，かつ，$w_1w_2\cdots w_n$ におけるいずれかの単語 $w_i(1\leq i\leq n)$ を w_i の同義語または類義語 x_i に置き換えた $w_1w_2\cdots w_n$ が意味をなさなくなるか，全く異なる意味になる，あるいは，不自然になるとき，単語列 $w_1w_2\cdots w_n$ は非構成性 MWE であると近似できる[3]．例えば，「赤の他人」は“全く知らない人”の意味では「真紅の他人」に，また「顔を売る」は“アピールする”の意味では「顔を販売する」に置き換えることができないため非構成性MWE であるとする[4]．この判断は基本的に内省によっている．非構成性 MWE には

1 例えば，I. A. Sag らは，WordNet1.7 での見出しの約 41%が MWE であること[1]，田辺らは，日本語の述部における文末表現を構成する助動詞，終助詞相当の MWE の出現比率が約 42%であること[4]を報告している．

2 例えば，認知心理学分野で，人の言語理解や言語産出において定型表現が意味のある 1 つのまとまりとして想起，処理されているとの N. Jiang らなどの研究報告がある[5][6]．

3 JMWEL に収録された MWE における n の最大値は 18 である．

4 単語の置換不可能性（non-substitutability）がコロケーションのもつ重要な性質の 1 つであることは D. Manning らにも指摘されている[10]．

表 1 に示すような種類がある.

2.2.2. 要素間の強い共起性

構成する単語間で共起性が強い表現を採録した.この種の表現は,構文・意味解析において係り先を優先的に決定して解析の曖昧さを低減する処理や語の出現を予測する種々の処理に有効である.形式的には,単語列 $w_1w_2\cdots w_n$ がまとまった構文・意味・談話上の機能をもち,かつ,$w_1w_2\cdots w_n$ におけるいずれかの単語 $w_i(2\leq i\leq n)$ について条件付後方出現確率 $pf(w_i|w_1\cdots w_{i-1})$ が,あるいは,単語 $w_j(1\leq j\leq n-1)$ について条件付前方出現確率 $pb(w_j|w_{j+1}\cdots w_n)$ が相対的に高いという確率的な特異性を持つとき,単語列 $w_1w_2\cdots w_n$ は単語間共起性の強い MWE であると考える.例えば,「手を拱く」,「ぐっすり眠る」は,pb(手|を拱く),pf(眠る|ぐっすり)が大きいと判断して単語間共起性の強い MWE であるとする.この基準は内省によって判断しているが,結果の妥当性は WEB 上の大量日本語コーパスを用いて統計的に推定されている[9].単語間共起性の強い MWE には**表 2** に示すような種類がある.

2.3. 編成

JMWEL は,以下のように分割して編纂・管理・公開している.以下,**1**〜**11** が文法機能別の部分レキシコン,**12**〜**18** がトピック別の部分レキシコンである.**19** は現在構築中である.

1.名詞性複単語表現レキシコン　JMWEL_nominal:「無二の親友」,「あれやこれや」,「愚の骨頂」などの約 23,600 表現

2.動詞性複単語表現（1 類）レキシコン　JMWEL_verbal (class 1):「手を結ぶ」,「意味がある」,「沽券に関わる」など,『名詞』＋「が,を,に」＋『動詞』の形式約 35,800 表現

3.動詞性複単語表現（2 類）レキシコン JMWEL_verbal (class 2):「骨の髄までしゃぶる」,「ゼロからやりなおす」,「目から鱗が落ちる」など 1 類,3 類以外の動詞的な句約 17,000 表現

4.動詞性複単語表現（3 類）レキシコン　JMWEL_verbal (class 3):「放り出す」,「飲んだくれる」,「秋めく」などの複合動詞的な句約 3,700 表現

5.形容詞性複単語表現レキシコン　JMWEL_adjective:「頭が痛い」,「性格がきつい」,「途方も無い」などの約 5,200 表現

6.形容動詞性複単語表現レキシコン　JMWEL_adjective verbal:「願ったり叶ったり」,「足手纏い」,「判で押した様」などの形容動詞性の句約 2,600 表現

7.連用修飾複単語表現レキシコン　JMWEL_adverbial:「思いもよらず」,「気を引き締めて」,「心を鬼にして」などの連用修飾句（副詞的な句）約 16,300 表現

8.連体修飾複単語表現レキシコン　JMWEL_adnominal:「世に言う」,「筋の通った」,「得も言われぬ」などの連体修飾句（連体詞的な句）約 16,300 表現

9.談話指標的表現レキシコン　JMWEL_discourse marker:「そうは言っても」,「とはいえ」,「驚くべき事に」など,文頭の談話指標的あるいは文副詞的な約 1,300 表現

10.文末表現（終助詞,助動詞性表現）レキシコン JMWEL_post-predicative:「〜かもしれない」,「〜てもよろしい」,「〜たところだ」,「〜なければなりません」,「〜で頂けませんでしょうか」など,話者の態度や相互行為情報,判断情報,（広義の）モダリティー情報等を与える助述（文末）表現,約 4,650 種

11.関係表現（格助詞,副助詞,接続助詞性表現）レキシコン　JMWEL_postpositional:「〜における」,「〜のいかんにかかわらず」,「〜の甲斐あって」,「〜ところの」,「〜を励みに」,「〜を機に」,「〜かの如く」,「〜に従って」,「〜もそこそこに」などの助詞的表現約 2,700 種

12.慣用句レキシコン　JMWEL_idiom:「油を売る」,「真っ赤なウソ」,「足が遅い」などの約 4,500 表現

13.格言・諺・成句・決まり文句レキシコン JMWEL_ proverb saying cliché:「河童の川流れ」,「義を見てせざるは勇無きなり」,「清水の舞台から飛び降りる」などの約 4,000 表現

14.オノマトペ表現レキシコン　JMWEL _ onomatopoeic:「グラリ」,「カラカラと」,「ガッツリ食う」などの擬声・擬態・擬音語,及び,それらを伴う典型表現約 19,500 種

15.四字熟語レキシコン　JMWEL_four character word:「切磋琢磨」,「支離滅裂」,「魑魅魍魎」などの約 3,500 表現

16.慣用的不完全句レキシコン　JMWEL_incomplete phrase:「病は気から」,「棚からボタ餅」,「蟹の甲より年の功」,「石の上にも三年」など,独立してよく使われる構文的に不完全な句(省略を含む句)約 470 表現

17.クランベリー型表現レキシコン　JMWEL_cranberry:「しがみつく」,「後ろめたい」などのクランベリー形態素（候補)を含む約 180 表現

18.呼びかけ・応答・挨拶・独言・間投表現レキシコン JMWEL_call response greeting monologue interjection:「参ったなあ」,「どういたしまして」,「あらマア」,「オット」,「本当？」などの約 1,100 表現で,＜驚き＞,＜疑問＞,＜困惑＞など,発話者の感情 27 種と重み付きで対応付けられている

19.用例文と英訳付き複単語表現レキシコン JMWEL_with J/E sample sentences:複単語表現約 6,500 に対して用例文とその英訳(案)が記載されている.例えば,慣用句「油を売る」には「彼は勤務中に油を売ってばかりいる. He is always_messing around_while on his duty.　彼はよくあの居酒屋で油を売る. He often_wastes time in idle talk_ at that pub.」などと記載している.

3. 動詞性複単語表現レキシコン JMWEL_verbal

動詞性複単語表現（1 類）辞書 JMWEL_verbal（class1）は,日本語文の最小基本型とも云える　(1)『名詞』+「を」+『動詞』,　(2)『名詞』+「が」+『動詞』,　(3)『名詞』+「に」+『動詞』 のいずれかの形式の書き言葉動詞性 MWE 約 35,800 を収録した計算機用辞書のプロトタイプである.（ただし,『サ変名詞』+「を」+「する,遣る,行う,実

行する」,『サ変名詞』+「が」+「できる」の形式の表現は一部を除き収録対象外としている.)

いっぽう，動詞性複単語表現（2 類）辞書 JMWEL_verbal（class2）は，上記 1 類と 3 類動詞性 MWE（複合動詞的表現）を除く書き言葉動詞性 MWE 約 17,000 を収録した計算機用辞書のプロトタイプである.

表現の採録基準は,前述の如く非構成性と要素語間の強い共起性であるが,自由結合句に比較的近いコロケーションから典型的慣用句,典型的決まり文句に亘るかなり広い編集となっている.

3.1. JMWEL_verbal の記載情報

本辞書は，Microsoft Excel で作られた xls 形式のファイルで提供される．xls ファイルの各 1 行に 1 表現を対応付け，A～K 欄に各種情報を記載している．例えば，「労に報いる」という表現に対して与えた情報を A～K 欄の順に列挙すれば以下のようになる．(欄の区切りを・・・で，空データをφで示す.)

class1・・・ろうにむくいる・・・ろう-に-むくいる・・・労-に-報いる・・・VP_a3・・・[*Nni]*V30・・・<adnom. modifier-no>*・・・むくいる・・・報いる・・・φ・・・φ

A～K 欄の情報は，概要，以下の通りである．(詳細は解説書[11][12]を参照のこと.)

A 欄（種別）：動詞性表現の区分を記す.
class1：1 類，class2：2 類，class3：3 類

B 欄（見出し）：平仮名ベタ書き見出しを与える．末尾の活用語は終止形（一部，命令形）で収録している.

C 欄（分かち書き）：形態素分かち書きを示す．形態素には単語，接頭語，接尾語，接頭造語要素，接尾造語要素がある．形態素間の区切りはハイフン「-」(明確な区切り）あるいはアンダースコア「_」(弱い区切り）で示している.

D 欄（異表記）：片仮名表記，漢字表記，送り仮名の有無など，表記の多様さを一種の正規表現で記載している．例えば，「行(な)う」は「行なう」,「行う」の可能性,「(在/有)る」 は「在る」,「有る」の可能性を示す.

E 欄（形態種別）：形態上の種別を VP_α_βの形式にコード化して情報を与える．VP は動詞句（Verb Phrase）を意味する．α部は，例えば，英数字列 a1 で表現が『名詞』＋「を」＋『動詞』の形式であること,d7 で『名詞』＋「が」+『名詞』＋「を」＋『動詞』の形式であることを示す.
β部は表現末尾に助動詞「られる」,「させる」,「ない」,「ぬ」などや形式的自立語,「する」,「ある」などが用いられている場合に,それらを英小文字綴り rareru, saseru, nai, nu, suru, aru などで表わしている.

F 欄（構文構造）：係り受けの修飾子，被修飾子の対を括弧[ ]で括った 2 項句表示で構文構造を与える．即ち，句α（の主辞）が句β（の主辞）に係って出来た句の構造記述を，α，βの構造記述 a，b を使って[ab]とする．構造記述のベースとして自立語を品詞記号で，機能語(および相当語)を英小文字の綴りで与える．(ただし，単語の品詞記号や英字綴りを括弧[ ]で括ることはしていない.) 例えば，「顔-を-揃える」の構造記述は[[*Nwo]*V30]とする．ここで，N は「顔」が名詞であること，V30 は「揃える」が終止形の動詞であることを表す品詞記号． wo は,「を」が機能語（格助詞）であることを意味する． アステリスク「*」は，直後の N「顔」が「元気な顔」のような内部修飾を，V30 の「揃える」が「皆が揃える」のような内部修飾を受ける可能性があることを示している.このようにアスタリスク「*」は後接する句の表現内での独立性を示し，表現中にギャップが生じる可能性がある事を示すものである．機能動詞「する」,「やる」,「ある」,「なる」，機能性形容詞「ない」などは例外的に活用形を含めて英小文字で suru， yaru，aru，naru，nai と表記している.
並列構造は，括弧表現< >または《 》で，並列される要素は括弧( )で表わす．例えば，「見(栄/得)-も-外聞-も-捨てる」の構造記述は[<(Nmo(wo))(Nmo(wo))>*V30]とする．ここで，(wo)は，係助詞「も」が深層ではヲ格で使われていることを示す.

G 欄（前方文脈条件）：例えば,「目-に-会う」は,「つらい-目-に-会う」のように「目」に対する修飾句を必須的に要求するが，修飾句を表現レベルでは特定しにくいので，連体修飾句が文頭側に必須であることを G 欄に<adnom. modifier>*と記載している.

H 欄，I 欄（主動詞部）：収録表現の末尾動詞部はすべて終止形（または命令形）である．終止形以外をカバーするには末尾の主動詞部を活用変化させればよいが,本辞書では,すべての活用形を見出し化しておくことは行わず,H,I 欄に末尾主動詞部を抜き出して再録するに留めている.これにより，必要な活用形は，一般の動詞辞書等を用いれば比較的容易に導出できる.

J 欄（活用）：一体性（fixedness）の特に強い表現，例えば，格言，諺，決まり文句，一部の古語表現などは，末尾を活用変化させて用いられることは殆どない.この種の表現には J 欄に「活用不要」などの記載をしている．この欄は表現の“強い硬さ”の標識と見做すことができる.

K 欄（語釈）：慣用句，諺，格言，決まり文句の意味が難解と思われる表現 500 種程度（1 類の場合）にはユーザーの便宜のため語釈を入れている.

4. おわりに

本論文では,筆者の一人が中心となって編纂した日本語複単語表現レキシコン(JMWEL)についてその概要と現在の状況を述べ，2017 年 11 月に言語資源協会（GSK）より公開された動詞性複単語表現（1 類および 2 類）サブレキシコンについては，やや詳細に説明を加えた．本研究の発端は，機械翻訳のためにフレーズを一括認識・理解する意味ベースの日本語文解析システムを開発することであった．意味の取扱いについては，現在もなお問題山積であるが，言語表現サイドから改めて意味の問題に切り込むための

一次資源として JMWEL は有効であろうと考えている．ワープロ，機械翻訳のように比較的表層レベルで成果の見込めそうな処理には JMWEL のより直接的な利用が考えられる．

自然言語は，言わば広大無辺の大海であり，多彩な言語表現の宝庫である．JMWEL がどこまで専門分野，方言を除く書き言葉日本語特異表現の全体像に迫ることができているか，今後，ユーザー諸賢による評価とさらなる改善を仰ぎたい[5]．また，JMWEL がこれからの日本語処理研究を支える言語資源のプロトタイプとして役立てられることを期待したい．

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] Corrigan, R., Moravcsik, E. A., Ouali, H. and Wheatley, K. M. (eds.): Formulaic Language, vol.1, Distribution and historical change, John Benjamins Publishing Company (2009).

[3] Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (eds.). : Longman Grammar of Spoken and Written English, Harlow: Pearson Education Limited (1999).

[4] 田辺利文，本田聖晃，高橋雅仁，小山泰男，吉村賢治，首藤公昭：文末表現の取り扱いについて，FIT2006，pp.241-244 (2006).

[5] N. Jiang, T. M. Nekrasova,: The processing of formulaic sequences by second language speakers, Modern Language Journal, 91, pp.433-445 (2007).

[6] 鳥羽素子：日本人英語学習者の語彙連想とライティング力との関係，都市文化研究，Vol.18，pp46-57 (2016).

[7] 首藤公昭，田辺利文：日本語の複単語表現辞書 : JDMWE，自然言語処理，Vol.17，No.5，pp.51-74 (2010).

[8] K. Shudo, A. Kurahone, and T. Tanabe: A Comprehensive Dictionary of Multiword Expressions, Proceedings of the 49th Annual Meeting of the ACL: pp.169-177 (2011).

[9] T. Tanabe, M. Takahashi, and K. Shudo: A lexicon of multiword expressions for syntactically precise, wide-coverage natural language processing, Computer Speech and Language, Vol.28, No.6, pp.1317-1339, Elsevier (2014).

[10] Manning, D., Schutze, H.: Foundations of Statistical Natural Language Processing, MIT Press (1999).

[11] 首藤公昭 : 動詞性複単語表現（1 類）辞書 : JMWEL_verbal (class1) v3.2 --- ガ格，ヲ格，ニ格を介した動詞と名詞のコロケーション集（慣用句等を含む） --- 解説書 (2017)..

[12] 首藤公昭 : 動詞性複単語表現（2 類）辞書 : JMWEL_verbal (class2) v3.2 --- 述語動詞と種々の語とのコロケーション集 --- 解説書 (2017).

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

表 1 非構成性 MWE の例

種類	例
意味上の非構成性を持つ表現	赤の他人，顔を売る，頭が切れる
形態・構文上での構成性が不備，あるいは不明瞭な表現	とはいえ，ありがとう，お疲れ様
一部の支援動詞構文	批判を加える，計画を立てる
一部の複合語	打ち拉がれる，袋叩き
四字熟語	一生懸命，一心不乱
慣用的な比喩表現	命の限り，血の雨が降る

表 2 単語間共起性の強い MWE の例

種類	例
共起性の特に強い表現	風前の灯，ずぶの素人，手を拱く
格言，諺，故事成句の類	急がば回れ，初心忘る可からず，石の上にも三年
擬声，擬音，擬態語を伴う表現	ぐっすり眠る，ポッカリと空く，クルクル回る
その他共起性が比較的強いと思われる表現	肩の荷を下ろす，景気が上向く，メリハリの利いた
概念に固有の固定的言い回し	情報検索，女流作家，機械翻訳

5 JMWEL は，K. Church の疑問 [13]に対する一つの回答試案と位置づけられる．