日本語複単語表現レキシコンJMWELの概要：動詞性表現を中心に

日本語複単語表現レキシコン JMWEL の概要
－ 動詞性表現を中心に －

首藤 公昭（福岡大学名誉教授）
田辺 利文（福岡大学）
髙橋 雅仁（久留米工業大学）

An Overview of
a Lexicon of Japanese Multiword Expressions: JMWEL

Kosho Shudo (Fukuoka University, professor emeritus)
Toshifumi Tanabe (Fukuoka University)
Masahito Takahashi (Kurume Institute of Technology)

要旨

コロケーション，決まり文句，慣用句，準慣用句などの長単位表現とその派生表現，計約 140,000 の見出しからなり，平仮名べた書き見出しのほかに，形態素分かち書き，構文機能，構文構造，内部修飾可否情報，文脈条件，呼応情報，語釈などを与えた日本語複単語表現レキシコン JMWEL の概要を動詞性表現を中心に紹介する．

１．はじめに

自然言語におけるコロケーション，慣用句，決まり文句など，単語の境界を越えた長単位表現は，従来，計算言語学 Computational Linguistics, CL や自然言語処理 Natural Language Processing, NLP の分野では例外的言語現象とみなされ，必ずしも十分な対応がなされてこなかった．しかし，近年，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が改めて認識され，(Sag et al. 2002) がこの種の表現を複単語表現 Multiword Expression, MWE と名付けて NLP における重要性を指摘したのを発端に，それらの表現のコーパスからの自動抽出，レキシコン開発等々，コンピュータ処理に向けた種々の基礎研究が各国で行われるようになって今日に至っている[1].

また，この種の表現は言語の獲得・認知の観点からも重視されるようになり，言語学分野でも定型言語 Formulaic Language (Corrigan et al. 2009, Jiang et al. 2007)，単語連鎖 Lexical Bundles (Biber et al. 1999)，構文文法 Construction Grammar (Fillmore et al. 1988)などの枠組みで，話し言葉，書き言葉の両面から盛んに研究が行われている．

本稿では，筆者の一人が 1960 年代からフレーズ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日本語複単語表現レキシコン Japanese MWE Lexicon，JMWEL の概要と現状を下記の動詞性表現部分レキシコン（１），（２）を中心に紹介する．(Tanabe et al. 2014, 高橋ほか 2018, Shudo et al. 2011, 首藤ほか 2010)　JMWEL の収録見出し数は，異なりで現在 140,000 件程度である．JMWEL の動詞性表現部分レキシコンには，その主要部分として

（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）－，（2）日本語動

[1] 例えば、SIGLEX-MWE(http://multiword.sourceforge.net/)にあるワークショップの一覧をみると 2018 年には LAW-MWE-CxG2018 が開催されている．

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（慣用句等を含む）－ が有る．

２．日本語複単語表現レキシコン JMWEL

２．１ JMWEL の特徴

（1）収録表現は，コロケーション，慣用句，準慣用句，決まり文句，格言，諺，一部の複合語，四字熟語，不完全句，挨拶・呼びかけ・応答表現など，日本語における特異表現を幅広くカバーしている．

（2）表現の構文機能，形態・構文構造を与えている．

（3）必ずしも隣接しない語の共起もデータ化している．（例えば，慣用句「手を広げる」には「手を/外国にまで/広げる」など，ギャップの可能性を記載.）

（4）語の長距離呼応をデータ化している．（例えば，「たった一つも/世の中に存在し/ない」など）

（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」など）

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特異性を持つ単語列を MWE として抽出・収録した．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．

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．ここでは，単語列 $w_1w_2…w_n$ がまとまった構文・意味・談話上の機能をもち，かつ，$w_1w_2…w_n$ におけるいずれかの単語 w_i $(1≤i≤n)$ をその同義語または類義語 x に置き換えた $w_1w_2…w_{i-1}xw_{i+1}…w_n$ が無意味になるか，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 $w_1w_2…w_n$ は非構成性 MWE であると近似する[2]．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．この判断は基本的に内省によっている．例えば，非構成性 MWE には表 1 に示すような種類がある．

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

[2] このような単語の置換不能性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（Manning et al. 1999）でも指摘されている．

２.２.２　要素間の強い共起性

表現を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 $w_1w_2…w_n$ がまとまった構文・意味・談話上の機能をもち，かつ，$w_1w_2…w_n$ におけるいずれかの単語 w_i $(2≤i≤n)$ について条件付後方出現確率 $pf(w_i|w_1…w_{i-1})$が，あるいは，単語 w_j $(1≤j≤n-1)$ について条件付前方出現確率 $pb(w_j|w_{j+1}…w_n)$ 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 $w_1w_2…w_n$ は単語間共起性の強い MWE であるとする．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，3.2 で述べる如く，収録結果の妥当性は 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．単語間共起性の強い MWE には，例えば，表 2 に示すような種類がある．

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

２.３　JMWEL の編成

対象表現が多岐にわたるため，JMWEL は，以下のように分割して編集・管理している．以下の 1〜9 は自立語性表現部分レキシコン，10，11 は機能語性表現（複合辞的表現）部分レキシコン，12〜18 は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』の形式の句約 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」などの形容詞句約 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，文接続詞的，文副詞的な句約 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，テンス，アスペクト，モダリティ，ポラリティ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」などの典型的慣用句約 4,900 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. オノマトペ共起表現レキシコン JMWEL _ onomatopoeic：
「グラリ」,「カラカラと」,「ガッツリ食う」などの擬態・擬音語とそれらを伴う典型表現約 34,500 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，句に纏まらない表現約 470 種

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)を含む表現約 180 種

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 5,000 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”.」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）レキシコン JMWEL_verbal (class1)は，日本語文の最小基本型とも言える　(1)『名詞』+「を」+『動詞』，(2)『名詞』+「が」+『動詞』，(3)『名詞』+「に」+『動詞』 の三つの形式の書き言葉動詞性 MWE 約 35,800 を収録したレキシコンである．（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）レキシコン 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 を収録したレキシコンである．

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

本レキシコンは，Microsoft Excel で作られた xlsx 形式のファイルとして作成されている．xlsx 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，概略，以下の通りである．

A 欄（種別）：動詞性表現の種別を記す．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．活用語尾は一部の例外を除き，切り離さない[3]．

D 欄（異表記）：片仮名表記，漢字表記，送り仮名の有無など，表記の多様さを正規表現類似の形式で記載している．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与える．VP は表現が動詞句（Verb Phrase）であることを意味する．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，それらを英小文字ローマ字綴り rareru, saseru, nai, nu, suru, aru などで表わす．

[3] 本レキシコンには，形態素解析機との整合を取る際に有効な情報として，ハイフン，アンダースコアで区切った単語候補，複合語候補のリストを添付している．

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記述を与える．即ち，句 α（の主辞）が句 β（の主辞）に係って出来た句 αβ の構造記述を，α，β の構造記述 a，b を使って[ab]とする．基本構成要素の構造記述は，自立語を品詞記号で，機能語(および相当語)を英小文字のローマ字綴りで与える． 文節内の語の接続も，便宜上，2 項句構造として記述する．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．アスタリスク * は，直後の N「顔」が「元気な顔」のような連体修飾を，V30 の「揃える」が「皆が揃える」のような連用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における独立性を示し，表現中にギャップが生じる可能性がある事を示している． 並列構造は，括弧< >または《 》で，並列される要素は括弧( )で表わす．例えば，「見栄-も-外聞-も-捨てる」の構造記述は [<(Nmo(wo))(Nmo(wo))>*V30] とする．ここで，mo(wo)は，係助詞「も」が深層の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」に対する連体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）：収録表現の末尾主動詞部は終止形（一部，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，本レキシコンでは，すべての活用形に応じて見出し化しておくことはせず，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． 一般の動詞辞書等を用いれば，必要な活用形はこの情報で容易に導出できる．

J 欄（活用）：一体性の特に強い表現，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．

K 欄（語釈）：慣用句，諺，格言，決まり文句でそ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．

３．２　JMWEL_verbal(class1)の統計的性質

200 億文からなる日本語 WEB コーパスにおける単語 1～7 グラムの出現頻度を求めた Google の大規模データ GSK2007-C（工藤ほか 2007），(以降，GoogleN グラムデータと略記する）との比較によって，JMWEL_verbal(class1)の統計的性質を調べた[4]．詳細は（田辺ほか 2018，Tanabe et al. 2014）に譲るが，主要な調査結果は以下の 2 点である．以下では，対象複単語表現を $w_1w_2w_3$ と記す．ここで，w_1，w_2，w_3 は，それぞれ，『名詞』，『格助詞（「を」，「が」，「に」のいずれか)』，および『動詞』とする．

１．本レキシコン収録表現 $w_1w_2w_3$ の前部分列 w_1w_2 の 14,075 表記の内，10,548 種が GoogleN グラムデータの $w_1w_2w_3$ の w_1w_2 として出現していた．GoogleN グラムデータ上でそれら 10,548 個の w_1w_2 ごとに，続く動詞の出現頻度を求めた結果，本レキシコンにおける動詞 w_3 が GoogleN グラムデータ上で出現頻度第 1 位である場合が 4,983 件であり，当該表現 w_1w_2 の (4,983/10,548)*100=47.24 % で条件付出現確率が最大の動詞 w_3 が選ばれていると推定できた．「ちょっかい-を-出す」，「熱戦-を-繰り広げる」，「アクション-を-起こす」などがこれらに該当する．同様に，第 2 位の場合は 1,495 件で 14.17%，3 位

[4] 調査対象の JMWEL_verbal(class1)は 2010 年時のバージョンである．

は 786 件で 7.45%，4 位は 433 件で 4.11%であった．20 位までの結果をグラフ化して図 1(a)に示す．このことから，本レキシコンに収録されている表現は，条件付確率 $p(w_3|w_1w_2)$の高いものほど多いという傾向が確認された．

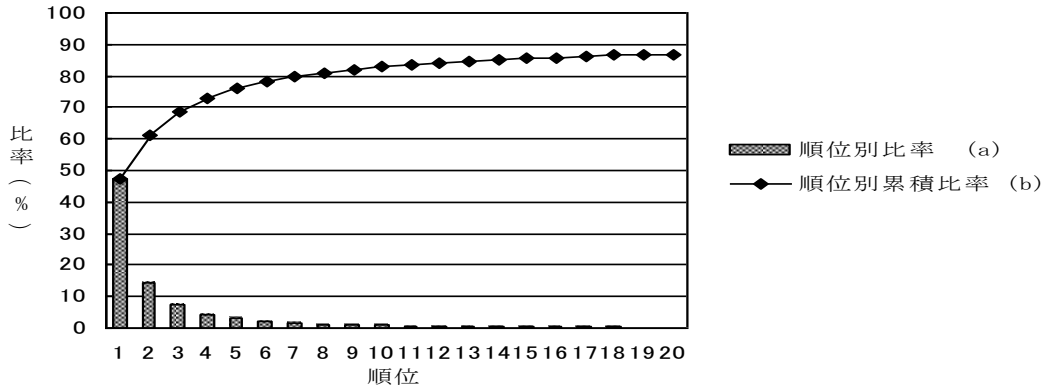

図 1　『名詞』+『格助詞（を，が，に）』+『動詞』型表現の GoogleN グラムデータ
における動詞の出現頻度順位別の動詞採録率(a)，および，順位別の動詞採録累積比率(b)

２．GoogleN グラムデータにおける上記と同じ形式の表現 $w_1w_2w_3$ において，w_1w_2 に続く w_3 に関する正規化エントロピー$H_f(w_3|w_1w_2)$を次式によって求めた．ここで N は動詞 w_3 の種類の数である．

$$H_f(w_3 \mid w_1w_2) = -\left(\sum_{w_3} pf(w_3 \mid w_1w_2) \log pf(w_3 \mid w_1w_2)\right) / \log_2 N$$

次に，GoogleN グラムデータの w_1w_2 を，得られた $H_f(w_3|w_1w_2)$の昇順に並べたうえで 20 区間に分割し，それぞれの区間において本レキシコンの w_1w_2 型表現(計 10,548 件)が含まれる比率を求めた．各区間の含有比率をグラフ化して図 2(a)に，各区間の平均エントロピーを図 2(b)に示す．結果として，本レキシコン w_1w_2 型表現の各区間における含有率は，区間 1 の場合 (1,262/5,542)*100=22.8 %，区間 2 の場合は 22.5%，区間 3 では 20.5%であり，区間 4 以降でも順次低くなっていることが観察された．このことから，本レキシコン収録表現における前部分列 w_1w_2 は，続く動詞 w_3 に関する正規化エントロピー $H_f(w_3|w_1w_2)$が小さいほど，すなわち，後接する動詞部のパープレキシティが小さいものほど多く採録されているという傾向が見られた．本レキシコンにある「墓穴-を（-掘る）」「難色-を（-示す）」「凶弾-に（-倒れる）」などが区間 1(平均エントロピーは 0.27)に含まれていた[5]．エントロピーの大きい表現は解析の曖昧さ低減や予測にあまり有効ではないため，通常の単語単位の処理に任せるのが妥当であると考えており，ほぼ期待された結果である．

基本的には文頭側から文末側へ向かって人の文理解が進むと考えれば，上記の検証は有意味であろうと考えている．また，JMWEL のすべての部分レキシコンにおいて，ほぼ同一基準で表現が採集されているので JMWEL 全体も上記と大差のない傾向を有しているものと考えている．

[5] これらの表現は，いずれも GoogleN グラムデータ上で出現頻度第 1 位であった．

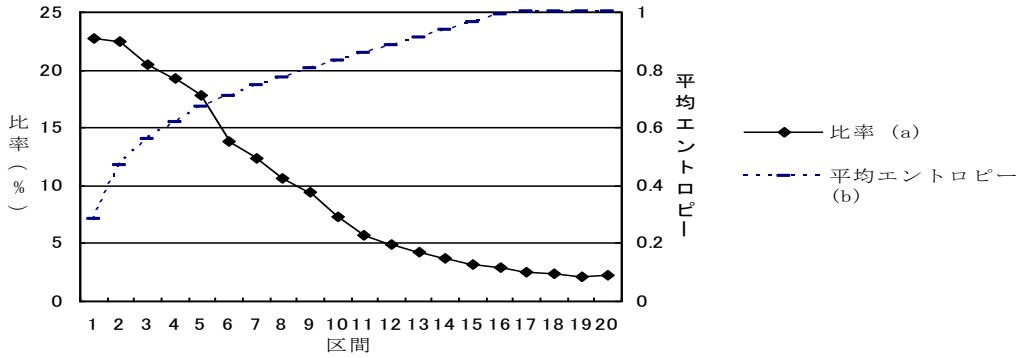

図２　『名詞』＋『格助詞（を，が，に)』 型表現の GoogleN グラムデータにおける後続動詞(正規化) エントロピー区間別採録率(a)，および，各区間の平均エントロピー(b)

４．JMWEL の応用

日本語処理において意味的な纏まりをもつ JMWEL の収録表現を処理単位とすることには大きな利点がある．例えば，JMWEL には，「手に付かず」，「散歩に出る」，「ことにする」という表現が，それぞれ，連用修飾複単語表現，動詞性複単語表現（1 類），文末表現（終助詞，助動詞性表現）として各レキシコンに収録されており，また，それぞれ，連用修飾句 AdvP，動詞句相当表現 VP，助動詞相当表現 Aux であり，構造記述は [[Nni][V12zu]], [[Nni]V30], [[Nni]suru]であることが各レキシコンに記載されている．さらに，「手に付かず」に対しては前方文脈条件としてガ格の後置詞句が必要であることが指定されている．そこで，例えば，入力文「彼は仕事が手に付かず散歩に出ることにした」に対してこれらの情報を用いた一つの構文解析過程のイメージを図 3 に示す．この例のように、MWE を単位的に扱うことで，実質文節数を削減すると同時に文意に近い解析の可能性が高まる[6]．

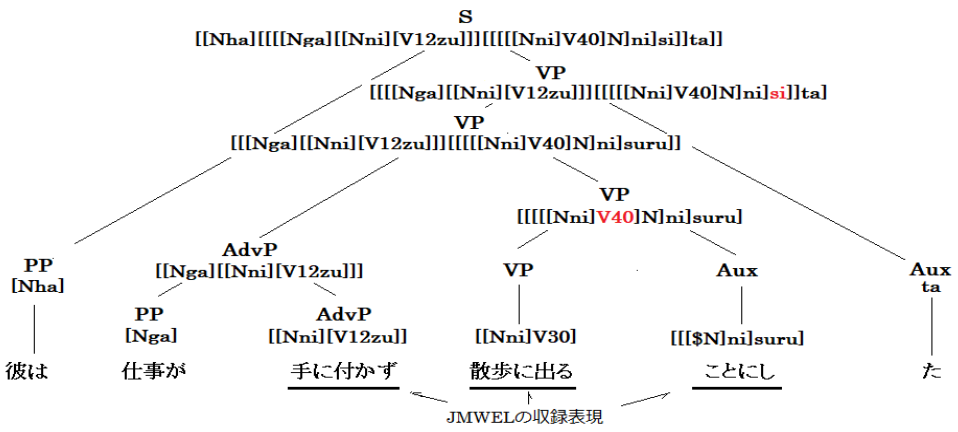

図３　JMWEL による構文解析のイメージ

また，上記のそれぞれの表現に，英訳情報，例えば，“as SUB is unable to get down to doing SUB’s N”，“(to) go out for a walk”，“(to) decide to”を与えたとすると，図 3 の解析にほ

[6] MWE レキシコンを用いる日本語構文解析の手法については，特許第 5379318 号がある．

ぼ平行した形で図 4 のように意味に基づいた日英翻訳が行える可能性が生じる．
その他，日本語の音声認識，日本語による知的対話システム，日本語ワープロの高度化，日本語教育など，JMWEL には広範な応用領域が想定される．

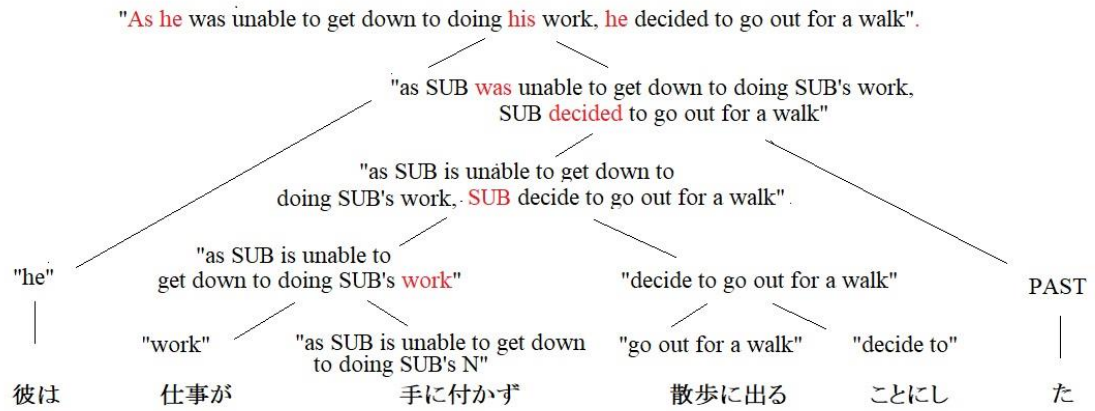

図４　JMWEL による日英機械翻訳のイメージ

５．おわりに

JMWEL は，膨大な日本語の単語 n グラム(2≤n≤18) 集合から，纏まった構文・意味・談話機能を持ち，非構成性，あるいは，要素語間の高い親和性を持つ n グラムだけを掬い取った部分集合の試案である．自然言語の意味処理を試みる際，通常，語類や意味素性による語の共起ルールが作られるが，それでは捉えられない慣用句的表現，決まり文句的表現が思いのほか多く，また，その方法では語の共起度合いの強弱が捉えにくい．本稿で紹介した日本語複単語表現レキシコン JMWEL は．そのような基本認識に基づいて編纂された．
意味の取扱いについては現在なお問題山積であるが，JMWEL は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有効ではないかと考えている．機械翻訳のように表層レベルの処理である程度の成果が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．
自然言語は言わば言語表現の大海であり，JMWEL の表現収録が十分網羅的であるとは言い切れないが，専門分野，方言を除く書き言葉日本語における特異表現については一つの言語資源プロトタイプとして機能するであろうと考えている[7]．JMWEL がこれからの日本語処理用，日本語研究用の基礎的言語資源として活用され，さらに充実されることを期待したい[8]．

[7] JMWEL は，K. Church の疑問(Church 2011)に対する現時点の一つの回答試案と位置づけられる．
[8] JMWEL の利用については関連サイト http://jefi.info を参照されたい．

文　献

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan (eds.) (1999). "Longman Grammar of Spoken and Written English", *Harlow: Pearson Education Limited*.

Kenneth Church (2011). "How Many Multiword Expressions do People Know?", *Proceedings of the MWE workshop(MWE2011)*, ACL, pp.137-144.

Roberta Corrigan, Edith A. Moravcsik, Hamid Ouali and Kathleen Wheatley (eds.) (2009). "Formulaic Language, vol.1, Distribution and historical change", *John Benjamins Publishing Company*.

Charles J. Fillmore, Paul Kay and Mary Catherine O'Connor (1988). "Regularity and Idiomaticity Grammatical Construction: The Case of Let Alone" *Language* 64, pp.501-538.

Nan Jiang and Tatiana M. Nekrasova (2007). "The Processing of Formulaic Sequences by Second Language Speakers", *The Modern Language Journal*, 91:3, pp.433-445.

工藤拓・賀沢秀人 (2007). 「Web 日本語 N グラム第 1 版」言語資源協会.

Christopher D. Manning, Hinrich Schűtze (1999). "Foundations of Statistical Natural Language Processing", MIT Press.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger (2002). "Multiword Expressions: A Pain in the Neck for NLP" *Proc. of the 3rd CICLING*, pp.1-15.

Kosho Shudo, Akira Kurahone and Toshifumi Tanabe (2011). "A Comprehensive Dictionary of Multiword Expressions", *Proceedings of the 49th Annual Meeting of the ACL*, pp.169-177.

首藤公昭・田辺利文 (2010). 「日本語の複単語表現辞書：JDMWE」自然言語処理, 17:5, pp.51-74.

高橋雅仁・田辺利文・首藤公昭 (2018). 「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－」言語処理学会第 24 回年次大会発表論文集, pp.428-431.

Toshifumi Tanabe, Masahito Takahashi and Kosho Shudo (2014). "A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing", *Computer Speech and Language*, 28:6, Elsevier, pp.1317-1339.

田辺利文・高橋雅仁・首藤公昭 (2018). 「日本語動詞性複単語表現(1 類)レキシコンの統計的性質」言語処理学会第 24 回年次大会発表論文集, pp.619-622.

関連 URL

日本語処理研究工房ことばの森　　https://jefi.info