E-022

日本語慣用句の機械処理用レキシコン：JMWEL_idiom
A Lexicon of Japanese Idiomatic Expressions for NLP: JMWEL_idiom

田辺 利文* 髙橋 雅仁† 首藤 公昭‡
Toshifumi Tanabe Masahito Takahashi Kosho Shudo

1. はじめに

日本語慣用句レキシコン JMWEL_idiom は，
1. 通常，慣用句とされている表現
2. 慣用句とはされていないが機械翻訳等で意味の構成性に問題があると思われる表現（準慣用句）
3. それらから派生する典型的な表現

の計約 4,950 種を与えた計算機処理用レキシコンである．（以降，これらの表現を慣用句と総称する．）
収録表現は，不特定多数の慣用句事典類[5][6][7][8][9][11][13][14][15][16][18][20][21]のほか，新聞記事，雑誌，小説，テレビ，ラジオ放送文などの生データから内省によって採集した書き言葉表現であり，市販の慣用句辞典類の見出し表現の多くはカバーされている．
本稿では，JMWEL_idiom の概要を報告する．
本レキシコンの特徴は，
1. 漢字，カタカナ表記や送り仮名の有無など，表記の多様性を記載している
2. 表現の構文機能と形態・構文構造を記載している
3. 内部修飾可能性 internal modifiability を表現ごとに記載している
4. 構文的に不完全な句にも対応している

などの点にある．

2. 記載情報

本レキシコンは，Microsoft Excel で作成した xlsx ファイルに収録されており，その各 1 行に 1 個の対象表現を対応させ，見出し，分かち書き，異表記，構文機能・表現末尾情報，構文構造・内部修飾可能性情報，前方文脈条件，後方文脈条件，連体・連用・動詞化形式情報，語釈情報（一部）等をこの順に B～K 欄に与えたものである．
以下，各欄の情報について説明する．

2.1 種別（A 欄）

当該表現が慣用句，準慣用句，あるいは，それらの関連表現である事を「idiom」と記している．

2.2 見出し（B 欄）

音韻処理に配慮して慣用句の平仮名ベタ書き見出しを与える．漢字が複数の読みを持っている場合，可能な読み毎に見出しを与える．例えば，「九死に一生を得る」には，「きゅうしにいっしょうをえる」，「きゅうしにいっしょうをうる」を別見出しとする．

2.3 分かち書き（C 欄）

B 欄の平仮名表記に対し，単語，接辞（接頭語，接尾語，接頭造語要素，接尾造語要）を単位とした分かち書き結果を与える．造語要素とは，造語能力が比較的強く，単独で用いられることのない形態素で，多くの場合，音読みの一漢字である．本レキシコンでは，接頭語，接尾語，接頭造語要素，接尾造語要素を表 1 の様に整理・分類している．
日本語では形態素区切りが判然としない場合が多いので，明らかな区切りを「-」，可能性のある区切りを「_」で与えている．また，単語であっても，その一部が異なった字種で表記可能な場合は字種の変わり目に「_」を入れている．例えば，「りゅういん-が-さがる」の「溜飲」は一単位ではあるが，「りゅう飲」と表記されることがあるので「りゅう_いん」と弱く区切っている．この情報と D 欄の漢字情報「溜_飲」とから「溜飲」，「溜いん」，「りゅう飲」，「りゅういん」という 4 つの表記が導出できる．
「ぜんあく」，「しろくろ」などの対比的意味の語の組み合わせは並列句とみなして「ぜん-あく」，「しろ-くろ」と区切っている．
原則として，活用語尾は語幹から切り離していない．ただし，形容動詞活用語尾「な」，「に」，「なる」，「たる」などは，助動詞「だ」，「なり」，「たり」の活用変化形とみなして語幹と切り離している．

2.4 異表記（D 欄）

片仮名表記，漢字表記，送り仮名の有無など，日本語特有の表記の多様さをコンパクトに記載した欄である．すなわち，表記，n_1, n_2, …, n_m がいずれも使用可能であるとき，$(n_1/n_2/\cdots/n_m)$と括弧内に記号「/」で区分して記載し，表記 n が無くてもよいとき(n)と記載している．例えば，「(思/想/惟/懐)い-も-及ば-ない」における「(思/想/惟/懐)い」の部分は「思い」，「想い」，「惟い」，「懐い」の 4 つの可能性が有ることを表わす．また，「真(っ)-赤-な-嘘」の「真(っ)」の部分は「真」と「真っ」の 2 つの可能性，さらに，「右-も-左-も-(分(か)/解/判)ら-ない」の「(分(か)/解/判)ら」は，「分から」，「分ら」，「解ら」，「判ら」の 4 つの可能性があることを意味する．これらと C 欄の分かち書き情報を加えれば，「わから」，「分から」，「分ら」，「解ら」，「判ら」がカバーされる．

* 福岡大学, Fukuoka University
† 久留米工業大学, Kurume Institute of Technology
‡ 福岡大学名誉教授, Fukuoka University, prof. emeritus

表 1　接頭語，接尾語，接頭造語要素，接尾造語要素の分類

1. 接頭語(P)	1.1 名詞，動詞連用形，形容動詞語幹に前接する接頭語	
	1.2 用言に前接する接頭語	
2. 接頭造語要素 (Q)	2.1 名詞，動詞連用形，形容動詞語幹に前接する接頭造語要素	2.1.1 名詞を与えるもの 2.1.2 形容動詞語幹を与えるもの
	2.2 用言に前接する接頭造語要素	
3. 接尾語(S)	3.1 名詞，動詞連用形，形容動詞語幹に後接する接尾語	3.1.1 サ変名詞以外の名詞を与えるもの 3.1.2 サ変名詞を与えるもの 3.1.3 形容動詞語幹を与えるもの 3.1.4 形容詞を与えるもの 3.1.5 動詞を与えるもの
	3.2 用言に後接する接尾語	3.2.1 名詞を与えるもの 3.2.2 形容動詞語幹を与えるもの 3.2.3 形容詞を与えるもの 3.2.4 動詞を与えるもの
4. 接尾造語要素 (R)	4.1 名詞，動詞連用形，形容動詞語幹に後接する接尾造語要素	4.1.1 サ変名詞以外の名詞を与えるもの 4.1.2 サ変名詞を与えるもの 4.1.3 形容動詞語幹を与えるもの
	4.2 用言に後接する接尾造語要素	4.2.1 名詞を与えるもの 4.2.2 形容動詞語幹を与えるもの

表 2　慣用句の構文機能とその記号

E 欄(α部)の記号	記号の意味	表現例
AdjP	形容詞句　Adjective Phrase	「愛想-が-(良/善/好)い」
AdjVP	形容動詞（語幹）句　Adjective Verb Phrase	「(顎/アゴ)-が-落ちる-様」
AdjVS	形容動詞文　Adjective Verb Sentence	「事実-は-小説-より-奇-なり」
AdnP	連体修飾句　Adnominal Phrase	「当(た)り-障り-の-無い」
AdvP	連用修飾句　Adverbial Phrase	「明け-ても-暮れ-ても」
AdvP/incomplete	不完全連用修飾句	「百-里-の-道-も-一-歩-から」
DM/SA	談話指標・文接続詞 Discourse Marker/Sentence Adverbial	「何-は-とも-(有/在)れ」
NP	名詞句　Noun Phrase	「赤-の-他人」
NP/incomplete	不完全名詞句	「(噂/ウワサ)-を-すれ-ば-(陰/影)」
NP/dynamic	動的名詞句（「をする」で動詞化）	「骨-肉-の-争い」
NP/haikai	俳諧・短歌調名詞句	「(我/吾)(が)-物-と-思え-ば-軽し-傘-の-雪」
NP/sahen	サ変名詞句（「する」，「をする」で動詞化）	「群雄-割拠」
NPS	名詞述語文　Nominal Predicate Sentence	「風邪-は-万病-の-(元/本/基)」
VP	動詞句　Verb Phrase	「身-(銭/ゼニ)-を-切る」

2.5 構文機能等（E 欄）

E 欄は，α_βの形式で構文的機能（α）と表現末尾の種別（β）を表示した欄である．α部は文脈自由文法の非終端記号に相当し，表 2 の左欄に英字で示す記号である．
慣用句の後続語に係る機能は原則として末尾の単語の機能に従う．しかし，例えば，「定規」は単純な名詞であるのに，「杓子-定規」は「な」を送ると連体修飾句となり，形容動詞の働きをするなどの現象も見られる．この種の情報は後述の I 欄で与える．

β部は，表現末尾に接続されている助動詞，「れる」，「られる」，「せる」，「させる」，「ない」や助詞「て」などのローマ字綴り表示である．

2.6 構文構造（F 欄）

2.6.1 係り受け構造

F 欄では，修飾子，被修飾子の対を括弧[ ]で括った句表示で表現の係り受け構造を記述する．すなわち，句α，句βの構造記述，a，b を使って，α（の主辞）がβ（の主辞）を修飾して出来た句αβの構造記述を[ab]と与える．また，文節内部の語の接続も，便宜上，係り受けと同じ括弧 [ ]で 2 項句構造表示をしている．部分句に句構造文法の非終端

記号に相当するものは与えていない．要素単語の構造記述は，自立語を品詞記号で，機能語を英小文字によるローマ字綴りで与える．例えば，「朝起きは三文の徳」という表現の F 欄の構造記述は[[[NV22]ha(ga)][[[NR]no]N]] とし，図 1 の構文構造を与える．

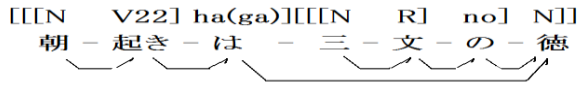

図 1 「朝起きは三文の徳」の構文構造

ここで，N は「朝」，「三」，「徳」が名詞であること，V22 は「起き」が動詞の中止型連用形であること，ha(ga)は「は」が深層のガ格で使われた（係）助詞であること，R は「文（もん）」が接尾造語要素であること，no は「の」が（連体格）助詞であることを意味する．この様に係助詞，副助詞が格助詞抜きで用いられているとき，その深層の格を括弧を使って(ga)などと与えている．また，活用自立語には活用形を数字 2 桁で表示しており，上位桁は活用形（語幹＝0，未然形＝1，連用形＝2，終止形＝3，連体形＝4，仮定形＝5，命令形＝6），2 桁目は活用形の下位種別を表す．

C 欄に与えたハイフン「-」による区切り単位が F 欄のアノテーション単位と 1 対 1 に対応している．

機能動詞の「する」，「なす」，「す」，「ある」，「なる」および機能性形容詞「ない」などは活用形を含めて綴りを英小文字でローマ字表記した．例えば，「足_枷-に-なる」の構造は [[*Nni]naru] とする．構造表記中のアスタリスク「*」は後述するように，それに続く「足枷」が「重い足枷になる」のように連体修飾される可能性があることを示している．

2.6.2 並列構造

並列構造は，括弧< >または《 》で，被並列要素は括弧( )で表わしている．例えば，「泣く-子-と-地頭-に-は-勝て-ぬ」の表示 [[[<([V40N])to(N)>ni]ha]*[V12nu]] は図 2 の構造を意味する．

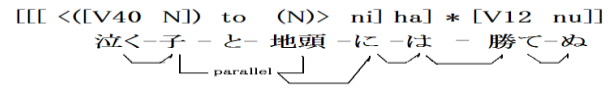

図 2 「泣く子と地頭には勝てぬ」の構文構造

ここで，「泣く-子」と「地頭」が並列されていることをそれぞれ( )で括って([V40N])，(N)と表示し，「泣く-子-に-は-勝て-ぬ」*-and-*「地頭-に-は-勝て-ぬ」とほぼパラフレーズできる分配型並列句であることを括弧< >で表示している．V40 は動詞の連体形，V12 は動詞のズ接続未然形を意味する．アスタリスク「*」については 2.6.4 で述べる．

一般に，A，B，C を句とし，x を「および」，「あるいは」，「か」，「と」，「に」などの並列マーカーとするとき， A[<(B)x(C)>]は<([AB])x([AC])>に，[<(A)x(B)>]C は<([AC])x([BC])>に意味を保存してほぼパラフレーズできる．

また，「盆-と-正月-が-一緒-に-来-た-様」の構造記述，[[[[《(N)to(N)》ga][[Nni]V23]]ta]you] は図 3 の構造を意味する．

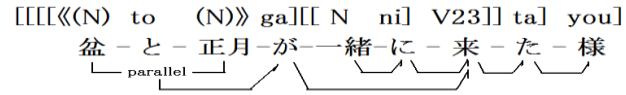

図 3 「盆と正月が一緒に来た様」の構文構造

ここで，V23 は動詞のタ接続連用形を意味する．この例の並列句「盆-と-正月」は「一緒に」という語句によって硬く結合されており，意味上，「盆-が-一緒-に-来-た」*-and-*「正月-が-一緒-に-来-た」と分配的パラフレーズができない束ねの並列句[10]である．

この様に< >は分配的な並列句 distributive coordination，《 》は束ねを代表とする硬い並列句 rigid coordination を意味する．

2.6.3 不完全句

慣用句や決まり文句には，文脈自由文法の句として纏まっていないものがある．本レキシコンでは，この種の表現を不完全句 incomplete phrase と名付け，その構造記述に空要素記号 null constituent symbol 「$」を用いている．例えば，「(暑/アツ)-さ-寒-さ-も-彼岸-迄」には，構造記述 [[<(A00S)(A00S)>mo(ga)][[Nmade]$]] によって図 4 の構造を与える．

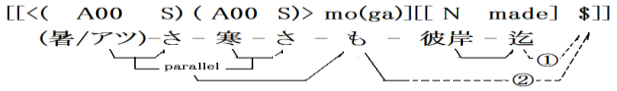

図 4 「暑さ寒さも彼岸迄」の構文構造

ここで，「$」は句を成立させるための空要素記号である．この例では$によって「続く」などの述語が暗黙のうちに想定されていると考える．この表現はこの様な潜在的な句構造を持っており，①，②の係り先が空のまま慣用されているとみなす．この場合，①，②の係りを完結させる [[[<(暑さ)(寒さ)>も] [[彼岸迄]続く]]だろう]といった用法や①，②を未完のままフレーズ化した[[[[<(暑さ)(寒さ)>も] [[彼岸迄]$]]の]季節]などの用法が有り得る．

一般に，日本語のよく使われる（右開放型）不完全句には，

1. 連用修飾句＋連用修飾句
2. 連用修飾句＋名詞句

の形式がある[1]．いずれも文頭側の連用修飾句の係り先（修飾先）が欠落している形式である．上例の「(暑/アツ)-さ-寒-さ-も」と「彼岸-迄」は，いずれも連用修飾句であ

[1] 不完全句には右開放型のほか，左開放型，俳諧型がある．

り，1. のパタンである．2. のパタンには，例えば，図 5 の「鬼-の-居-ぬ-間-に-洗濯」などがある．

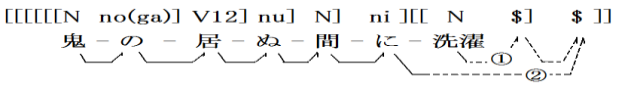

図 5 「鬼の居ぬ間に洗濯」の構文構造

2.6.4 内部修飾可能性

構造記述内に記されたアスタリスク「*」は内部修飾可能性，言い換えると後接句の独立性を示すもので，アスタリスクの直後の句（の主辞）が内部修飾を受ける可能性があることを意味する．例えば，「顔-色-を-(覗/窺)う」の構造記述 [[*[NN]wo]*V30] は，図 6 の様な拡張表現「彼の顔色を彼女はそっと窺う」の可能性を与えている．ここで，V30 は動詞終止形を意味する．

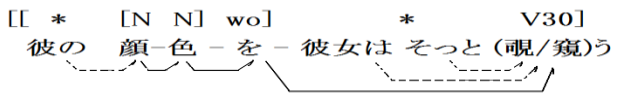

図 6「彼の顔色を彼女はそっと窺う」の構文構造

構造記述にアスタリスクが多いほど，各要素語の独立性が強い表現，すなわち，通常句（自由結合句）に近い表現である．

従来の機械翻訳等では，慣用句を単語として扱うため，柔軟性に対応できない場合が多い．例えば，“仕事をせずに時間をつぶす”意味の慣用句「油を売る」で「油をいつもの店で売る」などと内部修飾によるギャップが生じると，慣用句としての意味は認識できなくなる場合が多いが，本レキシコンでは，構造記述 [[wo]*V30]によって「売る」が連用修飾を受ける場合も有ることが記されており，慣用句としての柔軟性が確保されている[1]．

テキスト上で本レキシコンの見出しと文字面で一致した表現があった場合，F 欄でアスタリスクを付されていない語句に対する内部修飾句が存在すれば，その表現は慣用句ではない．すなわち，アスタリスクが付されていないいずれの句にも内部修飾句が存在しないことが慣用句であるための必要条件である．このことは，慣用句の多義選択に利用できる．例えば，慣用句「(棚/タナ)-に-上げる」の構造は [[Nni]V30] とされており，N，V30 にアスタリスクが付されていないため，「壁の棚に上げる」，「棚に高く上げる」などは慣用句の意味ではないと判定できる．アスタリスクの付された句には内部修飾句が存在してもしなくても慣用句の可能性がある．

[1] ただし，連用修飾句であっても「油-を-5-リッター-売る」は慣用句ではないから，修飾句の意味による多義選択ルールを定めておく必要があるが，これは今後の課題である．

2.7 前方文脈条件（G 欄）

例えば，「(羽目/ハメ)-に-なる」，「(目/眼)-に-(会/遭/逢/遇)う」は，それぞれ，「苦しむ-(羽目/ハメ)-に-なる」，「つらい-(目/眼)-に-(会/遭/逢/遇)う」のように連体修飾句を要求する．G 欄にはこの種の条件が<adnom. modifier>*と記載されている．

この種の条件には，ほかに例えば，以下のような種類があり，それぞれ，「*」の有無によって制約(必須)条件と選好条件とが区別されている．

<adnom. modifier-no> ：「の」による連体修飾句をとる
<noun concatenation> ：名詞が連接する
<adv. modifier-wo> ：ヲ格による連用修飾句をとる
<adv. modifier-ga> ：ガ格による連用修飾句をとる
<adv. modifier-ni> ：ニ格による連用修飾句をとる

2.8 後方文脈条件（H 欄）

例えば，副詞性表現「ひとつ-と-し-て」は後方の否定句と呼応する必要があり，H 欄に後方文脈条件として「ない」を記載している．

この種の条件にも制約(必須)条件と選好条件とがある．

指定される呼応表現や条件には，「か」(疑問)，「ても」，「とも」，「うと」，「なんて」，「とは」，「などと」，「ない」(否定)，<end of sentence>（文末にあること），などがある．

2.9 連体・連用・動詞化情報（I 欄）

収録表現が連体修飾句化，連用修飾句化，動詞化されて用いられる場合が有る．I 欄はその可能性と変化の仕方を三つ組，α-β-γ で与える．ここで，α，β，γ は，それぞれ，連体修飾句化，連用修飾句化，動詞化させる際に使うことのできる後接語句の集合である．例えば，形容動詞的な表現「前-後-不覚」に対して{na,no}-{ni}-{ninaru}と記載し，「前-後-不覚-な」，「前-後-不覚-の」で連体修飾，「前-後-不覚-に」で連用修飾，「前-後-不覚-になる」と動詞化できることを示す．

後接語句としては次のような表現をローマ字で記載している．

・連体修飾： 「たる」，「なる」，「な」，「の」
・連用修飾： 「に」，「と」，「で」，「ε」
・動詞化： 「する」，「をする」，「とする」，「になる」

「ε」は空表現であり，後接語句なしで連用修飾できることを意味する．

典型的な形容動詞は{na}-{ni}-{ninaru}というパタンに対応する．

2.10 備考欄（J 欄, K 欄）

例えば，慣用句「腰-を-据える」に対して副詞的によく使われる「腰-を-据え-て」も慣用句とみなして見出しに収録している．この種の派生表現の J 欄には「派生形」という表示を入れている．

K 欄には，ユーザーの便宜のため，一部の表現に対して語釈を与えている．

3. おわりに

自然言語にはコロケーション，慣用句，準慣用句，決まり文句のような単語境界を越えた長単位の特異表現が数多く使われているが，従来の自然言語処理 NLP では，例外的言語現象とみなされ，十分な対応がなされて来なかった．しかし，今世紀に入り，[12]の指摘によって，ようやくこの種の複単語表現 MWE: Multiword Expression が重視されるようになり，欧米を中心に研究が盛んになっている．いっぽう，言語学においても人間の言語獲得・認知メカニズムへの関心から構文文法 Construction Grammar[3]，定型言語 Formulaic Language[2][4]，語彙連鎖 Lexical Bundles[1]といった枠組みが提唱され，近年，種々の研究がなされるに至っている．

筆者の一人は 1960 年代に始めたフレーズに基づく機械翻訳の研究を通じて，語の共起を語類や意味属性によって規定する方式に加えて，語の共起を表層で規定する大規模レキシコンが不可欠であることを認識し，総括的な日本語複単語表現レキシコン JMWEL の開発を続けてきた．JMWEL の総見出し数は，現在，約 14 万件である．[17][19]

本稿で紹介したレキシコン JMWEL_idiom は，JMWEL における非構成的日本語表現群の中核を担う部分レキシコンである．

本レキシコンが我が国の今後の自然言語処理 NLP 技術，日本語研究の進化の一助となれば幸いである[1].

参考文献

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

[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] Fillmore, C., Kay, P., and O'Connor, M. K, “Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone”, Language, vol.64, No.3, pp.501-538 (1988).

[4] Jiang, N., Nekrasova, T. M., “The processing of formulaic sequences by second language speakers”, The Modern Language Journal, 91:3, pp.433-445 (2007).

[5] 金田一春彦, 金田一秀穂（監修）, “新レインボー小学国語辞典”, 学研 (2005).

[6] 金田一秀穂（監修）, “小学生のまんが慣用句辞典”, 小学館 (2005).

[7] 松村明（編）, “大辞林第 3 版”, 三省堂 (2006).

[8] 松村明（監修）, “大辞泉第 2 版”, 小学館 (2012).

[9] 宮地裕（編）, “慣用句の意味と用法”, 明治書院 (1982).

[10] 水谷静夫, 田中幸子, “語の並列結合子”, 計量国語学, No.63, pp.19-36 (1972).

[11] 旺文社（編）, “成語林-故事ことわざ慣用句”, 旺文社 (1993).

[12] Sag, I. A., et al. “Multiword Expressions: A Pain in the Neck for NLP”. Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics, CICLING2002: pp.1–15 (2002).

[13] 新村出（編）, “広辞苑第 5 版”, 岩波書店 (1998).

[14] 白石大二（編）, “国語慣用句大辞典”, 東京堂出版 (1977).

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

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

[17] 高橋雅仁, 田辺利文, 首藤公昭, “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 (2018).

[18] 竹田晃（編）, “四字熟語・成句辞典”, 講談社 (1990).

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

[20] 山田忠雄, 柴田武, 酒井憲二, 倉持保男, 山田明雄, 上野善道, 井島正博, 笹原宏之（編）, “新明解国語辞典第 7 版”, 三省堂 (2011).

[21] 米川明彦, 大谷伊都子（編）, “日本語慣用句辞典”, 東京堂出版 (2005).

[1] JMWEL の利用に関しては関連サイト http://jefi.info を参照されたい.