Vol.2009-NL-191 No.3
Vol.2009-SLP-76 No.3
2009/5/21

日本語の慣用的表現辞書について

首藤公昭[†]　田邊利文[†]　高橋雅仁[††]

日常の自然言語文には構成性（compositionality）に問題のある相当数の慣用句あるいは慣用句的な複単語表現（Multi-Word Expression；MWE）が使われており，構文・意味解析の大きなネックとなっている．また，強い語の結合によって成り立ち，一括して取り扱うことが処理効率の上で望ましいと思われる常套句や常套句的表現も数多い．筆者らは日本語処理を目的として，これらの日本語 MWE 候補を網羅した辞書の構築を行ってきたが，最近，初版の概要が定まったので，自立語相当表現に限定して報告する．

On a Dictionary of Japanese Multiword Expressions

Kosho Shudo[†]　Toshifumi Tanabe[†]　and　Masahito Takahashi [††]

NLP technology has been suffering from the fact that there used so many non-compositional (idiomatic) and/or probabilistically bound (collocational) multiword expressions in daily documents, however, the clear overall picture of them has not been explained yet.　This paper presents the overview of 89,000-head-line dictionary of Japanese multiword expressions, manually developed to remedy the above problem. Its remarkable feature is the extensiveness of entries, i.e. head lines, their notational variants, syntactical functions, internal structures (trees) and derivative forms.

1.　はじめに

日常英語の機械処理で問題となる複単語表現（Multi-Word Expression; MWE）をその種類とともに考察した文献 1)がきっかけとなって，自然言語処理（Natural Language Processing; NLP）における MWE 処理の重要性が，近年，改めて認識されるようになった．これを受け，(国際) 計算言語学会 (Association for Computational Linguistics; ACL) は 2003 年以降，MWE に関するワークショップをほぼ毎年開催しており，そこでは非構成的（non-compositional）な MWE を統計的に自動評価・抽出する方法等が活発に議論されている．しかし，最近の研究でも Multiword Verb，Multiword Noun，Verb Particle Construction，Verb Noun Construction などの特定の構文構造のみを対象とする研究が多く，それぞれに必ずしも充分な成果が得られているとも言い難い．現状では，いずれの国の NLP においても表現の多様性を十分に踏まえて実際に表現リストを提示したり，処理に利用したという研究は未だ報告されていない様である．

筆者らは日常の自然言語を対象とする NLP のためには人の内省によって問題のある MWE の候補を出来るだけ網羅的に資源化しておくことが不可欠であると考え，古くから日本語 MWE の収集・整理を行ってきた．本稿ではその現状を報告する．

本辞書は，慣用句（イディオム），常套句（決まり文句），連語，コロケーション，成句，語結合，機能動詞結合，支援動詞構文，クランベリー表現，四字熟語，格言，諺，擬態・擬音・擬声語（オノマトペ），強い共起性表現，複合語（一部），呼びかけ表現, 応答表現等などの複合表現を日本語処理を想定して総括的に整理･提示しようとする試案である．

本稿ではこれらの表現を「慣用的表現」あるいは単に「MWE」と総称する．

本辞書の主な特徴は，収録表現の網羅性が比較的高いこと，異表記（表記揺れ）情報，文法機能情報，文法構造情報，派生形情報等が収録されていること，人の内省によって編纂されていることなどである．

2.　関連研究

日本語 MWE に関する研究としては，古くから国語学の領域で人の利用を目的として慣用句辞典等の編纂が種々行われてきた（文献 2)-12)）．しかし，これら個々の研究には表現，表記の多様性や構造，用法の体系的記述が十分ではない場合が多く，NLP 向きとは言い難い．

NLP の立場における日本語 MWE の研究では，機能語（付属語）性 MWE を収集・

† 福岡大学工学部
Fukuoka University, Faculty of Engineering
†† 久留米工業大学
Kurume Institute of Technology

Vol.2009-NL-191 No.3
Vol.2009-SLP-76 No.3
2009/5/21

整理し，単語的に扱う“拡張文節モデル”を提案した文献 13)-14)が比較的古いほうだと思われる．その後の機能語性 MWE の研究には助動詞，終助詞と同様に日本語文末で用いられる MWE の意味体系を考察した文献 15)や，機能語性表現を階層的に整理する方法を提案した文献 16)などがある．

他方，NLP における概念語（自立語）性 MWE の研究には，［名詞＋格助詞＋動詞］型の述語性慣用句を対象として日英機械翻訳を考察した文献 17)や, 約 20,000 個の NLP 用日本語 MWE を収集・整理・公開した文献 18)があり，最近では市販の数種の慣用句辞典から約 3,400 個の慣用句を収集して考察を加えた文献 19)がある．

また，機能語性，概念語性を合わせた約 72,000 個の MWE を単語の共起情報として用いることで仮名漢字変換の正解率向上を試みた研究に文献 20)がある．

しかし，これまでの NLP における MWE の研究では，未だ表現，表記の多様性や機能, 構造記述等の点が不十分と言わざるを得ない. 本研究は，これらの問題を軽減し，将来の日本語処理の高度化に資すべく，文献 18)，20)の概念語性表現データを抜本的に修正・拡張して再提示するものである．

3. 採録した表現

筆者らは，雑誌記事，新聞記事，小説，随筆，事典・辞書類など，広範な文書から次の様な概念語性 MWE を収集・整理してきた．

3.1 慣用句（イディオム）性の表現

要素単語から全体の意味を規則で導くことが難しい，即ち慣用句（イディオム）性（non-compositionality）があると思われる表現，例えば，「赤−の−他人」,「耳−を−貸さ−ない」,「手−を−抜く」,「足−が−出る」,「首−が−回ら−ない」,「顔−を−売る」,「気−を−取(り)−直し−て」,「気−が−利く」等々を採録した．また通常，慣用句とは呼ばれないが，機械処理において構成性（compositionality）に問題が生じる可能性のある表現も出来るだけ網羅することを心がけた. この意味で支援動詞構文(SVC), 一部の複合語，派生語が含まれている．例えば,「一票−を−投じる」,「批判−を−加える」,「(磨/研)き−を−(掛/懸)ける」,「伝票−を−切る」,「計画−を−立てる」,「辞書−を−(引/曳/牽)く」,「(バカ/馬鹿/莫迦)−を−(言/云)う」,「練り−歩く」,「打(ち)−拉が−れる」,「積(み)−立てる」,「顔−を−する」,「ウロウロ−する」,「大学−を−出る」,「要求−を−(飲/呑)む」等々である．一般に，この種の表現は，纏まった文法･意味上の機能を持つ句（あるいはその様な句の列）であって，いずれかの要素単語を同意語あるいは下位概念の語（列）で置き換えたとき，意味をなさなく（不自然に）なるか，類似の（下位の）意味にならないという性質を持つ．例えば,「真紅−の−他人」,「耳−を−貸与−し−ない」,「手−を−引−き−抜く」, … ,「一票−を−投げる」,「批判−を−足す」, … ,「要求−を−飲用−する」などは，少なくとも慣用句の意味を保存していない．採否の判断にはこの性質も援用した．

3.2 単語間共起確率の高い表現

纏まった文法･意味上の機能を持つ句（あるいはそのような句の列）$w_1w_2w_3…w_n$で，いずれかの要素単語w_iについて，条件付確率$p_f(w_i|w_1…w_{i-1})$あるいは$p_b(w_i|w_{i+1}…w_n)$が相対的に高く，$w_1…w_{i-1}$に続く単語のエントロピー$H_f(W|w_1…w_{i-1})$あるいは$w_{i+1}…w_n$の前単語エントロピー$H_b(W|w_{i+1}…w_n)$が相対的に低いと思われる表現，例えば,「警鐘−を−鳴らす」,「手−を−こまぬく」,「腰−を−抜かす−程−驚く」,「故郷−を−(思/想)う」「故郷−を−出る」などを収録した．p_f(鳴らす|警鐘−を)，p_b(手|を−こまぬく)，p_f(眠る|グッスリ)などは大きく，H_f(W|警鐘−を)，H_b(W|を−こまぬく)，H_f(W|グッスリ)は小さいと判断できる．語の共起性が強く，NLPに有効なMWEの一種はこの種の表現であろうという仮定に基づいている．オノマトペとその派生形についても動詞との共起を出来るだけ網羅的にデータ化した．例えば,「ユルユル−と−動く」,「グラグラ−揺れる」,「グッスリ−眠る」,「クルクル−回る」,「ポッカリ−と−空く」などである．

3.3 常套句（決まり文句）的な表現

多くの場合，**3.1**，**3.2** と重複するが，例えば,「風前−の−灯」,「付きっ−切り」,「矢−継(ぎ)−早」,「禍−転じ−て−福−と−なす」,「雲−一つ−無い」,「時−は−金−なり」,「其れ−は−然う−と」,「オット−ドッコイ」,「程度−の−差−こそ−有れ」,「(眼/目)−に−も−(止/留)まら−ぬ−早−(技/業)」,「右−肩−上(が)り−に」,「(言/云/謂)わ−ず−も−がな」等々の様に，一体性の強い表現（単語境界の位置に他の単語列が現れることが殆ど無いと思われる表現）も意識して収集されている．

A	B	C	D	E	F	G	H	I
いまだかって	いまだ−かって	未だ−(嘗/曾)(っ)て	D		DD			否定
いまだかつて	いまだ−かつて	未だ−(嘗/曾)(つ)て	D		DD			否定
いまだかってない	いまだ−かって−ない	未だ−(嘗/曾)(っ)て−無い	Ya	aeb	nai			
いまだかつてない	いまだ−かつて−ない	未だ−(嘗/曾)(つ)て−無い	Ya	aeb	nai			
いまだしのかん	いまだし−の−かん	未だし−の−感	Mk		KnoM	No-De		
いまだしのかんあり	いまだし−の−かん−あり	未だし−の−感−(有/在)り	Yk	vb20	V'	X-De		
いまだしのかんがある	いまだし−の−かん−が.ある	未だし−の−感−が−(有/在)る	Yv	vb2	aru			
いまだしのかんのある	いまだし−の−かん−の.ある	未だし−の−感−の−(有/在)る	Tv	vb25	aru			
いまだに	いまだ−に	未だ−に	D		Dni			否定
いまだもって	いまだ−もって	未だ−以て	D		DD			否定

図 1　データの形式

Vol.2009-NL-191 No.3
Vol.2009-SLP-76 No.3
2009/5/21

以上の 3 種の性質を兼備する表現は非常に多く，境界は必ずしも明確ではない．
辞書ファイルは 9 個の欄（A 欄～I 欄）からなる図 1 の形式をとる．

3.4 表現の長さ

本辞書における表現のグラム数と収録数の関係を図 2 に示す[a]．

表現の長さ（グラム数）	収録表現の割合(%)
1	2.44
2	18.26
3	41.17
4	23.39
5	8.86
6	3.29
7	1.58
8	0.60
9	0.23
10	0.10
11	0.03
12	0.03
13	0.01

図 2　表現の長さと集録表現数の関係

4. 収録した情報

4.1 平仮名ベタ見出し（A 欄）

見出しは平仮名（音）表記に基づいている．例えば，「良い」は「よい」と「いい」に，「得る」は「える」，「うる」に，「言う」は「いう」，「ゆう」に適宜読み分けて別見出しとする．また，「もーてぃべーしょん」，「もてぃべいしょん」，「もちべーしょん」なども別見出しとする．
見出し総数は，現在約 89,000 件である．

a) 14 グラム以上は少数なので省略する．1 グラム表現には **4.7** の派生形によってMWEが与えられている．

4.2 構成単語間の境界（B 欄）

ハイフン「-」およびドット「.」で語境界を示す．ドットはこの位置で別の単語列（例えば副詞）が使われる可能性を示す．従って，ドットが記されていない「細大-漏らさ-ず」，「尻-切れ-トンボ」などが一体性の強い表現である．格助詞「に」，「の」との機能・用法上の類似性から，形容動詞の連用形語尾「に」，「と」，と同連体形語尾「な」，「たる」，「なる」だけは分離している．漢字列表現の分割には紛らわしい場合が多いが，ここでは表記の多様性を簡潔に表現することを重視した．そのため，字種が変化する可能性のある所に区切りを入れた．例えば，「ごくろうさま」は B 欄で「ご-くろう-さま」と区切り，C 欄の漢字情報「御-苦労-様」から異表記「御苦労さま」，「ご苦労さま」，「御苦労様」，「ご苦労様」が生成できるようにした．逆に，「営業車」では「車」＝「しゃ」には接尾語性があるとも考えられるが，表記が一体的と考えられるため，切り離していない．

4.3 字種，表記の揺れ情報（C 欄）

字種と表記の揺れ情報を同時に与える．例えば，「組(み)-付ける」などの括弧は文字の任意性，「(良/好/善)い」などの括弧と斜線の組み合わせは文字の選択肢を与える．B 欄，C 欄を合わせることで，殆ど全ての異表記に対応できる．例えば，B 欄「き-の-いい-やつ」，C 欄「気-の-(良/好/善)い-(奴/ヤツ)」から，次の 24 種の表記が得られる．
「きのいいやつ」，「きのいい奴」，「きのいいヤツ」，「きの良いやつ」，「きの良い奴」，「きの良いヤツ」，「きの好いやつ」，「きの好い奴」，「きの好いヤツ」，「きの善いやつ」，「きの善い奴」，「きの善いヤツ」，「気のいいやつ」，「気のいい奴」，「気のいいヤツ」，「気の良いやつ」，「気の良い奴」，「気の良いヤツ」，「気の好いやつ」，「気の好い奴」，「気の好いヤツ」，「気の善いやつ」，「気の善い奴」，「気の善いヤツ」

4.4 文法的な機能と種別（D 欄）

表現全体の文法的な機能を以下の様に記号化して記載する．（括弧内に見出しの概数を示す．）

C:　接続詞性表現，(1,000)
D:　副詞性(連用修飾)表現，(7,000)
T:　連体詞性(連体修飾)表現，(6,200)
M:　名詞性表現，(7,900)
Ms:　サ変名詞性表現，(500)
Md:　サ変以外の動的名詞性表現，(2,200)
Mk:　形容動詞的名詞性表現，(4,600)
Yv:　動詞性表現，(48,500)
Ya:　形容詞性表現，(4,500)
Yk:　形容動詞，準形容動詞性表現，(3,300)

Vol.2009-NL-191 No.3
Vol.2009-SLP-76 No.3
2009/5/21

Yo:　擬態・擬音・擬声表現，(600)
また，意味・語用論的機能の種別として，
-P:　格言，諺，(2,300)
-Self:自問，独り言表現，(200)
-Call:呼びかけ表現，(150)
-Grt:　挨拶表現，(200)
-Res:　応答表現，(200)
などを記載する．これらは文解析に不可欠な情報である．

4.5 述語への係り構造（E 欄）

表現に述語が含まれる場合，その表層格パタン等の修飾構造，約 80 種を以下の va1 のようにコード化して与える．

[[N+p]+P] 型
名詞＋「を」＋動詞　: va1 ex. 「異を唱える」
名詞＋「が」＋動詞　: va2 ex. 「異臭がする」
名詞＋「が」＋形容詞　: aa2 ex. 「歴史が浅い」
名詞＋「が」＋形容動詞　: ka2 ex. 「霊験があらたか」
名詞＋「に」＋動詞　: va3 ex. 「数に入れる」
名詞＋「に」＋形容詞　: aa3 ex. 「児戯に等しい」
名詞＋「に」＋形容動詞　: ka3 ex. 「基本に忠実」
名詞＋「で」＋動詞　: va4 ex. 「論理で押す」
名詞＋「で」＋形容詞　: aa4 ex. 「それで良い」
名詞＋「から」＋動詞　: va5 ex. 「不況から脱出する」
名詞＋「から」＋形容詞　: aa5 ex. 「理想からほど遠い」
⋮

[[[[N+p]+N]+p]+P] 型
名詞＋「の」＋名詞＋「を」＋動詞　: vb1 ex. 「尊敬の念を抱く」
名詞＋「の」＋名詞＋「が」＋動詞　: vb2 ex. 「化けの皮が剥げる」
名詞＋「の」＋名詞＋「が」＋形容詞　: ab2 ex. 「肩の荷が重い」
名詞＋「の」＋名詞＋「が」＋形容動詞　: kb2 ex. 「頭の中が真っ白」
名詞＋「の」＋名詞＋「に」＋動詞　: vb3 ex. 「玉の輿に乗る」
⋮

[[[P+N]+p]+P] 型
用言連体形＋名詞＋「を」＋動詞　: vc1 ex. 「危ない橋を渡る」
用言連体形＋名詞＋「が」＋動詞　: vc2 ex. 「見る目が変わる」
用言連体形＋名詞＋「が」＋形容詞　: ac2 ex. 「立つ瀬が無い」
⋮

[[N+p]+[[N+p]+P]] 型
名詞＋「も」＋名詞＋「も」＋動詞　: vd1 ex. 「性も根も尽きる」
名詞＋「も」＋名詞＋「も」＋形容詞　: ad1 ex. 「根も葉も無い」
名詞＋「に」＋名詞＋「を」＋動詞　: vd2 ex. 「死中に活を求める」
名詞＋「に」＋名詞＋「が」＋形容詞　: ad4 ex. 「枚挙に暇が無い」
⋮

[[P+p]+P] 型
用言連用形＋「て」(「で」) ＋動詞　: ve1 ex. 「切って落とす」
用言仮定形＋「ば」＋動詞　: ve2 ex. 「打てば響く」
⋮

[P+P] 型
用言連用形＋動詞　: ve3 ex. 「巧く行く」
⋮

[[[[N+p]+V]+p]+V] 型
名詞＋「を」＋用言連用形＋「て」(「で」)
＋　動詞　: ve5 ex. 「尻尾を巻いて逃げる」
名詞＋「に」＋用言連用形＋「て」(「で」)
＋　動詞　: ve7 ex. 「額に汗して稼ぐ」
⋮

[A+[[N+p]+P]] 型
副詞＋名詞＋「を」＋動詞　: vec ex. 「どっかと腰を据える」
副詞＋名詞＋「が」＋動詞　: vee ex. 「どっと疲れが出る」
⋮

4.6 付加的構造情報（F 欄）

表現に用言とその係り構造が含まれる場合，この欄には一般形で $(\alpha-)^*\beta^*$ と正規表現される英字列を記載している．α はE欄に補うべき係り要素がある時にこれを表す．β は述部が複合動詞であったり，助動詞，助詞等を含んでいること，あるいは連

Vol.2009-NL-191 No.3
Vol.2009-SLP-76 No.3
2009/5/21

体修飾をしていることなどの情報を与える.述部が単一の用言の場合はβは空とする.例えば,「先−に−述べ−た−様−に」では,「先−に−述べる」に対応した係りの構造va3がE欄に与えられるが,「述べ−た−様−に」の構造記述をF欄でVtayouniと与える.

表現が用言とその係り構造を有しない場合,品詞列レベルの構造記述を F 欄に与える.機能語や機能語相当表現は小文字のローマ字表記とする.例えば,「酒−は−百薬−の−長」には MhaMnoM と記す.品詞記号は次の通りである.

M:名詞
V:動詞
K:形容動詞
D:副詞
T:連体詞
P:接辞

活用語で,連用形,終止形,命令形を特に明記すべき場合は,それぞれ,「V'」,「V!」,「V"」と記す.最末尾表現の活用が表現全体の活用であると考えられる.

D,E,F 欄を総合し,若干の処理を施せば,表現全体の大まかな木構造が求められる.この意味で本辞書はコーパスに独立な一種の(選択的)ツリーバンクと見なすこともできる.

4.7 派生形(G 欄)

形容動詞性(様態)表現(D,Mk,P,Yk,Yo)の場合,

＜連体修飾形＞{−＜連用修飾形＞{−＜動詞形＞}}

の形式で派生形を与える.

例えば,「(我/吾)−関せ−ず」という表現では,「(我/吾)−関せ−ず−の」,「(我/吾)−関せ−ず−と−(言/云/謂)う」,「(我/吾)−関せ−ず−と−(言/云/謂)っ−た」で連体修飾,「(我/吾)−関せ−ず−と」,「(我/吾)−関せ−ず−で」と連用修飾句が派生することを NoToiuToitta−ToDe と記す.同様に,擬態語「フラフラ」には,「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」と動詞化することを NoSitaTosita−EToSiteTosite−SuruTosuru と記す.また,同じ擬態語でも「グングン」では連用句としての「グングン」,「グングン−と」以外の派生は不自然なので,G 欄は X−ToE と記す.(E は空列,X は派生ナシの意.)この様に,これらの派生パタンは多彩で,約 300 種にのぼる.X−ToE などのコードは様態表現の細密化した品詞の表記と考えることができる.この種の派生形を別見出しとすれば,見出し数は約 110,000 件となると推定される.

4.8 文頭側条件(H 欄)

表現が存立するための制約的な条件として文頭側コンテクストを与える.例えば,「割れ−に−なる」は単独では用いられず,「元本−割れ−に−なる」のように,文頭側に名詞による連体修飾語が必要であることを＜名詞連接＞と記す,などである.

4.9 文末側条件(I 欄)

H 欄と同様,文末側コンテクストを与える.例えば,「如何−と−も」は文末側に「〜難い」などの困難性を表す表現を必要とすることなどである.

5. 考察

収録表現群の性質の一端を探るため,文献 21)のGoogle Nグラムデータ(以降G−Nグラムと略記する.)との照合を試みた.対象とした表現は動詞性表現Yvのうち,[名詞＋格助詞＋動詞]型で,格助詞を「を」,「が」,「に」に限定したものとし,動詞部は単独の動詞,2 動詞からなる複合動詞,[サ変名詞＋する]型動詞(終止形)に限定した.これらの見出し数は 29,389 個であり,B, C欄の情報で展開した対象表記数は 82,125 個である.これらのうち,[名詞＋格助詞]部分の表記数は 13,806 個で,その内 12,120 個がG−Nグラムにおける 2, 3 グラムデータに一致した.これらの表記を前部分列とするG−Nグラムの 3,4,5 グラムデータの中から,格助詞の直後に動詞(終止形)が出現するもの 1,194,293 個に着目し,前部分列ごとに,各動詞の出現頻度を求めた[b].

その結果,辞書データの動詞がG−Nグラムで出現頻度第 1 位である場合が 5,787 件であり,対象とした前部分列表記w_1w_2の 47.7%=(5,787/12,120)*100 に対して **3.2** で述べた$p_f(w_3|w_1w_2)$が最大の動詞部w_3が選ばれていると推定できた.「ちょっかい−を−出す」,「熟戦−を−繰り−広げる」,「アクション−を−起こす」などはこれらに該当する.同様に,第 2 位の場合は 1,699 件で 14.02%, 3 位は 877 件で 7.24%, 4 位は 482 件で 3.98%,等々であった.20 位までの結果をグラフ化して図 3(a)に示す.収録表現は高い条件付き確率のものほど多いという図 3(a)が示す傾向は至当なものと思われる.

図 3(a)を累積の比率に改めたグラフを図 3(b)に示す.これから,例えば,本辞書では,対象とする前部分列の約 80%に頻度 8 位までの動詞が,約 86%に 20 位までの動詞が選ばれていることなどが分る.G−Nグラムデータでは高い頻度順位であるのに,本辞書で選ばれていない動詞が相当多いが,これらは格助詞に続く動詞のエントロピーが大きく,絞り込みが難しい場合であると考えられる[c].また,図 3(b)の外挿によれば,前部分列の 10%強に対して,後接する動詞がG−Nグラムでは同環境に現れていないと推定できる.例えば,本辞書に在る「才知−に−長ける」,「轢き−逃げ−を−働く」はG−Nグラムに存在しない.このことは,200 億文もの大規模Webコーパスに基づくG−Nグラムであっても,かなりの表現が捕捉出来ていない可能性を示唆している.Zipfの法則におけるロングテール部に対する内省による表現収集の重要性を示すものと考え

b) G-Nグラムデータの動詞性の判定には文献 22)のIPADIC動詞辞書(verb.dic)およびサ変名詞辞書(noun.verbal.dic)を用いた.
c) 現在,G-Nグラムデータ上で実際にエントロピーを計算するに至っていない.

Vol.2009-NL-191 No.3
Vol.2009-SLP-76 No.3
2009/5/21

られる．

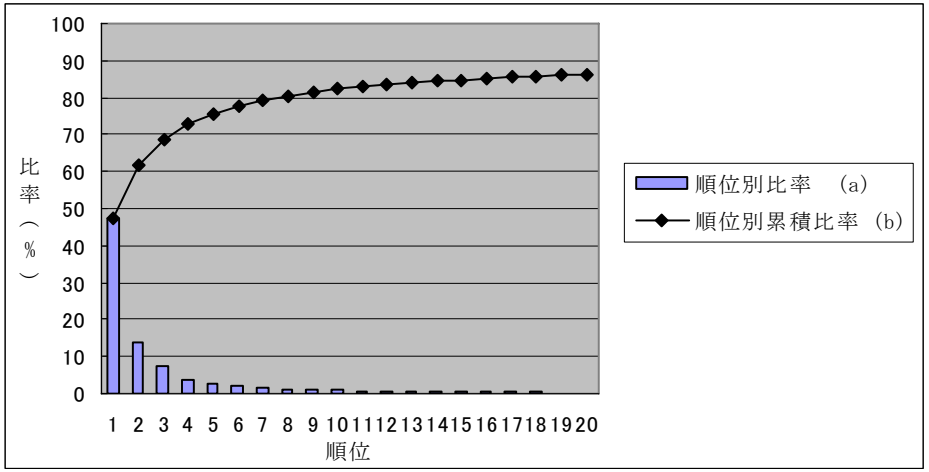

図 3 ［名詞+格助詞+動詞］型表現の Google N グラムによる動詞の出現頻度順位別比率(a)と順位別累積比率(b) −（格助詞は「を」,「が」,「に」に限定）

上記 1,194,293 表現の出現頻度の合計は 1,389,568,825 であるのに対し，本辞書データ 82,125 個の出現頻度の合計は 374,718,334 であり，本辞書の表現は G−N グラムの出現数の約 27%をカバーしている．いっぽう，動詞のバリエーションは G−N グラムで平均 98.5=1,194,293/12,120 個であるのに対し，本辞書データでは平均 5.95=82,125/13,806 個にすぎない．従って，約 1/17 の動詞の種類で G−N グラムにおける出現数の 1/4 以上をカバーしていることが分る．

その他，上記の型に限定しない本辞書データが一般の新聞紙上でどの程度使われているかの人手による検証も随時行ってきた．一例を示せば，2009 年 4 月 21 日の日本経済新聞朝刊第 1 面と最終面に掲載された 264 文中に，本辞書の表現は 219 箇所に出現していた．平均 100 文当たり 83 箇所に使われていることになる．このように，日常の文書ではイディオム性あるいは強い共起性を持つ比較的少数の MWE が相当高頻度で用いられていることが推定される．

イディオム性データの妥当性は本辞書を利用するシステムの意味構成ルールが明確でない時点で検証することは難しいが，筆者らは，本データは少なくとも市販の慣用句辞典等に収録され，日常現代語の文書に出現するものは，ほぼ網羅出来ており，さらに，弱いイディオム性の表現群もかなり収録されていると考えている．

6. おわりに

本辞書の表現の見出し数は現在，約 89,000 であるが，G 欄の派生形を加えれば約 110,000 表現，さらに B，C 欄からの異表記を加えれば，770,000 表現程度を内包している．

本辞書が想定している基本的な利用領域は日本語の構文・意味解析であるが，応用タスクとしては，

１．フレーズ・ベース訳出を行う機械翻訳，音声翻訳システム
２．予測機能を整備した仮名漢字変換システム
３．言語モデルを整備した音声認識システム
４．日本語読み上げ，仮名振りシステム
５．日本語教育システム
６．難しい表現を易しく言い換えるなどの言い換えシステム

等が考えられる．

当面の課題としては，例えば次の点が挙げられる．

ⅰ．表現のカバレジの詳細な検証．
ⅱ．意味上の多義性の有無情報の付与．
ⅲ．「です」,「ます」調表現等の充実．
ⅳ．標準的な表現への言い換え情報（含，decomposability 情報）の付与．
ⅴ．詳細な変化形情報の付与（文献 23)）．
ⅵ．条件付き確率，条件付きエントロピー推定値の付与．
ⅶ．異表記間の優先度情報の付与．
ⅷ．古語，現代語の区別情報の付与．

本辞書は，日本語の日常使用者が持っている言語モデルの一端を，一言でいえば「語の慣用」という観点から提示する試みであり，叩き台である．未だ不備な点が多く，今後の改良，補強等が不可欠と思われるが，そのためにも多方面での利用とフィードバック情報が期待される．

謝辞 本研究に至るきっかけを頂いた元九州大学教授，故栗原俊彦氏，データの収集に御協力頂いた江崎斗志子氏，武内美津乃氏，高丘満佐子氏をはじめとする多くの方々，貴重な助言や励ましを頂いた元九州芸術工科大学長，故吉田将氏，元言語処理学会会長，現 JAIST 教授，島津明氏，本研究の方向付けの段階でお世話になった元京

Vol.2009-NL-191 No.3
Vol.2009-SLP-76 No.3
2009/5/21

都大学総長，現国立国会図書館館長，長尾真氏に深甚の謝意を表します.

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) 新村出編，広辞苑 第 6 版，岩波書店 (2008).
3) 松村明監修，大辞泉，小学館 (1998).
4) 松村明編，大辞林 第 3 版，三省堂 (2006).
5) 尾上兼英監修，成語林−故事ことわざ慣用句，旺文社 (1993).
6) 三省堂編修所編，故事ことわざ慣用句辞典，三省堂 (1999).
7) 白石大二編，擬声語擬態語慣用句辞典，東京堂出版 (1992).
8) 竹田晃，四字熟語・成句辞典，講談社 (1990).
9) 田島諸介，ことわざ故事・成語慣用句辞典，梧桐書院 (2002).
10) 米川明彦，大谷伊都子編，日本語慣用句辞典，東京堂出版 (2005).
11) 藤田保幸，山崎誠編，複合辞研究の現在，和泉書院 (2006).
12) グループ・ジャマシイ編，日本語文型辞典，くろしお出版 (2007).
13) 首藤公昭，楢原斗志子，吉田将，日本語の機械処理のための文節構造モデル，電子通信学会論文誌，62-D-12 (1979).
14) 首藤公昭，文節構造モデルによる日本語の機械処理に関する研究，福岡大学研所報, 45 (1980).
15) K. Shudo, T. Tanabe, M. Takahashi, K. Yoshimura, MWEs as Non-propositional Content Indicators, Proc. of the 2nd ACL Workshop on MWE (2004).
16) 松吉俊，佐藤理史，宇津呂武仁，日本語機能表現辞書の編纂，自然言語処理,14-5 (2007).
17) 奥雅博, 日本語慣用表現の分析と日英翻訳への適用, 情報処理学会研究報告, 87-NL-62 (1987).
18) 首藤公昭，日本語における固定的複合表現，昭和 63 年度文部省科学研究費特定研究（Ｉ）「情報ドクメンテーションのための言語の研究」報告書，(1989).
19) 佐藤理史，基本慣用句五種対照表の作成，情報処理学会研究報告，07-NL-178 (2007).
20) 小山泰男，安武満佐子，吉村賢治，首藤公昭，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 (1998).
21) 工藤拓，賀沢秀人，Web 日本語 N グラム第 1 版，言語資源協会 (2007).
22) 浅原正幸，松本祐治，ipadic version 2.7.0 ユーザーズマニュアル，奈良先端科学技術大学院大学 情報科学研究科 (2003).
23) 安武満佐子，小山泰男，吉村賢治，首藤公昭，固定的共起表現とその変化形，言語処理学会第 3 回年次大会発表論文集, (1997).