Vol.2012-IFAT-105 No.4
Vol.2012-NL-205 No.4
2012/1/20

日本語表現辞書 JDMWE の統計的性質

田辺利文[†]　高橋雅仁[††]　首藤公昭[†]

本論文では，日本語複単語表現辞書 JDMWE の概要，および JDMWE が収録している表現の統計的性質の一端を Google 共起頻度データ LDC2009T08 との比較により明らかにする．

Statistical Properties of the JDMWE: Japanese Dictionary of Multiword Expressions

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

This paper presents an overview of the comprehensive dictionary (JDMWE) of Japanese multiword expressions, and we illustrate some important statistical properties of the JDMWE by comparing the dictionary with a large-scale Japanese N-gram frequency dataset, the LDC2009T08, generated by Google Inc.

1. はじめに

近年，Web に代表される文書情報の爆発的な増加に伴い，より精度の高い自然言語処理の必要性が認識されている．特に，意味を考慮した処理システムの需要が増加していることは Semantic Web 研究等の発展を見ても明らかである．筆者らは，意味を考慮した言語データ処理においては，複単語表現（Multiword Expression: MWE）を組み込んだシステム構築が不可欠であると考えている．MWE の研究動向は，2002 年の論文「Multiword Expressions: A Pain in the Neck for NLP」[2]を皮切りに（国際）計算言語学会（Association for Computational Linguistics: ACL）を中心に MWE workshop がほぼ毎年開催されるなど，世界的に重要性が認識されてきた．MWE は頻繁に自然言語に現れることも報告されており，Sag らは，WordNet1.7 での見出しの約 41%が MWE であること [2]，また田辺らは，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 [5]などから，MWE の適切な取り扱いが自然言語処理の質を向上させる上で必要不可欠であることを示している．

これまで首藤らは，非構成(イディオム)性，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 として収集してきた [4]．ここでの非構成性とは，概略，表現を構成している単語の通常の意味から表現全体の意味を構成するのが難しい性質，要素語間の強い共起性とは，要素単語相互の確率的な縛りが強い性質を意味する．日本語では，自立語性 MWE として「油を売る」「ぐっすり眠る」「手をこまぬく」などが，機能語性 MWE として，「によって」「に関して」「なければならない」「かもしれない」などがある．首藤らは，主として人の内省に基づき収集した約 104,000 表現の日本語の自立語性 MWE を JDMWE(Japanese Dictionary of Multiword Expressions) として編纂しており [4]，JDMWE の特徴として(1)一般的なイディオムや決まり文句などに限定しないこと，(2)多様な構文構造をカバーすること，(3)異表記や派生形も網羅すること，(4)構文的柔軟性を表現できる構文構造情報をもつこと，などが挙げられる．

本論文では JDMWE の概要，および JDMWE が収録している表現の統計的性質の一端を Google 共起頻度データ LDC2009T08 との比較により明らかにする．

2. 採録表現

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，語の共起に何らかの特異性が認められ，構文・意味・談話上の一定の働きを持つ MWE を，主として編者の内省によって収集・整理した．共起の特異性は，基本的なものとして次

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

Vol.2012-IFAT-105 No.4
Vol.2012-NL-205 No.4
2012/1/20

にあげる(1)非構成(イディオム)性，および，(2)要素語間の強い共起性，の 2 種に注目した．

2.1 非構成性 MWE

要素単語の標準的な機能から表現全体の構造・意味を規則で導くことが難しい，いいかえると形態・構文・意味上の非構成性(non-compositionality)を持つ表現，あるいは構成性は成立しているが適用すると過生成(overgeneration)をもたらすと思われる表現を非構成性 MWE として収録した．形式的には，単語列 $w_1w_2\cdots w_i\cdots w_n$ $(2\leq n\leq 18)$そのものがまとまった構文・意味・談話において一定の機能をもっており，かつ，$w_1w_2\cdots w_i\cdots w_n$ におけるいずれかの単語 w_i を，w_i の同意語または類義語である w'_i に置き換えて $w_1w_2\cdots w'_i\cdots w_n$ としたとき，意味をなさなくなるか，または，全く異なる意味になるようなとき，単語列 $w_1w_2\cdots w_i\cdots w_n$ は非構成性 MWE であると定義できる[a]．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，それぞれ非構成性 MWE であるといえる．その表現が非構成性 MWE であるかどうかは内省によって判断している．非構成性 MWE は表 1 に示すような種類に分けることができる[b]．

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, 気が利く
形態･構文上での構成性が不備, あるいは不明瞭な表現[c]	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	支離滅裂, 一心不乱
慣用的な比喩表現	命の限り, 血の雨が降る
その他意味の構成性に問題があると思われる表現	扇風機を回す, 頭が良い

a) JDMWE に収録されている表現は 18 グラム（「天は人の上に人を創らず人の下に人を創らず」）が最長である．
b) 必ずしもこれらの種類は互いに排他的ではない．
c) クランベリー表現もこのカテゴリに入る．クランベリー表現(cranberry expression)とは，例えば，「cranberry」の「cran」,「おだをあげる」の「おだ」のような不明語(クランベリー語)を含む表現をさす．

2.2 単語間共起性の強い MWE

表現の中には，その表現を構成する単語間で共起性が強い性質をもつものがある．このような語の共起性の強い表現は，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．このような表現を単語間共起性の強い MWE として収録した．形式的には，単語列 $w_1w_2\cdots w_i\cdots w_n$ $(2\leq n\leq 18)$そのものがまとまった構文・意味・談話において一定の機能をもっており，かつ，$w_1w_2\cdots w_i\cdots w_n$ におけるいずれかの単語 w_i について，条件付後方出現確率 $p_f(w_i|w_1\cdots w_{i-1})$あるいは条件付前方出現確率 $p_b(w_i|w_{i+1}\cdots w_n)$が相対的に高いという確率的な特異性(probabilistic idiosyncrasy)を持つと思われるとき，単語列 $w_1w_2\cdots w_i\cdots w_n$ は単語間共起性の強い MWE であると定義できる．例えば，「手を拱く」の p_b(手|を拱く),「ぐっすり眠る」の p_f(眠る|ぐっすり)などは比較的大きいと判断し，それぞれ単語間共起性の強い MWE であるとする．その表現が単語間共起性の強い MWE であるかどうかは内省によって判断している．単語間共起性の強い MWE は表 2 に示すような種類に分けることができる[d].

2.3 JDMWE に収録されている MWE の内訳

JDMWE に収録された MWE は，非構成性 MWE と単語間共起性の強い MWE とに大別できるが, 非構成性と強い単語間共起性をあわせ持つ MWE も存在する. そこで，JDMWE に収録された見出しがどちらの性質を持っているかのおおよその分布を調査した結果を図 1 に示す．その結果，JDMWE の見出しの約 38%が非構成性を持ち，見出しの約 92%が強い単語間共起性を持つものであることが推測された．JDMWE に収録された MWE は，非構成性または強い単語間共起性の少なくとも一方を持つものであることを考えると，非構成性と強い単語間共起性をあわせ持つ MWE の割合は約 30%であること，非構成性のみを持つ MWE の割合は約 8%であること，強い単語間共起性のみを持つ MWE の割合は約 62%であることが推測される．

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, 願ったり叶ったり, 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, 疑惑を生む, 女流作家, 機械翻訳

d) 必ずしもこれらの種類は互いに排他的ではない．

Vol.2012-IFAT-105 No.4
Vol.2012-NL-205 No.4
2012/1/20

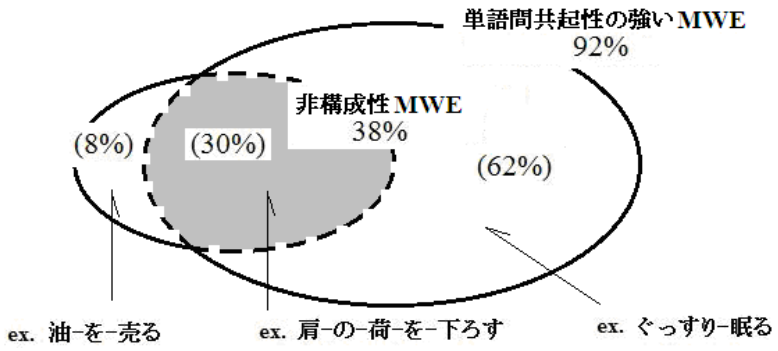

図 1 JDMWE に収録された MWE の分布

3. JDMWE の統計的性質

3.1 概要

表現の非構成性の程度を一般的に算出する基準は存在しないため，非構成的 MWE に関しては，JDMWE に収録したことの妥当性を統計的に評価することは困難である．一方，単語間共起性の強い MWE に関しては，統計的性質として条件付出現確率やエントロピーなどの値を算出することで客観的に評価できる．条件付出現確率やエントロピーを算出するためのリソースとしてサイズの十分に大きいコーパスを用いることが必要である．本研究では，日本語 Web N グラムである Google 共起頻度データ LDC2009T08[3]（これ以降「GoogleN グラム」と略記する）をリソースとして用いることにし，JDMWE との照合を試みた．GoogleN グラムは 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大規模データである．

対象とした表現は[名詞 w_1 + 格助詞 w_2 + 動詞 w_3] 型の動詞性表現で，格助詞 w_2 を「を」,「が」,「に」に，動詞部 w_3 を単独の動詞，[動詞+動詞]型複合動詞，あるいは[サ変名詞+する] 型動詞のそれぞれ終止形に限定した．GoogleN グラムにおいても上記と同じタイプの動詞性表現[名詞 w_1 + 格助詞 w_2 + 動詞 w_3]を抽出する．そのため，[名詞 w_1 + 格助詞 w_2]部分の表記を前部分列(2 グラム)とする 3,4 グラムデータのうち，名詞，動詞部の品詞制約をも満たしたものを，GoogleN グラムにおける動詞性表現とみなしてこれを用いることにした[e]．

e) GoogleN グラム上の品詞制約には浅原ら[1]の IPADIC 名詞辞書(noun.dic)，動詞辞書（verb.dic）およびサ変名詞辞書（noun.verbal.dic）を用いた．また，動詞部 w_3 は単独の動詞の終止形だけでなく，[動詞+動詞]型複合動詞，[サ変名詞+する]型動詞のそれぞれ終止形も含むため，w_3 は 1 グラムだけでなく 2 グラムの場合も含む．

3.2 動詞性表現と前部分列のバリエーション

GoogleN グラムにおける[名詞 w_1 +格助詞 w_2 + 動詞 w_3] 型の動詞性表現は 3,336,358 個であり，これらの前部分列 w_1w_2 の表記数は 110,822 個であった．

一方，JDMWE における[名詞 w_1 +格助詞 w_2 + 動詞 w_3] 型の動詞性表現の見出し数は 29,644 個，字種や表記のゆれ情報で展開した対象表記数は 82,983 個で，これらの前部分列 w_1w_2 の表記数は 14,075 個であった．

3.3 JDMWE に含まれる動詞性表現 w_1w_2 w_3 における動詞 w_3 の選択

JDMWE における動詞性表現の前部分列[名詞 w_1+格助詞 w_2]14,075 表記の内，10,548 個が GoogleN グラムにおける動詞性表現の前部分列[名詞 w_1 + 格助詞 w_2]に一致した[f]．これらの前部分列 w_1w_2 ごとに，各動詞の出現頻度を GoogleN グラムで求めた結果，JDMWE の動詞が GoogleN グラムで出現頻度第 1 位である場合が 4,983 件であり，対象とした前部分列表記 w_1w_2 の 47.24%=(4,983/10,548)*100 に対して条件付出現確率 $p_f(w_3|w_1w_2)$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」,「熱戦を繰り広げる」,「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 1,495 件で 14.17%，3 位は 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 2(a)に示す．このことから，JDMWE に収録されている表現は高い条件付確率のものほど多い傾向が示された．この傾向は，限られた形式の表現に対する条件付後方出現確率のみに関するものであるが，採録基準は全体に共通しており，条件付前方出現確率に関しても，また，その他の形式の表現に関しても類似した結果が得られるのではないかと推測している．

図 2(a)を累積の比率に改めたグラフを図 2(b)に示す．これから，例えば，JDMWE では，対象とする前部分列 w_1w_2 の約 80%に対して頻度 8 位までの動詞 w_3 が選ばれ，w_1w_2 の約 87%に 20 位までの動詞 w_3 が選ばれていることなどが分かる．また，図 2(b)を 20 位以降も考慮すれば，前部分列の 10%強に対して，後接する動詞が GoogleN グラムでは同環境に現れていないことが推定できる．例えば，JDMWE に存在する「才知に長ける」,「轢き逃げを働く」は GoogleN グラムに存在しない[g]．このことは，200 億文規模の WEB コーパスであっても，かなりの表現がとらえきれない可能性を示唆しており，Zipf の法則におけるロングテール部に対する表現収集の難しさを示すものと考えられる[h]．

f) 14,075 個の w_1w_2 を mecab0.96 で形態素解析したところ,1 グラムが 15 個，2 グラムが 11,829 個，3 グラムが 1,975 個，4 グラムが 221 個，5 グラムが 32 個，6 グラムが 3 個となった．そのため，w_1w_2 が 2 グラムの場合に限定して考えると，w_1w_2 の約 89.2%(=10,548/11,829*100)が GoogleN グラムに存在しているといえる．
g) 「才知に富む」は GoogleN グラムにおける出現頻度 43 で本辞書にも採録されている．一方 GoogleN グラムには「轢き逃げを告白する」が出現頻度 25 で存在するが，JDMWE には採録されていない．
h) GoogleN グラムには出現頻度カットオフが設けられており，出現頻度が 20 未満の N グラムは存在しないため，カットオフが大きすぎるという問題にも起因する．

Vol.2012-IFAT-105 No.4
Vol.2012-NL-205 No.4
2012/1/20

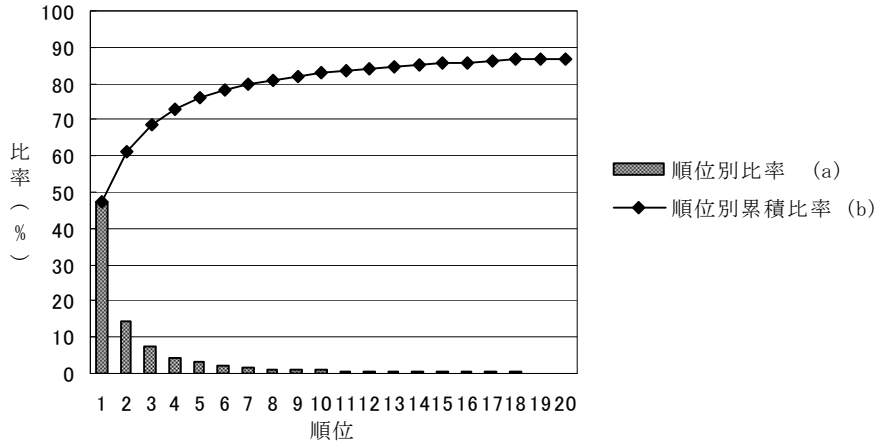

図 2 [名詞+格助詞+動詞]型表現の GoogleN グラムにおける動詞の出現頻度順位別の動詞採録率(a)と順位別の動詞採録累積比率(b)（格助詞「を」,「が」,「に」に限定）

3.4 JDMWE に含まれる動詞性表現 w_1w_2 w_3 における w_1w_2 の選択

[名詞 w_1 + 格助詞 w_2 + 動詞 w_3] 型の動詞性表現において，条件付出現確率 $p_f(w_3|w_1w_2)$が比較的大きい表現が採録されていることが図 2 で示されたが，条件付出現確率 $p_f(w_3|w_1w_2)$が比較的大きい場合でも，$p_f(w_3|w_1w_2)$が均一になっていない，言い換えるとエントロピーが小さい表現 w_1w_2 を優先して採録するほうが効果的である．

ここで GoogleN グラムにおける動詞性表現の，前部分列[名詞 w_1+格助詞 w_2]に対して後に続く動詞 w_3 に関する（正規化）エントロピー$H_f(w_3|w_1w_2)$を次の式のように与える．ここで N は[名詞 w_1＋格助詞 w_2]の後に続く動詞 w_3 の種類の数であるとする[i]．

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

次に，$H_f(w_3|w_1w_2)$の大きさの順に 20 区間に分け，（正規化）エントロピーの低い方から区間 1, 区間 2, … , 区間 20 として，それぞれの区間において JDMWE の[名詞＋格助詞]型表現（計 10,548 件）が含まれる比率を求めた[j]．各区間の比率をグラフ化して図 3(a)に，各区間の平均エントロピーを図 3(b)に示す．

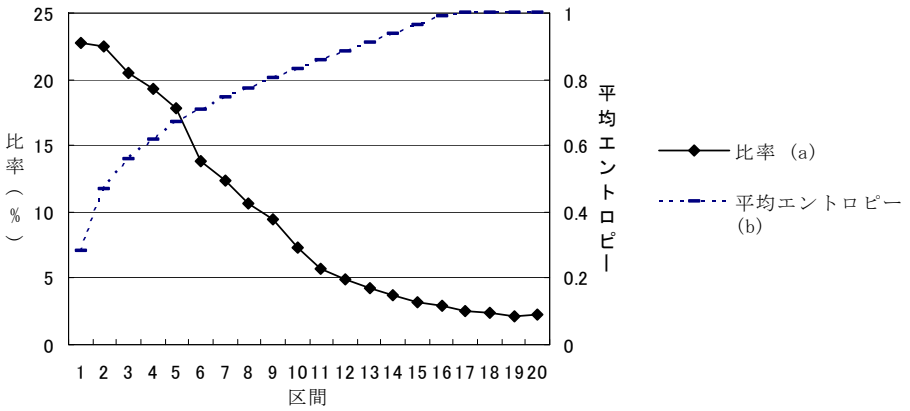

図 3 [名詞＋格助詞]型表現の GoogleN グラムにおける後続動詞(正規化)エントロピー区間別採録率(a)と各区間の平均エントロピー(b)（格助詞「を」,「が」,「に」に限定）

その結果，JDMWE に含まれる[名詞＋格助詞]型表現の割合は，区間 1 において 22.8%=(1,262/5,542)*100，区間 2 においては 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測された．このことから，JDMWE における動詞性表現[名詞 w_1 + 格助詞 w_2 + 動詞 w_3]における前部分列[名詞 w_1+格助詞 w_2]は，動詞部の（正規化）エントロピーが大きいほど，すなわち，後接する動詞部のパープレキシティが大きくなるほど，少なく採録されているという傾向が見られる．後に動詞が続く[名詞＋格助詞]型表現として，区間 1(平均エントロピーは 0.27)には，本辞書にある「墓穴を」「難色を」「凶弾に」などが観測された[k]．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といえる．また，区間ごとの平均エントロピーに視点を移すと，図 3(b)から，区間 18～20 では，それぞれ平均エントロピーは 1 であった[l]．エントロピーが 1 である[名詞＋格助詞]型表現は

i) エントロピーの最大値 $\log_2 N$ による正規化は，動詞が低頻度多種類で出現することによる影響を減らすためである．

j) 1 つの区間に属する[名詞＋格助詞]型表現の数は 5,542(=110,822/20)個となる．

k) JDMWE にある「墓穴を掘る」が出現頻度 44,197,「難色を示す」が 14,126,「凶弾に倒れる」が 2,835 で，それぞれ GoogleN グラムで出現頻度第 1 位で観測された．

l) エントロピーが 1 になる場合は，GoogleN グラムにおいて[名詞＋格助詞]型表現に続く動詞が 1 種類しか観測できなかった場合か，[名詞＋格助詞]型表現の後に続く動詞ごとに決まる条件付確率 $p_f(w_3|w_1w_2)$がすべて等確率であった場合のいずれかである．

Vol.2012-IFAT-105 No.4
Vol.2012-NL-205 No.4
2012/1/20

21,311 個観測され，そのうち動詞が 1 種類であったものが 21,081 個，2 種類であったものが 230 個であった．後に 1 種類の動詞が続く[名詞＋格助詞]型表現として「アマドコロが」「ダイカストに」「歯齦に」「巻添えを」などが[m]，また，後に 2 種類の動詞が続く[名詞＋格助詞]型表現として「毛じらみが」「ミゼットが」などが観測された[n]．

まとめると，図 2，図 3 から JDMWE の[名詞＋格助詞＋動詞]性表現は，条件付確率 p(動詞|名詞＋格助詞)が比較的大きく，かつ，動詞部のばらつきが比較的少ないものが選ばれている傾向がうかがえる．これらは限られた形態の表現に関するものではあるが，収録表現の一般的傾向と大差ないものと考えている．

4. おわりに

本論文では，日本語複単語表現辞書 JDMWE，および JDMWE への収録基準である，非構成性，強い単語間共起性について概説し，[名詞 w_1 + 格助詞 w_2 + 動詞 w_3] 型の動詞性表現を対象に，Google 共起頻度データ LDC2009T08 との比較を行い条件付後方出現確率，エントロピーの算出結果から JDMWE の統計的性質の一端を示した．JDMWE における表現の選定は基本的に内省に基づくもので，表現の網羅性を目指したために構成性が認められそうな表現や共起の排他性がそれほど高くない表現も採録されている可能性があるが，それに反して比較結果から，確率的側面に関しては表現の選定に大きな瑕疵は見られないと考えている．

参考文献

1) 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).
2) 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).
3) 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).
4) 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).
5) 田辺利文，本田聖晃，高橋雅仁，小山泰男，吉村賢治，首藤公昭: 日本語文末表現の取り扱いについて, FIT2006, pp.241-244 (2006).

m) それぞれ動詞を含めると「アマドコロがある」「ダイカストに該当する」「歯齦に終わる」「巻添えを食う」で，それぞれの出現頻度は 21，27，20，40 であった．ここで「巻添えを食う」は JDMWE に収録されている．
n) それぞれ動詞を含めると「毛じらみが(うつる/いる)」，「ミゼットが(ある/走る)」であり，それぞれ同じ出現頻度 27，30 で観測された．