日本語動詞性複単語表現(1 類)レキシコンの統計的性質

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

1. はじめに

最近の自然言語処理研究では，高性能な計算機環境，および Word2Vec に代表される意味情報を柔軟に計算できるアルゴリズムの登場など，意味をとり扱う研究が非常に活発な状況になっている[11]．筆者らは，意味を考慮した自然言語データ処理においては，定型表現を十分に考慮したシステムの構築が不可欠であると考えている．定型表現に関する研究としては，自然言語処理分野では[6]で提唱された Multiword Expression(MWE, 複単語表現)があるが，ほぼ同時期に言語学で Formulaic Language[3]や Lexical Bundles[2]などの概念が提唱されるなど，研究分野の垣根を越えて世界的に定型表現の重要性が認識されている．本論文は日本語 MWE を中心にとり扱う．

これまで著者の 1 人は，非構成性，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収集・整理し，日本語複単語表現レキシコン JMWEL(Japanese MWE lexicon)を編纂してきた[7][8][9][10].

本論文では JMWEL の中心的サブレキシコンの 1 つである日本語動詞性複単語表現(1 類)レキシコンに収録されている表現の統計的性質の一端を Google 共起頻度データ LDC2009T08 との比較によって明らかにする．なお，本論文で挙げている統計データは，過去にも報告している[7][10]が，今回，2017 年 11 月の GSK による公開を機に，改めて紹介するものである．

2. 日本語動詞性複単語表現(1 類)レキシコン

著者の 1 人は日本語複単語表現レキシコン(JMWEL: Japanese MWE Lexicon)を古くから開発してきており，現時点では見出し数は異なりで 12 万強に達している．

JMWEL の見出しの採録基準は基本的に次の 2 種の特異性のうち少なくとも 1 つを有することである．1 つは，例えば，『油を売る』の意味が「油」，「を」，「売る」という要素語の通常の意味から導くことが難しいという性質(非構成性，イディオム性)，他の 1 つは，例えば，「こまねく」という動詞は『手をこまねく』以外にはほとんど使われないというような性質(決まり文句性，要素語間の強共起性)を言う．JMWEL 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，(2)多様な構文構造をカバーする，(3)異表記や派生形もできるだけ網羅する，(4)構文的柔軟性を表現できる構文構造情報を与える，などが挙げられる．JMWEL の記載情報などの詳細は[7][8][9][10]を参照されたい．

JMWEL は，見出しの文法機能（相当品詞）で分割された 11 種のサブレキシコン，および，トピックで分割された 8 種のサブレキシコンに分けて管理・公開されている．文法機能が動詞性の日本語動詞性複単語表現レキシコンは 1 類，2 類，3 類の 3 種類に分割されている[1]が，本論文では JMWEL の主要なサブレキシコンの 1 つである日本語動詞性複単語表現(1 類)レキシコンについて述べる．このレキシコンには[名詞+ 格助詞(が，を，に）＋動詞]の形式の動詞句，例えば『手を結ぶ』，『意味がある』，『沽券に関わる』などの約 36,000 表現が収録されており[2]，2017 年 11 月に言語資源協会(GSK)を通じて公開されている[4][9][3]．本論文では，これ以降，日本語動詞性複単語表現(1 類)レキシコンを「本レキシコン」と書くことにする．

3. 本レキシコンの統計的性質

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，JMWEL 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，要素語間強共起性の MWE に関しては，統計的性質が客観的に評価できる．要素語間の相関尺度としては種々のものが考えられるが，ここでは，文末方向の遷移確率と正規化エントロピーを Google の日本語 Web N グラム共起頻度データ LDC2009T08[5]（これ以降「GoogleN グラムデータ」と略記する）を用いて算出した．GoogleN グラムデータは 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．

対象とした本レキシコンの表現は[名詞 w_1 + 格助詞 w_2 + 動詞 w_3] 型の動詞性表現であり，格助詞 w_2 を「を」，「が」，「に」に限定し，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．

---

[1] 文法機能で分割したサブレキシコンと，トピックで分割されたサブレキシコンには同一の見出しが採録されていることも少なくない．例えば『油を売る』は，日本語動詞性複単語表現(1 類)レキシコン，および日本語慣用句レキシコンのそれぞれに採録されている．

[2] [サ変名詞 + を +(する or 遣る or 行う or 実行する)],[サ変名詞 + が + できる] の形式の表現は一部を除き JMWEL 全体としても収録対象外としている．

[3] 日本語動詞性複単語表現(1 類)レキシコン(JMWEL_verbal(class1) v3.2)は，GSK2017-C(http://www.gsk.or.jp/catalog/gsk2017-c/)で，日本語動詞性複単語表現(2 類)レキシコン(JMWEL_verbal(class2) v3.2)は，GSK2017-D(http://www.gsk.or.jp/catalog/gsk2017-d/)で公開されている．

GoogleN グラムデータにおいても上記と同一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，いいかえると GoogleN グラムデータにおいて，名詞部には品詞が名詞である文字列が，動詞部には品詞が動詞である文字列がそれぞれ与えられているとみなせるものを，GoogleN グラムデータ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[4].

3.1. 本レキシコンの表現 $w_1w_2w_3$ における前部分列 w_1w_2 のバリエーション

GoogleN グラムデータにおける[名詞 w_1 +格助詞 w_2 + 動詞 w_3] 型の複単語表現は 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，本レキシコンにおける[名詞 w_1 +格助詞 w_2 + 動詞 w_3] 型の動詞性表現のうち，検証に用いた見出し数は 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.

3.2. 本レキシコンの表現 $w_1w_2w_3$ における動詞 w_3 の選択

本レキシコンに含まれる表現の前部分列[名詞 w_1+格助詞 w_2]14,075 表記の内，10,548 個が GoogleN グラムデータ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[5]．これらの前部分列 w_1w_2 ごとに，各動詞 w_3 の出現頻度を GoogleN グラムデータで求めた結果,本レキシコンの動詞が GoogleN グラムデータで出現頻度第 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 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 JMWEL 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 1(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，本レキシコンでは，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 1(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 グラムデータでは同環境に現れていないことが推定できる．例えば，本レキシコンに存在する『才知に長ける』,『轢き逃げを働く』は GoogleN グラムデータに存在しない[6]．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[7].

3.3. 本レキシコンの表現 $w_1w_2w_3$ 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 1 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解析の際にも効果的である.

[4] GoogleN グラムデータ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.

[5] 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 グラムデータに存在しているといえる.

[6] 『才知に富む』は GoogleN グラムデータにおける出現頻度 43 で本辞書にも採録されている．一方 GoogleN グラムデータには『轢き逃げを告白する』が出現頻度 25 で存在するが，本レキシコンには採録されていない.

[7] GoogleN グラムデータ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，カットオフが大きいという問題にも起因する.

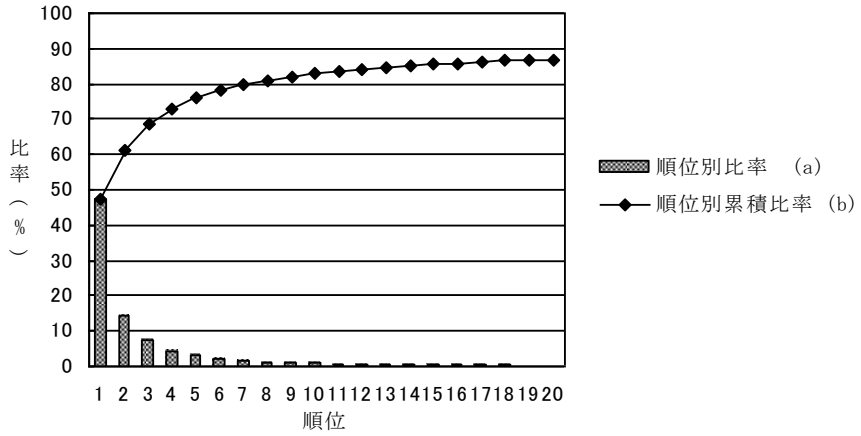

図 1 [名詞+格助詞+動詞]型表現の GoogleN グラムデータにおける動詞の出現頻度順位別の動詞採録率(a), および, 順位別の動詞採録累積比率(b) (格助詞「を」,「が」,「に」に限定)

ここで GoogleN グラムデータ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$H_f(w_3|w_1w_2)$を次式によって求めた．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[8].

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

次に，得られた $H_f(w_3|w_1w_2)$を昇順に並べ，区間 1,区間 2, … , 区間 20 と 20 区間に分割して，それぞれの区間において本レキシコンの[名詞 w_1＋格助詞 w_2]型表現(計 10,548 件)が含まれる比率を求めた[9]．各区間の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.

その結果，本レキシコンに含まれる[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察された．このことから，本レキシコンの動詞性表現[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，続く動詞部の正規化エントロピーが小さいほど，すなわち，後接する動詞部のパープレキシティが小さくなるほど，多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」などが観察された[10]．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 2(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[11]．エントロピーが 1 である[名詞＋格助詞]型表現は 21,311 個観察され，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[12]，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」などが観察された[13].

[8] 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.

[9] 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.

[10] 本レキシコンにある『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 グラムデータで出現頻度第 1 位で観測された.

[11] エントロピーが 1 になる場合は，GoogleN グラムデータにおいて[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.

[12] 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』は本レキシコンに収録されている.

[13] 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.

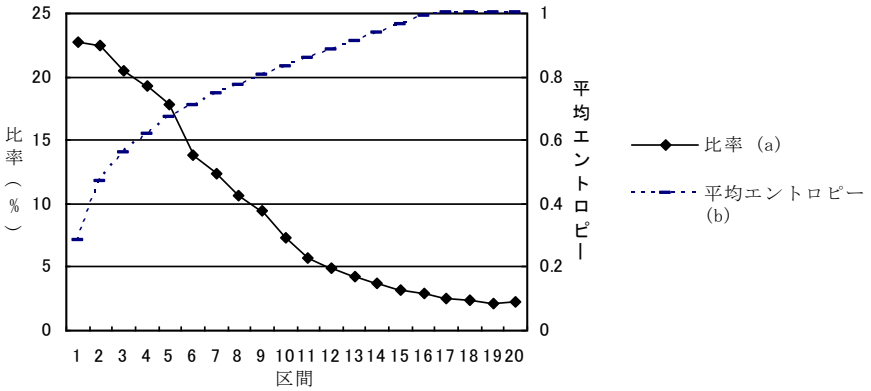

図 2 [名詞＋格助詞]型表現の GoogleN グラムデータにおける後続動詞(正規化)エントロピー区間別採録率(a), および, 各区間の平均エントロピー(b)(格助詞「を」,「が」,「に」に限定)

まとめると，図 1，図 2 から本レキシコンの[名詞＋格助詞＋動詞]型表現は，[名詞＋格助詞]部から[動詞]部への遷移確率 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という傾向がうかがえる．この結果は 1 類という限られた形態の動詞性表現に関するものではあるが，JMWEL の表現採録基準はほぼ共通であるから JMWEL 全体の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，日本語動詞性複単語表現(1 類)レキシコン JMWEL_verbal (class1)に収録されている[名詞 w_1 + 格助詞 (を，が，に) 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，正規化エントロピーの算出結果から JMWEL の統計的性質の一端を示した．JMWEL 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．しかし内省に基づく表現選定のメリットは，要素語間の共起性の高い表現のみならず，エントロピーの低い表現，すなわち解析の際のあいまいさ低減に効果的である表現が自然に収集されやすいことであり，本論文で示した検証実験によりそれらの性質が示された．本レキシコンと日本語動詞性複単語表現(2 類)レキシコンは GSK により既に公開されている．将来の自然言語処理の発展のために JMWEL が役立つことを期待したい．

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

[3] 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).

[4] GSK 特定非営利活動法人 言語資源協会 (2018).

[5] 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).

[6] 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).

[7] 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).

[8] 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).

[9] 髙橋雅仁, 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－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. (2018).

[10] 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).

[11] https://code.google.com/archive/p/word2vec/