

## 日本語呼びかけ・応答・挨拶・独言・間投表現レキシコンの性質と応用

田辺利文\* 高橋雅仁\*\* 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

『油を売る』『赤の他人』『金魚のフン』などの、近年、構成性に基づいて意味を扱うことが難しいイディオムや語の強い結合によって成り立つ決まり文句など特異性のある複数の単語からなる表現を複単語表現(Multi-Word Expression:MWE)と名付け、機械処理の立場からその全体像を考察した Sag らの論文 [5] がきっかけとなり、自然言語処理における MWE の重要性が広く認識されるようになった。著者の 1 人は、古くから新聞記事、雑誌記事等から日本語の複単語表現の収集を行い、日本語複単語表現レキシコン(以下、JMWE: Japanese MWE Lexicon と呼ぶ)としてまとめ、広く研究用に公開している [6][7][8][9]。

本論文では、JMWE の現状を紹介するとともに、『本当?』、『参ったなあ』『今に見てろ』『おととい来やがれ』等の表現に発話者の感情情報が付与された、JMWE のサブレキシコンの 1 つである日本語呼びかけ・応答・挨拶・独言・間投表現レキシコンにスポットをあて、見出しに記載された感情情報をもとにレキシコンの性質の一端を述べたのち、現時点での課題とこれからの応用について紹介する。

## 2. 日本語複単語表現レキシコン JMWE

著者の 1 人は日本語複単語表現レキシコン(JMWE: Japanese MWE Lexicon)を古くから開発してきており、現時点では、見出し数は基本形で 10 万件を超え、一般の日本語処理用レキシコンとしても有効なレベルに達している [6][7][8][9]。

JMWE の見出しの採録基準としては基本的に次の 2 種の特異性に着目している。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。これらの性質のうち少なくとも 1 つを有するものを JMWE の見出しとして採録している。採録に関しては、新聞記事、雑誌記事等をもとに内省に基づいているものの、これまでの研究で、(1) 収録表現は高頻度で文に現れるものが多く辞書は相対的にコンパクトにまとまっていること、(2) ビッグデータにも現れない表現をも収録していること、などから収録方法が妥当であることが示されている [6][7][9]。JMWE 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で構成される。JMWE の全体、サブレキシコンや記載情報などの詳細は [6][7][8][9] を参照されたい。

## 3. 日本語呼びかけ・応答・挨拶・独言・間投表現レキシコン

## 3.1. レキシコンの記載情報

日本語呼びかけ・応答・挨拶・独言・間投表現レキシコンは、JMWE のトピックで分割されたサブレキシコンの 1 つで、基本的に JMWE に収録された見出しのうち、呼びかけ、応答、挨拶、独言、間投表現としての機能をもつ口語表現をまとめたものである。見出しには、例えば、『あーおもしろい』『今に見てろ』『おととい来やがれ』『冗談じゃない』『すぐこれだ』『そうなんだよね』『そんなアホな』『なんということでしょう』『ほんとですかあ』『もう一度言ってみろ』『いやだ』などの多彩な表現が現時点で約 1,100 件収録されている。

日本語呼びかけ・応答・挨拶・独言・間投表現レキシコンの記載情報としては、一般的な JMWE の記載情報には含まれていない、種別情報および意味用法がそれぞれの見出しに与えられているのが特徴である。次にそれぞれの詳細を述べる。

**種別情報** : Call(呼びかけ), Res(応答), Self(独言), Lgh(笑い), Grt(挨拶), Fil(間投表現), Sht(叫び)の 7 種のうちの 1 つ。

**意味用法** : <疑問>, <驚き>, <理解>, <感謝>, <肯定>, <否定・拒否>, <納得・了解>, <同情>, <詫言・謝罪>, <合槌・単純応答>, <困惑>, <不満>, <喜び・満足>, <共感>, <寂しさ>, <つらさ・痛さ・苦しさ>, <感動>, <鬱陶しさ>, <悲しさ>, <羨ましさ>, <面白さ>, <しくじり>, <怒り・不快>, <恨み>, <怖さ>, <依頼>, <応援> と極めて詳細に設定された 27 種の感情情報(感情タグと呼ぶ)のそれぞれに 0(無い)から 9(きわめて強い)までの 10 レベルの強さを与えている。意味用法とは見出しに対応した 27 次元のベクトルである(以降、これを感情ベクトルと呼ぶ)。

感情タグの体系やレベルの設定も人手で設計していることなども本レキシコンの特徴である。例えば、『本当かなあ』の意味用法は、<疑問>が 4、<否定・拒否>が 1 のように設定されている<sup>2</sup>。

<sup>1</sup> 同一の見出しで複数の種別情報をもつものも存在する。例えば、『こまったなあ』は Res(応答)と Self(独言)の 2 つの用法がある。また種別情報を、以降、カテゴリと呼ぶことにする。

<sup>2</sup> レキシコン中での記載は、意味の強さが 0 の場合には空欄となっており、

本レキシコンの感情ベクトルを自然言語処理システムに組み込むことで、柔軟な対話処理、具体的には発話の強弱の調整や、状況に応じた言い換えも可能になるなど、これまでに類をみない、繊細な感情情報処理システムが実現できると考えられる。

### 3.2. 感情の強さ

見出しに与えられた感情ベクトルをもとに、本レキシコンの性質をうかがうことができる。感情ベクトルを構成するそれぞれの感情タグには0~9と10レベルの数値が与えられているため、感情ベクトルの絶対値(大きさ)が計算できる。絶対値が大きい見出しほど感情が「強い」ものと解釈できる。本レキシコンに採録されている見出しに対する感情ベクトルの大きさを算出した結果、例えば、『なにっ』が14.3、『だいじょうぶ』が13.3、『おもてにでろ』が12.7、『おぼえている』が10.7、『おとといきやがれ』が9.9、また、『こまっちゃう』は14.0となった。この結果から、おおむね強い感情をもつ見出しにはベクトルの絶対値が高くなるように値が設定されていると考えることができる。

同様に、カテゴリごとについても感情ベクトルの絶対値を計算する。表1に、それぞれのカテゴリに対して、そのカテゴリに属する見出し数、感情の強さの平均、感情の強さの分散<sup>3</sup>、最小値、最大値、最大値を取った見出しの例を示す。

表1から、感情の強さの平均は、Res(応答)に属するものが高く、それにSelf(独言)とSht(叫び)が続き、Fil(間投表現)やGrt(挨拶)が低いという結果となった。Grtの強さが低いのは、一部を除いて、感情があまりこもっておらず、感情伝達の上では挨拶という行為が平準化していると考えられる。また、Self(独言)は、例えば『こまっちゃう』はSelfの強さの最大値である14.04を取っているなど、強い感情をもつものが観察された。しかし『こまっちゃう』は機能的には独言であるため、感情は強いものの、その感情を自分自身で消化しようとしていると捉えることもできる<sup>4</sup>。Call(呼びかけ)は、強さの平均はあまり高くはないものの、強さの分散が大きいことが分かった。Callに属する見出しで強かったものとしては『だいじょうぶ』が13.3、『おもてにでろ』が12.7、『おぼえている』が10.7、『おとといきやがれ』が9.9と観察されたが、その反面、感情の強さが0である見出し(後述する)も少なくないことが観察された。さらにすべてのカテゴリにおいて、感情の強さの最小値はいずれも0であることが観察された。

空欄でない場合には1~9までの整数の数値が記述されている。また、見出しと種別情報が同一でも、複数の意味用法をもつものも存在する。例えば『そうですか』の種別情報がRes(応答)になるものは意味用法が2つ存在し、1つは、<疑問>が7、<否定・拒否>が2であり、もう1つは、<理解>が2、<納得・了解>が2、<合槌・単純>が5である。-

<sup>3</sup> ここでの分散は標本分散を用いている。

<sup>4</sup> 『こまっちゃう』は、Res(応答)の種別情報をもつものも別見出しとして収録されている。

### 3.3. 感情タグ体系の設定

日本語呼びかけ・応答・挨拶・独言・間投表現レキシコンで定義された27種の感情タグには、粒度が細かいことから似た意味をもつ感情タグが存在しうること容易に推測される。そこで本研究では、感情タグごとの相関係数を算出することを試みた。本レキシコンでは感情タグは列ごとに記載されているため、それぞれの見出しにおける感情タグの値(0~9)を要素とする列ベクトルを用いる<sup>5</sup>。

感情タグ間の相関係数を求めた結果、相関係数の絶対値が高かったものから、<怒り・不快>と<不満>の間が0.54、<納得・了解>と<理解>の間が0.50、<否定・拒否>と<怒り・不快>の間が0.42、<合槌・単純>と<理解>の間が0.42と算出された<sup>6</sup>。相関係数の値が負のものの絶対値の最大値をとったものは<疑問>と<共感>の間の-0.16であった。そこで具体的に見てみると、感情タグ<怒り・不快>と<不満>にかかわる表現として、『じょうだんじゃないよ』では、<怒り・不快>の値は6、<不満>は8であり、『じょうだんをいうな』では、<怒り・不快>の値は5、<不満>は0であった。このように、若干の依存関係はあるものの、具体的な見出し例から、細分化された感情タグを設定することには特に問題はないと考えることができる<sup>7</sup>。

### 3.4. 言い換えの可能性

日本語呼びかけ・応答・挨拶・独言・間投表現レキシコンを用いた応用の1つとして言い換えが挙げられる。見出しに対応した感情ベクトルを用いて、2つの見出し間のコサイン類似度を計算することが可能である。コサイン類似度が1に近い場合には、それらの見出し同士は意味的に似ているものと解釈できる可能性がある。今回、コサイン類似度の算出結果が1になる見出しのペアが複数存在することが確認された。このようなペアを使うことで工学的な応用が可能となりえる。

例えば、カテゴリがSelf(独言)である『うれしいな』とRes(応答)である『よかったです』は、類似度が1になったペアの1つであり言い換えが可能である。このようなポジティブな意味を含むペアの場合には、独言ではなく、相手への応答という形式に言い換えて、ポジティブな気持ちをダイレクトに相手に伝えることが考えられる。

<sup>5</sup> この列ベクトルの要素数はレキシコンの見出し数となる。

<sup>6</sup> 相関係数をrとすると、rは-1から1の間の値をとり、一般的には、 $0.7 \leq |r| \leq 1$ の場合にはかなり強い相関があり、 $0.4 \leq |r| \leq 0.7$ の場合にはやや相関があり、 $0.2 \leq |r| \leq 0.4$ の場合には弱い相関があり、 $0 \leq |r| \leq 0.2$ の場合にはほとんど相関はない、と言われている。

<sup>7</sup> 『じょうだんをいうな』と『じょうだんをいえ』は、共に<怒り・不快>と<不満>の値がそれぞれ5、0になっている。これらの表現は同じ意味として使われることから感情ベクトルも同一になっている。この場合、単語単位で意味を認識、合成しようとする否定表現の有無により、表現同士が同じ意味であることの認識がしにくくなってしまったため、表現の意味の認識には、単語単位ではなくMWEを単位とみなす考え方が重要であることを示唆している。

カテゴリ	表現数	強さの平均	強さの分散	最小値	最大値	最大値を取った見出しの例
Call(呼びかけ)	108	4.28	20.01	0	13.30	だいじょうぶ
Res(応答)	505	7.24	5.87	0	14.35	なにっ
Self(独言)	224	5.76	10.58	0	14.04	こまっちゃう
Lgh(笑い)	19	3.30	5.76	0	6.40	うふふ
Grt(挨拶)	136	1.69	7.39	0	8	ごめんなさい
Fil(間投表現)	11	1.45	5.70	0	6	なんといいますかー
Sht(叫び)	42	5.73	13.26	0	12.25	やったー

表1 カテゴリごとの、感情の強さに関するデータ

それとは逆に、ネガティブな意味を含む表現の場合でも類似度が1になるペアがあり、言い換えが可能なものとして具体的にはRes(応答)である『じょうだんも休み休み言え』とSelf(独言)『いやだなあ』が観察されたが、このようなネガティブケースである場合には、応答という形式を用いずに代わりに独言の形式を用い、間接的に、それとなく気持ちを伝えるような発話をするということが考えられる。

その一方、類似度が1になるペアであっても言い換えが可能であるとは限らない。Call(呼びかけ)である『これはなに』は<疑問>が8で、Res(応答)である『さあね』は<疑問>が4であるためこれらのペアは類似度が1になるが、意味的に正しく言い換えることは出来ない。感情ベクトルは疎であるものが多いこともあり、このように言い換えできないペアが多数散見されている。また、コサイン類似度は、感情ベクトルの方向がどれだけ一致しているかの指標であり、感情の強さ(感情ベクトルの大きさ)には無関係である。そこで、コサイン類似度が1である表現で言い換えできるペアに対して、それらの見出しの感情の強さも考慮できれば、相手に対して、意味は変えないまま感情の強さを変化させるようなより高度な言い換えも可能となる。

#### 4. 関連研究

感情分類の体系は研究者によって大きく異なる。日本語において、表現と感情との対応関係が示されているものとして感情表現辞典[2]がある。感情表現辞典は、<喜>、<怒>、<哀>、<怖>、<恥>、<好>、<厭>、<昂>、<安>、<驚>の10種類の感情を表した意味タグ(感情タグ)を定義し、感情を持つ表現(複単語表現を含む)がどの感情タグをもつかの対応関係が示されている。また、ラッセルの円環モデル[4]は、「すべての感情は、快(pleasure)–不快(unpleasure)、覚醒(arousing)–眠気(sleepy)を2軸とする平面に円環上に並んで存在する」というもので、すべての感情は2次元ベクトルとして表現できることを示しており、28種類の感情が円環状にどのように配置されるかが示されている。感情情報が2次元ベクトルで表現されることから計算機での取り扱いには適切であるが、表現と感情との対応関係は明確に示されておらずモデルの提案に留まっている。

本研究で用いる日本語呼びかけ・応答・挨拶・独言・間投表

現レキシコンは電子化されており、見出し約1,100、感情タグが27種類、感情の強さを10レベルで設定しているため、詳細でありながら計算機で利用しやすいという特徴をもつ。

#### 5. おわりに

本論文では、日本語の複単語表現を収録したJMWELに触れるとともに、そのサブレキシコンの1つである日本語呼びかけ・応答・挨拶・独言・間投表現レキシコンを紹介した。本レキシコンは、従来の感情分類体系には見られない、人手で設計した詳細な感情タグとレベル設定が大きな特徴である計算機用レキシコンである。

今後の課題としては、まずは感情表現辞典やラッセルの円環モデルに挙げられるような既存の感情分類との対応関係の整理が考えられる。対応関係が階層的に表現できれば都合がよい。

次に、感情の強さの粒度の妥当性の評価が必要となる。それぞれの感情タグには10レベルで感情の強さを表記しているが、例えば、10レベルの数値を、0~4(小)、5~9(大)の2レベルに分けたのち、感情タグの粒度が異なることでの違いを評価することが考えられる。

次はベクトル演算の可能性を示すことである。見出しに感情ベクトルが対応していることから、例えば、表現eの感情ベクトルを $V(e)$ とすると、表現 $\alpha$ 、 $\beta$ に対して、 $V(\alpha)+V(\beta)=V(\gamma)$ となる表現 $\gamma$ が存在すれば、表現 $\alpha$ と $\beta$ を含む文に対して、 $\alpha$ と $\beta$ の代わりに $\gamma$ で言い換えられるか等を調査する。

本論文で提案した27種の感情は、日本語呼びかけ・応答・挨拶・独言・間投表現レキシコンの見出しにのみ存在するものではなく、一般的な単語やMWEにも含まれる。MWEに含まれる感情タグとしては、例えば、動詞性複単語表現『手に余る』には<鬱陶しさ>、『目に余る』には<怒り・不快>、『手を結ぶ』には<納得・了解>などが、また動詞性以外の複単語表現、例えば『火の車』には<つらさ・痛さ・苦しさ>、『荷が重い』には<困惑>、『願ったり叶ったり』には<喜び・満足>などが考えられる。このように感情情報を含むMWEは少なくなく、JMWELに登録されている見出しに対して27種類の感情タグによる感情ベクトルを与えることができれば、より柔軟な感情情報処理が可能になるだろう。

その一方で、本レキシコンでの感情タグは27種類に設定されているが、感情の粒度は27種類で良いのかという問題がある。本レキシコンに収録されている表現の感情ベクトルの大きさが0である見出しは少なくなく、そのような見出しとしては、Callの『いってきます』『いってらっしゃい』『こらっ』『すみませんが』、Filの『うーん』、Grtの『おつかれさまです』『こんにちは』『じゃあね』、Lghの『おほほ』『がっはっは』、Resの『おそれいました』『そいつはありがたい』『まいったか』、Selfの『おー』『さーてね』『そうねえ』『しめたっ』、Shtである『おっとっ』『えいやっ』があるようである。このような表現は意味がないものと考えて、文の解析処理中に除去することも考えられるが、このような表現にも何らかの意味があると考えると、既存のどれかの感情タグが示すわずかな意味が入っているか、または既存の感情タグには存在していない感情タグが必要になる、などが考えられる。感情タグの種類は研究者ごとにまちまちであり、粒度の細かいものとしては、感情表現辞典のあとがきによれば、「表現を素材として見た感情の諸相は「慶・喜・謝・幸・誇・楽・明・爽・快・温・浮・恍・躍・弾・笑・怒・憤・屹・不愉快・煮・癩・不機嫌・悲・淋・怖・恥・辱・好・厭・嫌・憎・悔・妬・惑・鬱・惰・苦・焦・苛・緊・昂・動・安・驚・衝・呆」という四六種ほどの広がりが見られる」とある。つまり27種類の感情タグであっても多いとは断言できない。本論中でも述べたが、類似度が1になるペアであっても言い換えが可能でないケース、例えばCall(呼びかけ)である『これはなに』は<疑問>が8で、Res(応答)である『さあね』は<疑問>が4の場合などがあるため、言い換えできる条件として、新たな感情タグを追加することなどして新規に感情ベクトルを定義した上で類似度が1にならないようにすることなどが考えられる。

さて、構想中ではあるが、上に列挙した感情ベクトルの大きさが0になる見出しとして、例えば、『こらっ』『おほほ』『がっはっは』『さーてね』『おっとっ』には本レキシコンでは定義されていない、<おかしさ>を意味する感情タグが潜んでいるように見受けられる。他にも、主観的ではあるが『おととい来やがれ』『なんということでしょう』『カンラカラカラ』『えらいこっちゃ』『およよ』『うえー』『おっとどっこい』などにも潜んでいると考えている。このタイプの感情タグは主観的な側面が強だけでなく発話状況にも依存するため感情タグの付与や値の設計は困難を極めることが予想されるが、著者の知る限り<おかしさ>を感情タグとして定義しているものはなく、適切に実現できれば予想だにしないような斬新なシステムが登場するかもしれない。このように、表現の意味を再考しつつ感情の粒度についても再検討する必要もある。

応用としては、対話システムへの組み込みが考えられる。対話システムへの組み込みには、対人関係のモデルも組み込むことを考えており、ラッセルの円環モデルと類似したモデルとして対人円環モデルがある[1][3][10]。対人円環モデルの考え方は、温和-冷淡(親密性)、支配-服従(支配性)を2軸とする平面に対

人特性がマッピングされる、というものであり、ラッセルの円環モデルと同様に低次元のベクトルで表現可能である。対話の状態は固定的ではなく、対話が進むにつれて対人特性が円環上で変化することを仮定した上で、27次元の感情ベクトルが円環上のどの位置に対応するかが分かれば、発話をモデルに組み込むことも可能であると考えている。

対話システムへの組み込みの延長線上にあるのが、対話ロボットへの実装である。現在、1人暮らしの高齢者向けに、会話ができるペットロボットのニーズなどが高まってきているが、感情情報付きの会話・口語表現を収集した言語資源が十分に整備されていないことから、現状では、高齢者とペットロボットなどで潤滑に会話のやりとりを実現するのは難しいと考えており、本論文で紹介した日本語の呼びかけ・応答・挨拶・独言・間投表現レキシコンを用いることで、繊細な感情認識にも耐えうるような対話ロボットシステムの構築も目指している。

## 参考文献

- [1] 橋本泰央, 小塩真司: 対人円環モデルに基づいた IPIP-IPC-J の作成. 心理学研究, 87, pp.395-404. (2016).
- [2] 中村明: 感情表現辞典. 東京堂出版. (1993).
- [3] 大淵憲一, 堀毛一也: パーソナリティと対人行動. 誠信書房. (1996).
- [4] J.A.Russell: A circumplex model of affect. Journal of Personality and Social Psychology, Vol.36, pp.1161-1178, (1980).
- [5] Ivan. A. Sag et al.: Multiword Expressions: A Pain in the Neck for NLP. Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing, Volume 2276, pp 1-15. (2002).
- [6] Kosho Shudo, Akira Kurahone, and Toshifumi Tanabe: A Comprehensive Dictionary of Multiword Expressions, Proceedings of the 49th Annual Meeting of the ACL, pp.169-177, (2011).
- [7] 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).
- [8] 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-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).
- [9] Toshifumi Tanabe, Masahito Takahashi, and Kosho Shudo: A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing. Computer Speech and Language, 28-6, pp.1317-1339. Elsevier. (2014).
- [10] Wiggins, J. S., Trapnell, P., and Phillips, N: Psychometric and geometric characteristics of the Revised Interpersonal Adjective Scales (IAS-R). Multivariate Behavioral Research, 23, pp.517-530. (1988).