日本語の複単語表現データ

田辺 利文　高橋 雅仁*　吉村 賢治　首藤 公昭

福岡大学工学部電子情報工学科
*久留米工業大学工学部情報ネットワーク工学科

{ tanabe, yosimura, shudo }@tl.fukuoka-u.ac.jp
* taka@cc.kurume-it.ac.jp

1 はじめに

従来、各種言語ルールの構築に障害となる慣用句などの複数単語からなる固定的な言い回しが余り重要視されずに来た観があるが、近年になってこれら複単語表現、MWEs: Multi-Word Expressions に真面目に対処することが自然言語処理において不可欠であることが広く認識されるようになってきた(Sag et al., 2002)。これらに対処するためにはまず、どのような表現が MWE とみなされるべきかが重要となる。筆者らは日本語に関して基本的な MWE 候補をまとめて既に報告しているが(首藤ら, 1989)、未だ網羅的、体系的な日本語 MWE データベースは存在しない。本稿では筆者らが公開を目標に現在構築を進めている日本語 MWE データベースの概要を紹介する。

2 MWE データベース

2.1 収集した表現とその属性

日本語のような膠着言語の意味処理においては、どのように文の基本単位を捉えるかが特に重要な問題となる。我々は広範な領域の大規模日本語データに基づき 1970 年代から人手によって意味上の単位と考えるべき表現の収集・整理を行ってきた。（首藤他， 1988、首藤, 1989）我々が収集してきた表現は次に示す 3 つの性質のうち少なくとも 1 つを持つと考えられる長単位表現（単語列）と言うことが出来る。

f_1: 熟語性
f_2: 語彙的一体性
f_3: 確率的束縛性

熟語性とは、構成性原理の成り立ちにくさを意味しており、構成している単語の通常の意味から全体の意味を構成するのが難しいことを指す。語彙的一体性とは、分離しにくさ（要素単語の間への他の単語の割り込みにくさ）を、確率的束縛性とは、要素単語相互の確率的な共起しやすさを意味する。従って、収集した各表現は基本的にこれらの性質の有無や程度を表す 3 つ組(f_1 f_2 f_3)によって性格付けされるが、現在はこれらの性質の有無の情報を 2 値(1.0)で表すに留めている。有無の判断は収集者の内省によっている。たとえば、熟語的であり、分離可能で、確率的に強くは結び付いていない表現、“骨・を・折る”（「苦労する」に近い意）は(100)、構成的かつ分離可能で確率的に強く結びついていると考えられる表現、“口・を・つぐむ”は(001)と分類される。ここで、表現例中の記号‘・’は通常の単語境界を表す。また、以下では各長単位表現の例にこれらの分類コードを記す。

(Sag et al., 2002)は英語における複単語表現、MWE (Multi-word Expression)に関して基本的な考察を行っている。彼らは MWE の種類として fixed expression 、 decomposable idiom 、 institutionalized phrase、non-decomposable idiom、light verb construction などを挙げているが、これらを上記の枠組みで表現すれば、それぞれ(110)、(100)、(001)、(0x1)、(0x1)に対応すると考えられる。ただし、x は 1、0 のいずれもあり得ることを意味する。以下にこれらの表現の例を示す。

- **fixed expression (110)**
  構成性原理が成り立たず、かつ、構成している単語が分離できない表現。
  例）“ 赤の他人 ”、“ 前向き ”

・**decomposable idiom (100)**
語彙の解釈を例外的なものにまで広げれば構文構造に沿って意味が構成できる表現。
例）“ 腕を上げる ”、“ 手を広げる ”
・**institutionalized phrase (001)**
構成性は成り立つが、強い共起性をもつ表現。
例）“ 塩コショウ ”、“ 機械翻訳 ”
・**non-decomposable idiom (0x1)**
語彙の解釈をどう変えても構成的に意味が生成できない表現。
例）“ 顔に泥を塗る ”
・**light verb construction(0x1)**
名詞と機能動詞の結合。一般に名詞に対して動詞が決まっている。
例）“ 連絡をとる ”、“ 考慮に入れる ”

このように f_1 、f_2 、f_3 は MWE が資格として持つべき基本的な属性となっている。そこで本稿では筆者らの収集した日本語長単位表現を日本語の複単語表現、MWE と呼ぶ。

2.2 相当する文法カテゴリー

これまで収集した MWE は基本形約 65,000 個であり、大きく機能語的 MWE と概念語的 MWE とに分けられる。機能語的 MWE には関係表現と助述表現とがある。関係表現とは、格助詞、接続助詞およびそれらに相当する表現であり、助述表現とは、時制、相、話者の態度、判断否定など、広義の様相情報を与える助動詞、終助詞、およびそれらに相当する表現である。概念語的 MWE の文法カテゴリーは下記のとおりである。ここでの名詞的表現の属性は(11x)であり、(00x)である複合名詞や固有名詞は収集対象としていない。それぞれの類に属する表現の概数と表現例を以下に示す。

機能語的 MWE:
・**関係表現**<1,000>:“に・つい・て(110) ”; “に・よっ・て(110) ”; “と・とも・に(110) ”; “に・おける(110) ”; etc.
・**助述表現**<1,500>: “の・で・は・ない(100)”; “た・ところ・だ(110)”; “た・ばかり・の・ところ・だ(110) ”; “う・と・し・て・いる(100)”; “て・もらう(100)”; “べき・で・ない(100)”; “の・は・よく・ない(100)”; “なけれ・ば・なら・ない(111)”; “て・は・なら・ない(111)”; etc.

概念語的 MWE:
・**名詞的表現**<10,000>:“赤・の・他人(110) ”; “鶴・の・一声(110) ”; etc.
・**サ変名詞的表現**<1,700>:“もらい・泣き(110) ”; “ラッパ・飲み(110) ”; etc.
・**動詞的表現**<34,000>: “かみ・締める(110) ”; “煮・詰める(110) ”; etc.
・**形容詞的表現**<4,300>: “怒り・っぽい(010) ”; “注意・深い(110) ”; etc.
・**形容動詞的表現**<2,000>:“一巻・の・終わり(110) ”; “筋書き・通り(010) ”; etc.
・**副詞的表現**<5,200>: “悪く・する・と(110) ”; “うっとり・と(010) ”; etc.
・**連体詞的表現**<2,600>: “他愛・の・無い(101) ”; “ 断固・たる(010) ”; etc.
・**接続詞的表現**<300>: “その・結果(110) ”; “それ・は・さて・おき(111) ”; etc.
・**格言・諺（文）**<1,300>:“急が・ば・回れ(111) ”; “春眠・暁・を・覚え・ず(111) ”; etc.
・**格言・諺（不完全文）**<900>: “病・は・気・から(110) ”; “馬・の・耳・に・念仏(111) ”; etc.

2.3 MWE の出現頻度について

日常のテキストに MWE がどの程度生起するかについては、たとえば、(Sag et al., 2002)によると英語文(WordNet 1.7)における MWE の含有率は 41%であったと報告されている。
EDR 日本語コーパス(EDR, 1996)からランダムに抽出した 9210 個の述部（動詞、形容詞述語文節）について筆者らが行った調査によると、述部に使われた助述表現の 42%が上記 MWE であることが分かった。

2.4 変化形情報

一般に MWE には硬直性（rigidity）の高いものが多いが、その度合いは千差万別である。たとえば、“ 石橋を叩いて渡る ” は “ 石橋を叩いても渡らない ” と変形させて用いられるかもしれない。しかし、“ 叩いて渡った石橋 ” という表現は普通、「慎重に行う」という熟語的意味とは無関係である。従って、通常文の解析処理等に用いる資源としては前記の基本形 65,000 表現では不十分であると同時に、全ての文法的変化形を MWE として許すのでは誤った意味解析の原因となりうる。このように、可能な変化形だけをいかに忠実にデータ化しておくかが極めて重要な課題となる。筆者らは変化形情報として以下の 9

種 m_1 ~ m_9 を辞書中、各見出し表現に対して記載することでこの問題に対処している。

m_1: 連体修飾の可否

名詞を修飾する語句については、必須的に要求される場合、禁止される場合、その他の場合がある。また修飾句の形態に制約のある場合とない場合がある。さらに、修飾語句に対する意味上の制約も規定しておく必要がある。
例）○[必須]<感情>+顔をする
　　×　業を煮やす
　　○　<人>の+手に余る
（<X>は概念 X を表す語を表す。）

m_2: 連用修飾の可否

述語を修飾する語句についても、必須的に要求される場合、禁止される場合、その他の場合が考えられる。また、動詞が本来取らなかった格が新たに取られる場合や、その逆の場合もある。これらについても通常の動詞に対する各要素の規定と同様のことを行なっておく必要がある。
例）○[必須]　<人>に+バトンを渡す
　　○　<人>と+手を切る

m_3: 助動詞の付加の可否

述語性の慣用表現の場合、特定の助動詞等が必須的に要求される場合や禁止される場合がある。
例）○[必須]　手がつけられない
　　×　頭が切れている

m_4: 副助詞の付加の可否

副助詞等の付加も場合によって許されたり許されなかったりする。
例）○　気が気ではない
　　×　喉からは手がでる

m_5: 助詞の削除の可否

含まれる助詞を取り去っても慣用表現として働くものがある。助詞の削除が可能であるか否かを明らかにする必要がある。また、削除した結果の表現と複合動詞、複合形容詞等との区分も、少なくとも取扱い上明確にしておくことが必要である。
例　○　気味が悪い

m_6: 助詞の置換の可否

「副助詞の付加」と類似の現象として格助詞を副助詞で置き換え得るかどうかも表現ごとに異なる。
例）○　諦めを付ける⇔諦めが付く

m_7: 語順の変更の可否

格助詞等の順序を入れ換えてもよい場合と悪い場合がある。
例）○　横から口を出す⇒口を横から出す
　　×　足が棒になる⇒棒に足がなる

m_8: 倒置による体言化の可否

含まれる格要素としての名詞を末尾に回して連体被修飾語にかえる事が許される場合と許されない場合がある。
例）○　足が棒になる⇒棒になった足
　　×　足が出る⇒出る足
　　（「赤字が出る」の意）

m_9: 受身化・使役化の可否

受身化や使役化が許される場合と許されない場合がある。
例）○　金が物を言う⇒金に物を言わせる
　　×　へそが茶をわかす⇒へそに茶をわかさせる

このように、我々の MWE データベースは単なる MWE 基本候補のリストではなく、各表現固有の特異性を詳細に記述した言語資源となっている。(Shudo et al., 1980, 1988; 首藤, 1989; 安武他, 1997)。データベースは現時点では未完成であるが、部分的成果を用いた考察、予備的実験は(Koyama et al., 1998; 岩瀬ら, 2001; Shudo et al., 2004)等で報告している。

2.5　その他の収録情報

以上のほか、各 MWE 見出しには以下の情報が付与される。

１．切れ目情報
２．字種情報

３．構成要素の文法カテゴリー情報
４．表現の文法カテゴリー情報（2.2 参照）
５．頻度情報
６．その他

主として１～２は形態素解析、３～４は構文解析処理とのリンクのために必要な情報、５はタスクに応じた資源の分割に有効であると考えている。６は、たとえば動詞相当表現に対する格情報など、単語辞書と同等な利用を考える際に必要となる諸情報であるが現在未採録である。

3 おわりに

本稿では筆者らが公開を目標に現在構築を進めている日本語 MWE データベースの特徴、概要を紹介した。変化形情報をはじめとした情報の付与を完成させることが現在のもっとも重要な作業課題である。

本データベースを利用した機械処理における大きな課題は、たとえば、MWE として登録されている表現が実際熟語として用いられているのか、それとも語句本来の意味として用いられているのかを判別する問題である。たとえば、“水をさす”は「邪魔をする」に近い熟語としてもまた、本来の意味でも使われる。この問題の本質的な解決には文脈、知識にかかわる今後の研究の蓄積を待たねばならないが、この種の表現であるか否かの区別をデータベース中に記述しておくことは必要であろう。また、MWE データベースには既存の機械処理システムとのインターフェース機能も望まれる。たとえば、MWE “水をさす”の非 MWE による言い換え「邪魔をする」を与えておくことによって既存の機械処理システムとリンクさせられれば、既存システムの本質的な能力向上にも貢献できると考えられる。

参考文献

岩瀬修, 森元逞, 首藤公昭. 2000. 連語を組み込んだ統計言語モデル. 電子情報通信学会第34回音声言語情報処理研究会: SP2000-113: pp.109-114.

首藤公昭, 楢原斗志子, 吉田将. 1979. 日本語文の標準形変換に関する考察. 電気関係学会九州支部連合会大会論文集, p329

首藤公昭, 吉村賢治, 武内美津乃, 津田健蔵. 1988. 日本語の慣用的表現について– 語の非標準的用法からのアプローチ – 自然言語処理研究会 NL-66-1: pp.1-7.

首藤公昭. 1989. 日本語における固定的複合表現. 文部省科学研究費補助金特定研究(Ⅰ), 課題番号 63101005.

宮崎正弘, 池原悟, 横尾昭男. 1993. 複合語の構造化に基づく対訳辞書の単語結合型辞書引き，情報処理学会論文誌，Vol.34, No.4, pp.743-754

森田良行, 松木正恵. 1989. 日本語表現文型. アルク.

安武満佐子, 小山泰男, 吉村賢治, 首藤公昭. 1997. 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: pp449-452.

EDR(日本電子化辞書研究所). 1996. EDR電子化辞書 . http://www.iijnet.or.jp/edr/

Iwan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. *Multiword Expressions: A Pain in the Neck for NLP*. The Proc. of the 3rd CICLING: pp.1-15.

Kosho Shudo, Toshifumi Tanabe, Masahito Takahashi and Kenji Yoshimura. 2004. *MWEs as Non-propositional Content Indicators*. The Proc. of the Workshop on Multiword Expressions at 42nd Annual Meeting of the ACL: pp.32-39.

Kosho Shudo, Toshiko Narahara and Sho Yoshida. 1980. *Morphological Aspect of Japanese Language Processing*. The Proc. of the 8th COLING: pp.1-8.

Satoshi Shirai, Satoru Ikehara and Tsukasa Kawaoka. 1993. *Effects of Automatic Rewriting of Source Language within a Japanese to English MT System*. Fifth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-93: pp.226-239

Yasuo Koyama, Masako Yasutake, Kenji Yoshimura and Kosho Shudo.1998. *Large-Scale Collocation Data and Their Application to Japanese Word Processor Technology*. The Proc. of the 17th COLING: pp.694-698.