日本語慣用句用例データベースの構築法

尾嶋 憲治 † 佐藤 理史 ‡ 宇津呂武仁 *

† 京都大学工学部電気電子工学科
‡ 名古屋大学大学院工学研究科電子情報システム専攻
* 京都大学大学院情報学研究科知能情報学専攻
{ojima,utsuro}@pine.kuee.kyoto-u.ac.jp, ssato@nuee.nagoya-u.ac.jp

1. はじめに

我々が日頃用いている日本語には、「慣用句」と呼ばれる表現が多く含まれている。たとえば、次のような例文を考えよう。

(1) よい材料がすぐに 手に入った。
(2) 美しい絵が 目に入った。
(3) 彼は 石橋を叩いて渡る 性格だ。

ここで、例文 (1) の「手に入る」、例文 (2) の「目に入る」、例文 (3) の「石橋を叩いて渡る」が慣用句と呼ばれている表現であり、それぞれ、おおよそ「入手できる」、「見える」、「用心深い」という意味を表す。

しかしながら、我々は、このような慣用句の振舞いを十分に知っているとはいい難い。たとえば、次のように、「手に」と「入る」の間に副詞が入ることは可能なのだろうか。

(4) 手に すぐに 入る なら、それが欲しい。

あるいは、次の文の「目に入る」は、慣用句としての意味ではなく、文字通りの意味と解釈されるのが普通であるが、これは、どのように説明されうるのであろうか。

(5) ゴミが 目に入った。

現在、慣用句に関する情報を得るための資料として広く利用可能なものは、国語辞典および慣用句辞典である。しかしながら、これらの辞書には、一つの慣用句に対して、高々一つの用例が書かれているにすぎず、その慣用句が実際にどのように使われているかを知るには、あまりにも情報が少なすぎる。また、「目に入る」のように、慣用句としての意味以外に、文字通りの意味を持つ場合もある慣用句に対して、その事実や、例文 (5) のような負例（慣用句の解釈をとらない用例）を明示的に示しているものは、ほとんど皆無である。

機械翻訳などの応用を考えれば、テキスト中に現れる慣用句を正しく認定することは、自然言語処理において不可欠であることは明らかである。しかしながら、現在広く使われている日本語処理ツールでは、慣用句の認定処理は場当たり的で、体系的な処理が実現されているとは言い難い。

慣用句を認定する処理を適切に設計するためには、それぞれの慣用句の振舞いを十分に知ることが必要である。たとえば、

- もし、慣用句が固定的であり、分離や変形等の操作を受けないのであれば、慣用句をひとまとまりの単位として処理するのが適切であろう。一方、分離・変形等の操作を受けるのであれば、その点を考慮する必要が生じる。
- もし、慣用句が文字通りの意味を持たないのであれば、慣用句の認定は存在の有無の判定となる。一方、文字通りの意味の可能性を持つのであれば、そのような曖昧性を解消することが慣用句の認定には不可欠となる。

このような背景より、我々は、慣用句の機械処理の研究を進めるためには、日本語慣用句の用例データベースを作成することが不可欠と考え、これを実現することに取り組んでいる。その第一ステップとして、用例データベースの基本設計を行ない、その構築法を明らかにすることを行なった。ここでの構築法とは、データベース構築の手順を明確にし、そこで生じる各種の問題点について現実的な解を与え、最終的に作業マニュアルという形で明文化したものを指す。本稿では、これらの内容について述べる。

2. 用例データベースの基本設計

2.1 ３つのねらい

日本語慣用句用例データベースを設計するにあたり、まず、最初に考えたのは、つぎの３点である。

(1) 個々の慣用句の用例が観察できるような資料を作成する。

言語資料には、(a) ひとまとまりのテキストに各種の情報を付与したもの（コーパスと呼ばれることが多い）と、(b) ある特定の言語単位や言語現象に着目してその用例（文単位であることが多い）を集めたもの、の２種類が存在する。ここでの我々の興味は、慣用句に限定されているので、後者（用例集）の形態をとるのが妥当である。但し、慣用句の解釈は、それを含む文内で定まらない場合もあるので、より広い文脈を見ることができるような機能が必要である。

(2)　慣用句の認定等の機械処理を実現する際に役立つ資料を作成する。

人間のための用例集であれば、正例（慣用句としての意味で使われている用例）だけを集めるのが一般的であるが、機械処理の実現においては、正例と負例（表層上は慣用句とも考えられるが、文字通りの意味で使われている例）の区別が重要となる。そのため、正例だけでなく負例も収集の対象とする。

(3)　比較的整ったテキストから用例を収集する。

用例は色々なテキストから収集すべきであるが、まずは、標準的と考えられる用例から重点的に収集するという戦略も妥当であろう。著作権等の権利関係の問題も勘案し、今回は、新聞記事から用例を収集することとした。

2.2　３つの問題に対する基本方針

以上のような基本方針下で用例データベースを作成することを考えた場合、特に大きな問題となる点は、次の３点である。

(1)　対象とする慣用句の範囲をどのように定めるか。

基本方針としては、よく使われる慣用句や誰もが知っている慣用句を収集対象としたいと考えている。しかしながら、何を慣用句とみなすかということは、それほど明確ではなく、また、基本語彙に対応するような基本慣用句のリストもほとんどない [1)]。

そこで、まずは、この問題を棚上げし、対象とする慣用句が与えられたとき、その慣用句に対する用例データをどのように作成するかという問題に専念することにした。

(2)　「１つの慣用句」の範囲をどのように定めるか。

この問題は、どのような範囲の異形や変形を同一慣用句とみなすかという問題である。形と意味に共通性が見られることが、同一慣用句とみなす基本であり、異表記、活用、分離、接辞の付加等がその具体例となる。この範囲を明確に定めることが、作業マニュアルの一つの大きな柱となる。

(3)　語義をどのように設定するか。

正例だけでなく、負例も集めるということは、用例に、正負のラベルを振るということである。先に示した「目に入る」を例にとると、正例：「見える」、負例：「（物理的に物が）目に入る」を区別することである。これは、「『目に入る』に２つの語義があり、どちらの語義で使われているかを判断し、その情報を付与する」とみなすことができる。

慣用的意味をとる可能性がある表現（以下では、誤解のおそれがない場合は、便宜的に「慣用句」と呼ぶ）が、１つの慣用的語義と１つの文字通りの意味をもつ場合は単純であるが、複数の慣用的語義を持つ慣用句も存在する。そのため、単に正負のラベルを付与するのでは不十分で、語義ラベルを付与するのが望ましい。

このような語義ラベル付与のためには、語義のセットを定義する必要があり、それをどのように定めるかが大問題となる。

ID		C1010
見出し語		手が届く
主要情報	語構成	手-が-届く
	型	名詞-格助詞-動詞
	品詞	動詞相当
	活用語	届く

図 1　主要情報の付加例【手が届く】

我々は、この問題に対して、現実的な方針を採用した。まずは、既存の辞書に基づいて語義セットを定め、そのセットを用いて用例に対する語義ラベル付与の作業を行なう。その過程で、もし、そのセットに不備があることがわかった場合は、それを改訂する。

2.3　用例データベースの基本構造

以上のような検討の後に、作成する用例データベースの基本構造を設計した。用例データベースは、項目の集合から構成される。ここで、項目とは１つの慣用句（見出し語）に対するデータの総体を表しており、(1) 慣用句ID、(2) 見出し語、(3) 主要情報、(4) 語義情報、(5) 表示、および、(6) 用例、の６つの要素から構成される。これらの詳細については、次節以降で述べる。

3.　用例データベース作成手順

用例データベースの作成手順を定める際、日本語学に関して、専門のトレーニングを受けていなくとも作成作業が行なえるようなマニュアルを作成することを目標においた。一つの慣用句に対する一連のデータを作成する具体的な作成手順は、以下の通りである。

3.1　見出し語の選定

与えられた慣用句に対して、次の３冊の辞書に掲載されている表記を、その慣用句の見出し語（代表的な表記）として選定する。

- 大辞林（第二版）　［三省堂］
- 広辞苑（第五版）　［岩波書店］
- 大辞泉（増補・新装版）　［小学館］

これらの辞書の表記が一致しない場合は、そのなかで最も一般的に使用されていると考えられる表記を選定する。

3.2　主要情報の付与

主要情報は、次の４つの情報を付与する。

(1)　語構成：見出し語（代表表記）に対して、語の区切りに “-” を挿入した形式で示す。

(2)　型：構成する語の品詞を “-” でつないだ形式で示す。

(3)　品詞：慣用句全体が、どのような品詞相当となるかを記述する。

(4)　活用語：慣用句の末尾が活用する語である場合、その語を記述する。

図１に付与例を示す。

3.3　語 義 収 集

語義収集には、3.1 節に示した３つの辞書を用いる。辞書により語義の線引きが異なる場合は、全体としてなるべく細分化されるように各辞書から語義を収集する。文

01L	（実際に）目が回る。
02I	非常に忙しい。「目が回るほどに忙しい」
03G	目がくらむ。めまいがする。

図 2 語義の収集例【目が回る】

字通りの語義があると認められる場合は、辞書の掲載の有無に関わらず、それを一つの語義として取り上げる。辞書に例文が掲載されている場合、例文もあわせて収集する。

こうして収集した各語義に対して、語義 ID を付与する。語義 ID は、通し番号 2 桁と英字 1 文字で構成する。英字としては、次の 3 種類を用いる。

L その語義が、文字通りの意味と判断できる場合 (Literal)

I その語義が、慣用的な意味であると判断できる場合 (Idiomatic)

G 上記 2 つに当てはまらない場合 (Gray)

語義の収集例を図 2 に示す。

3.4 表示の選定

その慣用句を表しうる「表示」（表記と読みの組）[2] を収集する。表記、あるいは、読みのいずれかが異なる場合は、別の表示とする。慣用句の末尾の語が活用する場合、その語が基本形となる場合のみを表示として採用する。次のような異形は、表示のバリエーションとみなす。

(1) いわゆる異表記
- 漢字表記とかな表記
  - 「頭にくる」（頭に来る）
  - 「折り合いを付ける」（折り合いをつける）
- 漢字表記の異なるもの
- 送り仮名が異なるもの

(2) 助詞の交替
- 格助詞「が」→「の」「は」「も」
  - 「気の弱い」（気が弱い）
- 格助詞「を」→「は」「も」
  - 「手は汚す」（手を汚す）
- 格助詞「に」→「には」「にも」
  - 「気にもなる」（気になる）

(3) 助詞の交替に加え、名詞、形容詞の語幹に接辞がつくことで、相当品詞が変わるもの
- 「気の弱さ」（名詞相当；気が弱い）
- 「お気に入り」（名詞相当；気に入る）

(4) 否定の接尾辞・助動詞の交替
- 「歯が立たず」（歯が立たない）

上記に示すように、助詞の交替は異表示とみなすが、助詞の欠落は異表示とはみなさない（同一慣用句とはみなさない）。
- 「ご機嫌取り」（機嫌を取る）–「を」が欠落。
- 「気弱さ」（気の弱い）–「の」が欠落。

こうして収集したそれぞれの表示に、表示 ID を付与する。表示 ID の上位 2 桁は読み ID を、下位 2 桁は表記 ID を表す。なお、採用した「表示」の相当品詞が見出し語と異なる場合は、相当品詞を明記する。表示の選定例を図 3 に示す。

手を汚す	0101	てをよごす	手を汚す
（動詞相当）	0102	てをよごす	手をよごす
	0201	てはよごす	手は汚す
気が弱い	0101	きがよわい	気が弱い
（形容詞相当）	0201	きのよわい	気の弱い
	0301	きのよわさ	気の弱さ（名詞相当）

図 3 表示の選定例

3.5 用例収集

各表示に対し、用例収集ツール（文字列による全文検索、および、形態素列検索）を利用して、用例を収集し、用例 ID を付与することにより整理する。用例 ID は、項目 ID-表示 ID-xxx（3 桁の数字）で構成する。収集を行う際には、各表示に含まれる活用語の活用も考慮し、活用した形も用例として集める。用例は、CD-毎日新聞 95 版の記事データから文単位で収集する（すべての文には、あらかじめ記事番号と文番号を付与した）。なお、用例が 50 件以上含まれる場合には、ランダムに 50 件を選ぶ。

3.6 語義ラベル付与

収集した各用例に対し、どの語義で用いられているかを判定し、3.3 節で定めた語義 ID を用いてラベルづけする。必要な場合は、前後の文脈も考慮し、最も適切な語義を決定する。語義のラベルづけができない場合、次に定める特殊ラベルのいずれかを付与する。

B 対象文字列が、語義を決定する単位として適切でない場合。
- 「相手を抜く」（手を抜く）

N 対象文字列が、先行する語との係り受け関係により、見出しと一致しないと認められる場合。
- 「その気になる」（気になる）

Y 対象文字列の読みが、該当する見出しの読みと異なる場合。（このラベルに該当する例は見つかっていないが、存在する可能性がある。）

U 対象文字列が、設定した語義セットには含まれず、新しい語義を持つと考えられる場合。ただし、実際に付与する場合は語義の分類 (L, I, G) に準拠した UL, UI, UG のラベルを用いる。

X 対象文字列は語義を決定する単位として適切ではあるが、語義が判定できない場合。

3.7 備考の付与

慣用句の直後に否定の接尾辞、助動詞が現れる場合、備考欄に「否定」と記述する。慣用句が末尾に否定の接尾辞・助動詞を含んでいる場合は、さらにその直後に否定の接尾辞・助動詞が現れている場合に、備考欄に「否定」を記述する。

4. 用例データベース作成の実際

4.1 第 1 サイクル

我々は、2 節に述べた基本方針を定めた後、手探りの状

態で、用例データベースの作成を開始した。まず、21 個の述語型慣用句（表 2）を選び、その用例データを作成する作業を通して、データベースの細部と作業マニュアルを確定させていくことを行なった。同時に、作業に必要な各種ツールの作成を行なった。

このサイクルで作成したデータベースの概要を表 1 に示す。データベースのマスターデータは、XML 形式のファイルであり、1つの用例データは、(1) 用例 ID、(2) テキスト、(3) 対象文字列の範囲、(4) 語義ラベル、(5) 備考、から構成されている。ここで、対象文字列の範囲は、仮定する語の単位（文法）に依存するが、現在は、IPA 体系 [3] を採用している。

実際にデータベースを見る場合は、XML 形式のマスターデータをコンバータで HTML 形式に変換する。HTML 形式のデータベースでは、毎日新聞のオリジナルデータへのハイパーリンクを提供しており、その用例が現れた前後の文脈を簡単に見ることができる。

4.2 第2サイクル

次に、第1サイクルで確定させた作業マニュアルと各種ツールを用いて、31 個の慣用句の用例データを作成することを行なった。このサイクルの目的は、作業マニュアルの問題点を洗い出すことである。第1サイクルで対象とした慣用句とはタイプが異なる慣用句を積極的に選んだ（表 2)。このサイクルでは、次のような問題が発生しため、マニュアルの増補を行なった。

(1) 3つの辞書のいずれにも、慣用句が見出しとして掲載されていない。例：「難色を示す」
→このような場合は、慣用句辞典を調べ、見出し語を選定する。

(2) 見出しとなる表現が、3つの辞書で一致しない場合。例：「右に出るものがいない」（大辞林)、「右に出る者がない」（広辞苑）
→多数決をとる。同点の場合は、大辞林＞広辞苑＞大辞泉の順に優先する。

(3) 負の用例しか収集できなかった場合。例：「たかも知れる」（「高が知れる」）－すべての用例が「〜たかも知れない」の形をとる。
→用例データとして残す。

(4) 対象文字列の範囲を自動的に決定するのに失敗する場合。例：「穴を開け、...」（穴をあける）－「開け（あけ)」を「開け（ひらけ)」と誤解析する。
→用例データをチェックし、手作業で修正する。

5. おわりに

これまでの研究により、慣用句用例データベースを作成するための構築法（作業手順）は、かなり明確になってきたと考えている。次のサイクルは、第3者に用例データ作成の作業を依頼し、作業マニュアルのさらなる問題点を洗い出し、作業マニュアルの第1版を確定したいと考えている。また、これと並行して、2.2 節で棚上げした問題、すなわち、データベースに収録する慣用句のリストの策定を行なっていく予定である。

表 1 作成したデータベースの概要

		サイクル	
		第 1	第 2
項目数		21	31
語義数	総計	37	47
	xxL	7	9
	xxI	27	34
	xxG	3	4
	1 項目あたり	1.8	1.5
複数語義を持つ項目数		12	11
表示数	総計	53	69
	1 項目あたり	2.5	2.2
用例数	総計	765	825
	xxL	33	79
	xxI	635	534
	xxG	26	105
	B	37	105
	N	22	1
	U	11	0
	X	1	1
	1 表示あたり	14.4	12.0
	1 項目あたり	36.4	26.6

表 2 対象とした慣用句

第1サイクル (21 個)		
足を洗う	頭が痛い	頭に来る
後味が悪い	折り合いをつける	顔に泥を塗る
気が弱い	機嫌を取る	気に入る
気になる	手が届く	手に入れる
手を抜く	手を焼く	手を汚す
歯が立たない	火に油を注ぐ	耳にする
目が届く	目が回る	目にする
第2サイクル (31 個)		
胡座をかく	汗を流す	頭が固い
穴をあける	息を引き取る	石橋を叩いて渡る
一日千秋の思い	一本取られる	掛け値なし
肩身が狭い	金に糸目をつけない	借りてきた猫
軌道に乗る	気に食わない	漁夫の利
口を揃える	桁が違う	私腹を肥やす
十人十色	関の山	高が知れる
力を入れる	難色を示す	猫に小判
日が浅い	腑に落ちない	右に出る者がいない
道が開ける	見て見ぬ振り	胸をふくらませる
役に立つ		

謝辞 本研究では、CD-毎日新聞 95 版の記事データを使用した。本研究の一部は、独立行政法人情報通信機構からの受託研究「円滑なコミュニケーションのための言語処理基盤に関する基礎研究」の支援を受けた。

参 考 文 献

1) 宮地裕 編. 慣用句の意味と用法. 明治書院, 1982.
2) 佐藤理史. 異表記同語認定のための辞書編纂. 情報処理学会自然言語処理研究会, 2004-NL-161, pp.97-104, 2004.
3) 浅原正幸, 松本裕治. ipadic version 2.6.3 ユーザーズマニュアル. 奈良先端科学技術大学院大学, 2003.