

# 日本語複合辞用例データベースの作成

Development of An Example Database of Japanese Compound Functional Expressions

土屋 雅稔\*, 宇津呂 武仁†, 松吉 俊†, 佐藤 理史‡, 中川 聖一‡

\* 豊橋技術科学大学  
情報メディア基盤センター

† 京都大学  
情報学研究科

‡ 名古屋大学  
工学研究科

‡ 豊橋技術科学大学  
工学部情報工学系

## 1. はじめに

複合辞とは、幾つかの語が複合してひとまとまりの形となって非構成的な意味を持ち辞的な機能をはたす表現である<sup>1)</sup>。この複合辞に対して、それと同一表記をとり、内容語としての構成的な意味をもつ表現が存在することがある。例として、次の2つの文を翻訳する場合を考える。(1) 私は彼について話した。→ I talked about him. (2) 私は彼について走った。→ I run following him. この2つの文には「について」という表現が共通して現れているが、文(1)の「について」は前置詞 about と翻訳されるのに対して、文(2)の「について」は動詞 follow の現在分詞に翻訳される。このように区別して翻訳するには、文(1)中の「について」をひとかたまりの複合辞として検出すると同時に、文(2)中の「ついて」は内容語として用いられていることを検出して、それぞれの文の構造を正しく区別しておく必要がある。

しかし、既存の解析系における複合辞の取り扱い是不十分である。例えば、形態素解析器 JUMAN と構文解析器 KNP の組み合わせは、形態素解析時には機能表現を検出していない<sup>☆</sup>。構文解析時に、解析規則に記述された特定の形態素列が現れると、直前の文節の一部として処理したり、直前の文節からの係り受けのみを受けけるように制約を加えて、複合辞である可能性を考慮した解析を行っている。文(1)、(2)はいずれも、直前の文節からの係り受けのみを受けけるように制約を加えて構文解析され、それぞれの文の構造は区別されない。他の例として、形態素解析器と構文解析器 CaboCha の組み合わせを考える<sup>☆☆</sup>。その際、構文解析器は、京都テキストコーパス<sup>2)</sup>から機械学習したモデルを用いるとする。この場合、形態素解析用辞書に「助詞・格助詞・連語」と登録されている複合辞は、形態素解析時に検出される。また、「ざるを得ない」などの表現は直前の文節の一部としてまとめられ、機能的な表現として解析される。文(1)、(2)はともに「助詞・格助詞・連語」の「について」と解析され、2つの文の構造は区別されない。

以上より、複合辞は、日本語の文構造を把握する時に重要な役割を果たしているにも関わらず、従来の自然言

語処理における複合辞の取扱いは不十分であることがわかる。このような現状を改善するには、複合辞である可能性がある形態素列が現れたときに、非構成的な複合辞として用いられているか、その形態素列本来の構成的な意味で用いられているかを区別する検出器が必要である。本稿では、そういった検出器を作成するための基礎資料として、複合辞と同一の形態素列が、非構成的な意味の複合辞として用いられている用例だけでなく、その形態素列本来の構成的な意味で用いられている用例を含み、かつ、それぞれの複合辞について十分な数の用例を含む複合辞用例データベースを作成する手順を提案する。

## 2. 複合辞リストの作成

### 2.1 現代語複合辞用例集

現代語複合辞用例集<sup>3)</sup>(以下、**用例集**と呼ぶ)は、以下の5種類に分類された125項目(図1)の解説からなっている。

#### A. 基本的に活用しない「助詞的複合辞」

**接続辞類** 基本的に節を受けて、複文前件を形成するもの(「～とはいえ」など36項目)。

**連用辞類** 基本的に名詞を受けて、述語にかかる成分を形成するもの(「～について」など45項目)。

**連体辞類** 名詞や節などをを受けて、連体修飾句を形成するもの(「～といった」など2項目)。

**文末辞類** 文末に付加されて、話し手のコミュニケーション上のさまざまな気持ちを示すもの(0項目)。

#### B. 述語の部分に付加されて活用する「助動詞的複合辞」

(「～つもりだ」など42項目)

本稿では、この5分類に、前の文を後ろの文に関係付ける働きをする**接続詞類**を加えて、複合辞を6つに分類する。

用例集の収録対象には、日本語表現文型(450表現)と日本語文型辞典<sup>4)</sup>(965表現)に収録されている複合辞の一覧対照表を作成した上で、日本語表現文型に収録されている複合辞を基本とし、その中でも1つの複合形式として熟合度が高く、また一般性も高いと判断される複合辞が選ばれている。したがって、用例集に収録されている複合辞は、他の複合辞リストよりも少ないが、主要な複合辞は網羅されていると考えられる。また、用例集には、用例データベースの作成時の判定基準として有用な、丁寧に記述された解説文と比較的多数の用例(平均16.6文/項目)が含まれている点も好都合である。この2点より、本稿では、用例データベースを作成する時に参照す

<sup>☆</sup> <http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html>

<http://www.kc.t.u-tokyo.ac.jp/nl-resource/KNP.html>

<sup>☆☆</sup> <http://chasen.naist.jp/hiki/ChaSen/>

<http://chasen.org/~taku/software/cabocha/>

◇ A56 〜にとって・〜にとり

**接続** 名詞（名詞節を含む）に付く。

**意味・用法** 「A にとって B」という形で文の内容を規定する形で用いられ、「A にとって B」が係っていく文の内容として述べられる個別的な判断・とらえ方を表す主体を表す。

**用例** (1) 技術的な問題（拡大・縮小や、ゆがみ、雑音など）はいろいろありますが、コンピュータにとって「原理的に不可能」とはいえませんが、（野崎昭弘「人工知能はどこまで進むか」）

**文法** 「にとり」という言い方も、いささかこちないがなお可能である。連体修飾の言い方としては、「にとる」とそのまま連体形にしては用いられないが、「にとつての」という形なら可能である。「にとりまして」という丁寧の形も取れる。とらえ方を表す主体という立場を強調した言い方として (17)(18) のように「〜にとってみれば」という形もある。

図 1 用例集の項目例

るリストとして、用例集を用いる。

## 2.2 収録対象とする複合辞の体系化

用例集では、意味的・機能的に似通った、ある範囲の異形は 1 つの項目でまとめて説明されている。例えば、図 1 の項目には、「〜にとって」や「〜にとり」などの異形が区別されずに、1 つの項目にまとめられている。

このような複合辞も全て明示的に列挙・体系化するために、項目を表記等に着目して細分した小項目という単位を設ける。それぞれの小項目には、以下のような意味を持つ 4 桁の枝番号と項目 ID を組み合わせた小項目 ID を付与する。

1 桁目:助詞の挿入や脱落および交替、同意語の交替などによって、表記の一部が異なっている異形を区別。

2 桁目:文体を区別。

0 = 常体, 1 = 敬体, 2 = 口語体

3 桁目:以下の表現を区別。

0 = 基本形, 1 = 連体修飾形,

2 = 否定の変化形, 3 = 否定形

4 桁目:1 桁目～3 桁目がまったく同じである複数の小項目を区別するための一意な番号。

例えば、図 1 の項目 (A56) は、以下の 5 つの小項目に分割される。

A56-1000: にとって

A56-1010: にとつての (←「〜にとって」の連体修飾形)

A56-1100: にとりまして (←「〜にとって」の丁寧形)

A56-2000: にとり

A56-3000: にとってみれば

ただし、1 つの小項目は、常に 1 種類の表記にだけ対応するわけではなく、複数の表記に対応する場合もある。「〜におうじ」と「〜に応じ」のように平仮名と漢字の違いは、1 つの小項目にまとめている。また、連用辞類に属する表現「〜について」などは、助詞「は」「も」が後続することによって提題助詞的または副助詞的に働くことがあるが、これらの表現を個別の小項目に区別することとはなかった。

最終的に、用例集に収録されている 125 個の項目から、(連続しない) 複数の要素の呼応という特別の形をとっている「A66 〜といい〜といい」および「A67 〜といわず〜といわず」を除いた 123 項目を 337 個の小項目に分割

用例 ID: A56-1000-003

収集元 ID: MNP-950115192-6

テキスト (下線部がターゲット文字列):

大阪・関西にとって試金石だと思う。

判定ラベル: F

備考: (なし)

図 2 用例データベース中の用例

し、収集対象として選定した。

## 3. 複合辞用例データベースの設計と作成

### 3.1 用例データベースの仕様

用例データベースは、項目、小項目、用例の 3 つの単位から構成されている。

項目は、(1) 見出し語と、(2) 項目 ID および (3) 1 つ以上の小項目からなる。見出し語と項目 ID は、用例集の項目に完全に準拠している。例えば、図 1 に準拠した項目では、見出し語は「〜にとって・〜にとり」、項目 ID は A56 である。

小項目は、(1) 小見出し語、(2) 小項目 ID および (3) 用例 (複数) からなる。小見出し語は、この小項目の可能な表示 (表記と読みの組) のリストである。

用例は、(1) 用例 ID、(2) 収集元の記事 ID、(3) テキスト、(4) ターゲット文字列、(5) 判定ラベルおよび (6) 備考からなる。図 2 に例を示す。用例 ID は、小項目 ID と用例を識別するための一意な自然数 (3 桁) からなる。収集元の記事 ID は、この用例のテキストを収集した記事を表す。本データベースの作成にあたっては、研究用に広く利用できること、および、大量のテキストが収集できることの 2 点を考慮して、毎日新聞を収集元とした。ターゲット文字列は、文字列のみに基づいて判断すると複合辞である可能性がある部分であり、テキストは、ターゲット文字列を含む文である<sup>☆</sup>。判定ラベルは、ターゲット文字列が文中において果たしている働きを手で判定した結果を表す。

### 3.2 判定ラベル体系

本データベースでは、表 1 に示した 6 種類の判定ラベルを設定している。

任意の文とターゲット文字列が与えられたとき、ター

<sup>☆</sup> 文は、句点を手がかりとして機械的に分割した。

表 1 判定ラベル体系

| 判定ラベル | 判定単位 | 読み  | 内容 vs 機能 | 用法            | 例文                                 |
|-------|------|-----|----------|---------------|------------------------------------|
| B     | 不適切  |     |          |               | (3) 震災直後にとりあえずスタッフを出動させることができ、…    |
| Y     | 適切   | 不一致 |          |               | (4) 法律上は困難でも、もう少し組織的に救援活動に参加する …   |
| C     | 適切   | 一致  | 内容的      | 内容的用法         | (5) まな板にとっていいに納豆のタタキを作りみそ汁の実にする …  |
| F     | 適切   | 一致  | 機能的      | 用例集で説明されている用法 | (6) 受験などでは倍率が上がったところで入学金があがることはない、 |
| A     | 適切   | 一致  | 機能的      | 接続詞的用法        | (7) ところで、全国の桜の名所では近年、樹勢の衰えが目立ち、…   |
| M     | 適切   | 一致  | 機能的      | その他の機能的用法     | (8) 浜ノ島はあと一步のところ勝ち星に結び付かず負け越した、    |

ゲット文字列の用法を判定することができる場合と、判定できない場合がある。本データベースの作成にあたっては、ターゲット文字列が1個以上の語、複合辞または慣用表現からなる列であるとき、そのターゲット文字列は判定単位として適切であり、用法を判定することができる。この条件を満たしていない用例には、ターゲット文字列が判定単位として不適切であることを表す判定ラベル B を付与する。例えば、文 (3) は、そのターゲット文字列が助詞「に」と副詞「とりあえず」の一部からなっているため、判定ラベル B が付与される。

判定ラベル Y は、ターゲット文字列の読みが、判定対象となっている小項目の読みと一致していないことを表す。例えば、「～うえは (A14-1000)」の用例として文 (4) を判定する場合、ターゲット文字列の読み「じょうは」と小項目の読み「うえは」が一致していないので、判定ラベル Y を付与する。

判定ラベル C は、ターゲット文字列に内容的に働いている語が含まれていることを表す。例えば、文 (5) のターゲット文字列中の動詞「とる」は本来の意味で内容的に働いているので、判定ラベルとして C を付与する。

判定ラベル F, A, M は、ターゲット文字列が機能的に働いているとき、その機能を区別するためのラベルである。判定ラベル F は、ターゲット文字列が用例集で説明されている用法で働いていることを表し、判定ラベル A は、ターゲット文字列が接続詞的に働いていることを表す。判定ラベル M は、これら以外の機能的な働きをしていることを表す。例として、「～ところで (A22-1000)」の用例として文 (6) ～ (8) を判定する場合を考える。文 (6) のターゲット文字列は、用例集で説明されている通りに逆接の働きをしているので、判定ラベル F を付与する。文 (7) のターゲット文字列は、文頭で接続詞的に働いているので、判定ラベル A を付与する。文 (8) のターゲット文字列は、形式名詞「ところ」を含めて機能的に働いているので、判定ラベル M を付与する。

### 3.3 作成手順

用例データベースの作成手順の概略は以下の通り。

- (1) 新聞記事から 50 文を収集。
  - (2) 作業による判定ラベル付与。
  - (3) 別の作業による判定ラベルの検証。
  - (4) 用例集の用法で用いられている用例の割合を検査。
  - (5) 用例集の用法で用いられていない用例の割合を検査。
- 最初に、毎日新聞 (1995 年) から複合辞が用いられて

いる可能性があるターゲット文字列を含む文を 2 通りの方法で収集する。第 1 に、既存の形態素体系と整合しないような文も収集するために、小見出し語を含む文を無条件に収集する。第 2 に、文中の活用語を 1 つずつ基本形に置き換えた文を生成し、その文に対して小見出し語が含まれているか調べて、小見出し語の末尾形態素が活用して用いられている文を収集する。収集された文が 50 文以上になった場合は、均等に 50 文を取り出して、判定ラベル付与の対象とする。

次に、収集された文を対象として作業による判定ラベル付与を行い、その結果を別の作業によって検証する。

続いて、検証を行った 50 個の用例に、判定ラベル F が付与された用例が 10 個以上含まれているか調べる。含まれていなかった場合は、判定ラベル F が付与された用例が、なるべくバランス良く含まれるように、以下の手順で 50 個の用例を補充する。用例集に記述されている接続制約を満たすターゲット文字列を含む文を 40 個、満たさないターゲット文字列を含む文を 10 個、新聞記事から収集し、判定ラベル付与と検証を行う。

更に、検証を行った 50 個の用例に、判定ラベル F 以外の判定ラベルが付与された用例が 10 個以上含まれているか調べる。含まれていなかった場合は、判定ラベル F の用例以外の判定ラベルの用例が、なるべくバランス良く含まれるように、以下の手順で 50 個の用例を補充する。用例集に記述されている接続制約を満たすターゲット文字列を含む文を 20 個、満たさないターゲット文字列を含む文を 30 個、新聞記事から収集し、判定ラベル付与と検証を行う。

## 4. 複合辞用例データベースの評価

### 4.1 基本的な統計

毎日新聞 (1995 年) からの用例収集結果を表 2 に示す。用例集の項目を単位とすると、114 項目について 50 個以上の文が収集された。小項目を単位とすると、187 小項目について、50 個以上の文が収集された。

毎日新聞 (1995 年) から 50 文以上が収集された 187 小項目について、人手による判定ラベル付与と検証を行った。187 小項目の 9350 用例の内、判定ラベル F が付与された用例 (用例集の用法と判定された用例) は 6271 個 (67.1%) だった。187 小項目を、判定ラベル F が付与された用例の出現率  $r$  によって分類した結果を表 3 に示す。出現率  $r$  は次式によって定義する。

| 表 2 新聞記事から収集された文数 |           |           |
|-------------------|-----------|-----------|
|                   | 項目数       | 小項目数      |
| 50 ≤ 文数           | 114 (93%) | 187 (55%) |
| 0 < 文数 < 50       | 9 (8%)    | 117 (35%) |
| 文数 = 0            | 0 (0%)    | 33 (10%)  |
|                   | 123       | 337       |

| 表 3 判定ラベル F の出現率        |          |
|-------------------------|----------|
| 出現率 $r$                 | 小項目数     |
| $r = 100\%$             | 61 (33%) |
| $80\% < r < 100\%$      | 38 (20%) |
| $20\% \leq r \leq 80\%$ | 55 (29%) |
| $r < 20\%$              | 33 (18%) |
| 計                       | 187      |

| 表 4 補充した小項目における判定ラベル F の出現率 |          |
|-----------------------------|----------|
| 出現率 $r$                     | 小項目数     |
| $r = 100\%$                 | 40 (31%) |
| $80\% < r < 100\%$          | 37 (28%) |
| $20\% \leq r \leq 80\%$     | 43 (33%) |
| $r < 20\%$                  | 12 (8%)  |
| 計                           | 132      |

| 表 5 作業者間の判定の一致度 (全 184 小項目) |                       |       |          |                       |                          |                    |  |
|-----------------------------|-----------------------|-------|----------|-----------------------|--------------------------|--------------------|--|
| 判定内容                        | 判定ラベル                 | 平均値   |          | 小項目数                  |                          |                    |  |
|                             |                       | $P_a$ | $\kappa$ | $0.8 < \kappa \leq 1$ | $0.67 < \kappa \leq 0.8$ | $\kappa \leq 0.67$ |  |
| ターゲット文字列は判定単位として適切である       | Y or C or F or A or M | 0.97  | 0.77     | 126 (69%)             | 11 (6%)                  | 47 (26%)           |  |
| ターゲット文字列は機能的に働いている          | F or A or M           | 0.93  | 0.73     | 120 (65%)             | 19 (10%)                 | 45 (25%)           |  |
| ターゲット文字列は用例集の用法で働いている       | F                     | 0.96  | 0.85     | 144 (78%)             | 14 (8%)                  | 26 (14%)           |  |

$$r = \frac{\text{判定ラベル F が付与された用例数}}{\text{用例数}} \quad (1)$$

表 3 より, 99 小項目は判定ラベル F 以外の判定ラベルが付与された用例が不足しており, 33 小項目は判定ラベル F が付与された用例が不足している. そのため, これらの 132 小項目について, 毎日新聞 (1991 年~1999 年) から補充を行った. 補充した結果を表 4 に示す.

#### 4.2 判定ラベル付与作業の一致度

2 人の作業者が独立に判定ラベル付与作業を行った場合に, 判定ラベルが一致する割合を検討する. 一致度を測る尺度としては,  $\kappa$  値が良く用いられる.

$$\kappa = \frac{P_a - P_e}{1 - P_e} \quad (2)$$

ここで,  $P_a$  は 2 人の作業者の判定が実際に一致した割合,  $P_e$  は 2 人の作業者の判定が偶然に一致する確率である. Carletta<sup>5)</sup> は,  $\kappa > 0.8$  の場合の判定結果は完全に信頼でき,  $0.67 < \kappa < 0.8$  の場合の判定結果は概ね信頼できると報告している.

新規作業者に, 補充収集の必要がなかった 55 小項目から特に判定が困難な 2 小項目を除いた 53 小項目と, 補充収集を行った 132 小項目から特に判定が困難な 1 小項目を除いた 131 小項目の判定を依頼し, 筆者らが検証した判定ラベルと比較した. 結果を表 5 に示す. 表より, 多くの小項目 (75% ~ 85%) について, 作業者による判定結果は信頼できることがわかる.

$\kappa \leq 0.67$  となった小項目について, 判定結果が一致しない原因を手で分析した. その結果, 一致度を下げている原因 (慣用表現や用例集には掲載されていない複合辞など) を特定できている, 判定ラベル付与の作業マニュアルに方針を明示することにより, 一致度を改善することができるという見通しを得ている.

## 5. おわりに

本論文では, 現代語複合辞用例集に収録されている複合辞を対象として, 複合辞用例データベースの作成手順を提案した. 複合辞とは, 複数の形態素がひとかたまりとなって, ひとつの機能語相当語として働く表現である. データベースの仕様と作成手順の設定にあたっては, 複合

辞と同一の形態素列が構成的な本来の意味で用いられている用例と, 非構成的に複合辞として用いられている用例の双方が適度に収集されるように配慮した. さらに, 実際に複合辞用例データベースを作成して, このようなデータベースが作成可能であることを示した. 現在, このデータベースを利用して複合辞検出器を作成中である<sup>6),7)</sup>.

このデータベースを分析すると, 新聞上において用法が極端に偏っている表現があることが分かる. そのため, 今後の課題として, 新聞記事以外の言語資源を利用して用例を補充することが考えられる. また, 日本語話し言葉コーパス<sup>8)</sup>と比較することにより, 新聞上での複合辞の使われ方と, 話し言葉での複合辞の使われ方を比較することができる.

本研究で作成したデータベースは, 筆者らのウェブサイトにて公開する予定である.

**謝辞** 本研究の一部は, 次の研究費による: 文部科学省 科学研究費 基盤研究 (A) 「円滑な情報伝達を支援する言語規格と言語変換技術」(課題番号 16200009), 京都大学-NTT コミュニケーション科学基礎研究所共同研究「グローバルコミュニケーションを支える言語処理技術」.

## 参 考 文 献

- 森田良行, 松木正恵: 日本語表現文型, NAFL 選書 5, アルク (1989).
- 黒橋禎夫, 長尾真: 京都大学テキストコーパス・プロジェクト, 言語処理学会第 3 回年次大会発表論文集, pp. 115-118 (1997).
- 国立国語研究所: 現代語複合辞用例集 (2001).
- グループ・ジャマシイ: 日本語文型辞典, くろしお出版 (1998).
- Carletta, J.: Assessing Agreement on Classification Tasks: The Kappa Statistic, *Computational Linguistics*, Vol. 22, No. 2, pp. 249-254 (1996).
- 土屋雅稔, 宇津呂武仁, 佐藤理史, 中川聖一: 形態素情報を用いた日本語機能表現の検出, 言語処理学会第 11 回年次大会発表論文集, pp. 584-587 (2005).
- 高木俊宏, 注連隆夫, 土屋雅稔, 内元清貴, 松古俊, 宇津呂武仁, 佐藤理史: 機械学習を用いた日本語機能表現のチャンキング, 言語処理学会第 12 回年次大会論文集 (2006).
- 前川喜久雄: 『日本語話し言葉コーパス』の概観 ver.1.0 (2004).