日本語機能表現用例コーパスの作成

Developing a Corpus of Example Sentences of Japanese Functional Expressions

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

*豊橋技術科学大学 情報処理センター

† 京都大学 情報学研究科

‡豊橋技術科学大学 工学部情報工学系

1. はじめに

日本語には，複数の語が定型的・複合的に使われ，ひとかたまりの表現として機能的な関係を表す表現が多数存在する．このような表現は機能表現と呼ばれ，日本語文の構造を理解するために非常に重要である．機械翻訳や構文解析の段階で機能表現を適切に取り扱うには，自動的に機能表現を検出するシステムが必要だが，そのような検出システムの訓練や評価には，機能表現のコーパスが必要である．また，機能的な用法を網羅した資料としても，コーパスは有用である．そのような背景のもとで，我々は，機能表現を研究するための基礎資料として機能表現用例コーパスの作成を行っている．本稿では，その作成手順と作成状況について報告する．

機能表現には，その構成要素の意味から全体の意味を構成的に説明できるような表現と，その構成要素の意味から全体の意味を構成的には説明できないような非構成的な表現とが存在する．非構成的な意味を持つ機能表現リストの 1 つとして，現代語複合辞用例集 [1] がある．この用例集の各項目には，各表現の用例と用法の説明文が含まれている．このように整備された用例集があると，コーパスの作成作業も進めやすいので，本研究では，現代語複合辞用例集に収録されている機能表現の用例コーパスを作ることにする．

ここで，このコーパスを，現代語複合辞用例集に収録されている機能表現の検出器の訓練・評価用データとして利用する場合，(a) 用例集に収録されている表現が用例集で説明されている用法で用いられている文と，(b) 用例集に収録されている表現と同一の文字列または形態素列が含まれるが，用例集に説明されている用法以外の用法で使われている文，の 2 種類の文が必要である．しかし，現代語複合辞用例集には，(b) のような文は含まれていない．また，さまざまな書籍から例文を引用しているため，利用許諾の範囲が曖昧な文も含まれてしまっている．そこで本研究では，新聞記事から，現代語複合辞用例集に収録されている表現と同一の文字列または形態素列を含む文を収集して，機能表現用例コーパスを作成する．

2. 対象とする機能表現

2.1 機能表現と複合辞

機能表現とは，幾つかの語が複合してひとまとまりの形で辞的な機能を果たす表現である．辞的な機能には，以下の 6 つがある．

- 基本的に節を受けて，複文の前件を形成する (接続助詞型)
- 前件の体言を後件の用言に関係付ける (連用助詞型)
- 前件の体言を後件の体言に関係付ける (連体助詞型)
- 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 (終助詞型)
- 前件の用言に付加的なニュアンスを加える (助動詞型)
- 前の文を後ろの文に関係付ける (接続詞型)

森田ら [2] は，機能表現の中でも特に「単なる語の連接ではなく，表現形式全体として，個々の構成要素のプラス以上の独自の意味が生じている」表現を**複合辞**と呼び，個々の構成要素の意味から構成的に表現形式全体の意味を説明できるような機能表現とは区別している．

既存の機能表現リストとしては，以下のものがある．

- 現代語複合辞用例集 [1]
  125 項目，308 表現
- 日本語表現文型 [2]
  52 意味大分類，210 意味小分類，450 表現
- 日本語文型辞典 [3]
  115 意味分類，965 表現，2169 用法

本稿では，現代語複合辞用例集で説明されている機能表現を対象として，機能表現用例コーパスの作成を行う．

2.2 現代語複合辞用例集

現代語複合辞用例集 (以下，用例集と呼ぶ) は，日本語表現文型において複合辞として立項されている表現を基本とし，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表現を，さしあたりの研究対象として列挙したものである．用例集では，複合辞は以下の 5 つに分類されている．

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 (36 項目,67 表現,「〜とはいえ」など).
   - **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 (45 項目,123 表現,「〜について」など).
   - **連体辞類**　名詞や節などを受けて，連体修飾句を形成するもの (2 項目,3 表現,「〜といった」など).
   - **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**」(42 項目,115 表現,「〜つもりだ」など)

◇ **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・とらえ方をする主体を表す.
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．(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，いささかぎ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,「にとっての」という形でなら可能である.「にとりまして」という丁寧の形も取れる．とらえ方をする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

ただし，文末辞類は収録されていない．残る 4 種類の複合辞が，図 1 のような体裁で解説されている．それぞれの項目は,「A56 ～にとって・～にとり」という見出しと,「接続」「意味・用法」「文法」「ノート」といった説明文,および用例からなっている.

多くの複合辞には，(少なくとも形式的には) 自立語が含まれており，複合辞と同一の形態素列が内容的用法で用いられている場合がある.

素晴らしい彫刻を手にとってじっと観察してみる.

このような区別を説明文だけで記述することは大変難しい．したがって，複合辞 (または機能表現) を定義するには，具体的な用例をできるだけ多数掲載することが必要である．用例集の場合は，1 項目当り平均して 16.6 文が掲載されている.

2.3　機能表現リストの作成

用例集の 1 つの項目は，表現の一部が異なるような複数の表現を同時に解説している．第 1 に，見出しに含まれている機能表現の表記部分は，括弧や中黒を利用して,表記が異なる複数の表現を記述している．例えば，図 1 の項目 (A56) の見出しには,「～にとって」「～にとり」という 2 つの表現が含まれている．第 2 に,「意味・用法」「文法」「ノート」といった説明文でも，連体修飾形「～にとっての」や丁寧形「～にとりまして」といった表現が追加されている．用例集の説明文で 1 つの項目にまとめて説明されている表現としては，以下のようなものが見られる．(1) 連体修飾形 (～にとって →～にとっての),(2) 丁寧形 (～にとって →～にとりまして),(3) 口語形 (～ては →～ちゃ),(4) 否定の変化形 (～にとどまらない →～にとどまらず),(5) 否定形 (～ことがある →～ことがない)．これらの表現を，表層文字列によって機械的に区別・列挙して，本コーパスの対象とする機能表現リストを作成する．1 つの項目内で説明されている複数の表現を区別するために，各桁が以下のような意味を持つ 4 桁の枝番号を導入する.

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表現を区別
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:一意な番号 $(0, 1, 2, \ldots)$ を与える

例として，図 1 の項目 (A56) を，機能表現リストの形式に展開した結果を以下に示す.

見出し: A56 ～にとって・～にとり
A56-1000: にとって
A56-1010: にとっての　(←「～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

つまり，図 1 の項目 (A56) には，5 つの複合辞が含まれている．なお，このリストでは，それぞれの複合辞を単純な文字列として表しているが，あくまで，これらの文字列が用例集に説明されている用法で用いられている場合を対象として考えている．また，一部の複合辞については，その複合辞の説明文に記述は存在しなかったが，他複合辞との対称性を考慮して丁寧形や否定形を補った (以下の下線部分).

見出し: B17 ～までもない・～までのこともない
B17-1000: までもない
B17-1100: までもありません
B17-2000: までのこともない
B17-2100: までのこともありません

補った表現は現時点で 31 表現であり，全体では 339 表現となっている.

3.　機能表現用例コーパスの作成

3.1　作業対象となる文の収集

新聞記事から，以下の 2 つの方法のいずれかで収集された文を，機能表現リストに登録されている機能表現が用いられている可能性がある文として収集する.

表 1 「〜となると (A6–1000)」「〜にかけては (A44–1000)」の用例に対する用法の類別

	機能/内容	用例集の用法か?	類別結果例
(1)	f	o	<S id="950121193" class="f" bunrui="o">しかし風邪という特定の感染症に特効がある<F>となると</F>事は重大だ。</S>
(2)	g	−	<S id="950418032" class="g">脳死移植の適応基準は心臓、肝臓に次ぐもので、国会で継続審議中の臓器移植法案成立へ向けての条件整備の一つ<F>となると</F>みられる。</S>
(3)	g	−	<S id="950830315" class="g">汚職事件でもトップは「うまくやれ」とのあいまいな表現で、いざ<F>となると</F>部下に責任を取らす。</S>
(4)	f	x	<S id="951105160" class="f" bunrui="x"><F>となると</F>「たつ」とはなにか?</S>
(5)	f	o	<S id="950413334" class="f" bunrui="o">人心をつかむこと<F>にかけては</F>動物的な勘をもっている。</S>
(6)	f	x	<S id="950113264" class="f" bunrui="x">小学生のころから中学生<F>にかけては</F>、時間の流れが非常に緩やかだった。</S>
(7)	g	−	<S id="950104186" class="g">「私は、私が感じた烈を演じていくつもり」とあまり気<F>にかけては</F>いない。</S>

(a) 文字列一致による収集 機能表現の文字列を含む文を無条件に収集する.「〜として (A62–1000)」の収集例を以下に示す.

助手として働く
彼はきちんとしている
財布を落として困っている

(b) 基本形を考慮した収集 機能表現の末尾形態素が活用して用いられている場合を収集する. 以下に,「〜つつある (B35–1000)」の収集手順 (1) ~ (3) を示す.

(1) MeCab[4] を用いて文を形態素解析する.
台風/は/本土/を/北上/し/つつ/あっ/た
(2) 文中の活用している語の 1 つだけを基本形に置き換えた文を生成.
台風/は/本土/を/北上/する/つつ/あっ/た
台風/は/本土/を/北上/し/つつ/ある/た
(3) 機能表現の文字列「つつある」と一致し, かつ, 一致部分の先頭と末尾の位置が形態素区切りとなっている部分が検出されれば, この文を収集する.
台風/は/本土/を/北上/し/つつ/ある/た

3.2 機能的用法・内容的用法の類別

収集された文に対して, 最初に機能的か否かの区別を行い, 機能的に用いられている場合には用例集の用法で用いられているか否かの区別を行う. すなわち, 以下のような 2 段階の区別を行う.

- 機能的用法である (f)
  - 用例集の用法である (o)
  - それ以外の用法である (x)
- 内容的用法である (g)

この区別を実際のタグで表現した例を, 表 1 に示す. 用例集では「〜となると (A6–1000)」の用法は,「A となると B」の形で「A という事態になった場合には B」という関係を示す, と説明されている. 表中の文 (1) は, 文中の「となると」が用例集の用法で使われている例である. それに対して, 文 (2) は, 動詞「なる」が自立語としての動詞の働きをしていて, 内容的用法と分類される例である. また, 文 (3) のような場合は,「いざとなると」をひとかたまりとして慣用表現と考え, 内容的用法に含める. 文 (4) では, 文頭の「となると」は直前の文を受けて接続詞的な働きをしている. この場合, 機能的な働きをしているとは言えるが, 用例集で説明されている用法の範囲には含めない.

接続助詞「て」を末尾に含む表現 (「〜に関して (A46–1000)」「〜について (A53–1000)」など) の多くは, 直後に助詞「は」「も」を補うと,「〜に関しては」「〜についても」といった形で提題助詞的または副助詞的な働きをするようになる. しかし, 用例集には, このような表現は明示的には収録されておらず, 助詞の後続によって意味が顕著に変わる表現だけが個別に収録されている. たとえば,「〜にかけて (A59-1000)」は,「A から B にかけて」の形で時間的・空間的な範囲を示す機能表現である. それに対して,「〜にかけては (A44–1000)」は,「A にかけては」の形で高く評価されるような事柄 A を取り立てる働きをする表現であり,「〜にかけて (A59-1000)」とは全く意味が異なっている. そのため,「〜にかけては」という文字列を含む文の用法を判定する場合は,「〜にかけては (A44–1000)」の用法で用いられている文 (5) と,「〜にかけて (A59-1000)」に助詞「は」が後続して現れている文 (6) を区別する必要がある.

3.3 コーパスの作成手順

ここでの機能表現用例コーパスは, 機能表現検出器 [5] の訓練および評価用データとしての利用を想定しているため, 各表現について, その表現が用例集の用法で用いられている文 (以下, 正例と呼ぶ) と, それ以外の用法で用いられている文 (以下, 負例と呼ぶ) がバランス良く含まれていることが必要である. そこで, 正例が p 個以上, 負例が n 個以上含まれているような m 個の文からなるコーパスを以下の手順により作成する.

(1) 新聞記事から m 個の文を収集.
(2) 作業者による用法判定.
(3) 別の作業者による検証.
(4) 正例が p 個以上含まれているか検査. 含まれていない場合は, 正例である可能性が高い文を補充してステップ 2 に戻る.
(5) 負例が n 個以上含まれているか検査. 含まれていない場合は, 負例である可能性が高い文を補充し

表 2 新聞記事からの文の収集結果 (全 337 表現)

	表現数
50 ≤ 文数	187 (55%)
0 < 文数 < 50	117 (35%)
文数 = 0	33 (10%)

表 3 正例の出現回数 (正例・負例を未補充の段階，全 127 表現)

	表現数	例
正例数 = 50	44 (35%)	～に至っては (A49–1000)
40 < 正例数 < 50	29 (23%)	～て仕方がない (B33–5000)
10 ≤ 正例数 ≤ 40	41 (32%)	～にあって (A38–1000)
正例数 < 10	13 (10%)	～得る (B39–1000)

てステップ 2 に戻る．ただし，負例が存在しないことが確認できた場合は，そのまま終了する．

最初に，3.1 節の方法で新聞記事から機能表現が用いられている可能性がある文を収集する．m 個以上の文が収集された場合には，均等に m 個を取り出して，用法判定の対象とする．収集された文が m 個未満の場合は，対象とする新聞記事データを拡大して，m 個の文を確保する．次に，収集された文を対象として作業者による用法判定を行い，その上で別の作業者による検証を行う．検証段階で問題が発見されると，判定基準を明文化したり，判定作業のやり直しを行う．

続いて，検証された m 個の文に，p 個以上の正例が含まれているか検査する．含まれていなかった場合は，用例集の説明文に記述されている接続制約を利用して，正例である可能性が高い文を補充し，用法判定 (ステップ 2) に戻る．最後に，検証された m 個の文に n 個以上の負例が含まれているか検査する．含まれていなかった場合は，接続制約を満たさない文を負例である可能性が高い文として補充する．ただし，負例が存在しないことが確認できた場合は，補充は行わずに終了する．

3.4 コーパスの作成状況

本コーパスの作成では，$m = 50, n = p = 10$ とパラメータを設定した．文収集の対象データとしては，1995 年の毎日新聞記事データ [6] を使った．作業対象の 337 表現☆ を，収集された文数によって分類すると表 2 となる．

前節の作業手順に従えば，文数が 50 に満たない 150 表現について，更に多くの新聞記事データを参照して，十分な数の文を確保する必要がある．しかし，本稿執筆時の段階では，文が 1 つも収集されなかった 33 表現だけを作業対象外とし，50 文に満たない 117 表現についても，用法判定を行うことにした．新聞記事から文を収集できた 304 表現の内，現段階で用法判定が完了している表現は 290 表現，さらに検証まで完了している表現は 127 表現である．検証済みの 127 表現を，正例の数によって分類すると表 3 となる．正例が 10 個未満の 13 表現，および負例が 10 個未満の 73 表現については，文を補充する作業を実施中であり，3.3 節の手順を全て終了した表現は 41 表現である．

☆ 「～といい～といい (A66–1000)」および「～といわず～といわず (A67–1000)」については，文の収集手順が複雑になるため，今回は作業対象外とした．

表 4 正例の割合 (正例・負例を未補充の段階，全 127 表現)

分類		50 文中の正例の割合			
		100%	90%以上	50%以上	50%未満
助詞型	接続辞類	1	2	13	17
	連用辞類	15	9	8	5
	連体辞類	0	1	1	1
助動詞型		28	11	8	7
計		44	23	30	30

次に，検証済の 127 表現についての統計量を報告する．用例集では，複合辞は機能的な観点から 5 種類に分類されている (2.2 節)．検証済の 127 表現を，50 文中に占める正例の割合と，用例集における機能的な分類で分類すると，表 4 となる．表より，助動詞型および連用辞類に属する表現の場合は，新聞記事から収集された文の過半数が正例である．それに対して，接続辞類に属する表現の場合は過半数が負例である．

4. おわりに

本稿では，現代語複合辞用例集に収録されている機能表現を対象として，機能表現用例コーパスの作成手順および作成状況について報告した．機能表現を取り扱うとき，どの範囲の表現を同じ表現と見なすか，という問題がしばしば発生する．松吉 [7] は，機能表現を分類するための階層的な体系を提案し，その体系に基づく辞書を作成している．この辞書では，2.3 節で区別したいくつかの表現が同じものとして扱われている．

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

1) 国立国語研究所: 現代語複合辞用例集 (2001).
2) 森田良行, 松木正恵: 日本語表現文型, NAFL 選書 5, アルク (1989).
3) グループ・ジャマシイ: 日本語文型辞典, くろしお出版 (1998).
4) 工藤拓: 形態素解析器 MeCab. http://chasen.org/~taku/software/mecab/.
5) 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 C3–1 (2005).
6) 毎日新聞社: CD-毎日新聞'95 データ集, 日外アソシエーツ (1996). http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html.
7) 松吉俊, 佐藤理史, 宇津呂武仁: 機能・意味・形態にもとづく助詞型機能表現の分類, 言語処理学会第 11 回年次大会 B2–2 (2005).