階層構造による日本語機能表現の分類

松吉 俊[†]　佐藤 理史[‡]　宇津呂 武仁[†]

[†]京都大学大学院 情報学研究科, [‡]名古屋大学大学院 工学研究科
matuyosi@pine.kuee.kyoto-u.ac.jp, ssato@nuee.nagoya-u.ac.jp, utsuro@i.kyoto-u.ac.jp

1. はじめに

日本語の解析システムは、1990 年代にそれまでの研究が解析ツールとして結晶し、現在では、各種の応用システムにおいて、それらの解析ツールが解析モジュールとして利用されるようになってきている。解析ツールを利用した応用システムの理想的な構成は、与えられた文を解析する解析ツールと、その後の処理を直列につなげた図 1 に示すような構成[☆]であり、応用に特化しない言語解析処理を解析ツールがすべて担うことが期待される。

しかしながら、実際は、現在の解析ツールは、いくつかの言語現象の扱いが不十分であり、それぞれの応用モジュールが、やむを得ず、必要に迫られた範囲内でそれらの処理を行なっているのが現状である。そのような言語現象の具体例は、おおきく、以下の 4 種類に分類できる。

(1) 表記の問題
(2) 単位の問題 (複合語・複合辞・慣用句の認定の問題)
(3) 外部情報源 (たとえば、一般の国語辞典) とのインタフェースの問題
(4) 異形式同意味の問題

本研究の目的は、機能表現に関してこれらの問題を解決するための基礎となる**日本語機能表現辞書**を作成することである。

本論文は、以下のように構成される。まず、第 2 章で、機能表現の定義と日本語機能表現辞書の設計について述べる。次に、第 3 章において、機能表現を分類するための階層構造と、それを記述する形式について説明する。第 4 章で、辞書作成の作業手順について説明し、現状を報告する。第 5 章で、関連研究について述べ、最後に、第 6 章でまとめを述べる。

2. 日本語機能表現辞書の設計

2.1 機能表現

機能表現とは、おおよそ、文中においてなんらかの機能を持つ表現ということができよう。しかしながら、機能表現全体に対して、「なんらかの機能」をより精密に書き下すことはほとんど不可能であるので、われわれは、機能表現の定義として、次のような定義を採用する [1)]。

> 表 1 のいずれかの機能を持つ語・表現を、**それぞれの機能型に属する機能表現**と呼び、その総称として**機能表現**という用語を用いる。

表 1 機能と L^3ID (全 8 種類)

機能	機能型	L^3ID
前件を後件の用言に関係付ける	格助詞型	P
前件を後件の節に関係付ける	接続助詞型	Q
前件を後件の体言に関係付ける	連体助詞型	D
前の文を後ろの文に関係付ける	接続詞型	C
前件に付加的なニュアンスを与える	助動詞型	M
前件を名詞化する	形式名詞型	N
前件を取り立てる	とりたて詞型	T
前件を話題化する	提題助詞型	W

2.2 辞書に求める要件

われわれは、日本語機能表現辞書を作成するにあたり、辞書に次の 3 つの要件を設定した。

要件 1 機能表現の出現形を網羅する見出し体系をもっていること

要件 2 個々の機能表現に対して、文法情報や属する意味カテゴリーが記述されていること

要件 3 関連する機能表現間の関係が明示されていること

要件 1 を設定した理由は、すべての可能な機能表現の出現形を、計算機に誤りなく認識させたいからである。要件 2 を設定した理由は、解析システムなどの自然言語処理システムに対して、個々の機能表現についての情報をこの辞書から出力することを想定しているからである。要件 3 を設定した理由は、この辞書を、異形式同意味の判定や言い換えに利用したいと考えているからである。

2.3 辞書の設計方針

前節の要件を満たす辞書を作成するにあたり、設計方針を以下のように定めた。

見出し体系 9 つの階層をもつ**階層構造**(3 章参照)

辞書の形式 XML 形式

付加情報 以下に挙げる情報を記述する

- **表現の左 (もしくは右) に来る境界の境界 ID** [2),3)]
  左 (もしくは右) 接続に相当
- **否定形** 意味の観点から見た否定の表現
- **核** 表現構成の中心的な核の形態素の位置
- **抑制** 不必要な表現であるとして、外部への出力を抑制するかどうか

[☆] ここでいう応用モジュールとは、言語表現そのもの (言語構造) を対象とするものではなく、言語表現が伝える情報 (情報構造) を対象とするものである。

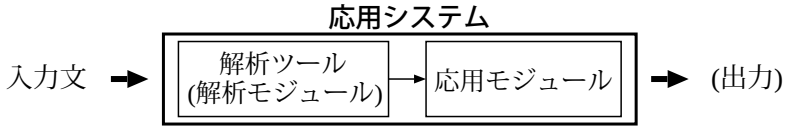

図 1 応用システムの理想的な構成

意味カテゴリー 全 103 種類
文体 常体、敬体、口語体、堅い文体の 4 種類
難易度 やさしい方から A1、A2、B、C、F の 5 段階 [4)]
例文 機能表現を含む文
慣用表現 機能表現を含むもの
文献への参照 文献名および参照ページ

見出し体系として採用した階層構造については、次章で詳しく説明する。辞書の形式として XML 形式を採用した理由は、XML 形式は、階層構造を表現するのに都合が良く、また、プログラムにより、他の形式への変換が容易であるからである。

3. 階層構造による分類

機能表現には、異形をもつものが多い。例えば、「なければならない」、「なくてはならない」、「なくてはならず」、「なければいけない」、「なけりゃならない」、「なければならぬ」、「ねばならん」などは、互いに異形の関係にある。

人間用の辞書であれば、代表的な表現のみを見出しとして採用し、それらに対する異形を本文もしくは例文の中で例示するのみで十分であろう。なぜならば、日本語の母語話者は、ある表現がどの代表的な表現の異形であるのかを判断することができ、逆に、ある表現に対してその異形を類推することができるからである。実際、機能表現に関する、言語学や日本語教育学の文献は、このような記述形式をとっている [5),6)]。

しかしながら、計算機のための、機能表現の出現形を網羅する見出し体系として、上と同じ方式を採用することはできない。なぜならば、ある機能表現からその異形をすべて生成する操作は、個々の表現にかなり依存しているため、誤りなく可能な異形のみを生成する少数の規則群を設定することは、ほとんど不可能であるからである。そこで、本研究では、可能な機能表現の出現形すべてを見出しとして列挙することにした。

すべての出現形を列挙するにあたり、その集合になんらかの構造を導入すると見通しが良くなる。そこで、われわれは、松吉ら [1)] の分類を発展させ、すべての可能な機能表現の出現形を体系的に分類することができる階層構造を作成し、それを見出し体系として採用することにした。

この階層構造は、9 つの階層をもつ。これらの階層に関する情報を、まとめて表 2 に示す。それぞれの階層において、表 2 に示す区分観点によって機能表現を区分する。これらの階層を、次の 3 つのグループに分けて、詳しく説明する。

(A) 階層 L^1、階層 L^2、階層 L^3
(B) 階層 L^4、階層 L^5、階層 L^6
(C) 階層 L^7、階層 L^8、階層 L^9

3.1 グループ A の階層

グループ A の階層では、機能表現集合を大きく区分する。ここでの区分は、区分の順番に多少の違いはあるが、言語学において、一般的に行なわれているものである。

まず、階層 L^1 において、機能表現を構成する要素 (主に、内容語) の異なりによって、機能表現を区分する。そして、階層 L^2、階層 L^3 において、それぞれ、意味、機能の異なりによって、機能表現を下位区分する。

3.2 グループ B の階層

グループ B の階層では、機能語が交替したものや音が縮約したものなど、一般に「異形」と認識されるものを区分する。活用形の異なりや表記のゆれによる異形を含まない、この「異形」を、本論文では、**変異体**と呼ぶことにする。そして、それを生成する変形操作を**変異**と呼ぶことにする。

機能表現の形態に関する文献を調査するとともに、実際に用例中に存在する機能表現の形態を観察した結果、変異の主なものは、次の 3 種類であることが分かった。

(i) 機能語の交替
(ii) 縮約、脱落、連濁など、音韻的な変化
(iii) とりたて詞の挿入

本論文では、(ii) の変異を**音変異**、(iii) の変異を **T 変異**と呼ぶことにする。

われわれは、上の 3 種類の変異に対して、機能語の交替 > 音変異 > T 変異という優先順位を設け、L^3 単位の機能表現を、階層 L^4 で機能語の交替の観点から区分し、階層 L^5 で音変異の観点から下位区分し、階層 L^6 で T 変異の観点からさらに下位区分する。このような優先順位を設けた理由は、機能語の交替が、可能な音変異に影響し、音変異が、可能な T 変異に影響すると思われるからである。

上記の 3 種類以外の変異は、個々の機能表現に強く依存するマイナーな変異である。マイナーではあるが、これらの変異は、可能な音変異に影響を与えると思われるので、機能語の交替と同様に扱う。

3.3 グループ C の階層

グループ C の階層では、機能表現集合を細かく区分する。

階層 L^7、階層 L^8 において、それぞれ、活用形の異な

表 2 9 つの階層に関する情報

階層	区分観点	ID		表 3 における区分例	XML 形式による表現法 (図 2 参照)	現在の単位数
		文字種	長さ			
L^1	構成語	数字	4	(1) と (2)	ENTRY 要素の集合	241
L^2	意味	数字	2	(1) と (3)	ENTRY 要素	278
L^3	機能	英字 (8 種)	1	(1) と (4)	P、Q、D、C、M、N、T、W 要素	390
L^4	機能語の交替	数字	1	(5) と (6)	PHRASE 要素の集合	572
L^5	音韻的な変化	英字 (29 種)	1	(5) と (7)	PHRASE 要素	895
L^6	とりたて詞の挿入	英字 (17 種)	1	(8) と (9)	PHRASE 要素の TINSERT 属性	1479
L^7	活用形	数字	2	(10) と (11)	FORM 要素	6028
L^8	「です」「ます」の有無	英字 (2 種)	1	(10) と (12)	NORMAL、DESUMASU 要素	8249
L^9	表記	数字	2	(10) と (13)	SPELLING 要素	13690

表 3 表示と機能表現 ID

	表示	意味カテゴリー	機能表現 ID
(1)	〈 **にたいして, にたいして**〉	**対象**	jA2000801P1xx01n01
(2)	〈について, について〉	対象	jA2000901P1xx01n01
(3)	〈 にたいして, にたいして 〉	割合	jB0000802P1xx01n01
(4)	〈 にたいして の, にたいして の〉	対象	jA2000801D1xx42n01
(5)	〈 **なければならない, なければならない**〉	**当為**	jA2015501M1xx04n01
(6)	〈 なく ては ならない, なく ては ならない 〉	当為	jA2015501M2xx04n01
(7)	〈 なけ りゃ ならない, なけ りゃ ならない 〉	当為	jA2015501M1bx04n01
(8)	〈 **といっても, といっても**〉	**逆接確定**	jA2009701Q1xx24n01
(9)	〈 と は いっても, と は いっても 〉	逆接確定	jA2009701Q1xh24n01
(10)	〈 **ことができる, ことができる**〉	**可能**	jA2017501M1xx43n01
(11)	〈 ことができ れ, ことができ れ〉	可能	jA2017501M1xx07n01
(12)	〈 ことができ ます, ことができ ます〉	可能	jA2017501M1xx43s01
(13)	〈 ことが 出来 る, ことができる 〉	可能	jA2017501M1xx43n02

り、「です」「ます」の有無によって、L^6 単位の機能表現を区分する。そして、最後に、階層 L^9 において、表記の異なりによって機能表現を下位区分する。最終的に得られる区分単位 (L^9 単位) は、表示 (表記と読みの組)[4] である。

3.4 機能表現 ID

L^9 単位である機能表現の表示に対して、一意な機能表現 ID を設定する。この機能表現 ID は、次の 3 つから構成される 18 文字の記号列である。

プリフィックス “j”

難易度 “A1”、“A2”、“B0”、“C0”、“F0” のいずれか

L^1ID～L^9ID の列 表 2 の ID の欄参照

表示と機能表現 ID の例を表 3 に示す。

3.5 XML 形式による階層構造の表現法

上記の階層構造を、日本語機能表現辞書において、XML 形式で表現する。XML 形式による階層構造の表現例を図 2 に示す。

それぞれの階層は、表 2 の「XML 形式による表現法」の欄に示す方法によって表現する。表記と読みは、“.” によって短単位の形態素に区切り、とりたて詞が挿入可能な位置は、“..” で示す。この形式の階層構造から、必要に応じて、9 層の入れ子構造をもつ、より素直な形式の階層構造を自動生成することが可能である。

4. 作業手順と現状

日本語機能表現辞書作成にあたっての作業手順は、次の通りである。

(1) 言語学の文献など、機能表現のリストを入手する
(2) リストの一つの機能表現を既存の階層構造の適切な場所に挿入する
(3) その機能表現の異形と思われる表現を半自動的に生成する
　具体的には、表現の末尾に応じたテンプレートを適用し、表現の幹の部分を適切に置換する
(4) 実際に存在しない表現を消去する、付加情報を記述するなど、整理を行なう
(5) 上の作業を繰り返す

現在は、「日本語表現文型」[5] にとりあげられている機能表現を対象として、分類・整理を行なっている。すでに整理が終わったものは、「格助詞相当のもの」、「接続助詞相当のもの」、および、「助動詞相当のもの」で、計 309 エントリーである。階層構造に存在する各区分単位☆の数を表 2 の「現在の単位数」の欄に示す。

なお、付加情報 (2.3 節参照) のうち、境界 ID と慣用表現はまだ記述していない。

5. 関 連 研 究

首藤ら [7)~9)] は、関係表現 (助詞相当語)1000 エントリー、助述表現 (助動詞相当語)1500 エントリーを収集し、それらを意味に基づいて分類することにより、言い

☆ 抑制 (2.3 節参照) が指定されているものを除く

```
<?xml version="1.0" encoding="EUC-JP" ?>
<ENTRIES>
 <ENTRY ID="0001" MID="01" DIFFICULTY="A2">
    :
 </ENTRY>
    :
 <ENTRY ID="0008" MID="01" DIFFICULTY="A2">
  <P>
   <PHRASE SUBID="1" PHONETIC="x" BASE="に. たいし. て"
      TINSERT="x">
    <FORM TAIL="て" LEFTBOUNDARIES="3100"
       RIGHTBOUNDARIES="2200;2500">
     <NORMAL CORE="2/3">
      <SPELLING>に. たいし. て</SPELLING>
      <SPELLING>に. 対し. て</SPELLING>
     </NORMAL>
     <DESUMASU CORE="2/4">
      <SPELLING>に. たいし. まし. て</SPELLING>
      <SPELLING>に. 対し. まし. て</SPELLING>
     </DESUMASU>
    </FORM>
   </PHRASE>
   <PHRASE SUBID="1" PHONETIC="h" BASE="に. たいし. ちゃ"
      STYLE="colloquial" TINSERT="x">
       :
   </PHRASE>
   <PHRASE SUBID="2" PHONETIC="x" BASE="に. たいし"
      TINSERT="x">
       :
   </PHRASE>
  </P>
    :
  <D>
   <PHRASE SUBID="1" PHONETIC="x" BASE="に. たいし. て. の"
      TINSERT="x">
       :
   </PHRASE>
       :
  </D>
    :
 </ENTRY>
 <ENTRY ID="0008" MID="02" DIFFICULTY="B">
    :
 </ENTRY>
    :
</ENTRIES>
```

図 2 XML 形式による階層構造の表現例

換えなどの応用に役立てている。しかしながら、異形についての大規模な整理は行なっていないようである。

兵藤ら [10] は、機能語辞書において、可能な異形 (13882 エントリー) をスロットを用いて記述する方法を提案している。しかしながら、それは、表現のある部分文字列に対して交替可能な文字列を列挙しているだけであり、異形の体系的な整理を行なっているとは言いがたい。

EDR 日本語単語辞書には、助詞相当語 82 エントリー、助動詞相当語 49 エントリーが登録されているが、異形に関する情報は記載されていない [11]。

日本語話し言葉コーパスにおいては、助詞相当句 80 エントリー、助動詞相当句 92 エントリーが長単位として認定されている [12]。丁寧形や異形態などの観点から異形の分類を行なっているが、リストとしての規模が小さい。

6. おわりに

本論文では、現在、われわれが作成している、自然言語処理のための日本語機能表現辞書について報告した。この辞書では、機能表現を体系的に分類するために、9 つの階層をもつ階層構造を見出し体系として採用した。辞書は、XML 形式で記述されている。

「日本語表現文型」のエントリーの整理が終わり次第、「日本語文型辞典」[6] のエントリーの整理にとりかかる予定である。また、佐藤 [2),3)] に従い、各機能表現に対して、境界 ID を記述する予定である。

本研究の一部は、次の研究費による；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)、21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」17 年度 若手リーダーシップ養成プログラム研究費、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) 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 回年次大会発表論文集, pp. 384–387 (2005).
2) 佐藤理史: 境界認定の提案: (1) コンセプトと実現法, 情報処理学会研究報告 2004-NL-164, pp. 25–32 (2004).
3) 佐藤理史: 境界認定の提案: (2) 背景と思想, 情報処理学会研究報告 2004-NL-164, pp. 33–40 (2004).
4) 佐藤理史: 異表記同語認定のための辞書編纂, 情報処理学会研究報告 2004-NL-161, pp. 97–104 (2004).
5) 森田良行, 松木正恵: 日本語表現文型 用例中心・複合辞の意味と用法, アルク (1989).
6) グループ・ジャマシイ編: 教師と学習者のための日本語文型辞典, くろしお出版 (1998).
7) 首藤公昭: 文節構造モデルによる日本語の機械処理に関する研究, 福岡大学研究所報第 45 号 (1980).
8) Shudo, K., Narahara, T. and Yoshida, S.: Morphological Aspect of Japanese Language Processing, *Proceedings of the 8th COLING*, pp. 1–8 (1980).
9) Shudo, K., Tanabe, T., Takahashi, M. and Yoshimura, K.: MWEs as Non-propositional Content Indicators, *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrating Processing (MWE-2004)*, pp. 32–39 (2004).
10) 兵藤安昭, 村上裕, 池田尚志: 文節解析のための長単位機能語辞書, 言語処理学会第 6 回年次大会発表論文集, pp. 407–410 (2000).
11) 日本電子化辞書研究所: EDR 電子化辞書 2.0 版 仕様説明書, 第 2 章 日本語単語辞書 (2001).
12) 小椋秀樹, 山口昌也, 西川賢哉, 石塚京子, 木村睦子: 『日本語話し言葉コーパス』の形態論情報の概要 ver. 1.0, 国立国語研究所 (2004).