大規模日本語機能表現辞書の階層性を利用した機能表現検出*

長坂 泰治† 宇津呂 武仁‡ 土屋 雅稔§
筑波大学第三学群工学システム学類†，筑波大学大学院 システム情報工学研究科‡，
豊橋技術科学大学 情報メディア基盤センター§

1 はじめに

機能表現[1]とは,「にあたって」や「をめぐって」のように，2 つ以上の語から構成され，全体として 1 つの機能的な意味をもつ表現である．一方，この機能表現に対して，それと同一表記をとり，内容的な意味をもつ表現が存在することがある．そのような表現においては，機能的な意味で用いられている場合と，内容的な意味で用いられている場合とを識別する必要がある．我々はこれまでに，現代語複合辞用例集 [国研 01](以下，用例集) 中の代表的複合辞一覧に基づいて，それらの派生形である 337 種類の機能表現を規定し，その用例データベース (日本語複合辞用例データベース [土屋 06, 土屋 07a]，以下，用例データベース) を作成した．また，それらの用例データベースを訓練事例として，機械学習により機能表現の検出・係り受け解析を行う方式を提案した [土屋 07b, 注連 07]．また，機能表現の異形の語構成パターンを網羅することにより，16,801 種類の出現形を網羅した階層的辞書 (日本語機能表現一覧 [松吉 07b]，以下，機能表現一覧) を作成した．ここで，[土屋 07b, 注連 07] の機械学習による機能表現検出においては，一つの表現あたり 50 例程度の訓練用例に対して，人手で機能的・自立的等の用法判定を行う必要がある．しかし，機能表現一覧の全機能表現 16,801 種類に対して，それだけの規模の作業を行うことは容易ではない．そこで，本稿では，機能表現一覧の階層性を利用し，階層において下位に位置する機能表現について，用法が類似するより上位の表現に言い換えた後，用法判定を行う方式を提案する．そして，人手による用法判定済コーパスを用いた統計的調査を行い，その妥当性を検証する．

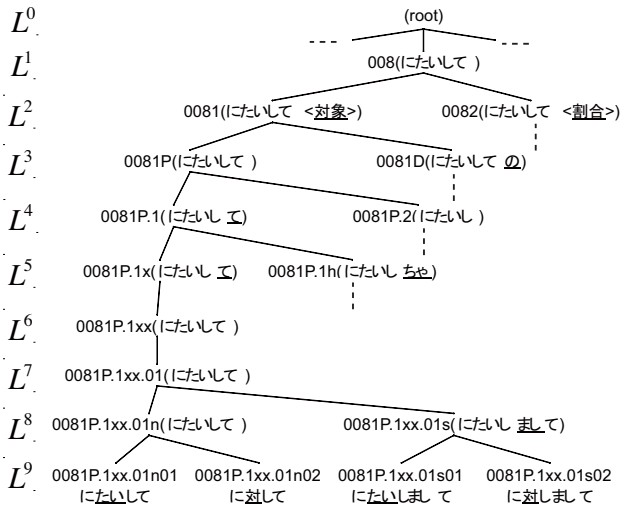

図 1: 機能表現辞書階層構造の一部

2 階層的日本語機能表現辞書

機能表現一覧 [松吉 07b] の 9 階層における機能表現の分類法，および機能表現数を表 1 に，階層構造の一部を図 1 に，それぞれ示す[2]．

3 機能表現用法判定済コーパス

これまでに，毎日新聞 1995 年テキストから選定したテキストから構成される用例データベース [土屋 06] において，337 種類の代表的な機能表現に限定して，表 2 の判定ラベルを人手で付与した[3]．また，機能表現チャンキング [土屋 07b] や統計的係り受け解析 [注連 07] において訓練・評価データとして利用する目的で，京都テキストコーパス [黒橋 97] 中の機能表現に対して表 2 の判定ラベルを人手で付与した [土屋 07a]．本研究においては，現在，これらの用法判定済コーパス，および，それを拡張したコーパスにおいて，用例データベースの 337 表現に限定せず，機能表現一覧の 16,801 表現 (あるいはその中の代表的表現) に対する用法判定を行う作業を進めている[4]．

機械学習により機能表現の検出を行う手法 [土屋 07b,

*Detecting Japanese Functional Expressions based on a Large Scale Hierarchical Lexicon

†Taiji Nagasaka, College of Engineering Systems, Third Cluster of Colleges, University of Tsukuba

‡Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba,

§Masatoshi Tsuchiya, Information and Media Center, Toyohashi University of Technology

[1]機能表現は，複数形態素からなる複合辞と一つの形態素からなる機能語から構成されるが，本稿では，複合辞と同等の意味で機能表現という用語を用いる．

[2]現時点では，用例集の 125 項目中，9 項目，用例データベースの 337 表現中，18 表現は，機能表現一覧には含まれていない．

[3]判定ラベル B のうち，形態素解析器 ChaSen(`http://chasen.naist.jp/hiki/ChaSen/`) による形態素解析結果の形態素区切りと交差する位置に機能表現の候補となる文字列が存在する場合は，判定ラベルを除外して扱う．

[4]機能表現一覧においては，あらゆる機能的用法が網羅されているため，今後作成する用法判定済コーパスにおいては，機能的用法を表す二つの判定ラベル F，M は，判定ラベル F に統合する予定である．

表 1: 機能表現辞書の 9 つの階層

階層		分類数	表現数 合計 (L^9 表現数)	助動詞型以外	助動詞型
L^1	見出し語	—	341 (488)	281	207
L^2	意味	88	435 (488)	281	207
L^3	派生 (格助詞型, 接続助詞型, 連体助詞型, 接続詞型, 助動詞型, 形式名詞型, とりたて詞型, 提題助詞型)	8	555	348	207
L^4	機能語の交替	—	774	492	282
L^5	音韻的変化	38	1,187	633	554
L^6	とりたて詞の挿入	18	1,810	659	1151
L^7	活用	—	6,870	659	6211
L^8	「です/ます」の有無	2	9,722	895	8827
L^9	表記のゆれ	—	16,801	1360	15411

表 2: 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				(1) 不平等条約を盾にとり、ゆすりに等しい権利を主張している。
Y	適切	不一致			(2) 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	(3) まな板にとってていね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	(4) 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	(5) 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	(6) 浜ノ島はあと一歩のところで勝ち星に結び付かず負け越した.

注連 07] においては，この人手付与済判定ラベルの情報を訓練用例として，形態素解析結果の形態素列をチャンキングして判定ラベル復元することにより機能表現の検出を行う.

4 代表的機能表現への言い換えによる大規模機能表現検出

本節では，機能表現一覧の階層において下位に位置する機能表現について，用法が類似するより上位の表現に言い換えた後，用法判定を行う方式の基本的考え方について説明する．本研究においては，階層の上位に位置する代表的表現は，L^3 階層もしくは L^4 階層相当の 1,000 表現程度の規模とする．以下では，助動詞型以外の機能表現の場合と，助動詞型の機能表現の場合に分けて，例を示す.

まず，助動詞型以外の機能表現の例として，とりたて詞型,「限定」の意味の表現「にかぎりまして」の場合，階層中では L^8 階層以下に位置するが，この表現の場合は，L^3 階層 ID を共有し階層中では L^3 階層以下に位置する表現「にかぎって」に言い換えた後，用法判定を行う.

(L^8 階層の表現) 彼 にかぎりまして，それはありません.
↓
(L^3 階層の表現) 彼 にかぎって，それはありません.

一方，助動詞型の場合は，活用を考慮して，活用形を保存したまま，代表的表現に言い換える．例として,「不可能」の意味の表現「わけにもいかず」の場合，活用した形であるため L^7 階層以下に位置し，その基本形は，L^6 階層以下に位置する「わけにもいかない」となる．これはさらに，L^3 階層 ID を共有し階層中では L^3 階層以下に位置する表現「わけにいかない」に言い換えられ，さらに,「わけにもいかず」と同じ活用形を持つ，L^7 階層以下の表現「わけにいかず」に言い換えた後，用法判定を行う.

(L^7 階層の表現) 帰る わけにもいかず，待った.
↓
(L^6 階層の表現) 帰る わけにもいかない，待った.
↓
(L^3 階層の表現) 帰る わけにいかない，待った.
↓
(L^7 階層の表現) 帰る わけにいかず，待った.

5 用法判定済コーパスを用いた調査

表 2 の判定ラベルを人手で付与した用法判定済コーパスを用いて，前節の手法の基本的な考え方が成り立つかどうかの調査を行った.

5.1 調査対象コーパス

用法判定済コーパスとしては，3 節で述べた用例データベース [土屋 06]，および，京都テキストコーパス中の機能表現に対して用法判定を行ったもの [土屋 07a] を用いた．現在，これらのコーパスに対して，用例データベースの 337 表現に限定せず，機能表現一覧の 16,801 表現 (あるいはその中の代表的表現) に対する用法判定を行う作業を進めているが，本稿執筆時点では，前者のコーパスに対して，L^3 階層に出現する 555 表現を対象として判定ラベルの付与が済んでいる．以下では，これらのコー

表 3: 用法判定済コーパスの文数・表現数・判定ラベル分布

コーパス	文数	表現数			判定ラベル分布 (表記単位)					
		トークン	タイプ (表記)	タイプ (ID)	F	A	M	C	Y	B
毎日新聞 (1995 年)	2743	8403	365	505	6900 (82.1%)	117 (1.5%)	355 (4.2%)	692 (8.2%)	27 (0.3%)	312 (3.7%)
京都テキストコーパス	14568	16736	285	389	12618 (75.4%)	205 (1.2%)	2550 (15.2%)	1137 (6.8%)	33 (0.2%)	193 (1.2%)
合計	17311	25139	461	618	19518 (77.6%)	322 (1.3%)	2905 (11.6%)	1829 (7.3%)	60 (0.2%)	505 (2.0%)

パスにおいて，用例データベースにおける表現 ID を持つ 337 表現，および，L^3 階層に出現する 555 表現のいずれにおいても，L^9 階層の 16,801 表現での ID に変換したうえで，調査を行った．表 3 に，これらのコーパスにおける文数，表現数，判定ラベル分布を，表 4 に，機能表現ごとの判定ラベル F，M の出現率の分布を，それぞれ示す[5]．

5.2 調査手順・結果

4 節で述べた言い換え方式の妥当性を検証するために，言い換えの対象となる表現対の間で判定ラベル分布の差を測定した．具体的には，$i = 4 \sim 9$ として，L^{i-1} 階層まで共通の機能表現一覧 ID を持ち，L^i 階層で異なる ID を持つ機能表現集合の組を収集し，それらの組の間で判定ラベル分布の差を測定した．機能表現集合の総頻度に下限値を設け，下限値を越える表現集合の組数，および，機能表現集合の対の間の判定ラベル分布の差が一定以上の組数を表 5 に示す．また，判定ラベル分布の差が一定以上の機能表現集合の組の抜粋を，判定ラベル分布 (FM/A/CYB の三分割) とともに表 6 に示す．

まず，L^8 階層まで共通の機能表現一覧 ID を持ち，L^9 階層で異なる ID を持つ機能表現集合の組では，仮名・漢字表記が異なるが，表 5 から分かるように，仮名・漢字表記の違いにより，判定ラベル分布の違いが生じる場合が相当数あることが分かる．そこで，以降の調査では，仮名表記の機能表現と漢字表記の機能表現は，異なる集合として扱った．より上位の階層では，L^7 階層における活用形の違いにより，判定ラベル分布の違いが生じる場合があるが，本研究の方式では活用形の違いは保持したまま言い換えを行うので，影響はないと言える．一方，最上位での，L^4 階層における機能語の交代の違いの場合は，判定ラベル分布の違いが生じる場合が相当数ある．したがって，本研究の方式においては，代表的表現としては，L^4 階層の表現を用いることとする．

[5] L^9 階層の 16,801 表現での ID に変換する際に，一つの表記に対して複数の ID が対応することがあり，その多義性の解消は現時点では行っていない．表 3 において表記単位でのタイプ数と ID 単位でのタイプ数が異なるのはこのためである．

表 4: 機能表現ごとの判定ラベル F,M の出現率の分布 (頻度 ≥ 50)

出現率 x	合計	助動詞型以外	助動詞型
x= 100%	31 (30.1%)	17 (27.4%)	14 (34.1%)
95% <x< 100%	23 (22.3%)	10 (16.1%)	13 (31.8%)
5% ≤x≤ 95%	48 (46.6%)	34 (54.9%)	14 (34.1%)
x< 5%	1 (1.0%)	1 (1.6%)	0 (0%)
計	103 (100.0%)	62 (100.0%)	41 (100.0%)

$$x = \frac{\text{判定ラベル F,M が付与された機能表現候補数}}{\text{機能表現候補数}}$$

表 5: 各階層でラベル分布の差が一定以上の表現組数

L^{i-1} 階層まで共通 ID，L^i 階層で分岐	仮名/漢字表記	頻度下限	表現組数 (助動詞型以外/助動詞型)	
			L^i 階層の各表現組の頻度総和が下限以上	L^i 階層の各表現組間のラベル分布の差が一定以上
$i = 4$	仮名	10	**19/2**	**4/1**
	漢字		**9/0**	**3/0**
$i = 5$	仮名		2/3	0
	漢字		0	0
$i = 6$	仮名	5	0/4	0/1
	漢字		0/1	0
$i = 7$	仮名		0/26	0/1
	漢字		0/8	0/1
$i = 8$	仮名		0	0
	漢字		0	0
$i = 9$	—		**12/10**	**8/5**

また，この判定ラベル分布の違いの調査とは別に，L^5 階層における音韻的変化において，前接形態素に対する制約の違いが生じる場合がある．774 組中 100 組程度において，「てならない」「でならない」のように，語頭が無声・有声のみの違いがある場合，前接する活用語の活用型が制限される．したがって，代表的表現に言い換える場合も，この制約の範囲内で代表的表現を選択する必要がある．

5.3 代表的表現への言い換え方式の設計

前節の調査結果をふまえて，4 節で述べた言い換え方式において，代表的表現および言い換えの際の制約を以下

表 6: 各階層においてラベル分布の差が一定以上の表現組

L^i 階層で分岐	代表的表記	判定ラベル分布 FM	A	CYB	代表的表記	判定ラベル分布 FM	A	CYB
助動詞型以外								
$i=4$	にとって	295 (97.4%)	5 (1.6%)	3 (1.0%)	にとり	18 (64.3%)	0 (0%)	10 (35.7%)
	にあたって	40 (97.6%)	0 (0%)	1 (2.4%)	にあたり	24 (61.5%)	0 (0%)	15 (38.5%)
	とすると	22 (56.4%)	5 (12.8%)	12 (30.8%)	とすれば	40 (80.0%)	10 (20.0%)	0 (0%)
	について	972 (99.5%)	1 (0.1%)	4 (0.4%)	につき	45 (75.0%)	0 (0%)	15 (25.0%)
	と言うと	1 (10.0%)	0 (0.1%)	9 (90.0%)	と言えば	19 (70.4%)	0 (0%)	8 (29.4%)
	に応じて	51 (68.0%)	0 (0%)	24 (32.0%)	に応じ	16 (30.8%)	0 (0%)	36 (69.2%)
	とは言っても	9 (60.0%)	4 (26.7%)	2 (13.3%)	と言って	2 (2.8%)	2 (2.8%)	68 (94.4%)
$i=9$	というと	44 (89.8%)	0 (0%)	5 (10.2%)	と言うと	1 (10.0%)	0 (0%)	9 (90.0%)
	ほか	15 (75.0%)	0 (0%)	5 (25.0%)	他	0 (0%)	0 (0%)	13 (100.0%)
	といって	22 (51.2%)	6 (13.9%)	15 (34.9%)	と言って	2 (2.8%)	2 (2.8%)	68 (94.4%)
	うえに	10 (58.8%)	0 (0%)	7 (41.2%)	上に	5 (5.7%)	0 (0%)	83 (94.3%)
助動詞型								
$i=4$	とよい	48 (66.6%)	0 (0%)	24 (33.3%)	ばよい	106 (100.0%)	0 (0%)	0 (0%)
$i=6$	てもよい	58 (100.0%)	0 (0%)	0 (0%)	てよい	168 (88.0%)	0 (0%)	23 (12.0%)
$i=7$	ていい	113 (96.6%)	0 (0%)	4 (3.4%)	てよかっ	1 (6.3%)	0 (0%)	15 (93.7%)
	得る	50 (50.0%)	0 (0%)	50 (50.0%)	得	45 (22.2%)	0 (0%)	158 (77.8%)
$i=9$	うる	10 (100.0%)	0 (0%)	0 (0%)	得る	50 (50.0%)	0 (0%)	50 (50.0%)
	てよい	45 (97.8%)	0 (0%)	1 (2.2%)	て良い	3 (60.0%)	0 (0%)	2 (40.0%)
	がいい	3 (13.6%)	0 (0%)	19 (86.4%)	が良い	0 (0%)	0 (0%)	6 (100.0%)

のように設計する．まず，L^4 階層に存在する機能表現 774 種類の集合を F^4 とする．そして，F^4 の各表現について，語頭を有声音化した表現が L^5 階層に存在する場合は，各表現につき一つずつ集めた集合を $P(F^4)$ とする．また，$F^4 \cup P(F^4)$ の各表現について，表記の一部が漢字となった表現を集めた集合を $K(F^4 \cup P(F^4))$ とする．そして，最終的に，以下の集合 $F^4_{p,k}$ を代表的表現の集合とする．

$$F^4_{p,k} \quad = \quad F^4 \cup P(F^4) \cup K(F^4 \cup P(F^4))$$

また，4 節で述べた言い換え方式において，機能表現一覧の階層中の下位の表現を代表的表現に言い換える際には，以下の制約を課す．

- 機能表現の語頭の無声・有声の制約により前接する活用語の活用型が制限される場合は，この制限を保持する．
- 機能表現の仮名表記・漢字表記の違いを保持する．

6 まとめと今後の課題

本稿では，機能表現一覧 [松吉 07b] の階層性を利用し，階層において下位に位置する機能表現について，用法が類似するより上位の表現に言い換えた後，用法判定を行う方式を提案した．そして，人手による用法判定済コーパスを用いた統計的調査を行い，その妥当性を検証した．今後は，4 節および 5.3 節の提案方式の実装を進める．関連研究として，[松吉 07a] では，機能表現一覧において意味的等価クラスを設定し，その範囲での言い換えを行う方式を提案している．この方式では，本稿で述べた用法判定結果の情報が付与済であるとして言い換えを行っている．[松吉 07a] では，意味的に等価な多様な表現への言い換えが目的であるのに対して，本稿では，用法が最も近い代表的表現への言い換えが目的である．

謝辞: 本研究に関して協力して頂いている京都大学情報学研究科知能情報学専攻 松吉俊氏に感謝する．

参考文献

[国研 01] 国立国語研究所: 現代語複合辞用例集 (2001).

[黒橋 97] 黒橋禎夫, 長尾眞: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).

[松吉 07a] 松吉俊, 佐藤理史：体系的機能表現辞書に基づく日本語機能表現の言い換え, 言語処理学会第 13 回年次大会論文集, pp. 899–902 (2007).

[松吉 07b] 松吉俊, 佐藤理史, 宇津呂武仁：日本語機能表現辞書の編纂, 自然言語処理, Vol. 14, No. 5, pp. 123–146 (2007).

[注連 07] 注連隆夫, 土屋雅稔, 松吉俊, 宇津呂武仁, 佐藤理史：日本語機能表現の自動検出と統計的係り受け解析への応用, 自然言語処理, Vol. 14, No. 5, pp. 167–197 (2007).

[土屋 06] 土屋雅稔, 宇津呂武仁, 松吉俊, 佐藤理史, 中川聖一：日本語複合辞用例データベースの作成と分析, 情報処理学会論文誌, Vol. 47, No. 6, pp. 1728–1741 (2006).

[土屋 07a] 土屋雅稔, 注連隆夫, 松吉俊, 宇津呂武仁, 佐藤理史, 中川聖一：機能表現を考慮した日本語係り受け解析器学習のためのコーパス作成, 言語処理学会第 13 回年次大会論文集, pp. 510–513 (2007).

[土屋 07b] 土屋雅稔, 注連隆夫, 高木俊宏, 内元清貴, 松吉俊, 宇津呂武仁, 佐藤理史, 中川聖一：機械学習を用いた日本語機能表現のチャンキング, 自然言語処理, Vol. 14, No. 1, pp. 111–138 (2007).