大規模階層辞書を利用した日本語機能表現の集約と解析*

長坂 泰治† 宇津呂 武仁† 松吉 俊‡ 土屋 雅稔§
筑波大学大学院 システム情報工学研究科†，奈良先端科学技術大学院大学 情報科学研究科‡
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．例えば，文 (1) と文 (2) には「にあたって」という表記の表現が共通して現れている．

(1) 出発する にあたって，荷物をチェックした．

(2) ボールは壁 にあたって，跳ね返った．

文 (1) では，下線部はひとかたまりとなって,「機会が来たのに当面して」という機能的な意味で用いられている．それに対して，文 (2) では，下線部に含まれている動詞「あたる」は，動詞「あたる」本来の内容的な意味で用いられている．こ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，(日本語機能表現一覧 [松吉 07]，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，[長坂 08] 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 (以下，派生的表現) について，用法が類似するより上位の表現 (以下，代表的表現) に言い換えた後，用法判定を行う方式を提案した．

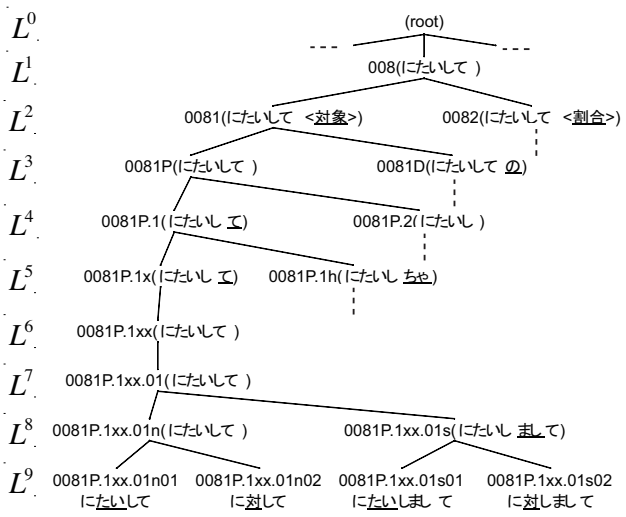


図 1: 機能表現辞書階層構造の一部

一方，本論文では，[長坂 08] の提案をふまえて，機能表現一覧中の情報のうち，特に文体の情報に注目し，代表的表現および派生的表現の区別を整理する．さらに，毎日新聞 1995 年分のテキストデータ中において，機能表現一覧の機能表現の出現頻度調査を行い，[土屋 07b, 注連 07] の機械学習による機能表現検出において必要となる訓練事例 (出現頻度 50 以上) が存在する機能表現の規模を推定する．

2 階層的機能表現辞書

機能表現一覧 [松吉 07] は，9つの階層構造をなしており，各階層は，表 1 に示されるような観点によって分類されている．同表に，各階層における機能表現数が示されており，図 1 に階層構造の一部をそれぞれ示す．

また，機能表現の文体に着目し，文体ごとの機能表現の振る舞いについて述べる．文体とは，機能表現一覧中の表現に付与されている情報であり，常体，堅い文体，口語体，敬体の 4 種類がある．表 2 にそれぞれの文体における表現例を示す．

---

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

†Taiji Nagasaka, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba,
‡Suguru Matsuyoshi, Graduate School of Information, Nara Institute of Science and Technology,
§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．

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: 文体の種類

文体	表現例
常体	について
堅い文体	につき
口語体	についちゃ
敬語体	につきまして

3 代表的表現への集約

3.1 基本的な考え方

[長坂 08] で提案した代表的表現への集約方式においては，階層の上位に位置する代表的表現は，L^4 階層相当の 1,000 表現程度の規模とする．そして，機能表現一覧において，代表的表現を除く表現はすべて，言い換えの対象の表現となる．本研究では，これらの表現を派生的表現と定義する．派生的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．
- 助動詞型の機能表現の場合には，言い換え前後で活用形を保持する．

3.2 文体ごとの機能表現数

[長坂 08] では，助動詞型の場合，派生的表現に 3,000 表現を，代表的表現 500 表現に集約した．一方，助動詞型以外の場合，文体を常体および堅い文体で見てみると，代表的表現が 353 表現であるのに対して，派生的表現が 94 表現と少ないため，集約方式の効果はほとんど得られない．そこで，本論文では，助動詞型以外の機能表現のうち，常体および堅い文体の表現は，敬体や口語体の派生的表現に対する代表的表現を除いて，集約方式の対象表現から外す．

以上をふまえて，各体ごとに，代表的表現，派生的表現の数を示したものを，図 2[2] に示す．両図の節点の数字は，機能表現の表記数を表している．また，枝の数字は，端点における体の間で言い換え組を形成する表現の数である．

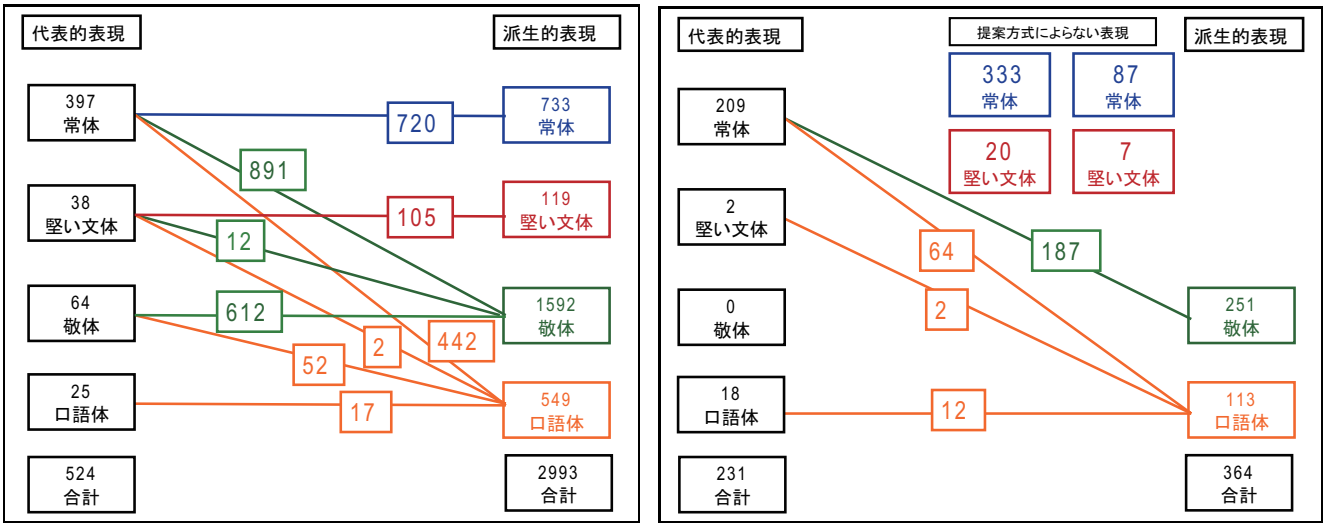
4 新聞記事における機能表現数の分布

毎日新聞 1995 年の一年分の中で，50 回以上出現する表現を調べた結果を表 3 に示す．表 3 の中には，機能表現一覧の中で，同一の表記に対して複数の ID が存在する表現があるので，そういった表現は表記単位の数字を出している．表 4 には，複数の ID を持つ機能表現の表記数の一覧を示す．

表 3 より，全 602 表現のうち，常体の機能表現が 467 表現と大きい割合を占めていることがであること分かる．表 3 の中で，代表的表現と派生的表現の関係にある組の数および例を表 5 に示す．ここで，[土屋 06] の用例データベースにおいては，機能的用法および自立的用法の両方が新聞記事中で適度にバランスして出現する機能表現の範囲は，収録された全機能表現のうちの約 3 分の 1 程度であることが判明している．したがって，上記の全 602 表現中で，二つの用法が適度にバランスして出現しており，機械学習によって用法判定のための検出器の学習を行う必要があるものは，200 表現程度であると推定される．

これらの表現から，毎日新聞 1995 年の中で，すでに各表現につき 50 用例以上用法判定した表現を取り除くと，441 表現になる．そして，各表現 50 用例の用法判定をする必要があるが，441×50=22,050 箇所になる．さらに，これからすでに用法判定した分を差し引くと，20,145 箇

[2]助動詞型では，活用形を基本形に限定した数値を算出している．



(a) (助動詞型基本形)　　(b) (助動詞型以外)

図 2: 代表的表現への集約における文体ごとの機能表現数

表 3: 毎日新聞 1995 年において，50 回以上出現する機能表現数の分布

	助動詞型 (基本形)		助動詞型以外			
	代表的表現	派生的表現	代表的表現	派生的表現	その他	合計
常体	164	38	87	0	178	467
堅い文体	8	3	0	0	9	20
敬体	14	42	0	1	0	57
口語体	7	37	1	13	0	58
合計	193	120	88	14	187	602

所になる．[土屋 06] では，1 表現 50 箇所あたりの用法判定にかかる一人の作業量は 12 分と報告されている．したがって，約 27 時間となる．

以上の人手判定作業を経て，さらに，機械学習のための用法判定済み訓練データの整備を別途行えば，代表的表現への集約を行う用法判定方式，および，集約を行わない用法判定方式の定量的評価が可能となる．

また，表 6 には，毎日新聞 1995 年分を用いて代表的表現の用法判定のための検出器の教師あり学習を行った場合に，集約方式によって検出対象となる派生的表現のうち，毎日新聞 1995 年分に 50 回以上出現しない表現の数を示す．これらの合計 1,213 表現は，毎日新聞 1995 年分に 50 回以上出現しないため，毎日新聞 1995 年分の範囲では，用法判定のための検出器の教師あり学習ができない．しかし，代表的表現への集約方式を用いれば，代表的表現の集約を介することにより，用法判定が可能となる．

5　関連研究

[松吉 08] においては，機能表現一覧 [松吉 07] 中の機能表現を対象として，意味を保存する言い換えが可能な機能表現の分類を規定している．一方，本論文では，機能表現の用法判定の性能を保ったまま，代表的表現への言い換えを行うという，より緩い制約のもとでの機能表現の言い換えが目的である．また，代表的表現への言い換えを介した機械翻訳の研究としては，内容語と口語的な機能表現を扱った [山本 01, 山本 02]，機能表現一覧 [松吉 07] の機能表現を対象とした [坂本 09] がある．

6　まとめ

本稿では，機能表現に文体を導入して，新聞記事で出現する機能表現の，文体ごとに見られる特徴について，報告した．そして，50 回以上出現する機能表現の各用例文に対して人手で用法判定を行うための作業量の見積もりについて報告した．今後は，人手による判定作業を実行する．また，その作業結果を受けて提案方式の実装，

表 4: 複数の ID を持つ機能表現表記数 (機能表現一覧全体/毎日新聞 1995 年分で頻度 50 以上)

	助動詞型 (基本形)		助動詞型以外			
	代表的表現	派生的表現	代表的表現	派生的表現	その他	合計
常体	57/44	13/5	27/19	0/0	136/73	233/141
堅い文体	6/3	14/0	0/0	0/0	12/5	32/8
敬体	2/0	72/8	0/0	64/1	0/0	136/9
口語体	7/4	33/13	3/1	35/9	0/0	78/27
合計	72/51	132/26	30/20	99/10	148/78	479/185

表 5: 毎日新聞 1995 年において，50 回以上出現する代表的表現と派生的表現の組数の分布および組の例 (助動詞型基本形)

文体の組み合わせ	組数	組の例
常体の代表的表現-常体の代表的表現	25	てよい-てもよい
常体の代表的表現-敬体の代表的表現	30	なければならない-なければならないです
常体の代表的表現-口語体の代表的表現	20	ても仕方がない-ても仕方ない
合計	75	

表 6: 毎日新聞 1995 年に 50 回以上出現する代表的表現に対する派生的表現のうち，毎日新聞 1995 年に 50 回以上出現するものを除いた表現の数

文体	助動詞型 (基本形)	助動詞型以外
常体	347	8
堅い文体	52	2
敬体	537	78
口語体	164	25
合計	1100	113

および評価を行う．

参考文献

[国研 01] 国立国語研究所: 現代語複合辞用例集 (2001).

[松吉 07] 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).

[松吉 08] 松吉俊, 佐藤理史：文体と難易度を制御可能な日本語機能表現の言い換え, 自然言語処理, Vol. 15, No. 2, pp. 75–99 (2008).

[長坂 08] 長坂泰治, 宇津呂武仁, 土屋雅稔：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, 言語処理学会第 14 回年次大会論文集, pp. 837–840 (2008).

[坂本 09] 坂本明子, 宇津呂武仁, 松吉俊：日本語機能表現の集約的英訳, 言語処理学会第 15 回年次大会論文集 (2009).

[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).

[山本 01] 山本和英, 白井諭, 坂本仁, 張玉潔：SANDGLASS: 両言語換言機構を基軸とする音声翻訳, 言語処理学会 第 7 回年次大会発表論文集, pp. 221–224 (2001).

[山本 02] 山本和英：換言と言語変換の協調による機械翻訳モデル, 言語処理学会第 8 回年次大会発表論文集, pp. 307–310 (2002).