研究活動・成果の総括：日本語機能表現班
大規模階層辞書を用いた日本語機能表現解析体系の研究

宇津呂武仁　(班長：筑波大学大学院システム情報工学研究科)*
鈴木敬文　(協力者：筑波大学大学院システム情報工学研究科)
島内蘭　(協力者：筑波大学大学院システム情報工学研究科)
阿部佑亮　(協力者：筑波大学大学院システム情報工学研究科)
松吉俊　(協力者：奈良先端科学技術大学院大学情報科学研究科)
土屋雅稔　(協力者：豊橋技術科学大学情報メディア基盤センター)

Final Progress Report: 'Japanese Functional Expressions' Group

Takehito Utsuro　(University of Tsukuba)
Takafumi Suzuki　(University of Tsukuba)
Ran Shimanouchi　(University of Tsukuba)
Yusuke Abe　(University of Tsukuba)
Suguru Matsuyoshi　(Nara Institute of Science and Technology)
Masatoshi Tsuchiya　(Toyohashi University of Technology)

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，このような日本語機能表現の解析の課題に対して，これまでに，国立国語研「現代語複合辞用例集」 [国研 01] に収録されている 125 機能表現の異表記を展開した 300 表現について，新聞記事中の用例に対して機能的用法・内容的用法を判別した用例データベース [土屋 06] を作成・公開した．また，機能的・内容的用法の自動判別ツールを作成し，係り受け解析ツールとの統合により，複合辞としての機能的用法を考慮した係り受け解析を実現した [注連 07]．また，日本語機能表現の全表記を網羅した辞書として，日本語機能表現の全表記約 17,000 を網羅的に収録した「つつじ」 [松吉 07, 松吉 08][2] が公開されたのを受けて，17,000 表現全てを対象とした機能的・内容的用法の判定方式を提案した [長坂 08].

*utsuro @ iit.tsukuba.ac.jp

[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]http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/

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	45/128/199	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

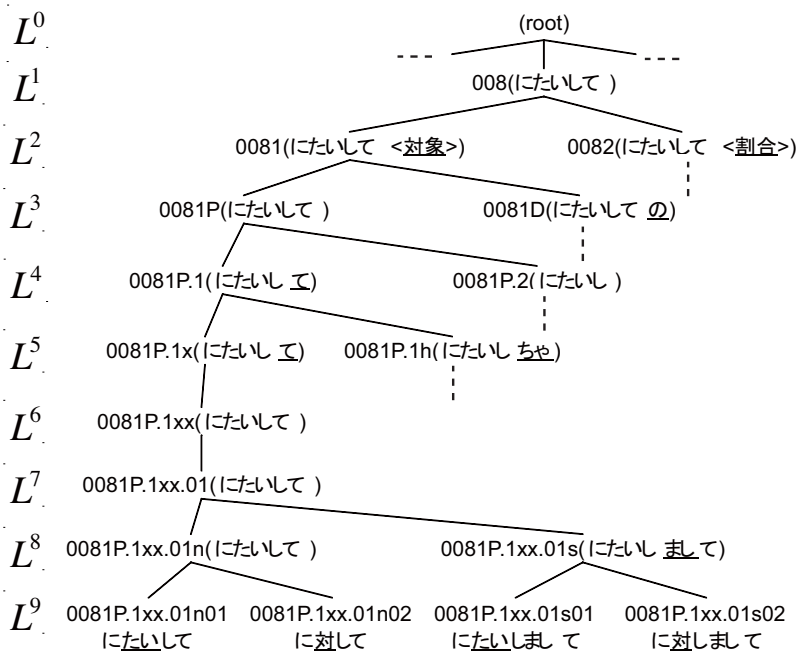

図 1: 機能表現辞書階層構造の一部

本研究では，この提案方式をふまえて，日本語機能表現の全表記約 17,000 を網羅的に収録した辞書「つつじ」の階層的構造および言語学的特性を活用して，網羅的な日本語機能表現の解析，および，日本語機能表現の集約的翻訳の枠組みを実現した.

2. 日本語機能表現一覧「つつじ」

代表的な機能表現の規模を超えて機能表現の表記を網羅的に列挙した辞書を設計・編纂することを目的として，日本語機能表現一覧「つつじ」 [松吉 07] が編纂された.「機能表現一覧」においては，日本語における機能表現の表記を網羅することを目的として，機能表現の構成要素の組み合わせとして，機能表現の異形を階層的に収録している．表 1 および図 1 に示すように，全体としては，形態に基づいて，全機能表現の表記の集合が 9 つの階層構造によって構成されている．階層の上位に

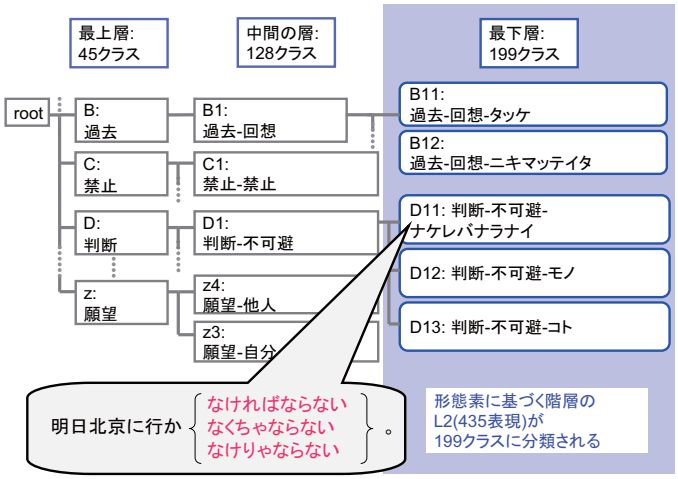

図 2: 日本語機能表現一覧「つつじ」: 意味的等価クラスの一部

は，341 種類の機能表現を見出し語として配置し，意味の違い，機能表現末尾の活用，機能表現の各構成要素の音韻的変化，とりたて詞の挿入，口語表現・敬語表現の言い換えなどによる異形として，16,801 表現が収録されている．また，機能表現の意味的な分類は，図 2 に示す 3 階層の体系によって構成されている [松吉 08]．この階層の最下層に位置する全 199 個の各意味的等価クラスに属する機能表現は，一定の文脈のもとで言い換え可能であるとされている．また，機能表現の文体については，常体，敬体，口語体，堅い文体の 4 種類の文体を区別して，各表現に付与している．

3. 派生関係及び用例を利用した日本語機能表現の解析

3.1 概要

[鈴木 10, 鈴木 11] において,「つつじ」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) の用例を参照して，用法判定を行う方式を提案した．

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 [長坂 08]．そして,「機能表現一覧」において，代表的表現を除く表現を派生的表現と定義する．ただし，代表的表現を選定する際には，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，活用形を保持する．

この提案方式においては，前後の形態素の品詞が代表・派生間において不変の場合には，代表的表現と派生的表現の間で用法の傾向に相関がある，という点に注目する．さらに，前後の形態素品詞に加え，代表的表現と派生的表現の間で，機能表現の表記を構成する形態素列の品詞パターンの間に派生関係があるという特性を利用する．提案方式に基づいて，派生的表現の用法の分析を行った結果，代表的表現の表記の用法判定済み用例集合 (約 38,000 例) を参照して，派生的表現の表記の用法判定を行うことにより，80%以上の用例の用法を正しく判定できることが分かった．

3.2 派生的な表現の解析方式

以下では，代表的表現の表記の用法判定済用例集合 S_c^{tr} を参照して，派生的表現の表記の用法判定を行う方式について述べる．

3.2.1 機能表現表記照合個所の表現形式

まず，一文中で，機能表現表記と文字列照合する個所を $e = \langle f, l, r \rangle$ (ただし，f は機能表現表記，l は機能表現表記の先頭の文字位置，r は末尾の文字位置) によって表現する[3]．このとき，評価用の文において機能表現表記 f_{ts} と照合した個所を $e_{ts} = \langle f_{ts}, l_{ts}, r_{ts} \rangle$ とし，e_{ts} に前接する形態素を m_{-1}^{ts}，後接する形態素を m_{+1}^{ts} とする．一般には，f_{ts} の可能性としては，派生的表現の表記 f_d の場合，および，代表的表現の表記 f_c の場合の二通りが考えられる．ここで，f_{ts} が派生的表現 f_d の場合には，f_d の代表的表現 f'_c の用例が，用法判定済用例集合 S_c^{tr} 中の機能表現表記照合個所の一つ $e_{tr} = \langle f'_c, l_{tr}, r_{tr} \rangle$ となる．一方，f_{ts} が代表的表現 f_c の場合には，f_c 自身の用例が，用法判定済用例集合 S_c^{tr} 中の機能表現表記照合個所の一つ $e_{tr} = \langle f_c, l_{tr}, r_{tr} \rangle$ となる．いずれの場合も，e_{tr} に前接する形態素を m_{-1}^{tr}，後接する形態素を m_{+1}^{tr} とする．

ここで，次節の解析手順においては，評価用の文における用法判定対象個所の単位として，相互に重複して連続する複数の機能表現表記から構成される列をひとまとめとして，機能表現表記列の用法判定を一括して行う．具体的には，評価用の文において，連続する 2 個の機能表現表記の文字列のうちの少なくとも一部が重複するような機能表現表記列 $E = e_i, \ldots, e_k$ (すなわち，機能表現表記列 $E = e_i, \ldots, e_k$ 中における連続する任意の 2 個の機能表現表記の組 $e_j,\ e_{j+1}$ において表記の文字列の少なくとも一部が重複する: $l_j < l_{j+1} < r_j < r_{j+1}$) をひとまとめとする．

3.2.2 解析手順

まず，評価用の文における用法判定の単位である機能表現表記列 $E = e_i, \ldots, e_k$ に対して，以下の条件**「前後形態素が類似する用法判定済用例の存在」**の成否を判定する．

> **「前後形態素が類似する用法判定済用例の存在」**
> $E = e_i, \ldots, e_k$ 中で，少なくとも一つの機能表現表記照合個所 $e_{ts} = \langle f_{ts}, l_{ts}, r_{ts} \rangle$ に対して，機能表現表記 f_{ts} に対応する機能表現表記照合個所 e_{tr} が用法判定済用例集合 S_c^{tr} 中に存在する．さらに，前接形態素 m_{-1}^{ts} と m_{-1}^{tr}，および，後接形態素 m_{+1}^{ts} と m_{+1}^{tr} の間で，それぞれ，品詞大分類[4] が一致する．

そして，この成否に応じて，下記の手順 (I) もしくは (II) を行う．

(I) **「前後形態素が類似する用法判定済用例の存在」**が成り立たない場合，機能表現表記列 $E = e_i, \ldots, e_k$ 中の全ての機能表現表記が内容的用法であると判定して終了する．

(II) **「前後形態素が類似する用法判定済用例の存在」**が成り立つ場合，以下を行う．

(II-i) 条件「機能表現表記列 E において，最長の表記となる照合個所 e_{ts} がただ一つである．さらに，e_{ts} に対して，用法判定済用例集合 S_c^{tr} 中の対応する機能表現表記照合個所 e_{tr} (複数個所の場合もあり得る) を参照することにより，e_{tr} に対する用法判定結果 l_{tr} が一意に決まる.」が成り立つならば，機能表現表記列 E に対して,「e_{ts} の用法は l_{tr}，E 中のその他の機能表現表記の用法は内容的用法」を採用して終了する．その他の場合は，(II-ii) を行う．

---

[3]ただし，機能表現表記 f としては,「機能表現一覧」 [松吉 07] における一文字表記の機能語は除外する．
[4]IPAdic (`http://sourceforge.jp/projects/ipadic/`) を用いる．

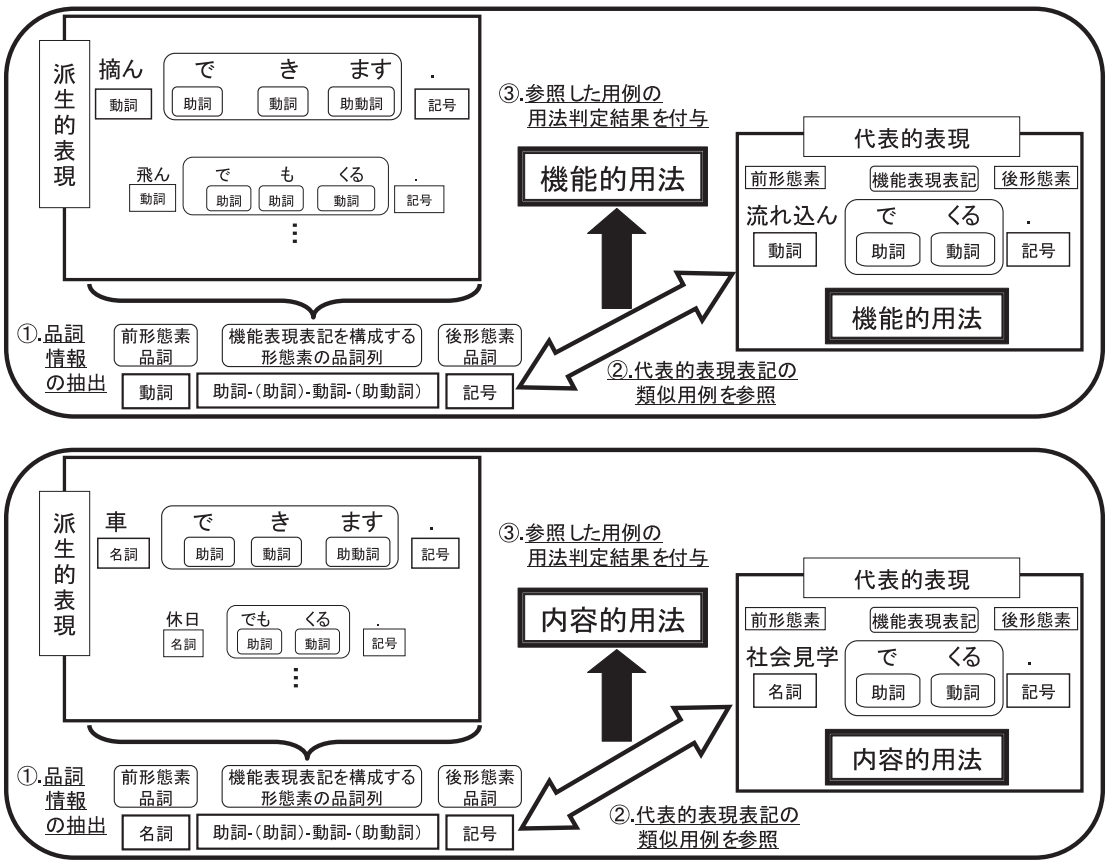

図 3: 模式図: 「代表的表現の表記の用例」を参照して「派生的表現の表記の用例」の用法を判定

(II-ii) 条件**「前後形態素が類似する用法判定済用例の存在」**において,「前後形態素の品詞大分類の一致」の代わりに「前後形態素の品詞細分類が一定以上の基準で類似する」を課し，(II-i) と同様の手順を行う．機能表現表記列 E に対する用法判定結果が一意に決まらない場合には，(II-iii) を行う．

(II-iii) 条件**「前後形態素が類似する用法判定済用例の存在」**において,「前後形態素の品詞大分類の一致」の代わりに「機能表現表記を構成する形態素の品詞列が一定以上の基準で類似する」を課し，(II-i) と同様の手順を行う．機能表現表記列 E に対する用法判定結果が一意に決まらない場合には,「不正解」と判定し終了する．

以上の手順にしたがって，派生的表現の表記の用法が機能的用法であると判定した例の模式図を図 3 上半分に，内容的用法であると判定した例の模式図を図 3 下半分に，それぞれ示す．

3.4 評価

代表的表現の表記の用法判定済用例としては，毎日新聞 1995 年の 1 年分から収集して人手で機能表現表記の用法判定を行った約 38,000 用例を参照することとする．評価対象としては，同じく毎日新聞 1995 年の 1 年分のうち，機能的用法と内容的用法として適度な割合で新聞記事内に出現する代表的表現に対して，用例数が 10 例以上となる派生的表現を中心に収集した 1,882 用例，及び，機能的用法に偏って新聞記事内に出現する代表的表現に対して，用例数が 50 例未満となる派生的表現を中心に収集した 916 用例の計 2,798 用例 (243 表現) を評価対象とする．

表 2: 派生関係及び用例を利用した機能表現の解析: 評価結果

(a) 代表的表現の用例を参照する手法

類型		割合 (%)	
「3.2.2 節の手順 (II)」前後の形態素の品詞もしくは機能表現表記を構成する形態素の品詞列の条件を満たす代表的表現の用法判定結果を採用し正解		71.6	
「3.2.2 節の手順 (I)」前後の形態素の品詞が一致する代表的表現が存在しないため，内容的用法と判定し正解		10.9	
不正解	適切な作例をすることにより正解可能	13.2	17.5
	作例しても正解不可能	4.3	
合計		100	

(b)「代表的表現の用例+左・右接続接続情報を参照する手法

類型		割合 (%)	
「3.2.2 節の手順 (II)」において用法判定済用例の一つとして左・右接続情報を追加して正解		77.0	
「3.2.2 節の手順 (I)」により正解		10.0	
不正解	適切な作例をすることにより正解可能	8.1	
	作例しても正解不可能	4.9	13.0
合計		100	

評価結果を表 2(a) に示す．また，3.2.2 節のいずれかの手順における判定結果が「不正解」となる場合について，代表的表現の適切な用例を作成して用法判定済用例集合 S_c^{tr} に追加した場合に，正解可能か否かの分析を行った結果も併せて示す．この結果から分かるように,「適切な用例の作例なしで正解」となる割合は約 82%，作例を許す場合は約 95%である．

また，表 2(b) には，用法判定済用例集合 S_c^{tr} に対して，用法判定済用例の一つとして，左・右接続情報 [松吉 07, 松吉 08] を追加した場合の評価結果を示す．左・右接続情報とは，機能表現表記の用法が機能的用法である場合の情報である．左接続情報は，直前に接続可能な形態素の情報を示しており，右接続情報[5]は，機能表現表記を構成する末尾の形態素の情報を示したものである．これらは「機能表現一覧」 [松吉 07] において，各機能表現ごとに定義されており，53 種類の左接続情報，および，51 種類の右接続情報が掲載されている．これらの左・右接続情報を追加した場合,「適切な用例の作例なしで正解」となる割合は，約 87%に改善する．

4. 日本語機能表現の集約的翻訳

日本語には 16,000 種類以上の機能表現の異形が存在する．従来の機械翻訳ソフトは，日本語機能表現の異形に対して個別に訳語を割り当てる手法を用いていると考えられるが，この手法では全ての異形を網羅することが困難である．そのため，日本語入力文中に，翻訳規則が未定義の機能表現の異形が存在した場合に，その表現を正しく翻訳できないという問題を抱えていた．そこで，本研究では，日本語機能表現の異形を網羅的に機械翻訳するために，類似する意味を持つ日本語機能表現を予め 1 つのクラスにまとめ，各クラスに対して 1 つの集約的な翻訳規則を作成する手法を提案した．機能表現の意味クラスとしては,「つつじ」の意味的等価クラス (199 クラス) を用いた．

以上の考え方に基づき，[坂本 09b, Sakamoto09a] では，日本語学習者向けの機能表現用例集 [グループ・ジャマシイ 98]，及び新聞記事テキストから，各意味的等価クラスに含まれる機能表現が出現した例文が十分な数収集できた 91 クラスを対象として，それらの機能表現の集約的英訳可能性を検証した．その結果，49 クラスについては，1 クラスに対して 1 規則で英訳可能となったが，その他の

[5]右接続情報に加えて，IPAdic を用いて形態素解析を行った場合の形態素列の情報を参照することにより，機能表現表記の直後に接続可能な形態素の情報が得られる．

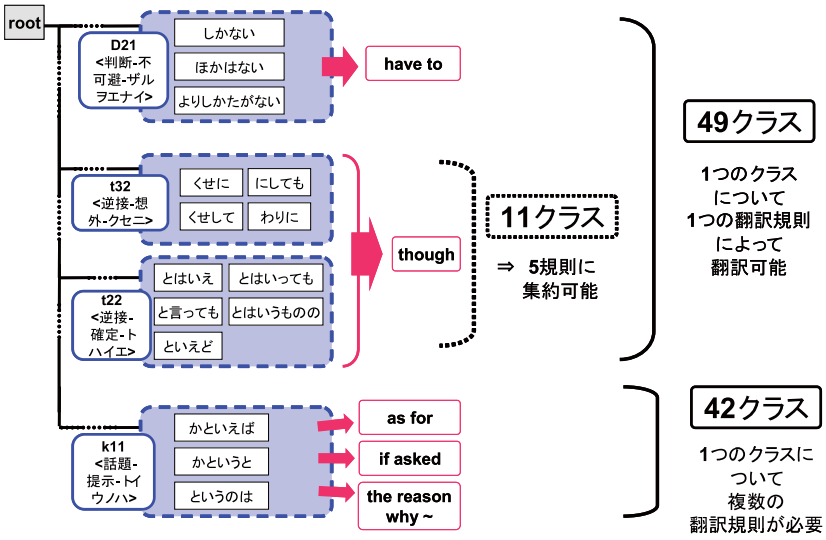

図 4: 集約的英訳可能性に基づく意味的等価クラスの粒度の再編

42 クラスについては，1 クラスに対して複数の英訳規則が必要であることが明らかになった (図 4). 同様に，[劉 10] においては，日本語学習者向けの機能表現用例集 [グループ・ジャマシイ 98] の中国語訳を日中対訳コーパスとして利用し，日本語機能表現の集約的中国語訳規則の作成・評価を行った. 一方，[Nagasaka10, 島内 11] では，日英対訳特許文を対象として，日本語機能表現の集約的英訳規則の作成および評価を行った．この研究では，NTCIR-7 の特許翻訳タスクで配布された 1,798,571 件の日英対訳特許文対に対して統計的機械翻訳モデルを適用することによりフレーズテーブルを学習し，日英対訳機能表現対を獲得するために用いた．特許文の場合は，使用される機能表現の意味範囲が狭く，その種類も少ないので，翻訳規則作成が容易である点が大きな利点となる．対象として，「つつじ」の 199 意味的等価クラスの中で，91 意味的等価クラスに属する日本語機能表現について，翻訳規則を作成し，その中の意味的等価クラス 12 個に属する日本語機能表現について評価を行なった結果，96.6%の正解率を得ることが出来た.

5. おわりに

本研究では,「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派生的表現について，用法が類似するより上位の代表的表現の用例を参照して，用法判定を行う手法を実現した．また,「つつじ」の意味的等価クラスを利用して，日本語機能表現の集約的翻訳を実現した．今後の課題としては，以下が挙げられる．まず，提案方式では，代表的表現の用例，派生的表現の用例のいずれについても，できるだけ多くの用法の用例を収集し，用法判定結果を付与した用例集合を蓄積することが性能改善の鍵を握る．そこでは，大規模な未解析テキストコーパスを情報源として，機能表現表記の前後の形態素の品詞のバリエーションをできるだけ多く収集し，サンプリングして用法判定結果を付与することが最も効果的である．また，新聞記事を対象として構築した代表的表現の表記の用法判定済み用例集合を参照して，新聞記事以外の多様なジャンルのテキスト中の機能表現表記の用法判定を行うタスクにおいて，提案方式の有効性を評価する必要がある．さらに，機能表現を考慮した文解析を高度化する目的においては，機能表現の検出・係り受け解析と格構造解析を統合する方式を確立することが必要である．情報抽出・テキストマイニング・評判抽出・質問応答・含意認識等の応用の観点からは，機能表現が担う多様なアスペクト・モダリティの同定が不可欠であり，これまでの研究成果 (例えば，[江口 10]) をふまえて，多方面の応用における発展が期待される．一方，日本語機能表現の集約的翻訳を組み込んだ機械翻訳手法を実現するためには，多義性を持った機

能表現の意味的曖昧性を解消する方式の確立が不可欠である．

参考文献

[グループ・ジャマシイ 98] 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).

[国研 01] 国立国語研究所: 現代語複合辞用例集 (2001).

[劉 10] 劉颯, 長坂泰治, 宇津呂武仁, 松吉俊：意味的等価クラスを用いた日本語機能表現の集約的日中翻訳規則の作成と分析, 言語処理学会第 16 回年次大会論文集, pp. 194–197 (2010).

[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).

[江口 10] 江口萌, 松吉俊, 佐尾ちとせ, 乾健太郎, 松本裕治：モダリティ，真偽情報，価値情報を統合した拡張モダリティ解析, 言語処理学会第 16 回年次大会論文集, pp. 852–855 (2010).

[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).

[Nagasaka10] Nagasaka, T., Shimanouchi, R., Sakamoto, A., Suzuki, T., Morishita, Y., Utsuro, T. and Matsuyoshi, S.: Utilizing Semantic Equivalence Classes of Japanese Functional Expressions in Translation Rule Acquisition from Parallel Patent Sentences, *Proc. 7th LREC*, pp. 1778–1785 (2010).

[Sakamoto09a] Sakamoto, A., Nagasaka, T., Utsuro, T. and Matsuyoshi, S.: Identifying and Utilizing the Class of Monosemous Japanese Functional Expressions in Machine Translation, *Proc. 23rd PACLIC*, pp. 803–810 (2009).

[坂本 09b] 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, pp. 654–657 (2009).

[島内 11] 島内蘭, 阿部佑亮, 鈴木敬文, 宇津呂武仁, 松吉俊：特許文における日本語機能表現の集約的英訳規則の作成と評価, 言語処理学会第 17 回年次大会論文集 (2011).

[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).

[鈴木 10] 鈴木敬文, 宇津呂武仁, 松吉俊, 土屋雅稔：代表・派生関係を利用した日本語機能表現の解析, 情報処理学会研究報告, Vol. 2010, No. (2010–NL–199) (2010).

[鈴木 11] 鈴木敬文, 宇津呂武仁, 松吉俊, 土屋雅稔：代表・派生関係および用例を利用した日本語機能表現の解析, 言語処理学会第 17 回年次大会論文集 (2011).

[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).