自然言語処理における日本語機能表現の解析*

宇津呂 武仁† 松吉 俊‡ 土屋 雅稔§ 鈴木 敬文† 島内 蘭†
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 (以下，**機能的用法**) 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 (以下，**内容的用法**) で用いられている．このような表記においては，機能的用法として用いられている場合と，内容的用法として用いられている場合とを識別する必要がある．

本稿では，日本語文の自然言語処理において，機能表現となり得る表記に対して，機能的用法か内容的用法かを同定する解析を中心として，これまでに筆者らが行った研究を紹介する．本稿で紹介する研究は，主として，以下に類別される．

I. 研究の対象とする機能表現の表記の一覧を規定する．日本語機能表現の解析に関する研究の初期の段階においては，代表的な機能表現およびその派生形 (合計数百表記程度) を研究対象としたが (2 節)，研究が進むにつれて，日本語における機能表現の表記を網羅する方向へと研究の焦点が移った (3 節).

II. I. で規定した機能表現の表記に対して，文 (1) および (2) に示すような用例を収集し，機能的用法，内容的用法の区別を人手で付与した用例データベースを作成する (2 節).

III. 機能表現となり得る表記の用例に対して，その表記が機能的用法として用いられているのか，それとも，内容的用法として用いられているのか，の区別を自動解析 (識別，あるいは，機能表現の自動検出) する (4 節).

IV. 機能表現を含む文に対して，機能表現の検出を行った上で，文全体の係り受け解析を行う (5 節).

---

*Analysis of Japanese Functional Expressions in Natural Language Processing

†Takehito Utsuro, Takafumi Suzuki, Ran Shimanouchi, Graduate School of Systems and Information Engineering, University of Tsukuba,

‡Suguru Matsuyoshi, Graduate School of Information Science, 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．

◇ **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: 現代語複合辞用例集の項目例

V. 機能表現の自動解析の応用例の一つとして，類似する意味を持った多数の派生形を集約し，英語・中国語等の外国語における代表的な訳語に翻訳する (6 節).

以下の各節においては，これらの各項目について紹介する.

2 日本語複合辞用例データベース

日本語の機能表現を列挙したリストとしては，従来より,「日本語表現文型」 [森田 89],「日本語文型辞典」 [グループ・ジャマシイ 98] 等が知られていた．これに対して,「現代語複合辞用例集」 [国研 01] は，この二つのリストに収録されている複合辞の一覧対照表を作成した上で,「日本語表現文型」に収録されている複合辞を基本とし，その中でも一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複合辞 125 項目を選定し収録している．図 1 に示すように，それぞれの項目は,「A56 〜 にとって・ 〜 にとり」というような見出しと,「接続」「意味・用法」「文法」「ノート」といった説明文，および用例を含む.

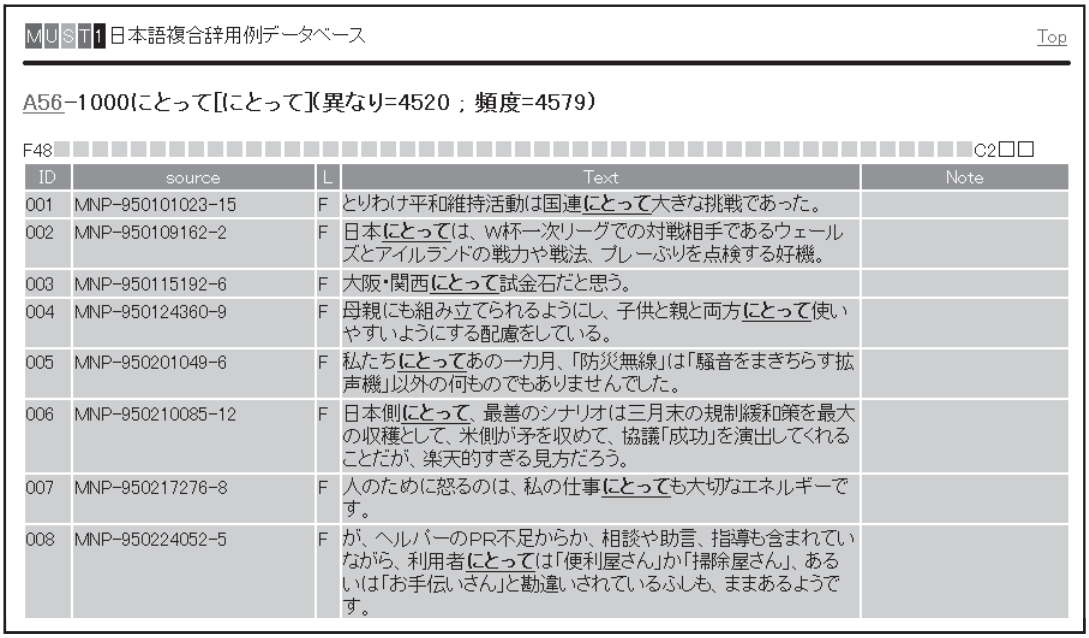
これらのリストをふまえて，我々は，機能表現となり得る表記の用例のデータベースに収録する複合辞の表記のリストを選定した．具体的には，現代語複合辞用例集を基礎資料とし，125 項目のうちの 123 項目の代表的複合辞の派生形である 337 種類の機能表現を規定し，その用例データベース (日本語複合辞用例データベース [土屋 06][2]) を作成した．このデータベースにおいては，機能表現となり得る 337 種類の各表記に対して，毎日新聞 1995 年から最大 50 用例を収集し，図 2 に示すように，各表記が機能的用法であるか内容的用法であるかの区別を人手で付与した.

3 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][3] 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

[2] http://nlp.iit.tsukuba.ac.jp/must/

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



MUST1 日本語複合辞用例データベース Top

A56−1000にとって[にとって](異なり=4520 ; 頻度=4579)

F48 C2□□

ID	source	L	Text	Note
001	MNP−950101023−15	F	とりわけ平和維持活動は国連にとって大きな挑戦であった。	
002	MNP−950109162−2	F	日本にとっては、W杯一次リーグでの対戦相手であるウェールズとアイルランドの戦力や戦法、プレーぶりを点検する好機。	
003	MNP−950115192−6	F	大阪・関西にとって試金石だと思う。	
004	MNP−950124360−9	F	母親にも組み立てられるようにし、子供と親と両方にとって使いやすいようにする配慮をしている。	
005	MNP−950201049−6	F	私たちにとってあの一カ月、「防災無線」は「騒音をまきちらす拡声機」以外の何ものでもありませんでした。	
006	MNP−950210085−12	F	日本側にとって、最善のシナリオは三月末の規制緩和策を最大の収穫として、米側が矛を収めて、協議「成功」を演出してくれることだが、楽天的すぎる見方だろう。	
007	MNP−950217276−8	F	人のために怒るのは、私の仕事にとっても大切なエネルギーです。	
008	MNP−950224052−5	F	が、ヘルパーのPR不足からか、相談や助言、指導も含まれていながら、利用者にとっては「便利屋さん」か「掃除屋さん」、あるいは「お手伝いさん」と勘違いされているふしも、ままあるようです。	

(a) 機能的用法の用例

012 MNP−950326072−6 C 女性がふと漏らしたり、本の中の女性ならではの言葉を毎回一つ取り上げ、それをコラム(七百三十五字)のタイトルにとって、ひずみやごまかしを直視しながら一九八五年四月号からまとめてきた。

048 MNP−951128259−12 C 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のである。

(c) Group MUST, 2005. Generated by subentry2html at Mon Oct 9 10:03:08 2006.

(b) 内容的用法の用例

図 2: 日本語複合辞用例データベースの用例の抜粋

表 1: 「機能表現一覧」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	文法機能 (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

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 3 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，

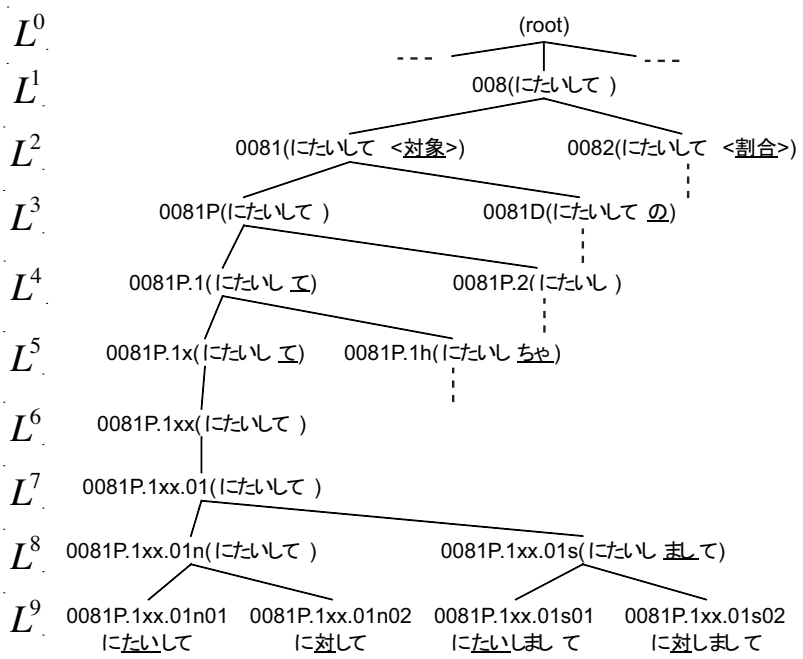


図 3: 日本語機能表現一覧「つつじ」: 形態に基づく階層構造の一部

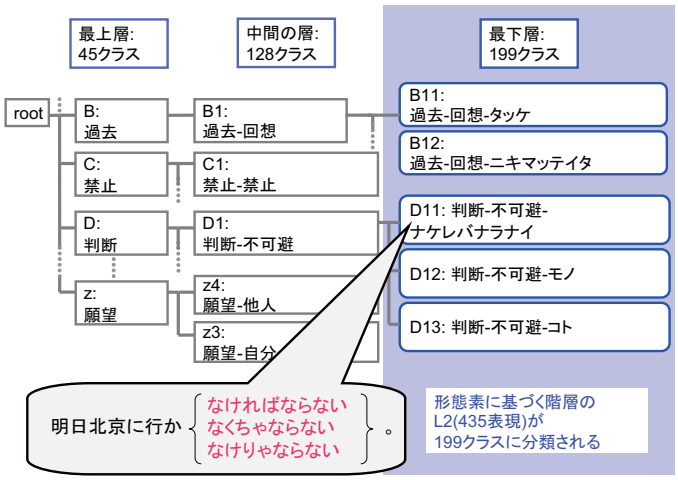


図 4: 日本語機能表現一覧「つつじ」: 意味的等価クラスの一部

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 4 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．表 2 にそれぞれの文体における表現例を示す．

4 機能表現の検出

2 節で述べた機能的用法・内容的用法を自動識別 (機能表現を自動検出) する際に用いる基本的な知識源は，人手によってあらかじめ機能的用法・内容的用法の区別を付与した用例集合である．特

表 2: 日本語機能表現一覧「つつじ」: 文体の種類

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

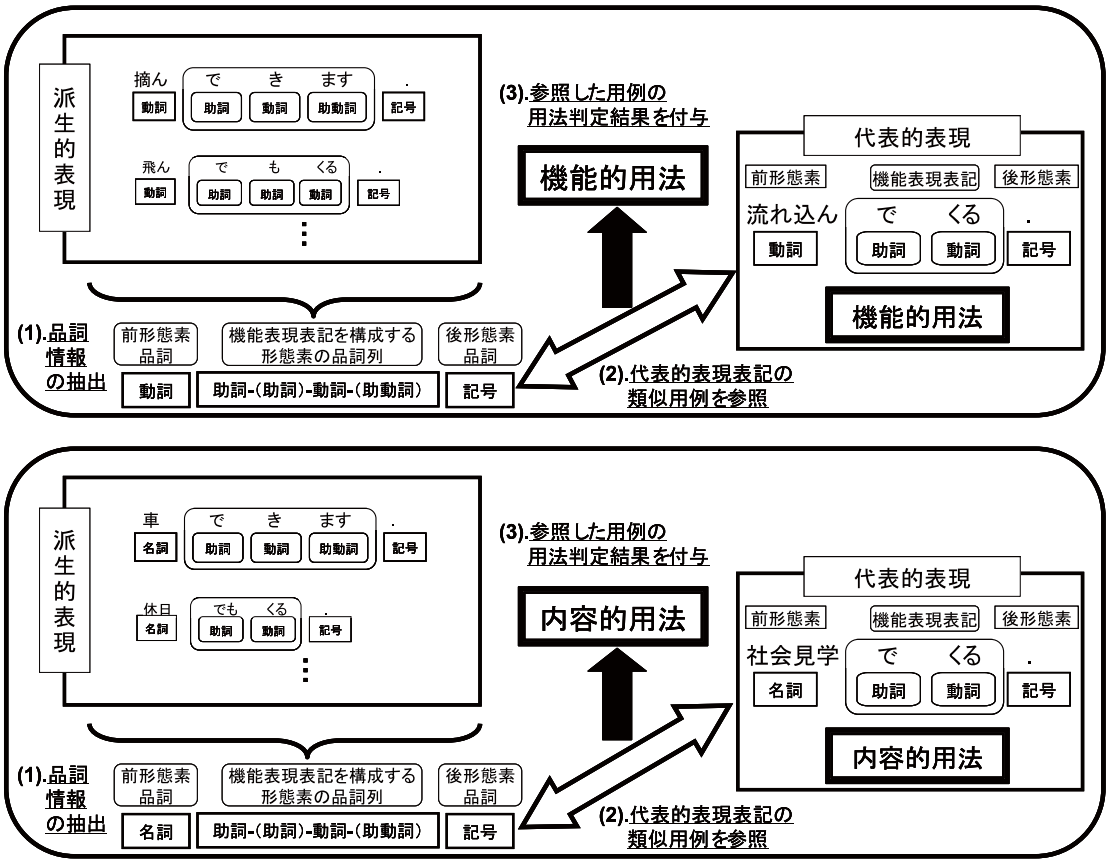


図 5: 模式図: 「代表的表現の表記の用例」を参照して「派生的表現の表記の用例」の用法を判定

に，それらの用例において，機能表現となり得る表記を構成する形態素の情報 (品詞・活用形等)，および，その前後に位置する数形態素ずつの情報 (品詞・活用形等) が重要な手がかりとなる．

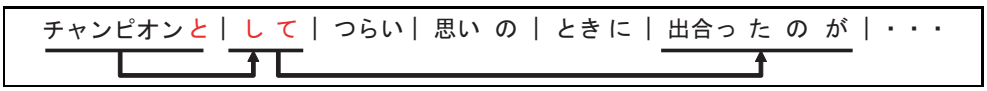
我々は，それらの手がかりを最適な形で統合して機能的用法・内容的用法の自動識別を行う方式のプロトタイプとして，人手によってあらかじめ機能的用法・内容的用法の区別を付与した用例集合を訓練事例として，機械学習により機能表現の検出・係り受け解析を行う方式を提案した [土屋 07, 注連 07]．この方式では，機能的用法・内容的用法が適度に分布しており，相対的に用法の自動識別が容易でない 59 表記を対象として F 値で 93%の性能を達成した．

ここで，[土屋 07, 注連 07] の機械学習による機能表現検出においては，一つの表記あたり 50 例程度の訓練事例を収集して，人手で機能的用法・内容的用法の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, 鈴木 10] 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) の用例を参照して，用法判定を行う方式を提案した．この方式では，階層の上位に位置する代表的表現は，表 1

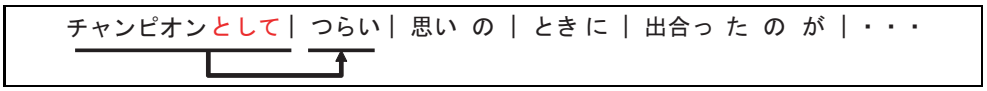
(a) 「機能表現を考慮しない係り受け解析」による失敗例

(b) 「機能表現を考慮した係り受け解析」による成功例

図 6: 機能表現を含む文節の係り元同定の改善例 (助詞型-連用辞類)



(a) 「機能表現を考慮しない係り受け解析」による失敗例



(b) 「機能表現を考慮した係り受け解析」による成功例

図 7: 機能表現を含む文節の係り先同定の改善例 (助動詞型)

および図 3 の L^4 階層相当の 1,000 表現程度の規模とした．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した．ただし，代表的表現を選定する際には，以下の制約を課した．

- 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．

この方式にしたがって，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を図 5 上半分に，内容的用法であると判定した例を図 5 下半分に，それぞれ示す．上半分の例においては，派生的表現の表記の前後の形態素の品詞は，いずれも,「動詞」および「記号」となる．また，派生的表現の表記を構成する形態素列の品詞情報は，派生のタイプによって差異があるが，それらを包含する品詞列パターンは「助詞-(助詞)-動詞-(助動詞)」と表現することができる．そして，これらの情報と十分に類似する前後文脈，および，機能表現表記を構成する形態素列を持つ代表的表現の表記の用例を検索し，その用法判定結果を参照することにより，これらの派生的表現の表記の用法は,「状態が継続している」意味をもつ機能的用法であると判定できる．同様に，下半分の例においては，派生的表現の表記の用法は，派生的表現の表記を構成する動詞の内容的用法であると判定できる．

5　機能表現を考慮した係り受け解析

日本語文の係り受け解析においては，機能表現となり得る表記が機能的用法なのか，それとも，その表記を構成する内容語本来としての内容的用法なのか，を正しく識別することにより，後段の係り受け解析等の文解析の性能を改善できる場合がある．具体的には，機能表現となり得る表記が機能的用法であると識別できる場合には，それらの形態素列を一文節中の機能語列とみなして，一文の係り受け解析を行うことにより，解析性能が改善する場合がある．逆に，それらの表記が，その表記を構成する内容語本来としての内容的用法の場合には，その主辞となる内容語を中心として一つの独

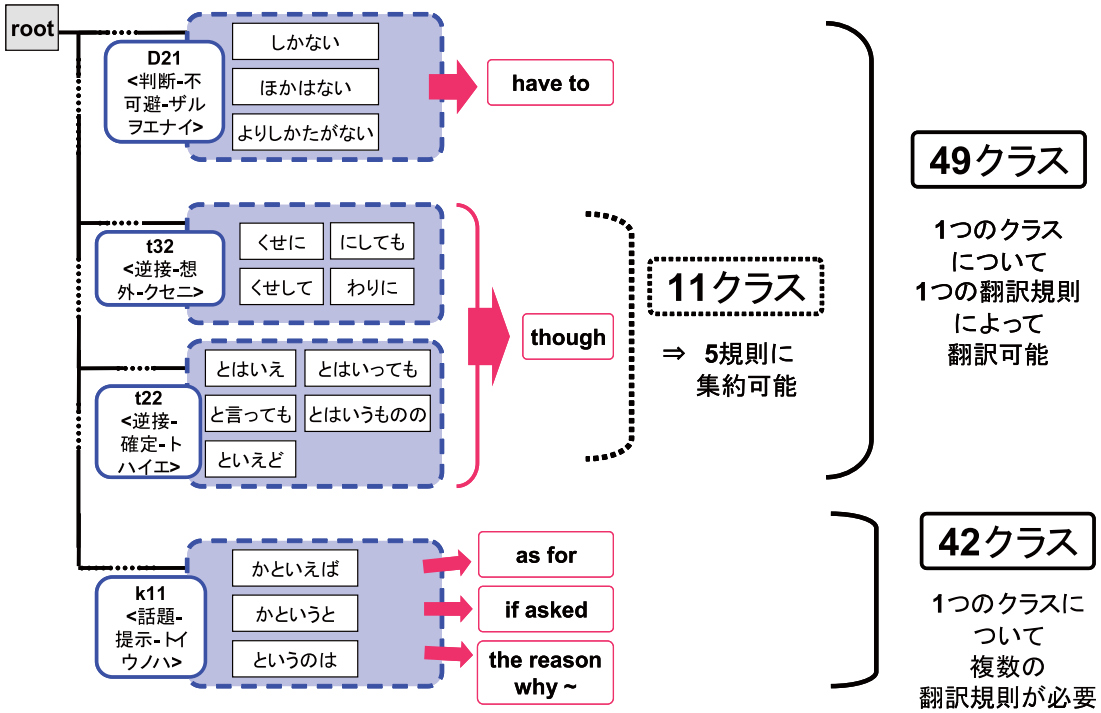


図 8: 集約的英訳可能性に基づく意味的等価クラスの粒度の再編

立した文節を構成し，一文の係り受け解析を行うのが適切である．以上の考えに基づいて，[注連 07] においては，形態素解析の後段で機能表現の検出を行った後，係り受け解析を行うことにより，係り受け解析の性能が改善できることを示した．

この方式によって，機能表現を含む文節の係り元の推定が改善された例を図 6 に示す．機能表現「として」を含む図 6 の文の場合，機能表現「として」の構成要素である形態素「と」および「して」を非構成的な複合辞として扱わず，二つの形態素の列として扱うと，図 6 (a) に示すように,「チェチェン進行を」という文節が，誤って最も近くの動詞文節「して」に係ってしまう．それに対して,「として」は非構成的な複合辞であるとして，一つの形態素として扱った場合，図 6 (b) に示すように,「チェチェン進行を」の係り先を正しく推定することができる．

一方，機能表現を含む文節の係り先の推定が改善された例を図 7 に示す．機能表現「として」を含む図 7 の文の場合，機能表現「として」の構成要素である形態素「と」および「して」を非構成的な複合辞として扱わず，二つの形態素の列として扱うと，図 7 (a) に示すように，構成形態素の一つである動詞「する」の連用形「し」が，最も近くの動詞文節との間で並立構造を構成すると誤判定される．それに対して,「として」は非構成的な複合辞であるとして，一つの形態素として扱った場合，図 7 (b) に示すように，動詞文節の並列構造は構成されず，機能表現「として」を含む文節の係り先を正しく判定できる．

6 日本語機能表現の集約的翻訳

3 節で示したように，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，[坂本 09, 島内 10, Nagasaka10, 劉 10] 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，3 節で述べた「機能表現一覧」の意味的等価クラス (199 クラス) を用いた．例えば，[坂本 09] では，日本語学習者向けの機能表現用例集，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 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．例えば,「判断-不可避-ザルヲエナイ」という意味ラベルのクラスは “*have to*” という代表的訳語に英訳可能であり，このクラスには,「しかない」,「ほかはない」,「よりしかたがない」等の表現が属していた．

7　おわりに

以上，本稿では，日本語文の自然言語処理において，機能表現となり得る表記に対して，機能的用法か内容的用法かを同定する解析を中心として，これまでに筆者らが行った研究を紹介した．今後，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必要がある．また，機械翻訳等の応用の観点からは，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．一方，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．

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).

[森田 89] 森田良行, 松木正恵：日本語表現文型, NAFL 選書, 第 5 巻, アルク (1989).

[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).

[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, pp. 654–657 (2009).

[島内 10] 島内蘭, 長坂泰治, 坂本明子, 宇津呂武仁, 松吉俊：日英特許翻訳における日本語機能表現の集約的英訳可能性の調査, 言語処理学会第 16 回年次大会論文集, pp. 611–614 (2010).

[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).

[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 07] 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).