日本語機能表現の集約的英訳*

坂本 明子† 宇津呂 武仁† 松吉 俊‡

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‡

1 はじめに

機能表現とは，文 (1) の「について」や文 (2) の「にちがいない」，文 (3) の「とはいえ」ように複数の語が一つの助詞・助動詞・接続詞のように振舞う表現を指す [土屋 06]．機能表現は，その語を構成する複数の構成要素を合わせた意味ではなく，表現全体で 1 つの意味を持つのが特徴である．

機能表現は日英機械翻訳ソフトで正しく翻訳できないことがある．本稿では，始めに現在市販されている翻訳ソフトが抱える，日本語機能表現の異型と多義性にまつわる問題を提示する．次に，異型の問題を解消するための方向性を先行研究である機能表現を網羅的に扱う類語辞書と似た意味を持つ話し言葉を代表形にまとめてから翻訳するアプローチを引用しながら解説する．続いて，日本語機能表現を網羅的に機械翻訳するための提案手法として，大規模日本語機能表現階層辞書を用いた日本語機能表現の集約的な日英機械翻訳手法を提案する．最後に，提案手法の実現可能性に関する調査結果を述べ，今後の展望について述べる．

格助詞型	(1)	農村の生活**について**調べている．
助動詞型	(2)	これは天狗の仕業**にちがいない**．
接続詞型	(3)	手紙を出した**とはいえ**，返事が来るとは限らない．

2 機械翻訳による日本語機能表現の英訳

2.1 日本語機能表現

以下に，機能表現の国語学分野と自然言語処理分野における機能表現研究の経緯を説明する．

国語学分野の [森田 89, 国研 01] が日本語機能表現の網羅的な体系を作成したのを受けて，自然言語処理分野においても機能表現が研究されるようになった経緯がある．[松吉 07] の辞書とは独立して，[土屋 06] では [国研 01] で列挙された 125 個の見出し語だけでなく，その活用形を含めた 337 表現に対して，最大 50 文ずつの用例を文字列照合を用いて収集し，機能的な用法と内容的[1]な用法の人手判定ラベルを付与した．その後，機能表現を入力文中から検出する手法 [土屋 07] が提案され，更に，[注連 07] では，機能表現検出のための係り受け解析を提案した．また，[松吉 07] が，日本語機能表現を各表現の構成要素の組み合わせとして階層的に網羅した辞書を作成した．この辞書は [土屋 06] の用例データベースを受けて，辞書に収録する機能表現の範囲を拡張することを目指したものである．また，後に [松吉 08] は，辞書内で言い換え可能な表現ごとに機能表現を分類し，言い換え可能な機能表現群ごとに意味等価クラスラベルを付与した．さらに，[長坂 09] では，[松吉 07] の機能表現一覧に対応した機能表現検出の枠組みを提案している．

以上の機能表現研究の先行研究により日本語機能表現を網羅的に取り扱うことが容易になったことを踏まえ，本研究では日本語機能表現を網羅的に機械翻訳することを試みる次第である．

2.2 機械翻訳による日本語機能表現の英訳

ここでは，機能表現が持つ 2 つの特徴と，それらにまつわる機械翻訳上の問題について解説する．

機能表現は，ほぼ同じ意味を持つ見出し語表記の異型が複数存在することが特徴である．現行の翻訳ソフトでは，異型が辞書に登録されてないことが原因で出力エラーを起こすことがある．

例えば，文 (4) と文 (6) では機能表現が正しく解析されて英語訳を得られているのにたいし，文 (8) では，機能表現の異型が翻訳ソフトの辞書に登録されていないために解析できず，出力エラーの原因となっている．

*Machine Translation of Japanese Functional Expressions into English through Canonical Expressions

†Akiko Sakamoto, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba,

‡Suguru Matsuyoshi, Graduate School of Information, Nara Institute of Science and Technology,

[1] 機能語が助詞・助動詞・接続詞を指すのに対し，内容語は名詞・動詞・形容詞・副詞を指す．

入力 1　(4)　教師はどの生徒に対しても公正で**なくてはならない**.

MT 出力 1　(5)　A teacher must be fair to every student.

入力 2　(6)　教師はどの生徒に対しても公正で**なくちゃならない**.

MT 出力 2　(7)　A teacher must be fair to every student.

入力 3　(8)　教師はどの生徒に対しても公正で**なけりゃならない**.

MT 出力 3　(9)　A teacher is fair to every student, and textbfif he is なけりゃ, there is. （ no ）

機能表現のもう一つの特徴は，複数の意味を持つ表記が存在することである．現行の翻訳ソフトでは，複数ある意味のうち 1 つにしか対応できてない場合がある．

文 (10) の「によって」は行為者を表していると捉えることができるので，機械翻訳出力文 (11) の”by”は日本語の意味を正しく反映した翻訳といえる．一方，文 (12) の「からして」は,「明日の天気」という判断材料として考慮すべき状況を指しているので，少なくとも行為者を表す by では翻訳できないはずだが，機械翻訳では「によって」に対して”by”という 1 種類の翻訳規則しか用意していないために，同じ「からして」という文字列が異なる意味を持つ場合の文脈に応じた訳し分けができていない．

今回の調査では，機能表現の多義性の問題は取り扱い対象から除外し，今後の課題とした．

入力 1　(10)　これらの聖典はヨーロッパの宣教師たち**によって**もたらされた.

MT 出力 1　(11)　These sacred books were brought **by** propagators from the Europe.

入力 2　(12)　行くか行かないかはあしたの天気**によって**決めよう.

MT 出力 2　(13)　Will decide whether do not go whether go **by** the tomorrow's weather.

3　階層的機能表現辞書を用いた日本語機能表現の集約的英訳

本稿では，機能表現の異型を網羅的に翻訳するためのアプローチを提案する．本手法の先行研究は，日本語の機能表現を網羅的に扱う辞書と，日本語の話し言葉の表記の揺れを集約的に英訳するアプローチである．

3.1　階層的日本語機能表現辞書

3.1.1　形態素に基づく階層構造

[松吉 07] は，日本語の機能表現の異型を，機能表現の構成要素の組み合わせとして階層的に収録している．これにより，日本語機能表現の網羅的取り扱いが可能になった．

この辞書の階層の上位には，[土屋 06] において作成された 337 表現を配置し，機能表現末尾の活用だけでなく，機能表現の各構成要素の音韻的変化や，とりたて詞[2] の挿入，口語的な表現と敬語表現の差し替えなどによる異型を機械的に展開した後に，実際に日本語として使用できるものだけを人手で残した 16771 表現が収録されている．

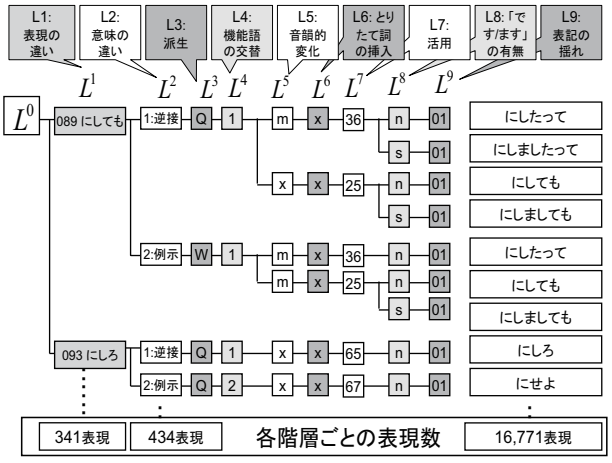

図 1: 形態素に基づく階層構造

3.1.2　意味等価クラスに基づく階層構造

また，[松吉 08] は，上記の辞書に収録された見出し語間の類似度に応じて，3 段階のクラス分けを行った．

この最下層に位置する全 199 個の各意味等価クラスに属する機能表現群は，日本語文中で言い換え可能であることが確認されている．

この研究で階層辞書に意味等価クラスが付与されたことにより，日本語機能表現の言い換え候補を網羅的に取り扱うことが可能になった．

3.2　代表的表現への言い換えによる Sand-Glass 機械翻訳

[山本 01, Yamamoto04] では，文 (14)〜文 (19) のように換言することで日本語話し言葉の表記揺れを代表形に

[2] とりたて詞の一例に,「でも」,「しか」,「さえ」がある．[グループ・ジャマシイ 98] によれば,「朝はコーヒー**しか**飲まない.」のように,「ひとつの事だけを取り上げて，他を排除する」際に用いられる．

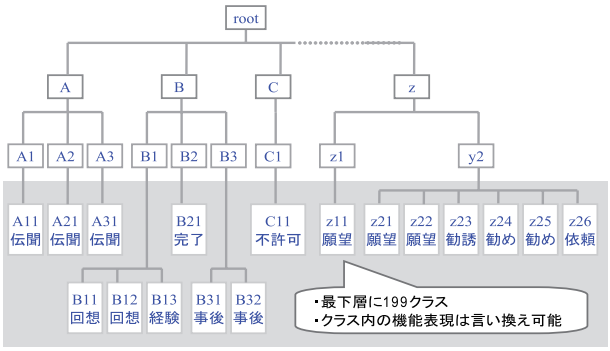

図 2: 意味等価クラス

集約し，少ない翻訳規則で様々な入力文を翻訳することに成功した．この研究で扱われているのは，内容語と機能表現のうちの口語的なものだけであるが，本稿では，口語的な機能表現だけでなく，全ての文体の機能表現を網羅的に取り扱う．

換言前	(14) 御安心ください
	↓
換言後	(15) ご安心ください
換言前	(16) ではいかがいたしましょうか
	↓
換言後	(17) ではどうしましょうか
換言前	(18) 風邪じゃないかと思うんですけど
	↓
換言後	(19) 風邪でしょう

3.3 意味等価クラスを用いた日本語機能表現の集約的英訳

本研究では，先行研究である日本語機能表現一覧の意味等価クラスの粒度を，日英翻訳用に再調整し，調整後のクラスごとに翻訳規則を定めることにより，日本語機能表現を網羅的に集約的英訳する手法を提案する．

集約的に英訳する様子を図 3 に例示する．図 3 では，始めに入力文中の機能表現を代表形に言い換えた後，代表形に対する翻訳規則を用いて英語に翻訳し，さまざまな機能表現の異型に対して入力解析エラーを起こすことなく翻訳することができる．

4 集約的英訳可能性の調査

既存の辞書の意味等価クラスの粒度を日英翻訳用のクラスとして再調整する際には，図 4 に示した 3 つの場合が予測される．

既存の意味等価クラスの粒度が日英翻訳用には粗すぎる場合には，意味等価クラスを下位分類し，各下位集合

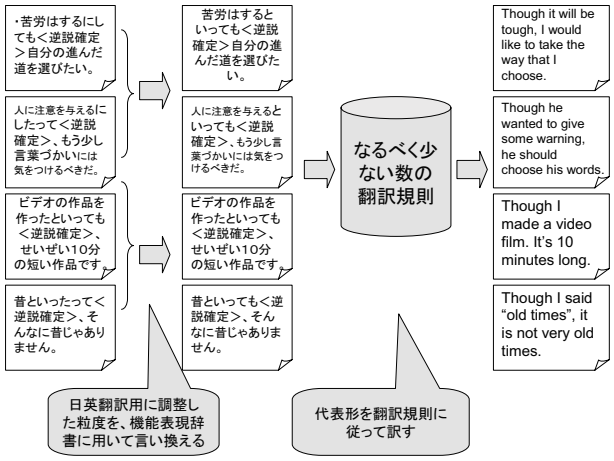

図 3: 代表的表現への集約を経由する機械翻訳

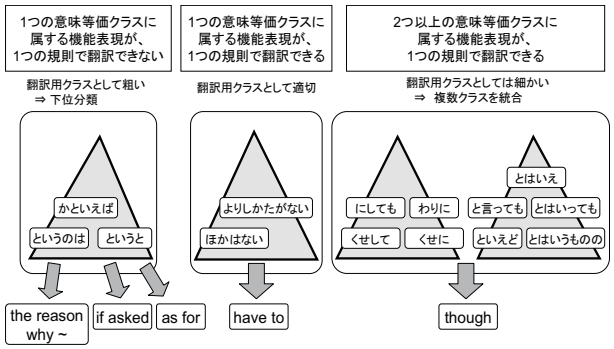

図 4: 意味等価クラスの粒度の再編

に対して翻訳規則を設定する必要がある．もし既存の意味等価クラスの粒度が日英翻訳用としても適切である場合には，1 クラスに収録された機能表現を用いた例文は，全て同じ翻訳規則で翻訳できる．さらに，1 クラス 1 規則で翻訳できるクラスの間で，共通の翻訳規則を使えるクラスがあれば，それは既存の意味等価クラスが日英翻訳用としては細かすぎたということなので，同じ規則が使えるクラスを統合する．

4.1 調査用例文の収集

4.1.1 日本語文型辞典

機能表現の用例文を集めるためのコーパスには，日本語文型辞典 [グループ・ジャマシイ 98] の電子テキスト版を用いる．この辞典は日本語学習者向けに機能表現の用例を約 8000 文収録している．

4.1.2 例文の収集方法

まず，コーパスに収録された例文のうち，日本語機能表現一覧の表記と文字列照合したものを，例文候補として取得する．ここでまず，195 クラスについて文字列照合した例文を 1 文以上得ることができた．

次に，日本語文型辞典 [グループ・ジャマシイ 98] の例文の収録欄の見出し語が，文字列照合した機能表現表記とほぼ一致する例文数が 5 文以上存在するクラスの数を集計したところ，76 クラスとなった (表 1 中,「1 クラス中の照合例文 5 文以上 (ケース 1)」)．日本語文型辞典 [グループ・ジャマシイ 98] の例文の収録欄の見出し語が文字列照合する機能表現表記と一致する場合には，例文中の機能表現の用法が機能表現一覧の表記が持つ用法と一致する場合が多く，英訳対象の例文として適切な場合が多い．

さらに，単なる文字列照合などのノイズが起こりにくいと予測される機能表現を含んだクラスを選び，すべての文字列照合結果に対して，例文中の機能表現の用法が機能表現一覧の表記が持つ用法と一致する例文数を集計したところ，新たに 16 クラスについて，1 クラスあたり 5 文以上の例文を収集することができた (表 1 中,「1 クラス中の照合例文 5 文以上 (ケース 2)」)．

以上の結果より，本論文において，集約的英訳可能性の調査が可能なクラス数を，合計 92 クラスとした．

表 1: 得られた例文の集計

意味等価クラスの全数	199
文字列照合した例文が存在したクラス	195
1 クラス中の照合例文 5 文以上 (ケース 1)	76
1 クラス中の照合例文 5 文以上 (ケース 2)	16
集約的英訳可能性の調査が可能なクラス	92

4.2 集約的英訳可能性の調査および結果

調査用例文が得られた 92 クラスについて，1 クラスから 5 文ずつ例文を抽出し，1 クラス 1 規則で翻訳できるか調査した．

その結果，下位分類が必要なクラスは 42 クラス，1 クラス 1 規則で翻訳可能なクラスは 50 クラスあり，50 クラス中の 11 クラスを計 5 規則に集約できることが分かった．

表 2: 集約的英訳可能性の調査結果

1 クラス 1 規則では翻訳できないクラス	42
1 クラス 1 規則で翻訳できるクラス	50
1 クラス 1 規則で翻訳できるクラスのうち，他のクラスと翻訳規則を共有するクラス	11

5 まとめと今後の課題

本稿では，既存の大規模日本語機能表現階層辞書の意味等価クラスの粒度を日英機械翻訳向けに再調整することにより，日本語機能表現を集約的英訳する手法提案した．

また，この手法の実現可能性について調査するために，全 199 個の意味等価クラスのうち 92 クラスについて調査用例文を取得した．調査の結果，42 クラスは日英翻訳用に下位分類する必要があり，50 クラスは 1 クラスにつき 1 つの翻訳規則で翻訳でき，日本語言い換え用のクラスを日英翻訳用にも使えることが明らかになった．更に，50 クラス中の 11 クラスは，5 規則に集約して翻訳できることも判明した．

今後は，今回用いたコーパスのほかにも，新たなコーパスを導入するなどして調査用例文を増やし，今回未調査のクラスについて調査を行う．また，新たな例文収集結果も踏まえて，下位分類した意味等価クラスへ翻訳クラスを付与したり，下位分類する必要の無いクラスを上位統合するための調査を行う．さらには，[松吉 04] で指摘されているような，同一の見出し語を複数の意味に翻訳するための代表形への言い換えを参考にしながら，多義な見出し語の曖昧性解消の方法も構築する．

謝辞: 日本語文型辞典 [グループ・ジャマシイ 98] の電子テキスト版の使用を許可して頂いた，筑波大学大学院人文科学研究科 砂川有里子教授に感謝する．

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).

[松吉 04] 松吉俊, 佐藤理史, 宇津呂武仁：機能表現「なら」の機械翻訳のための言い換え, 情報処理学会研究報告, Vol. 2004, No. (2004–NL–159), pp. 201–208 (2004).

[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).

[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 09] 長坂泰治, 宇津呂武仁, 松吉俊, 土屋雅稔：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, 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 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).

[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).

[Yamamoto04] Yamamoto, K.: Interaction between Paraphraser and Transfer for Spoken Language Translation, *Journal of Natural Language Processing*, Vol. 11, No. 5, pp. 63–86 (2004).