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*

島内 蘭† 長坂 泰治† 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，以下の例文の「について」,「にちがいない」,「とはいえ」ように複数の語が一つの助詞・助動詞・接続詞のようにふるまう表現を指す [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．

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

日本語機能表現には，非常に多様な異形が多く存在するが，現状の日英機械翻訳ソフトにおいて，それらの異形をすべて網羅的に正しく翻訳することは容易ではない [坂本 09]．本稿では，原言語における類似の表現を，代表的な表現に言い換えた後，機械翻訳の言語変換部を適用するという SandGlass 翻訳方式 [山本 01] を採用する．そして，[坂本 09] では，日本語機能表現を網羅的に列挙した大規模日本語機能表現階層辞書 [松吉 07, 松吉 08] を利用して，日本語機能表現の日英翻訳を対象として，この SandGlass 翻訳方式を適用することにより，日本語機能表現の集約的な日英機械翻訳手法を提案している．

[坂本 09] では 1 意味的等価クラス内の日本語機能表現を 1 規則で翻訳できる可能性がある 49 意味的等価クラスを示している．そこで，本研究ではまず最初に対象とする意味的等価クラスのサンプルとして，この 49 クラスを取り上げる．なお，翻訳規則作成のためには，目的言語側の訳が不可欠である．この際，目的言語側の訳が利用できない場合には翻訳作業を行う必要があり，[坂本 09] では日常会話文に対して目的言語側の訳を作成したうえで，翻訳規則の作成を行っている．一方，本研究では，NTCIR-7 の特許翻訳タスク [Fujii08] で配布された 1,798,571 件の日英対訳特許文対を用いて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．特許文の場合は使われる機能表現の意味範囲が狭く，その種類も少ないので，翻訳規則の作成が容易になると考えられる．

2 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．[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，機能的な用法と自立的な用法の人手判定ラベルを付与した．また，[松吉 07, 松吉 08] は，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 (日本語機能表現一覧「つつじ」[1])．また，後に [松吉 07, 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的等価クラスラベルを付与した．

3 階層的日本語機能表現辞書

3.1 形態に基づく階層構造

[松吉 07, 松吉 08] 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，図 1 に示すように，日本語機能表現の網羅的取り扱いが可能になった．

この辞書には，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

*Machine Translation of Japanese Functional Expressions into English through Canonical Expressions in JE Patent MT

†Ran Shimanouchi, Taiji agasaka, Akiko Sakamoto, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba,

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

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

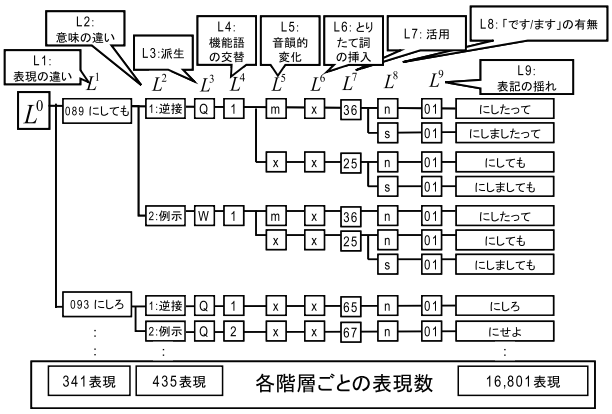

図 1: 形態に基づく階層構造

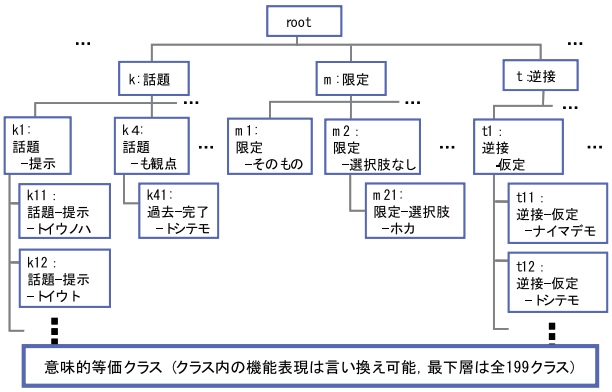

図 2: 意味的等価クラス

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 16,801 表現が収録されている．

3.2 意味的等価クラスに基づく階層構造

また，[松吉 07, 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，図 2 に示す 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 個の各意味的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とされている．[松吉 07, 松吉 08] において，機能表現の階層辞書に対して，意味的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となった．

4 意味的等価クラスを用いた日本語機能表現の集約的英訳

[坂本 09] では，先行研究である日本語機能表現一覧の意味的等価クラスの粒度を，日英翻訳用に再編し，再編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し英訳する手法を提案している．集約するという考え方は，似た意味を持つ文を代表形に言い換えてから翻訳するという [山本 01] に基づいている．

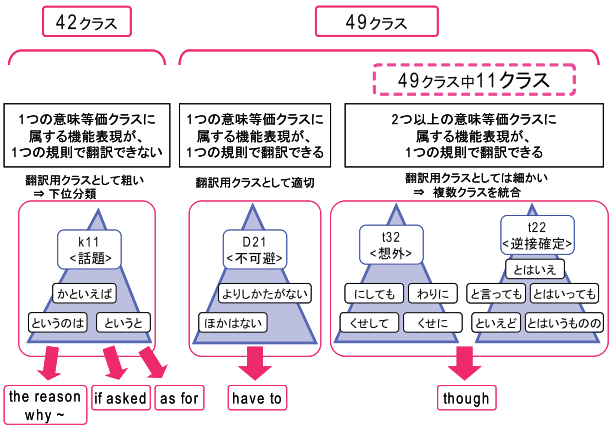

図 3: 日英翻訳の観点からの意味的等価クラスの粒度の再編

既存の辞書の意味的等価クラスの粒度を日英翻訳用のクラスとして再編する際には，図 3 に示した 3 つの場合が予測される．既存の意味的等価クラスの粒度が日英翻訳用には粗すぎる場合には，意味的等価クラスを下位分類し，各下位集合に対して翻訳規則を設定する必要がある．一方，既存の意味的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存在するならば，それは既存の意味的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．

[坂本 09] においては，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ている．この辞典は日本語学習者向けに機能表現の用例を約 8,000 文収録している．このコーパスにおいては，199 個の意味的等価クラスのうち，91 クラスについて，1 クラス 5 文以上の例文を収集することができた．これらの 91 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否かの調査を行った．その結果，図 3 に示すように，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 49 クラスあり，49 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．

以上の [坂本 09] の成果をふまえて，そこで本稿では，まず最初に対象とする意味的等価クラスのサンプルとして，この 49 意味的等価クラスをとりあげる．

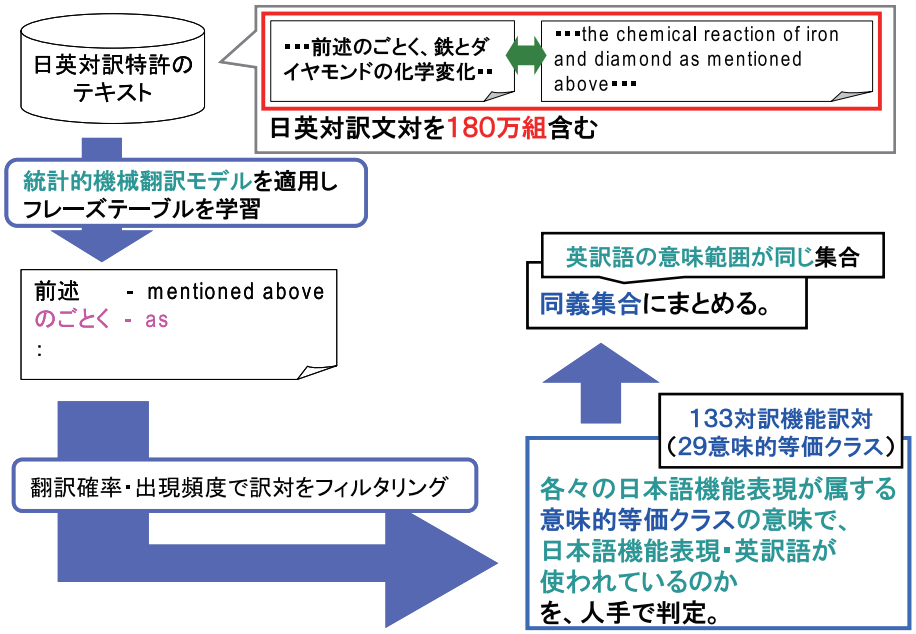

図 4: 日英対訳特許文からの日本語機能表現の集約的英訳規則獲得手順

表 1: フレーズテーブルに含まれる日本語機能表現のエントリの例

日本語機能表現	英訳語
のごとく	as
ほか	in addition to
以外	except
だけでなく	not only

5 対訳特許テキストを利用した集約英訳規則の獲得

5.1 統計的機械翻訳モデルのフレーズテーブル

日英対訳特許文対に対して，句に基づく統計的機械翻訳モデル [Koehn03] のツールキットである Moses を適用することにより，句の日英対応及びその確率を記載したフレーズテーブルを作成する．フレーズテーブルにおける日本語機能表現のエントリの例を表 1 に示す．

5.2 集約的英訳規則の獲得

日本語機能表現の集約的英訳規則の獲得手順を図 4 に示す．

まず，前節で作成したフレーズテーブルから，大規模的階層機能表現辞書「つつじ」 [松吉 07, 松吉 08] に収録されている機能表現のエントリを抽出する．ただし，対訳特許文対における日本語機能表現の出現頻度の下限を 20，対訳特許文対における日本語機能表現および英訳語が句対応していると判定された頻度の下限を 10，フレーズテーブルにおける日英翻訳確率 $P(f_e \mid f_j)$ の下限を 0.05 とする．次に，これらの日本語機能表現および英訳語のエントリのうち，4 節で述べた 49 個の意味的等価クラスに含まれる日本語機能表現の表記，および，語義に該当する組のみを抽出した．その結果，29 個の意味的等価クラスに含まれる日本語機能表現を含む，133 組の日英機能表現対訳対が抽出された．最後に，133 組の日英機能表現対訳対を，29 個の各意味的等価クラスごとに分割し，各クラスにおいて，英訳語の意味・用法が同義となる「同義集合」へのまとめ上げをおこなった．

以上の手順の結果，29 個の意味的等価クラスのうち，26 クラスについては,「同義集合」の数は一つとなったが，残りの 3 クラスについては，一クラス中の「同義集合」数がそれぞれ二つに分割された．これらの 3 クラスおよび 6 個の「同義集合」を図 5 に示す．また，これらの 133 組の日英機能表現対訳対に含まれる日本語機能表現の種類数は 72 表現であった．以上の集計結果を表 3 に示す．また，意味的等価クラス「m21(語義は「限定-選択肢」)」に属する三組の日英対訳機能表現対の例文を表 2 に示す．

このように，本研究の方式により，72 種類の日本語機能表現の英訳規則を 32 個に集約することができた．

6 おわりに

本研究では，日英対訳特許テキストと，既存の大規模日本語機能表現階層辞書の意味的等価クラスを用いること

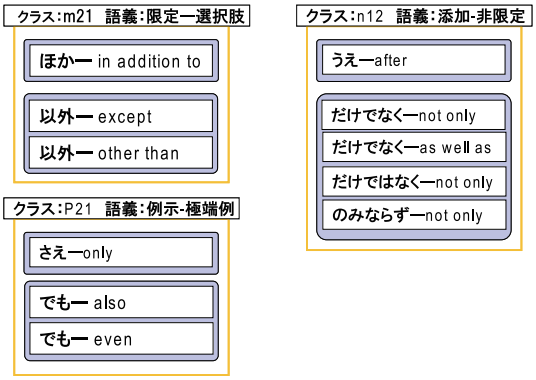

図 5: 同義集合が 2 個に分割された意味的等価クラス

表 2: 意味的等価クラス「m21(語義は「限定-選択肢」)」中の機能表現の対訳例文

日英機能表現対訳対	例文	意味
ほか, in addition to	沈殿室には、汚水の供給用の配管22、 処理水の流出部24のほか、 中央上部に円筒形のセンターウェル26が備えられている。 the settling chamber 18 accommodates a cylindrical center well 26 in its upper central portion in addition to a sewage supply pipe 22 and a treated water effluent portion 24	添加
以外, except	この粉体インク媒体を使用し、粉体インク担持体の バイアス電圧を70Vにし、25mAのパルスを印加した 以外は、実施例1と同様な方法で粉体インクを付着させ、 印字を行った。 printing was done in the same manner as embodiment 1 except using this ink transfer medium making the bias voltage of the toner holding means 70 V and applying a pulse of 25 mA to it	除外
以外, other than	この方法では、目標層以外の層の溝からの回折光は、 焦点がずれているために、・・・ in this method since diffracted lights from grooves other than the target layer are out of focus・・・	

により，日本語機能表現を集約的に英訳する翻訳規則を獲得する手法を提案した．全 199 個の意味的等価クラスのうち 49 クラスを対象として本手法を適用したところ，日英対訳特許テキストからは，29 クラスに対する翻訳規則が獲得された．また，29 クラスのうち 26 クラスについては，1 クラスあたり一つの英訳規則に集約可能であった．今後は，今回未調査のクラスについて本手法を適用するとともに，獲得された集約的英訳規則の評価を行う．また，他ジャンルの文書に対して本手法を適用し，集約的英訳規則の獲得を行う．さらに，今後，日中対訳特許文書といった他の言語対の対訳特許文書が利用可能となった場合には，[劉 10] 等，日英以外の言語対を対象とする日本語機能表現翻訳規則作成の研究成果をふまえて，対象となる目標言語の範囲を拡大することも可能である．

表 3: 集約的英訳規則数および対訳機能表現数

	意味的等価クラス中の「同義集合」の数		
	1	2	合計
意味的等価クラス数	26	3	29
「同義集合」数	26	6	32
日本語機能表現の数	65	7	72
日英対訳機能表現数	122	11	133

参考文献

[Fujii08] Fujii, A., Utiyama, M., Yamamoto, M. and Utsuro, T.: Overview of the Patent Translation Task at the NTCIR-7 Workshop, *Proc. 7th NTCIR Workshop Meeting*, pp. 389–400 (2008).

[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).

[Koehn03] Koehn, P., Och, F. J. and Marcu, D.: Statistical Phrase-Based Translation, *Proc. HLT-NAACL*, pp. 127–133 (2003).

[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 (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).

[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).

[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).

[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).

[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).