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*

劉 颯† 長坂 泰治‡ 宇津呂 武仁‡ 松吉 俊§

筑波大学 第三学群工学システム学類† 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 (例えば，日英機械翻訳ソフトに関しては，[坂本 09b])．この問題に対して，本研究では，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．我々は，これまでに，日英翻訳に関しては，[坂本 09b,Sakamoto09a] において，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英訳手法を提案した．[坂本 09b,Sakamoto09a] 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．この日英翻訳に関する研究成果に基づき，本稿では，日中翻訳を対象として，日本語機能表現の集約的中国語訳規則を作成する．

ここで，翻訳規則作成のためには，翻訳規則作成の対象とする原言語表現に対して，目的言語側の訳の情報が不可欠である．この際，あらかじめ目的言語側の訳の情報が付与された言語資源が利用できない場合には，翻訳規則作成過程において翻訳作業を行う必要がある．この問題に対して，[坂本 09b,Sakamoto09a] では，日本語文型辞典 [グループ・ジャマシイ 98] に収録された日本語学習者向け日常会話文を対象として，日本語機能表現の英訳情報を作成したうえで，日英翻訳規則の作成を行った．一方，本稿では，日本語文型辞典 [グループ・ジャマシイ 98] の中国語訳 [グループ・ジャマシイ 01] 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を作成する．

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．

*Utilizing Semantic Equivalence Classes of Japanese Functional Expressions in Japanese to Chinese Machine Translation

†Sa Liu, College of Engineering Systems, Third Cluster of Colleges, University of Tsukuba

‡Taiji Nagasaka, 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/

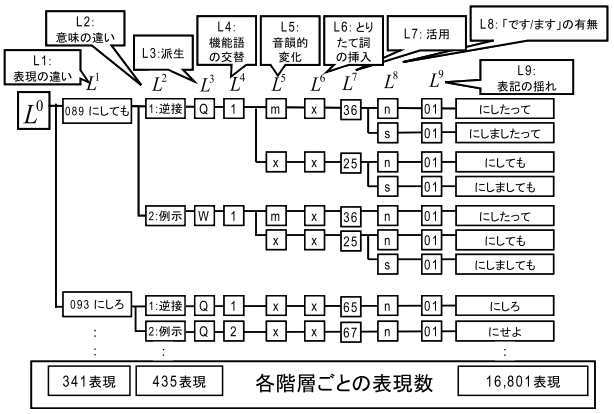

図 1: 形態に基づく階層構造

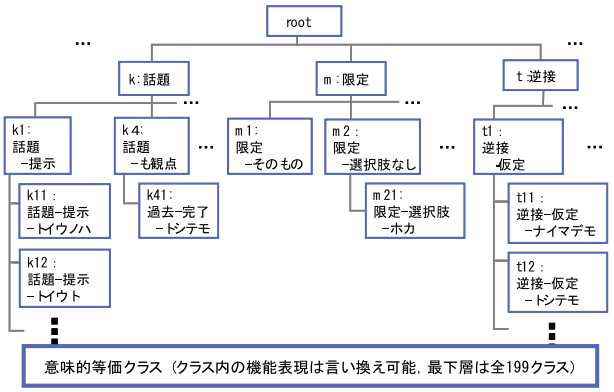

図 2: 意味的等価クラス

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．

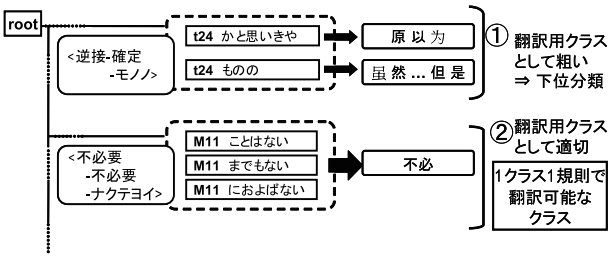

図 3: 日中翻訳の観点からの意味的等価クラスの粒度の再編

4 意味的等価クラスを用いた日本語機能表現の集約的中国語訳規則の作成

[坂本 09b, Sakamoto09a] における集約的日英翻訳規則作成の場合と同様に，本稿における集約的日中翻訳規則の作成においても，日本語機能表現一覧の意味的等価クラスの粒度を，日中翻訳用に再編し，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中国語訳する．

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 2 つの場合が予測される．日中翻訳の観点において，既存の意味的等価クラスの粒度が粗すぎる場合には，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，日中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．

本稿では，日本語文型辞典 [グループ・ジャマシイ 98] の中国語訳 [グループ・ジャマシイ 01] を日中対訳コーパスとして利用し，この意味的等価クラス再編の作業を行った．この辞典には，日本語学習者向けに，日本語機能表現の用例を約 8,000 文が収録されている．まず，199 個の意味的等価クラスのうち，クラス中の日本語機能表現の例文が十分な数，上記日中対訳コーパスから収集可能な 50 クラスを対象として，上述の意味的等価クラス再編の作業を行った．その結果，1 クラスあたりの翻訳規則数が 1 個となるクラスが 20 クラス，1 クラスあたりの翻訳規則数が複数個となるクラスが 30 クラスとなった．

次に，次節において，これらの翻訳規則の評価を行うにあたって，既存の日中機械翻訳ソフトとの比較を行うために，機能的用法の多義性，および，機能的用法・内容的用法の間の曖昧性が観測されにくい表現の例文を評価文として選定する [長坂 10, Sakamoto09a]．ここで，機

表 1: 1 クラスあたりの翻訳規則数が 1 個となる 13 クラス

意味的等価クラス	機能表現例
A21(伝聞-ぼかし-トヤラ)	とやら，とか
c21(仲介-経由-ヲツウジテ)	をつうじ，をつうじて，をとおし，をとおして，を通し，を通した，を通して，を通しての，を通しました，を通じ
G32(意志-計画-マデ)	までのこと，まで
H21(自然発生-抑制不能-テシカタガナイ)	てしかたがない，てしかたない，てしょうがない，てたまらない，でしかたがない，てならない，でしょうがない，てしようがない，でいけない
H31(自然発生-強制-ナイデハオカナイ)	ずにはいられない，ずにはおかない，ないではいられない，ないではおかない
I21(推量-高確実性-ニチガイナイ)	にきまっている，にきまってる，にちがいない，にはちがいない
I23(推量-高確実性-ノダロウ)	ことだろう，だろう，のだろう，んだろう
m21(限定-選択肢なし-イガイ)	以外
o21(同時性-対比-タトオモウト)	たかと思うと，たかと思えば，たかと思ったら，たと思うと，たと思ったら，だかと思うとたと思えば，だかと思えば，だと思うと，だと思えば
t11(逆接-仮定-ナイマデモ)	ないまでも，ぬまでも
t22(逆接-確定-トハイエ)	といえど，といえども，といって，とはいいながら，とはいうものの，とはいえ，とは言え，と言って
x11(疑問-問いかけ-カ)	かしら
y61(否定-不可能-ワケニハイカナイ)	わけにいかない，わけにいくまい，わけにはいかない，わけにはいくまい，わけにもいかない，わけにもいくまい，わけにいかない

能的用法の多義性については，日本語機能表現一覧「つつじ」において複数の語義が登録されているか否かによって判定した．また，機能的用法・内容的用法の間の曖昧性については，[長坂 10] において，新聞記事 1 年分において 50 回以上出現する機能表現表記については，その用法を人手で判定することにより，機能的用法・内容的用法の間の曖昧性の有無を判定した．また，その他の機能表現表記については，内省に基づいて，機能的用法・内容的用法の間の曖昧性の有無を判定した．以上の結果，評価文が利用可能な意味的等価クラスの数は，1 クラスあたりの翻訳規則数が 1 個となるクラスについて 13 クラス，1 クラスあたりの翻訳規則数が複数個となるクラスについて 22 クラスとなった．このうち，本稿の範囲では，次節における評価において，1 クラスあたりの翻訳規則数が 1 個となるクラスとしては，13 クラスのうち 8 クラスのみを評価対象とした．1 クラスあたりの翻訳規則数が複数個となるクラスについては，22 クラス全体を評価対象とした．

このうち，表 1 には，1 クラスあたりの翻訳規則数が 1 個となる 13 クラスについて，各クラスの機能表現例を示す．また，図 4 には，1 クラスあたりの翻訳規則数が多数得られる意味的等価クラス J33(進行-継続-テクル) における，機能表現「てくる」の複数の中国語翻訳規則を示す．

5 評価

前節で述べたように，1 クラスあたりの翻訳規則数が 1 個となる 8 クラス，および，1 クラスあたりの翻訳規則

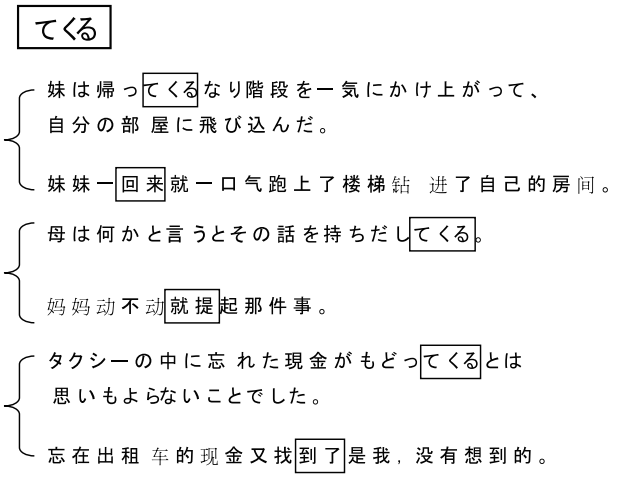

図 4: 意味的等価クラス J33(進行-継続-テクル) における複数の中国語訳の例

数が複数個となる 22 クラスについて，提案手法による翻訳規則の評価を行った．いずれのクラスに対しても，まず，日中対訳コーパス (日本語文型辞典 [グループ・ジャマシイ 98] の中国語訳 [グループ・ジャマシイ 01]) の用例集合を，翻訳規則作成用の訓練文および評価文に分離した．その際，前節で述べたように，評価文に含まれる日本語機能表現としては，機能的用法の多義性，および，機能的用法・内容的用法の間の曖昧性が観測されにくい表現ができるだけ多種類含まれるように，機能表現の選定を行った．

次に，1 クラスあたりの翻訳規則数が 1 個となる 8 クラスについては，各クラスの翻訳規則をそのまま用いた．

表 2: 日本語機能表現の中国語翻訳規則の評価結果

	クラス数	翻訳規則数	訓練文数	評価文数	翻訳精度 (%)	
					提案手法	MT ソフト
1 クラス 1 翻訳規則	8	8	125	77	**92.2**	67.5
1 クラス複数翻訳規則	22	79	1176	173	**92.4**	65.3

一方，1 クラスあたりの翻訳規則数が複数個となる 22 クラスについては，提案手法の翻訳規則の性能の上限を見積もるために，各評価文について，中国語の知識のない筆者の一人が最も適切と思われる翻訳規則を手動で選択するという評価法を用いた．その際，各評価文の機能表現が属する意味的等価クラス中の複数の翻訳規則のうちの一つを選択することとし，翻訳規則の日本語側用例文のみを参照して最も適切と思われる翻訳規則を手動で選択した．

また，比較対象の既存の日中機械翻訳ソフトとして，エキサイト日中翻訳[2] を用い，翻訳結果の中国語文のうち，日本語側の機能表現に相当する表現部分のみを評価対象として，翻訳精度の評価を行った．

以上の評価における翻訳規則数，訓練文数，評価文数，および翻訳精度を表 2 に示す．この結果から分かるように，1 クラスあたりの翻訳規則数が 1 個となる 8 クラス，および，1 クラスあたりの翻訳規則数が複数個となる 22 クラスともに，提案手法の翻訳精度が日中機械翻訳ソフトの翻訳精度を上回った．このうち，1 クラスあたりの翻訳規則数が 1 個となる 8 クラスについては，自動で翻訳規則の適用が可能なため，提案手法の翻訳規則を既存の機械翻訳手法に容易に組み込むことが可能であり，有用な評価結果であると言える．一方，1 クラスあたりの翻訳規則数が複数個となる 22 クラスについて，特に翻訳誤りとなった評価文の多くは，中国語において日本語側では不要な訳し分けを必要とする事例であった．また，提案手法によって正解となった事例についても，今後，意味的等価クラス内で複数翻訳規則間の自動選択手法を確立する必要がある．

6 おわりに

本稿では，日本語機能表現の集約的中国語訳規則を作成しその翻訳精度を評価した．評価実験の結果，既存の日中機械翻訳ソフトの翻訳精度を上回る精度を達成した．本研究の成果を実用化するためには，今後，機能的用法の多義性，および，機能的用法・内容的用法の間の曖昧性を解消するモジュールを作成し，それらの多義性解消モジュールと連動して日中翻訳規則を適用する必要がある．

また，日英の言語対においては，[坂本 09b] の日本語機能表現集約的英訳規則の成果をふまえる形で，[島内 10] において，日英特許翻訳に特化した日本語機能表現の集約的英訳規則の作成について，一定の成果が得られている．そこで，今後，日中対訳特許文書が利用可能となった場合には，日中特許翻訳に特化して，日本語機能表現の集約的中国語訳規則の作成において，本稿の研究成果を応用することが可能であると考えられる．

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).

[グループ・ジャマシイ 01] グループ・ジャマシイ（編）: 徐一平 (訳): 教師と学習者のための日本語文型辞典中国語訳簡体字版, くろしお出版 (2001).

[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).

[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 10] 長坂泰治, 宇津呂武仁, 松吉俊, 土屋雅稔：階層的機能表現辞書に基づく日本語機能表現の分析と検出, 言語処理学会第 16 回年次大会論文集 (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 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 (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).

[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).

[2] http://www.excite.co.jp/world/chinese/