日本語機能表現の集約的英訳における意味的等価クラスの利用

坂本明子 (日本語機能表現班協力者：筑波大学大学院システム情報工学研究科)*
宇津呂武仁 (日本語機能表現班班長：筑波大学大学院システム情報工学研究科)
長坂泰治 (日本語機能表現班協力者：筑波大学大学院システム情報工学研究科)
松吉俊 (日本語機能表現班協力者：奈良先端科学技術大学院大学情報科学研究科)

Utilizing Semantic Equivalence Classes in Machine Translation of Japanese Functional Expressions

Akiko Sakamoto (University of Tsukuba)
Takehito Utsuro (University of Tsukuba)
Taiji Nagasaka (University of Tsukuba)
Suguru Matsuyoshi (Nara Institute of Science and Technology)

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 02] を採用する．そして，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] 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する．さらに，この日本語機能表現の日英翻訳手法を適用するための条件として，用法・意味の曖昧性のない機能表現を対象とする必要があることを述べ，さらに，用法・意味の曖昧性のない機能表現を同定した結果を示す．最後に，提案手法の評価結果を述べ，今後の展望について述べる．

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，機能的な用法と内容的[1] な用法の人手判定ラベルを付与した．ま

*sakamoto @ nlp.iit.tsukuba.ac.jp
[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．

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 (日本語機能表現一覧「つつじ」[2])．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的等価クラスラベルを付与した．

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．

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

ここでは，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[3]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 16801 表現が収録されている．

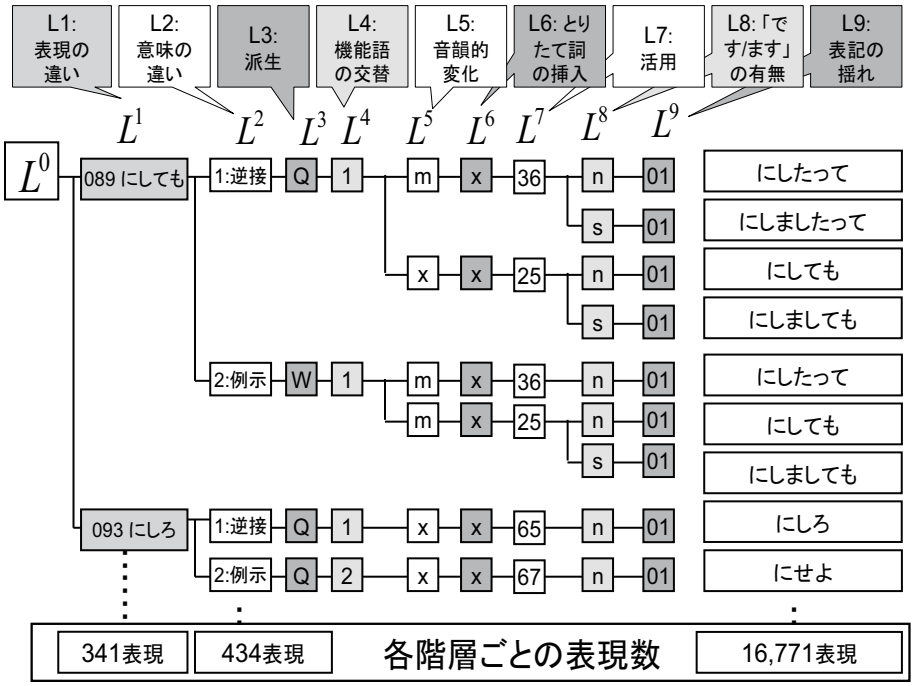

図 1: 形態素に基づく階層構造

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

[3] 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．

3.1.2 意味的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ことが確認されている．

この研究で階層辞書に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になった．

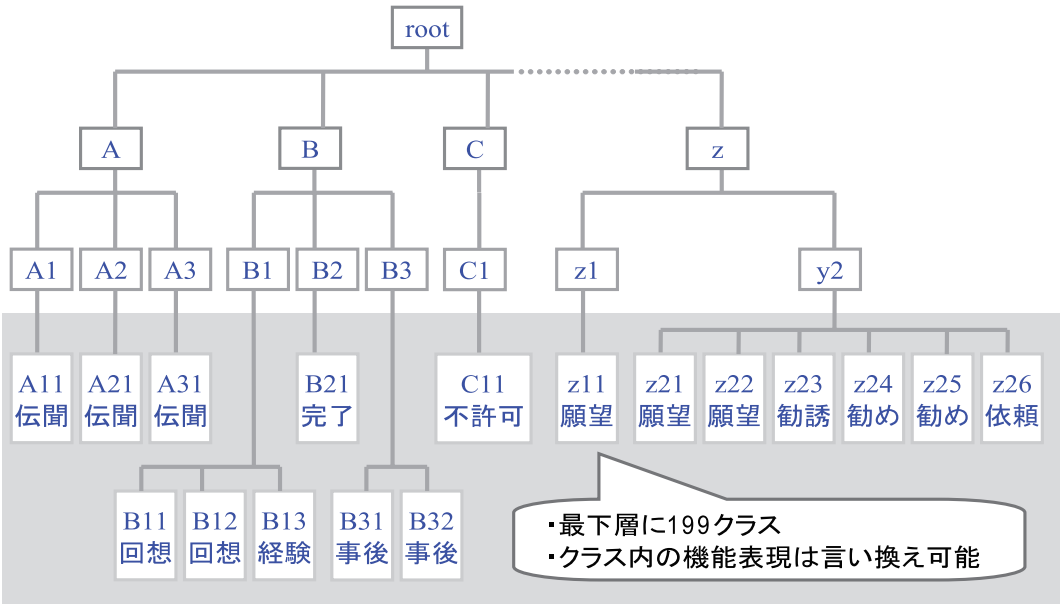

図 2: 意味的等価クラス

3.2 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

本研究では，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 02] を参考にしたものである．

4. 集約的英訳規則の作成

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調整する際には，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．もし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あれば，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．

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．

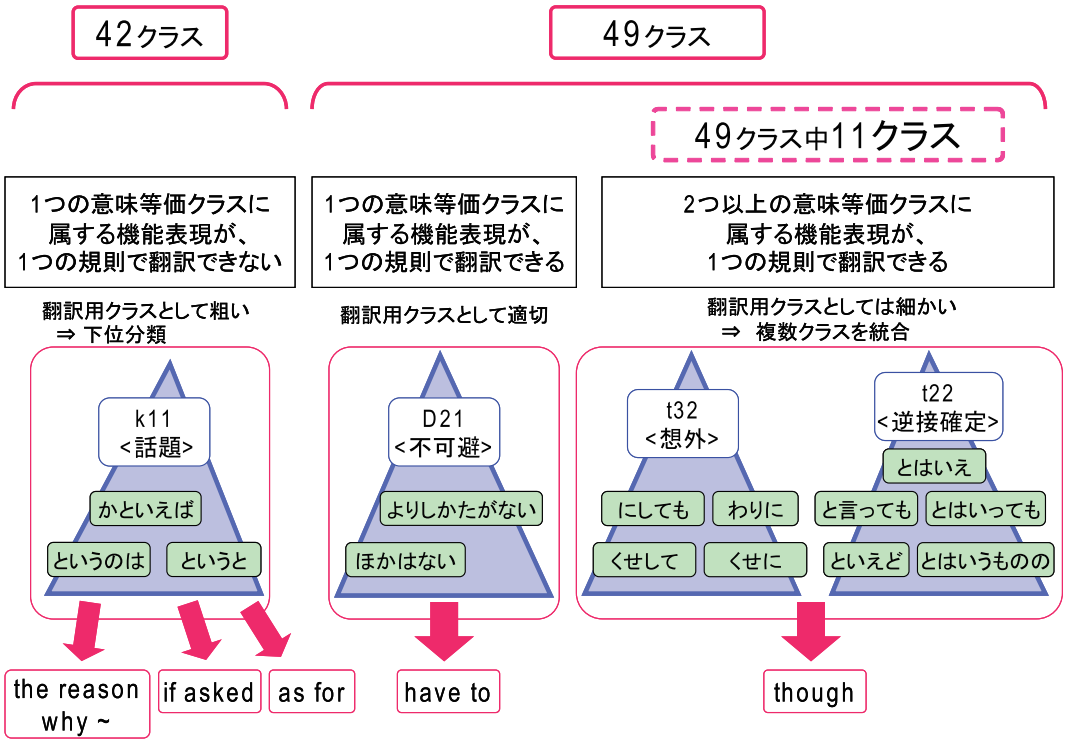

図 3: 既存の意味的等価クラスの粒度の再編

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．

5. 用法・意味の曖昧性のない機能表現の同定

1 つの機能表現表記が 1 つの用法・意味しか持たない場合，つまり用法・意味の曖昧性のない場合には，見出し語が特定できれば用法・意味も確定するので，機械的に翻訳規則を適用すればよい．一方，1 つの機能表現表記が複数の用法・意味を持つ場合，すなわち用法・意味の曖昧性を持つ場合には，何らかの方法で機能表現の用法・意味を特定してから翻訳規則を適用する必要がある．したがって，前節で作成した翻訳規則の適用する前段階として，個々の機能表現表記の用法・意味の曖昧性の有無を判定し，用法・意味の曖昧性がある場合には，用法・意味の曖昧性を解消しなければならない．

そこで，本節では，前節で求めた，1 クラス 1 規則で翻訳可能な 49 クラス中の機能表現に対して，用法・意味の曖昧性の有無を判定し，用法・意味の曖昧性のない表現と，用法・意味の曖昧性のある表現の数を推定する．

機能表現表記の用法・意味の曖昧性のうち，機能的用法/自立的用法の曖昧性，および，機能的用法の多義性がない機能表現表記の例を表 1 (a) に示す．一方，機能表現表記に対して，機能的用法/自立的用法の曖昧性 [土屋 07, 長坂 09] がある場合の例を表 1 (b) に示す．また，機能表現表記に対して，機能的用法の多義性がある場合の例を表 1 (c) に示す．

このうち，機能的用法の多義性の有無の判定においては，日本語機能表現一覧「つつじ」において同一の表記が複数のエントリを持つか否かによって判定を行った．また，機能的用法/自立的用法の曖昧性の有無の判定においては，毎日新聞 1995 年版 (約 130 万文，427Mbytes) およびブログコーパ

表 1: 機能表現における用法・意味の曖昧性の有無の例

(a) 機能的用法における多義性なし・機能的用法/自立的用法の曖昧性なし

	機能表現	例文	用法・意味
(1)	ことができる	彼は英語を話す[ことができる]。	機能的用法, 意味分類 = 「可能」

(b) 機能的用法における多義性なし・機能的用法/自立的用法の曖昧性あり

	機能表現	例文	用法・意味
(2)	とはいえ	状況は改善している[とはいえ]、まだ安心できない。	機能的用法, 意味分類 = 「逆接確定」
(3)	とはいえ	状況が改善した[とはいえ]ない。	自立的用法

(c) 機能的用法における多義性あり

	機能表現	例文	用法・意味
(4)	ために	世界平和の[ために]国際会議が開かれる。	機能的用法, 意味分類 = 「目的」
(5)	ために	雨の[ために]彼の到着が遅れた。	機能的用法, 意味分類 = 「理由」

表 2: 1 クラス 1 規則で翻訳可能な 49 クラスにおける機能表現の用法・意味の曖昧性の有無 (L^2 / L^9 エントリ数)

機能的用法における多義性なし			機能的用法における 多義性あり
機能的用法 /自立的用法の 曖昧性なし	機能的用法 /自立的用法の 曖昧性あり	新聞記事 1 年分・ ブログコーパス中の 出現頻度 20 未満	
42 / 2752	22 / 749	33 / 2188	
97 / 5689			69 / 690
166 / 6379			

ス (260Mbytes) 中から，L^2 エントリ 166 表現について，各 20 文を選定して，機能的用法/自立的用法の曖昧性の有無を判定した．さらに，L^9 エントリ 6379 表現については，それぞれ上位の L^2 エントリの機能的用法/自立的用法の曖昧性の有無を継承すると仮定して，機能的用法/自立的用法の曖昧性の有無の判定を行った．以上の結果を集計したものを表 2 に示す．この結果より，前節の翻訳規則が無条件に適用できる機能表現表記の割合を推定できた．

6. 評価

本節では，1 クラス 1 規則で翻訳可能な 49 クラスについて，4 節で作成した翻訳規則を評価する．

評価対象となるのは，作成した英文のうちの機能表現部分のみである．また，評価においては，翻訳の精度を以下の 3 段階で評価した.「正解: 英文の機能表現は日本語の機能表現の意味を保持する」,

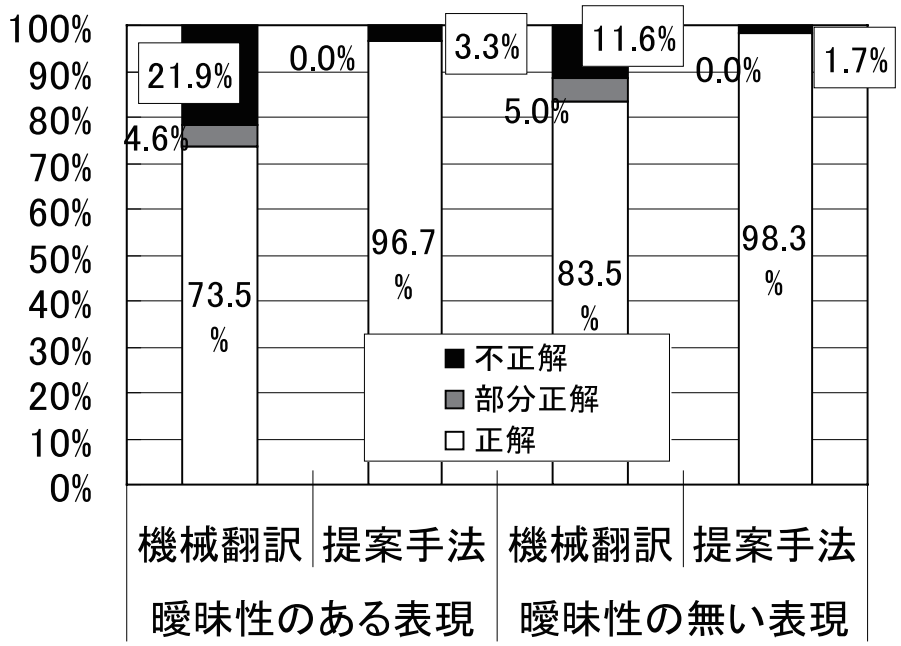

図 4: 機能表現翻訳規則の評価結果

「部分正解: 英文の機能表現は日本語文の機能表現の意味を多少保持する.」,「不正解: 英文の機能表現は日本語の機能表現の意味を保持しない」.

評価対象となるのは，規則作成に用いていない計 272 文である．これらの文を，用法・意味の曖昧性がある機能表現が出現する 151 文，および，用法・意味の曖昧性のない機能表現が出現する 121 文に分けて評価を行った．表 3 に，意味的等価クラス「b11 (対象)」，および,「D11(当為)」について，用法・意味の曖昧性のない機能表現を対象として行った評価の例を示す.

評価の集計は図 4 の通りである．機能表現表記に対して，用法・意味の曖昧性がない場合，既存の機械翻訳ソフトの「正解」の割合は 83.5%であるのに対し，提案手法では 98.3%に向上する．このことから，提案手法により，既存の翻訳ソフトに比べて機能表現部分の翻訳精度が向上することがわかる.

7. 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的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を提案した.

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 個の意味的等価クラスのうち 91 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，49 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

表 3: 翻訳規則の例

b11 (対象)	規則作成	(1) その事件**に関して**学校から報告があった。 (2) 農村の生活様式**について**調べている。
	規則評価	(3) これ**に関して**はわかりません。 ○: I don't have any idea **about** this. (4) 入力ラベルが得られる毎に、各単語**に関する**スコアが更新される。 ×: For every label input, the score **about** each word is updated.
D11(当為)	規則作成	(5) 学生は勉強する**べき**だ。 (6) 履歴書は自筆のもので**なくてはいけない**。
	規則評価	(7) 他人の私生活に干渉する**べき**ではない。 ○: You **should** not intrude into someone's life. (8) 一致協力して問題解決に当たら**ねばならない**。 ○: We **should** cooperate to solve the problem.

意味的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な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．

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, pp. 328–331 (2009).

[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).

[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 02] 山本和英：換言と言語変換の協調による機械翻訳モデル, 言語処理学会第 8 回年次大会発表論文集, pp. 307–310 (2002).