体系的機能表現辞書に基づく日本語機能表現の言い換え

松吉　俊 †, ‡　　佐藤　理史 ‡

† 京都大学大学院 情報学研究科, ‡ 名古屋大学大学院 工学研究科

1. はじめに

内容的・機能的という観点から、日本語の表現は、大きく 2 つに分類できる。さらに、「表現を構成する語の数」という観点を加えると、表 1 のように分類できる。ここで、**複合辞**とは、「にたいして」や「なければならない」のように、複数の語から構成されているが、全体として 1 つの機能語のように働く表現のことである。われわれは、機能的というカテゴリーに属する機能語と複合辞を合わせて**機能表現**と呼んでいる。日本語には、多くの機能表現が存在し、日本語を表現豊かなものにしている。われわれは、この機能表現を研究対象とする。

機能表現は、意味という観点から分類できる。すなわち、同じような意味を持っている機能表現を集めて、類義表現の集合を作ることができる。例えば、「とすぐに」、「たとたん」、「たかとおもったら」、「と同時に」、「や否や」などは、形態的にはかなり異なっているが、共通して「前件と後件がほぼ同時に起こった」という意味を表しているので、1 つの集合にまとめることができる。

同じ集合に属する 2 つの機能表現は、同等な意味を持っているので、ある文脈においては置換可能であることが期待される。したがって、類義表現の集合を作成し、それらを利用すれば、機能表現の言い換えが実現できると考えられる。

このような背景により、われわれは、類義表現の集合を生成することができる機能表現辞書を作成した [1]。本研究では、この機能表現辞書に基づいて機能表現を言い換えるシステムを提案する。

2. 機能表現の言い換え

乾ら [2] は、語彙・構文的言い換えを次の 6 つに分類した。

(1)　節間の言い換え
(2)　節内の言い換え
(3)　内容語の複合表現の言い換え
(4)　機能語/モダリティの言い換え
(5)　内容語句の言い換え
(6)　慣用表現の言い換え

本研究は、機能表現を言い換え対象とするので、われわれの言い換えは、乾らの分類の (4) に相当する。乾らは、(4) に関して、「語彙的な性格が強く、局所的な情報を参照するだけで言い換えられるものも多い」と述べている。

表 1　日本語表現の分類

	一語	二語以上
内容的	内容語 (名詞、動詞、形容詞など)	複合名詞、複合動詞、慣用表現
機能的	**機能語**(助詞、助動詞、接続詞、形式名詞)	**複合辞**

したがって、大部分の機能表現は、局所的に類義表現と置換することにより、その言い換えが実現できると考えられる。

一方、機能表現には、異形を持つものが多い。例えば、「なければならない」、「なければなりません」、「なくてはならない」、「なけりゃならない」、「なくちゃならない」、「なければならん」、「ねばならん」などは、互いに異形である。それゆえ、実際に言い換えシステムを構築するときには、機能表現の異形について考慮する必要がある。

2.1 先行研究

飯田ら [3] は、機能表現の解説文や例文から、279 個の言い換え規則を人手で作成している。土屋ら [4] は、機能表現を含む文とその機能表現を言い換えた文の対のデータを作成し、そこから 642 個の言い換え規則を半自動的に生成している。これらの研究で作成された言い換え規則は、ある機能表現と別の機能表現が言い換え可能であることを示す個別的なものである。このような個別的な規則の集合を用いる手法では、数多く存在する機能表現の異形を言い換えるために、膨大な量の言い換え規則を作成しなければならない。

Tanabe ら [5]、Shudo ら [6] は、助動詞型機能表現の列を言い換えるために、論理的類似規則と語用論的類似規則を導入している。彼らの研究は、格助詞型、接続助詞型などの機能表現を扱っておらず、また、異形についての言及はない。

2.2 本研究の提案手法

本研究では、異形を考慮して機能表現を言い換えるために、個別的な言い換え規則に従うのではなく、体系的な機能表現辞書が提供する類義表現の集合を利用する。

本論文では、言語単位を次のように定義する:

文節は、内容語部と機能語部からなる
内容語部は、1 つ以上の内容語からなる
機能語部は、0 以上の機能表現からなる

われわれが提案する言い換えシステムは、入力として 1 つの文節を受け、出力として、それと同等な意味を持つ文節 (代替文節) の順位付きリストを生成する。このと

表 2 9 つの階層

階層	区分観点	ノード数
L^1	構成語	341
L^2	**意味**	**434**
L^3	機能	551
L^4	機能語の交替	769
L^5	音韻的な変化	1,182
L^6	とりたて詞の挿入	1,805
L^7	活用形	6,857
L^8	「です」「ます」の有無	9,705
L^9	**表記**	**16,771**

き、機能表現辞書から提供される類義表現の集合に基づいて、文節の機能語部に存在する機能表現をその類義表現に置換する。

本研究の方針は、次の 2 つである。

- できるだけ多くの代替文節を生成する
- より自然な表現が上位にくるようにする

上位 5 位以内に、望ましい代替文節が得られることを目指す。

以下、機能表現辞書について述べた後、すでにわれわれが実装した、1 つの機能表現からなる機能語部を言い換えるシステムについて説明する。

3. 機能表現辞書

われわれは、自然言語処理での使用を想定した日本語機能表現辞書を作成した [1)]。この辞書は、次の 2 つの特徴を持っている。

(1) 抽象度の異なる複数の機能表現リストを含む
(2) 機能表現に関するさまざまな情報を外部システムに提供することができる

3.1 機能表現リスト

われわれが作成した機能表現辞書は、階層構造に基づいて、機能表現およびその異形を分類する。この階層構造は 9 つの階層を持ち、各々の階層において、表 2 の「区分観点」に示す観点により、機能表現を分類する。

それぞれの階層における機能表現ノードの集合は、同じ抽象度を持つ機能表現のリストとして利用することができる。例えば、見出し語のリストがほしいときには、L^1 の機能表現ノードの集合を利用すればよい。また、意味が異なるものを区別した見出し語のリストがほしいときには、L^2 の機能表現ノードの集合を利用すればよい。同様に、異形も含めた、機能表現のすべての表記のリストがほしいときには、L^9 の機能表現ノードの集合を利用すればよい。

この辞書は、次の 2 つのリストの機能表現とその異形をすべて含む。

(1) 「日本語表現文型」[7)] の助詞と同様の働きをする表現 (並立助詞の働きをするものは除く) と助動詞と同様の働きをする表現、計 412 項目
(2) 「使い方の分かる類語例解辞典 新装版」[8)] の助詞、助動詞およびその連接形、計 368 項目

各階層のノード数を表 2 の「ノード数」の欄に示す。この表は、機能表現辞書が、16,771 表現からなる表記のリストを含むことを示している。

3.2 機能表現に関する情報

この辞書に記述されている情報のうち、言い換えに関連する情報は、左接続、意味カテゴリー、難易度、文体の 4 種類である。

3.2.1 左 接 続

一般に、ある機能表現の左に接続できる語の集合 (左接続) は、それを構成する語列の先頭の要素に対する左接続より制限が強い。例えば、「てくれませんか」の先頭の要素は、接続助詞「て」であり、これは、左接続として広く用言をとることができるが、「てくれませんか」の左接続は、動詞のみである。

機能表現の左接続を、先頭の要素の左接続と同じであると仮定することは、解析系においては特に問題にならない。なぜならば、機能表現に接続できない語が接続した形で文が入力されることはないということを想定して、解析を行なうからである。

その一方で、生成系においては、上記を仮定することは、大きな問題となる。なぜならば、実際には接続できない語が左接続に記述されていた場合、その語と機能表現が接続したものを、システムが、文法的に正しい表現であるとして出力してしまうおそれがあるからである。

このような誤りを避けるためには、生成時においても文中の機能表現を認識し、辞書に機能表現として適切な左接続を記述する必要がある。

3.2.2 意味カテゴリー

「日本語表現文型」[7)] における意味分類を参考にして、同じような意味を持っている機能表現の集合として、88 の意味カテゴリーを導入した。

辞書においては、L^2 の機能表現ノードに、それが属する意味カテゴリーを記述し、下位のノードにそれを継承させた。

3.2.3 難 易 度

表現の分かりやすさに基づいて、機能表現に 5 段階の難易度 (やさしい方から A1、A2、B、C、F)[9)] を記述した。難易度を記述する際には、「日本語能力試験出題基準」[10)] における「< 機能語 > の類」の級を参考にした。

3.2.4 文 体

表現の文体に基づいて、機能表現に、常体、敬体、口語体、堅い文体の 4 種類のいずれかを記述した。

3.3 類義表現集合の生成

上記の機能表現辞書の特徴より、ある意味カテゴリーに属する機能表現の表記のリストを生成すれば、それを類義表現の集合として利用することができる。

それぞれの意味カテゴリーに属する L^2 の機能表現ノードの数を表 3 に示す。この表から、< 逆接確定 > や < 推量 > などの意味カテゴリーに、多くのノードが属していることが分かる。この事実より、これらの意味カテゴ

表 3 意味カテゴリーに属する L^2 の機能表現ノードの数

L^2	意味カテゴリー (意味カテゴリー数)
15	逆接確定 (1)
14	推量, 強調, 否定, 状況 (4)
13	理由 (1)
12	– (0)
11	同時性, 対象 (2)
10	感嘆, 限定, 自然発生 (3)
9	依頼, 疑問, 並立, 話題 (4)
8	意志, 願望, 逆接仮定, 順接仮定, 想外 (5)
7	当為, 勧め, 継起, 仲介 (4)
6	伝聞, 起点, 順接確定, 添加 (4)
5	判断, 不可能, 不必要, 対比, 立場, 極端例, 主体 (7)
4	許可, 不許可, 不可避, 勧誘, 継続, 事後, 終点, 相関, 付帯, 放置 (10)
3	可能, 順接限定, 非限定, 因状況, 回想, 完了, 基準, 根拠, 目的, 同格, 不満, 比況, も観点, 内-授与, 着継続, 程度, 相手 (17)
2	範囲, 否定意志, 定義, 割合, 相応, 事前, 順接必要, 発継続, 不均衡, 状態, 内-受益, 他-授与, 習慣, 目標, は観点, 不明確, 無意味 (17)
1	否定推量, 回避, 傾向, 経験, 最中, 場合, 適当, 反復, 無視, 名詞化 (10)

リーに属する機能表現に対して、異形の言い換えを越えた、興味深い言い換えを実現できることが期待される。

4. 機能表現言い換えシステム

提案する言い換えシステムの全体像を図 1 に示す。このシステムの入力は、次の 2 つである。

i. 機能語部が 1 つの機能表現からなる文節
ii. 出力を制限する難易度条件・文体条件

システムの出力は、代替文節の順位付きリストである。

言い換えシステムは、(1) 文節解析、(2) 類義表現列挙、(3) 内容語部と機能語部の接続、(4) フィルタリングの 4 つの部分からなる。

4.1 文節解析

まずはじめに、システムは、入力文節を、内容語部と機能語部に分割する。

本研究では、解析を簡略化するために、内容語部は 1 語であるという仮定をおく。解析手順を以下に示す。

(1) 形態素解析器 MeCab☆を用いて、入力文節を形態素解析する
(2) 最初の形態素を内容語部 (の語) とする
(3) 2 つめ以降の形態素の表層形をまとめ、それを機能語部の表記とする
(4) 文字列完全一致で機能表現辞書を引き、その表記に対応する機能表現 ID を得る
(5) 機能表現 ID から、意味カテゴリー、左接続などの情報を得る

機能語部の曖昧性解消を行なっていないため、機能表現 ID が複数得られることがある。この場合は、各々の機能表現 ID に対して、以下の処理を独立に行なう。

4.2 類義表現列挙

次に、システムは、機能表現辞書を参照することにより、機能語部の機能表現の類義表現を列挙する。例えば、意味カテゴリー < 同時性 > に属する「や否や」が入力された場合、「とすぐに」、「たとたん」、「たかとおもったら」、「と同時に」など、同じ意味カテゴリーに属する機能表現が、表記単位 (L^9 単位) で 97 表現出力される。

入力に難易度条件や文体条件が設定されていた場合、その条件を満たす類義表現のみを列挙する。

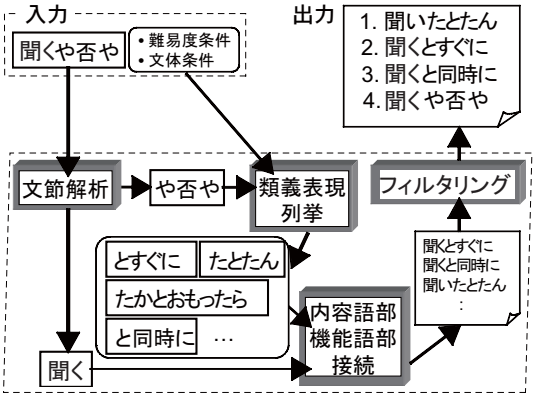

図 1 言い換えシステムの全体像

4.3 内容語部と機能語部の接続

システムは、内容語と、類義表現の集合に属する機能表現を接続し、代替文節を生成する。

一般に、機能表現は、たとえ同じ意味カテゴリーに属する表現であったとしても、それぞれ異なる左接続を持つ。例えば、「や否や」と「たとたん」は、同じ < 同時性 > という意味カテゴリーに属しているが、前者の左接続は動詞の基本形であるのに対し、後者の左接続は動詞の連用タ接続である。また、「にちがいない」と「にほかならない」は、同じ < 推量 > という意味カテゴリーに属しているが、前者の左接続は名詞、動詞、形容詞であるのに対し、後者の左接続は名詞のみである。

したがって、機能表現を類義表現に置換することによって言い換えを行なう場合、内容語部と新しい機能語部を適切に接続する必要がある。

内容語と機能表現の接続には、以下に示す 4 種類がある。

4.3.1 単純接続可能

内容語が機能表現の左接続に含まれる場合、それらを単純に接続する。例えば、「聞く」と「とすぐに」は、単純接続可能である。

4.3.2 活用形の変更が必要

活用形を除いて、内容語が機能表現の左接続に含まれる場合、内容語の活用形を変更することにより、それらを接続する。例えば、「聞く」と「たとたん」を接続するには、活用形の変更が必要であり、まず、活用形変化表を参照して「聞く」を「聞い」に活用させた後、それらを接続する。

4.3.3 語の挿入が必要

内容語と機能表現が、間に語を介せば接続可能な場合、その語を挿入し、全体を接続する。例えば、「作成」と「さ

☆ http://mecab.sourceforge.jp/

表 4 入力「聞くや否や」に対する出力
(i: 条件なしの順位、j: 難易度 B 以下の順位)

i	j	代替文節	出現回数
1	1	**聞いた途端**	49
2	2	**聞いたとたん**	21
3	3	**聞いた途端に**	19
3	-	聞くや	19
5	4	**聞くとすぐ**	15
6	-	聞くなり	14
7	5	**聞いたとたんに**	10
8	6	**聞くとすぐに**	7
8	6	**聞くと同時に**	7
10	-	聞くそばから	2
11	-	聞くが早いか	1
11	-	聞くやいなや	1
11	-	聞くや否や	1

いに」を接続するには、語の挿入が必要である。「作成」の形態素情報と、「さいに」の左接続から挿入語選択表を参照して「する」を得、それを介して全体を接続する。

現在、システムが挿入する語は、次の 5 語である。

「する」 名詞-サ変接続を動詞化する
「なる」 形容詞を動詞化する
「こと」 用言を名詞化する
「の」 名詞を連体化する
「な」 名詞-形容動詞語幹を連体化する

4.3.4 接続不可能

上のいずれにも当てはまらない場合、内容語と機能表現は接続不可能であると判定し、システムは何も出力しない。例えば、「聞く」と「だとたん」(「たとたん」の「た」が有声化した表現) は、接続不可能である。

4.4 フィルタリング

最後に、システムは、毎日新聞 1991-2005 年版 (計 15 年分) における出現回数に基づいて代替文節をフィルタリングし、順位付けする。具体的には、単純な文字列照合により、新聞 15 年分における、代替文節の出現回数を数え、その出現回数が多い順に代替文節を出力する。このとき、出現回数が 0 回のものは出力しない。

4.5 出 力 例

「聞くや否や」を入力したときの、言い換えシステムの出力を表 4 に示す。i 欄に条件なしの場合の順位を、j 欄に難易度条件を B 以下に設定した場合の順位を示す。

「見てくれるか」を入力したときの、言い換えシステムの出力を表 5 に示す。i 欄に条件なしの場合の順位を、j 欄に文体条件を常体のみに設定した場合の順位を示す。

これらの表より、提案システムが、望ましい代替文節を出力できることが分かる。また、条件を設定することにより、その出力を適切に制御できることが分かる。

5. おわりに

本研究では、体系的構造を持つ機能表現辞書に基づいて機能表現を言い換えるシステムを提案した。このシステムは、表記単位で 16,771 の機能表現に対して、類義表現を出力することができる。

表 5 入力「見てくれるか」に対する出力
(i: 条件なしの順位、j: 常体のみの順位)

i	j	代替文節	出現回数
1	-	見て下さい	859
2	-	見てください	237
3	1	**見てくれるか**	23
4	2	**見てもらえるか**	12
5	3	**見てくれないか**	8
6	-	見てちょうだい	5
7	4	**見てもらえないか**	4
8	-	見て下さるか	3
9	-	見ていただけますか	2
9	-	見てくれますか	2
9	-	見てくれませんか	2
9	-	見てもらえませんか	2
13	-	見ていただけないですか	1
13	-	見てはもらえないでしょうか	1
13	-	見て下さいませ	1
13	-	見て下さいませんか	1
13	-	見て頂戴	1

今後の課題は、複数の機能表現からなる機能語部を言い換えるシステムの実装とその評価である。

本研究の一部は、次の研究費による: 科学研究費補助金 基盤研究 (A) 「円滑な情報伝達を支援する言語規格と言語変換技術」(課題番号 16200009)。

参 考 文 献

1) 松吉俊, 佐藤理史, 宇津呂武仁: 階層構造による日本語機能表現の分類, 言語処理学会第 12 回年次大会, pp. 408–411 (2006).
2) 乾健太郎, 藤田篤: 言い換え技術に関する研究動向, 自然言語処理, Vol. 11, No. 5, pp. 151–198 (2004).
3) 飯田龍, 徳永泰浩, 乾健太郎, 衛藤純司: 言い換えエンジン KURA を用いた節内構造および機能語相当表現レベルの言い換え, 第 63 回情報処理学会全国大会予稿集第二分冊, pp. 5–6 (2001).
4) 土屋雅稔, 佐藤理史, 宇津呂武仁: 機能表現言い換えデータからの言い換え規則の自動生成, 言語処理学会第 10 回年次大会発表論文集, pp. 492–495 (2004).
5) Tanabe, T., Yoshimura, K. and Shudo, K.: Modality Expressions in Japanese and Their Automatic Paraphrasing, *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS)*, pp. 507–512 (2001).
6) Shudo, K., Tanabe, T., Takahashi, M. and Yoshimura, K.: MWEs as Non-propositional Content Indicators, *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrating Processing (MWE-2004)*, pp. 32–39 (2004).
7) 森田良行, 松木正恵: 日本語表現文型 用例中心・複合辞の意味と用法, アルク (1989).
8) 遠藤織枝, 小林賢次, 三井昭子, 村木新次郎, 吉沢靖: 使い方の分かる類語例解辞典新装版, 小学館 (2003).
9) 佐藤理史: 異表記同語認定のための辞書編纂, 情報処理学会研究報告 2004-NL-161, pp. 97–104 (2004).
10) 国際交流基金, 財団法人日本国際教育協会: 日本語能力試験出題基準【改訂版】, 凡人社 (2002).