山形方言から共通語への翻訳システム

佐藤守†, 横山晶一††, 西原典孝††

†山形大学大学院理工学研究科, ††山形大学工学部
E-mail: †m01833@eiefs.yz.yamagata-u.ac.jp, ††{yokoyama,nisihara}@yz.yamagata-u.ac.p

1 はじめに

方言とは特定の地域に限って使われる言語である. 一つの国語または民族語が地域によって独自の発達を示し, 音韻・語彙・文法の上で標準語と異なった姿を見せたとき, その個々の言葉および, その言葉を一つの要素として含むその地域の言語全体を言う. また, なまりとは, 普通地方性の強い発音をさす. 現在, 地域社会で使われる日本語は, 共通語に適度に方言が入り込んだものであると言える [1].

方言の研究は従来, 社会言語学や民俗学的な観点から取り扱われ, 機械処理の観点から取り扱われることは数少ない. 方言は社会言語学などではその存在が認められながらも, 日本語教育では閑却されてきた. 山形大学教育学部の高木らのグループでは, 定住型外国人たちが日本語習得の過程で, 共通語でその基礎が築かれるところへ方言が入り込むことによる混乱を防いだり, 地域社会で日本語学習を開始するような場合では, 方言と共通語とが無意識に混交し, 適切な使い分けができなくなったりするのを防ぐために「定住型外国人を対象にした "地域共通語" 教材開発に関する研究 [1]」が行われている.

方言の機械処理による研究意義は、方言と共通語の比較によって, 日本語の性格が明らかになり翻訳の研究にも有用であることである. 我々のグループは既に方言から共通語への翻訳システムのプロトタイプを作成し報告している [2]. 前回の段階での変換処理方式は, 辞書登録してある語彙に直接置き換える方式 (単語直接変換方式) と, 有声化した語彙を規則的に清音に変換する方式 (文字コード変換方式) と, 結合価による変換方式 (多義語翻訳) の3つであった. 今回はそれに例外処理を加えて, 形態素数や品詞が変化する変換を可能にした. 本稿では山形方言を共通語に翻訳する際に得られた知見を報告する.

2 村山方言概説

村山方言は, 山形市を中心とした, 内陸部の中央に位置する村山地方の方言である. この地方は,17 世紀以降小藩が分立し, 変動の激しい治世が続いたこともあって, 音韻や文法で南奥羽方言的ではあるが, 語彙では北奥羽方言的な要素をもちあわせたりする. このような理由で村山方言は, 山形県の方言である庄内方言, 置賜方言とは語彙や文法で異なる面がある [3].

2.1 音声面での特色

音声面での特色は, 語中や語尾にあるカ行, タ行がガ行, ダ行に変化する有声化,3 音節からなる形容詞が促音化したり, 発音の段階で「い」を「え」,「し」を「す」,「ち」を「つ」,「ひ」を「し」,「ゆ」を「い」,「せ」を「へ」と発音するものが多い [4].

(例 1) このかぎあがいげんと渋い。(このかきあかいけれど渋い。)
(例 2) このみがん、やっすいな。(このみかん、やすいな。)

表 1: なまり音と通常音の対応関係

なまり音	通常音	例語
い	え	えす (石), えしゃ(医者)
し	す	ずすん (地震), すんぶん (新聞)
ち	つ	つ (血), つつ (乳)
ひ	し	しと (人), しま (暇)
ゆ	い	いめ (夢), いび (指)
せ	へ	へなか (背中), へんべえ (煎餅)

2.2 村山方言文法

2.2.1 文末表現・述語の文法的カテゴリー

(a) **否定表現**
否定形は, 共通語の「ない」がなまって変化した「ね」を用いて表現する [5].
(例 3) すいが、かねが?(すいか、食べないか?)

(b) **希望表現**
希望を表す場合は, 助動詞「だい」を用いる. これは, 共通語の助動詞「たい」が有声化したものである [5].
(例 4) あしたスキーさいぎっだい!(あしたスキーに行きたい!)

(c) **意志・推量・勧誘表現**
意志・推量・勧誘を表す場合は, 終助詞「べ」を用いる. また, 同意を求める場合も「べ」が用いられる [5].
(例 5) 七日町さ行ぐんなら、一緒に行ぐべ。(七日町に行くなら、一緒に行こう。)

2.2.2 気づかない方言

村山方言には, 共通語と同形で意味・用法が異なるものが存在する. そのため, 方言と気づかずに (共通語だと思って) 使用している. このような言葉を「気づかない方言」と考えて研究しているグループもある [1].

(例 6) バスがらおちた。(バスから降りた。)

2.2.3 形式的対応のずれ

村山方言と共通語では, 文法・構造に大きな違いは見られないが, わずかに語彙の対応のずれや単語と句の対応などの形式的対応のずれが見られる.

(a) **語彙の対応のずれ**
共通語で「(学校に) 入学する」,「(仕事が終って) 帰る」,「(食べ物を) 食べる」などの動詞を村山方言では「あがる」で表す. これは, 共通語と村山方言の単語が 1 対 1 に対応しておらず, 意味する内容にずれがあるということを示している.

(b) **単語と句の対応 (層のずれ)**
村山方言の「うるがす」という語は, 共通語では「水につけて柔らかくする」という句に対応する. 他にも,「つかす」という語は, 共通語では「お世辞を言う」という句に対応する例が見られる.

(c) **構造の違い (構造のずれ)**
例えば, 記号の中に数字や文字が記述してあるときの読み方である. 村山方言では, 記号の中に記述してある数字や文字を先に読む. これを記号表現を用いずに書き起こすと以下のようになる.

まる 1　(接頭詞 + 名詞 : 共通語)
1 まる　(名詞 + 接尾詞 : 村山方言)

は句における構造のずれである.

3 翻訳システムと実行例

3.1 処理概要

村山方言から共通語への翻訳処理の流れを図 1 に示す. 次の 4 ステップから構成される:

[ステップ 1] 形態素解析 (part-of-speech tagger):入力された文書を単語に分割し, 品詞を付与する.
[ステップ 2] 語彙トランスファ(vocabulary transfer): 辞書引きにより語彙変換を行う.
[ステップ 3] 意味解析 (semantic analysis):原言語の文書が持つ意味内容に合致する語彙変換を行う.
[ステップ 4] 形態素構造生成 (morpheme structure generation):変換された単語を連結し, 文書を生成する.

ステップ 1 の形態素解析には茶筌 [6] を用いる. ステップ 2 の語彙トランスファでは語彙辞書を参照して一意の訳語を持つ語彙についてのみ変換する. 活用する品詞の場合は, 活用形定義辞書を参照して語尾変換を行う. ステップ 3 の意味解析では一意に定まらず多義の訳語を持つ語彙について, 結合価辞書を参照して変換を行う. 本来, 機械翻訳システムは, ある言語から別の言語に変換するのが目的であるため, 辞書引きした後に構文解析などのプロセスが必要となる [7]. しかし, 同じ日本語内での翻訳で, 語順に相違が見られないため, 構文解析などのプロセスは省ける. ステップ 2〜ステップ 3 の処理を全ての形態素について行った後に, ステップ 4 で共通語の形態素構造を生成し出力する.

3.2 茶筌のカスタマイズ

茶筌を, 村山方言の形態素解析器として機能するようにカスタマイズする. その際に, 従来の共通語解析器としての機能を損失しないようにする.
単語登録, および連結表を作成する際に,vgramtools for Japanese(version 0.01)[8] を使用する. このツールは, 茶筌の日本語辞書を作成する script 群である.

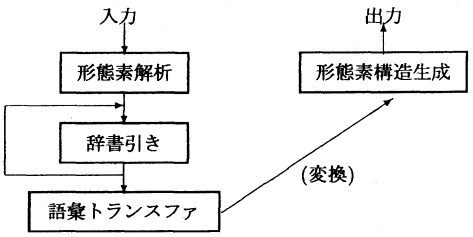

図 1: 村山方言から共通語への翻訳処理の流れ

3.2.1 意味素性の付加

茶筌の辞書作成の際に, 名詞に意味素性を付加する. 意味素性とは, 名詞の持つ意味を表したものである. 具体的には, 品詞名に「名詞-一般-(*)」のように (*) の部分に意味素性が付加されるように,「一般」の下位層に意味素性名を追加した.
カスタマイズした茶筌で「びっき」を解析すると,

```
>びっき
>びっき　ビッキ　びっき　名詞-一般-動物
```

となる. このように, 名詞に意味素性を付加することで, 結合価を用いた多義語翻訳が可能となる.

3.3 辞書

3.3.1 語彙辞書

各方言語彙に対応する共通語語彙などが記述されている辞書である. この辞書は品詞ごとに用意されており, 共通パラメータは「方言語彙-見出し語」,「方言語彙-読み」,「分類番号」,「訳語数」,「共通語語彙-見出し語」,「共通語語彙-読み」である. 活用語の場合は, これに「基本形-見出し語」,「基本形-読み」,「語幹-見出し語」,「語幹-読み」の 4 つが加えられる. また, 結合価に関与する名詞・助詞については「意味素性」が加えられる.

表 2: 名詞語彙辞書

方言語彙	読み	分類番号	訳語数	共通語	読み	意味素性
いぎ	イギ	1	2	息 雪	イキ ユキ	生理 自然物
おなご	オナゴ	1	1	女	オンナ	-
かぎ	カギ	0	0	-	-	-
びっき	ビッキ	1	1	蛙	カエル	-

(a) **分類番号**
方言語彙を分類するための番号. 共通語が規則的 (有声化, 促音化など) に変化したものを 0, 不規則的に変化したものを 1, 単語が句になったり, 品詞が異なったものに変化したりと例外的に変化したものを 2 とする. 表 3 に分類番号とそれに対応する例語を示す.

表 3: 分類番号表

分類番号	変化規則	方言語彙	共通語語彙
0	規則的に変化	こだづ へなか つつ	こたつ せなか ちち
1	不規則的に変化	べご びっき	牛 蛙
2	例外的に変化	こわい	疲れた

(b) **訳語数**
語彙辞書に登録されている方言語彙に対する共通語語彙数を指す.

3.3.2 結合価辞書

村山方言の動詞についての結合価が記述されている辞書である. 結合価とは, ある動詞を基準とした場合に, 共に出現する名詞・格助詞の情報をまとめたものである. 多義語を翻訳する際に参照する. 具体的には, 語彙辞書に記述されている方言語彙で訳語数が 2 以上かつ分類番号が 1 のものを翻訳する際に参照する. この辞書を作成する際には, 板坂の研究成果 [9] を参考にした.

3.4 処理方式

本システムの処理方式は, 処理のレベルに応じて, 次の四つに分類される.

1. 単語直接変換方式
2. 文字コード変換方式
3. 結合価による変換方式
4. 例外処理

以下に, それぞれの方式について述べる.

3.4.1 単語直接方式

最も単純な直接変換方式は, 計算機で自然言語を扱うための文法理論がまだ提案されていなかった初期の機械翻訳システムで用いられた方式であり, 単語レベルの語彙辞書による訳語への置き換えによる変換からなる.
例えば, 次のような村山方言文の翻訳を考える.

(文 1) びっきばへめろ。

まず, 解析系では, 形態素解析器 (茶筌) が単語を認識し, 原形に直す. 結果は次のような単語の並びとなる.

形態素解析

びっき(名詞),ば(助詞),へめろ(動詞, 一段, 命令形),。(記号)

次に, おのおのの単語を語彙辞書を参照して, 目標言語 (共通語) の単語に変換する.

形態素解析

蛙(名詞),を(助詞),捕まえろ(動詞, 一段, 命令形),。(記号)

最終的に, 全ての単語を連結して, 目標言語の文を生成する.

(文 2) 蛙を捕まえろ。

この単語直接変換方式は, 表層的な単語レベルの解析しか行わないため, 文の構造的な情報や意味情報を訳文に反映させることができない.

3.4.2 文字コード変換方式

この変換方式では, 音声面での特徴を利用する.

(a) **濁音 → 清音変換**
濁音は清音より文字コードの 2 バイト目が 1 多い値になっていることから, 濁音の 2 バイト目の文字コードを 1 減らせば清音に変換することが可能である. この処理を全ての文字について行うことで, 有声化した語彙を共通語に変換する.

(b) **なまり音 → 通常音変換**
この場合は, 有声化のように全てのなまり音の文字コードをある値だけずらすことによって通常音に変換するといった体系化ができない. よって, なまり音の場合はハッシュを用いる. ハッシュとは連想配列のことで, 配列の添字に数字ではなく文字列をキーとして使用することが可能である. つまり,(キー, 値) = (なまり音, 通常音) というデータ形式で格納しておくことで, 規則的になまり音を通常音に変換可能となる.

3.4.3 結合価による変換方式

訳語数が 2 以上で分類番号が 1 の方言語彙の変換方式について述べる. まず, 以下の 2 文の翻訳を考える.

(文 3) いぎが降る。
(文 4) いぎば吐ぐ。

結合価

降る: N[NAT/PLA/PRO]_+V
吐ぐ: N[ANI/HUM]_+N[PHE] ば+V

名詞「いぎ」は, 表から分かるように訳語数が 2 の多義語である. 文中に含まれる動詞の結合価によって適切な共通語に変換する. 文 3 の場合は, 動詞「降る」の結合価をみると 'N[NAT/PLA/PRO]_' にマッチする. 文 4 の場合は, 動詞「吐ぐ」の結合価をみると 'N[PHE] ば' にマッチする. 表のいぎをみると, 意味素性が自然物, 生理の場合は「雪」,「息」となる.

(文 5) 雪が降る。
(文 6) 息を吐く。

3.4.4 例外処理

この処理方式は,1 単語レベルの置き換えでは変換できない方言語彙に適用される.2.2 で述べた形式的対応のずれが生じているものについては, この処理方式で変換する.

(文 7) なにつかしてる。

文中の動詞「つかす」は「お世辞を 言う」という句に変換されるが,「お世辞を」の 2 形態素は修飾部なので,「なに」と「つかす」の間に挿入する.「つかす」の活用に合わせて「言う」を活用させる.

(文 8) なにお世辞を言ってる。

表 4: 実行例

	入力	出力
(1)	明日東京さ行ぐ。	明日東京に行く。
(2)	北さ行ぐ。	北へ行く。
(3)	このかぎあがいげんと渋い。	このかきあかいけれども渋い。
(4)	あのおなごめんこい。	あの女かわいい。
(5)	このみがんやっすい。	このみかんやすい。
(6)	高校さあづまらね。	高校にあつまらない。
(7)	あいづじぇねねーおどご。	あいつお金なきおとこ。
(8)	びっきばへめろ。	蛙を捕まえろ。
(9)	びっきばへめろずー。	蛙をつかまえろ！
(10)	さるのけっつ真っ赤だごど。	さるのおしり真っ赤だこと。
(11)	やろこ、この前の運動会でびりけっつだった。	男の子、この前の運動会で最下位だった。
(12)	じゃあ、1まるを佐藤君解いて。	じゃあ、まる1を佐藤君解いて。
(13)	なにつかしてるか。	なにお世辞を言ってるか。
(14)	そだなごどしゃね。	そんなこと知らない。

3.5 実行例

本システムは, 村山方言から共通語への翻訳結果に加えて, 村山方言と翻訳後の共通語の形態素解析結果をオプションで表示する. 形態素解析結果では, 茶筌と同様に「見出し語」「読み」「基本形」「品詞名」「活用型名」「活用形」の6つを出力する. 最初の形態素解析結果は原言語 (村山方言) のものであり, 次に表示される形態素解析結果は目的言語 (共通語) のものである. ここでは, 形態素解析結果を除いた翻訳結果のみを表4に示す. 入力は原言語 (村山方言), 出力は目的言語 (共通語) である.

4 おわりに

方言を一種の言語とみなして機械処理の観点, すなわち翻訳システム構築という観点から研究を行った結果, 次のような示唆が得られた.(1) 気づかれない方言は, 共通語語彙と同形であるが意味は異なる. このことは, 昔使われていた言葉が, 表層的変化はないが深層的変化によって現在の共通語となったものが存在することを示す.(2) 縮約形や促音化などは, 共通語の話し言葉にも見られる現象であるが, 促音化については規則性が発見できた. (3) 村山方言と共通語の翻訳システムを構築することで, 次のようなことが明らかになった.

(3-1) 語順の変化がほとんど見られないことから, 構文解析のプロセスは省ける.

(3-2) 活用語, 主に動詞であるが, 村山方言とそれに対応する共通語で活用の仕方が異なるため, 後接する語によって活用語尾を変化させる必要がある. つまり, 活用語については単なる置き換え処理だけでは変換できず, それに活用語尾処理を加える必要がある.

(3-3) 現段階での語彙辞書の語彙数を見ても分かるように, 方言語彙は全部で数千単位ではないかと推測できる.

このことから, 方言語彙と共通語語彙の対応関係と活用語の活用の仕方が分かれば, その地方の方言を理解でき, 方言話者にもなれるのではないかと推測できる. つまり, 方言語彙を全て収録した辞書を完備した方言翻訳システムでの学習が, 十分効果的であるといえる. また, 村山方言と山形県内の他地方の方言 (庄内方言 [10]) との比較を行ったところ, 語彙の相違は見受けられるが, 構造の相違は見られない. すなわち, 本システムに他地方の語彙辞書を完備すれば, その地方の方言も共通語に翻訳可能であることが分かる.

参考文献

[1] 高木裕子:定住型外国人を対象にした “地域共通語” 教材開発に関する研究, 平成 10 年度-平成 12 年度科学研究費補助金基盤研究 (B)(2) 研究成果報告書 (2002)

[2] 佐藤守, 横山晶一, 西原典孝:方言から共通語への翻訳システムに関する基礎的研究, 第 150 回自然言語処理研究会 (2002.7) NLC2002-22

[3] 平山輝男他:山形県のことば, 明治書院 (1997)

[4] 尾花沢市「昔を語る会」:尾花沢の方言 (1992)

[5] 山形地域語研究会:山形ことばを学ぼう, 山形地域語研究会 (2001)

[6] 形態素解析システム「茶筌」, 奈良先端科学技術大学院大学

[7] 横山晶一:自然言語の理解,AI-情報処理から知能処理へ-, アスキー (1988)

[8] vgramtools for Japanese(version 0.01) , 奈良先端科学技術大学院大学

[9] 板坂智裕:村山方言解析システム作成のための基礎的研究, 山形大学卒業論文 (1999)

[10] 横山晶一, 安野克彦:方言の機械処理に関する予備的考察, 電子情報通信学会技術報告,NLC95-45(1995)