

## 年代差を考慮した方言翻訳システム

柴田 直由<sup>†</sup> 横山 晶一<sup>††</sup> 井上 雅史<sup>††</sup>

<sup>†, ††</sup> 山形大学大学院理工学研究科情報科学専攻

〒992-8510 山形県米沢市城南4-3-16

E-mail: †tfw98479@st.yamagata-u.ac.jp, ††{yokoyama, mi}@yz.yamagata-u.ac.jp

**あらまし** 方言は、その方言を扱う地方の人間との会話をより円滑にし、より良い人間関係を築くためには必要不可欠なものである。方言理解の補助手段として機械翻訳の利用は有効である。ここでの機械翻訳は、他国語間の翻訳レベルではなく同じ日本語内での翻訳であるため、一般の機械翻訳に比べてそれほど困難ではないと考えられる。我々のグループでは既に共通語から方言(特に村山方言)へ、品詞ごとの置き換えで、ある程度自然な訳文が得られる翻訳システムを作成している。本稿では、年代差を考慮することで、従来のシステムよりも精度の高い訳文生成が可能となるシステムの構築を目指した。出力結果に対してアンケートを取ったところ、従来のシステムよりもよい、約84%の正答率が得られた。

**キーワード** 村山方言、共通語、翻訳システム

## Dialect Translation System considering Age Difference

Naoyoshi SHIBATA<sup>†</sup>, Shoichi YOKOYAMA<sup>††</sup>, and Masashi INOUE<sup>††</sup>

<sup>†, ††</sup> Graduate School of Science and Engineering ( Informatics ), Yamagata University,

4-3-16, Jonan, Yonezawa-shi, Yamagata, 992-8510, Japan

E-mail: †tfw98479@st.yamagata-u.ac.jp, ††{yokoyama, mi}@yz.yamagata-u.ac.jp

**Abstract** The dialect is necessary for the smooth communication with people who live in the country side and usually use the dialect. It is not so difficult for the machine translation between the standard Japanese and a dialect because the translation is limited inside in Japanese. We have already constructed the machine translation system from the standard Japanese to Murayama dialect in Yamagata prefecture which generates relatively natural outputs using the replace of each part of speech. This paper describes the construction of the more accurate system considering the age difference. The result of a questionnaire shows about 84% accuracy rate.

**Key words** Murayama dialect, Common language, Translation system

## 1. はじめに

方言とは、一定の地域社会において使用されている日常生活語である。また方言とは、その地域の人々との会話をより円滑にし、よりよい人間関係を築くためには必要不可欠なものである。そこで、方言を理解する有効な手段として、方言機械翻訳システムを利用した方言学習が挙げられる。

言語学では、方言学という学問体系が確立しており[1-3]、音韻やアクセント、語彙、文法など、様々な面から研究が行われている[4]。一時共通語の普及を図るために、方言を圧迫していた時代もあったが、近年では地方の活性化のためにあえて方言を利用する場面も増え、マスメディアで取り上げられることも多々ある。

方言機械処理の研究目的は以下の通りである。

- (1)方言は、語彙や文法の面で古い日本語を保存していると言われており[5]、規則的な解明が現代語に対する示唆を与えることが期待される
- (2)方言は主に話し言葉として伝えられてきた。そのため、方言文法を作成することは、話し言葉文法を作成することに寄与する
- (3)その方言を使う地域の人々とより良い人間関係を築くための(方言を理解するための)、有効な手段となる可能性がある

以上の考えから、我々は方言を電子的に扱う研究に着目した[4]。その後、平成18年度には、共通語から村山方言への機械翻訳システムのプロトタイプ[6]が構築され、平成20年度には付属語の拡張[7]による精度の向上、平成21年度には方言辞書の拡張[8]による精度の向上が図られた。

これらの研究で、助詞や助動詞といった付属語、または文末表現の方言はどの年齢層でも使用されているが、名詞や動詞といった方言は若い年齢層になるにつれ使用されなくなっており、訳文生成の精度が落ちるという傾向があることが判明している。

以上を踏まえ、本研究では、先行研究である”共通語から村山方言への機械翻訳システム”について、年代差という問題に着目し、年代別に訳し分けが可能になるよ

うシステムの改良・拡張を行った。具体的には、新しいパラメータとして年齢情報を追加し、システムの出力もそのパラメータに応じて出力ができるように拡張した(後述)。

改良・拡張した本システムの出力についてアンケートによる評価を実施したところ、約84%の正答率が得られ、従来より精度の高い方言翻訳文の生成が可能となった。よって、方言機械翻訳において年代差を考慮することは有効であることが示された。

本稿の構成は次のようになっている。まず次節では、村山方言の特徴について概説する。3節においては、年代差を考慮した翻訳システムの処理概要を述べ、実行例を示す。4節では、いくつかの文についてアンケート評価を行い、本システムの有効性について検証する。5節は総括である。

## 2. 村山方言概説

山形県は、庄内・最上・村山・置賜の四地方にわかれ、それぞれの地域で特有の方言が使用されている。

村山方言は、山形市を中心とした、山形県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、歴史的な背景から、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持ち合わせている[9,10]。このため村山方言は、山形県の他の地方の方言である庄内方言(北奥羽方言)、置賜方言(南奥羽方言)とは語彙や文法で異なる面がある。

### 2.1 音声面での特徴

#### (1)有声化(濁音化)

村山方言では、語中や語尾にあるカ行・タ行に濁音が付き、ガ行・ダ行の濁音に変化する。これは東北地方全体に見られる特徴であり、共通語よりも濁音が多くなるので、これが「ズーズー弁」と呼ばれる要因の1つになっている。

(例) スイカが食べたい。 ⇒ すいがん食だ**い**。

明日晴れ**だ**と**い**いな。 ⇒ 明日晴れ**だ**と**い**いな。

## (2)促音化

村山方言では、3音節からなる形容詞は促音化することが多い。また、語尾がラ行の動詞の場合、その語尾が促音化することが多い。

(例) この商品はたかい。⇒ この商品たっかい。  
何見ると思う? ⇒ 何見っど思う?

## 2.2 助詞

### (1)主格「が」、「は」

村山方言では、主格を表す「は」、「が」が省略されることが多い。

(例) 雪が止んだ。⇒ ゆぎ\_\_止んだ。  
ここはどこだろう? ⇒ ここ\_\_どこだべ?

### (2)場所・方向「へ」、「に」

共通語で場所や方向を表す「へ」、「に」などを、村山方言では区別せずに「さ」で表す。

(例) 海へ向かう。⇒ 海さ向かう。  
山に登る。⇒ 山さ登る。

### (3)受け手「に」

受身文の場合、受け手は共通語では「に」を用いるが、村山方言では「から」、または有声化した「がら」を用いる。

(例) 母に怒られた。⇒ 母がらごしゃがった。  
車に跳ねられた。⇒ 車がら跳ねらった。

この他にも、目的「を」が変化した「ば」、逆説仮定「なら」が変化した「ごんたら」、逆説限定「けど・けれども」が変化した「げんと・げんとも」といった助詞が存在する。

## 2.3 文末表現・述語の文法的カテゴリー等

### (1)否定表現

共通語の否定形を、村山方言では助動詞「ない」が訛った「ね」を用いて表す。接続は共通語と同様に動詞の未然形につく。五段動詞中で最後に「る」がつくものや、可能の意味を含む場合の「れる・られる」の後ろにつく場合、「～らね」ではなく「～んね」とラ行が撥音化する。

(例) 行かない。⇒ 行かね。  
登れない。⇒ 登らんね。

### (2)推量・意思・勧誘表現

村山方言では助動詞「べ」を用いて表す。最後が「る」で終わる動詞と接続する場合、ラ行が促音化する。同意を求める場合も、同じく「べ」を用いる。

(例) 彼は来るだろうか。⇒ 彼来っべが。  
よし、勉強しよう。⇒ よし、勉強すっべ。  
買い物に行こう。⇒ 買い物に行ぐべ。

### (3)義務・当然表現

共通語の「～(し)なければならない・なくてはならない」という義務・当然を表す句を、村山方言では「～(さ)んなね」で表す。

(例) 勉強しなくてはならない。⇒ 勉強さんなね。  
行かなければならない。⇒ 行かんなね。

### (4)依頼表現

共通語の「～(し)てくれる?」は、村山方言では「～(し)てける?」のように表す。一段動詞と五段動詞の中で「く」で終わる動詞の場合、「～でける?」のように「て」が有声化する。共通語では他に「～(し)てくれない?」や「～(し)てくれ」等様々な形があるが、村山方言でもそれぞれに対して様々に変化する。

(例) 一緒に来ててくれる? ⇒ 一緒に来ててける?  
引いてくれない? ⇒ 引いてけね?  
財布を探しててくれ。⇒ 財布ば探しててける。

### (5)受身表現

村山方言で受け身を表す場合、共通語と同様に「れる・られる」を用いるが、動詞との接続部分が促音化、撥音化することが多い。2.2の(3)で述べた、受け手「に」とセットでよく使われる。

(例) 母に怒られた。⇒ 母がらごしゃがっだ。  
車に跳ねられた。⇒ 車がら跳ねらった。

その他にも、希望、禁止、尊敬など、数多くの言い回しや表現が存在する。

## 2.4 年代による方言の違い

先行研究より、若い世代、特に20代以下の年代の多くは、名詞や動詞、形容詞といった方言語彙を知らないということがわかっている[8]。一方で、活用形方言語彙の場合、語幹は共通語と同じ語を取るが、活用部分や活用型が村山方言文法に倣うようである。例えば共通語「持つ(タ行五段動詞)」の場合、村山方言では「たがぐ(ガ行四段動詞)」と言うが、若い世代では「持つ(ダ行四段動詞)」というように、活用部分が有声化し活用型も変わる。

以上の理由により、訳文の精度を向上させるには、年代による方言語彙の訳し分けを行う処理が必要であると考えられる。

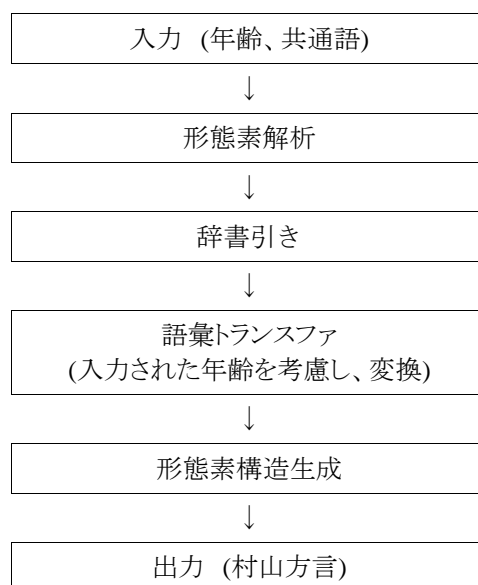


図1 共通語から村山方言への翻訳処理の流れ図

## 3. 翻訳システムと実行例

### 3.1 処理概要

処理の流れは以下の4ステップで構成される。

#### step1. 形態素解析

入力された文章を単語に分割し、品詞を付与する。ここでは、形態素解析器に茶筌2.4.4[11]を、形態素解析用辞書に NAIST-JapaneseDictionary0.4.3[12]を使用している。

#### step2. 語彙トランスファ

年代を考慮した辞書引きにより、語彙変換を行う。

#### step3. 意味解析

原言語の持つ意味内容に合致するように、語彙変換を行う。

#### step4. 形態素構造生成

変換された形態素を連結し、訳文を生成する。

同じ日本語内の翻訳であり、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略できる。基本的なシステムの流れは、従来のシステムを踏襲した。翻訳処理システムの全体の流れを以下の図に示す。

### 3.2 語彙辞書

方言辞書に記述する基本パラメータは、「共通語語彙－見出し」「共通語語彙－読み」「方言語彙－見出し」「方言語彙－読み」である。活用語の場合はこれに加え、「基本形－見出し」「基本形－読み」「語幹－見出し」「語幹－読み」をパラメータとして記述してある。今回新たに、年代別に翻訳が可能になるよう、「年齢1」「年齢2」という年齢パラメータを基本パラメータとして追加した。この年齢パラメータは筆者(柴田)の主観によって与えた値だが、今回行ったアンケート調査に掲載した単語については、アンケート結果からより適切な値になるように修正してある。辞書のデータ形式は、各値をカンマで区切ったデータ形式の CSV(Comma Separated Value)形式でまとめている。現在は1800語程度の語彙が登録されている。

例として、名詞辞書の一部を表1に示す。なお、太字部分が今回新たに追加したパラメータを表している。年齢1は、この変換が適用される下限年齢を示し、年齢2は上限年齢を示す(次節参照)。この表中の年齢2パラメータに書かれてある200という値は、正の無限大を意味している。

表1 名詞辞書の一部

共通語 語彙一 見出し	共通語 語彙一 読み	方言語 彙一見 出し	方言語 彙一読 み	年齢1	年齢2
氷柱	ツララ	ぼんだ ら	ボンダ ラ	50	200
同じ	オナジ	おんな す	オンナ ス	30	200
尻尾	シッコ	おっぱ	オッパ	30	200

### 3.3 年代別語彙変換処理

共通語を入力する前に、年齢を入力するコマンドラインを追加した。辞書引きから語彙変換を行う際、この入力された年齢と、語彙辞書に記載されている年齢1、年齢2のパラメータをそれぞれ比較し、「年齢1 ≤ 入力年齢 < 年齢2」という条件を満たす、共通語語彙に対応する方言語彙を検索する。条件に一致する方言語彙が存在するならば、その方言語彙に変換し、存在しない場合は、変換を行わずにそのまま共通語として出力し、次の語彙変換処理に移る。

年代別に出力される例(この例はアンケートで比較的评价が高かった例である)を以下に示す。

#### (例1\_1)

入力年齢：23

共通語：昨日 教えた じゃないか。

村山方言：昨日 教えだっけ べしたん。

#### (例1\_2)

入力年齢：35

共通語：昨日 教えた じゃないか。

村山方言：昨日 おしえだっけ べしたん。

#### (例1\_3)

入力年齢：52

共通語：昨日 教えた じゃないか。

村山方言：きんな おしえだっけ べしたん。

「昨日」、50歳以上では「きんな」、また、共通語動詞「教える」を30歳未満ではそのまま「教える」、30歳以上では「おしえる」と変換されるよう年齢パラメータを与えている。そのため上記の例では、各例文のように入力された年齢に適した方言に訳している。

また、このシステムで与えている年齢パラメータは、今回のアンケート調査で得られた結果から、方言使用有無の境界線と思しき差が見られた単語について主観的に推定した値である。

### 4. アンケート評価

山形県立寒河江高等学校の1、2年生とそのご家族の計1200人を対象に、紙媒体で以下のようなアンケートを実施した。その中で回答が得られた255人分のデータを対象に評価・統計を行った。回答者の内訳については以下の表2に示す。在校生とその家族を対象としているため、10代以下と40代、50代に回答者が集中しているが、その他の年代も存在している。

上記の高校は筆者(柴田)の母校であり、村山地方の各市町から生徒が集まってくる。方言の年代差、地域差の調査を行う対象としてふさわしいと考えたため、以上を対象として選択した。

表2 アンケート回答者の内訳

	10代 以下	20 代	30代	40 代	50 代	60 代	70代 以上	年代 無回 答	計 (人)
	77	5	7	89	43	8	17	9	255
男性	32	1	2	29	26	3	6	0	99
女性	44	4	5	60	17	5	10	0	145
性別 無回 答	1	0	0	0	0	0	1	9	11

ここでは、共通語名詞「昨日」を50歳未満ではそのまま

#### 4.1 アンケート内容

共通文26文について、それぞれ比較的高齢の方(ここでは50代以上)が使用していると思われる方言1と、比較的若い世代が使用していると思われる方言2の2つの出力を用意し、方言1、2の両方について「その方言が正しいかどうか」を問い、「Yes/No」で判断してもらおう。「No」である場合や掲載した翻訳文以外の言い回しや表現、語彙等がある場合は、「その他」の欄に任意に答えてもらい、出力された方言が間違っている場合には、任意で出力文を添削して頂くというものである。このアンケートには、本システムが出力した翻訳文を使用した。

また、アンケートを取る目的として、「この研究を進める中で、出力された方言の正しさを確かめる必要が出てきた。方言は最近、世代によって使われ方が変わってきており、方言の使い方を高校生やご家族の方などの幅広い年代の方に評価して頂く必要があると考えたため、協力して頂きたい。」といった内容をアンケート協力者に明示した。

アンケート内容の一例と記入例を以下に示す。

##### 記入例

共通語 ( 昨日 教えた じゃないか。 )  
 方言1 ( ~~きんな~~ おしえだっけ べしたん。 ) (Yes・No)  
 方言2 ( 昨日 教えだっけ べしたん。 ) (Yes・No)  
 その他 ( 「べしたん」だけでな、く「べした」とも言う。 )

図2 アンケートに記載した記入例

##### 設問1

共通語 ( 昨日 教えた じゃないか。 )  
 方言1 ( きんな おしえだっけ べしたん。 ) (Yes・No)  
 方言2 ( 昨日 教えだっけ べしたん。 ) (Yes・No)  
 その他 ( \_\_\_\_\_ )

図3 アンケートに掲載した質問の一部

#### 4.2 評価と考察

総合評価として、各設問の正答率の平均値を計算し、約84%という結果が出た。この正答率は次の方法で算出した。方言1または方言2のどちらかが「Yes」だった場合は「Yes」として「Yes」の数をカウントし、どちらも無回答だ

った場合は「無回答」として「無回答」の数をカウントする。最終的に以下の式(1)の計算を行い、正答率を求める。

$$\frac{\text{「Yes」の総数}}{\text{データの総数} - \text{「無回答」の数}} \times 100 \quad (1)$$

アンケートでは、方言1の評価から、方言語彙(特に名詞・動詞・形容詞)の使用について年代差が見られた。特に、以下に記した設問2の内容が、年代差がよく現れた例である。この設問では、方言1の正答率が50代以上では80%を超えるのに対し、30代以下では30%を下回る傾向が見られた。逆に、方言2では30代以下の正答率が70%を超える傾向が見られ、方言語彙の使用有無の区切りがはっきりと分かれた。

##### 設問2

共通語 ( ちよつと 財布を 探してくれ。 )  
 方言1 ( ちえつと 財布ば たねでけろ。 )  
 方言2 ( ちよつと 財布ば 探してけろ。 )

図4 設問2の内容

逆に、以下に記した設問20の内容は、年代差が現れなかった例である。この設問では、方言1、方言2の両方の正答率どの年代を通して70%程度と高かった。これは、共通語形容詞「重い」の村山方言「おもだい」、動詞「持つ」の村山方言「だがぐ」を、年代問わず使用しているからだと考えられる。

##### 設問20

共通語 ( 無理して 重いものを 持たなくても  
 大丈夫だよ。 )  
 方言1 ( 無理して おもだいものば たががたたって  
 大丈夫だよ。 )  
 方言2 ( 無理して 重いものば 持だねたって  
 大丈夫だよ。 )

図5 設問20の内容

また、正答率が低かった例として、以下の設問9がある。この設問では年代差があまり見られず、正答率も一番低く63%程であった。正答率が低くなった理由としては、次の2つが考えられる。まず、「後ろ前」と「みっともない」という語彙を、村山方言ではそれぞれ2通りの語彙で置き換えることができるため、方言1の内容を話す人とそうでない人に分かれたことが1つの要因として考えられる。詳しく説明すると、「後ろ前」という共通語は「うっしょまえ」と「かっちゃえ」という方言に、「みっともない」という共通語は「めんくさい」と「みだぐない」という方言に置き換えられ、計4通りの組み合わせが生じたために意見が分散したからだと考えられる、ということである。また、全体的に「うっしょまえ」という方言語彙の添削率が高かったため、この語彙があまり使われていないことが2つ目の要因として考えられる。

#### 設問9

共通語 (後ろ前に 着てたら みっともないでしょ。)

方言1 (うっしょまえに 着てたら めんくさいべしたん。)

方言2 (うっしょまえに 着てたら みっともないべしたん。)

図6 設問9の内容

アンケート全体を通して、20代以下では方言語彙の使用頻度がかなり低いことが判明した。しかし、方言特有の言い回しや文末表現については、どの年代・地域でもほとんど共通して使用されているようだ。このことから、言い回しや文末表現といったものが、村山方言の特徴として確固たる地位を築いていることがわかる。

今回のアンケートでは年代や地域に偏りが生じてしまい、全体的な年代差・地域差を見積もる情報が少ないので、さらに広くデータを収集するために Web 上でアンケートを行っている。アンケート内容は今回行ったアンケートと全く同じ内容で、研究室のホームページよりアクセスして頂く形となっている。現段階での書き込み件数は未だ0件であるため、書き込み件数を増やすためには Web ページのデザイン構成を練り直さねばならない。

また、アンケート調査に掲載した方言単語については、アンケート結果より年齢パラメータの閾値を推定し、より

適切なパラメータになるように修正することができたが、アンケートに掲載していない単語については未だに主観的に与えた値になっている。そのため、今後この掲載されなかった単語のパラメータについて、どのように適切な値を求めるかが問題になってくる。

## 5. おわりに

本研究により、文末表現、方言特有の言い回しを増強するとともに、年代別に語彙の訳し分けが可能になるようシステムを拡張することで、生成訳文の違和感を抑え、より柔軟な訳文を生成することが可能になった。

一方アンケートより方言の活用語尾の変化、有声化や撥音化などは人によって様々であり、共通語語彙1語彙に対してニュアンスによって方言語彙が様々に対応することがわかった。これが方言の翻訳を困難にしている要因であると考えられる。

以上の問題を克服し、方言らしさが現れるような翻訳または出力が可能になったシステムを利用することで、より効果的に方言を学習することが可能になると言える。

以下より、システムの問題点と今後の方針について述べる。

アンケートより、村山方言では細かなニュアンスの違いが非常に多く、方言語彙にあたる共通語の明確な決まりが存在しないため、一意的な翻訳が不可能に近いという問題が出た。この問題を解消するためには、今後、訳文候補を多数出力するような処理(例えば、生成訳文の上位3つを出力・提示するような処理)等をシステムに組み込むことが必要になると考えられる。

本システムでは形態素解析の結果が正確であるということ为前提に翻訳を行っている。そのため、形態素解析性能が優れている和布蕪[13]にシステムを移し変えることで、訳文生成の精度向上が図れると考えられる。

また、本システムは、出力された訳文を方言話者に納得して頂けるような訳文になるよう、より自然で柔軟な訳文の生成を目指している。そこで、現在注目を浴びているフレーズベース統計的翻訳(PB-SMT)の導入を考えている。これは PB-SMT の手法が、語順が似た言語間で

は高精度な翻訳が可能になるという利点を持つ[14]ため、従来よりも高精度で柔軟な翻訳が可能になると考えられるからだ。この手法を導入する際、大量の対訳データをいかにして集めるかが問題となってくるため、2言語間で翻訳に利用できるリソースが少ない状況の場合に有効な手法であるピボット翻訳(中間言語を利用した2言語間での翻訳)[15]の導入も考えられる。また、方言とは話し言葉であるため、出力を文字だけでなく音声出力を加えることにより、より方言らしい出力が可能になると考えられるほか、より学習効率が上がるのではないかと考えられる。

以上を踏まえ、今後より柔軟で方言らしい出力ができるようなシステムの構築や、方言から共通語に翻訳するシステムの構築に努めていきたい。

## 謝辞

アンケート調査にあたり、対応にあたって下さった山形県立寒河江高等学校大滝潤二校長、志田浩雄教諭、また、アンケートにご協力下さった皆様方に心から感謝します。本当にありがとうございました。

## 文 献

- [1] 藤原与一: 方言学の方法, 大修館書店(1977)
- [2] 北原保雄 監修、江端義夫 編: 方言・朝倉日本語講座10, 朝倉書店(2002)
- [3] 佐々木冠 他: 方言の方法・シリーズ方言学2, 岩波書店(2006)
- [4] 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)pp39-46
- [5] 飯豊毅一 ほか(編): 方言概説・方言講座学1, 国書刊行会(1986)
- [6] 尾形真美, 横山晶一: 共通語から村山方言への機械翻訳システム, 言語処理学会第13回年次大会発表論文集(2007)pp923-926
- [7] 工藤翔陽: 付属語に着目した村山方言翻訳システム, 山形大学卒業論文(2008)

- [8] 兼子真弓, 高橋誠, 横山晶一: 共通語から山形村山・置賜方言への機械翻訳システム, 平成21年度第6回情報処理学会東北支部研究会(2010)
- [9] 森下喜一: 標準語引・東北地方方言辞書, 桜楓社(2008)
- [10] 平山輝男: 山形県のことば・日本のことばシリーズ6, 明治書院(1997)
- [11] 奈良先端科学技術大学院大学: 形態素解析システム「茶釜」Ver.2.2.4
- [12] Hideki YAMANE, Masayuki ASAHARA: 形態素解析用辞書「NAIST Japanese Dictionary」Ver.0.4.3
- [13] 京都大学情報科学技術研究科-日本電信電話株式会社コミュニケーション科学基礎研究所共同ユニット: 形態素解析システム「和布蕪」、<http://mecab.sourceforge.net/>: 「和布蕪とは」参照(2010年12月1日アクセス)
- [14] 林健一、綱川隆司、宮尾祐介、辻井潤一: 句に基づく機械翻訳のための構文情報を利用したデコーディング, 言語処理学会第14回年次大会発表論文集(2008)pp41-44
- [15] Hau Wu, Haifeng Wang: Pivot language approach for phrase-based statical machine translation, *Machine Translation*(2007)Volume.21, Number 3, pp165-181