方言学習と方言変換ソフトの改善方法の考察

平成 19 年 2 月 2 日

情報電子工学科
三間 優

目 次

1 はじめに **1**

2 標準語と共通語と方言 **1**

3 日本語テキスト変換フィルタについて **3**

4 方言学習について **4**

5 web 上の新潟弁変換サイトについて **6**

5.1 単語等の変換 6

5.2 変換結果についての考察 9

5.3 文章変換 9

5.4 実験と考察 10

6 まとめ **18**

参考文献 **20**

概要

方言はまだ高齢者がよく使用しており、高齢者にとっては標準語で書かれている文章などは読みにくいのではないかと考えられる。また文章をコンピュータに読ませる場合、標準語の文章よりも方言の文章を読ませた方が聞き取りやすいのではないかと考えた。その他に、若い人達も方言を知っていた方が高齢者の人達とコミュニケーションを取りやすいのではないかと考えた。
現在、標準語の文章等を新潟の方言 (新潟弁) を使用したものに変換する Web サイトがある。また osaka という日本語テキストを大阪弁に変換できるフィルタなどもある。単語や語尾を置換するだけなので特定の表現を必ず同じ言葉にする。しかし、ひらがなの多い文章等を変換させてみると、まだ正確に文章を変換することができないという問題点があることがわかった。また変換しなくてもよいところも変換されてしまう可能性があることなどもわかった。
そこで本研究では、方言学習が重要と考える理由を挙げて、方言学習の意義を考えた。またコンピュータによる方言学習の方法にはどのようなものがあるのか簡単にまとめてみた。さらに、文章を解析し、単語に分割してそれぞれの品詞を表示できる ChaSen(茶筌) というソフトを用いて、変換サイトの問題点を改善する方法について考察し、ChaSen を利用した Perl のプログラムを作成して、より正確に文章等を新潟弁に変換することのできるフィルタの作成を検討した。

1　はじめに

新潟県には多くの方言があり、まだ高齢者を中心によく使われている。現在 Web 上には、共通語を各都道府県の方言に変換するページがある。しかし、文章などを変換させてみると、違う言葉を意味する方言に変換されるなどの問題点がある。そこで本研究では、Web 上の変換ページの問題点を解消するための方法を実験などをして考察していき、標準語を新潟県の方言 (新潟弁) に変換する精度を向上させて、より正確に変換ができるような変換フィルタを作成することを目標とする。

2　標準語と共通語と方言

まず、標準語と共通語と方言の説明をしていこうと思う。

標準語

「標準語」は、Standard language の訳語であり、音韻・語彙・語法などすべての面で国語の規範として尊重され、教育・法令などの公用語として用いられる言語である。「標準語」という用語は、明治 23 年 (1890) に岡倉由三郎が最初に使った。

標準語は理想であり、人為的につくられるものである。したがってきびしい規範である。言いかえれば標準語はその言語の価値を高めるためのものである。

共通語

「共通語」は、Common language の訳語であり、規範性をもつ「標準語」という用語と分別するために使用される語である。一般に「国内に方言差があっても、それを超えて異なった地方の人々が意志を通じあうことのできる言語」のことである。日本では東京語 (特にその山の手言葉) がこれにあたる。

共通語は現実であり自然の状態である。したがってゆるい規範である。言いかえれば共通語は現実のコミュニケーションである。

「共通語」は原義的には、異なった言語間のコミュニケーションに使われる第三の言語のことを指すものである。例えば、インドネシア各地で通用するマレー語、東アフリカにおけるスワヒリ語等である。英語は世界の多くの国で共通語として機能している。したがって、日本での用法はその原義に照らして、ややレベルを異にした使い方である。

方言

「方言」は、ある限られた地域に使われる、共通語とは異なる語彙・発音・語法である。訛り (なまり) や俚言 (りげん) のことでもある。語彙とは、ある言語、ある地域・分野、ある人、ある作品など、それぞれで使われる単語の総体のことである。

訛りとは、ある地方特有の発音や標準語・共通語とは異なった発音のことである。
俚言とは、その土地特有の単語や言い回しのことであり、俚語とも言う。

言語は変化しやすいものなので、地域ごと、話者の集団ごとに必然的に多様化していく傾向があり、発音や語彙、文法に相違が生じる。そのために、差異の程度が別の言語までには広がっておらず同じ言語の変種と認められるものの、部分的に他の地域の言葉と異なった特徴を持つようになったものを方言と呼ぶ。また、方言には同一地域内にあって、社会階層の違いによって異なる変種もある。日本語の方言は日本語の地域変種のことである。

日本では方言という語は標準語とは異なる地方ごとの語彙や言い回しなどを指して使う場合も多いが、この様な語彙の事は「俚言」といい、方言の一構成要素である。日本語の各方言はもっぱら口頭の表現に使われ、文字に書き表わされる事は、方言詩や民話集などの例を除けば、非常に少ない。そのため、方言は非常に失われやすい存在といえる。

- 方言の長所
  同じ地域内の人との会話がしやすくなる。特に高齢者の方と市役所の公務員の方が会話をする際、垣根をつくらない。
- 方言の短所
  他の地域または他の県の人は、意味が全くわからないことがある。標準語と方言の語があまりにも異なる場合は、全く意味がわからない。

新潟の方言について説明する。

新潟県は一般に、上越・中越・下越・佐渡に分けられるが、方言区画上、佐渡は西日本方言、下越の阿賀野川以北は東北方言であり、県の代表的方言は、下越に属する新潟市よりも、中越に属する県中央部の長岡市や三条市あたりとなる。その代表的方言の特徴を以下に示す。

発音の特徴

- 「イ」と「エ」を区別しない
- 「火事」を「クヮジ」と歴史的仮名遣いのまま発音
- ダ行音をラ行音に発音することが多い
- アクセントは、「ハシラ」、「ウサギ」、「スズメ」などの 3 拍語に頭高が多い

文法の特徴は以下の通りである。

- ハ行動詞「払った」は「ハロータ」、形容詞「良く」は「ヨー」のようにウ音便
- 「見る」、「出る」の命令は「ミレ」「デレ」
- 「読む」の推量は「ヨムロー」
- 終助詞では、穏やかなものに「ノー」、強いものに「コテ」、「ネカ」がある

3　日本語テキスト変換フィルタについて

ここでは本研究をしようと思うきっかけになった「osaka」というソフトウェアを紹介する。

osaka とは、日本語テキストを大阪弁に変換するフィルタのことである。簡単なアルゴリズム (慣用表現の置き換えだけ) で作られているため、その表さえ書き換えてしまえば、任意の方言のフィルタを作ることが可能である。

osaka の特徴を以下にまとめた。

- 日本語文 (EUC コード) を大阪弁に翻訳する
- ソースは簡単なアルゴリズム (慣用表現の置き換えだけ) でつくられている
- ソースの表を書き換えるだけで任意の方言用フィルタがを作成できる

次に osaka の使い方を説明する。使用法は、例えばファイル hoge を大阪弁にして表示するだけであれば、kterm 上で、

```
% osaka < hoge
```

とする。すると、変換結果が kterm 上に表示される。その出力をたとえばファイル hoge.osk にする場合は、

```
% osaka < hoge > hoge.osk
```

のようにする。

ここからは、実際に osaka で文章を変換させるとどのようになるかを示す。変換させる文章は、「お国ことばを知る 方言の地図帳」という書籍に書いてある共通語による昔話「桃太郎」の一部分を使用した。「桃太郎」の文章を以下に示す。

むかし むかし あるところに
おじいさんと おばあさんが ありました。
おじいさんは 山へ しばかりに
おばあさんは 川へ せんたくに 行きました。
おばあさんが せんたくを していると
川上から 大きな ももが
どんぶらこ どんぶらこと 流れてきました。

この文章を osaka を使って変換させた結果を以下に示す。

むかし むかし あるんやトコに
おじいはんと おばあはんが おました。
おじいはんは 山へ しあほりに
おばあはんは 川へ せんたくに 行きたんや。
おばあはんが せんたくを しとると
川上から 大きな ももが
どんぶらこ どんぶらこと 流れてきたんや。

上の変換結果から、単なる単語や語尾などを置換するだけなので、文脈によらず特定の表現は、必ず同じ言葉に変換されることがわかる。そのため、例えばまだ漢字をあまり使用しない子供が、ひらがなで文章を書いてそれを変換させた場合、意味がわからない文章に変換されてしまうかもしれないということが考えられる。短所と長所を以下に示す。

短所 短所は、「あるところに」の部分が「あるんやトコに」となったり、「行きました」の部分が「行きたんや」になるなど、文によっては不自然な文に変換されてしまうことがあるということである。

長所 長所は、大阪に住んでいる人にとっては普段使っている言葉に近くなるため文章が理解しやすくなることである。また他県の人にとっては、大阪弁の簡単な勉強にもなるということである。

4 方言学習について

ここでは、方言学習の意義やコンピュータによる方言学習の方法を考えていく。

方言学習が重要と考える理由を以下にまとめた。

1. まだ方言を使う人 (特に高齢者) がかなりいるので、他地方や他県の出身者がそのような人と会話するときに意味がわからず、驚き戸惑ってしまう可能性がある
2. 若い人はその地域にいてもあまり方言を使わないため、方言の意味がよくわからない
3. 方言は文字としては、特別な本以外にはあまりでてこない
4. 使われなくなった方言はなくなってしまうかもしれない

以上の理由から方言学習をする意義があると考えられる。

1. のように他地方や他県の人の方言学習の方法について考えてみる。このような人達は、その場ですぐに見ることができる日常持ち運べるもの必要だと考えられる。例えば携帯電話に新潟弁を標準語に変換する辞書の仕組みを導入できれば便利だと考える。またその地域の方言による日常会話の辞書もあれば便利だと考えられる。

2. の場合を考える。その地域にいる若い人は幼い頃からその地域の人達の方言を聞いているため、リスニングはできる (言われればわかる) と考えられる。そのため日常持ち運べなくてもよいと考えられる。この場合は、標準語の言葉が新潟弁ではどうなるかを調べられる仕組みの他にアクセントやイントネーションも調べられる仕組み等が携帯電話等にあれば便利と考えられる。

3. は文字データとしてたくさん残すということが必要だと考える。

4. は除雪器具の方言など現在使われなくなった方言があるため、それらがなくならないように残すということが必要である。高齢者の方に直接聞いたり、古い文献を調査するなどして、どれが使われていて、どれが使われていないかを分類することが必要と考えられる。

コンピュータによる方言学習を簡単に考えてみることにする。以下に実際に存在するものなどをまとめる。

1. 辞書のようなもの
2. 文章を変換するもの
3. クイズのようなもの

1. の場合は標準語から新潟弁を調べる場合と新潟弁から標準語を調べる場合の 2 つの場合が考えられる。

長所 知りたい言葉をすぐに調べることができる。

短所 持ち運べる機器に現在その機能は実装されていないと思われる。

2. も標準語を新潟弁に変換する場合と新潟弁を標準語にする場合の 2 つの場合が考えられる。osaka や本稿で取り挙げている Web 上の変換サイトなどは前者である。

長所 長い文章でもすぐに変換してくれる。

短所 文章が不自然になることがある。

3. は問題が出題されて、その問題の答えを入力または選択するというものである。web 上のサイトにもそのようなサイトがある。

長所 子供は楽しく学習できるかもしれない。

短所 問題が同じだと単調でつまらなくなる。

5 web 上の新潟弁変換サイトについて

5.1 単語等の変換

Web 上の変換サイトで単語を変換させるとどのようになるのか調べてみることにした。Web 上の変換サイトの特徴を以下に示す。

- 標準語を新潟弁に変換する
- 単語でも長い文章でも変換できる
- HTML ソースを変換させることもできる
- 無料で利用することができる

変換させる単語は、「最新 ひと目でわかる 全国方言一覧辞典」という本の中に載っていたものを選んだ。また二つの短い文も変換させることにした。変換させる標準語の単語とその単語の新潟の方言を以下に示す。方言の表記は、「最新 ひと目でわかる 全国方言一覧辞典」に書いてあったカタカナによる発音表記とその単語の新潟の方言を以下に示す。方言の表記は、「最新 ひと目でわかる 全国方言一覧辞典」という本に書いてあったカタカナによる発音表記とした。

標準語	方言
君	ンナ
こんなに	コンゲ
大変に	バーカ
弟	オジ
なまける	ノメシコク
怠け者	ノメシ
指名する	カケル
寒い	サーブェ
つらら	カネッコリ
暑い	アッチェ
うろこ	コケラ
おしゃべり	シャベッチョ
かかと	アクト
かがむ	コゴム
かんたん	ジョーサネァ
きのこ	コケ
こんばんは	オバンデス
神社	オミヤ
すてる (捨てる)	ブチャル
ちょっと	チットバカ
しまった (失敗したときなどに発する感嘆の言葉)	アッキャー
招く	ヨブ
まっすぐ	マッツグ
わんぱくもの	キカンボ
はずかしい	ショーシー
ばんめし (晩飯)	ヨーハン
びっくりする	タマゲル
じゃがいも	ニドエモ
うそつき	ウソコキ
あかんぼう (赤ん坊)	ボ
かえる (蛙)	ゲァール
かえる (孵る)	ミヨケル

単語等を変換するために利用したサイトは、「全国方言コンバータ」、「方言変換道場」、「さるでもわかる 新潟弁辞典 音声解説付」、「新潟弁ナビ (~ Nandeyanen! ver2.0 ~)」の 4 つのサイトである。「新潟弁ナビ (~ Nandeyanen! ver2.0 ~)」は、見附市を中心に、長岡市 (旧長岡市域・旧栃尾市)・三条市の 3 市で広く使われている方言に変換するというサイトである。

各サイトに以下のように、変換させる単語等を貼り付けて変換させてみた。

> 君、こんなに、大変に、弟、なまける、怠け者、指名する、寒い、つらら、暑い、あつい、うろこ、おしゃべり、かかと、かがむ、かんたん、きのこ、こんばんは、神社、すてる、捨てる、ちょっと、しまった、招く、まっすぐ、わんぱくもの、はずかしい、ばんめし、びっくりする、じゃがいも、うそつき、赤ん坊、あかんぼう、かえる、蛙、卵からヒナがかえる。かえるを連れて家にかえる。

各サイトの変換結果を以下に示す。

1. 「全国方言コンバータ」の変換結果
   んな、こんげ、ば~か、おじ、のめしこく、のめし、かける、さーぶえ、かねっこり、あっちぇ、あっちぇ、うろこ、しゃべっちょ、あくと、こごむ、じょーさねぁ、こけ、おばんです、おみや、ぶちゃる、ぶちゃる、ちっとばか、あっきゃー、よぶ、まっつぐ、きかんぼ、しょーしー、よーはん、たまげる、にどえも、うそこき、ぼ、ぼ、げぁーる、げぁーる、卵からヒナがげぁーる。げぁーるを連れて家にげぁーる。

2. 「方言変換道場」の変換結果
   んな、こんげ、ば~か、おじ、のめしこく、のめし、かける、さーぶえ、かねっこり、あっちぇ、あっちぇ、うろこ、しゃべっちょ、あくと、こごむ、じょーさねぁ、こけ、おばんです、おみや、ぶちゃる、ぶちゃる、ちっとばか、あっきゃー、よぶ、まっつぐ、きかんぼ、しょーしー、よーはん、たまげる、にどえも、うそこき、ぼ、ぼ、みよける、げぁーる、卵からヒナがみよける。みよけるを連れて家にみよける。

3. 「さるでもわかる 新潟弁辞典 音声解説付」の変換結果
   君、こんがに、大変に、おじ、のめしこきる、怠け者、指名する、さぁべ、かねこおり、暑っちぇ、あつい、うろこ、おしゃべり、かかと、かがむ、じょーさない、こけ、こんばんは、神社、すてる、びちゃっる、ちーとばかし、あっささ、招く、まっすぐ、わんぱくもの、しょーしい、ばんめし、たんまげたする、にどイモ、てんぼこき、赤ん坊、あかんぼう、かえる、蛙、卵からヒナがかえるこてさ。かえること連れて家にかえるこてさ。

4. 「新潟弁ナビ (~ Nandeyanen! ver2.0 ~)」の変換結果
   君、こんげんに、ばかげた、弟、なまける、のめしこき、指名する、さぶい、つらら、あっちぇ、あつい、うろこ、さべっちょう、あくと、かがむ、かんたん、きのこ、なじらね、神社、ぶちゃる、ぶちゃる、ちっと、しまった、招く、まっすぐ、わんぱくもん、しょうしい、ばんめし、びっくりする、にどいも、うそつき、坊、あかんぼう、かえる、ぎゃく、卵からヒナがかえる。かえるをひっぱって家にかえる。

5.2 変換結果についての考察

変換の問題点を以下に示す。

- 変換しなくてもよいところも変換されてしまう可能性がある

  「なまけもの」は「怠け者」だけではなく動物の名前で「ナマケモノ」があるため、動物を意味する文を書いた場合、違う意味の方言に変換されてしまう。この問題を解消する方法として、カタカナによる「ナマケモノ」を正規化しておくということが考えられる。

- 名詞と動詞を判断することができない

  上のように、「かえるを連れて家にかえる」という文の場合、文として一番ありえるのは、「蛙をつれて家に帰る。」という意味だと考えられるが、変換させてみると、「げぁーるを連れて家にげぁーる。」となっている。また、他のサイトの変換では「みよけるを連れて家にみよける。」となっている。このように、現在の Web 上にある一般的な方言変換サイトでは、名詞と動詞を判断せずに変換しているため文が不自然になってしまう。

- 漢字を含む語を変換させる場合、たくさん登録しなければならない

  日本語には、正書法がないため、上の「怠け者」の場合は変換できたが、他に「怠けもの」、「なまけ者」、「なまけもの」の合わせて 4 通りの書き方ができるため、それらも変換できるようにしなければならない。しかし全ての語で同じようにするのは、非常に難しい。

5.3 文章変換

ここでは、Web 上の変換サイトで少し長い文を変換させるとどのようになるのかを調べてみることにする。変換に使用する文は、本稿 3.2 と同じ文を使用した。変換させるために利用したサイトは「さるでもわかる 新潟弁辞典 音声解説付」と「新潟弁ナビ (~ Nandeyanen! ver2.0 ~)」である。

変換結果は以下の通りである。

1. 「さるでもわかる 新潟弁辞典 音声解説付」の変換結果

   むかし むかし あるところに
   おじいさんと おばあさんが あったてば。

おじいさんは 山へ しばっかに
おばあさんは 川へ せんたくに 行きましたて。
おばあさんが せんたくこと しとると
川上から 大きな ももが
どんぶらこ どんぶらこと 流れてきましたて。

2. 「新潟弁ナビ (~ Nandeyanen! ver2.0 ~) 」の変換結果

むかし むかし あるところに
じさと ばさが ありましたて。
じさは 山へ しばかりに
ばさは 川へ せんたくに 行きたがーて。
ばさが せんたくを していると
川上から でっこい ももが
どんぶらこ どんぶらこと 流れてきたがーて。

上の 2 つの変換結果から、osaka と同じように特定の表現は必ず同じ言葉に変換されていることがわかる。

変換サイトのいいところは、どんな文章であっても「あったてば。」や「ありましたて。」のように新潟弁風の文章に変換できるということである。

変換サイトの問題点は、単語等を変換させたときと同じことであるが、変換の精度が不十分であるということである。この例文では、1. サイトの変換結果で「しばかり」の「ばかり」の部分が、範囲を限定する意を表す「ばかり」の新潟弁「ばっか」に誤変換されてしまっている。

5.4 実験と考察

ここでは、5.2 でわかった問題点を改善する方法を考察していく。まず実験で使用する ChaSen というソフトを説明する。

ChaSen

ChaSen(茶筌) は、奈良先端科学技術大学院大学の松本研究室が開発し、公開している日本語形態素解析システム。ChaSen は、単語の区切りを調べるのが難しい日本語を単語分割することができる。また分割した単語の品詞を詳細に表示することができる。UNIX,MS-Windows,Linux などで動作させられる。

chasen の使い方を説明する。ChaSen は基本的に以下のように使う。

1. ファイル (file) に日本語文を書いて、それを食わせる
   % chasen file
2. 標準入力から日本語文を食わせる
   % echo "私は大学生です。" | chasen
3. 直接端末から日本語入力する
   % chasen とした後、

   私は大学生です。

   のように入力する。終了は「Ctrl」キーと「D」を押す。

ここで、「-F オプション」について説明しようと思う。これは、各形態素の表示形式 (出力フォーマット) を指定するオプションである。「-F オプション」を使用するときは、

```
% echo "私は大学生です。" | chasen -F "format"
```

のように chasen の後ろに「-F "format" 」を追加する。format の部分には出力フォーマットを指定するための変換文字を入れる。最新版である『茶筌』version 2.3.3 で使用できる変換文字の一覧を以下に示す。なお一覧中には古いバージョンでは使用できない変換文字も含まれている。

変換文字	機能
%m	見出し (出現形)
%M	見出し (基本形)
%y, %y1	読みの第一候補 (出現形)
%Y, %Y1	読みの第一候補 (基本形)
%y0	読み全体 (出現形)
%Y0	読み全体 (基本形)
%a	発音の第一候補 (出現形)
%A	発音の第一候補 (基本形)
%a0	発音全体 (出現形)
%A0	発音全体 (基本形)
%rABC	ルビつきの見出し (“A 漢字 B かな C” と表示)(※ 1)
%i, %i1	付加情報の第一候補
%i0	付加情報全体
%Ic	付加情報 (空文字列か “NIL” なら文字 c)(※ 1)
%Pc	各階層の品詞を文字 c で区切った文字列
%Pnc	1 ~ n(n:1 ~ 9) 階層目までの品詞を文字 c で区切った文字列
%h	品詞の番号
%H	品詞文字列
%Hn	n(n:1 ~ 9) 階層目の品詞 (なければ最も深い階層)
%b	0(旧版との互換性のみ)
%BB	品詞細分類 (なければ品詞)
%Bc	品詞細分類 (なければ文字 c)(※ 1)
%t	活用型の番号
%Tc	活用型 (なければ文字 c)(※ 1)
%f	活用形の番号
%Fc	活用形 (なければ文字 c)(※ 1)
%c	形態素のコスト
%S	解析文全体
%pb	最適パスであれば “*”, そうでなければ “ ”
%pi	パスの番号
%ps	パスの形態素の開始位置
%pe	パスの形態素の終了位置 +1
%pc	パスのコスト
%ppiC	前に接続するパスの番号を文字 C で区切り列挙
%ppcC	前に接続するパスのコストを文字 C で区切り列挙
%?B/STR1/STR2/	品詞細分類があれば STR1，なければ STR2(※ 2)
%?I/STR1/STR2/	付加情報が “NIL” でも “”(空文字列) でもなければ STR1，そうでなければ STR2(※ 2)
%?T/STR1/STR2/	活用があれば STR1，なければ STR2(※ 2)
%?F/STR1/STR2/	%?T/STR1/STR2/と同じ

変換文字	機能
%?U/STR1/STR2/	未知語なら STR1\，そうでなければ STR2(※ 2)
%U/STR/	未知語なら”未知語”，そうでなければ STR(%?U/未知語/STR/と同じ)(※ 2)
%%	% そのもの
.	フィールド幅の指定
1-9	フィールド幅の指定
\n	改行文字
\t	タブ
\\	\ そのもの
\’	’ そのもの
\"	" そのもの

※ **1** ipadic では,「行く (いく/ゆく)」のように形態素が複数の読みを持つ場合，その読みを「{ イ/ユ } ク」のように，半角のブレースとスラッシュを使って表している．通常の読みの出力 (出力フォーマットの %y) では，その第一候補である「イク」が出力され，%y0 を使うと読み全体である「{ イ/ユ } ク」が出力される．

※ **1** A,B,C,c が空白文字の時は何も表示しない．

※ **2** ‘/’ には任意の文字が使える．また，括弧 “(){}[]<>” を用いることもできる．以下に例をあげる．

ChaSen を使用して文を解析すると結果はどのようになるのかを示す。次のような文を単語に分割できる分けることができる。

私は学校に行く。

これを、例えば 2. のような使い方で ChaSen に「-F オプション」付けて使った場合、解析結果は、以下のように表示される。

```
% echo 私は学校に行きます。| chasen -F "%m\t%H\n"
私 名詞
は 助詞
学校 名詞
に 助詞
行き 動詞
ます 接尾辞
。 特殊
```

EOS

このように文が単語に分割されて、しかもその分割された単語の品詞が何であるかも表示することができる。この ChaSen を使って、Web 上の変換サイトの問題点を解消する方法を考えていこうと思う。5.2 のところでも述べたが、現在 web 上にある方言変換ページで、次のような文を変換させる場合、

かえるを連れて家にかえる。

前の「かえる」は「蛙」を意味し、後ろの「かえる」は「帰る」を意味すると仮定する。これを変換した場合、一般的な Web 上の変換サイトでは次の二つのいずれかになってしまう。

げぁーるを連れて家にげぁーる。

または

みよけるを連れて家にみよける。

上の文では後ろの「かえる」が生き物の「蛙」を意味する方言「げぁーる」に変換されてしまっている。下の例文では、前の「かえる」が、動詞のふかするという意味の方言「みよける」に変換されている。このように、Web 上の変換サイトでは名詞と動詞を区別せずに変換しているため、文が不自然になってしまう。

そこで上の文を chasen を使って文を使って解析すると、

```
% echo かえるを連れて家にかえる。| chasen -F "%m%y%H\n"
かえるかえる名詞
をを助詞
連れてつれて動詞
家いえ名詞
にに助詞
かえるかえる動詞
。。特殊
EOS
```

上の場合の解析結果の見方を説明する。この場合は、「%m」、「%y」、「%H」の 3 つの変換文字があるので結果は、見出し (元の単語)、読み、品詞の順に表示される。「\n」が一番後ろについているために単語の解析結果ごとに改行される。

結果の 1 行目を説明する。一つ目の「かえる」は見出し (元の単語)、二つ目の「かえる」は読み、最後は「名詞」と品詞が表示されている。結果から「前の「かえる」と後ろの「かえる」の品詞が名詞と動詞に区別されているのがわかる。

この解析結果は ChaSen の古いバージョンである ChaSen-1.51 で解析したときの結果である。本研究室のコンピュータには、ChaSen-1.51 と ChaSen-2.3.3 が入っており、本来はバージョンの新しい方を使ったほうがよいのだが、ChaSen-2.3.3 で上の文を解析した場合、前の「かえる」を名詞ではなく動詞と判断してしまう。そのため本研究では、ChaSen-1.51 を使っていくことにした。しかし ChaSen を使用して文を解析するだけでなく、辞書構造を改良する必要がある。

辞書構造について考えていこうと思う。

Web 上の変換サイトの辞書構造では、たとえば、ひらがなの「かえる」であれば「げぁーる」または「みよける」に変換するように登録されているため、名詞や動詞の区別をせずに「かえる」という単語を全て「げぁーる」または「みよける」にしてしまう。これは名詞と動詞を分けて登録していないためと考えられる。

そこで、名詞と動詞の両方に「かえる」を登録しておく辞書構造である必要がある。

5.2 でも述べたように、Web 上の変換サイトでは、漢字を含む語は、たくさん辞書に登録しなければならないが、ChaSen の解析結果の読みの部分を利用するため、ひらがなだけを登録しておくだけでよい。

この ChaSen の出力を使って、どうすればできるのかを研究室の Perl のプログラム (test2.pl) を元に説明する。

研究室の Perl のプログラムは、あらかじめファイルとして変換させる文を書いておき、それをプログラムに食わせて実行させるというものである。

まずプログラム実行時に指定したファイルの文を読み込む。読み込むときに、行の先頭に「#」が付いている場合は、その行は読み込まない。

次に ChaSen を実行する。実行するときは-F オプションを付けて実行する。本プログラムでは-F オプションは、"%m %y %H\n"を使用している。「%m」は見出し (元の単語)、「%y] は (単語の) 読み、「%H」は (単語の) 品詞を表示させる変換文字である。「\n」は単語の解析結果ごとに改行させる変換文字である。ファイルから 1 行を読み込み、それを解析する。解析結果を分けて$org、$yomi、$hinshi という変数に入れる。$org は chasen の解析結果の見出し (元の単語) の部分入れる変数、$yomi\verb は読みの部分を入れる変数、$hinshi は品詞の部分を入れる変数である。

その次に単語の解析結果の品詞の部分を名詞、動詞、接尾辞の順で調べて、そこにその単語の読みと同じ単語が登録されている場合は、その単語の方言が s に追加され、なけれ

ば元の単語がs に追加される。あらかじめ Perl のソースにハッシュとして名詞、動詞、接尾辞にわけて、変換させる単語 (ひらがな) とその新潟弁を登録しておく。

最後に改行をして終了する。フローチャートを以下に示す。

そのプログラムによる結果を以下に示す。Perl ソースプログラムの名詞のハッシュのところに「かえる」を「げぁーる」となるように登録し、動詞のハッシュのところに「かえる」を「けーる」となるように登録しておくと、下のようになる。

```
% perl test2.pl < bun.memo
げぁーるを連れてうちにけーる。
```

- 結果の考察
  前の「かえる」と後ろの「かえる」が区別されて、「げぁーる」と「けーる」になっているので、名詞と動詞に区別されないという問題は解消されていることがわかる。しかし、この場合の後ろのひらがなの「かえる」は動詞の「帰る」と仮定したものであるので、本当に解消されたとは言い切れない。

そこでプログラムを変更せずに、プログラムに食わせるファイルの文を次のような文に変更して考えてみることにする。

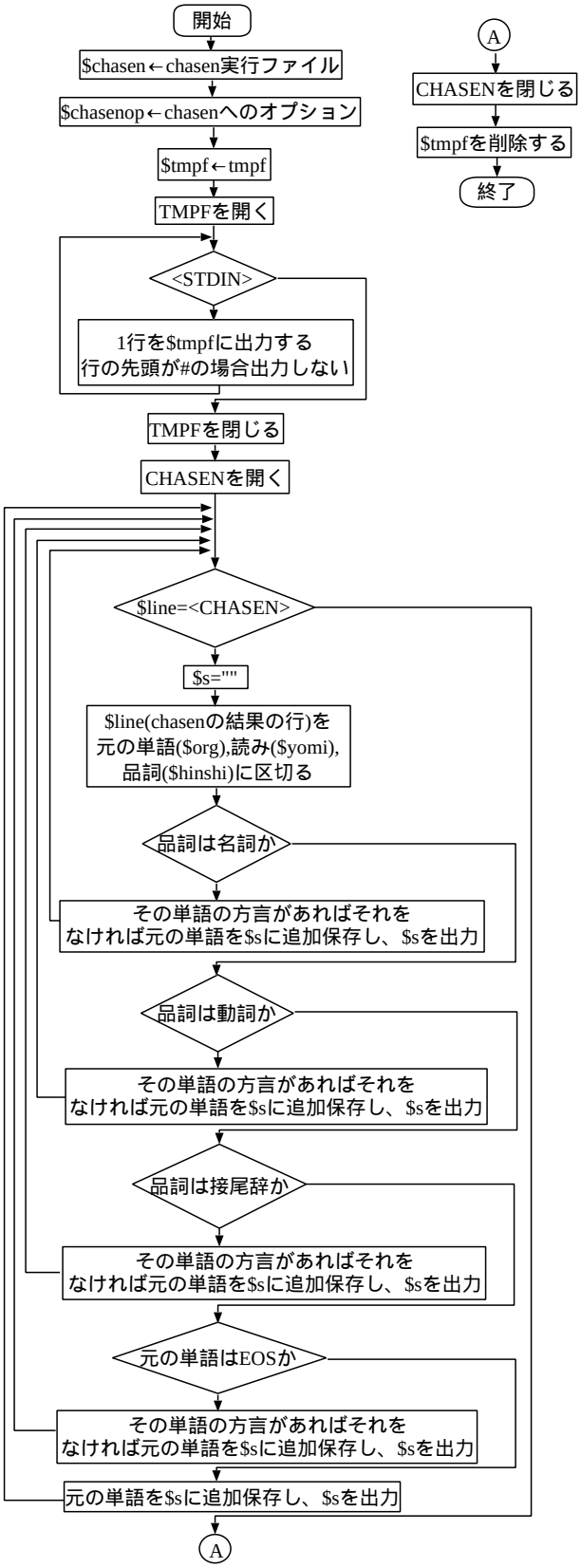
```
卵からヒナが孵る。
```

この文でプログラムを実行した場合の結果を以下に示す。

```
% perl test2.pl < bun.memo
卵からヒナがけーる。
```

- 結果の考察
  結果をみると、「孵る」が「帰る」と同じ方言「けーる」に変換されているのがわかる。しかし、新潟弁では、「孵る」には孵化するという意味の「みよける」という方言がある。その単語も登録しなければならないのだが、動詞の「かえる」はすでに「けーる」に変換させるように登録されているため、動詞の「かえる」のところに「みよける」を登録できないという問題がある。その他にも動詞の「かえる」は「返る」、「変える」、「換える」などがあるためその場合も同じようになってしまうと考えられる。

次に 5.2 の変換しなくてもよいところが変換されるという問題を解消するための方法を考えていきたいと思う。



%

Fig. 1 Perl のプログラムのフローチャート

たとえば、「あの動物はなまけものです。」という文は、Web 上の変換サイトやここまでの研究室のプログラムでは、

あの動物はのめしです。

または

あの動物はのめしこきです。

ように誤変換してしまうと考えられる。

そこで、変換しない言葉を登録する場所 (固有名詞) を作っておいて、そこにカタカナの「ナマケモノ」を入れて、「なまけもの」がカタカナの場合変換させないようにするということが考えられる。このように無変換にするという部分を追加したプログラムで以下のように

あの動物はナマケモノです。

と「なまけもの」のところをカタカナにした文で実行した場合、

```
% perl test2.pl < bun2.memo
あの動物はナマケモノです。
```

- 結果の考察
  「のめしこき」に変換されることはなくなる。しかし、動物の「ナマケモノ」は、普通ひらがな表記でもいいいためカタカナだけにする場合問題があるかもしれないため、これも本当に問題が解消されたとはいえないと考えられる。

6 まとめ

今回の研究では、標準語と共通語と方言、方言学習、日本語テキスト変換フィルタ、web 上の新潟弁変換サイト、ChaSen 等について調べ、web 上の変換フィルタの問題点を改善する方法を考察し実験をした。

標準語と共通語と方言について調べてみて、標準語というのは人為的につくられるきびしい規範で、共通語はゆるい規範で、現実のコミュニケーションであるということがわかった。方言とは、新潟弁の「なまけもの (人)」を意味する「のめし」や「のめしこき」

のような共通語と異なる語彙のことだけではなく、語彙・発音・語法を全て含めたものであるということがわかった。

方言学習については、重要性をいくつか挙げることで方言学習の意義を考えてみた。またコンピュータによる方言学習についても簡単にまとめてみた。

日本語テキスト変換フィルタは、本研究をしようと思うきっかけになった大阪弁変換フィルタである。どんな文章であっても大阪弁に変換してくれるため、お遊びと考えると非常に面白いソフトである。しかし、文章を自然な大阪弁の文章にしたい場合は、本研究の「桃太郎」の変換結果のように文章が不自然になるなど、問題があるということがわかった。

web 上の新潟弁変換サイトも osaka と同じように、お遊びと考えると非常に面白く、新潟弁を味わうことができるのだが、自然で正しい文章であるか見てみると、ひらがなの多い文章では、osaka のように不自然な文章になってしまうという問題等があることがわかった。

ChaSen は、文を単語に分けて、各単語の品詞なども表示できる優れたソフトであるということがわかった。またいろいろなオプションがあるということもわかった。本研究の実験では、ChaSen の解析結果により、文章中の 2 つの同じ単語が 2 つある場合の品詞を区別し、その結果を Perl のプログラムで利用した。

web 上のサイトの問題点を改善する方法の実験をやってみて、ChaSen を使用すれば一時的に解消できるということがわかった。しかし、今回のプログラムは、名詞「かえる (蛙)」と動詞の「かえる (帰る)」を区別して変換させただけであって、他の「孵る」、「返る」、「変える」、「換える」などもあるためプログラムを改善する必要がある。またカタカナの「ナマケモノ」を動物のことを指すとして、「のめしこき」に変換させないようにしたが、これも問題があると考えられる。5.4 で ChaSen を使用すると、漢字を含む語も、複数登録しなくてもよくなると書いたが、動詞の「帰る」の場合は、文章では「帰った」、「帰らない」というように使ったりもするため、今度は動詞を複数登録しなければならなくってしまう。

今回実験をしてみて、どの問題点も一時的に解消されただけであるということがわかった。そのため、より正確に文章等を方言に変換するには、さらにプログラムを改善する必要があると考えられる。

参考文献

[1] YAHOO!JAPAN 辞書 大辞林 三省堂

http://dic.yahoo.co.jp/dsearch?p=%E6%A8%99%E6%BA%96%E8%AA%9E&enc=UTF-8&stype=1&dtype=0&dname=0ss

[2] YAHOO!JAPAN 辞書 大辞林 三省堂

http://dic.yahoo.co.jp/dsearch?p=%E5%85%B1%E9%80%9A%E8%AA%9E&enc=UTF-8&stype=1&dtype=0&dname=0ss

[3] 方言 - Wikipedia

http://ja.wikipedia.org/wiki/%E6%96%B9%E8%A8%80

[4] 日本語の方言 - Wikipedia

http://ja.wikipedia.org/wiki/

[5] 佐藤 亮一: お国ことばを知る 方言の地図帳【新版】方言の読本 (小学館,2002)

[6] ぎふけんキッズ 岐阜県クイズ

http://www.pref.gifu.lg.jp/pref/kodomo/quiz/index.html

[7] SPACE ALC 日本一周 方言クイズ

http://www.alc.co.jp/jpn/teacher/hogen/nigata.html

[8] 江端義夫 加藤正信 本堂 寛: 最新 ひと目でわかる 全国方言一覧辞典 (学習研究社,1998)

[9] 全国方言コンバータ

http://www.netricoh.com/contents/variety/hougen

[10] 方言変換道場

http://www.mitene.or.jp/~hiro3/word.html

[11] さるでもわかる 新潟弁辞典 音声解説付

http://www2.icn.ne.jp/~sonmin/hougen/henkan.html

[12] 新潟弁ナビ (~ Nandeyanen! ver2.0 ~)

http://tutuga.hp.infoseek.co.jp/hogn/hognnmt.html