

# 形態素解析における関西弁の自動認識

廣川純也<sup>†1</sup> 深澤拓海<sup>†1</sup> 松村冬子<sup>†1</sup> 原田実<sup>†1</sup>

**概要:** 言語処理システムの基盤技術である形態素解析では既に高い精度が実現されている。しかし、方言やネットスラングなどの標準語から外れた日本語文を解析する際、それらの表現が形態素解析で用いる辞書に登録されていないため、正しく解析ができない場合がある。本研究では方言の中でも特に関西弁を含む日本語文の形態素解析の精度向上を目指し、形態素解析器 JUMAN に関西弁特有の語の表記や活用形、接続規則を追加することで、従来は未知語として処理されていた語の正しい解析を実現する。

**キーワード:** 形態素解析, 関西弁

## Morphological analysis of Kansai dialect

Junya Hirokawa<sup>†1</sup> Takumi Fukazawa<sup>†1</sup> Fuyuko Matsumura<sup>†1</sup> Minoru Harada<sup>†1</sup>

The outline: High accuracy has already been achieved in the morphological analysis of the standard Japanese sentences. However, it is not likely to be able to analyze correctly the Japanese sentences including dialect and net slang words, etc., because those words are not registered in the dictionary used in the morphological analysis. Aiming at the accuracy improvement of the morphological analysis of Kansai dialect, we added the iteration, conjugation, and juncture rule of the words which are peculiar to Kansai dialect and were previously recognized as unknown words to the dictionary of the morphological analysis system JUMAN.

**Key Words:** Morphological analysis, Kansai dialect

### 1. 背景・研究目的

ビッグデータの活用が活発になっている現在において、ソーシャルメディアは有力な情報源といえる。膨大なデータの中から有用なものをフィルタリングするためには、自然言語処理の技術は必須である。自然言語処理の基盤技術である形態素解析は、現状でも高い精度での解析を実現しているが、形態素辞書に登録されていない未知語は解析することができず、使用者の意図と異なる解析結果を出力する場合がある。このように解析できない未知語の中でも「おはよう」と小文字化したものや「すげえ」と母音変化するように標準語から外れているものを笹野、勝木ら[1],[2]は「崩れた日本語」と呼び、特に長音記号や撥音に対応する語を形態素解析用辞書に追加することで、その解析範囲を広げること成功している。

本研究では、ソーシャルメディアで多用されている方言のうち、特に「関西弁」に着目し、辞書の拡充による関西弁の形態素解析の実現を目指す。

### 2. 形態素解析の概略

本研究の提案手法を示すために、まず形態素解析の基本

的な処理の流れを以下に示す。

- 1 入力文に対し、形態素辞書を用いて各位置から始まる形態素の候補を全て検索する
- 2 全ての形態素の候補を列挙したラティス構造のグラフを作成する。
- 3 形態素同士の組み合わせの中で、コストが最も小さい形態素の並びを解析結果として決定する。

例えば、「旅行せえへん」という文が入力された場合、図1に示すラティス構造が作成され、幾つかの候補の中から最終的に赤線で示されている組み合わせを決定する。

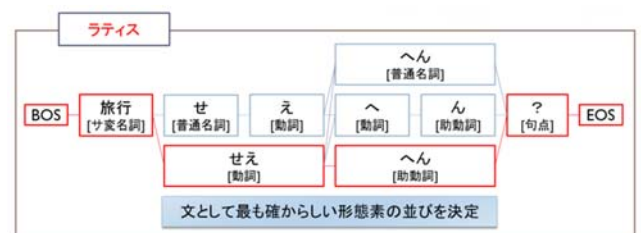


図1: ラティス構造の例

<sup>†1</sup> 青山学院大学理工学部情報テクノロジー学科

本研究では形態素解析器として JUMAN [3]を用いる。JUMAN では 3 の手順において形態素の並びを決定する際に形態素コストを計算し、最もコストが小さくなる並びを決定する。形態素コストは人手で設定した接続コストや単語生起コストを用いて決定される。この流れは MeCab などの機械学習を用いた形態素解析器においても同様である。

本研究では関西弁混じりの文章を認識するために関西弁特有の活用形や関西地方で用いられている言葉を形態素の候補として辞書に追加する。その際、これらの候補の形態素コストには対象語句を含む複数の文章を解析することで解析誤りが少なくなる値を設定した。

### 3. 関西弁解析のための JUMAN 辞書の拡充

#### 3.1 辞書拡充方法の概要

本研究では JUMAN 辞書の拡充により、関西弁を含む日本語文の形態素解析の精度向上を目指す。具体的には既存の形態素解析辞書に対し、関西弁特有の活用規則の追加、関西地方で使われている語で未知語に相当する形態素の追加、および追加された活用形の接続規則を追加することで関西弁の形態素解析を行う。本研究では、活用形や接続規則の追加においては標準語への悪い副作用を抑えるように配慮した。

本研究では関西弁の活用について「てらポチ&カリンの大阪弁講座」[4]を基準とする。ここで整理されている関西弁の活用形を、これらの活用形が標準語の活用形と競合して解析結果に大きな悪影響が出ない範囲でできるだけ多く登録した。同様に、コストに関しても人手で調整した値を既存のコストと類似している値を参考に、経験的に良い値を設定した。

#### 3.2 活用規則の追加

標準語と関西弁を比較した際、その活用型として新規に追加する必要があるものはなかった。一方、活用形は関西弁独自の変化があったため、本研究では活用形辞書 (JUMAN.katuyou) に関西弁用の活用形を新規に追加した。活用形とは動詞、形容詞などが活用して変化する種々の語形であり、未然形、意志形、連用形、基本形、連体形、仮定形 (条件形)、命令形がある。これらの活用形に関西弁に表れる活用変化を新規活用形としてその活用形が表れる活用型ごとに登録した。表 1 に新規に追加した 19 個の関西弁活用形と、それを追加した活用型の箇所を示す。なお挿入する箇所の活用型は表 1 下段にその略記の詳細を示した。

表 1: 新規活用規則の追加と追加した活用箇所

No	登録した新規活用形	追加した品詞箇所
1	大阪未然形特殊	サ変,カ変,ワ
2	大阪未然イ形	母
3	大阪未然エ形	母
4	大阪未然オ形	カ変
5	大阪未然ヤ形	母
6	京都未然形	サ変
7	京都未然イ形	サ変
8	大阪意思形	サ変
9	タ系大阪連用テ形	ワ,ワ文,イ形アウオ,イ形イ
10	タ系大阪連用ウテ形	ワ
11	タ系大阪条件形	ワ
12	タ系大阪条件ウ形	ワ
13	大阪命令形	母,カ,カ促,ガ,サ,タ,ナ,バ,マ,ラ,ワ,ワ文,カ変,サ変,ザ変,
14	大阪命令エ形	母
15	大阪サ変特殊命令形	サ変
16	大阪基本連用形	イ形アウオ,イ形イ
17	タ系大阪特殊連用アテ系	イ形アウオ,イ形イ,イ形イ特
18	タ系大阪特殊連用アテ系	イ形アウオ,イ形イ,イ形イ特
19	タ系大阪特殊連用アテ系	イ形アウオ,イ形イ,イ形イ特

略記	内容	略記	内容
母	母音動詞	ライ	子音動詞ラ行イ形
カ	子音動詞カ行	ワ	子音動詞ワ行
カ促	子音動詞カ行促音便形	ワ文	子音動詞ワ行文語音便形
ガ	子音動詞ガ行	カ変	カ行変格活用
サ	子音動詞サ行	サ行	サ行変格活用
タ	子音動詞タ行	ザ変	ザ変格活用
ナ	子音動詞ナ行	イ形イ	イ形容詞イ段
バ	子音動詞バ行	イ形イ特	イ形容詞イ段特殊
マ	子音動詞マ行	イ形アウオ	イ形容詞アウオ段
ラ	子音動詞ラ行		

関西弁動詞で最も追加が多かったのは命令形で、関西弁形容詞で最も追加が大きかったのは連用形だったため、追加した活用型箇所が多くなっている。以下に登録した新規活用形の説明と具体例を載せる。

#### 大阪未然形特殊

「する」などのサ行変格活用、「来る」などのカ行変格活用、「敵う」などの子音動詞ワ行に対して、否定を表す「〇〇へん」との接続を考慮し、活用形が「せえ」「けえ」「あ」など特殊に変化するものを大阪未然形特殊とした。

#### 大阪未然イ形

母音動詞の中でも、特に活用形がイ段に変化するもの

(特に「イイひん」と活用するもの)に対して、「〇〇ひん」の接続を考慮して大阪未然イ形とした。たとえば「着いひん」「見いひん」などが該当する。

#### 大阪未然エ形

母音動詞の中でも特に下一段活用する動詞に対して、「〇〇へん」の接続を考慮し、活用形がエ段に変化するもの(特に「エエへん」と活用するもの)を大阪未然エ形とした。たとえば「出えへん」「寝ねへん」などが該当する。

#### 大阪未然オ形

母音動詞の中でも特にカ行変格活用する動詞(「来(こ)おへん」など)に対して、「〇〇へん」の接続を考慮し、活用形がオ段に変化するもの(特に「オオへん」と活用するもの)を大阪未然オ形とした。「来る」という動詞に関しては上記の活用以外にも「来(け)えへん」「来(い)いひん」という未然形活用があるが、これらはそれぞれ大阪未然形特殊、大阪未然イ形に対応している。

#### 大阪未然ヤ形

母音動詞の中でも特に上一段活用や下一段活用する動詞に対して、「〇〇へん」の接続を考慮し、活用形が「ヤ」に変化するものを大阪未然ヤ形とした。たとえば「出やへん」「寝やへん」などが該当する。

#### 京都未然形

サ変動詞の中で、活用形が「シ」に変化するものを京都未然形とした。たとえば「勉強しん」などが該当する。

#### 京都未然イ形

京都未然形で対応できたもの以外で、「〇〇ひん」の接続を考え、活用形が「シイ」に変化するものを京都未然イ形とした。たとえば「勉強しいひん」などが該当する。

#### 大阪意志形

意志形の中でもサ変動詞の意志形「しょう」を解析するために作成した活用形である。「よ」が小文字のときのみ解析対象となる。たとえば「勉強しょう」などが該当する。

#### タ系大阪連用テ形

連用形の中でも活用形が「テ」に変化するものをタ系大阪連用テ形とした。この活用規則は追加した活用形の中で唯一動詞・形容詞どちらにも対応している。たとえば「思(おも)て」「赤かって」などが該当する。

#### タ系大阪連用ウテ形

タ系大阪連用テ形で対応できたもの以外で、活用形が「ウテ」に変化するものをタ系大阪連用ウテ形とした。たとえば「言(ゆ)うて」「買(こ)うて」などが該当する。

#### タ系大阪条件形

子音動詞ワ行の条件形は既存の条件形の活用「〇〇ったら」と違い、小文字の「つ」が省略されるものをタ系大阪条件形とした。たとえば「思(おも)たら」「違(ちご)たら」などが該当する。

#### タ系大阪条件ウ形

タ系大阪条件形で対応できたもの以外で、特に「言う」

の条件形「言(ゆ)うたら」に対応するために追加した。関西弁の動詞「言(ゆ)う」の活用語尾は「う」であり、大阪条件形に当てはめるために「ゆ+うたら」の接続を考慮しタ系大阪条件ウ形とした。

#### 大阪命令形

関西弁動詞の命令形は語幹の母音(原則イ段)を活用語尾にもつため、活用形が「イイ」に変化するものを大阪命令形とした。たとえば「売(い)い」「落(ち)い」などが該当する。

#### 大阪命令エ形

大阪命令形で対応できたもの以外で、活用形が「エエ」に変化するものを大阪命令エ形とした。母音動詞の中でも特に上一段活用や下一段活用する動詞に対応している。たとえば「食(べ)え」「考(か)ええ」が該当する。

#### 大阪サ変特殊命令形

サ行変格活用動詞の命令形「せえ」を解析するために作成した活用形である。たとえば「勉強せえ」が該当する。また同じくサ変変格活用動詞の命令形「しい」は大阪命令形に分類される。大阪未然形特殊と同じ表記となっているが、接続規則が異なるため新たに定義した。

#### 大阪基本連用形

関西弁形容詞の連用形は「〇〇い」の「い」を省略する。たとえば「偉い」の連用形は「偉-なる」であり、活用語尾が無活用になる。そのため、無活用(語幹なし)を大阪基本連用形として登録した。

#### タ系大阪特殊連用アテ形

大阪基本連用形で対応できたもの以外で、活用形が「アテ」に変化するものをタ系大阪特殊連用アテ形とした。たとえば「赤(あ)あて」「偉(偉)あて」などが該当する。

#### タ系大阪特殊連用イテ形

大阪基本連用形で対応できたもの以外で、活用形が「イテ」に変化するものをタ系大阪特殊連用イテ形とした。たとえば「美(美)しいて」「優(優)しいて」などが該当する。

#### タ系大阪特殊連用オテ形

大阪基本連用形で対応できたもの以外で、活用形が「オテ」に変化するものをタ系大阪特殊連用オテ形とした。たとえば「白(白)おて」「す(す)ごおて」などが該当する。

### 3.3 地方言葉の追加

その地方で使われている独特の言い回しや言葉を本研究では地方言葉と定義した。たとえば標準語の「いけない」と同義である大阪弁の「あかん」や「どら焼き」と同義である大阪弁の「三笠」などが該当する。これら地方言葉の多くは辞書に登録されていないため形態素解析において未知語として扱われることが多い。

そこで未知語として扱われる地方言葉を内容語辞書(ContentW.dic)に登録する。この際、あらかじめ内容語辞書に登録されていた語と同じ意味をもつ単語に関してはその見出し語に追加することで対応する。たとえば上記で挙

げた「どら焼き」と同義である「三笠」は図2のように登録した。

(名詞 (普通名詞 ((読み どら焼き)(見出し語 どら焼き  
 どら焼き みかさ 三笠)(意味情報 "代表表記:どら焼き/  
 どら焼き カテゴリ:人工物 ドメイン:料理・食事")))

図2: 内容語辞書に登録されている語と同義語の追加の例

また辞書に登録されていない語や同じ意味を持っている語が存在するが活用型が異なるものに関しては新語として登録した。例えば「ほかす」という語を登録する場合、図3のように追加する。語の意味は「捨てる」と同じであり、「捨てる」は既に内容語辞書には登録されている。しかし「ほかす」の活用型は子音動詞サ行であり「捨てる」の活用型と一致しない。そのため既に登録されている「捨てる」の定義を参照し、読みと見出し語と活用型を新たに設定する。以上の方法で同様の作業を行い、内容語辞書に計 123 個の新語を登録した。

(動詞 ((読み すてる)(見出し語 捨てる すてる)  
 (活用型 母音動詞)(意味情報 "代表表記:捨てる/すてる  
 付属動詞候補 (基本) 反義:動詞:捨う/ひろう"))

(動詞 ((読み ほかす)(見出し語 ほかす)(活用型 子音動  
 詞サ行)(意味情報 "代表表記:ほかす 付属動詞候補 (基  
 本) 同義:動詞:捨てる/すてる 反義:動詞:捨う/ひろう"))

図3: 内容語辞書に登録されている語と  
 活用型が異なる語の追加の例

### 3.4 接続規則の追加

活用形辞書に追加した活用形と内容語辞書に追加した語に接続規則を与えなければ形態素解析をすることは出来ない。JUMAN では接続規則辞書 (JUMAM.connect) に接続規則が記述されており、本研究では新たに活用形を追加する度に接続規則を定義した。

ここでは具体例として「ウチってあんま勉強せえへん」という文を考える。この文を関西弁の規則を適用する前のJUMANで解析すると図4のようになる。

ウチってあんま勉強せえへん  
 ウチ うち ウチ 名詞 6 普通名詞 1 \* 0 \* 0 "代表表記:内/うち 漢字読み:訓 カテゴリ:場所-機能;抽象物"  
 っ て っ て っ て 助詞 9 副助詞 2 \* 0 \* 0 NIL  
 あんま あんま あんま 副詞 8 \* 0 \* 0 \* 0 "自動獲得:テキスト"  
 勉強 べんきょう 勉強 名詞 6 サ変名詞 2 \* 0 \* 0 "代表表記:勉強/べんきょう カテゴリ:抽象物 ドメイン:教育・学習"  
 せえ せえ 名詞 6 普通名詞 1 \* 0 \* 0 "代表表記:瀬/せ 漢字読み:訓 カテゴリ:場所-自然"  
 @ せ せ せ 名詞 6 普通名詞 1 \* 0 \* 0 "代表表記:背/せ 漢字読み:訓 ~を~に構成語 カテゴリ:動物-部位; 場所-機能"  
 ええ える 動詞 動詞 2 \* 0 母音動詞 1 未然形 3 "代表表記:得る/える 反義:動詞:失う/うしなう"  
 へん へん へん 助動詞 5 \* 0 無活用型 26 基本形 2 NIL  
 EOS

図4: 既存の形態素解析

まず、この文を標準語で言い換えると「ウチ」は「私」、「あんま」は「あまり」、「せえへん」は「しない」に該当すると考えられるため、「私はあまり勉強しない」となる。

はじめに「ウチ」という言葉に「私」という意味を持たせるため、内容語辞書に登録を行う。既に内容語辞書には同じ意味をもつ「私」があるため、ここでは見出し語に追加することで対応した。

「せえへん」は標準語では「しない」に該当するため、その接続は「せえ+へん」だと考えられる。既存のJUMANではサ行変格活用動詞「する」が「せえ」と活用すると定義されていないため活用形辞書への登録を行う。具体的には「へん」と接続していることから「せえ」は未然形だと判断し、3.2 で述べたように新たな活用形として「大阪未然形特殊」を定義して図5のように登録をした。

「せえへん」はサ行変格活用動詞の大阪未然形特殊「せえ」と無活用型の助動詞「へん」が接続するため、接続規則辞書には大阪未然形特殊の後ろに助動詞「へん」が接続するように図6に示すような追加をした。

(サ変動詞  
 ((語幹 \* )  
 (基本形 する )  
 (未然形 さ )  
 (意志形 しよう )  
 (大阪未然形特殊 せえ) <--追加した活用形

図5: 大阪未然形特殊の追加

((((動詞 \* \* 大阪未然形特殊))  
 ((助動詞 \* 無活用型 \* へん))  
 )

図6: 接続規則辞書への新規規則追加

この規則を追加したことでサ行変格活用動詞の大阪未然形特殊の後に無活用型の助動詞の「へん」が接続できるようになり、図7に示すとおり入力文「ウチってあんま勉強せえへん」をJUMANが正しく解析できるようになった。

ウチってあんま勉強せえへん  
 ウチ うち ウチ 名詞 6 普通名詞 1 \* 0 \* 0 "代表表記:内/うち 漢字読み:訓 カテゴリ:場所-機能;抽象物"  
 @ ウチ うち ウチ 名詞 6 普通名詞 1 \* 0 \* 0 "代表表記:私/わたし 漢字読み:訓 カテゴリ:人"  
 っ て っ て っ て 助詞 9 副助詞 2 \* 0 \* 0 NIL  
 あんま あんま あんま 副詞 8 \* 0 \* 0 \* 0 "自動獲得:テキスト"  
 勉強 べんきょう 勉強 名詞 6 サ変名詞 2 \* 0 \* 0 "代表表記:勉強/べんきょう カテゴリ:抽象物 ドメイン:教育・学習"  
 せえ せえ する 動詞 2 \* 0 サ変動詞 16 大阪未然形特殊 9 "代表表記:する/する 付属動詞候補 (基本) 自他動詞:自:成る/なる"  
 へん へん へん 助動詞 5 \* 0 無活用型 26 基本形 2 NIL  
 EOS

図7: 本研究における辞書拡充後の形態素解析

同様に、接続規則辞書に新規の活用形や地方言葉ごとに計 72 の規則を新たに追加した。その際、新たな接続規則をつくる場合と既存の接続規則に追加する 2 通りの方法で追加を行っている。

新たな接続規則をつくる場合には、図 6 に示すように接続の前後の活用形をそれぞれ定義した。図 6 のような場合では全ての動詞の大阪未然形特殊の後ろには無活用型助動詞の「へん」が追加される接続規則を意味する。この方法では 11 個の規則を追加した。

既存の接続規則に追加する場合には、既に定義されている接続規則の中で類似する活用形を見つけ出しその前後に当該活用形を追加する。例えば「わろてんだ」という関西弁を解析するために接続規則を設定する場合を考える。既存の接続規則辞書には「どうなってんだろう」や「知ってんだ」といった、動詞の連用形の後ろに助動詞「んだ」が接続する規則が登録されている。「わろてんだ」は、「笑う」の関西弁である「わろう」と助動詞の「んだ」と接続した構文である。「わろて」は「わろう」がタ系大阪連用テ形に活用した語なので、タ系大阪連用テ形が助動詞「んだ」と接続するように接続規則辞書の図 8 が示す箇所にタ系大阪連用テ形を登録する。

```

((VerbBasicForm
 IAdjBasicForm
 (***) ダ列基本連体形)
 (***) ダ列タ形)
 (***) デアル列基本形)
 (***) デアル列タ形)
 (***) デス列基本形)
 (***) デス列タ形)
 (***) ヤ列基本形)
 (***) ヤ列タ形)
 (助動詞 * 無活用型 基本形)

(動詞 ** タ系連用テ形) ;06/07/16 「どうなってんだろう」
(動詞 ** タ系大阪連用テ形) ;15/10/01 「わろてんだ」
(接尾辞 動詞性接尾辞 * タ系連用テ形)
(形容詞 ** タ系大阪連用テ形)
(助動詞 * 助動詞ぬ型 タ系連用テ形))
 ((助動詞 * ナ形容詞 * んだ))
)
    
```

図 8: 既存の接続規則に拡張するように登録した規則

## 4. 評価実験

本研究の実験では、形態素解析器として JUMAN [2] を用いた。実験で用いるコーパスは毎日新聞 2000 年 1 月分の新聞記事から 9 万文、Heffernan らが作成した関西弁コーパス [5] から抽出した 1 万文を合わせた合計 10 万文とした。

### 4.1 関西弁認識の評価

本研究の提案する辞書拡充による関西弁の形態素解析の精度向上の有効性を確認するため、まず関西弁を含むテキストコーパスの作成を行った。具体的には、Heffernan Kevin らの関西弁コーパスの中で平仮名を 1 字以上含み、かつ全

表 2: 評価実験の Confusion Matrix

	関西弁だと解析	標準語だと解析
実際は関西弁	72	13
実際は標準語	14	1

表 3: 関西弁規則適用後の JUMAN による自動認識の精度

適合率	再現率	F 値
0.837	0.847	0.842

体が 20 文字以上で構成される文を正規表現で抽出し、1 万文の文集合を作成した。

提案する辞書拡充を適用した JUMAN と、適用しない既存の JUMAN の両方で上記のテキストコーパスを解析させ、その解析結果を比較することで評価する。両者の間で解析結果が異なる箇所を無作為に 100 箇所選び取り、新たな規則を適用した JUMAN の関西弁の識別能力の適合率と再現率を評価することで本研究の有効性を認識する。既存の JUMAN では関西弁に対応できる規則や新語が登録されていないため関西弁を含む文には全く対応できなかった。

関西弁規則適用後の JUMAN に対する実験結果は表 2 に示す通りになった。本研究では適合率とは関西弁だと解析した語が本当に関西弁である割合であり、再現率とは実際に関西弁であった語の中で関西弁と解析されたものの割合と定義する。関西弁規則適用後の JUMAN の実験結果は表 3 に示したように適合率は 83.7%、再現率は 84.7% となった。よって関西弁のテキストに対して既存の形態素解析よりも高い精度で認識したといえる。以下に正しく認識できた事例、誤って認識した事例を示す。

#### 正しく関西弁を認識できた事例

- ・あんまり観たりせえへん
- ・なんでかってゆうたら、やっぱ
- ・昨日買ったバックなんやけど
- ・すんごいおもんないと思う。
- ・それはほんまにすごい。

#### 誤って認識した事例

- ・よっしゃええやん
- ・やっぱな、純粹であまーい感じが (i)
- ・せやな。やっぱ女の子は
- ・どないしょ。これ全然だめだ
- ・あのころはよかった (ii)

誤って認識した事例のうち、(i)の文は「やっぱな」の「な」の部分形容詞「無し」の大阪基本連用形と認識されてしまい正しい解析ができていなかった。これは活用規則辞書

に大阪基本連用形を追加した事でそちらを優先してしまい、正しい品詞である終助詞「な」と接続できなかった為と考えられる。

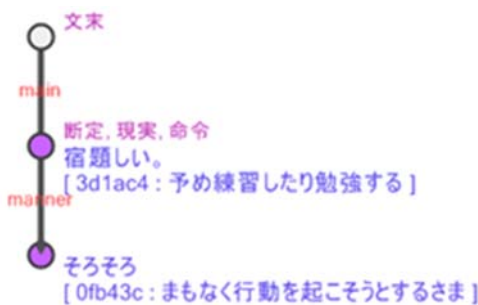
また、(ii) の文では「はよかった」の部分が形容詞「はやい」のタ形活用と解析されていた。これは関西弁において「早い」を意味する「はよい」を単語辞書に追加したことが原因だと考えられる。正しくは「は、よかった」と認識されることが自然だが、「はよい」を追加した事で「早かった(はよかった)」という活用が認められたため、形態素の数が少ないためにコストの小さい「早かった」が優先されてしまったと考えられる。

関西弁で使われる単語の追加や活用規則を追加した事で既存の形態素解析より高い精度で関西弁混じりの文を解析できることが認められた。一方で「あのころはよかった」など既存の JUMAN で認識できていた文が誤って認識された事例も認められた。そのため、大量の事例を元に規則の精選を行うことや単語のコストを設定する事でより高い精度での形態素解析が実現できると考えられる。

#### 4.2 関西弁の意味解析

原田研究室では形態素解析器 JUMAN を利用して日本語文における語の意味と、語と語の間の深層関係を正確に計算する意味解析システム SAGE を開発している。本研究により関西弁規則の追加を実現したことで意味解析システム SAGE も関西弁への対応が可能となった。

##### ①そろそろ宿題しい



##### ②そうするんちやう

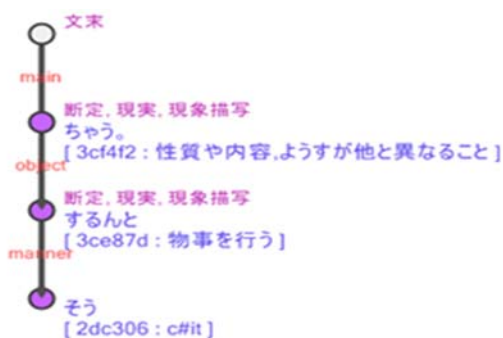


図9: 関西弁に対応した意味解析システム SAGE

図9の①は SAGE が入力文「そろそろ宿題しい」を意味解析した結果である。活用形辞書にサ行変格活用型の大阪命令形として「しい」を登録し、サ変名詞「宿題」の後ろに接続するように接続規則辞書に登録した事で JUMAN での解析が可能となった。SAGE でも関西弁の意味を解析することが可能となり、「しい」が命令を意味する動詞であることが判明したため「宿題しい」のモダリティに命令が追加されたことが分かる。

図9の②は SAGE が入力文「そうするんちやう」を意味解析した結果である。内容語辞書に子音動詞ワ行として「ちやう」を登録したことで JUMAN での解析が可能となった。ここで「ちやう」は新語として登録したため EDR 辞書にも「違う」と同じ意味として「ちやう」を登録した。この結果 SAGE も「ちやう」の正しい語意「性質や内容、ようすが他と異なること」を割り当てている。

#### 5. おわりに

本研究では、関西弁を対象として、形態素解析の辞書参照の際に関西弁特有の活用や表現を形態素候補として認識するための辞書拡充を行った。これにより既存の JUMAN よりも関西弁混じりの文に対して高い精度での形態素解析を実現した。

今後の課題としては、形容詞「ええ」などの既存の活用型に属さない単語の登録、「～はる」などの活用する敬語の登録を行う。

#### 6. 謝辞

本研究を進めるにあたり、日本語形態素解析システム JUMAN の VisualC++ のソースプログラムを提供くださった京都大学の河原大輔准教授、関西弁コーパスを提供くださった関西学院大学の Kevin Heffernan 教授、辞書編集および実験にあたりアドバイスを下さった東京工業大学の笹野遼平助教に深く感謝いたします。

#### 参考文献

- [1] 笹野遼平, 黒橋禎夫: 形態素解析における連濁および反復形オノマトペの自動認識, 言語処理学会, 第13回年次大会論文集, B4-05(2007.3)
- [2] 勝木健太, 笹野遼平, 河原大輔, 黒橋大輔: Web 上の多彩な言語バリエーションに対応した頑健な形態素解析, 言語処理学会, 第17年次大会 発表論文集(2011.3)
- [3] JUMAN version 7.01 : <http://nlp.ist.i.kyoto-u.ac.jp/index.php?cmd=read&page=JUMAN&alias%5B%5D=日本語形態素解析システム JUMAN>
- [4] うえぶこみ!友が来通信: <http://cameronjove.web.fc2.com/>
- [5] Kevin Heffernan: 関西弁コーパスの紹介, 総合政策研究, 関西学院大学総合政策学部研究会, No41, p.157-163(2012.7)

- [6] 小林聖也, 奥村紀之: 方言と標準語の違いを考慮した言語認識システムの開発, 人工知能学会全国大会第 23 回論文集, 1-4 (2009)
- [7] 利根川翔, 笥捷彦: 崩れた表記に対応する日本語形態素解析器の開発, 情報処理学会全国大会講演論文集, vol175, No2, pp.2115-2116(2013.3)
- [8] 颯々野学: JUMAN における形態素文法記述の問題点に関する考察, 情報処理学会第 53 回全国大会論文集, 第 2 分冊, 1L-7, pp.75-76(1996)
- [9] 利根川翔: 日本語形態素解析における崩れた表記への対応手法の提案と評価, 修士論文, 早稲田大学大学院 基幹理工学研究科 情報理工学専攻,(2012)
- [10] 小木曾智信, 市村太郎, 鴻野知暁: 近世口語資料の形態素解析の試み, 第 4 回コーパス日本語学ワークショップ 予稿集 (2013.9)
- [11] 黒橋禎夫, 河原大輔: 京都大学大学院 情報科学研究科 日本語形態素解析システム JUMAN version 7.0 マニュアル (2012)