自然言語処理 103−14
（1994. 9. 16）

口語的表現を含む日本語文の形態素解析の実現と評価

竹元義美　福島俊一

日本電気(株)
情報メディア研究所

日本語文章の口語的表現に対応した形態素解析手法を提案し、その評価結果を報告する。広い分野のテキスト処理を想定した場合に口語的表現の形態素解析は重要であるものの、その精度は十分とは言えなかった。本稿では、口語的表現を正しく形態素解析するために2つの手法を示す。1つは、口語特有の言い回しを辞書登録すること、もう1つは、通常は平仮名表記する語を意図的に片仮名表記するなど、表記を変えた強調表現を通常の表記に直して辞書検索することである(口語置換検索処理)。これらの手法を実現した結果、口語的表現を多く含むテキストの文節区切り精度が1.8%向上し、テキストのタイプによらず安定した高い精度を得ることができた。辞書登録では、話し言葉特有の語の登録によって、文節区切りに失敗していた話し言葉の88%を正しく解析できた。口語置換検索処理では、形態素解析に失敗していた意図的な片仮名表記の75%、強調表現で特殊文字を含む単語の79%を救済できた。

Implementation and evaluation of a morphological analysis method for colloquial Japanese text

Yoshikazu Takemoto　Toshikazu Fukushima

Information Technology Research Labs, NEC Corp.

This paper presents a new morphological analysis method for colloquial Japanese text, and describes its evaluation results. To enlarge application for natural language processing, it is necessary to deal with not only written language as before, but also colloquial language. This paper shows two techniques as the new method. One is to enter words peculiar to spoken language in dictionaries. The other is to replace words written in Katakana or special characters with usual writing and search through dictionaries for them. The two techniques can improve Bunsetsu-segmentation accuracy by 1.8% over a conventional method for text including colloquial expressions, and also accomplish stable accuracy for various types of text. The first technique can remove 88% of Bunsetsu-segmentation failures caused by spoken words. The second technique can remove 75% of failures caused by words written in Katakana expressly for emphasis, and 79% of failures caused by words written in special characters expressly for emphasis.

1 はじめに

自然言語処理技術は、ワープロ・機械翻訳・校正支援・情報検索・文字認識・音声認識など広く応用されている。ワープロは、専用機やフロントエンドプロセッサとして、オフィスから一般家庭にまで普及して一応の成功を収めることができた。しかしながら、ワープロ以外は、システムやアプリケーションとして製品化されているものの、高い実用性をもって普及しているとは言えないのが現状である[9]。

形態素解析[1][2][3]は、入力文章を単語辞書と照合することによって単語に認定するための処理である。上述した製品は、形態素解析を用いているものが多い。従来の形態素解析は、書き言葉を中心に研究されてきたため、口語的な文章への対応が十分に行われていないという問題があった[4][5][6]。しかし、自然言語処理システムの本格的な実用化のためには、入力文に対する頑健性が重要となる。したがって、形態素解析は口語的表現を含めた広範な文章を扱う必要がある。

口語的表現は、音声処理の場合とテキスト処理の場合とで次のような捉え方の違いがある。音声処理では、口語的表現として日常会話などの話し言葉における特有の言い回しにアクセントや発音を含めて処理する必要がある。また、助詞の省略など構文的な許容の広さも含めて検討されることが多い。一方、テキスト処理では、アクセント・発音は対象とせず、省略構文なども厳密に扱う必要はない。しかし、話し言葉特有の言い回しには対応しなければならない。さらに、音声処理では対象としないテキスト特有の表記による強調表現も扱わなくてはならない。ここでは、テキスト処理の観点で口語的表現を取り上げる[6][7]。

本稿では、口語的表現を含む日本語文の形態素解析手法を提案し、とその評価結果を述べる。以下では、2章で口語的表現とその形態素解析における問題点を示す。3章で口語的表現の形態素解析手法について述べ、4章でその手法の評価と考察について述べる。5章でまとめと今後の課題について述べる。

2 口語的表現が引き起こす問題点

口語的表現に対する処理を特別に行っていない形態素解析を、口語的表現が頻出するテキストを対象に実行し、口語的表現が引き起こした誤り箇所を分析した。分析には、週刊誌テキスト4冊分(432,917文字)から特に口語調エッセイや対話文を取り出したテキスト(15,881文字)を用いた。解析誤りとして、文節区切り誤りに着目した。解析誤りは、人手で作成した文節区切り正解と形態素解析による文節区切り結果を比較して抽出した。文節区切り精度は図1のようになり、文節区切り失敗は口語的表現によるものが大半であった。

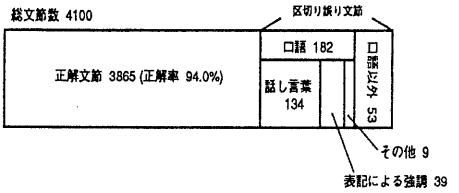

図 1: 従来の形態素解析による文節区切り結果

文節区切りに失敗した口語的表現を調査して以下に示す3項目に分類した。

1. 話し言葉に特有な言い回し
   - 音の簡便化 (例)"困っ ちゃう"
   - 特有な語の接続 (例)"なるほど ねえ"
2. 表記による強調表現
   - 意図的な片仮名表記 (例) "ガンバッ て"
   - 特殊な文字の挿入・追加
     (例) "だ ぁ～ いすき"
3. その他
   - 方言・古語
     (例) "してもうて","ござりまする"
   - 擬音語・擬態語
     (例) "ぐげげ～っ","どっしぇー"

項目1の話し言葉特有の言い回しとして、まず、音の簡便化が起こりやすい。例えば、書き言葉における"困っ てしまう"が話し言葉では"困っ ちゃう"に変化する。また、"なるほど ねえ"のように話し言葉に特有の語が接続しやすい。これらの例に対して話し言葉特有の言い回しに対応していない形態素解析を実行すると、解析に失敗して文節区切り誤りを引き起こす。

項目2の表記による強調表現とは、意図的に片仮名を用いるものや特殊文字("～","ぁ" など)を用いるものである。例えば、通常"がんばって"と表記するところを著者が意図的に"ガンバッて"と片仮名表記したとする。通常は辞書に"ガンバ(る)"は登録されていないので、"ガンバッ(未知語)て(名詞)"と解析されて"ガンバッ/て"のように文節区切り誤りを起こしてしまう。また、通常の表記"だいすき"の中に特殊文字を挿入して"だ ぁ～ いすき"と表記したとする。このとき"だ/ぁ/～/いす/き"のように特殊文字前後で単語分割に失敗することが多い。

また、表記による強調表現に関しては、次のような単語レベルでの問題もある。例えば、通常"ばつぐん"と表記するところを著者が意図的に"バツグン"と片仮名表記したとする。通常は辞書に"バツグン"は登録さ

	登録語	書き言葉との対応	例
助動詞相当語	ちゃう	接続助詞「て」＋補助動詞「しまう」	困っちゃう
	ちまう	接続助詞「て」＋補助動詞「しまう」	捨てちまう
	じゃう	接続助詞「で」＋補助動詞「しまう」	悩んじゃう
	じまう	接続助詞「で」＋補助動詞「しまう」	死んじまった
	てる	接続助詞「て」＋補助動詞「いる」	来てる
	でる	接続助詞「で」＋補助動詞「いる」	泳いでる
	じゃ	断定助動詞「だ」	冗談じゃない
助詞相当語	ちゃ	接続助詞「て」＋係助詞「は」	急がなくちゃ
	じゃ	格助詞「で」＋係助詞「は」	彼じゃ無理だ
		形容動詞連用形「で」	穏やかじゃない
	なきゃ	打ち消し助動詞「ない」仮定形＋接続助詞「ば」	行かなきゃ
	ねえ	終助詞「ね」	なるほどねえ

図 2: 登録した話し言葉の例

されていないから未知語となる。未知語処理で未知語は名詞と認定されてしまうことが多い。片仮名列をまとめて名詞とするような未知語処理は、文節区切りには成功する場合があるものの、例えば、“コンピューター”のような誤字を含む単語を判別する能力を持たない。“バツグン”のような単語は、文の理解の観点でも未知語とするよりは辞書中の単語と対応をつけたい。

項目3については件数が少なかったため、本稿では項目1,2について対処した。3節では項目1,2の対処方法を述べる。

3 口語的表現の形態素解析手法

2節で分類した口語的表現のうち、話し言葉特有の言い回しについては、従来行われているように[4]、辞書登録を行った。表記による強調表現については、新しい手法として、口語置換検索処理を提案する。

3.1 話し言葉特有の言い回しへの対処

週刊誌テキスト(2節で示した約43.3万文字)を対象とした文節区切り誤りの分析に基づいて話し言葉に特有な言い回しを整理した。そして、それらの話し言葉に特有な言い回しを竹下ら[4]と同様に助動詞相当語あるいは助詞相当語として辞書登録した。

例えば、“困っちゃう”の“ちゃう”は、“困ってしまう”の“て(接続助詞)しまう(補助動詞)”に相当する助動詞として辞書登録する。また、“急がなくちゃ”の“ちゃ”は、“急がなくては”の“て(接続助詞)は(係助詞)”に相当する助詞として辞書登録する。

登録した単語は、活用形を含めて151件である。登録した単語の例を図2に示す。

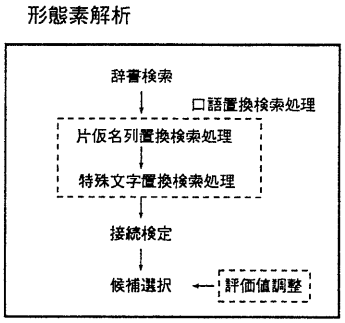

図 3: 口語置換検索処理を導入した形態素解析の流れ

3.2 表記による強調表現への対処

意図的な片仮名表記や特殊文字を用いた表記による強調表現はバリエーションが多様である。したがって、3.1節のような辞書登録で対処しようとするのは、登録しなければならない表記パターンが多すぎるため現実的でない。

そこで、形態素解析に3.2.1,3.2.2節で説明するような片仮名列置換検索処理と特殊文字置換検索処理とを導入する(以下、あわせて口語置換検索処理と呼ぶ)。

従来の形態素解析には、句読点などで区切られた区間ごとに一括して辞書検索を行い、単語の候補をすべて辞書から取り出してしまってから、接続検定・候補選択を行うような流れを想定している。そして、口語置換検索処理は、辞書検索処理の直後に組み込む。すなわち、図3に示すように、辞書検索→「片仮名列置換検索」→「特殊文字置換検索」→接続検定→候補選択という流れとする。口語置換検索処理は、辞書検索結果としての単語の候補を追加するように働く。そこで、評価値を調整して候補選択時の悪影響を抑える。

3.2.1 片仮名列置換検索処理

片仮名列置換検索処理は、辞書から単語が検索されていない片仮名列を平仮名列に置換して、辞書を再検索する処理である。以下に処理の適用条件・適用範囲などの詳細を述べる。

1. 置換範囲は、片仮名先頭から最長一致で単語をつないで4文字以下の片仮名単語しかとれていない部分とする。
2. 再検索では、置換範囲より前方の字種境界位置(平仮名から漢字や片仮名に変化した位置または句読点の直後など)から置換範囲の末尾までの各文字位置を先頭とする単語を検索する。検索された単語の末尾位置は置換範囲を越えてもよい。

3. 置換範囲の前方より再検索した場合、置換範囲に届かない単語は無効とする。

従来処理で検索された片仮名語でも短いものは信頼性が低い(例えば、“ゲンコツ”は“ゲン(固有名詞)コツ(名詞)”のように2文字の検索単語の組み合わせとして解析されてしまう)。そこで項目1では、4文字以下の単語が検索されていても再検索するようにした(この値にした理由は4.3.4節に示す)。項目2は、片仮名列が漢字・平仮名混じりの単語の一部になっている場合(例:“捨てゼリフ”が“捨てぜりふ”として再検索される場合)に備える。項目3は、置換範囲に届かない単語は通常検索で既に得られているので、二重に登録するのを防ぐためのものである。

3.2.2 特殊文字置換検索処理

特殊文字置換検索処理は、特殊文字を削除または置換して辞書を再検索する処理である。以下に処理の詳細を述べる。

1. “ぁ”,“ぃ”,…,“ぉ”,平仮名列に挿入・追加された“ー”, “〜”,文末の“っ”,“ッ”を特殊文字として扱う。
2. 特殊文字を削除して辞書を再検索する。
3. 直前の文字との音韻的規則に基づいて特殊文字を置換して再検索する。音韻的規則を以下に示す。
   - あ段の文字の直後の“ー”,“〜”は、“あ”に置換する。
   - い段の文字の直後の“ー”,“〜”は、“い”に置換する。
   - う段の文字の直後の“ー”,“〜”は、“う”に置換する。
   - え段の文字の直後の“ー”,“〜”は、“い”および“え”に置換する。
   - お段の文字の直後の“ー”,“〜”は、“う”および“お”に置換する。
4. 再検索は、削除・置換位置より前方の字種境界位置から削除・置換位置までの各文字位置を先頭とする単語を検索する。
5. 4で、削除・置換位置をまたがる単語、および、削除位置を末尾とする単語のみを有効とする。その際、削除位置を末尾とした単語については、直後についた特殊文字をその単語に含めて扱う。

項目2によって、例えば“ずーっと”を“ずっと”として辞書を再検索する。項目3によって、例えば“どーする”を“どうする”および“どおする”として、“おーきい”は“おうきい”および“おおきい”として辞書を再検索する。項目5によって、例えば“悪い〜っ!”の下線部分の特殊文字は直前の単語“い”に含めてしまう。

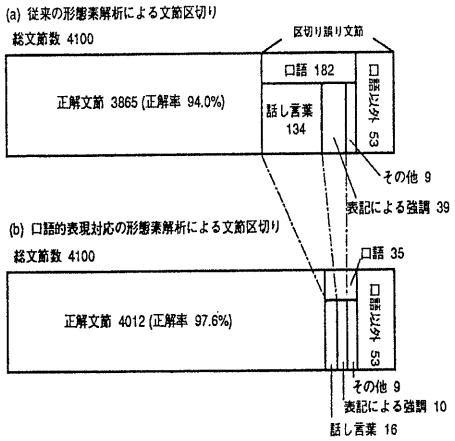

図 4: 文節区切り精度比較

3.2.3 評価値調整

口語置換検索処理は、辞書から検索される単語の候補を増やすことを目的とする。そのため、片仮名列置換検索処理を用いて検索された単語はすべて意図的に片仮名表記されているものという保証はない。例えば、入力文“ホンダの”に対して、“ホンダ(固有名詞)”と“の(格助詞)”の従来検索結果に、“ほん(名詞)”と“だの(並立助詞)”という単語候補を加えてしまう。

このように、口語置換検索処理結果の単語を含む候補が誤った解析結果となってしまうという文節候補選択時への悪影響が生じる。そこで、このような悪影響を抑えるために、口語置換検索処理で検索された単語は、従来処理で検索された単語より評価値を悪くするようにした。

4 評価

4.1節では、3節の辞書登録および口語置換検索処理の導入により口語的表現に対応した形態素解析を総合的に評価した。4.2,4.3節では、それぞれ辞書登録・口語置換検索処理の単独効果を評価した。

4.1 口語的表現対応の形態素解析の評価

まず、2節で分析に用いた小規模なテキストについて文節区切り誤りを調査した。図1の変化を図4に示す。文節区切りに失敗していた話し言葉特有の言い回しを含む134文節については、118文節(88%)を救済でき、誤りが16文節にまで減少した。文節区切りに失敗していた表記による強調表現を含む39文節については、29文節(74%)を救済でき、誤りが10文節にまで減少し

表 1: 文節区切り正解率比較

	A	B	C	D	E	B,D,E	A-E
総文節数	10,715	7,764	35,608	7,156	12,713	27,633	73,956
正解文節数(従来)	10,387	7,384	34,635	6,967	12,042	26,393	71,415
正解率(%)	96.9	95.1	97.3	97.4	94.7	95.5	96.6
正解文節数(口語的表現対応)	10,453	7,510	34,717	6,996	12,381	26,887	72,057
正解率(%)	97.6	96.7	97.5	97.8	97.4	97.3	97.4
正解文節数(辞書登録のみ)	10,452	7,475	34,767	7,009	12,354	26,838	72,057
正解率(%)	97.5	96.3	97.6	97.9	97.2	97.1	97.4

た。文節区切り精度は、全体として3.6%向上した。

次に、より多量のテキストを対象とした評価を行うために、前述の週刊誌テキストのうち3冊分(262,564文字)を対象として文節区切り精度を測定した。評価対象テキストを人手で以下のような5つのタイプに分類した。そして、従来の形態素解析と口語的表現に対応した形態素解析とによる文節区切り精度を調べた。結果を表1に示す。

A. 小説文・通常のエッセイ
B. 口語調エッセイ
C. 通常のレポート
D. 口語調レポート
E. 対話文・会話文

表1によれば、従来の形態素解析では、口語的表現を多く含むテキスト(特にB,E)の文節区切り精度は他のタイプのテキストに比べて低かった。しかし、口語的表現対応の形態素解析によって、テキストのタイプによらないほぼ安定した高い正解率を得ることができた。

口語的表現対応の形態素解析によって、特に口語的表現を多く含むテキスト(B,D,Eのようなタイプ)の文節区切り精度が従来の形態素解析より1.8%向上し、テキスト全タイプの文節区切り精度は0.8%向上した。

4.2 辞書登録の評価

4.2.1 辞書登録による単独効果

3.1節の辞書登録だけを行った形態素解析で文節区切りの精度を調べた(表1)。表1で従来の形態素解析の文節区切り精度を比べると、辞書登録だけでB,D,Eのタイプのテキストの文節区切り精度は1.6%向上し、テキスト全タイプの文節区切り精度は0.8%向上した。

4.2.2 対処できなかった例

図4で対処できなかった例(16文節、11件)を示す。"／"は文節区切りを表す。

1. 接続表の修正が必要なもの(1件)
   - 座骨／はだ(名詞)な、
2. 助詞などの省略表現(3件)
   - 電話 か(副助詞)／かって(副詞)
   - で(接続詞)／しょう(動詞)？ (→文頭に来たもの)　他 1件
3. 辞書登録が必要なもの(7件)
   - 優等生じゃ ね／え んだよ。(2件)
   - たまん／ない ね。　他4件

項目1は、"は(格助詞)"＋"だ(断定助動詞)"と"だ(断定助動詞)"＋"な(終助詞)"の接続表を"接続可"に変更すればうまくいく。しかし、この変更のために従来解析に悪影響も生じる。悪影響が多かったので今回は接続表の変更を行わなかった。項目2は、助詞や主語が省略されたために文節区切りに失敗した例である。話し言葉にありがちな現象なので対策を考えていきたい。項目3は、地道に辞書登録していくしかない。上の例では、下線の単語を登録することで対処できる。

4.3 口語置換検索処理の評価

4.3.1 口語置換検索処理の文節区切り

表1によると、テキストのタイプによっては辞書登録のみの方が口語的表現対応の形態素解析よりも文節区切り精度がよいものがある。また、全体としても辞書登録だけで文節区切りの精度は、辞書登録だけで口語的表現対応の形態素解析と同等のものが得られることがわかる。これは言い換えると、文節区切りの観点では、3.2節の口語置換検索処理の効果は小さかったことになる。

より詳細に分析するため、辞書登録のみを行った形態素解析と口語的表現対応の形態素解析とで解析結果が変化した箇所について、文節区切りとの関係を調べた。以下に、(辞書登録のみを行った形態素解析結果)→(口語的表現対応の形態素解析結果)の形式で示す。

A. 片仮名列

(a) 文節区切りが向上するもの

- ゲン(固有名詞)／コツ(名詞) → ゲンコツ(名詞)
- ガンバッ(未知語)／て(名詞) → ガンバ(動詞語幹)ッ(促音便)て(接続助詞)

(b) 文節区切りは変化しないもの

- バツグン(未知語)。 → バツグン(形容動詞)。
- ニラ(未知語)んで → ニラ(名詞)んで

(c) 文節区切りが悪くなるもの

- オバタリアン(未知語) → オバ(名詞)／タリア(固有名詞)ン(準体助詞)

B. 特殊文字

(a) 文節区切りが向上するもの

- ど(未知語)／ー(記号)／する。 → どーする(サ変動詞)。
- う(未知語)／ー(記号)／む(未知語)。 → うむ(五段動詞)。

B.特殊文字を含む部分で、従来解析と比べて、(b)文節区切りは変化しないもの、(c)文節区切りが悪くなるものはなかった。

上に示すように、口語置換検索処理の導入によって、文節区切り精度が向上する例もあれば、そうでない例もある。例えば、A.(b)で"バツグン"という単語は、従来は未知語と解析されていたのが、辞書中の単語と対応づけられて正しく解析されている。しかしながら、文節区切りは改善されていない。

また、例えば、A.(c)で"オバタリアン"という単語は辞書に未登録なため、従来解析では未知語となってしまう。これも従来解析では片仮名未知語処理によって未知語はまとめて名詞と同等に扱われるので文節区切りには失敗しない。一方、口語置換検索導入後の形態素解析では、片仮名列置換検索処理で得られた"おば(名詞)"および"ん(準体助詞)"と、従来検索で得られた"タリア(固有名詞)"とを解析結果として選んでしまい、"オバ／タリアン"のように文節区切りに失敗してしまう。しかし、2節で述べた未知語を誤りと捉える考えから、いずれも単語レベルでは解析誤りである。

B.(a)の"うーむ"は従来解析では未知語だった。口語置換検索処理の導入によって"うむ(動詞)"の解析結果を得る。これも文節区切りは改善されるといえども、解析結果は誤りである。

そこで次節では口語置換検索処理に関して単語レベルでの評価を行った。

表 2: 片仮名列置換検索処理の評価結果

	A.検索語				B.未知語	
置換検索導入前	正		誤		誤	
	325		13(11)		96(60)	
置換検索導入後	正	誤	正	誤	正	誤
	325	0	5(5)	8(6)	48(48)	48(12)

表 3: 特殊文字置換検索処理の評価結果

	A.検索語				B.未知語	
置換検索導入前	正		誤		誤	
	0		0		27(24)	
置換検索導入後	正	誤	正	誤	正	誤
	0	0	0	0	19(19)	8(5)

4.3.2 口語置換検索処理による単独効果

週刊誌テキストから片仮名あるいは特殊文字を含む箇所(句読点で区切られた範囲)を取り出して評価の対象とした。口語置換検索処理の導入前後の解析結果を比較・分析した。片仮名列は2節で分析に用いたテキストから434件、特殊文字は4.1節のB,D,Eのタイプのテキスト97,892 文字から27件を抽出した。片仮名列は、片仮名文字が始まって初めて片仮名以外の字種に変わるまでを1件として数える(例："スミマセン"で1件)。特殊文字は、特殊文字を含む単語を1件として数える(例："だぁーいすき"で1件)。

その結果を表2,3に示す。各表では、A.検索された単語またはその組み合わせでカバーされる箇所と、B.未知語が含まれる箇所とを分けた。また、2節で述べた考えから、Bは誤りとみなしている。()内の数字は、片仮名や特殊文字を含む箇所のうち、表記による強調表現の件数を示す。なお、強調表現でない特殊文字は、擬音語・擬態語に含まれるものである。

片仮名列置換検索の導入によって、以前は意図的な片仮名表記のために解析に失敗していた箇所71件のうちの53件(75%)が正しく解析できた。特殊文字置換検索の導入では、以前は特殊文字による強調表記のために解析に失敗していた箇所24件のうちの19件(79%)が正しく解析できた。

しかし、表記による強調表現にもかかわらず改善されなかった箇所23件(片仮名：18件、特殊文字：5件)が残った。これらについて次節で述べる。なお、以前は正しく解析されていた片仮名列が、片仮名列置換検索の導入によって解析に失敗する副作用はなかった。

4.3.3 口語置換検索処理で改善されなかった表記による強調表現

4.3.2節で表記による強調表現にもかかわらず改善されなかった23件の原因分類を示す。

A. 評価値調整が必要なもの(8件)

- ニラ(名詞)んで
- そーなん(サ変動詞)です。　　他 6件

B. 片仮名を平仮名に置換した文字列に特殊文字置換検索処理が必要なもの(5件)

- ビョーキ
- ズーッと　　他 3件

C. 平仮名以外の文字の直後に"～","ー"を持つもの(2件)

- 甘～い
- 鬼は外～っ！

D. 辞書登録が必要なもの(8件)

- キヨハラ
- スンマセン　　他 6件

Aは、単語候補はとれているが、解析結果が誤っているものである。より大きなテキストを評価して評価値の与え方をさらに検討する必要がある。Bは、例えば"ビョーキ"を"びょーき"に置換した結果に対して特殊文字置換検索処理を適用("びょうき"として辞書を再検索)するなどを検討する必要がある。Cは、今回導入した口語置換検索処理では漢字の直後の"～"は特殊文字とみなしていないので改善されなかった。しかし、例えば"夜９時～９時５４分"は、記号として扱いたい。平仮名文字以外でどのような文字の直後の"～","ー"は特殊文字とみなすべきか調べる必要がある。

4.3.4 従来解析への影響

本節では、口語置換検索処理の導入による従来解析の解析結果や処理効率への影響を考察する。

片仮名列置換検索処理で、3.2.1節の適用条件1中の下線部分の値は、これより大きくすると、より多くの片仮名列を処理対象とするため追加される単語の候補数が大きくなり、解析結果や処理効率に悪影響を与える可能性がある。一方で、小さくすると処理の適用洩れが生じる可能性がある。

まず、片仮名列置換検索処理で適用条件の値を「n文字以下」としてn=1,2,3...と変化させた形態素解析と従来の形態素解析との解析結果の差分を分析した。分析に用いたのはA～Eのタイプをすべて含むテキスト(週刊誌1冊分、82,235文字)である。図5に結果を示す。

図5に示すように、全片仮名列1,854件のうち、意図的片仮名表記は、216件であった。○印は、改善件数を

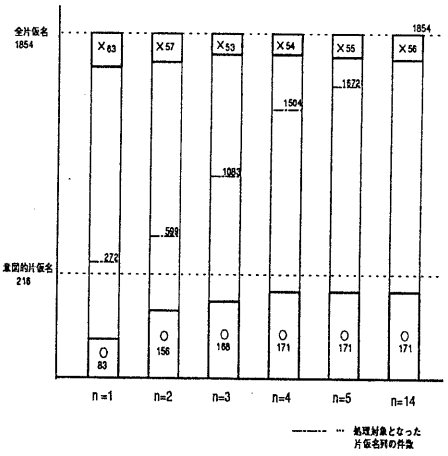

図 5: 適用条件1を変化させた場合の解析結果の変化

表す。これは、解析に失敗していた意図的な片仮名列が正しく解析されるようになるものである。テキスト中、片仮名単語の最大長は14だったので、n=14のときがすべての片仮名列を平仮名列に置換して辞書を再検索した場合に相当する。意図的片仮名総数216とn=14のときの改善件数171との差は、辞書に平仮名表記がないために改善できなかったものである。よって、n=14のときの改善件数と各nでの改善件数との差が、処理の適用洩れを表す。鎖破線は、処理対象となった片仮名列の件数(1件の片仮名列が1文字でも平仮名に置換されたなら1件として数える)を示す。

また、×印は、解析結果への悪影響を表す。これは、4.3.2節で述べた"オバタリアン"のように、未知語が不適切な検索語の組み合せで未知語でなくなってしまうものである。未知語は誤りとみなしたのでこのような例による解析精度の低下はない。しかし、ここでは従来解析の未知語処理への悪影響として捉えた。悪影響は、n=3で53件と最少になる。nが2以下で悪影響が多いのは、片仮名列を中途半端に平仮名列に変換して辞書を再検索するためである。nが4以上で悪影響が増えているのは、nが大きくなると再検索によって追加される単語の候補数が増え、候補選択に悪影響を与えるためである。

次に、nを変化させた場合のメモリおよび処理速度への影響を調べた。各nでの単語候補の追加総数、処理時間を表4に示す。適用文字総数は、処理が適用された文字(片仮名列およびその直前の文字列)の総数である(従来検索の場合はテキストの総文字数を示す。このうち片仮名は7,090文字)。候補増分比は、従来検索に対する単語候補数の増分の割合を表す。処理時間は、

表 4: 従来の単語候補リストへの追加単語候補数と処理速度

	従来検索	片仮名列置換検索処理					
n=	–	1	2	3	4	5	14
適用文字総数	82,235	1,905	3,540	6,405	9,079	10,262	11,504
追加候補総数	0	5,006	12,691	22,812	33,958	40,126	45,890
単語候補総数	173,963	178,969	186,654	196,775	207,921	214,089	219,853
候補増分比(%)	0	2.9	7.3	13.1	19.5	23.1	26.4
処理時間(秒)	132	135	136	138	139	140	144
時間増分比(%)	0	2.3	3.0	4.5	5.3	6.1	9.1

EWS4800/220(CPU：R3000A、30MHz)で測定した(特殊文字置換検索処理の時間も含む)。時間増分比は、従来検索に対する処理時間の増分の割合を表す。

片仮名列置換検索処理の適用条件1について、適用洩れがなく、かつ、余分な検索が少ないように、図5からn=4を選択することにした。n=4の場合、図5において、(○の件数－×の件数)が最大になっている。また、表4から、従来検索に比べて、単語候補数は19.5%増加、処理時間は7秒(5.3%)増加する。この程度の増加なら実用上は問題ないと言える。

同じテキスト中、特殊文字を含む単語は全部で53件、そのうち擬音語・擬態語を除く強調表現は31件であった。特殊文字置換検索処理による従来解析からの改善は27件、悪影響が15件であった。悪影響とは4.3.2節で述べた“うーむ”のような例である。また、追加される単語候補数は565であった。特殊文字を含む箇所は週刊誌1冊分でも片仮名列に比べると非常に少ない(適用文字総数：250)。候補増分比は0.3%程度なので、メモリ面での従来の解析に与える影響はほとんど無視できる程度と言える。処理時間も片仮名列置換処理を外して1秒(0.8%)増加する程度だった。よって、特殊文字置換検索処理も実用上の問題がないと言える。

5 おわりに

口語的表現を正しく形態素解析するために、辞書登録と口語置換検索処理の2つの手法を示した。これらの手法を実現した結果、口語的表現を多く含むテキストの文節区切り精度が1.8%向上し、テキストのタイプによらず安定した高い精度を得ることができた。辞書登録では、文節区切りに失敗していた話し言葉の88%を正しく解析できた。口語置換検索処理では、形態素解析に失敗していた意図的な片仮名表記の75%、強調表現で特殊文字を含む単語の79%を救済できた。

本稿で述べた手法のうち、辞書登録によるものは、実用システムである校正支援システムSt.WORDS[8]に既に組み込んで、その有効性を確認している。St.WORDSを利用している出版社では、週刊誌などの口語的表現を多く含むテキストを校正の対象としており、口語解析処理の強化が必要であった。今後さらに口語置換検索処理も組み込んでいきたい。ただし、その際、従来は未知語として検出されていた誤字・脱字箇所が、口語置換検索処理によって未知語として検出されなくなるような危険がある。これは、ユーザがシステムに不信感を抱く原因となる。校正支援システムに組み込む場合は、このような悪影響を防ぎ、ユーザを必要以上に惑わすことのないように考慮することも必要になる。

謝辞 評価用テキストを提供してくださった、株式会社 講談社に深く感謝致します。また、日頃から有益な助言をくださるNEC情報メディア研究所 山田洋志氏に感謝します。

参考文献

[1] 長尾眞 他,国語辞書の記憶と日本語文の自動分割,情報処理,Vol.19,No.6,1978

[2] 長尾眞監修,日本語情報処理,電子通信学会編,1984

[3] 宮崎正弘,係り受け解析を用いた複合語の自動分割法,情処論,Vol.25,No.2,1984

[4] 竹下敦 他,話し言葉に対する形態素解析情処42全大1C-3,1991

[5] 荻野紫穂,形式的でない表現における“ん”“ちゃ”“じゃ”“きゃ”の接続上の性質,計量国語学第19巻第1号,1993

[6] 竹元義美 他,口語的表現を含む日本語文の形態素解析,情処46全大1B-2,1993

[7] 竹元義美 他,口語的表現を含む日本語文の形態素解析の実現と評価,情処48全大1Q-1,1994

[8] 福島俊一 他,日本語文書校正支援システムSt.WORDS,情処45全大,6C-1,1992

[9] 長尾眞 他,自然言語処理のこれからの課題, 言語処理学会資料,1994