

筑波大学大学院博士課程

システム情報工学研究科修士論文

統計的言語モデルを用いた  
日本語活用誤りチェック

綿貫 雄太

(コンピュータサイエンス専攻)

指導教官 山本 幹雄

2005年1月

## 要旨

現在、情報技術を利用した語学教育システムが広く普及してきている。しかし自習用の課題の多くは × や選択、穴埋めなどの問題形式で決められた正解の入力を学習者に要求しており、入力自由度が少ないという欠点がある。作文などの自由入力の場合、計算機による有効な添削の手法がまだ確立されていないため、教師が直接添削する必要がある。こうした現状に対する解決策として考えられるのが、自由入力された文章に対して少なくとも低レベルの綴りや文法の誤りを自動的に指摘し、教師の負担を軽減するシステムである。しかし留学生など日本語の初級者の作文には局所的な誤りが数多くみられ、構文解析や形態素解析などの計算機による文書処理を大変困難にしている。またこれまで開発されてきたスペルチェッカは一般的な日本語を扱うユーザの作文に用いることを前提としているため、日本語の初級者の作文から誤りを正確に検出することはできない。当研究室ではこれらの問題を踏まえて、教師の添削支援および学習者の自習用を目的とした日本語の学習者向けのスペルチェッカを従来より開発してきた。システムは文字列の探索・変換をおこなうルールと文字単位で確率を計算する統計的言語モデル (*n*gram モデル) で構成されている。本研究では上記のスペルチェッカの「検出対象の拡大」および「検出性能の向上」を最終目標としている。これを実現するにあたって、今回は留学生の作文データを用いて日本語の初級者の活用誤りの分析を行い、その結果を元にスペルチェッカの検出対象を活用誤り全般に拡大して、詳細な評価をおこなった。さらに活用誤りの検出に用いる *n*gram モデルに逆向き *n*gram を導入した。本論文では留学生の日本語作文の特徴の詳細と逆向き *n*gram の導入によるスペルチェッカの評価結果も報告する。結論として、上記のスペルチェッカは出現回数の少ない誤りに対しても多くの場合確実に検出できること、逆向き *n*gram の導入は性能が低いルールによる過剰な検出の削減に効果があることを示す。

# 目次

<b>第1章</b>	<b>はじめに</b>	<b>1</b>
1.1	研究の背景と目的 . . . . .	1
1.2	本論文の構成 . . . . .	2
<b>第2章</b>	<b>日本語初級者の誤りの分析</b>	<b>3</b>
2.1	初級者の日本語作文の特徴 . . . . .	3
2.1.1	局所的な誤り . . . . .	4
2.1.2	平仮名表記・一般的でない表現 . . . . .	5
2.2	誤り要因としての日本語の活用 . . . . .	6
2.2.1	基本系とタ系 . . . . .	6
2.2.2	基本系語尾の活用規則 . . . . .	7
2.2.3	タ系語尾の活用規則 . . . . .	8
2.3	誤りの傾向の分析 . . . . .	9
2.3.1	分析方法 . . . . .	9
2.3.2	誤りの傾向(回数) . . . . .	10
2.3.3	誤りの傾向(パターン) . . . . .	12
2.3.4	まとめ . . . . .	13
<b>第3章</b>	<b>活用誤りチェック</b>	<b>14</b>
3.1	スペルチェックの概要 . . . . .	14
3.2	検出・訂正ルール . . . . .	15
3.3	統計的言語モデル . . . . .	17
<b>第4章</b>	<b>性能評価</b>	<b>19</b>
4.1	評価指標 . . . . .	19
4.2	ルール別の評価 . . . . .	20

4.2.1	性能が低いルール . . . . .	20
4.2.2	各ルールの性能 . . . . .	21
4.3	ルール集合に対する評価 . . . . .	25
<b>第 5 章</b>	<b>おわりに</b>	<b>26</b>
5.1	まとめ . . . . .	26
5.2	日本語学習への応用 . . . . .	26
	謝辞	28
	付録	29
	参考文献	43

# 第1章 はじめに

## 1.1 研究の背景と目的

現在、情報技術を利用した語学教育システムが広く普及してきている。計算機によって問題の出題や採点を自動化することで、教師の負担を軽減したり、学習者の自習を助けたりすることができるためである。これまで × や選択、穴埋めなどの問題を計算機に出題・採点させるシステムが開発され、日本語の教育現場で利用されている [1][2]。音声で提示された文章を文字に起こす形式の問題もあり [3]、効果を上げている。しかしこれらのシステムではあらかじめ決められた正解の入力を学習者に要求しており、正確な判定・評価が可能である反面、入力の自由度が少ないという欠点がある。

留学生による作文などの自由入力の場合、計算機による有効な添削の手法はまだ確立されていない。現状としては学習者が書いた作文に対して、日本語の教師が人手で添削して学習者に書き直させて、それをまた添削することを繰り返すといった手法がとられている。人手による添削は非常に手間がかかる作業であり、このことが現場で指導をおこなっている教師にとって大きな負担となっている。

こうした現状に対する解決策として考えられるのが、自由入力された文章に対して少なくとも低レベルの綴りや文法の誤りを自動的に指摘し、教師の負担を軽減するシステムである。このシステムは「スペルチェッカ」という形で以前からさまざまな手法が研究されてきた。しかしこれまでに開発されたスペルチェッカでは、日本語の初級者が書いた文章の誤りを正確に指摘することができない。いずれの手法も一般的な日本語を扱うユーザが書いた文章を対象としており、初級者特有の表現が使われる文章に対して利用することを想定していないためである。

そこで当研究室では、教師の添削支援および学習者の自習用を目的とした日本語の学習者向けのスペルチェッカを従来より開発してきた [4][5]。このシステムは対象とする誤りを限定するための検出ルールと、高い誤り捕捉率が期待できる統計的言語モデ

ルで構成されており、比較的容易にシステムの拡張および調整ができる特徴をもつ。スペルチェッカは開発当初から高い検出性能を示していたが、検出対象とした誤りのルールがやや少なく、それ以外の誤りは性能評価の基準とされていなかった。

本研究では上記のスペルチェッカにおける、日本語初級者の作文にみられる誤りの「検出対象の拡大」および「検出性能の向上」を最終目標としている。これを実現するにあたって、今回は筑波大学留学生センターよりご提供いただいた留学生の作文データ(91人分933文)による日本語の初級者の活用誤り(活用する品詞の語尾の誤り)の分析を行い、その結果を元にスペルチェッカの検出対象を活用誤り全般に拡大して、詳細な評価をおこなう。さらに活用誤りの検出に用いる統計的言語モデル(*n*gramモデル)に逆向き *n*gram を追加するなどして、検出性能を向上させることを目的とする。

## 1.2 本論文の構成

以下、第2章では今回の評価に用いた、留学生が実際に書いた作文の特徴について活用誤りの分析を中心に述べる。第3章では第2章の分析結果に基づいて開発したシステムについて述べる。第4章では性能評価の手法と、評価のためにおこなった実験の結果を述べ、システムの改良法の検討をおこなう。第6章ではまとめをおこない、今後の展望を述べる。

## 第2章 日本語初級者の誤りの分析

本研究の最終目標のひとつは「検出対象の拡大」である。これを実現するためには日本語の初級者がどのような綴りに対して、どういった誤りを、どの程度繰り返しているのかを、できるだけ詳しく知る必要がある。そこで本章では実際に留学生が書いた作文を分析し、その結果を用いて日本語の初級者の誤りの特徴を説明する。

### 2.1 初級者の日本語作文の特徴

まず初級者の日本語作文の特徴をつかむため、留学生の作文データの分析をおこなった。分析対象として、筑波大学留学生センターより提供していただいた91人分933文を使用した。作文データは、セリフのない漫画の内容を自由入力で説明させる問題の回答文をまとめたもので、猿の群れに盗まれた売り物の帽子を老人が取り返す様子を説明した文章が1行1文の形式でまとめられている。分析の結果から見えてきた主な特徴を以下に列挙し、具体例を図2.1に示す。数字付きの下線が引かれている部分が誤り箇所であり、列挙した特徴の数字と下線の数字とが対応している。

1. 局所的な誤りを数多く繰り返す  
例：活用の誤り、濁音の誤り、助詞の誤り等
2. 本来は漢字表記の単語が平仮名で書かれている  
例：「さる(猿)」「ぼうし(帽子)」「あたま(頭)」等
3. 一般的ではない特異な表現が使われる  
例：「お祖父」「古男人」「老者」等(いずれも老人の意)

図 2.1: 日本語初級者の誤りの例

<p><u>3</u>お祖父は<u>2</u>ぼうしをか<u>1</u>いて寝ました。 ひとり<u>1</u>(<u>の</u>)<u>3</u>古男人は木の<u>2</u>したにすわ<u>1</u>ています。 <u>3</u>老者は、<u>2</u>ぼうしをと<u>1</u>る、<u>2</u>さるたちは<u>3</u>老者<u>1</u>にまねて、<u>1</u>ぼうをすて<u>1</u>いました。</p>
---

検出対象を拡大するためには、こうした特徴のうち本研究で開発したスペルチェッカで検出できそうなものはどれなのかを検討する必要がある。そこで先ほど挙げた特徴について、まず局所的な誤りを、続いて平仮名表記と一般的でない表現を、順番に説明する。

### 2.1.1 局所的な誤り

ここでは先ほど挙げた特徴のうちの、「局所的な誤り」について簡単に説明する。日本語の初級者が書く文章は同じ綴り誤りを何度も繰り返しており、特に活用する品詞の語尾の誤りや、助詞の誤り、清音と濁音の誤りが多くみられる。それぞれの誤りの具体例を以下に示す。下線部分が誤り箇所である。

#### 活用の誤りの例

そのおじさんはぼうしをうています。

男の人は寝っています。

おきゃくさまこないのでおじさんがねむくてすぐねてしまたんです。

いちじかんあとで、おじいさんはおきった。

#### 助詞の誤りの例

老人は大きいきの下に帽子を売ります。

最後、おじいさん(が)勝ちました。

ひとり(の)古男人は木のしたにすわっています。

#### 濁音の誤りの例

さるがほうしをかぶりしました。

そこでじぶんがかぶっているほうしを地面にすてました。

それぞれの誤りの特徴を説明する。活用する品詞の語尾の誤りは、動詞や形容詞などの活用の規則を取り違えることで発生する。特に「て」と「って」、「た」を「った」を取り違えることが非常に多い(これについては次の節でさらに詳しく説明する)。助詞の誤りは、複数の単語の関係に見合った助詞を正しく選べないことで発生する。特に「で」を書くべき箇所に「に」を書いたり「が」や「の」を書くべき箇所に何も書かれていなかったりすることが多い。清音と濁音の誤りは、単順に濁点をつけ忘れてたり余分につけてしまったりするほか、間違って覚えた発音をそのまま文字に書き表すことでも発生する。今回分析した作文データでは「ぼうし」が多く登場するためか「ぼ」の濁点をつけ忘れていることが多い。

図 2.2: 初級者の作文の形態素解析例

老者はぼうしをもらいました。			
老	オイ	老	名詞-一般
者	シャ	者	名詞-接尾-一般
は	ハ	は	助詞-係助詞
ぼうし	ボウシ	ぼうする	動詞-自立 サ変・ - スル 未然形
を	ヲ	を	助詞-格助詞-一般
もらい	モライ	もらう	動詞-自立 五段・ワ行促音便 連用形
まし	マシ	ます	助動詞 特殊・マス 連用形
た	タ	た	助動詞 特殊・タ 基本形
。	。	。	記号-句点
EOS			

いずれの誤りも、正しい綴りとそれに対する誤りのパターンの組合せに、ある程度の規則性がみられる。この組合せをルールの形で表せば、検出ルールとして誤り検出のシステムに組み込み、スペルチェックの検出範囲を拡大させることが期待できる。

## 2.1.2 平仮名表記・一般的でない表現

ここでは「平仮名表記」「一般的でない表現」について簡単に説明する。本来なら漢字で表記される単語のほとんどを、日本語の初級者は平仮名で表記する。初級者の多くは漢字を使わない言語を母国語としており、日本語に使われている数多くの漢字を覚えて使いこなすことが難しいためである。また一般的な日本語には使われていないような特殊な表現を使うこともある。ある対象を指し示す正確な表現がわからないとき、関連する表現を覚えているものの中から選んで組み合わせるなどして書こうとするためである。

こうした表現は明確な誤りとは言いがたいが、いずれも構文解析や形態素解析などの計算機による文書処理をさまたげる原因となっている。実際に初級者の作文を形態素解析した場合の一例を図 2.2 に示す。下線部分が解析に失敗した箇所である。「老人」を表す特殊な表現「老者」は1文字ずつ別の単語として扱われ、「帽子」の平仮名表記「ぼうし」は動詞と判断されている。

## 2.2 誤り要因としての日本語の活用

日本語の初級者の作文にみられる「局所的な誤り」のうち、回数および種類が最も多いのが「活用の誤り」である。初級者にとって品詞ごとに異なる活用規則を正確に区別することが難しく、用いるべき規則を取り違えることで誤りが発生する。ここではそうした誤りの原因となっている日本語の活用について詳しく述べる。

### 2.2.1 基本系と夕系

日本語の活用は、活用語尾の違いに応じて基本系と夕系に大きく分けられ、その語尾の変化には一定の規則性がある [6]。そこで活用語尾の変化を一覧できるように、基本系・夕系それぞれの語尾の変化を表 2.1・表 2.2 にまとめた (より詳細な表は付録を参照)。表中の品詞の定義は次の通りである。

母音動詞 語幹の末尾が母音で終わる動詞

例：変える (kae)・着る (ki)

子音動詞 語幹の末尾が子音で終わる動詞

例：帰る (kaer)・切る (kir)

イ形容詞 名詞を修飾するとき、末尾が「い」になる形容詞

例：美しい・楽しい

ナ形容詞 名詞を修飾するとき、末尾が「な」になる形容詞

例：豊かな・幸せな

判定詞 名詞と結合して述語を作る語

例：だ・である・です

「タグ」の項目は作文データの分析にあたって誤り箇所につけたタグの、活用語尾との対応を示している。タグの形式については後ほど説明する。

表 2.1: 基本系語尾の活用規則

基本系語尾		基本形	命令形	意志形	条件形	連用形	連体形	
タグ		k	m	i	z	y	t	
母音動詞	kb	る	ろ	よう	れば	(無)	-	
子音動詞	ks	s	す	せ	そう	せば	し	-
		k	く	け	こう	けば	き	-
		g	ぐ	げ	ごう	げば	ぎ	-
		m	む	め	もう	めば	み	-
		n	ぬ	ね	のう	ねば	に	-
		b	ぶ	べ	ぼう	べば	び	-
		t	つ	て	とう	てば	ち	-
		r	る	れ	ろう	れば	り	-
w	う	え	おう	えば	い	-		
イ形容詞	ki	い	-	-	ければ	く(ず)	(な)	
ナ形容詞	kn	だ	-	-	-	に	な(の)	
判定詞		である	る	-	-	れば	り	-
です		す	-	-	-	-	-	

## 2.2.2 基本系語尾の活用規則

基本系語尾の活用規則を表 2.1 に示す。子音動詞の項目の英字は、ローマ字表記での語幹(形が変化しない部分)の「末尾」を表している。例えば「帰る(kaer + u)」の語幹の末尾は「r」であり、連用形は「帰り(kaer + i)」となる。母音動詞の連用形「(無)」は、語尾が無くなることを示している。例えば「変える(kae + ru)」の連用形は「る」が無くなって「変え(kae + (無))」となる。イ形容詞の連用形「(ず)」はイ形容詞系接尾辞「ない」の連用形の形式のひとつである。例えば「変えない(kae + nai)」の場合は「変えなく(kae + naku)」「変えず(kae + zu)」、「帰らない(kaer + anai)」の場合は「帰らなく(kaer + anaku)」「帰らず(kaer + azu)」となる。イ形容詞の連体形「(な)」は「大きい」「小さい」などごく限られた単語にのみ現れる活用である。ナ形容詞の連体形は「な」、判定詞の連体形は「の」であるが、後に「の」が続くと判定詞の連体形は「な」になるため、表中では「な(の)」の形で表記している。例えば「豊かだ」の連体形は「豊かな」、「学生だ」の連体形は「学生の」「学生なので」となる。

表 2.2: タ系語尾の活用規則

タ系語尾			基本形	条件形	連用形(テ)	連用形(タリ)
		タグ	k	z	t	r
母音動詞		tb	た	たら	て	たり
子音動詞	s(si)	ts1	した	したら	して	したり
	k(i)	ts2	いた	いたら	いて	いたり
	g(i)	ts3	いだ	いだら	いで	いだり
	m,n,b(n)	ts4	んだ	んだら	んで	んだり
	t,r,w(t)	ts5	った	ったら	って	ったり
イ形容詞		ti	かった	かったら	くて(いで・ずに)	かったり
ナ形容詞	だ	tn1	った	ったら	で	ったり
判定詞	である	tn2	った	ったら	って	ったり
	です	tn3	した	したら	して	したり

### 2.2.3 タ系語尾の活用規則

タ系語尾の活用規則を表 2.2 に示す。子音動詞の項目の英字は、ローマ字表記での語幹(形が変化しない部分)の「末尾の変化」を表している。具体的には語幹の末尾が( )で示した形に切り替わり、例えば「帰る(kaer + u)」の語幹の末尾は「t」に切り替わって連用形は「帰った(kaet + ta)」となる。表では活用語尾の変化をわかりやすくするため、日本語表記での活用部分を載せている。イ形容詞の連用形「(いで・ずに)」はイ形容詞系接尾辞「ない」の連用形の形式のひとつである。例えば「変えない(kae + nai)」の場合は「変えなくて(kae + nakute)」「変えないで(kae + naide)」「変えずに(kae + zuni)」、「帰らない(kaer + anai)」の場合は「帰らなくて(kaer + anakute)」「帰らないで(kaer + anaide)」「帰らずに(kaer + azuni)」となる。

## 2.3 誤りの傾向の分析

日本語の初級者の誤りに対して有効な検出をおこなうルールを用意するためには、これまで説明したような規則を持つ活用について、初級者がどういった誤りをどの程度繰り返しているのかを知る必要がある。ここでは作文データに出現する誤りを数えあげた結果を、種類ごとの誤り回数の傾向、および実際の誤りのパターンに分けて説明する。

### 2.3.1 分析方法

留学生の作文データに見られる誤りに対してタグ付けをおこない、その数を対応する正しい活用ごとに数えあげて表 2.3・表 2.4 にまとめた(より詳細な表は付録を参照)。「/」を挟んで並んでいる数字は、左が誤りの数、右が正しく書かれていた数を表している。その他の誤りについてもタグ付けと数えあげをおこない、その数を表 2.5 にまとめた。

実際の誤りには<\*\*\* \_\_/ \_\_>という形でタグ付けがなされた。「\*\*\*」は誤りの種類を示す記号列、「\_\_/」は原文の誤り部分、「/ \_\_」は正しい綴りを表している。誤りの種類を示す記号は、基本的に表 2.1・表 2.2 に使われている各項目のローマ字表記の頭文字で表される。例えば「老人は起きた」のように、夕系語尾 (takeigobi) のうち母音動詞 (boindousi) の基本形 (k<sub>i</sub>honkei) として「た」が現れるはずの箇所に誤りがあった場合、「老人は起き<tbkった/た>」のようにタグが付けられる。

表 2.3: 基本系語尾の誤り箇所・正しい箇所の数

基本系語尾			基本形	命令形	意志形	条件形	連用形	連体形
		タグ	k	m	i	z	y	t
母音動詞		kb	4/129	0/1	0/2	0/0	27/278	-
子音動詞	s	ks	1/3	1/1	0/0	0/0	1/9	-
	k		0/6	0/0	0/0	0/0	1/10	-
	g		0/0	0/0	0/0	0/0	0/0	-
	m		0/0	0/0	0/1	0/0	0/10	-
	n		0/0	0/0	0/0	0/0	0/0	-
	b		4/4	0/0	0/1	0/0	0/20	-
	t		0/2	0/0	0/0	0/0	0/4	-
	r		2/78	0/0	0/2	0/0	4/117	-
w	0/11	0/0	0/0	0/0	1/42	-		
イ形容詞		ki	5/64	-	-	0/0	4/14	0/2
ナ形容詞	だ	kn	0/5	-	-	-	0/2	0/6
判定詞	である		0/0	-	-	0/0	0/0	-
	です		1/93	-	-	-	-	-

### 2.3.2 誤りの傾向 (回数)

基本系語尾の誤り回数の分布を表 2.3 に示す。母音動詞は連用形「(ゼロ)」の誤りが目立って多い。子音動詞は誤りの回数そのものは多くはないが、正しく書かれていた回数に対する誤りの割合が大きい。一方で子音動詞の連用形「び」は登場回数の多さにも関わらず、すべての箇所が正しく書かれている。イ形容詞も子音動詞と同様に、正しく書かれていた回数に対する誤りの割合が大きい。ナ形容詞・判定詞については目立った誤りが見られなかった。

表 2.4: タ系語尾の誤り箇所・正しい箇所の数

タ系語尾		基本形	条件形	連用形(テ)	連用形(タリ)	
タグ		k	z	t	r	
母音動詞		tb	20/93	1/5	42/163	0/5
子音動詞	s(si)	ts1	2/8	0/1	4/13	1/0
	k(i)	ts2	0/3	0/0	2/10	0/0
	g(i)	ts3	0/0	0/0	1/2	0/0
	m,n,b(n)	ts4	1/7	0/0	15/59	0/1
	t,r,w(t)	ts5	17/83	0/0	76/238	7/6
イ形容詞		ti	1/18	0/0	2/10	0/0
ナ形容詞	だ	tn1	0/1	0/1	0/0	
判定詞	である	tn2	0/0	0/0	0/0	0/0
	です	tn3	1/13	0/0	0/0	0/0

表 2.5: その他の誤りの数

分類	誤り数
助詞	301
濁音(脱落)	52
濁音(挿入)	37
その他	402

タ系語尾の誤り回数の分布を表 2.4 に示す。母音動詞は基本形「た」と連用形「て」、子音動詞は基本形「った」と連用形「って」の誤り回数が非常に多い。それ以外の子音動詞やイ形容詞は、誤りの回数はそれほど多くはないが、正しく書かれていた回数に対する誤りの割合が大きい。ナ形容詞・判定詞についてはほとんど誤りが見られなかった。

表 2.5 は活用するもの以外の品詞のうち、特に誤りが多いものをまとめた。「助詞」は主に格助詞(述語とそれを補足する語との関係を示す助詞)の挿入・脱落・置換といった誤りを対象としている。「濁音(脱落)」は本来つくはずの濁点がついていない誤り、「濁音(挿入)」は本来つかないはずの濁点がついている誤りを対象としている。それ以外の、規則性が見出しにくい誤りは「その他」でまとめている。

表 2.6: 基本系語尾・濁音の誤り

誤り/正解	誤り数
り/(無)	19
い/(無)	5
ほ/ぼ	20
と/ど	7
ふ/ぶ	7
か/が	5
だ/た	9
で/て	6
が/か	4
ぞ/そ	4

表 2.7: タ系語尾の誤り

誤\正	た	て	して	んで	った	って
た	(93)	-	-	-	9	-
て	1	(163)	1	1	-	53
して	-	-	(13)	1	-	-
んで	-	3	-	(59)	-	3
った	14	-	-	-	(83)	2
って	2	24	3	-	-	(238)
んて	-	-	-	4	-	1
っで	-	1	-	1	-	1
だ	1	-	-	-	4	-
で	1	3	-	1	-	3

表 2.8: 助詞の誤り

誤\正	を	に	の	が	で	は	(無)
を	(-)	4	-	6	1	4	2
に	15	(-)	5	4	26	1	9
の	-	-	(-)	1	-	1	19
が	14	2	-	(-)	1	14	4
で	-	7	-	-	(-)	-	4
は	3	-	-	9	-	(-)	11
(無)	9	11	21	23	7	16	-

### 2.3.3 誤りの傾向 (パターン)

これらの誤りについて実際に留学生の作文データに出現した誤りのパターンまで区別した場合の、代表的な誤りをまとめたのが表 2.6~2.8 である (より詳細な表は付録を参照)。「/」を挟んで並んでいる文字は、左が誤りのパターン、右が正しい綴りを表している。表 2.6 に示した基本系語尾の誤りでは母音動詞の連用形 (語幹のみの形) に、子音動詞の連用形のつもりで「り」を加えてしまい「替えます」などとする誤りが多い。表 2.7 に示したタ系語尾の場合は「た」と「った」、「て」と「って」を取り違えて「起きった」「取た」「起きって」「取て」などとする誤りが多い。表 2.8 に示した助詞の誤りでは「で」を「に」とする置換誤りや「の」「が」の脱落誤りのほか、「を」を「に」「が」とする置換誤りや「に」「は」の脱落誤り、「の」「は」の挿入誤りが多い。表 2.6 に示した濁音の誤りでは「ぼ」の濁点の脱落誤りが多い。

表 2.9: 代表的な活用誤りの誤り箇所・正しい箇所のパターンと回数

		基本形		連用形	
		パターン	回数	パターン	回数
基本系語尾	母音動詞	-	4/129	り/e	19/278
夕系語尾	母音動詞	った/た	14/93	って/て	24/163
	子音動詞	た/った	9/83	て/って	53/238

表 2.10: 代表的な助詞・濁音の誤り

タグ	誤り/正解	誤り数
z	に/を	15
	が/を	14
	e/に	11
	e/の	21
	e/が	23
	に/で	26
	e/は	16
	が/は	14
	の/e	19
	は/e	11
+	ほ/ぼ	20
-	だ/た	9

### 2.3.4 まとめ

以上の分析結果から特に回数が多い誤りを選び、活用誤りを表 2.9 に、それ以外の誤りを表 2.10 にまとめた。これまでは夕系語尾の基本形と連用形を中心に検出ルールを組み込んでいたが、今回の分析で新たに基本系語尾の母音動詞の連用形に誤りが集中していることが判明した。全体的に見ると、品詞では母音動詞と子音動詞、活用形では基本形と連用形に誤りが集中しているが、これは日本語の初級者の作文には、上記以外の品詞や活用形が使われること自体がほとんどないためと考えられる。

## 第3章 活用誤りチェック

本研究のもうひとつの最終目標は「検出性能の向上」である。これを実現するためには日本語の初級者の誤りを検出するシステムを、より正確な検出ができるように改良する必要がある。そこで本章では従来より当研究室で開発されてきたスペルチェックの概要と、検出ルールおよび統計的言語モデルによる誤り処理の手順を示し、性能向上の手法として新たに導入された逆向き *ngram* を説明する。

### 3.1 スペルチェックの概要

日本語のスペルチェックは、これまでにさまざまな手法で開発がおこなわれてきた。例えばスペルチェック用の辞書をつくる手法としては大量の平仮名列を登録する手法 [7] が用いられている。単語や文字の出現確率を用いる手法としては単語 *ngram* [8]、平仮名の文字 *ngram* を利用したり [9]、各種のマルコフモデルを組み合わせたり [10][11]、確率的 LSA を導入する手法 [12] が用いられたりしている。

しかしこれらのシステムは初級者特有の誤りが数多く含まれる作文から、誤りを正確に検出・訂正することは非常に困難である。いずれも日本語話者が扱う一般的な日本語に対して用いることを前提に作られているためである。初級者の綴り誤りの具体的なパターンは無数に存在するため、辞書を作る手法で全ての誤りに対応しようとすると無理が生じる。初級者の作文は精度よく単語に分割することが難しいため、単語 *ngram* は適用できない。文字 *ngram* は、本来なら漢字で表記されるはずの平仮名表記の単語を誤りとして過剰に検出してしまう。

また初級者の作文に対して用いる日本語スペルチェックは、複雑な前処理を必要としないことが重要である。第2章で述べた初級者特有の誤りは形態素解析などの前処理も難しくしているためである。これらの問題を踏まえて、今回基本としたシステムは文字列の検索・変換をおこなうルールと文字単位で確率を計算する統計的言語モデルで構成されている [4][5]。誤り処理の手順を以下に示す。具体的な流れは図 4.1 のようになる。

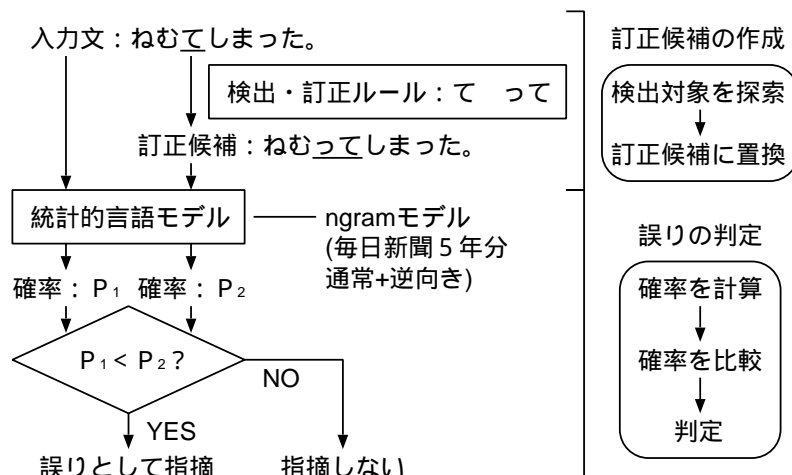


図 3.1: スペルチェッカの誤り処理の手順

1. 入力文に対してルールに記述された「対象」の検索をおこない、見つかった箇所の文字列を「候補」に変換した「訂正候補」を作成する。
2. 入力文と訂正候補の日本語らしさの確率を統計的言語モデルで比較し、訂正候補の確率が入力文の確率を上回ったら「対象」が誤りであると指摘し、必要ならば訂正する。

### 3.2 検出・訂正ルール

誤りの検出・訂正をおこなうためのルールは、表 2.1、表 2.2 に代表される留学生の誤りの傾向をもとに作成した。具体的には作文データに存在する誤りのうち、「その他」以外のいずれかに分類される誤りを出現回数に関わらずすべてルール化した。現在の作文データに 1 回しか出てこない誤りであっても、今後さらに大量の作文データを分析した場合に大量に出現する可能性があり、その誤りを検出するルールが非常に有用なものとなることが考えられるためである。

今回の評価にあたって導入した、基本系語尾・夕系語尾・濁音の誤りに対応するための全 125 種類のルールを表 3.1～3.3 に示す。助詞の誤りについてはこれまでの評価で検出性能が極めて低いことが確認されているため [4][5]、今回導入したルールには含まれていない。ルール中の「e」はその箇所に文字列がないことを示しており、「e」は文字列の脱落誤りを検出・訂正するルール、「 e」は文字列の挿入誤りを検出・訂正するルールを表す。

表 3.1: 基本系語尾の誤りの検出・訂正ルール(全 26 種類)

e	る、す	る、つ	る、り	e、い	e、る	e、る	ぶ、
	び	ぶ、む	ぶ、う	る、す	る、した	す、ろう	せ、
	させ	し、く	き、る	り、え	り、み	り、れ	り、e
	い、	の	い、く	て	く、い	く、ま	す
	です	です	ます	です	です	ます	

表 3.2: 夕系語尾の誤りの検出・訂正ルール(全 73 種類)

つ	た、	つ	た、	る	た、	て	た、	だ	た、	で	た、	た	が	た	ら、
っ	て	て、	で	て、	e	て、	る	て、	れ	て	て、	わ	て、		
れ	て、	こ	て、	い	て	て、	む	れ	て、	っ	で	て、	た	した、	
っ	た	した、	っ	て	して、	て	して、	う	たり	し	たり、	け	て	いて、	
っ	て	いて、	れ	て	いで、	ん	で	ん	だ、	ん	て	ん	で、	び	ん
で	ん	で、	e	ん	で、	て	ん	で、	む	ん	ん	で、	し	て	ん
っ	で	ん	で、	び	て	ん	で、	び	て	ん	で、	び	っ	て	ん
た	った、	だ	った、	る	った、	い	た	った、	し	た	った、	り	た	った、	て
っ	て、	い	て	っ	て、	り	て	っ	て、	で	っ	て、	ん	で	っ
っ	て、	ら	れて	っ	て、	る	して	っ	て、	る	っ	たり、	た	り	っ
り	たり	っ	たり、	り	たり	っ	たり、	い	か	った、	か	っ	て	く	て、
い	て	いで、	て	した	でした、	ま	した	ました、	ま	ました、	ま	た	ました、	ま	だ
ま	だ	ました、	て	した	ました、	で	した	ました、	ま	ましたら、	な	した、	e	して	

表 3.3: 濁音の誤りの検出・訂正ルール(全 26 種類)

ほ	ほ、	と	ど、	ふ	ぶ、	か	が、	け	げ、	し	じ、	た	だ、
ひ	び、	こ	ご、	つ	づ、	て	で、	ひ	び、	け	げ、		
だ	た、	で	て、	が	か、	ぞ	そ、	ぴ	ひ、	ず	す、	ど	と、
ば	は、	ぎ	き、	ご	こ、	ち	ち、	づ	つ、	ぶ	ぶ		

### 3.3 統計的言語モデル

統計的言語モデルとは、記号の集合の中に日本語など特定の言語を仮定したとき、その言語に対する特定の文字列の「受け入れやすさ」を確率で与える確率分布関数である。たとえば日本語のコーパス(モデルの構築に用いる大量のテキスト)で単語あるいは文字の出現頻度を統計的に算出したモデルの場合、日本語として受け入れやすい文字列に対しては高い確率を与え、日本語として受け入れにくい文字列に対しては低い確率を与える。

評価するスペルチェッカに組み込まれた言語モデルは *n*gram モデル [13] と呼ばれるもので「ある文字列における特定の文字の出現確率は直前の *n*-1 文字に依存する」という仮定に基づいている。例えば 3gram(trigram) モデルは、連続した 2 文字の「直後」に出現する 1 文字の出現確率によって構築されている。

今日 は いい天気です。

今回は逆向き *n*gram による確率を考慮に入れた場合の性能も評価した。これは通常とは逆の方向に文章を読みとって算出した出現確率を用いたモデルで、例えば逆向き 3gram モデルは、連続した 2 文字の「直前」に出現する 1 文字の出現確率によって構築されている。

今日 は いい 天気です。

*N* 文字からなる文  $S = c_1, c_2, \dots, c_N$  の確率は *n*gram モデルによる *S* の確率を文字当たりに幾何平均した値を用いる。 $c_i$  は文字を表す。

$$p(S) = \left\{ \prod_{i=1}^N p(c_i | c_{i-n+1}, c_{i-n+2}, \dots, c_{i-1}) \right\}^{1/N}$$

入力文を  $S_{in}$ 、訂正候補文を  $S_c$  とすると訂正候補文が実際に誤りを訂正しているかの判断は次式でおこなう。

$$\frac{p(S_c)}{p(S_{in})} > Th$$

*Th* は閾値で、上式を満たせば訂正候補文は誤りを訂正していると判断し、誤りを指摘する。実際の確率の比較は対数値でおこなっているため、閾値も対数値で設定している。例えば閾値を 2 に設定した場合、実際の閾値は  $e^2 = 7.389$  となる。

初級者の日本語作文に多い平仮名表記にも対応できるように、今回組み込んだ *ngram* モデルは独自の手順で作成した。以下にその手順を示す。

1. モデル作成のためのテキストとして  
毎日新聞の記事1年分(95年)および5年分(95~99年)を用意する。
2. 元の記事を平仮名に変換したテキストを、元の記事とは別に用意する。
3. 逆向き *ngram* を作成する場合は  
ここで1.と2.のテキストの文字の並び順を逆にする。
4. それぞれのテキストを文字に分割する。
5. 元の記事の文字分割テキストと  
平仮名に変換した文字分割テキストを連結する。
6. 連結した文字分割テキストをもとにして  
CMU/Cambridge SLM Toolkit[14]で *ngram* を作成する。  
モデルの大きさは4、ディスカウントは Good-Turing、カットオフはなし。

## 第4章 性能評価

この章ではスペルチェッカの性能評価に用いる指標を説明したのち、2.3節で作成したタグ付きデータを基準データとしておこなった性能評価の結果について考察する。評価項目は以下の通りである。

- ルール別の評価  
単一ルールで動作させたシステムの性能が最良になるように（具体的にはF値（後述）が最大になるように）閾値を設定した場合の性能をすべてのルールに対して個別に評価した。
- ルール集合に対する評価  
125種類すべてのルールで動作させたシステムにおける閾値の一律な変化に応じた性能の変化をグラフで表し、導入する *n*gram の種類の違いによる性能の違いを評価した。

### 4.1 評価指標

今回の性能評価の指標として用いる適合率・再現率は以下のように計算される。

$$\text{適合率 (正解率)} = \frac{\text{実際に誤りであった箇所}}{\text{システムが指摘した全ての誤り}}$$

$$\text{再現率 (捕捉率)} = \frac{\text{システムが指摘できた誤り}}{\text{文書中に存在する全ての誤り}}$$

一般にこの2つの値はトレードオフの関係にあり、3.3節で触れた閾値を変化させることでいずれかの値が上がるように調整することができる。閾値を高くすれば明らかな誤りだけを指摘できるため適合率は上がるが、誤りであっても指摘しないことが多くなるため再現率は下がる。閾値を低くすれば多数の誤りを指摘できるため再現率は上がるが、誤りでない箇所も指摘してしまうため適合率は下がる。

表 4.1: 最大 F 値が 0.2 未満のルール

<p>e る、つて る、い e、る e、る ぶ、む ぶ、す る、した す、          る り、e い、る い、い く、です ます、る た、て た、          で て、e て、た した、つて して、て して、e んで、          て んで、して んで、だ った、る った、る って、っ って、          られて って、る ったり、い かった、でした ました、な した、          e して、け げ、し じ、て で、</p>
--

## 4.2 ルール別の評価

ルール別の評価の結果をもとに、最大 F 値 0.2 を基準にして各ルールの振り分けをおこない、最大 F 値が 0.2 未満のルールを表 4.1 にまとめた。システムの性能を、実用に耐えうる水準に維持するためである。F 値とは適合率・再現率値をまとめて 1 つの値にしたものであり、両者の重みを同じとした場合は、両者の幾何平均で表される。最大 F 値が 0.2 以上の各ルールの適合率・再現率は表 4.2、表 4.3 にまとめた (より詳細な表は付録を参照)。

### 4.2.1 性能が低いルール

「る」を検出するほとんどのルールの性能が低い水準にとどまっているが、これは ngram モデルの作成に用いた新聞記事の内容と評価に用いた作文データの内容のずれによるものである。「る」に対する実際の検出結果の一部を以下に示す。

- 「る」の検出結果の例

木の上に五ひきさ\*るがいます。

さ\*るは、ぜんぶぼうしをかぶります。

おじいさんのぼうしをのんとんさ\*るに取された。

だから、さ\*るを叱って、自分のぼうしを返したかと尾もいます。

でもさ\*るたちはそのことがしたくなかった。

実際の検出結果は、このようにシステムが誤りと判断した箇所の先頭に誤りを示す印 (\*) を挿入する形式になっている。これを見ると、「さる」の「る」に対して過剰に検出をおこなっていることがわかる。作文データは老人と猿とのやりとりが大半を占めている一方、新聞記事は幅広い話題について書かれているため、結果として「さる」の出現確率が低く見積もられ、誤りと判断してしまうためと考えられる。

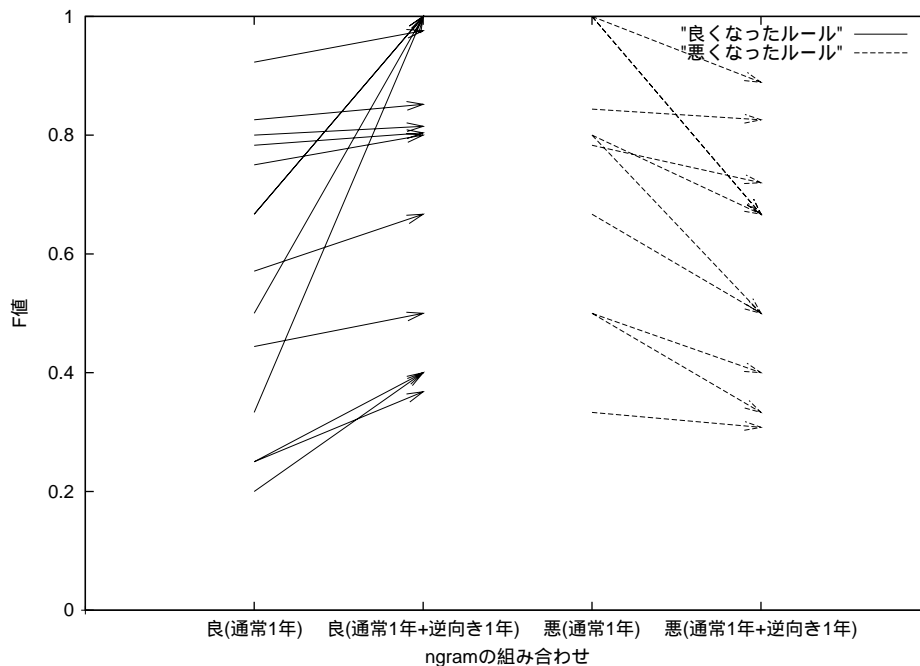


図 4.1: 各ルールの性能の変化

#### 4.2.2 各ルールの性能

最大F値が0.2以上の各ルールの適合率・再現率を表4.2、表4.3に示す。表4.2には逆向き ngram の導入によって性能が上がったルール、表4.3には性能が下がったルールをまとめた。またそれぞれのルールのF値の変化を図4.1に示す。表4.2、表4.3を見ると、性能が上がったルールの数が下がったルールの数を上回っている。また図4.1を見ると、F値の増加の割合が減少の割合を上回っている。こうしたことにより、全体として検出性能が向上していることがわかる。

代表的な活用誤りの検出ルールについて、通常の ngram1年分のみ・通常1年分と逆向き1年分を併用といった条件のもと、閾値の変化に対する性能の変化をグラフにして図4.2~4.5にまとめた。逆向き ngram の導入による性能の変化を見てみると、「って」「て って」はほとんど変化がみられないが、「た った」「で って」は確実に性能が向上しており、特に「で って」は誤りの指摘洩れが完全になくなっている。このことから、逆向き ngram の導入はもともと性能が高いルールよりも、これまで性能が低かったルールに対して有効であることがわかる。

表 4.2: 性能が上がったルール

ルール	通常 1 年			通常 1 年+逆向き 1 年		
	F 値	適合率	再現率	F 値	適合率	再現率
れり	0.333	20%(1/5)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
ったた	0.800	91%(10/11)	71%(10/14)	0.815	85%(11/13)	78%(11/14)
ってて	0.826	86%(19/22)	79%(19/24)	0.852	77%(23/30)	96%(23/24)
っでて	0.500	33%(1/3)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
たった	0.250	17%(4/23)	44%(4/9)	0.368	24%(7/29)	78%(7/9)
いたった	0.667	50%(1/2)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
てって	0.783	72%(45/62)	85%(45/53)	0.804	76%(45/59)	85%(45/53)
でって	0.667	50%(3/6)	100%(3/3)	1.000	100%(3/3)	100%(3/3)
ったって	0.750	60%(3/5)	100%(3/3)	0.800	100%(2/2)	67%(2/3)
まました	0.250	14%(1/7)	100%(1/1)	0.400	25%(1/4)	100%(1/1)
まましたら	0.200	11%(1/9)	100%(1/1)	0.400	25%(1/4)	100%(1/1)
ほぼ	0.923	95%(18/19)	90%(18/20)	0.976	95%(20/21)	100%(20/20)
でて	0.571	40%(6/15)	100%(6/6)	0.667	56%(5/9)	83%(5/6)
がか	0.444	40%(2/5)	50%(2/4)	0.500	50%(2/4)	50%(2/4)

表 4.3: 性能が下がったルール

ルール	通常 1 年			通常 1 年+逆向き 1 年		
	F 値	適合率	再現率	F 値	適合率	再現率
り e	0.844	73%(19/26)	100%(19/19)	0.826	70%(19/27)	100%(19/19)
びぶ	0.500	33%(1/3)	100%(1/1)	0.400	25%(1/4)	100%(1/1)
るて	0.333	20%(2/10)	100%(2/2)	0.308	18%(2/11)	100%(2/2)
けていて	0.500	33%(1/3)	100%(1/1)	0.333	20%(1/5)	100%(1/1)
りてって	1.000	100%(4/4)	100%(4/4)	0.889	80%(4/5)	100%(4/4)
んでって	0.800	100%(2/2)	67%(2/3)	0.667	67%(2/3)	67%(2/3)
いていで	1.000	100%(1/1)	100%(1/1)	0.667	50%(1/2)	100%(1/1)
またました	0.667	50%(1/2)	100%(1/1)	0.500	33%(1/3)	100%(1/1)
だた	0.783	64%(9/14)	100%(9/9)	0.720	56%(9/16)	100%(9/9)
どと	0.800	67%(2/3)	100%(2/2)	0.500	33%(2/6)	100%(2/2)
ぶぶ	1.000	100%(1/1)	100%(1/1)	0.667	50%(1/2)	100%(1/1)

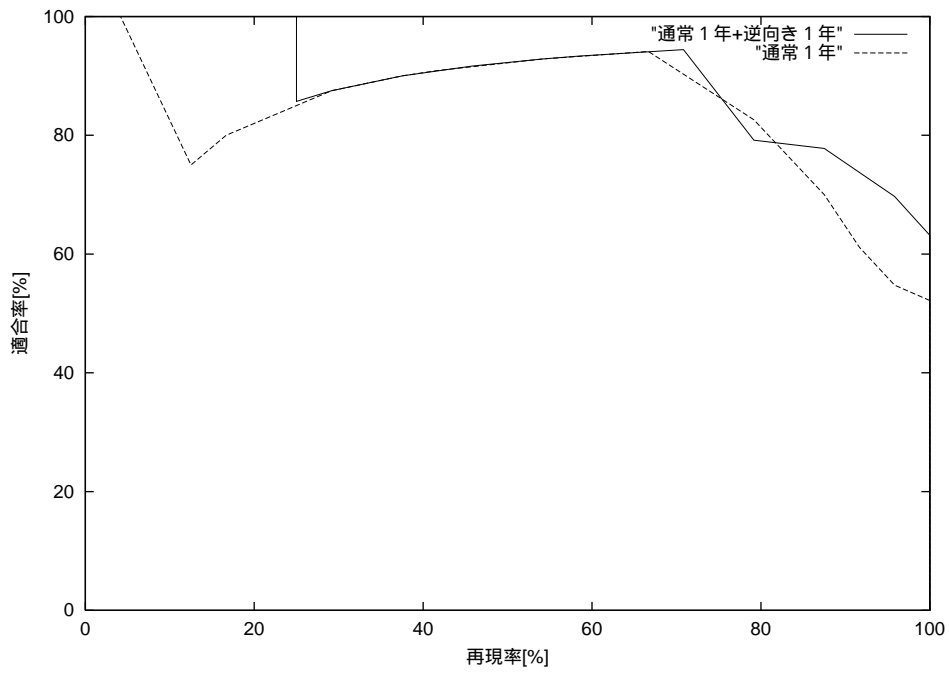


図 4.2: 「って て」の性能の変化

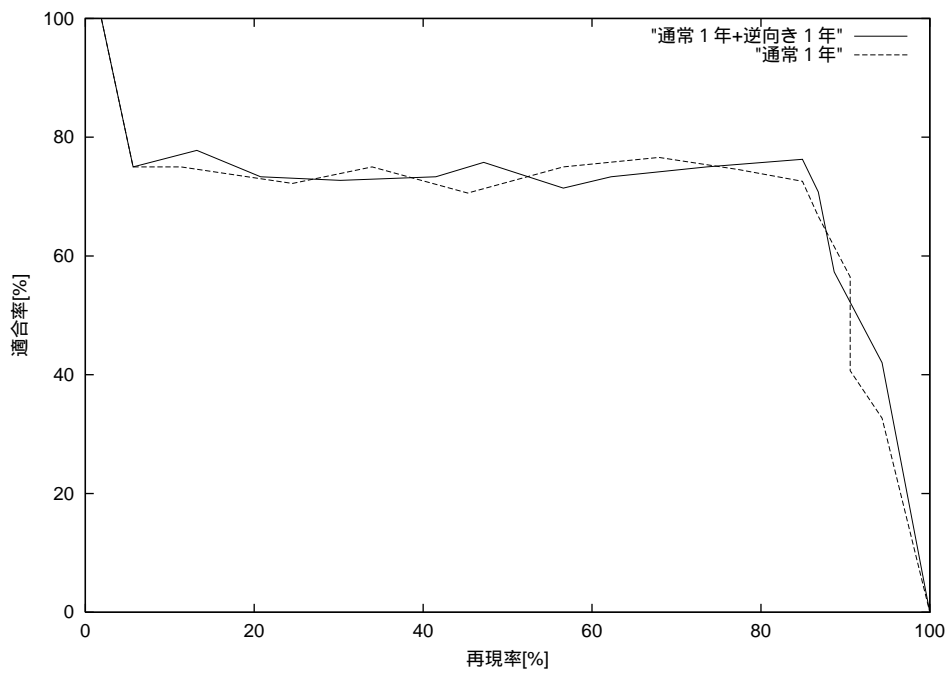


図 4.3: 「て って」の性能の変化

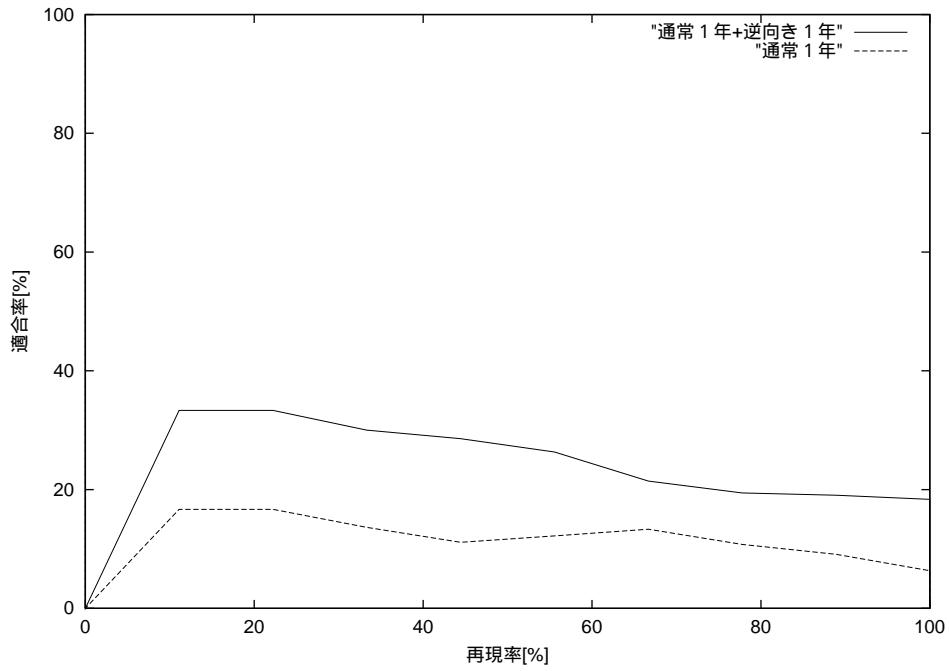


図 4.4: 「た った」の性能の変化

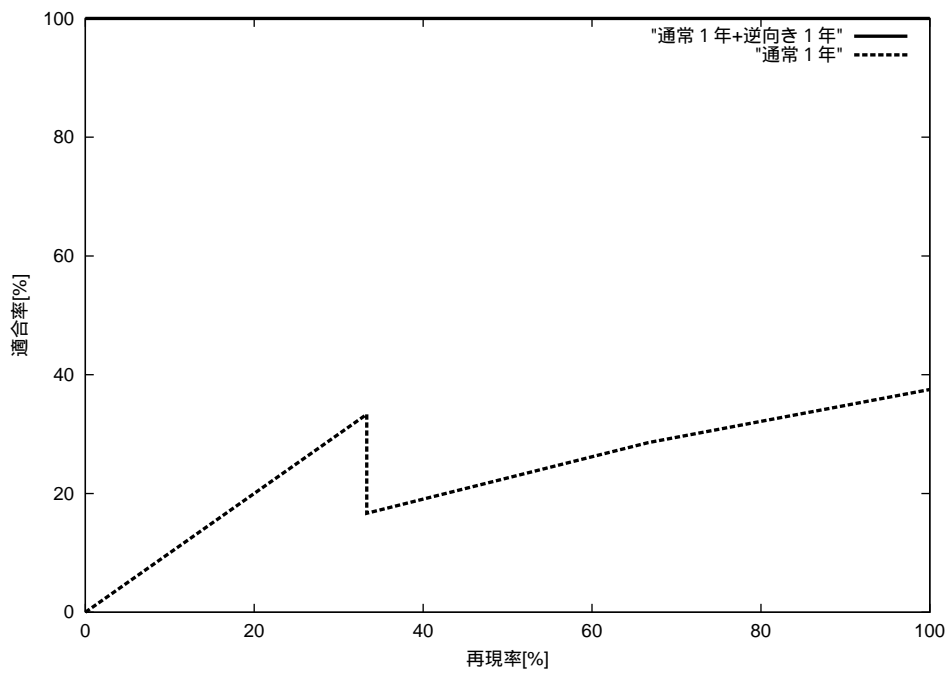


図 4.5: 「で った」の性能の変化

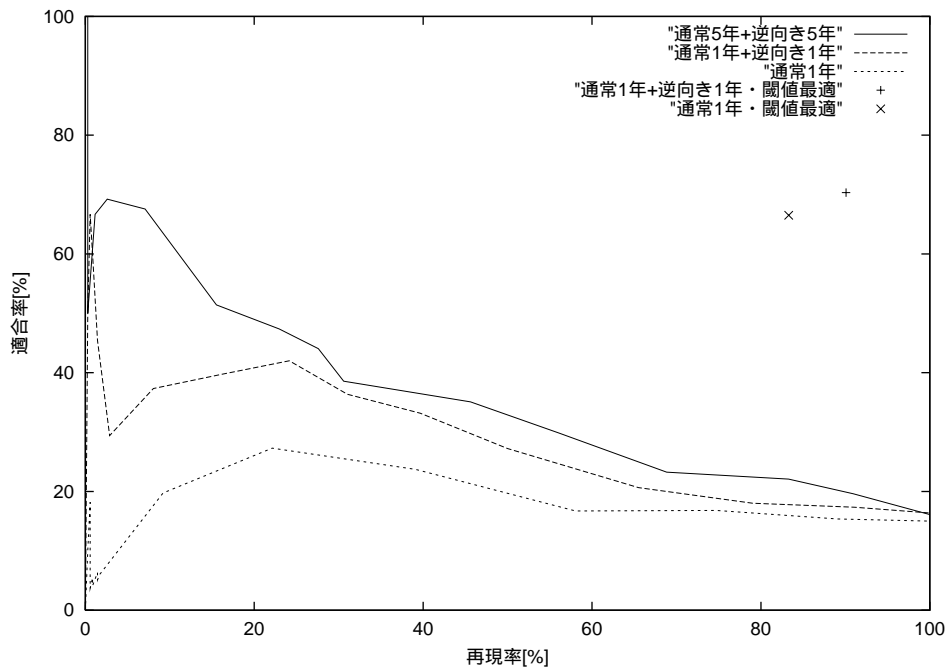


図 4.6: 逆向き ngram 導入による性能の変化

### 4.3 ルール集合に対する評価

通常の *ngram* 1 年分のみ・通常 1 年分と逆向き 1 年分を併用・通常 5 年分と逆向き 5 年分を併用といった条件のもと、すべてのルールの閾値を一律に変化させた場合の検出性能を図 4.6 に示す。適合率と再現率はトレードオフの関係にあるため、本来ならば右下がりのグラフになるはずだが、図 4.6 の左側は左下がりになっている。この理由は「た った」の性能の変化に見られるように、作文データ中のごく少数の誤りに対するルールの性能のグラフが図 4.6 と似た形状になることから、そうしたルールが多くの検出誤りを発生させるためであると考えられる。

逆向き *ngram* の導入や新聞記事の増量による性能の変化を見てみると、再現率が高い部分ほど性能の向上の度合いが低い。このことから、これらの改良は性能が低いルールに対する「過剰な検出の削減」に効果があると考えられる。図 4.6 の左側の 2 つの点は、それぞれ通常の *ngram* 1 年分のみ・通常 1 年分と逆向き 1 年分を併用といった条件のもと、各ルールに適切な閾値を設定した場合の検出性能を表している。なおこれについては、システムの実用性を考えて最大 F 値が 0.2 未満のルールは除いている。これを見ると、適合率・再現率ともに確実に向上していることが確認できる。

## 第5章 おわりに

### 5.1 まとめ

日本語初級者の活用誤りの傾向を留学生の作文データをもとに分析し、当研究室で開発されてきた日本語スペルチェッカの活用誤り全般についての評価実験を報告した。初級者の活用誤りの傾向として、これまでおもな検出対象としてきたタ系語尾の基本形「た」「った」と連用形の「て」「って」の取り違いによる誤りのほか、基本系語尾の連用形「(語尾無し)」「り」の混同による誤りが多く見られた。またその他の活用誤りについても非常に多くの種類が、それぞれ少数ではあるが確認された。スペルチェッカの活用誤り全般についての評価実験では逆向き *n*gram の導入による性能の向上がみられ、特にごく少数の誤りに対する過剰な検出が減って多くの種類の誤りを確実に検出できることが確認された。今後はより多くの作文データによる誤り傾向の分析、およびより確実な性能向上の手法の検討が課題となる。

### 5.2 日本語学習への応用

計算機による語学学習では、問題の提示による学習の支援だけでなく教師の問題作成の支援もおこなわれている [15]。今回評価したシステムを初級者の日本語学習に利用する場合、検出結果を利用した独自の出題方式を考えることができる。たとえば自由入力による回答に対して、実際の誤りを検出するとともに、誤りに対する正解となる正しい綴りやその誤りと同じ形の正しい綴りを多くの学習者の作文から見つけ出し、誤りと正解を混ぜ合わせたリストを作成すれば、どの綴りが誤りか、あるいは正解かを選ばせる選択問題として利用することができる。

この出題形式は初級者だけでなく学習言語にある程度慣れてきた中級者にも対応できる。初級者に対してはごく小さいリストを提示し、誤りの数または正解の数をあらかじめ教えることで、問題の難易度を抑えることができる。

問. 次の文の中から誤った使い方をしている「って」を1つ選びなさい。

この話は帽子を売っている人と5匹の猿の話です。  
男の人は寝っています。  
五ひきのさるは木の上にぼうしをかぶっています。

中級者に対してはある程度大きなリストを提示し、誤りの数または正解の数を学習者自身に考えさせることで、問題の難易度を引き上げることができる。

問. 次の文の中から誤った使い方をしている「って」を選びなさい。

木の下にぼうしをおじいさんは売ってしました。  
木の下に男の人はぼうしをかぶっています。  
男の人は寝っています。  
五ひきのさるは木の上にぼうしをかぶっています。  
男の人は笑っています。  
この話は帽子を売っている人と5匹の猿の話です。  
帽子を売っている人はマリアさんです。

このように自由入力による回答に対して、ただ添削をおこなうだけでなく誤りの検出結果をもとにした選択問題を出すことで、教師は学習者の誤りの傾向にあわせた指導が、学習者は誤りやすい綴りに重点を置いた学習ができるようになる。

# 謝辞

本研究を進めるにあたり、多大な助言と御指導を頂きました筑波大学 電子・情報工学系 山本幹雄 助教授に心からお礼申し上げ、厚く感謝致します。

また、様々な助言をくださり、本論文の審査をして頂きました筑波大学 電子・情報工学系 板橋秀一 教授と椎名 毅 教授に深く感謝致します。

最後になりましたが、日頃から様々な助力をいただきました、知能情報・生体工学研究室の皆様にも厚く感謝致します。

# 付録

## 日本語の品詞の活用表

基本系・タ系それぞれの語尾の変化を表 5.1・表 5.2 にまとめた。表中の品詞の定義は次の通りである。どの品詞の活用規則にもあてはまらない特殊な活用をする単語は、品詞の項目の一番下に示している。「タグ」の項目は作文データを分析するにあたって誤りの部分につけたタグの、活用語尾との対応を示している。

母音動詞 語幹の末尾が母音で終わる動詞

例：変える (kae)・着る (ki)

動詞性接尾辞 他の語の語幹に接続して、動詞を派生する接辞

(語幹接続)

動詞の語幹や他の動詞性接尾辞に接続するもの

例：させ・られ・な

(連用形接続・母音語幹)

動詞の連用形に接続するもののうち、語幹の末尾が母音で終わるもの

例：あり「うる」・あり「えた」

子音動詞 語幹の末尾が子音で終わる動詞

例：帰る (kaer)・切る (kir)

イ形容詞 名詞を修飾するとき、末尾が「い」になる形容詞

例：美しい・楽しい

イ形容詞系助動詞 イ形容詞と同じように活用する助動詞

例：～である「らしい」

イ形容詞性接尾辞 他の語の語幹に接続して、イ形容詞を派生する接辞

例：男「らしい」・たい・ばい

ナ形容詞 名詞を修飾するとき、末尾が「な」になる形容詞

例：豊かな・幸せな

ナ形容詞系助動詞 ナ形容詞と同じように活用する助動詞

例：ような・みたいな

ナ形容詞性接尾辞 他の語の語幹に接続して、ナ形容詞を派生する接辞

例：～的な・がちな・そうな

判定詞 名詞と結合して述語を作る語

例：だ・である・です

判定詞系助動詞 判定詞と同じように活用する助動詞

例：はずだ・つもりだ・のだ・わけだ・ものだ

## 誤りと正解の回数

留学生の作文データに見られる誤りの回数を、対応する正しい活用ごとに数えあげて表 5.3・表 5.4 にまとめた。

## タグが付けられた誤りのパターン

実際にタグが付けられた誤りのパターンをタグごとにまとめたのが表 5.10～5.17 である。表中の「e」はその箇所に文字列がないことを示しており、これが誤りとされている場合は文字列の脱落誤りを、正解とされている場合は文字列の挿入誤りを表す。

## 各ルールの性能

最大 F 値が 0.2 以上の各ルールの適合率・再現率を表 5.5～5.9 に示す。

表 5.1: 基本系語尾の活用規則

基本系語尾		基本形	命令形	意志形	条件形	連用形	連体形	
タグ		k	m	i	z	y	t	
母音動詞 動詞性接尾辞 (語幹接続) (連用形接続・母音語幹)		kb	る	ろ	よう	れば	(無)	-
子音動詞 動詞性接尾辞 (連用形接続 ・子音語幹)	s	ks	す	せ	そう	せば	し	-
	k		く	け	こう	けば	き	-
	g		ぐ	げ	ごう	げば	ぎ	-
	m		む	め	もう	めば	み	-
	n		ぬ	ね	のう	ねば	に	-
	b		ぶ	べ	ぼう	べば	び	-
	t		つ	て	とう	てば	ち	-
	r		る	れ	ろう	れば	り	-
w	う	え	おう	えば	い	-		
イ形容詞 イ形容詞系助動詞 イ形容詞性接尾辞		ki	い	-	-	ければ	く(ず)	(な)
ナ形容詞 ナ形容詞系助動詞 ナ形容詞性接尾辞 判定詞 判定詞系助動詞	だ	kn	だ	-	-	-	に	な(の)
	である		る	-	-	れば	り	-
	です		す	-	-	-	-	-
「くる」		kk	くる	こい	こよう	くれば	き(こず)	-
「する」		ks	する	しろ	しょう	すれば	し(せず)	-
「ます」		km	ます	-	ましょう	-	-	-

表 5.2: タ系語尾の活用規則

タ系語尾		基本形	条件形	連用形(テ系)	連用形(タリ系)	
タグ		k	z	t	r	
母音動詞 動詞性接尾辞 (語幹接続) (連用形接続・母音語幹)		tb	た	たら	て	たり
子音動詞 動詞性接尾辞 (連用形接続 ・子音語幹)	s(si)	ts1	した	したら	して	したり
	k(i)	ts2	いた	いたら	いて	いたり
	g(i)	ts3	いだ	いだら	いで	いだり
	m,n,b(n)	ts4	んだ	んだら	んで	んだり
	t,r,w(t)	ts5	った	ったら	って	ったり
イ形容詞 イ形容詞系助動詞 イ形容詞性接尾辞		ti	かった	かったら	くて(いで・ずに)	かったり
ナ形容詞 ナ形容詞系助動詞 ナ形容詞性接尾辞 判定詞 判定詞系助動詞	だ	tn1	った	ったら	で	ったり
	である	tn2	った	ったら	って	ったり
	です	tn3	した	したら	して	したり
「くる」		tkr	きた	きたら	きて	きたり
「する」		tsr	した	したら	して	したり
「ます」		tms	した	したら	して	したり
「ません」		tmn	でした	-	で	-

表 5.3: 基本系語尾の誤り・正解の数

基本系語尾		基本形	命令形	意志形	条件形	連用形	連体形	
タグ		k	m	i	z	y	t	
母音動詞 動詞性接尾辞 (語幹接続) (連用形接続・母音語幹)		kb	4/129	0/1	0/2	0/0	27/278	-
子音動詞 動詞性接尾辞 (連用形接続 ・子音接続)	s	ks	1/3	1/1	0/0	0/0	1/9	-
	k		0/6	0/0	0/0	0/0	1/10	-
	g		0/0	0/0	0/0	0/0	0/0	-
	m		0/0	0/0	0/1	0/0	0/10	-
	n		0/0	0/0	0/0	0/0	0/0	-
	b		4/4	0/0	0/1	0/0	0/20	-
	t		0/2	0/0	0/0	0/0	0/4	-
	r		2/78	0/0	0/2	0/0	4/117	-
w	0/11	0/0	0/0	0/0	1/42	-		
イ形容詞 イ形容詞系助動詞 イ形容詞性接尾辞		ki	5/64	-	-	0/0	4/14	0/2
ナ形容詞 ナ形容詞系助動詞 ナ形容詞性接尾辞 判定詞 判定詞系助動詞	だ	kn	0/5	-	-	-	0/2	0/6
	である		0/0	-	-	0/0	0/0	-
	です		1/93	-	-	-	-	-
「くる」		kkkr	0/1	0/0	0/0	0/0	0/8	-
「する」		ksr	0/10	0/0	0/1	0/1	1/64	-
「ます」		kms	2/244	-	0/0	-	-	-

表 5.4: タ系語尾の誤り・正解の数

タ系語尾			基本形	条件形	連用形(テ系)	連用形(タリ系)	
			タグ	k	z	t	r
母音動詞 動詞性接尾辞 (語幹接続) (連用形接続・母音語幹)			tb	20/93	1/5	42/163	0/5
子音動詞 動詞性接尾辞 (連用形接続 ・子音語幹)	s	ts1	2/8	0/1	4/13	1/0	
	k	ts2	0/3	0/0	2/10	0/0	
	g	ts3	0/0	0/0	1/2	0/0	
	m,n,b	ts4	1/7	0/0	15/59	0/1	
	t,r,w	ts5	17/83	0/0	75/238	7/6	
イ形容詞 イ形容詞系助動詞 イ形容詞性接尾辞			ti	1/18	0/0	2/10	0/0
ナ形容詞 ナ形容詞系助動詞 ナ形容詞性接尾辞 判定詞 判定詞系助動詞	だ	tn1	0/1	0/1	0/0	0/0	
	である	tn2	0/0	0/0	0/0	0/0	
	です	tn3	1/13	0/0	0/0	0/0	
「くる」			tkr	0/8	0/0	0/3	0/0
「する」			tsr	1/12	0/1	2/27	0/1
「ます」			tms	7/325	1/0	0/2	0/0
「ません」			tmn	2/6	-	0/0	-

表 5.5: 基本系語尾の誤り

タグ	誤り/正解	誤り数
kbk	e/る	2
	す/る	1
	って/る	1
kby	り/e	19
	い/e	5
	る/e	3
ksk	る/ぶ	2
	び/ぶ	1
	む/ぶ	1
	う/る	1
	す/る	1
	した/す	1
ksm	ろう/せ	1
ksy	させ/し	1
	く/き	1
	る/り	1
	え/り	1
	み/り	1
	れ/り	1
	e/い	1
kik	e/い	3
	る/い	1
	の/い	1
kiy	くて/く	3
	い/く	1
knk	ます/です	1
kmsk	です/ます	2

表 5.6: 夕系語尾の誤り (1)

タグ	誤り/正解	誤り数
tbk	った/た	14
	って/た	2
	る/た	1
	て/た	1
	だ/た	1
	で/た	1
tbz	たが/たら	1
tbt	って/て	24
	で/て	3
	んで/て	3
	e/て	2
	る/て	2
	れて/て	2
	わ/て	1
	ね/て	1
	こ/て	1
	いて/て	1
	むね/て	1
っで/て	1	
ts1k	た/した	1
	った/した	1
ts1t	って/して	3
	て/して	1
ts1r	うたり/したり	1
ts2t	けて/いて	1
	って/いて	1
ts3t	ねて/いで	1
ts4k	んで/んだ	1

表 5.7: タ系語尾の誤り (2)

タグ	誤り/正解	誤り数	
ts4t	んて/んで	4	
	び/んで	2	
	で/んで	1	
	e/んで	1	
	て/んで	1	
	むん/んで	1	
	して/んで	1	
	っで/んで	1	
	びて/んで	1	
	びで/んで	1	
	びって/んで	1	
	ts5k	た/った	9
		だ/った	4
る/った		1	
いた/った		1	
した/った		1	
りた/った		1	
ts5t	て/って	53	
	いて/って	4	
	りて/って	4	
	で/って	3	
	んで/って	3	
	った/って	3	
	る/って	1	
	っ/って	1	
	んて/って	1	
	っで/って	1	
	られて/って	1	
	るして/って	1	

表 5.8: タ系語尾の誤り (3)

タグ	誤り/正解	誤り数
ts5r	る/ったり	4
	たり/ったり	2
	りたり/ったり	1
tik	い/かった	1
tit	かって/くて	1
	いて/いで	1
tn3k	てした/でした	1
tmsk	ました/ました	2
	ま/ました	1
	また/ました	1
	まだ/ました	1
	てした/ました	1
	でした/ました	1
tmsz	ま/ましたら	1
tsrk	な/した	1
tsrt	e/して	2

表 5.9: 助詞の誤り (1)

タグ	誤り/正解	誤り数
z	に/を	15
	が/を	14
	e/を	9
	は/を	3
	の/を	3
	も/を	1
	え/を	1
	e/に	11
	で/に	7
	を/に	4
	と/に	2
	が/に	2
	き/に	1
	e/と	7
	を/と	1
	に/と	1
	e/の	21
	に/の	5
	e/が	23
	は/が	9
	を/が	6
	に/が	4
	の/が	1

表 5.10: 助詞の誤り (2)

タグ	誤り/正解	誤り数
z	で/から	1
	e/から	1
	に/で	26
	e/で	7
	を/で	1
	が/で	1
	へ/で	1
	e/は	16
	が/は	14
	を/は	4
	に/は	1
	の/は	1
	も/は	1
	と/も	1
	は/も	1
	でも/も	1
	の/e	19
	は/e	11
	に/e	9
	が/e	4
で/e	4	
と/e	3	
を/e	2	

表 5.11: 濁音 (脱落) の誤り

タグ	誤り/正解	誤り数
+	ほ/ぼ	20
	と/ど	7
	ふ/ぶ	7
	か/が	5
	け/げ	2
	し/じ	2
	た/だ	2
	ひ/び	2
	こ/ご	1
	つ/づ	1
	て/で	1
	ひ/ぴ	1
	け/げ	1

表 5.12: 濁音 (挿入) の誤り

タグ	誤り/正解	誤り数
-	だ/た	9
	で/て	6
	が/か	4
	ぞ/そ	4
	ぴ/ひ	3
	ず/す	2
	ど/と	2
	ば/は	2
	ぎ/き	1
	ご/こ	1
	ぢ/ち	1
	づ/つ	1

表 5.13: 基本系語尾の誤り検出ルールのパフォーマンス

タグ	ルール	通常 1 年			通常 1 年+逆向き 1 年		
		F 値	適合率	再現率	F 値	適合率	再現率
kbk	する	0.500	33%(1/3)	100%(1/1)	0.500	33%(1/3)	100%(1/1)
kby	り e	0.844	73%(19/26)	100%(19/19)	0.826	70%(19/27)	100%(19/19)
ksk	び ぶ	0.500	33%(1/3)	100%(1/1)	0.400	25%(1/4)	100%(1/1)
	うる	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
ksm	ろう せ	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
ksy	させ し	0.500	33%(1/3)	100%(1/1)	0.500	33%(1/3)	100%(1/1)
	く き	0.667	50%(1/2)	100%(1/1)	0.667	50%(1/2)	100%(1/1)
	え り	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	み り	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	れ り	0.333	20%(1/5)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
kik	の い	1.000	100%(1/1)	100%(1/1)	0.400	25%(1/4)	100%(1/1)
kiy	くて く	0.667	50%(3/6)	100%(3/3)	0.667	67%(2/3)	67%(2/3)
knk	ます です	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)

表 5.14: タ系語尾の誤り検出ルール性能 (1)

タグ	ルール	通常 1 年			通常 1 年+逆向き 1 年		
		F 値	適合率	再現率	F 値	適合率	再現率
tbk	った た	0.800	91%(10/11)	71%(10/14)	0.815	85%(11/13)	78%(11/14)
	って た	0.800	67%(2/3)	100%(2/2)	0.800	67%(2/3)	100%(2/2)
	だ た	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	で た	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
tbz	たが たら	0.667	50%(1/2)	100%(1/1)	0.667	50%(1/2)	100%(1/1)
tbt	って て	0.826	86%(19/22)	79%(19/24)	0.852	77%(23/30)	96%(23/24)
	んで て	0.500	33%(3/9)	100%(3/3)	0.500	33%(3/9)	100%(3/3)
	る て	0.333	20%(2/10)	100%(2/2)	0.308	18%(2/11)	100%(2/2)
	れて て	1.000	100%(2/2)	100%(2/2)	1.000	100%(2/2)	100%(2/2)
	わ て	0.400	25%(1/4)	100%(1/1)	0.400	25%(1/4)	100%(1/1)
	ね て	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	こ て	0.500	33%(1/3)	100%(1/1)	0.500	33%(1/3)	100%(1/1)
	いて て	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	むれ て	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	っで て	0.500	33%(1/3)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
ts1k	った した	0.286	17%(1/6)	100%(1/1)	0.286	17%(1/6)	100%(1/1)
ts1r	うたり したり	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
ts2t	けて いて	0.500	33%(1/3)	100%(1/1)	0.333	20%(1/5)	100%(1/1)
	って いて	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
ts3t	れて いで	0.400	25%(1/4)	100%(1/1)	0.400	25%(1/4)	100%(1/1)
ts4k	んで んだ	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
ts4t	んて んで	1.000	100%(4/4)	100%(4/4)	1.000	100%(4/4)	100%(4/4)
	び んで	0.667	50%(2/4)	100%(2/2)	0.667	50%(2/4)	100%(2/2)
	で んで	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	むん んで	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	っで んで	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	びて んで	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	びで んで	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	びって んで	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)

表 5.15: タ系語尾の誤り検出ルール of 性能 (2)

タグ	ルール	通常 1 年			通常 1 年+逆向き 1 年		
		F 値	適合率	再現率	F 値	適合率	再現率
ts5k	た った	0.250	17%(4/23)	44%(4/9)	0.368	24%(7/29)	78%(7/9)
	いた った	0.667	50%(1/2)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	した った	0.500	33%(1/3)	100%(1/1)	0.500	33%(1/3)	100%(1/1)
	りた った	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
ts5t	て って	0.783	72%(45/62)	85%(45/53)	0.804	76%(45/59)	85%(45/53)
	いて って	0.889	80%(4/5)	100%(4/4)	0.889	80%(4/5)	100%(4/4)
	りて って	1.000	100%(4/4)	100%(4/4)	0.889	80%(4/5)	100%(4/4)
	で って	0.667	50%(3/6)	100%(3/3)	1.000	100%(3/3)	100%(3/3)
	んで って	0.800	100%(2/2)	67%(2/3)	0.800	67%(2/3)	100%(2/2)
	った って	0.750	60%(3/5)	100%(3/3)	0.800	100%(2/2)	67%(2/3)
	んて って	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	っで って	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	るして って	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
ts5r	たり ったり	1.000	100%(2/2)	100%(2/2)	1.000	100%(2/2)	100%(2/2)
	りたり ったり	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
tit	かって いで	0.667	50%(1/2)	100%(1/1)	0.667	50%(1/2)	100%(1/1)
	いて いで	1.000	100%(1/1)	100%(1/1)	0.667	50%(1/2)	100%(1/1)
tn3k	てした でした	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
tmsk	ました ました	1.000	100%(2/2)	100%(2/2)	1.000	100%(2/2)	100%(2/2)
	ま ました	0.250	14%(1/7)	100%(1/1)	0.400	25%(1/4)	100%(1/1)
	また ました	0.667	50%(1/2)	100%(1/1)	0.500	33%(1/3)	100%(1/1)
	まだ ました	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	てした ました	0.500	33%(1/3)	100%(1/1)	0.500	33%(1/3)	100%(1/1)
tmsz	ま ましたら	0.200	11%(1/9)	100%(1/1)	0.400	25%(1/4)	100%(1/1)

表 5.16: 濁音 (脱落) の誤り検出ルール の性能

		通常 1 年			通常 1 年+逆向き 1 年		
タグ	ルール	F 値	適合率	再現率	F 値	適合率	再現率
+	ほ ぼ	0.923	95%(18/19)	90%(18/20)	0.976	95%(20/21)	100%(20/20)
	と ど	0.667	63%(5/8)	71%(5/7)	0.667	63%(5/8)	71%(5/7)
	ふ ぶ	1.000	100%(7/7)	100%(7/7)	1.000	100%(7/7)	100%(7/7)
	か が	0.600	60%(3/5)	60%(3/5)	0.600	60%(3/5)	60%(3/5)
	た だ	0.667	50%(2/4)	100%(2/2)	0.667	50%(2/4)	100%(2/2)
	ひ び	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	こ ご	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	つ づ	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	ひ ぴ	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	け げ	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)

表 5.17: 濁音 (挿入) の誤り検出ルール の性能

		通常 1 年			通常 1 年+逆向き 1 年		
タグ	ルール	F 値	適合率	再現率	F 値	適合率	再現率
-	だ た	0.783	64%(9/14)	100%(9/9)	0.720	56%(9/16)	100%(9/9)
	で て	0.571	40%(6/15)	100%(6/6)	0.667	56%(5/9)	83%(5/6)
	が か	0.444	40%(2/5)	50%(2/4)	0.500	50%(2/4)	50%(2/4)
	ぞ そ	1.000	100%(4/4)	100%(4/4)	1.000	100%(4/4)	100%(4/4)
	ぴ ひ	1.000	100%(3/3)	100%(3/3)	1.000	100%(3/3)	100%(3/3)
	ず す	1.000	100%(2/2)	100%(2/2)	1.000	100%(2/2)	100%(2/2)
	ど と	0.800	67%(2/3)	100%(2/2)	0.500	33%(2/6)	100%(2/2)
	ば は	1.000	100%(2/2)	100%(2/2)	1.000	100%(2/2)	100%(2/2)
	ぎ き	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	こ こ	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	ぢ ち	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	づ づ	1.000	100%(1/1)	100%(1/1)	1.000	100%(1/1)	100%(1/1)
	ぶ ぶ	1.000	100%(1/1)	100%(1/1)	0.667	50%(1/2)	100%(1/1)

## 参考文献

- [1] 劉 軼, 加藤伸隆, 馬目知徳, 伊丹誠, 伊藤紘二, ”状況と機能に応じた日本語の学習を支援するシステム”, 言語処理学会第3年次大会発表論文集, pp.173-176(1997).
- [2] 掛川淳一, 神田久幸, 藤岡英太郎, 伊丹誠, 伊藤紘二, ”日本語学習支援システムにおける作文診断処理系の試作”, 言語処理学会第6年次大会発表論文集, pp.163-166(2000).
- [3] 李相穆, 佐藤滋, 上原聡, ”ウェブ上での日本語書取学習支援システムの開発”, 言語処理学会第7年次大会発表論文集, pp.441-444(2001).
- [4] 森輝彦, ”統計的言語モデルを用いた日本語学習者向けスペルチェッカ”, 筑波大学大学院修士課程 理工学研究科修士論文 (2002).
- [5] 森輝彦, 山本幹雄, ”ngram モデルと単純な誤り検出ルールを用いた日本語スペルチェッカ”, 言語処理学会第8年次大会発表論文集, pp.140-143(2002).
- [6] 益岡隆志, 田窪行則, ”基礎日本語文法-改訂版-”, くろしお出版, pp.8-72(1992).
- [7] 白木伸征, 黒橋禎夫, 長尾眞, ”大量の平仮名列登録による日本語スペルチェッカーの作成”, 言語処理学会第3年次大会発表論文集, pp.445-448(1997).
- [8] 石場正大, 竹山哲夫, 青木恒夫, 兵藤安昭, 池田尚志, ”品詞 N-gram 統計情報を用いた日本語文書における誤り検出法について”, 電子情報通信学会技術研究報告, NLC97-49, pp.43-48(1997).
- [9] 新納浩幸, ”平仮名 N-gram による平仮名列の誤り検出とその修正”, 情報処理学会論文誌, Vol.40, No.6, pp.2690-2698(1999).
- [10] 荒木哲郎, 橋本憲久, 池原悟, ”スキップマルコフ連鎖モデルを用いた日本語の誤り検出, 訂正方法”, 電子情報通信学会技術研究報告, NLC99-78, pp.1-8(2000).

- [11] 荒木哲郎, 池原悟, 佐藤政伸, 榮代正男, ”マルコフ連鎖モデルを用いた日本語文の置換型, 挿入型及び脱落型誤りの検出・訂正法の改善”, 電子情報通信学会技術研究報告,NLC99-78,pp.1-8(2000).
- [12] 三品拓也, 山本幹雄, ”確率的 LSA と trigram モデルを用いた日本語スペルチェック”, 言語処理学会第 9 年次大会発表論文集,pp.183-186(2003).
- [13] 北研二, ”確率的言語モデル”, 東京大学出版会 (1999).
- [14] R.Rosenfeld, ”The CMU Statistical Language Modeling toolkit and its use in the 1994 ARPA CSR evaluation”, In Proc.ARPA Spoken Language System Technology Workshop, pp.47-50(1995).
- [15] 黒川和也, 宇津呂武仁, ”日本語機能表現呼応可能性規則の作成および機能表現練習問題作成支援における評価”, 言語処理学会第 9 年次大会発表論文集,pp.125-128(2003).