K-009

外国人の初級日本語学習における仮名表記と文法の初歩的誤りの検出方式
Error Detection of Basic Kana Spelling and Grammar in Foreigner's Japanese Language

杉野 勝也† 佐藤俊也† 絹川 博之†
Katsuya Sugino Toshinari Sato Hiroshi Kinukawa

1. はじめに

近年，コンピュータが教育分野で利用されるようになり外国人を対象とした日本語教育においても多く利用されるようになってきた．しかし，外国人日本語学習者が作成した文章を添削するシステムはほとんど見られず，日本語教師等の人手によって添削されているのが現状である．そのため，学習者が独学で文章作成を学習することは困難である．

そこで我々は外国人学習者が独学で文章作成を学習できることを目標として日本語学習支援システムを開発している．第一段階として，対象を初級日本語にしぼり，学習者の作成した文の誤りを検出，訂正する方式を研究している．

2. 外国人の学習する日本語

2.1 初級日本語

本研究では外国人のための初級日本語を研究対象にしているが，ここでの初級日本語とは，以下の通りである．

単語数	約 1000 語
複文	述語は 2 つまで． 例文：明日雨が降ると思います．
重文	接続助詞による接続文． 文は 2 つまで． 例文：今日は忙しいが，明日は暇だ
	動詞の羅列 動詞は 3 つまで 例文：本を読んだり，テレビを見たり，買い物をしたりします．
敬語	一般的な初級日本語では敬語を扱うが，本研究では扱わない．

このレベルの日本語を習得すれば，基本的な語彙や漢字を使って書かれた身近な話題の文章を読んで理解することができ，ややゆっくりと話される会話であれば内容がほぼ理解できる．

2.2 外国人学習の初級日本語

本研究では，実際に日本語を学習している外国人が作成した日本語を収集、解析して研究している[1][2][3][4] [5]．学習者には非漢字圏の学習者もおり，また，初級レベルにおいては正しい振り仮名を身に付けるために漢字を使わずに振り仮名だけで表記することがある．そのため，初級日本語学習者の日本語では振り仮名が多くなり，その中には誤りも多く見られる．また，動詞等の活用の誤りも多く見られる．

†東京電機大学 大学院 情報メディア学専攻
Tokyo Denki University, Graduate School

3. 外国人の初級日本語の誤り

収集した外国人の日本語を解析し，どのような誤りがあるかを調べた．以下に代表的な誤りを示す．

- 振り仮名の誤り
- 活用の誤り
- 指示詞の誤り

これらの中で，振り仮名の誤りが 71.0%を占め一番多かったので，我々はまず，振り仮名の誤りの検出，訂正方式を研究することにした．また，活用の誤りに関しても，検出，訂正を行った．

今回の研究では，意味を考慮しなければ検出できない誤り（「本を食べる」等）は対象外としているが，第二段階において研究開発する．

4. 文章作成支援システム

本研究では第一段階として文章作成支援システムを開発している（図 1 参照）．学習者が文章を入力すると，システムが誤り検出，訂正を行い，学習者に誤りの指摘と正解を提示する．このシステムの主な機能は「誤り検出」と「誤り訂正」で，「誤り検出」では振り仮名の誤り検出以外に文末に接続されている単語の品詞，活用の誤りも検出し，訂正可能な場合は訂正を行う．「誤り訂正」は，振り仮名の訂正を行う．

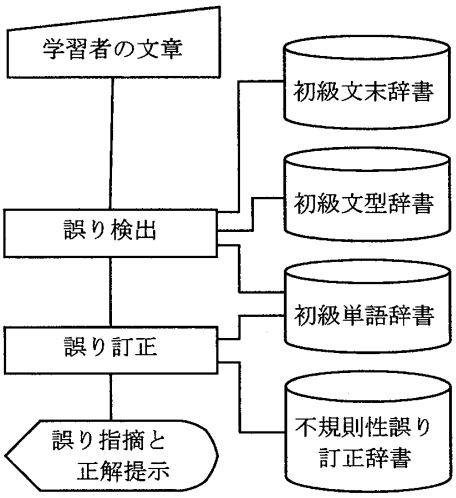

図 1 文章作成支援システム

文章作成支援システムでは以下に示す 4 種類の辞書を使用している．

- 初級文末辞書
  初級日本語でよく使われる文末を登録したもので，学習者が作成した文の文末を正誤判定する際に使用される．

・初級文型辞書
　初級日本語で扱う典型的な文型の形態素解析結果を登録した辞書で，学習者の文を形態素解析したものと比較するのに使われる．
・初級単語辞書
　初級日本語で扱う単語を登録した辞書で，誤り検出，誤り訂正両方で使用され，処理中の単語が正しいか否かを判断するときに利用される．
・不規則性誤り訂正辞書
　この辞書は誤り訂正で使用する辞書で，不規則な誤りを登録したものである．

なお，誤り訂正については別途報告する[4][5]ので，本稿では誤り検出について述べる．

5. 誤り検出方式

本方式では JUMAN[6]による形態素解析を利用して誤り検出を行っている．しかし，学習者が作成した誤りを含んだ文を，そのまま解析すると，単語の境界を誤り，誤解析が多くなることがある．そこで形態素解析の精度向上のために，形態素解析前に前処理を行っている．
（図 2 参照）

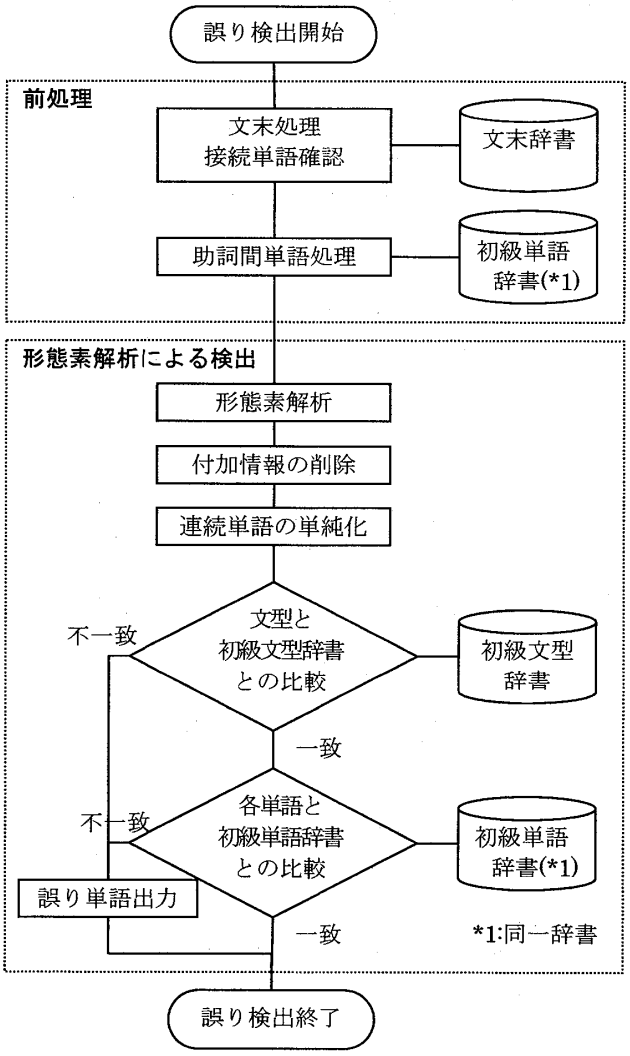

図 2 誤り検出方式

5.1 前処理

前処理では以下の処理を行う．
・文末および接続語の確認
・助詞間単語の確認

5.1.1 文末処理，接続単語確認

初級日本語学習者でも文末表現を間違うことは少なく，文末表現に誤りがあっても推測容易で，訂正可能であることが多い．また，文末に接続される単語の品詞，活用形は決まっている．そこで，文末を調べ，文末表現自体の正誤および文末に接続されている単語の品詞，活用形の正誤を調べる．訂正可能であれば訂正も行う．以下に検出，訂正方式を述べる．

(1) 学習者の文の文末が初級文末辞書と一致するかを調べる．一致する場合は(4)の処理へ移り，不一致の場合は(2)処理を行う．なお，文末の辞書検索は最終文字より前方に行う．

(2) 「誤字が 1 文字ある」場合の訂正を行う．文末の最終文字より辞書検索を行い，一致しない文字が出現した場合，その文字以外が一致するかを調べる．一致するものが辞書に一つだけ存在すれば，それを正しい文末とし，訂正を行う．辞書と一致しない場合，および，一致するものが複数ある場合，(3)の方式で訂正を試みる．

訂正例：いただきませんか
↓
いただけませんか

(3) 「脱字が 1 文字ある」場合の訂正を行う．文末の辞書検索で，一致しない文字が出現した場合，それを脱字と判断し，それ以外が一致するかを調べる．一致するものが辞書に一つだけ存在すれば，それを正しい文末とし，訂正を行う．一致しない，複数候補がある場合は訂正不可能と判断し，一致しなかった文字を誤りであると学習者に指摘する．

訂正例：とおも＿ています
↓
とおもっています

なお，訂正不可の場合は，この後の形態素解析等の誤り検索および誤り訂正は行わず，処理終了とする．

(4) 文末が初級文末辞書と一致した，または(2)(3)の処理により訂正できた場合，文末の接続単語が初級単語辞書と一致するか調べる．一致する場合は(7)の処理へ移り，一致しない場合は次の処理を行う．なお，接続単語の辞書検索も後方検索で行う．

(5) 文末訂正方式(2)(3)と同様の方式で接続単語を訂正する．訂正できない場合は，次の訂正処理を試みる．

(6) 「余分な文字がある」場合の訂正を行う．辞書検索で，一致しない文字が出現した場合，その文字を削除して検索を続ける．辞書に一致する単語が一つだけ存在すれば，それを正しい単語とし，訂正を行う．一致しない，あるいは複数候補がある場合は訂正不可能と判断し，学習者に指摘する．その場合，これ以降の誤り検索，誤り訂正は行わない．

訂正例：しょうくじする
↓
しょくじする

(7) 接続単語が初級単語辞書と一致した，または(5)(6)の処理により訂正できた場合は，接続単語の品詞が文末と接続できるものか，否かを調べる．接続不可である場合は，その誤りを学習者に指摘し，訂正は行わない．なお，誤りの有無に関わらず次の処理を行う．

誤り例：水泳ことがあります

(8) 次に接続単語の活用形が文末と接続できる形であるかを調べる．活用形が接続不可の場合は，その誤りを学習者に指摘し，正しい活用形に訂正して次の前処理 5.1.2 を行う．

訂正例：お酒を飲むすぎました
↓
お酒を飲みすぎました

なお，活用形の訂正は，初級単語辞書によって行う．初級単語辞書には各活用形が登録されているので，活用形の変換は容易である．

5.1.2 助詞間単語処理

初級日本語では，各名詞文節に必ず助詞がついていることが多い．

弟の学校に英語の先生がいます（助詞間は 1 単語）

以下の例文のように自立語のみの文節を含む場合もある．

私はきのう 買った 本を読む(助詞間に 3 単語ある)

そこで助詞の前に自立語があると仮定し，それらが基本単語辞書と一致するかを調べる．一致しない場合は，訂正を試みる．以下に誤り検出，訂正方式を述べる．

(1) 5.1.1 で調べた文末の接続単語の一つ前の文字を検出対象文字とする．

(2) 検出対象文字が以下に示す助詞と一致するかを調べる．

は，も，の，を，が，に，へ，で，と，や，か
から，まで，までに，より，でも，しか，とか

(3) 助詞と一致した場合，助詞の前の文字が自立語の最終文字として，基本単語辞書と一致するかを後方検索により調べる．一致しなかった場合は，先ほどの助詞を自立語の最終文字として，基本単語辞書と比べる．

(4) 初級単語辞書検索で一致しなければ 5.1.1 の(5)(6)と同様の方法で訂正を行う．訂正ができなかった場合は，これ以降の誤り検索は行わず，5.2 を行う．

(5) 初級単語辞書検索で一致した，または(4)で訂正できた場合は，一致，訂正した単語の前の文字を検出対象文字にして，(2)の処理を行う．これらの処理を文頭まで繰り返す．

5.2 形態素解析による検出

前処理後，形態素解析を行い，その結果と初級文型辞書との比較により誤り検出を行うが，それ以外も誤り検出の精度を上げるために各種処理を行う．以下にそれらの処理を示す．

- 形態素解析
- 付加情報の削除
- 連続名詞の単純化
- 初級構文辞書との比較
- 単語の正誤確認
- 誤り単語出力

5.2.1 形態素解析

前処理後の文を JUMAN[6]によって形態素解析する．

5.2.2 付加情報の削除

形態素解析後，形態素解析結果と初級文型辞書との比較を行うが，単純比較では，正常な結果が得られない場合が多い．以下にその例を示す．

文 1：私は本を読んだ．
文 2：きのう，私は本を読んだ．
文 3：私はゆっくり本を読んだ．
文 4：私は本を読んだよ．

上記の文は，意味はほとんど同じであるが，文型は全て異なる．単純比較をする場合，これらの文型を全て初級文型辞書に登録する必要があり，登録数が非常に多くなり，実用的とは言えない．文 2 から文 4 の下線部は，一種の付加情報で，無くても文の意味が通る単語である．そこで，形態素解析結果より，付加情報を削除することにした．名詞，動詞，形容詞，形容動詞，助詞と文末以外の品詞を削除する．名詞のうち，時を表すもの（例「きのう」）も削除する．この削除によって文 1 から文 4 は全て同一となる．

5.2.3 連続名詞の単純化

以下の 2 つの文は，文法上は同じなので，同じ文型として扱いたいが，形態素解析結果は異なる．

文 1：駅へ行った．
文 2：東京駅へ行った．

「へ行った」の前は，文 1 では名詞 1 語（「駅」）だが，文 2 では名詞 2 語（「東京」+「駅」）である．文 1 も文 2 も同じ文型として扱うため，名詞が連続する場合，それらをまとめ一つの名詞と扱うことにした．

5.2.4 文型と初級文型辞書との比較

5.2.2 および 5.2.3 を行った後の文型と初級文型辞書と比較する．一致した場合は 5.2.5 を行い，一致しなかったときは 5.2.6 を実施する．

5.2.5 各単語と初級単語辞書との比較

5.2.4 において，一致しても誤り単語を含んでいる場合がある．以下にその例を示す．

文 1：**わたしは がくせい です**　（正文）
文 2：**わたしは かくせい です**　（非文）
形態素解析の結果
がくせい：名詞（「学生」と判断）
かくせい：名詞（「覚醒」と判断）

文 2 の「かくせい」は「がくせい（学生）」の誤りであるが，形態素解析すると「覚醒」（名詞）と判断される.すなわち，文 1 の「がくせい」も文 2 の「かくせい」も同じ名詞と判断され，差異がなく，文 2 の誤りを検出することはできない．

「かくせい」を誤りと判断するために，形態素解析により抽出された各単語と初級単語辞書を比較し，一致しない場合，その単語が誤りと判断して 5.2.6 を実施する．

5.2.6 誤り単語出力

誤りと判断した単語を別途報告の誤り訂正処理[4][5]に渡す．

5.3 単語のみ対応

初級日本語学習者は文ではなく，単語だけの正誤確認，特に振り仮名の正誤確認をする場合がある．そこで単語のみを入力した場合でも誤りを検出できるようにした．以下に処理方法を示す．

(1) 学習者の文が以下の場合，単語と判断する．
- 漢字混じりの場合：文字数が 5 文字以下
- 漢字がない場合：　文字数が 10 文字以下

(2) 単語と判断した場合，初級単語辞書と比較し，一致した場合は誤りなし．不一致の場合は，誤りがあるとして，別途報告の誤り方式[4][5]によって訂正を行う．この場合，前処理，形態素解析，後処理は行わない．

6. 実験方法

6.1.1 使用データ

実験で使用したデータは，日本語学校に在籍している留学生のテスト結果より採取した．

6.1.2 データ数

採取した非文：175 文
実験に使用した非文：128 文．

採取した非文は 175 文であるが，その中には中級・上級の文章，対象外の文書もあり除外した．実験に使った非文は以下の誤りを含む初級日本語である．

- 振り仮名の誤り
- 活用の誤り
- 品詞の誤り
- 文型の誤り

7. 実験結果

以下に実験結果を示す．

誤り検出できた文：120 文
誤り検出できなかった文：8 文
誤り検出率：93.75%　(=120/128)

誤り検出できなかった文の例を以下に示す．

彼は家でいない（「に」の間違い）
助詞を別の助詞に間違っても文型上に違いがないため誤り検出できなかった．

ちか〜です（「近い」を「ちか」に間違った）
正文は「近いです」だが，このままでも「地下です」と解釈でき，誤りを検出できない．

いちほん（1 本），よんがつ（4 月）
数詞の正しい読み方を確認していないので，この誤りは検出できない．

8. 考察

誤り検出率は 93.75%であったが，まだまだ改良の余地はある．数詞の読み間違いは，誤り検出も訂正も比較的容易に対応できると思われるので，今後検討していきたい．また，助詞を別の助詞にした誤りは，動詞の結合価の情報が必要であるが，対応可能と思われる．数詞，助詞の誤りはよくあるだけに対応する必要があると思われる．

また，本研究の対象外の文であった誤りであるが「日本の大学」を「日本大学」としているものがあったが，本検出方式では 5.2.3 の処理を行うため，この誤りは検出できない．他にも今回使用したデータには含まれていない誤りで検出できないものがあると思われる．それらについても今後考慮していきたい．

9. おわりに

外国人学習者が作成した日本語の誤りの検出方式を検討したが，今後は，本検出方式と別途報告[4][5]の訂正方式を組み込んだシステムを開発し，評価，改良をしていく予定である．

現在は，外国人学習者のテスト結果をデータにして実験しているが，今後，外国人学習者がシステムを使えるようにしていきた．そのためには，誤りの検出，訂正機能以外にもユーザインタフェース等も考慮する必要がある．

謝辞

本研究を行うにあたり，学校法人 吉岡学園 千駄ヶ谷日本語学校に御協力を頂きました．この場を借りて御礼を申し上げます．

参考文献

[1]杉野勝也,絹川博之, “外国人の初級日本語文の誤り検出方式”, 第 7 回情報科学技術フォーラム(FIT2008)第 3 分冊 pp563-564 (2008).
[2]杉野勝也,絹川博之, “外国人の初級日本語単語の訂正方式”, 情報処理学会 第 71 回全国大会分冊 4 pp615-616 (2009).
[3]杉野勝也,絹川博之, “外国人の初級日本語文における振り仮名の誤り検出”, 第 8 回情報科学技術フォーラム(FIT2009)第 3 分冊 pp591-592 (2009).
[4]佐藤俊也,杉野勝也,絹川博之, “外国人の初級日本語文における振り仮名の誤り訂正”, 第 8 回情報科学技術フォーラム(FIT2009)第 3 分冊 pp593-594 (2009).
[5]佐藤俊也,杉野勝也,絹川博之, “外国人の初級日本語学習における仮名表記誤りの分類と訂正方式”, 第 9 回情報科学技術フォーラム(FIT2010)第 3 分冊 (2010).
[6]黒橋禎夫, 河原大輔, “日本語形態素解析システム JUMAN Version5.1”, 東京大学大学院情報理工学系研究科 (2005).
[7]益岡隆志,田窪行則, “基礎日本語文法 –改訂版-”, くろしお出版 (1992).
[8]益岡隆志, “24 週日本語文法ツアー”, くろしお出版 (1993).
[9]吉川 武時, “日本語文法入門”, アルク(1989).
[10] 千駄ヶ谷日本語教育研究所著, “コミュニケーション日本語 1, 千駄ヶ谷日本語研究所 (1999) .
[11] 千駄ヶ谷日本語教育研究所著, “コミュニケーション日本語 2, 千駄ヶ谷日本語研究所 (1999) .
[12] 千駄ヶ谷日本語教育研究所著, “コミュニケーション日本語 3, 千駄ヶ谷日本語研究所 (1999) .
[13] スリーエーネットワーク編著, “みんなの日本語 初級 I , 本冊”,スリーエーネットワーク(1998.)
[14] スリーエーネットワーク編著, “みんなの日本語 初級 II 本冊”, スリーエーネットワーク(1998.)
[15] スリーエーネットワーク編著, “みんなの日本語 初級 I , 翻訳・文法解説 英語版”,スリーエーネットワーク(1998.)
[16] スリーエーネットワーク編著, “みんなの日本語 初級 II 翻訳・文法解説 英語版”,スリーエーネットワーク(1998.)