福井大学 大学院工学研究科 研究報告 第56巻 2008年3月
Mem. Grad. Eng. Univ. Fukui, Vol. 56 (March 2008)

連文節に基づく辞書にない読みで入力された単語の検出アルゴリズム

－日本語初学者における日本語入力及び日本語学習支援システム－

劉 棋* 張 小剛* 黒岩 丈介* 小倉 久和* 小高 知宏† 白井 治彦**

The Detection Algorithm of a Word Input by a Different Reading Based on Compound Clauses

–Japanese Input in the Japanese Abecedarian and a Japanese Learning Support System–

Ki RYU*, Syougo TYO*, Jousuke KUROIWA*, Hisakazu OGURA*, Tomohiro ODAKA†, Haruhiko SHIRAI**

(Received January 31, 2008)

There are too many Japanese kanjis which have more than 2 ways to read. For a foreign Japanese learner, it is difficult to remember all the ways to read the kanji. When the foreign Japanese learners want to input a word which they don¨t know the way to read, they have to input the kanji by transferring it from another word which include the same kanji but another way to read, or by looking up a paper dictionary. It is unfavorable for them to cost so much in Japanese study. In order to solve these problems, we provide the inputers a Japanese input assistant system to give a suggestion when they input a word which is not defined in the dictionary of the system.In this study, both the matters in the process of studying Japanese and problems of Japanese input system for foreign learners were discussed. In order to solve these problems, we proposed an algorithm for present Japanese input and study-support systems. This algorithm distinguishes the spellings of compound clauses, which cannot be found in dictionaries, and what is more, the validity of it is tested and verified by simulation experiments.

Key words : Foreign Learners, Beginner, Japanese Input, Study-Support System, Algorithm

1. はじめに

現在, 日本への海外からの留学生はますます増加している. 更に, パソコンとインターネットが急速に普及することによって, 日本国内の外国人がパソコンで日本語を入力する機会が増えている. 特に, 留学生の場合は, ほとんどのレポート及び論文をパソコンを用いて日本語で作成する. 日本語を入力する際には, キーボードを用いる. キーボード以外にも, ライトペン, タッチパネル, 音声入力システムなど様々あるが, まだ, キーボードが最も有力な入力装置である. キーボードを用いた日本語入力方式には, 主にカナ文字入力とローマ字入力がある. この二種類とも文字の読み方による入力方式であり, 読みを分かっている人にとっては, 効率的な入力方式と言える [1]. しかしながら, 漢字は, その読み方を知らないと入力することができない. 更に, 漢字系言語を母国語としている人にとっては, 漢字を含む単語個々の母国語での読みは分かるが, その単語全体としての日本

*工学研究科知能システム工学専攻
†工学研究科原子力・エネルギー安全工学専攻
**工学部技術部
*Department of Human and Artificial Intelligent System
†Nuclear Power and Energy Safety Engineering Course
**Department of engineering

語読みを知らない単語が多数存在する．このような場合に，パソコンで所望の漢字単語を入力するためには，通常は母国語もしくは異なる日本語での読みで個々の漢字を入力し，目的の単語入力を完成する．こうのように，非常に労力の要る作業となる．つまり，読み方による入力方式は，留学生日本語学習者にとっては必ずしも便利な入力方式とは言えない．むしろ，日本語の漢字の読み方がよく分からない日本語初学者にとっては非常に難しい方法である．

例えば，「生物」を入力する際に，辞書に記載している読みである「せいぶつ」が分からないものとする．このような状況では，漢字「生」と漢字「物」を読み「なま」と「もの」で別々に入力し，目的の単語「生物」が得られる．また，単語「時間」の読み「じかん」が分かるが，単語「間」の読み「あいだ」を知らないとすると，「時間」を入力して，「時」を消して「間」を残す．このような辞書にない読みで入力するといった不適切な方法は大変手間がかかるだけではなく，その単語あるいは漢字の辞書読みが分からないまま，入力が終ってしまう．漢字の単独の読みあるいは単語の読みを覚えるまで，同じように手間のかかる方法で入力しなければならないため，日本語学習の面でも好ましい状況ではない．

日本語には漢字が複数の読みを持つという特徴があるので，留学生のような日本語初学者は漢字のすべての読みを覚えることは困難である．このような問題を解決するための一手法として，我々は日本語初学者が日本語を入力する際，単語漢字の本来の読みと異なる読みで入力した場合，その単語あるいは漢字の本来の読みを入力者に提示し，正しい読みを学習しながら記憶してもらい，後で同じ単語あるいは漢字を入力する際に，本来の読みで入力できるように支援するシステムを提案し，実装してきた [2]．

しかし，このシステムに使用されているアルゴリズムの判断対象は，最後に入力された二つの単語に制限したことが原因で，辞書にない読みで入力された単語を検出できないこともある．そういった問題を解決するため，本研究では入力された文章のうちの連文節を対象に，連文節の情報に基づく辞書にない読みで入力された単語を検出アルゴリズムを提案する．更に，シミュレーション実験により，本アルゴリズムの有効性を確認する．

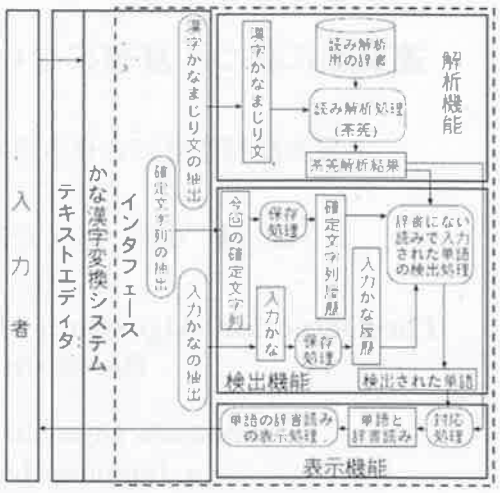

図 1: システムの詳細構成図

2. 日本語初学者における日本語学習支援システム

2.1 我々が実装した学習支援システム

2.1.1 システムの概要と構成

辞書にない読みで入力された日本語単語の検出・読み表示システムは，入力者に入力された単語について辞書読みで入力されたかどうかを確認する．辞書にない読みで入力された単語に対して，辞書読みを入力者に提示する．これにより，その辞書読みを覚える機会を入力者に与え，外国人日本語学習者の日本語入力及び日本語学習を支援するシステムである．

本システムの機能を図 1 に示す．我々が実装したシステムは，図 1 の点線で囲った部分に対応し，テキストエディタ，かな漢字変換システムとのインタフェースを介して動作する．図に示すように本システムは，「読み解析機能 (以下「解析機能」と呼ぶ)」，「辞書にない読みで入力された単語の検出機能 (以下「検出機能」と呼ぶ)」及び「単語の辞書読みの表示機能 (以下「表示機能」と呼ぶ)」，以上 3 つの機能を持つ．本システムは，UNIX 上で動作し，エディタ Emacs とかな漢字変換システム Canna と連携している．システムの主要部分は，Perl で記述した．

本システムでは，辞書にない読みで入力された単語を検出するために，「インタフェース」を介して必要なデータを集め，その一部を「解析機能」で分析し，保存する．そして，これらのデータと現時点までに入力されたデータを用いて，辞書にない読みで入力された単語を「検出機能」で検出し，「表示機能」を通して辞書読みを入力者に表示する．

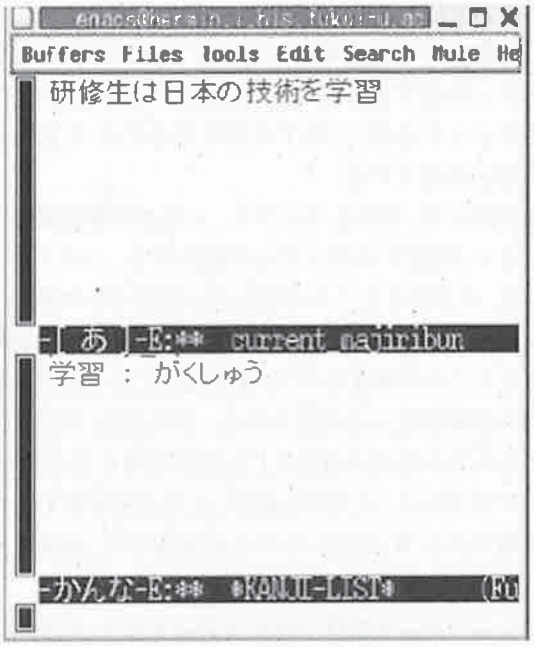

図 2: 単語の辞書読みを表示する画面

本システムを組み込んだ「Emacs」で日本語を入力する際，入力者は，入力した文字列を確定するために，リターンキーを押すと，本システムが起動される．確定された文字列には，辞書にない読みで入力された日本語単語を含んでいれば，図 2.のように「Emacs」の入力ウィンドウが二つに分かれ，辞書にない読みで入力された日本語の単語と，辞書読みが下のウィンドウに表示される．次の文の入力が始まったら，入力のさまたげにならないように下の表示ウィンドウは自動的に閉じられるようにしてある．

図 2 に示すように，本システムが起動されると，まずシステムは漢字仮名交じり文と確定された文字列 (以下，確定文字列と呼ぶ)，及び入力された読み (以下「入力かな」と呼ぶ) を受け取る．漢字仮名交じり文は，テキストエディタに入力されたすべての文字のことである．確定文字列は，入力者が確定操作をしてテキストエディタにわたす文字列とする．この時，キーボードから入力された記号列すべてではなく，漢字かなまじり文を生成するのに使われた文字列だけを用いる．エスケープ記号などの制御記号は含まない．「入力仮名」は，仮名漢字変換前の文字列である．

次に，システムは，それまでに入力され確定された漢字仮名交じり文に対し形態素解析を実行して，形態素の列を生成する．本システムでは，形態素解析に「茶筅」[3] を利用する．以下では，この形態素解析システムより得られた形態素のことを，簡単のため「単語」と呼ぶ．形態素解析の対象となる漢字かなまじり文は，エディタ

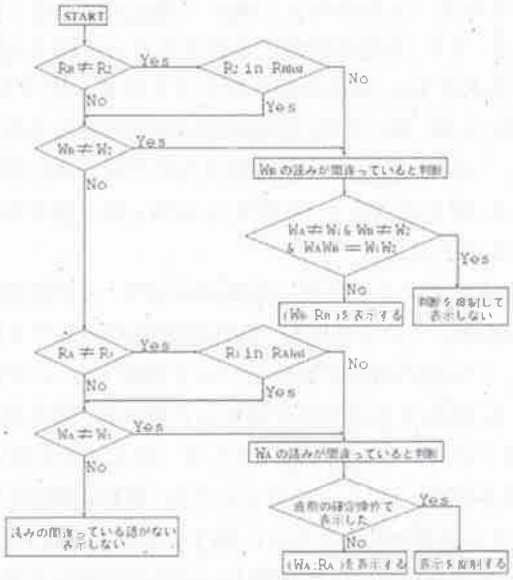

図 3: 最後の 2 語を対象とするアルゴリズムの処理の流れ

に入力されているすべての文字を対象としている．

最後に，システムは，単語ごとの辞書読みを入力かなと比較し，確定文字列の履歴を利用して入力文中の辞書にない読みで入力された単語を検出して，辞書読みを表示する．

2.1.2 システムに使われた辞書にない読みで入力された単語の検出アルゴリズムの説明及び問題点

前述のシステムを実現するためには，辞書にない読みで入力された単語を検出するアルゴリズムが必要となる．そのアルゴリズムは，入力者が，入力し変換した文字列を確定するためにリターンキーを押す (以下，確定操作と呼ぶ) たびに，検出処理を行う．これまでに用していた検出アルゴリズムは，以下の通りである．

この判断アルゴリズムでは，確定操作時における最後の 2 つの語（形態素）を対象とする．具体的には，茶筅が入力文を形態素解析して出力した最後の 2 つの語とその読み，および，入力者の入力した最後の 2 つの確定文字列と入力かな文字列である．

茶筅の出力する最後から 2 つ目の語を「語 A」(以下，記号で W_A と書く)，最後の語を「語 B」(W_B) とする．また，茶筅が出力する語 A と語 B の読みを「読み A」(R_A)，読み B」(R_B) とする．入力者の入力したかな文字列とそれを変換・確定した確定文字列についても，確定操作直後の最後とその 1 つ前を対象とする．最後から 1 つ前のものを「入力かな 1」(R_1)，「確定文字列 1」(W_1)，最

後のものを「入力かな２」(R_2),「確定文字列２」(W_2)とする．また，茶筅の辞書に登録されている語Ａの読みのリストをR_{Alist}，語Ｂの読みのリストをR_{Blist}とする．なお，W_AとW_1，W_BとW_2は基本的には対応するものであるが，この対応関係は一般には入力のプロセスに依存しており，例えば，W_1とW_2がともにW_Bの一部であったりすることがある．

読み判断アルゴリズムを図３に示す．まず最後の語Ｂの部分について判断し，語Ｂの部分が一致する場合には，１つ前の語Ａの部分について判断する．いずれの場合も，読みと入力かなが異なった場合は，例えば単語「明日」について茶筅は読みとして「あした」を出すが，「あす」「みょうにち」も正しいため，茶筅の辞書に登録されている語の読みリストにおける有無も考慮する．つまり，登録されている語Ａあるいは語Ｂの読みをすべて検索し，R_1あるいはR_2とマッチする読みがあるかどうかチェックする．W_BがW_1とW_2の連接W_1W_2と等しい場合は，R_1とR_2の連接R_1R_2のかな文字列についてもチェックする．

しかし，このアルゴリズムは最も後で入力された辞書にない読みで入力された一つの単語のみ表示するので，入力者にとって覚えやすいが，最後の２回の確定情報しか判断対象としないので，辞書にない読みで入力された単語を検出できないことがある．日本語初学者は日本語入力際，文法的に不適切な方法で入力することがよくある．格助詞とその後の文を区切りせず連続で入力することが，その代表例の一つとなる．例えば，「魚をさばく」を入力する際，さかな」の読み方を分からずに「きんぎょ（金魚）」「をさばく」の区切りで入力してから「金魚」の「金」を削除し，目的の入力を完了する．最後の２語を対象とする検出アルゴリズムでは判断対象を入力直後の最後の二つの単語に制限したため，W_A，W_Bの語の境界がW_1，W_2の境界とずれてしまい，システムは正しく動作しない．

2.2 本研究で提案した連文節に基づく辞書にない読みで入力された単語の検出アルゴリズム

これまでの検出手法としては，入力直前の文字のみ対象とし，仮名変換処理ごとに検出処理が行われるが，検出判断対象を最後の二つの単語にに制限したことが原因で，辞書にない読みでの入力を検出できないこともある．このような問題を解決するために，今回は，入力された文（連文節）に基づく検出アルゴリズムを提案する．

入力者が日本語文書を入力する際，いろいろな方法が考えられる．単語を入力する度に確定する入力方法もあるし，単語とその後の助詞を入力し，変換してから確定する入力方法もあるし，長いかな列を入力して一度に変換して確定する方法もある．また，読みの分からない単語あるいは漢字を入力する時，ほかの単語を借りて，入力することもある．このアルゴリズムでは，これらの状況に対応可能とする．

具体的には，次のようにする．入力者が確定操作をするごとに，検出アルゴリズムは動作する．このアルゴリズムは，入力するところと最も近い読点からの間の文を対象にし判断を行う．入力者の習慣によって，確定操作をするところが異なるので，確定操作による入力かなの分割は意味がないと考えられる．そのため，その文のすべての入力かな列を連接し，入力かな列Ｃを作成する．アルゴリズムは，入力かな列Ｃ，入力文を解析した結果の単語リストＷと読みリストＲを用いて，辞書にない読みで入力された単語を検出する．

アルゴリズムの処理手続きを要約すると，以下のような流れになる．

1. 句読点を使って，Emacs と Canna それぞれから入力された文章から最後の文（対象文Ａ）（入力した最新文）と入力読み（Ｃ）を取り出す．
2. 対象文Ａを ChaSen で解析し，単語リスト（Ｗ）と辞書読みリスト（Ｒ）を生成
3. Ｗの中の連続仮名を連結し，ひとつの文字列にする．（連続仮名が存在しない場合，単独の仮名文字をマッチング）
4. 作業３で作成した仮名文字（列）の中で，一番長いかつＣの中で同じ仮名文字（列）が一つしかない仮名文字（列）をマッチングし，それを使ってＡとＣを分割．（分割する仮名文字（列）は文の最初と最後にある場合，文から分割する仮名文字（列）を省く）
5. 作業４で分割された文の各節の中で一番長いかつＣの中で一つしかない仮名文字（列）をマッチングし，それを使ってさらにＡとＣを分割する（二次分割）
6. 分割された各節の中に仮名文字（列）が存在しないまであるいは文の中に仮名文字（列）が存在するが，対応するＣの中に同じ仮名文字（列）が一つではなく複数存在するまで，分割し続ける
7. 分割されたＡの各節の中に対応するＣの各節の中の各単語の辞書読みＲをマッチングする．なかった場合，正しくない読みで入力されたと判断し，その単語と正しい読みを入力者に表示する

例えば, 入力者は「強盗は車を盗み出してから速やかに逃走した」を入力しようとして, アルゴリズムは以下のような手続きで動く.

1. 句点で文を区きり, 「強盗は車を盗み出してから速やかに逃走した」の文を取り出し, 対応する入力読み (C), 辞書読みリスト (R) と単語リスト (W) を検出

   Canna から読み込んだ入力読み (C) : つよいぬすむはしゃをぬすみだしてからはやいやかににげるはしるした. ＊：□中の文字は入力後削除されるもの Emacs から読み込んだ入力結果を Chasen で解析し辞書読みリストを生成 (R) :「ごうとう」, 「は」, 「くるま」, 「を」, 「ぬすみだし」, 「てか」, 「ら」, 「すみやか」, 「に」, 「とうそう」, 「し」, 「た」 Emacs から読み込んだ入力結果を Chasen で解析し単語リストを生成 (W) : 「強盗」, 「は」, 「車」, 「を」, 「盗み出し」, 「てか」, 「ら」, 「速やか」, 「に」, 「逃走」, 「し」, 「た」.

2. 「てか」と「ら」, 「し」と「た」を連結し, 「てから」, 「した」にする

3. W 文の中の一番長いかつ C の中に一つしかない純かな文字列「てから」で対象文 W,C を分割. W は「強盗」, 「は」, 「車」, 「を」, 「盗み出し」(前半 W1 とする) と「速やか」, 「に」, 「逃走」, 「した」(後半 W2 とする) の 2 部分となる. C は「つよいぬすむはしゃをぬすみだし」(前半 C1 とする) と「はやいやかににげるはしるした」(後半 C1 とする) の 2 部分となる

4. 分割された C の各節の中で一つしかない純かなを使って, さらに C と W を分割. 分割された W 1 の中に C1 に一つしかない純かな「は」, 「を」を使って, W 1 を「強盗」, 「車」, 「盗み出し」と分割し, 対応的に C1 を「つよいぬすむ」, 「しゃ」と「ぬすみだし」と分割. 同様に, W2 中の C2 に一つしかない純かな「に」, 「した」を使って, W2 を「速やか」, 「逃走」, 「した」と分割し, 対応的に「はやいやか」, 「にげるはしる」と分割

5. 4 で分割された各漢字単語の入力読み (C) と対応的な単語リスト (W) の辞書読み (R) と比較し, 同じかどうかを判断する

   W 対応 C 比較 R 結果

   「強盗」→「つよいぬすむ」⇔「ごうとう」 ×

   「車」 → 「しゃ」 ⇔「くるま」 ×

   「盗み出し」→「ぬすみだし」⇔「ぬすみだし」○

   「速やか」→「はやいやか」 ⇔「すみやか」×

   「逃走」 →「にげるはしる」⇔「とうそう」×

   マッチングした結果としては「盗み出し」以外の単語は全部辞書読みではない読みで入力したことが考えられる.

6. C に対応する節に R でないかなであれば, 正しくない読み方で入力したと判断し, その漢字と正しい読みを表示する.

   ごうとう くるま すみやか とうそう

   強盗 , 車 , 速やか , 逃走

このアルゴリズムは入力文の前から単語単位で検出する仕組みなので, より自然的な検出手法となる. また, 入力文が長くても短く分割することができ, 検出精度が向上することができると考えられる.

3. 実験

実際日本語初学者が日本語入力する際, 不適切な入力方法での入力はどの程度で利用されるのか, またこのような入力方法は人力者の日本語能力とどうのような関係があるのかを, 実験を通して明らかにする. また, 提案したアルゴリズムの有効性を確認するため, 漢字圏出身の外国人日本語学習者である本学の留学生を対象に評価実験を行った. 実験目的は次の 2 点である.

(1) 日本語初学者を対象に日本語を入力する過程を観察し, 不適切な入力方法をどの程度行うのか, 更には, 不適切な入力方法は入力者の日本語能力とどうのような関係があるのかを明らかにする.

(2) 本研究で提案した連文節に基づく辞書にない読みで入力された単語を検出アルゴリズムの検出機能の効果を示す.

3.1 実験方法

5 人の外国人留学生学習者である本学の留学生の協力を得て, 300 − 400 文字で 50 句程度のテキストを入力する実験を行った. 5 人の日本語学習時間 100 時間から 1200 時間までそれぞれなので, 日本語能力も差がある.

実験では, 各入力者がテキストエディタで実験入力文を入力する. 入力する際, 人力者はどのような手法で入力したのかを観察し記録する. 図 4 で示している入力例文のように, 下線が引いてある文は不適切な入力で人力される可能性がある文であり, 不適切な入力方法で入力した場合はその回数を計算して, 最終的に入力者の全員

1. 日本人の朝食は主に和食、洋食となる
2. 日本ではごみを捨てるのに、一定のルールがある
3. 自動販売機はいたるところにある
4. この映画は最近話題となっている
5. 我々にとって地球は唯一のものとなる
6. 日本の街はどこを歩いても清潔と、外国人観光客は驚く
7. 来日した外国人は外国人登録の手続きをするべきである
8. 特筆すべき日本の店に「100円ショップ」がある

...

全文に対して不適切な入力は何割あったのかを計測

図 4: 入力文例

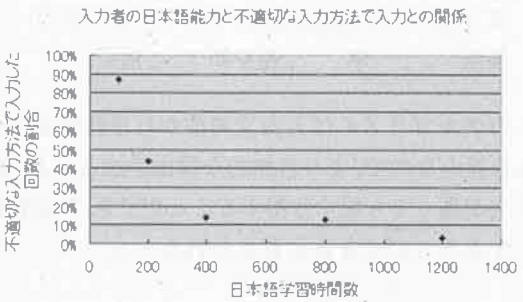

図 5: 入力者の日本語能力と不適切な入力方法で入力との関係

と比較しその日本語能力との関係を分析する. なお, 実験では, 外国人日本語学習者の日本語入力の実際の状況と近い結果を得るため, 単語やそれに含まれる個々の漢字の読みが分からない場合は入力者個人の習慣により, 紙製の辞書或は電子辞書（以下では両方とも括弧つきの「辞書」と表記する）で調べてもよいこととした.

また, 本アルゴリズムの有効性を示すために, シミュレーション評価実験を行った. 今回の実験は今までのシステム対応できなかった辞書にない読みで入力した単語の検出を重点として計画し, 手入力で実行した.

3.2 実験結果と評価

得られた結果は 5 人のもので, かつ, 数百文字という小さなデータであるから, 評価結果は断定的なものではない. しかし, この中には多くの外国人日本語学習者が抱えている日本語における問題が多数現れており, 彼らに対する日本語学習支援機能の充実の必要性を示していると考える.

また, 入力者である日本語初学者に入力によるストレスを極力に発生させないようにするため, 本実験で用いた実験用文の例文を図 4 に示すように日本での生活常識で構成し, 彼らの興味を引くものにした.

入力者の日本語能力と不適切な入力法での入力の関係を図 5 に示す. 図 5 の横軸は入力者の日本語学習時間数で, 縦軸は不適切な入力方法で入力した回数の割合である. 図の一番左側の日本語学習時間数 100 時間しかない入力者は不適切な入力方法で入力した回数は全体の 70 箇所のうち 61 箇所で, 全体の 90%近くとなり, ほぼ不適切な入力方法で実験文章を入力したと言える. それと逆に, 図の一番右側の日本語学習時間数 1200 時間の入力者の不適切な入力状況は全体の 70 箇所に対して 2 箇所しかなかったので, わずか全体の 3%である. 本実験では 5 人のデータを収集し, 断定的なものではないが, 入力者の日本語能力と不適切な入力方法で入力した回数の割合とは負の相関を有していると考えられる.

本研究で提案したアルゴリズムはまだシステムに実装レベルまで達していないため, 今回は実験で使われた対象文を手入力により入力し, うち不適切な入力が起こりうる 70 箇所を全部不適切な入力方法で入力した. そのうち, 66 箇所の不適切な入力が検出し, わずか 4 箇所の不適切な入力が検出できなかった. 不適切な入力を検出できなかった理由は, その四つの文ともいずれも長文であり, 本アルゴリズムは茶筅解析結果と入力仮名文字列を分割するのに, 使う純仮名単語のほとんどは「は」, 「が」, 「に」, 「を」のような単文字助詞である. しかし, 長い日本語文にそのような単文字助詞と同じ仮名が他の単語に存在する比率が高い. その場合, 細かく入力文を分割できない. まして, 一回も分割できないこともある. そういった場合では, 辞書にない読みで入力された単語の検出精度が低くなり, 検出できないこともある.

4. 考察と課題

本研究で提案したアルゴリズムを使用することによって, 比較する節を短く分割することができ, システムの検出精度が高くなり, 辞書にない読みで入力された単語のほとんどを検出することができると考えられる. また, 日本語初学者がよく使われる不適切な入力による辞書にない読みで入力された単語の検出ができ, より有効的に日本語初学者における日本語入力及び学習の支援になることが期待できる.

しかし, 本アルゴリズムは長い日本語の中の辞書にない読みで入力された単語を検出する時に, 分割できないことがあり, 検出精度が低くなって, 辞書にない読みで入力された単語を検出できないことがある. また, 本アルゴリズムを使用したシステムは入力文を文ごとに検出処理を行うため, 一文の中のすべての検出された辞書にない読みで入力された単語の辞書読みを表示するので, 入力者に負担をかけてしまい, 学習効果が落ちる恐れがある. そのため, 今までの検出処理の上にさらに入力文ごとに辞書にない読みで入力された単語の再検出する機能は今後の課題となる.

参考文献

[1] 土屋 順一：外国人のための日本語キーボード入力支援システムの母語別カスタマイズ，電器通信普及財団研究調査報告書第１５号，98-01044

[2] 張 小剛：日本語入力及び学習支援機能の検討と構成，福井大学大学院工学研究科博士論文，2007 年３月

[3] 奈良先端大学開発 形態素ツール「茶筅」：http://cactus.aistnara.ac.jp/lab/nlt/NLT.html