福井大学
工学部　研究報告
第41巻　第1号
1993年3月

マルコフ連鎖モデルによるかな漢字変換の曖昧性の解消効果

荒木哲郎*　芳永寛史*　稲波太志**　和田久信***

Evaluation of Disambiguity for Kana-Kanji Transformation of Non-Segmented Japanese Kana Sentences by Markov Model

Tetsuo ARAKI, Hiroshi YOSINAGA, Hutosi INAMI, and Hisanobu WADA

(Received Feb. 26, 1993)

In this paper, we describe a method of disambiguity for kana-kanji transformation of non-segmented Japanese kana sentences by 2nd-order markov model.

The experimental results using the newspaper articles show that this method is useful for disambibuity of the kana-kanji string candidates transformed from non-segmented kana bunsetsus, in case of bunsetsus being correct kana strings and ambiguous kana strings.

1　まえがき

日本語を計算機に入力する方法としては、現在区切りを一切入れないべた書き文入力方式が主流となっている。べた書きかな文の場合に、総当たり法でかな漢字変換等の処理により生成される、あらゆる単語候補列の組み合わせを考慮して解析を行うと、一般に探索木が爆発する問題が生じる。従来、漢字かな交じり文の単語並びに文節分割については高精度に行える［8］が，かな文については最長一致法による方法［1］、文節数最小法［2］、前後の連接文字を利用した方法［3］、格文法を用いる方法［4］，連語解析を用いる方法［5］、単語共起の関係を用いる方法［6］等があるが、同音語による曖昧さと分かち書き処理の曖昧さを同時に解決しなければならず、現在のところではまだ十分な精度を得るには至っていない。

本論文では、べた書きかな文の文節分かち書き処理に対しては、［8］［10］における仮文節切りの考え方に補正を行う方法で求められる文節境界を前提としており，そのべた書きかな

*電子工学科　**（株）日立製作所　***（株）松下電器産業

文節より、かな漢字変換によって得られる膨大な漢字かな候補列を、音節認識候補列の絞り込み［9］及び漢字かな候補絞り込み［１０］に対して、有効性が示されているマルコフ連鎖率モデルを用いて絞り込む方法を示すと共に、その効果を実験により把握する。更にかな漢字変換の絞り込み効果を定量的に調べるために、曖昧なべた書きかな文節の場合として、日本語音声認識における音響処理結果の曖昧な出力候補として与えられる音節ラテイス［7］から、マルコフ連鎖確率によって絞り込まれた１０位までの音節認識候補列［９］を用いて実験を行う。

第２章では、基本的な定義を行い、マルコフ連鎖確率辞書の飽和特性並びにシャノンの情報量を求め、２重マルコフ連鎖確率情報の持つ基本的な特徴を述べる。また日本語音声入力におけるかな漢字変換処理によって得られる漢字かな候補の生成方法ならびに文節単位の漢字かな交じり文節候補の絞り込み方法を示す。また３章では、２重マルコフ連鎖確率を用いた文節単位の漢字かな交じり文節候補の絞り込みの実験結果を示す。特に文節単位の漢字かなマルコフ連鎖確率を用いた絞り込みの効果を定量的に把握する為に、正しい音節候補列が特定された場合の音節列文節からかな漢字変換によって得られる漢字かな交じり文節候補の絞り込みと、曖昧な音節候補列から得られる漢字かな交じり文節候補の絞り込みの２つの場合について評価する。更に、マルコフ連鎖確率による絞り込みの方法として、漢字かな文字列で正規化する場合としない場合の評価ならびにかな漢字変換処理の際に用いる文節内の単語分割数と正解率の関係についても述べる。

2　マルコフ連鎖確率辞書の特性と音節漢字変換による漢字かな交じり文節候補の絞り込み方法

2．1　漢字かなマルコフ連鎖確率辞書の学習と特性

【定義１】日本語文$S = x_1 x_2 \cdots x_n$において、すべての文字x_i（$1 \leqq i \leqq n$）が音節文字であるときSを音節文、または全てのx_iが｛漢字、ひらかな、カタカナ、英数字、記号等）のいずれかの文字で表されるときSを漢字かな交じり文と呼び、nを文Sの長さと呼ぶ。Sの中の連続した文字の部分列$x_i x_{i+1} \cdots x_{i+q}$が自立語（詞）＝（名詞、動詞、形容詞、形容動詞、副詞、連体詞、接続詞、感動詞、形式名詞）と付属語（辞）＝（助詞、助動詞、接辞）から構成される一つの単位を文節Bを呼び、$B = 《x_i x_{i+1} \cdots x_{i+q}》$と書く。但し、Bには自立語は必ず含まれるが、付属語は必ずしも含まれるとは限らない。（定義終）

ここでは、［８］や［１１］の仮文節切りなどの考え方に従って正しく求められた文節境界をもとに以降議論する。

音節マトリックスから、マルコフ連鎖確率を用いて絞り込みを行った結果［９］、正解候補が第一位にくる場合を一意な音節候補列と呼び、また１０位内に入る音節候補を全て出力結果とする場合を、曖昧な音節候補列と呼ぶ。

一意な音節候補列及び、曖昧な音節候補列に対して、それぞれ以下のように単語辞書引きし

て、候補と存在した単語を順次接続して単語境界の整合がとれた文節候補列の集合を、各々一意な音節候補列から生成された漢字かな文節候補の集合と呼び**Y**で、また曖昧な音節候補列から生成された漢字かな文節候補の集合と呼び**Z**で表す。ここでは［９］で示したように日本語音声出力［８］で用いる１１０音節をかな表記として用いる。

【定義２】単語辞書が音節列をキー見出しとして、漢字かな表記の単語を読み出せるとき、文節候補$B = x_1 x_2 \cdots x_n$の中の部分音節列$c_i = x_k x_{k+1} \cdots x_{k+r}$をキーとして辞書引きに成功したとき、音節列$c_i$を「キー」とする漢字かな単語候補$q_j$の集合を$\{q_j(c_i)\}$と表す。Bを互い重なり部分がない音節部分列に分割$B = c_1 + c_2 + \cdots + c_s$したとき、全ての$c_i$に対して単語辞書に各単語候補$q(c_i)$が全て存在する場合，一連の単語候補列$q(c_1)\,q(c_2)\cdots q(c_s)$を文節当たりの漢字かな候補列と呼ぶ。その時のBの音節部分列への分割数sを単語分割数と呼び、文節における最小の単語分割数を最小分割数と呼ぶ。（定義終）

新聞記事７７日分を用いた漢字かなの２重マルコフ連鎖確率の辞書（文節単位）の学習の飽和特性をそれぞれ図１に、また０重、１重、２重のエントロピーをそれぞれ表１に示す。同図より約８０％～９０％の所でほぼ飽和していることがわかり，同表より２重連鎖マルコフ連鎖確率は１重に比べてかなりエントロピーが小さく絞り込み能力が高いと推察される。

【定義３】文節候補列を$B = x_1 x_2 \cdots x_n$とし、Bの確からしさを表すコストC（B）と書き、２重マルコフ連鎖確率によって次のように定義する。

$$C(B) = -\Sigma \log P(x_i | x_{i-2} x_{i-1}) \quad (1)$$

但し、$i<0$または$i>n$のときは、文節境界を表す空白文字bとする。（定義終）

２重マルコフ連鎖確率を用いた文節候補Bのコスト計算法の例を図２に示す。同図では文節文字列長による正規化を行う場合と行わない場合を示している。

【定義４】正解音節候補列から漢字かな候補列の集合**Y**または、曖昧な音節候補列から漢字かな候補列の集合**Z**に対して、定義３のコスト計算法によって求めたコスト値の小さい順（確率値が高い）にソートし、第ｎ位までに入る候補の中に正解候補が存在すると見なして正解候補を求めることを、マルコフ連鎖確率モデルによる絞り込みと呼ぶ。（定義終）

３　文節単位の漢字かな交じり文節候補の絞り込み実験

文節単位の漢字かな候補列の絞り込み実験では、文節単位の漢字かなマルコフ連鎖確率を用いた絞り込みの効果を定量的に把握する為に、正しい音節候補列が特定された場合の音節列文節から音節漢字変換によって得られる漢字かな交じり文節候補の絞り込み実験と、曖昧な音節候補列から得られる漢字かな交じり文節候補の絞り込みに分けて実験を行う。

３．１　実験の条件

（１）文の種類：新聞記事データ

（２）文節数：６１２６文節
（３）平均音節列長：５.４８
（４）その他
・単語辞書は１５万語
・候補順位は１０位まで

３．２　実験結果

（１）２重マルコフ連鎖確率による漢字かな交じり文節の絞り込み効果

一意な音節候補列（正解音節列の場合）からの漢字かな交じり文節の集合Ｙの絞り込み、及び曖昧な音節候補列からの漢字かな交じり文節の集合Ｚの絞り込みの結果を図３と図４に示す。

一意な音節列からの音節漢字変換による漢字かな交じり文節の絞り込み実験により、第１位内に８４％の正解漢字かな交じり文節が入り、１０位までの累積正解率は９４％であった。また曖昧な音節候補列からの音節漢字変換による漢字かな交じり文節の絞り込み実験では、第１位内正解率は６２％で、１０位内累積正解率は８７％であった。前者の値は、１５万語の単語辞書による膨大な単語候補を組み合わせた文節候補列に対して、２重マルコフ連鎖確率が持つ平均絞り込み能力を表しており、第一位正解率が８０％以上のかなり高い正解率を示すことがわかった。また後者との比較では、変換の対象となる音節列の個数が、１０倍の曖昧さを有しているにもかかわらず、１０位内累積正解率で見ると７％程度の差となっていることから、かなり高い絞り込み効果があることがわかった。一意な音節候補列の場合においてマルコフ辞書作成に用いた標本内データと音節列を取る場合，またそれ以外の各種標本外データを用いたときの１０位内に正解とならない文節の例を図５に示す。

（２）文節内の単語分割の最大数（最小分割数）

図６に文節内の単語分割数を変化させた時の１０位内累積正解率の変化を示す。音節漢字変換の際に用いる文節内の単語分割数の基準としては、最小分割数が通常良く用いられるが、分割数を色々変化させると、最小分割数に＋１の所で、累積正解率が飽和していることが今回の実験によって初めて明かにされた。

（３）コスト計算法における文節列長による正規化の効果

図７に一意な音節候補列（正解音節列の場合）からの漢字かな交じり文節の集合Ｙの絞り込みの場合おけるマルコフ連鎖確率値によるコスト計算方法として、漢字かな文字列で正規化する場合としない場合の絞り込みの結果を示す。また正規化しない場合に正解順位が高くなる例と、正規化した方が正解順位が高くなる例を表２にに示す。

両者における差異は余り見られず、今回の実験ではむしろ正規化しない場合の方が高い正解率を示した。

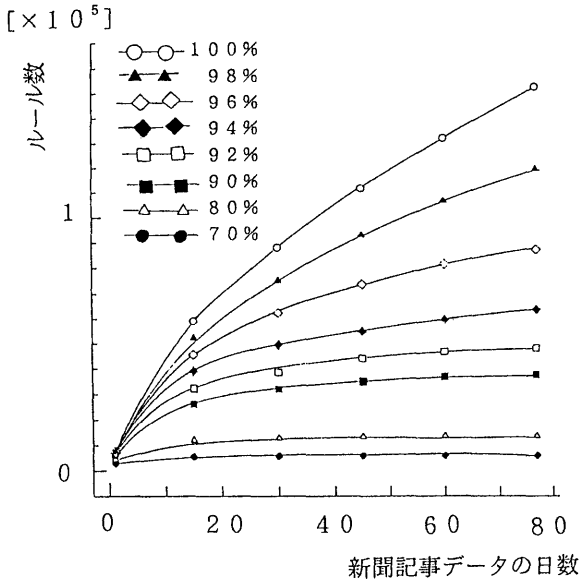

図1　文節単位の漢字マルコフ辞書（2重）の飽和特性

表　1　各種タイプのマルコフ連鎖確率辞書のエントロピー

辞書タイプ	エントロピー
0重	8.152142
1重	4.448533
2重	2.878640

$$P(\_\_大蔵省は\_\_) = (P_1(\_\_大) + P_2(\_大蔵) + P_3(大蔵省) + P_4(蔵省は) + P_5(省は\_) + P_6(は\_\_)) / 6$$

a)　漢字かな列長で正規化する場合の例

$$P(\_\_大蔵省は\_\_) = P_1(\_\_大) + P_2(\_大蔵) + P_3(大蔵省) + P_4(蔵省は) + P_5(省は\_) + P_6(は\_\_)$$

b)　漢字かな列長で
正規化しない場合の例

但し　P=-logpとする

図2　マルコフ連確率によるコストCの計算方法の例

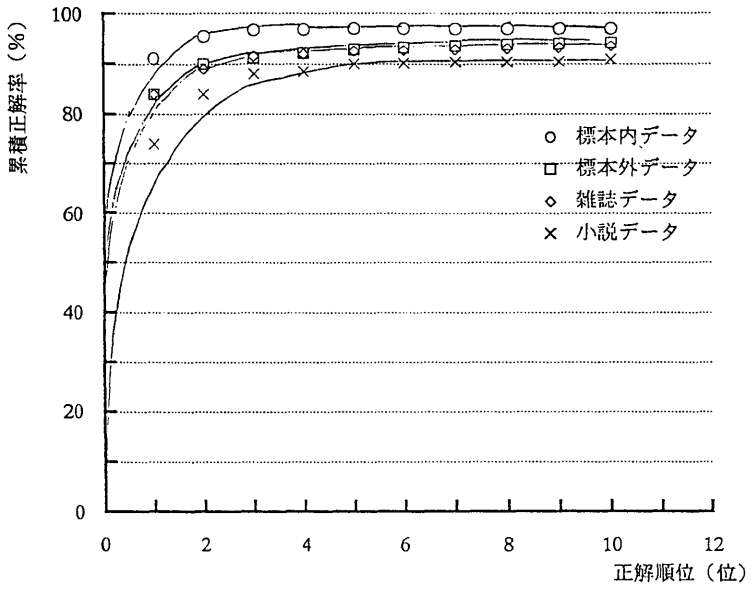

図３　一意な音節列からの漢字かな交じり文節候補の絞り込み結果

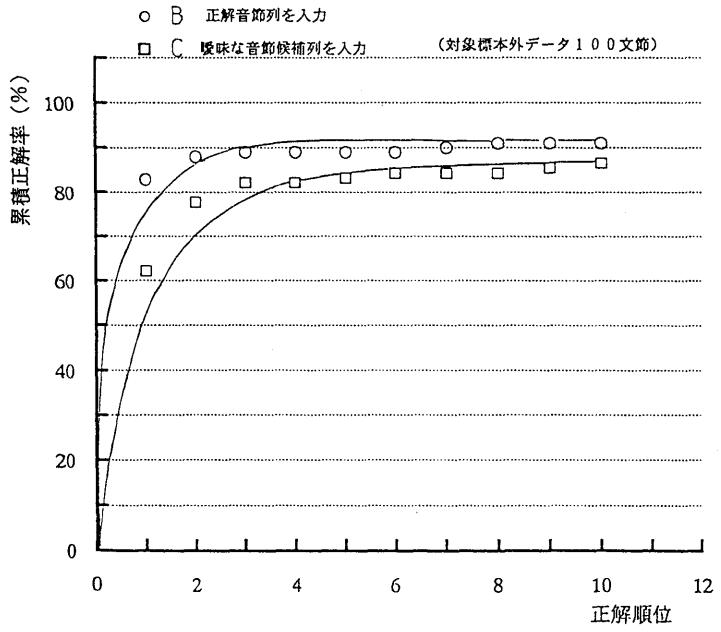

図4　入力音節列の違いによる漢字かな交じり文節候補の絞り込みの比較

大蔵省はこれによって在日外銀に関する法的根拠が明確になるほか在日外銀の国内活動がしやすくなり欧米諸国の間に出始めているわが国の金融制度に対する不満を和らげるのに役立つとみている
わが国経済の国際化に伴い日本に進出する外国銀行は急増している
大蔵省は邦銀に対しては銀行の経営基盤を安定させる目的で同準備金の積み立てを義務づけており大きな損失などが生じる時だけに取りくずしを認めている
大蔵省はこの規定を在日外銀に適用することで取引先などの在日外銀に対する信頼が高まるとみている
在日外銀の自己資本はその進出外銀の本支店合計の自己資本を計算基準にする
また国内での現地法人設立も出来るよう法体系を整備する
わが国の企業が海外でワラント債を発行するのは初めて
ワラント債は転換社債に似た社債だが償還期限が接近するまでワラントが行使されないことが多くしたがって株価が上がればすぐ株式に転換される転換社債に比べ自己資本の増加がなだらかという特徴がある
国内ではワラントと社債を分離できないが海外では分離できるので発行条件が国内より弾力的に決められる
しかも海外で発行すれば外貨建て債権のヘッジにも利用できるためワラント債は外貨建て債権の多い輸出企業にとっては株式への転換で社債が消滅してしまう転換社債よりも発行するメリットが大きい
所得税減税問題に関する首相の発言は次の通り
電力会社は施設が多くの都道府県にまたがるため事業税は本社所在地だけでなく発電施設のある都道府県に配分している
しかし国防総省はこれに対しほとんど関心を示さなかった
国防総省の装備品調達リスト
和田局長がわざわざ申し出なくても米国は日本の軍事転用可能な技術を以前から利用している
米国で最も脚光を浴びている軍需企業だ
わが国には戦闘の現場で実用性耐久性を確認した国産兵器など存在しない
戦車は砂漠の戦場ではじめてきたえあげられるものだという
では米国の軍需産業は何をねらっているのだろうか
それは日本が将来の防空システムの要についても米国の指示を仰いでいることを意味している
その半面外務国防法務内務教育などの閣僚は動いていない
更迭された南首相自身も朴政権下で副首相兼経済企画院長官を務め高度成長論者として知られた人である
こうみれば高度成長路線が残したひずみの是正に取り組みつつ不況下で沈滞する韓国の産業経済をなんとかして活性化しようとする期待と悲願が今回の改造からうかがえよう

（ａ）標本内データ新聞記事

いま日本は繁栄のきわみにいるのだろうか
その中でひとり繁栄に安住しようという気分も広がっている
鎖国の夢である
西欧の音響機器市場では日本勢が激しく競争している
音楽好きの若者が集まる反核集会は格好の宣伝の場だったらしい
ワシントンの米国務省内
エレクトロニクス通信など日本の汎用技術水準は急速に高まっているがそれが軍事技術に転用可能かどうか判断するモノサシを日本は持ち合わせていない
米国のコップ国務次官補代理デニシク商務次官補代理らの顔もみえる
しかし会議の合間をぬって米仏代表がひそかに協議したことはあまり知られていない
ミッテラン大統領の登場とともにフランスは対ソ高度技術輸出にきびしい態度に転じている
対ソ規制をめぐって西欧各国の同調を求める米国
ココムで外交の新しい方向を示そうとするフランス
両国の利害はかみ合ってきている
ココムでの米仏会談をようやく知った日本代表はあのときのことを連想する
長岡氏が昨年東南アジア中近東アフリカを回ったときまとめたものだ
あえて長岡氏が外務省の痛いところをついたのは日本の繁栄につれて途上国勤務が敬遠される現実を知らされたからだ
集まった大手商社の現地駐在代表は口口に訴えた
部長代理がそれを断るとこんどは人事部が意地になった
同じ部内から代わりを出せという
ナイロビはアフリカのなかでは比較的いい環境の勤務地なのだが勤務には条件がつけられていた
内に閉じ込もり勝ちな現代日本の心理状態をあらわしている
西独系化学会社の副社長をつとめるトレス氏
ところが会議になると悪いけどはずしてくれと言われる
西洋文化が積極的に取り入れられた
米国の対日貿易は赤字に
国振り文化の要素である傲慢さがはっきり出る
国際的孤立
米国からの技術導入が経済発展を支えた
日本国内に内向きの兆し
繁栄の中で鎖国の亡霊もさ迷う日本
しかし各国にとっては日本こそ巨大な黒船に映っていることに気がついていない
この認識ギャップが摩擦激化の背景にある
日本経済の底力は大きい
長崎市
オランダ政府も協力する姿勢だ
そこに出島商社を設立する
海外とのからみ合いを軸に発展してきた日本
そのからみ合いをもっと深くしないかぎりいまの繁栄はつかの間に終わるおそれがある

（ｂ）標本外データ新聞記事例

図５-１　１０位内に正解とならない文節の例（Ｙの場合）

親は我が子の行く末を案じ国は未来を担う次なる世代の行く末を案じる
子供たちがもっと読み書きの力をつけハイテク関係の仕事に就けるようになってくれればとアメリカは願っている
東欧諸国は市場経済という未知の航路をこぎゆく技術を若者たちの身に着けさせたい
日本は他人の研究成果を加工するのではなく自前の画期的な発明を送り出せる研究者の出現に期待をかけている
世界を襲った急激な変化は将来への不安を一段と高めた
自分たちの今後はどうなるのか
落ちこぼれたりしないだろうか
いま先進国で自国の教育システムに満足している国はないだろうとバンダービルト大学で公共政策と教育を講じるチェスターフィン教授は言う
もう一つ各国が知りたがっているのは教育で世界のトップに立つ方法だ
ある意味で各国間の競争は昔から教育改革の原動力の一つだったのである
ブッシュ政権も二０００年のアメリカと名づけた将来構想で数学と科学教育の見直しを呼びかけ今世紀末までにアメリカ教育レベルを世界一に引き上げるという不可能に近い公約を掲げている
この対局に位置する考え方教育改革で各国が協力し合うという考え方は残念ながらまだ日が浅い
だが教育関係はすでに他国の教育システムの長所に目を向けはじめている
たとえばニュージーランドで有効な読み書き教育の方法はアメリカでも役立つはずだ
ドイツの優れた職業訓練プログラムはきっと周辺の旧共産圏諸国の役に立つ
学校が十分機能しておらず教育不在が多くの社会的経済的問題の原因になっていることは世界中が認めておりその解決策を模索していると先ごろ設立されたあらゆる人のための教育に関する世界会議のシャロンピケットは言う
金の使い方が大切
では世界のどの国の学校がどの分野で最も優れているのか
そして優れている理由は
本誌は多くの専門家に話を聞き各国の得意分野を割り出した
たとえば読み書きはニュージーランドが数学と語学はオランダが日本の得意分野は科学ドイツは中等教育と教員養成スウェーデンは社会人教育だ
アメリカは異論もあろうが高等教育とくに大学院のシステムが優れている
幼児教育ではイタリアトスカーナ地方のレッジョエミーリアが草の根レベルのプロジェクトとして大成功をおさめ世界中のモデルとなっている
アメリカのピッツバーグは創造性と思考能力を育てる芸術教育に定評がある
一連の取材からみえてきたのはやる気になればできるという事実だ
たとえば日本ドイツオランダ
この三国は第二次大戦後国の復興と同時に学校制度の立て直しに取り組み今や数学や科学などの教育でトップに立った

（ｃ）雑誌記事データ例

客は夕方の散歩から帰ってわたしの書斎でわたしのそばに腰かけていた
昼間の明るさは消えうせようとしていた
窓の外には色あせた湖が丘の多い岸に鋭くふちどられて遠くかなたまで広がっていた
ちょうどわたしの末の男の子がおやすみを言ったところなので私たちは子供や幼い日の思い出について話し合った
子供ができてから自分の幼年時代のいろいろの習慣や楽しみがまたよみがえってきたよ
それどころか一年前から僕はまたちょう集めをやっているよ
お目にかけようか
とわたしは言った
彼が見せてほしいと言ったのでわたしは収拾の入っている軽い厚紙の箱を取りに行った
最初の箱を開けてみて初めてもうすっかり暗くなっているのに気づきわたしはランプを取ってマッチをすった
するとたちまち外の景色はやみに沈んでしまい窓全体が不透明な青い夜の色に閉ざされてしまった
わたしのちょうは明るいランプの光を受けて箱の中からきらびやかに光り輝いていた
私たちはそのうえに体をかがめて美しい形や濃い見事な色を眺めちょうの名前を言った
これはワモンキシタバでラテン名はフルミネア
ここらではごく珍しいやつだ
とわたしは言った
友人は一つのちょうをピンの付いたまま箱の中から用心深く取り出し羽の裏側を見た
妙なものだ
ちょうを見るくらい幼年時代の思い出を強くそそられるものはない
僕は小さい少年のころ熱情的な収拾家だったものだ
と彼は言った
その思い出が不愉快ででもあるかのように彼は口早にそう言った
その直後わたしが箱をしまってもどってくると彼は微笑して巻きたばこをわたしに求めた
悪く思わないでくれたまえとそれから彼は言った
君の収拾をよく見てなかったけれど
僕も子供のときむろん収拾していたのだが残念ながら自分でその思い出をけがしてしまった
実際話すのもはずかしいことだがひとつ聞いてもらおう
彼はランプのほやのうえでたばこに火をつけ緑のかさをランプに載せた
すると私たちの顔は快いうす暗がりの中に沈んだ
彼が開いた窓のふちに腰かけると彼の姿は外のやみからほとんど見分けがつかなかった
私は葉巻を吸った
外ではかえるが遠くから甲高くやみ一面に鳴いていた
友人はその間に次のように語った
僕は八つか九つのときちょう集めを始めた
初めは特別熱心でもなくただはやりだったのでやっていたまでだった

（ｄ）小説データ例

図５-２　１０位内に正解とならない文節の例（Ｙの場合）

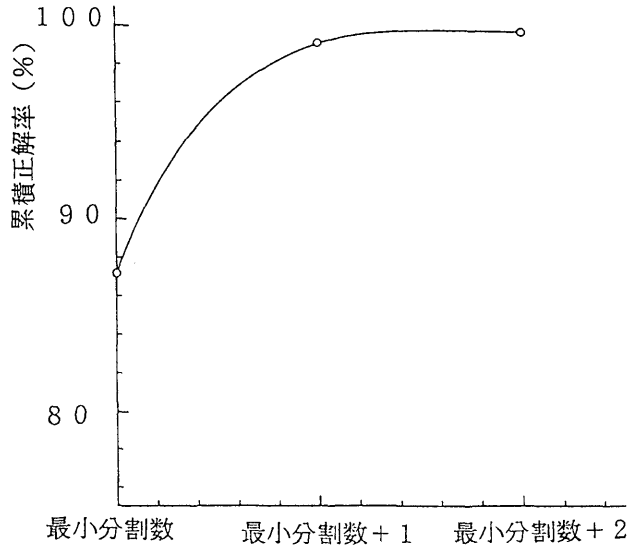

図6　文節内の単語分割数と累積正解率の関係
（標本内データ使用）

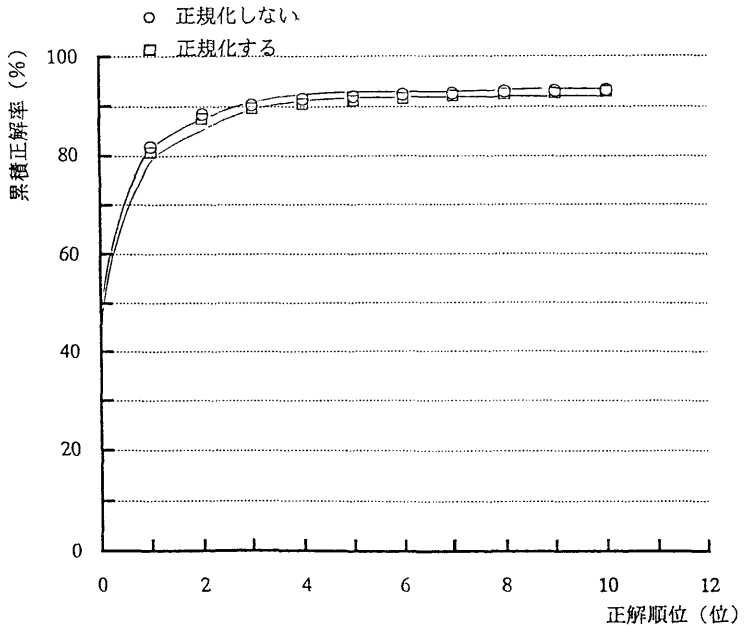

図7　漢字かな文字列長の正規化の有無による
絞り込み効果

表２-１　漢字かな文字列長の正規化有無による
正解順位の比較例
（※１１位は１０位内に正解のないこと示す）

音節列	漢字かな列	正規化有	正規化無
いま	いま	×(2位)	○(1位)
はんかくしゅうかいわ	反核集会は	○(3位)	×(11位)
こくむしょうない	国務省内	○(4位)	×(11位)
ものさしお	モノサシを	×(3位)	○(2位)
だいり	代理	○(2位)	×(3位)
たいそきせいお	対ソ規制を	○(5位)	×(8位)
しめそうと	示そうと	○(1位)	×(2位)
りがいわ	利害は	○(1位)	×(2位)
まわつたと	回ったと	○(1位)	×(2位)
ぶちょうだいりが	部長代理が	○(1位)	×(2位)
ことわると	断ると	○(7位)	×(11位)
なかでわ	なかでは	×(2位)	○(1位)
うちに	内に	○(1位)	×(3位)
しんりじょうたいお	心理状態を	○(9位)	×(10位)
あらわしている	あらわしている	×(3位)	○(1位)
せいどくけい	西独系	○(8位)	×(9位)
つとめる	つとめる	×(4位)	○(3位)
こりつ	孤立	○(1位)	×(3位)
きざし	兆し	○(1位)	×(2位)
さまよう	さ迷う	○(4位)	×(11位)
じくに	軸に	○(1位)	×(2位)
かぎり	かぎり	×(2位)	○(1位)
いまの	いまの	×(2位)	○(1位)
おそれが	おそれが	×(3位)	○(2位)
しんしゅん	新春	○(1位)	×(4位)
はいるが	入るが	○(1位)	×(2位)
いまのところ	いまのところ	×(2位)	○(1位)
きわめて	きわめて	×(2位)	○(1位)
てきとうとの	適当との	○(1位)	×(2位)
はじめる	始める	○(1位)	×(2位)
とこうの	渡航の	○(7位)	×(8位)
あきに	秋に	○(2位)	×(4位)
ざいせきのまま	在籍のまま	○(10位)	×(11位)
つまれるのが	積まれるのが	×(3位)	○(2位)
がいこくせいかつが	外国生活が	○(1位)	×(3位)
ごたいけんと	ご体験と	○(2位)	×(6位)
かえつて	かえって	×(2位)	○(1位)
もつとも	最も	○(1位)	×(2位)
くろじばかりが	黒字ばかりが	×(4位)	○(1位)
らつかんひかん	楽観悲観	○(3位)	×(11位)

表2-2　漢字かな文字列長の正規化有無による
正解順位の比較例

音節列	漢字かな列	正規化有	正規化無
しょげんしょうわ	諸現象は	○(6位)	×(8位)
はいいろの	灰色の	○(2位)	×(3位)
ものごと	物事	○(1位)	×(2位)
けいかいぎみに	警戒気味に	○(2位)	×(11位)
もつとも	最も	○(1位)	×(2位)
われわれが	われわれが	×(2位)	○(1位)
じんるいしゃかいわ	人類社会は	○(1位)	×(2位)
もさくてき	模索的	○(8位)	×(11位)
はいつた	入った	○(1位)	×(2位)
わくないに	ワク内に	×(3位)	○(2位)
けいざいじたいに	経済自体に	○(4位)	×(7位)
もつとも	もっとも	×(2位)	○(1位)
じんるいしゃかいの	人類社会の	○(1位)	×(2位)
ぶんやや	分野や	○(8位)	×(10位)
もろさお	もろさを	×(11位)	○(2位)
くるまざお	車座を	○(1位)	×(6位)
ないかくきしゃかいと	内閣記者会	○(4位)	×(6位)
おこない	行い	○(1位)	×(2位)
すすめてゆき	進めてゆき	○(1位)	×(2位)
ていれいの	定例の	○(2位)	×(5位)
てんまで	点まで	○(5位)	×(11位)
ろうしもんだいが	労使問題が	○(9位)	×(11位)
ていれいの	定例の	○(2位)	×(5位)
おこない	行い	○(1位)	×(2位)
けんとうあんお	検討案を	○(2位)	×(5位)
じどうふりかえけいやくお	自動振替契約を	○(10位)	×(11位)
むすべば	結べば	○(1位)	×(2位)
はじめる	始める	○(1位)	×(2位)
うんようしだいでわ	運用次第では	○(2位)	×(4位)
わくお	枠を	○(2位)	×(3位)
そうこうしに	総公司に	○(1位)	×(2位)
しそう	思想	○(1位)	×(2位)
がいこうてき	外交的	○(4位)	×(11位)
せんいきかくの	戦域核の	○(1位)	×(3位)

4　むすび

本論文では漢字かなマルコフ連鎖確率を用いた文節単位の漢字かな候補列の絞り込み実験により、以下の知見を得た。

（１）正解音節列（または曖昧な音節候補列）からの音節漢字変換による漢字かな交じり文節の絞り込み実験により、第１位内に８４％（または６２％）の正解漢字かな交じり文節が入り、１０位までの累積正解率は９４％（８７％）であり、高い絞り込み効果があることがわかった。

（２）かな漢字変換の際にどの範囲まで文節内の単語分割数を調べればよいかの目安として、最小分割数に＋１の所で累積正解率が飽和していることが今回の実験によって初めて明かにされた。これは、文節が自立語（名詞、動詞など）と付属語（助詞、助動詞など）の組み合わせ方に大きく関係していることを示しており、今後の文節構造の研究に大いに役立つものである。

（３）マルコフ連鎖確率値による絞り込みの方法として、漢字かな文字列で正規化する場合としない場合の評価を行った結果、両者における差異は余り見られず、今回の実験ではむしろ正規化しない場合の方が高い正解率を示した。このことは、マルコフ連鎖確率モデルを用いて、長さが異なる漢字かな交じり文節候補列を比較評価する上において、必ずしも正規化することが必要とは言えない事を示しており、今後マルコフ連鎖確率モデルを適用していく上で一つの指針となりうるもので、日本語文の構造との関係を含めて、今後更に検討を重ねていく必要があると思われる。

謝辞

本研究を進めるに当たってお世話になりました情報通信網研究所知識処理研究部の池原悟主幹研究員並びに自然言語処理研究グループの方々に感謝致します。

［文献］

(1)牧野,木澤:べた書き文の分かち書きとかな漢字変換-二文節最長一致法による分かち書き,
情処論,20,4,pp337-345　(1979)

(2)吉村,日高,吉田:文節数最小法を用いたべた書き日本語文の形態素解析,
情処論,24,1,pp40-46 (1983)

(3)栃内,伊藤,鈴木:前後連接文字を利用した同音語選択機能を有するかな漢字変換システム,
情処論,27,3,pp313-320 (1986)

(4)大島,阿部,湯浦,武市:格文法による仮名漢字変換の多義解消,
情処論,27,7,pp679-687 (1986)

(5)本間,山階,小橋:連語解析を用いたべた書きかな漢字変換,
情処論,27,11,pp1062-1067 (1986)

(6)内山,板橋:共起関係を利用した日本語複合名詞の分割,
情処自然言語処理研資,91-7,pp57-54 (1992)
(7)中津,好打:会話音声の機械認識における音響処理,
信学論,J61-D,4,pp261-268 (1978)
(8)宮崎, 大山:日本文音声出力のための言語処理,
情処論,27,11,pp1053-1061(1986)
(9)荒木,村上,池原:2重マルコフ連鎖モデルによる日本語文節音節認識候補の曖昧さの解消効果,情処論,30,4,pp467-477 (1989)
(10)村上,荒木,池原:日本文音節入力に対して2重マルコフ連鎖モデルを用いた漢字かな交じり候補の抽出精度,信学論,J75-DII,pp11-20 (1992)
(11)荒木,池原,土橋:2重マルコフ連鎖確率モデルを用いたべた書き日本語文の先頭位置推定法の評価,情処自然言語処理研究会予定, 94,(1993)