論　文

帰納的学習を用いたべた書き文のかな漢字変換

荒木　健治[†]　高橋　祐治[††]　桃内　佳雄[†]　栃内　香次[†††]

Non-Segmented Kana-Kanji Translation Using Inductive Learning

Kenji ARAKI[†], Yuji TAKAHASHI[††], Yoshio MOMOUCHI[†], and Koji TOCHINAI[†††]

あらまし　我々は，人間の言語および知識獲得能力の解明とその実現を目的として研究を行っている．このような研究はこれまでいくつか存在するが，工学的に有効なシステムを完成するまでに至っていないのが現状である．また，心理学の立場からの研究も存在するが，これらの研究は工学的にどう実現するかという問題は対象としていない．そこで，本論文では，この言語および知識獲得能力を工学的に実現する一つの試みとして，入力べた書き文とその漢字かな交じり文より帰納的に語の読みと表記を獲得し，その獲得状況および変換精度に基づく確実性の高い語より順に変換するという帰納的学習によるべた書き文のかな漢字変換手法を提案する．本手法は辞書が空の状態からでも学習機能により語を獲得することができるので，辞書を個人ごとに自動的に作成することが可能，個人用辞書とすることにより辞書容量を少なくできる，未登録語の自動登録が可能，初期辞書作成の労力を回避できるという利点がある．本論文では，更に本手法に基づく実験システムを作成し，異なる4分野の論文40編を用いて本手法の学習機能による適応性能を確認するための実験を行った．実験の結果，4分野すべてにおいて90％以上の精度が得られ，べた書き文のかな漢字変換における本手法の有効性が確認された．

キーワード　帰納的学習，かな漢字変換，適応，確実性，べた書き文，言語獲得

1. まえがき

我々は従来より人間が言語を獲得し，知識を学習していくメカニズムに興味をもち，このメカニズムを解明して工学的に実現することにより，環境より情報を学習し，次第に成長するシステムの開発を最終目的とする研究を進めている．このようなシステムを開発することにより，複雑で使いにくいシステムやプログラム作成の重労働といった高度情報化社会の種々の問題を解決し，人にやさしい柔軟なシステムを実現できると考えている．人間の幼児は，言語も知識ももたない状態から周囲のさまざまな環境より学習を行い，言葉を話し，種々の知識をもつ大人に成長する．従って，このような生得的な能力は確かに存在する．この生得的な能力に関する研究は，いくつか存在するが[1]～[3]，工学的に有効なシステムを完成するまでには至っていない．また，心理学の立場からの研究も多くある[4],[5]．しかし，これらの研究は，どのようにして子供が言語を獲得するのかを解明するという点では大きな意義があるが，そのような機構をどのようにして工学的に実現するのかという問題は，当然ながら対象としていない．

我々は，このような生得的能力を解明する基礎的な研究の第1段階として，既に日本語の漢字かな交じり文を対象として未知の文字列中より語を獲得するメカニズムについての考察を行い，形態素解析手法としての応用を行った[6]．これを我々は帰納的学習による形態素解析手法と呼ぶ．この帰納的学習による形態素解析手法では，漢字かな交じり文より共通部分と差異部分を多段階に抽出することにより語を帰納的に獲得する．本手法を用いることにより，辞書が空の状態から出発し，論文2編程度の学習によって約90％の語を正しく認識できることを実験により確認した．また，異なる4分野の40編の論文を用いた実験でも各分野へ迅速に適応し，その学習機能の有効性を確認した[6]．

本論文では，言語獲得，知識獲得能力実現のための基礎的な研究の第2段階として，べた書き文とその漢

† 北海学園大学工学部電子情報工学科，札幌市
Faculty of Engineering, Hokkai-Gakuen University, Sapporo-shi, 064 Japan
†† 北海道ソフト・エンジニアリング株式会社，札幌市
Hokkaido Soft Engineering Corporation, Sapporo-shi, 060 Japan
††† 北海道大学工学部電子情報工学専攻，札幌市
Faculty of Engineering, Hokkaido University, Sapporo-shi, 060 Japan

字かな交じり文から帰納的学習によりかな漢字変換に必要な語の表記と読みを獲得し，次いで獲得状況および変換精度に基づく確実性の高い順に多段階に変換を行う手法[7]〜[9]について述べる．更に，本手法の学習能力の有効性を確認するために異なる4分野の40編の論文を用いて行った適応能力評価実験結果および考察結果について述べる．実験の結果，すべての分野で約95％の変換率が得られ，有効性が確認された．また，本研究の第1段階として行った帰納的学習による形態素解析手法[6]が漢字かな交じり文という1種類のデータからの学習を行うのに対して，本手法では，べた書き文とその漢字かな交じり文という2種類のデータからその対応関係を推定する過程が必要となる．本論文では，この推定メカニズムに関する考察を行う．

本手法では学習機能による適応能力の高さにより構文情報，意味情報といった解析的な文脈情報を用いずに比較的高い変換率を実現している．これは，本手法が文章中の語の出現確率といった確率的な文脈情報を利用しているとも考えられる．従来の手法が単語の見出し語はあらかじめ人手により与えて辞書に登録しその語のもつ頻度情報などを学習するのに比べて，本手法は単語の見出し語そのものをも学習することができる．従って，辞書に登録される単語の見出し語のレベルからシステムがユーザあるいは対象分野に根本的に適応することが可能になる．

一方，形態素解析，かな漢字変換における未登録語の問題に対する研究もいくつかある[10],[11]が，これらの研究は与えられた語が既に辞書中にある程度存在することを前提とした上での少数の未登録語に対する処理であるため，大半の語が辞書に存在し，その語の情報を利用して未登録語を予測している．従って，これらの手法でも結局，初期辞書作成の労力は避けられない．これに対して，本手法は，辞書が空の状態からでも学習により語を獲得することができるので，辞書を個人ごとに自動的に作成することが可能，個人用として辞書容量を少なくできる，未登録語の自動登録が可能，初期辞書作成の労力を回避できるという利点がある．

2. 基本的考え方

本手法の前段階となる帰納的学習による形態素解析手法においては，生得的な能力を「二つの事物が同じか異なるかを判断する能力」と仮定して手法を開発し，実験を行った．その結果からこの仮定の正当性を確認[6]できた．従って，本論文で述べる，対応関係を有する2種類の未知文字列からの帰納的学習においても，この能力から語の表記と読みを獲得できなくては研究の目的に合致しない．ところで，人間は対応関係を有する2種類の記号列に対してどのような方法を用いてその対応関係を決定しているのであろうか．表1に未知文字列からの対応関係抽出の例を示す．

表1 未知文字列からの対応関係抽出の例（1）
Table 1 An example of the extraction of correspondence relations from unknown character strings (1).

入力1	$\alpha\ \theta\ \sigma\ \psi\ \delta\ \lambda\ \vartheta$	
入力2	$\Xi\ \Sigma\ \psi\ \delta\ \Upsilon\ \Phi\ \Theta$	
セグメント1	$\alpha\ \theta\ \sigma$	$\Xi\ \Sigma$
セグメント2	$\psi\ \delta$	$\psi\ \delta$
セグメント3	$\lambda\ \vartheta$	$\Upsilon\ \Phi\ \Theta$

表2 未知文字列からの対応関係抽出の例（2）
Table 2 An example of the extraction of correspondence relations from unknown character strings (2).

入力1	$\alpha\ \theta\ \sigma\ \psi\ \delta\ \lambda\ \vartheta$	
入力2	$\Xi\ \Sigma\ \psi\ \delta\ \Upsilon\ \Phi\ \Theta$	
セグメント1	$\alpha\ \theta\ \sigma$	$\Upsilon\ \Phi\ \Theta$
セグメント2	$\psi\ \delta$	$\psi\ \delta$
セグメント3	$\lambda\ \vartheta$	$\Xi\ \Sigma$

表1の入力1，入力2のような対応関係を有する二つの未知文字列を見た場合に人間は，まず二つの文字列に共通な部分を検出する．表1の例では，下線部（$\psi\ \delta$）に注目すると考えられる．そして，その両側の差異部分をその出現順に対応付ける．すなわち，表1のセグメント1，2，3のような対応関係を考える．このような抽出過程は，我々が生得的な能力として仮定している「二つの事物が同じか異なるかを判断する能力」を基本として考えることができる．しかし，二つの文字列に共通な部分は別として異なる部分は，表1以外に，表2のようなセグメントを考えることもできる．

表2では，対応関係が出現順ではなく出現順と逆順になっている．人間がこのような対応関係を想定することはあり得るので，ここで出現順の対応関係を取るか出現順と逆順の対応関係を取るかは，対象とする問題に依存したヒューリスティックスであると考えられる．このようなヒューリスティックスがどのようにして獲得されるのかは，非常に重要な問題である．し

表3　セグメントからのプリミティブ抽出の例
Table 3　An example of the extraction of primitive segments.

セグメント1	$\alpha\,\theta\,\sigma$	$\Upsilon\,\Phi\,\Theta$
セグメント2	$\theta\,\sigma\,\gamma\,\mu$	$\Phi\,\Theta\,\Sigma$
プリミティブ1	α	Υ
プリミティブ2	$\theta\,\sigma$	$\Phi\,\Theta$
プリミティブ3	$\gamma\,\mu$	Σ

かし，同時に非常に大きなテーマでもあるので，別のテーマとして研究を進め，ここでは，このヒューリスティックスは与えられたものとする．ヒューリスティックスを獲得するメカニズムについては，更に研究を進め別の機会に報告したい．

更に，このようにして抽出されたセグメントから共通部分を抽出することにより，プリミティブな単位に分解することができる．表3にセグメントからの共通部分の抽出によるプリミティブ抽出の例を示す．表3に示したように，セグメント1，2の対応関係にある二つの文字列の各々の共通部分（$\theta\,\sigma$, $\Phi\,\Theta$）をプリミティブ2として抽出し，その両側の差異部分（α, Υ），（$\gamma\,\mu$, Σ）をそれぞれプリミティブ1，プリミティブ3として抽出する．このように，共通部分と差異部分を分離することにより三つのプリミティブを得ることができる．この三つのプリミティブを合成することによりもとの二つのセグメントを得ることができるので，この三つのプリミティブがあれば，二つのセグメントは不要となる．このような抽出方法は，基本的には我々の仮定している「二つの事物が同じか異なるかを判断する能力」によって得られるものではあるが，その他にやはり上述したような出現順に対応関係があるというヒューリスティックスが使われている．このように実例より共通部分と差異部分を多段階に抽出することにより知識を獲得する手法は帰納的学習の一つである．

3.　帰納的学習によるべた書き文のかな漢字変換手法

2.で述べた立場に基づき，人間の言語および知識獲得能力に基づく手法の，べた書き文のかな漢字変換における語の獲得への応用を行った．本章では，その応用方法について述べる．

3.1　概　　要

図1に本手法の処理の流れを示す．図1に示すように本手法は，変換処理，学習処理，フィードバック学習処理の三つの処理からなる．べた書き文が入力されるとそれまでの学習によって獲得された語によって語の当てはめが行われる．この際には，確実性の高い語より順に変換が行われる．確実性は，学習処理におけるその語の獲得状況および変換に使用された際の精度によって決定される．ここで，変換結果に誤りが存在する場合には，人手によって校正が行われる．次に，学習処理を行う．学習処理は，このようにして得られた正しい漢字かな交じり文と入力されたべた書き文を用いて行われる．学習処理は，二つの段階からなる．まず，第1段階では，入力べた書き文と校正済みの変換結果より共通部分を手掛りに語候補を得る．次に，第2段階ではこの語候補同士の共通部分と差異部分を抽出することにより語を獲得する．最後にフィードバック学習処理では，変換結果と校正された変換結果を比較することにより正しく変換した語と誤って変換した語を決定し，正しく変換した語のゆう度を上げ，誤って変換した語のゆう度を下げる．このことにより，次回からの変換精度の向上を図ることができる．

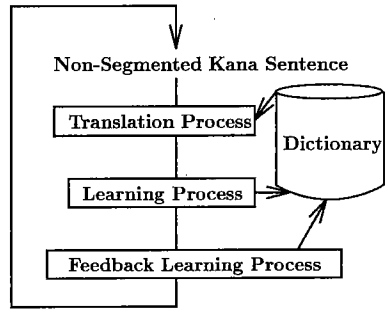

図1　処理の流れ
Fig. 1　Outline of process.

実際に行われる処理の流れは，図1に示されるように変換処理，学習処理，フィードバック学習処理の順に行われるが，変換処理では，確実性の高い語より順に多段階に変換が行われる．この確実性は，学習処理での語の獲得状況およびフィードバック学習処理でのゆう度の更新によって決定される．そこでまず，学習処理およびフィードバック学習処理について述べる．

3.2　学 習 処 理

3.2.1　語候補の獲得

入力べた書き文と校正済みの変換結果を用いて語候補を獲得する．この方法は，**2.**で述べた人間の二つの未知文字列からの対応関係の獲得過程に対する考察に基づいている．この処理で抽出される語候補をS1と

表4 語候補の抽出例
Table 4 Examples of the extraction of S1.

1．入力べた書き文	
われわれはしゅじゅのほうほうをおこなった．	
2．校正済みの変換結果	
我々は種々の方法を行なった．	
3．語候補の獲得結果	
差異部分	共通部分
（われわれ：我々）	（は：は）
（しゅじゅ：種々）	（の：の）
（ほうほう：方法）	（を：を）
（おこな：行な）	（った：った）

表5 対応関係を誤る例
Table 5 Examples of erros of the S1 extraction.

1．入力べた書き文	
そのすんぽうはびさいかされつつある．	
2．校正済みの変換結果	
その寸法は微細化されつつある．	
3．語候補の獲得結果	
差異部分	共通部分
（すんぽう：寸法）	（その：その）
*（び：微細化）	（は：は）
*（いかさ：）	*（さ：さ）
	（れつつある：れつつある）

＊は誤りを示す．

表記し，Segment Oneと呼ぶ．語候補を獲得する手順は以下のとおりである．

（1） 入力べた書き文と校正済みの変換結果の共通部分を決定する．

（2） 共通部分を区切りとして差異部分を抽出する．

（3） 出現順に従って差異部分の対応関係を決定する．

（4） すべての共通部分および差異部分をS1として抽出する．

表4にS1抽出の例を示す．表4で入力べた書き文と校正済みの変換結果の共通部分は下線を引いた部分である．この部分の対応関係がまず決定される．すなわち，表4の3.の右側の部分のようなS1を得る．次に共通部分の間の差異部分を抽出する．その対応関係は出現順なので，表4の3.の左側の部分のようなS1を得る．このようにして，語候補S1の表記と読みが獲得される．

表4のように対応関係が一意に決定できる場合は問題ないが，漢字の読みの中にひらがなで表記される語と同じ読みが存在する場合には，対応関係が多対多になり，誤った対応関係が抽出されてしまう場合がある．この例を表5に示す．

表5は左から右方向へ解析を行った結果である．このような誤ったS1から**3.2.2**で述べる方法で生成される語は，変換の際に誤りを引き起こすので，**3.3**で述べるフィードバック学習によってそのゆう度が下がる．すなわち，**3.3.3**で述べる語の階層の下位に所属することとなり誤り率は低下する．このような試行錯誤の過程は，人間の行う処理に近いと考えられる．しかし，誤った語のゆう度が下がるのは，徐々にであり，最初から誤った語が辞書中に存在しない方がシステムの変換精度が定常的に高くなるのは確かである．そこで，

表6 右から左方向の解析結果の例
Table 6 Result of right-to-left directional analysis.

1．入力べた書き文	
そのすんぽうはびさいかされつつある．	
2．校正済みの変換結果	
その寸法は微細化されつつある．	
3．語候補の獲得結果	
差異部分	共通部分
（すんぽう：寸法）	（その：その）
（びさいか：微細化）	（は：は）
	（されつつある：されつつある）
4．採用されるS1	
差異部分	共通部分
（すんぽう：寸法）	（その：その）
	（は：は）

システムの効率を考え，ここでは双方向解析を導入する．これは，更に右から左方向の解析を行い，二つの方向の解析結果が一致しなければ語候補としないというものである．双方向解析はヒューリスティックスの一つで，このようなヒューリスティックスも当然何らかの方法で獲得されたものであるが，ここでは与えられるものとする．

表5を右から左方向に解析した例を表6に示す．実際には，表6の3.の語候補の獲得結果には誤りは含まれてはいない．しかし，システムはこの段階では，左から右の解析結果と右から左の解析結果のどちらが正しいかを判断することはできないので，表5の3.と表6の3.の共通部分のみをS1として決定する．従って，両者で異なるS1は，ここでは獲得されないが，あいまいさの生じない別の文脈で獲得されると考えられる．

3.2.2 語候補からの語の獲得

次に，語候補として抽出されたS1より，更に共通部分と差異部分を抽出し，語を獲得する．これは，S1

表7　語候補からの語の獲得の例
Table 7　Examples of extraction of CS and RS.

1．S1	(れんぞくおんせい：連続音声) (おんせい：音声)
2．CS	(おんせい：音声)
3．RS	(れんぞく：連続)

表8　読みおよび表記の一致する条件
Table 8　Conditions of extraction of CS and RS.

1．文字数による制限	
$D=0$	3文字以上表記が一致
$D\neq 0$	2文字以上表記が一致
2．重複の状況	
一方が他方を含む．	

D = 読みの文字数 − 表記の文字数

が複数の語から構成される可能性があるので，S1 をそれを構成する語に分解するために行う処理である．この処理は，**2.** で述べたセグメントからのプリミティブの抽出に基づくものである．このときに抽出された共通部分をCSと表記し，Common Segmentと呼ぶ．また，差異部分をRSと表記し，Remained Segmentと呼ぶ．また，S1がCS，RSに分解されたとき，分解されたS1 は辞書中から削除する．表7に，語候補からの語の獲得の例を示す．

読みおよび表記が一致する条件を表8に示す．表8の1で $D=0$ の場合は表音文字の場合，$D\neq 0$ の場合は表意文字の場合を近似的に表している．表8の2の条件は本処理が複数の語から構成されるS1を一つの語から構成されるS1 に分解することを目的としているためである．従って，本手法では複合語を分割する際には少なくとも一つその複合語を構成する単語が既知でなければならない．しかし，すべての複合語が常にそのような状況にあるとは限らないので，未分割の複合語が存在する可能性があるが，その複合語を構成する語が一つでも獲得されると本処理により分割される．

3.2.3　1字1音語の処理

上述したような処理を行って語を抽出すると漢字とかなが1字ずつ交互に出現する「取り引き」のような語は，(と：取)，(り：り)，(引：ひ)，(き：き) のように抽出される．このような1字1音語の連続する場合には，変換処理でも常に変換されない．これは，**3.4.1** で述べるように変換処理では，読みが1字の語は両側が確定済みの場合にしか当てはめないという条件のためである．このためこのように漢字とかなが1字ずつ交互に出現する語は永遠に獲得されず，また当てはめられることもない．そこで，このように1字1音の語が連続する場合には抽出後，1語にまとめるというヒューリスティックスを与える．

3.3　フィードバック学習処理

3.3.1　手　　順

フィードバック学習では，変換結果および校正済みの変換結果より正しく変換した語のゆう度を上げ，誤って変換した語のゆう度を下げる．以下にフィードバック学習処理の手順を示す．

（1）　変換結果および校正済みの変換結果の共通部分と差異部分を決定する．

（2）　共通部分を正変換とし，差異部分を誤変換とする．

（3）　正しく変換した語の正変換度数を1増加させる．

（4）　誤って変換した語の誤変換度数を1増加させる．

3.3.2　ゆう度評価関数と頻度の更新方法

3.4.2 で述べるように変換処理で複数の候補が重複した場合には，ゆう度評価関数（CEF：Credibility Evaluation Function）を用いて最適な候補を決定する．このゆう度評価関数は，以下の式によって与えられる．

$$CEF = AF + \alpha \times CF - \beta \times EF \qquad (1)$$

ここで，AFは出現頻度（Frequency of Appearance），CFは正変換度数（Frequency of Correct Translation），EFは誤変換度数（Frequency of Erroneous Translation），α，β は係数である．

式（1）は，頻繁に出現し，精度も高く，誤りが少ない語がゆう度が高くなり，優先されて選択されるという意味である．式（1）でAFは，変換処理で語が使用された場合および学習処理でS1として抽出されたときに増加する．S1がCSおよびRSに分解されたときはCSのAFはCSを生成するために使用されたS1のAFの和とし，RSのAFはRSを含んでいるS1のAFとする．また，正変換度数CFについては，フィードバック学習処理で正変換と決定された変換に用いられた語のCFを1増加させる．また，誤変換度数EFについては，誤変換と決定された変換に用いられた語のEFを1増加させる．

前述したようにゆう度評価関数は，同一階層の中で

表9 語の階層の条件
Table 9 Rules for update of word ranks.

Rank	条件	
	獲得状況	正変換率
MS	CSより	CR ≧ 95%
CS	S1の共通部分	40% ≦ CR < 95%
S1	入力べた書き文と校正済み変換結果より	CR ≧ 40%
RS	S1の差異部分	CR ≧ 40%
LS	LS以外の各階層より	CR < 40%

当てはめ候補が重複した場合に最適な候補を決定する場合に用いられる．従って，候補が上位の階層に存在する場合には最適な候補の選択にゆう度評価関数は用いられず，その候補が所属する階層によって決定される．この所属階層は主に，語の抽出状況に基づいており，語の抽出方法はヒューリスティックスに基づいているので，本手法ではヒューリスティックスがゆう度評価関数に優先して用いられている．

3.3.3 語の階層

3.2で述べたように語はその獲得状況によって，S1, CS, RSの三つに階層化されている．更に，本手法で用いる語は各語の正変換率によって二つの階層を追加し，表9に示す計5階層に分類される．まず，CSの中で正変換率が95% 以上のものを最も確実性の高い語として辞書に登録する．このような語をMSと表記し，Most Certain Segmentと呼ぶ．また，各階層で正変換率が40% 以下の語を最も確実性の低い語として辞書に登録する．このような語をLSと表記し，Less Certain Segmentと呼ぶ．LSは，使用の結果再度正変換率が40% を超えた場合には，もとの階層に復帰する．正変換率（CR : Rate of Correct Translation）の定義式を以下に示す．

$$CR = \frac{CF}{CF + EF} \tag{2}$$

表9に各階層の獲得状況および正変換率による条件を示す．

3.4 変換処理

3.4.1 手順

上述のようにして得られた語を用いて変換を行う．変換処理は確実性の高いものより行われる．従って，MS, CS, S1, RS, LSの順に変換が行われる．この際には，まず読みが2文字以上の語のみを対象として変換を行う．確実性の低い，読みが1字の語については，2文字以上の語の処理終了後変換を行う．また，1字語は確実性が低いので両側が確定済みである場合についてのみ変換を行う．

3.4.2 複数の変換候補からの最適な候補の決定方法

変換の際に同一箇所に複数の変換候補が出現する場合には，その中から最適な候補を決定しなければならない．最適な候補を決定する場合には，**3.3.2**の式（1）に示すゆう度評価関数CEFを用いる．このゆう度評価関数が最大のものを選択するが，ゆう度評価関数の値が同一の場合には，以下の順に評価する．

（1） 誤変換度数の最小のもの．
（2） 正変換度数の最大のもの．
（3） 出現頻度の最大のもの．
（4） 一致した文字数の最大のもの．
（5） 一致した位置が最も左のもの．
（6） 辞書の登録順が最新のもの．

4. 適応能力評価実験

4.1 実験手順

本手法に基づく実験システムを作成し，本手法の適応能力を評価するための実験を行った．用いた資料は，表10に示す情報処理，機械工学，応用化学工学，人文科学各分野の論文各10編，計40編，259,187文字である．辞書は空の状態から始め，1文ごとに図1に示す処理を行った．なお，本実験では式（1）の係数α, βはそれぞれ2, 8に，式（2）における語の所属階層の更新を行う時期をCFとEFの和が3以上として行った．これらの値は，予備実験の結果，最適とされた値である[12]．この予備実験では，表10の（1）の1〜5までの論文5編を用い，αを2に固定し，βの値を変化させて極大値を決定した．また，$CF + EF$は3, 5, 10の各場合について行った．また，この予備実験でも辞書の初期状態は空である．

本実験においてα, βはそれぞれ2, 8であるので，誤変換が正変換の4倍の重みをもつ．すなわち，式(1)である語のＣＥＦの値はその語の変換を1度誤ると4回正変換を行わないともとに戻らない．

4.2 実験結果および考察

変換率の定義を以下に示す．

$$\text{正変換率} = \frac{\text{正変換文字数}}{\text{総文字数}} \tag{3}$$

$$\text{誤変換率} = \frac{\text{誤変換文字数}}{\text{総文字数}} \tag{4}$$

表10　実験に用いた資料
Table 10　Data for the experiment.

（1）分野：情報処理　出典：情報処理学会論文誌				
No	論文名	著者名	巻名	文字数
1	高集積マイクロコンピュータに適したマイクロプログラム制御方式	前島他	Vol.23 No.1	10,946
2	ＣＯＢＯＬマシンとその設計思想 － ハードウェア構成について －	山本他	Vol.23 No.1	8,184
3	フーリエ変換を用いたテクスチュアの構造解析	松山他	Vol.23 No.2	7,517
4	日本語文入力用カタカナ語検出規則とオンライン国語辞典の一分析	木村	Vol.23 No.2	8,485
5	インテリジェント・コンソール －ＯＳの機能拡張の一方法 －	有田	Vol.23 No.3	9,074
6	ポータブル画像処理ソフトウェア・パッケージＳＰＩＤＥＲの開発	田村他	Vol.23 No.3	9,269
7	グラフィック・ディスプレイ・ターミナルのための端末作画システム	高藤他	Vol.23 No.4	6,279
8	オペレーティング・システムのファームウェア化対象選定法	長岡他	Vol.23 No.4	7,179
9	プログラム階層構造の生成，処理，文書化能力を有するテキスト・エディタ	酒井他	Vol.23 No.4	7,698
10	パステストに本質的な分岐に着目した網ら率尺度の提案	中所他	Vol.23 No.5	9,306
	情報工学分野の文字数の合計			83,937
（2）分野：機械工学　出典：北海道大学工学部研究報告				
No	論文名	著者名	巻名	文字数
1	格子乱流中における垂直平板の流力特性	有江他	第106号	4,875
2	暖房用ストーブの燃焼性能に関する研究（第１報）	園田他	第106号	7,274
3	暖房用ストーブの燃焼性能に関する研究（第２報）	園田他	第106号	5,321
4	任意に調節可能な座標線密度をもつ流れ場内格子点網の創成法	飯田他	第107号	5,964
5	円弧切欠きと荷重端の干渉について	岸田他	第108号	2,090
6	換気回数の低い室内における都市ガスの燃焼と一酸化炭素の発生	伊藤他	第105号	7,097
7	チェーンソーの振動におよぼす切削条件の影響	金内他	第104号	6,764
8	境界層剝離の近似的な推定法	知名他	第 98号	4,046
9	任意形状を有する四辺形膜の自由振動	入江他	第 96号	3,896
10	疲れ強さにおよぼす加工効果および残留応力の影響 － 低温燃焼の場合 －	秦他	第 97号	8,163
	機械工学分野の文字数の合計			55,490
（3）分野：応用化学工学　出典：北海道大学工学部研究報告				
No	論文名	著者名	巻名	文字数
1	ニオブ酸カリの凍結に関する研究	小平他	第112号	3,226
2	混合粒子から成る気固系流動層の粒子混合と分級	千葉他	第102号	6,769
3	有機アルミニウム化合物存在下での環化重合	横田他	第102号	5,048
4	高分子のメカノケミストリー（Ⅰ）（綜報）	相馬	第102号	9,811
5	高分子のメカノケミストリー（ＩＩ）（綜報）	相馬	第108号	10,459
6	混合粒子系噴流層の流動特性	上牧他	第 98号	4,317
7	炭酸ガスの有効利用に関する研究	杉岡他	第105号	4,847
8	3－シジクロヘキシルボリルアクリル酸エチルのアルコール中での光化学反応	徳田他	第104号	6,578
9	2－Alkylbenzimidazolの改良合成法	高田他	第 96号	4,447
10	金属酸化物と銅からなる混練触媒のキャラクタリゼーションおよびメタノールリフォーミング反応	小林他	第102号	5,100
	応用化学工学分野の文字数の合計			60,602
（4）分野：人文科学　出典：「美と宗教の発見」（梅原　猛著）				
No	論文名	頁数		文字数
1	明治百年における日本の自己認識	pp. 65 - 76		8,400
2	明治百年における日本の自己認識	pp. 77 - 83		5,173
3	明治百年における日本の自己認識	pp. 83 - 92		6,078
4	美学におけるナショナリズム	pp. 93 -103		5,852
5	美学におけるナショナリズム	pp.103 -112		6,541
6	浄土教的感情様式について	pp.258 -262		3,552
7	浄土教的感情様式について	pp.263 -275		7,803
8	創価学会の哲学的宗教的批判	pp.276 -282		4,698
9	創価学会の哲学的宗教的批判	pp.282 -286		3,507
10	創価学会の哲学的宗教的批判	pp.287 -296		7,574
	人文科学分野の文字数の合計			59,158
	実験に用いた資料（1），（2），（3），（4）の合計			259,187

表11 全資料学習後の辞書登録語数
Table 11 The number of registered words in the dictionary after learning of all data.

階層	語数
MS(Most Certain Segment)	954
CS(Common Segment)	1,705
S1(Segment One)	4,939
RS(Remained Segment)	2,826
LS(Less Certain Segment)	689
合計	11,113

表12 変換結果の例
Table 12 Examples of the translation results.

1．正しいテキスト
ソフトウェアのテストを効率的かつ効果的に行うために，プログラム内の全分岐方向のテスト実行をめざす分岐テストに従来から用いられていた網ら率尺度は，品質が過大に評価されたり，冗長なテストケースが選択されやすいという欠点があった．

2．変換結果
[ソフトウェア]<の>[テスト]<を>[効率]+的+<かつ>(効果)+的+<に><行>[うため]<に>,[プログラム]<内><の>+全+[分岐](方向)<の>[テスト][実行]をめざす[分岐][テスト]<に><従来>(からも)+値+[いられ](ていた)<網><ら><率>[尺度]<は>,[品質]が(か(第二)(評価)[されたり],<冗長><な>[テスト]<ケース><が>[選択]<され>やすい[という]+け+<って>んが[あった].

[]: MS, (): CS, <>: S1, +: RS, |: LS
注）変換結果で下線部は誤変換を示す．

$$\text{未変換率} = \frac{\text{未変換文字数}}{\text{総文字数}} \tag{5}$$

ここで，正変換文字数とは正変換に使用された文字数と未変換であるが，結果として正しかった文字数を含んでいる．これは．本実験では未変換はすべてひらがなで出力されるので，漢字かな交じり文で本来ひらがなの場合には正しい結果となる場合を含むということである．また，誤変換文字数とは誤って変換された文字数である．また，未変換文字数とは変換対象とならなかった部分で結果として誤った文字数である．これは，上述したように本実験では未変換はすべてひらがなで出力されるので，漢字かな交じり文で本来ひらがなでない部分の未変換が誤りとなる場合を含むということである．

表11に全資料学習後の辞書登録語数を示す．図2に各変換率の推移を示す．図2で縦線の実線は，各分野の境界を表している．図2に示すように正変換率は辞書が空の状態から始めても10,000文字程度の学習で約90% まで上昇している．また，図2の実線で示した各分野の変化点では一時的に正変換率は落ちるが，すぐに回復し，各分野の終わりには約95% の正変換率を示している．このことは，本手法の学習機能の有効性を示していると考えられる．また，情報処理，機械工学，応用化学工学という工学分野の論文を学習した後に人文科学という全く異なる分野の資料に対してもすぐに適応している．この場合には，逆に分野の変化による一時的な落ち込みが見られない．これは，辞書が十分に大きくなり分野の変化に対して適応が早くなったことと人文科学の最初の論文が比較的長い論文であったために，1文ごとの学習で最初の論文の終わりまでにシステムが既に適応してしまったためと考えられる．

図2のグラフで最初の情報処理分野の4番目の論文で正変換率の低下が見られるのは，この論文が日本語の表記をテーマとした論文であり，日本語の誤りが例として含まれているためである．しかし，論文4の中を章ごとに見ると次第に正変換率は上昇している[13]ので，解析型における静的な手法では変換が困難な資料も更に同様の資料の学習を続けることにより他の論文の正変換率と同程度にまで上昇するものと考えられる．

表12に変換結果の例を示す．表12の例は情報処理分野の10編目の論文の一部なので，精度が約95% と高く，誤りは3箇所のみである．これらの誤りの原因は，複数の変換候補が重複して出現した場合の最適な候補の決定を誤った結果である．本手法では，**3.3.3** で述べたように五つの階層に語を分類し，これらの階層の確実性の高いものより語を変換し，同一の階層中では式（1）に示すゆう度評価関数により最適な候補を決定する．このような誤りは，本手法のように確実性を用いている限りいくらかは必ず残る．一方，人間の場合には，確率的情報を用いるのと同時に前後の意味的な文脈情報より正しい語を決定していると考えられる．従って，今後は他の語との共起関係を学習により獲得することにより，意味的な文脈情報を学習するメカニズムについての考察を進め，よりいっそうの精度の向上を図る必要がある．このような文脈情報の導入は，変換精度の向上に寄与するばかりでなく，誤り方が人間の誤り方に近くなるので，誤りに対するユー

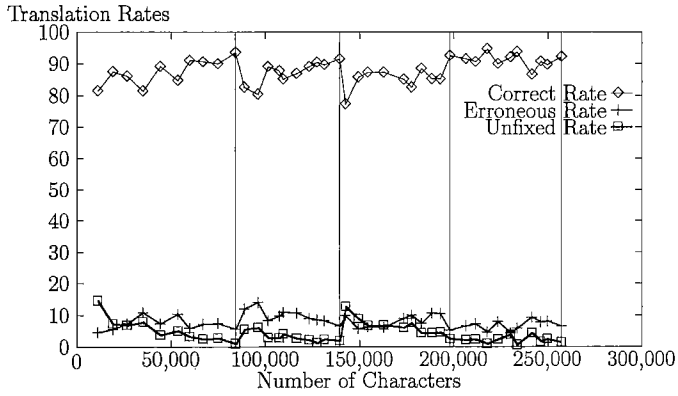

図2　変換率の推移
Fig. 2　Changes in the translation rates.

ザの不快感の軽減にもつながるものであり，この点での改善は非常に重要であると考えられる．

次に，誤り内容の分類とその本手法による改善の可能性について述べる．なお，以下の文中の記号の意味は表12と同様である．

(1) 未変換

例えば，

あせちれん2 < 7) >< も >，< γ− >< アルミナ >< から >< 上 >< と > [同様] < な > [メカノ][ケミカル] < な > [手法] < で > [重合] < する > [ことが][見い出][された].

における「あせちれん2」である．未変換の語は，出現状況にあいまいさのない文からは学習により獲得されるので，次第に変換される．従って，本手法では解決可能な誤りである．

(2) 同音異義語

例えば，

[インタフェース] < で > [必要] < に >< 応 >< じて > < 容易 > < する >．

における「容易」である．正しくは「用意」である．本研究の目的である人間の言語および知識獲得能力の解明とその工学的応用という観点から，本研究では同音異義語だけを特別に処理するという知識を与えるのではなく，このような知識を生得的な能力より獲得することを目的としている．従って，現時点では同音異義語も全く異なる単語候補として扱っている．そのために，もしこの前後の文章中で交互に「容易」と「用意」が出現するような場合には，本手法を用いてもさほど改善されない．これは，本手法が誤りを学習してその語の優先度を下げるという方法をとっているためである．従って，本手法が有効となるのは，どちらかの語がある特定の範囲で片寄って出現する場合である．本手法は，このような状況が多いことを前提として開発されている．また，一般に同音異義語はそれと共起する語によってその識別が可能であるので，本手法の枠組みの中でも前後の語を含むセグメントを残すという方法で交互に出現する同音異義語を正しく変換できる可能性がある．すなわち，「容易」と「用意」が出現する前後の語を含めてセグメントS1として学習されている場合である．しかし，複数の語を含むS1は次第に分解されるので，このように有効なS1を選択し，そのまま残す必要がある．この問題については今後の課題とする．

(3) 語の一部が他の語で当てはまることによる誤り

例えば，

(自動)うん < 点 >[のための][コンソール][メッセージ]

における「うん < 点 >」である．この誤りの原因は，「運転」が未登録の語であった場合と「運転」より「点」が優先されてあてはめられた場合の二つが考えられる．いずれの場合にも誤りが起こることにより「点」より「運転」の階層およびゆう度が上昇するために，次第に「運転」が当てはめられるようになる．しかし，「点」が他の部分で頻繁に正しく当てはめられると正変換度数

も上昇するので階層もゆう度も下がらない．また，一般に短い語は長い語に比べて出現頻度が高いので，長い語より優先して当てはめられる場合が多くなる．これを防ぐためには，最長一致のヒューリスティックスを用いる必要がある．本手法では，**3.4.2**で述べるようにゆう度を評価する場合の非常に遅い段階でしか最長一致のヒューリステックスを用いていない．これは，本研究が人間の言語および知識獲得能力の解明とその工学的応用を目的とし人間の生得的能力を「二つの事物が同じか異なるかを判断する能力」と仮定して，本手法を構築しているためである．すなわち，最長一致法を優先して用いると本手法の有効性を確認することが困難となるために，できるだけその導入を制限している．より実用的なシステムを構築する場合には，最長一致法の導入はその精度の向上に有効である．しかし，本研究の主旨からは最長一致のヒューリスティックスを本手法の枠組みの中で学習機能により自動的に獲得できるようにする方向で研究を進めることが本質的な解決方法と考えられる．

また，本手法は初めに単語ではなく複数の単語を含むセグメントを学習するので，「容易」と「用意」が出現する前後の語を含めてセグメントとしてＳ１に登録されている可能性がある．この際に，「容易」と「用意」で前後の語が異なれば上述のような交互に同音異義語が出現する場合でも正しく変換できる可能性がある．これも広い意味で文脈を学習していると考えられる．しかし，本手法では漢字とかなの連続したS1は抽出されず，しかも学習が進むにつれ複数の語を含むＳ１は一つの語に分解され，そのＳ１は削除される．従って，変換に有効な複数の語を含むセグメントをどのようにして合成するのか，またそのようなセグメントをどのようにして残すのかが今後の課題となる．

また，本手法において，分野１を学習し，次に分野２を学習し，再度分野１を学習した場合の精度について考察する．いわゆる学習の振動という問題である．このようなデータを処理する場合には，本手法は学習に基づく手法であるので，最初分野１を学習したときと全く同じ精度で即座に再度，分野１を変換することはできない．しかし，分野２を学習した際に完全に分野１を忘れるわけではなく，そのときに学習した語が下位の階層に残っているので，最初分野１に適応した期間よりも短い期間ですみやかに分野１に適応できる．これは，図２に示すように工学から文学という対象分野の大きな変化に対しても辞書が十分に飽和していれば分野の変化による正変換率の減少がなく，すみやかに適応していることからも推測できる．また，このような学習の振動は特定分野の専門語いよりも，むしろ種々の分野に頻繁に出現する平易な語いでよく起こるものと考えられる．そこで，平易な語いとして「している」を取り上げ，各分野の中間に位置する各分野の５番目の論文での変換結果の推移を調査した．結果を以下に示す．

- 情報処理分野の５番目の論文

＜知＞[ている]→＜し＞[ている]→＜し＞[ている]→をし[ている]→＜し＞[ている]→＜し＞[ている]

- 機械工学分野の５番目の論文

出現せず

- 応用化学工学分野の５番目の論文

＜知＞[ている]→＜知＞[ている]→＋にし＋[ている]→＜知＞[ている]→＜知＞[ている]→＜知＞[ている]→＜知＞[ている]→＜し＞[ている]→＜し＞[ている]→＜し＞[ている]→＜し＞[ている]→＜とし＞[ている]→＜し＞[ている]→＜し＞[ている]→＜し＞[ている]→＜し＞[ている]→＜知＞[ている]→(開始)[ている]→＜し＞[ている]→＜し＞[ている]

- 人文科学分野の５番目の論文

＜し＞[ている]→＜知＞[ている]→＜し＞[ている]→＜とし＞[ている]→どし[ている]

この実験結果より「＜し＞[ている]」と「＜知＞[ている]」の間で学習の振動が起きていることがわかる．すなわち，情報処理分野で一度正しく「＜し＞[ている]」と学習されたにもかかわらず，応用化学工学分野では「し」に対する「知」のゆう度が増大し，「＜知＞[ている]」という誤変換が頻繁に起きている．「し」と「知」の選択は，「し」と「知」が同一階層Ｓ１に存在するので，ゆう度評価関数を用いて行われる．この誤りを防ぐためには，誤変換度数の係数βを大きくすることが考えられる．しかし，そうすると今度は「知」が当てはまるべき箇所で「し」があてはまるという誤変換が増大する．従って，このような誤りの抜本的改善のためには解析的な文脈処理を行わねばならない．本手法のような確率的手法では，この点に関しては最適な値を実験的に探す以外に方法がない．しかし，人文科学分野では再び正変換の数が増大している．これは，再び「知」より「し」のゆう度が大きくなったためである．このように変換は文章中の語の出現状

況に依存するので，これも広い意味で文脈処理を行っていると考えることができる．しかし，構文情報，意味情報を利用した解析的な決定的手法における文脈処理に比べて，誤りを学習するという本手法の性質のために，非常に短期的な変化には追従できない．従って，本手法は一定の傾向で出現する同音異義語の処理に有効である．多くの同音異義語は分野あるいはユーザによってその出現傾向は一定であると考えられるので，本手法は同音異義語の処理にも有効であると考えられる．また，前述したように本手法の枠組みの中でも前後の語を含むセグメントを合成し，そのセグメントを残すという方法で交互に出現する同音異義語を正しく変換することが可能であるが，この問題については今後の課題とする．

5. むすび

人間の言語および知識獲得過程の考察より得られた仮説に基づく語の獲得手法および確実性を用いたべた書き文のかな漢字変換手法を提案し，その学習機能による適応性能を評価する実験を行った．実験の結果，異なる四つの分野において90% 以上の精度が得られ，べた書き文かな漢字変換における本手法の有効性が確認された．

また，かな漢字変換では個人あるいは分野を限定すれば3,000語程度の辞書で十分であり [14], [15]，頻度上位500語程度を用いれば個人あるいは特定分野で使用する語の35%程度を網羅する [16] ことが確認されている．従って，最初に非常に精度の高い少数の語で変換し，以後確実性の高い語より順に変換するという本手法は有効であると考えられる．また，個人あるいは特定分野にシステムを適応させることにより精度が向上することも確認されている [14], [15]．従って，現在市販されているかな漢字変換システムのようにすべてのユーザが数万語の辞書をもつことは，不要な単語候補や同音異義語の増加によるかな漢字変換の精度の低下および処理量の増大という点から決してよいことではない．これに対して，我々は従来より，あらかじめ人手により辞書に登録された語を対象に，その統計情報を用いて個人あるいは特定分野で使用される語を更新する手法 [14], [15] を提案している．しかし，このような手法でも結局，初期辞書作成および新語登録の労力は避けられない．これに対して本論文で提案する手法は，入力べた書き文と校正済みの変換結果より帰納的学習により語を獲得できるので，個人あるいは特定分野の10,000文字程度の入力べた書き文および校正済みの変換結果が得られれば，それらより個人用あるいは特定分野用の辞書を自動的に作成できる．以後は，本実験結果に示すように学習機能を用いて個人あるいは分野に動的に適応できるので，登録語数を限った小規模な辞書で高い精度のかな漢字変換を行うことが可能である．

今後は，更に科学的立場から人間のヒューリスティックスの獲得メカニズム，文脈情報などの意味を獲得するメカニズムについての考察を進め，かな漢字変換はもとより機械翻訳システム，質問応答システムなどへの工学的な応用を図る予定である．

また，実例からの帰納的学習によって言語および知識を獲得するのとは別に，人間の場合でも言語あるいは知識をメタ的に与えられることは日常的に起こっている．従って，実例からの帰納的学習とメタ的な知識の教示とのかかわり合いについても考察を進める必要がある．

謝辞 北海学園大学在学中実験等に御協力頂いた佐藤隆之氏（現，NEC（株）），富居広樹氏（現，コアシステム（株））に感謝します．また本研究の一部は，文部省科学研究費補助金（奨励研究A第06858039号）によって行われている．

文　献

[1] 赤間　清, “未知言語環境における帰納的学習のモデル,” 情処学論, vol.28, no.5, pp.446–454, 1987.
[2] 赤間　清, “知識はいかに獲得されうるか,” 認知科学の発展, vol.1, pp.161–188, 講談社, 1988.
[3] 錦見美貴子, 中島秀之, 松原　仁, “一般学習機構を用いた言語獲得の計算機モデル,” 認知科学の発展, vol.5, pp.143–185, 講談社, 1991.
[4] 小林春美, “アフォーダンスが支える語彙獲得,” 月刊 言語, vol.21, no.4, pp.37–45, 大修館書店, 1992.
[5] H.H. Clark and E.V. Clark “Psychology and Language,” Harcourt Brace Jovanovich, Inc., 1977.
[6] 荒木健治, 栃内香次, “帰納的学習による語の獲得および確実性を用いた語の認識,” 信学論 (D-II), vol.J75-D-II, no.7, pp.1213–1221, July 1992.
[7] 荒木健治, 栃内香次, “帰納的学習によるべた書き文のかな漢字変換,” 信学技報, NLC91-38, 1991.
[8] 高橋祐治, 荒木健治, 桃内佳雄, “帰納的学習によるべた書き文のかな漢字変換の有効性,” 情処研報, vol.93(93-NL-93), no.1, pp.31–38, 1993.
[9] 荒木健治, 高橋祐治, 桃内佳雄, 栃内香次, “帰納的学習によるべた書き文のかな漢字変換手法の適応能力の評価,” 信学技報, NLC94-3, 1994.
[10] 吉村賢治, 日高　達, 吉田　将, “日本語科学技術文における専門用語の自動抽出システム,” 情処学論, vol.27, no.1, pp.33–40, 1986.

[11] 吉村賢治, 武内美津乃, 津田健蔵, 首藤公昭, "未登録語を含む日本語文の形態素解析," 情処学論, vol.30, no.3, pp.294–301, 1989.
[12] 高橋祐治, 荒木健治, 桃内佳雄, "帰納的学習によるべた書き文のかな漢字変換における尤度の評価について," 平4北海道連大, pp.409–410, 1992.
[13] 荒木健治, 栃内香次, "帰納的学習によるべた書き文かな漢字変換における適応能力の評価," 情処第44回全大, vol.3, pp.183–184, 1992.
[14] 栃内香次, 斉藤 康, "適応型変換辞書を用いるかな漢字変換," 情処学論, vol.24, no.2, pp.209–220, 1983.
[15] 栃内香次, 岡沢好高, "適応型変換辞書方式かな漢字変換システムの性能測定," 情処学論, vol.26, no.4, pp.733–739, 1985.
[16] 荒木健治, 栃内香次, "多段階分割法によるべた書き日本語文のかな漢字変換," 情処学論, vol.28, no.4, pp.412–421, 1987.

(平成6年9月8日受付, 7年7月10日再受付)

荒木 健治 (正員)

昭57北大・工・電子卒. 昭63同大大学院博士課程了. 工博. 同年, 北海学園大学工学部電子情報工学科助手. 平成元年同講師. 平成3年同助教授. 1992.9〜1993.8スタンフォード大学CSLI客員研究員. 自然言語処理, 音声言語情報処理, 機械学習の研究に従事. 情報処理学会, 日本認知科学会, 人工知能学会, 言語処理学会, IEEE, ACL, AAAI, ACM各会員.

高橋 祐治 (正員)

平2北海学園大・工・電子情報卒. 平4年同大大学院修士課程了. 同年より北海道ソフト・エンジニアリング(株)に勤務. 自然言語処理に関する研究に従事. 情報処理学会会員.

桃内 佳雄 (正員)

昭40北大・工・精密卒. 昭42同大大学院修士課程了. 同年(株)日立製作所入社. 昭47年北大大学院博士課程単位取得退学. 昭48北大大学院情報工学専攻助手, 昭59講師, 昭61助教授. 昭和63北海学園大学工学部電子情報工学科教授. 自然言語の理解と生成に関する研究に従事. 工博(北大). 情報処理学会, 日本認知科学会, 計量国語学会, 言語処理学会各会員.

栃内 香次 (正員)

昭37北大・工・電気卒. 昭39同大大学院工学研究科電気工学専攻修士課程了. 現在, 北大・工・電子情報工学専攻教授. 主として音声情報処理, 自然言語処理の研究に従事. 工博. 情報処理学会, 日本音響学会各会員.