仮名漢字変換における最近使用語優先学習方式のモデル化

下村秀樹[†] 酒井貴子[††] 並木美太郎[†††]
中川正樹[†††] 高橋延匡[†††]

仮名漢字変換において，最近の変換に使用した語を優先的に扱う学習方式（最近使用語優先学習方式）は，変換性能の向上に大きな効果がある．本研究では，計算機システムの仮想記憶管理におけるワーキングセットの概念を用いて，この学習方式をモデル化し，実際の文章でそれを検証した．モデル化では，学習効果の現れる必要条件が満たされる確率として，学習語生存確率の概念を導入した．この確率は，文章中での単語使用間隔の確率分布と「最近」の定義範囲の長さから計算できる．次に，仮想記憶のようにバッファに単語を登録する方式で最近使用語優先学習を実現し，それを仮名漢字変換に組み込み，単語使用間隔の分布の測定，学習語生存確率の実測を行った．さらに，学習語生存確率と変換精度の関係も調べた．その結果，学習語生存確率は提示したモデルから解析的に求めることができるとともに，バッファの登録可能語数からその近似値が求められることを確認した．また，学習語生存確率と変換率（欲しい変換結果が第1候補である割合）は，正の相関関係にあることも明らかになった．したがって，文章の単語使用間隔の分布をもとにして，学習語生存確率を求めれば，最近使用語優先学習が変換精度に与える効果をかなり解析的に議論することができる．

A Model of Recently Used Word Preference Method in Kana-to-Kanji Translation

Hideki Shimomura,[†] Takako Sakai,[††] Mitarou Namiki,[†††]
Masaki Nakagawa[†††] and Nobumasa Takahashi[†††]

In kana-to-kanji translation the employment of the recently used word preference (RUWP) learning method, which gives high priority to the words used in recent translation, has a substantial effect on the success rate of the translation. This paper presents and verifies the model of this learning method by an analogy to the concept of working set model for the virtual memory management of computer systems. The model introduces learned word survival probability (LWSP) as the probability that the necessary condition for the learning to be effective is satisfied. This probability is derived from the distribution of the intervals of using the same word in text and the length of the period defined as "recent". The RUWP method was implemented with the recently used word buffer and built into a kana-to-kanji translation system to obtain the interval distribution and LWSP in real text and to verify the relation between LWSP and the translation rate. This verification confirms that LWSP is analytically derivable from the model and its lower bound can be estimated from the buffer size. The verification has also shown that LWSP is positively correlated with the correct translation rate.

1. はじめに

仮名漢字変換において，変換精度を高めることは重要である．従来から，2文節最長一致法[1]，文節数最小法[2]などを基本に，文法情報や意味情報，学習情報（一連の変換でどのような単語を使用したかなどの情報）を利用して変換精度を高める研究が行われてきた．その成果の多くは市販の日本語ワードプロセッサにも搭載されるなど，実用化されている．

我々は，仮名漢字変換技術の現状を認識し，より変換精度の高い変換処理を実現することを目的として研究を行っており，その一環として市販ワードプロセッサの変換性能比較調査を行った[3]．その結果，学習情報の利用方法が変換性能の差に大きな影響を及ぼすことが明らかになった．また，学習方式としては，ほと

† NEC 情報メディア研究所
Information Technology Research Laboratories, NEC Corporation
†† 日立製作所システム開発研究所
Systems Development Laboratory, Hitachi Ltd.
††† 東京農工大学工学部電子情報工学科
Department of Computer Science, Faculty of Technology, Tokyo University of Agriculture and Technology

んどの機種が最近使用した語を含む変換結果を優先する方式（以下，「最近使用語優先学習方式」と呼ぶ）を採用していることがわかった．

さて，最近使用語優先学習方式を仮名漢字変換に利用する場合，次の要因が変換精度に影響を与えると思われる．

（1）どれくらい前に使用された語までを優先するか（「最近」の範囲の定義）

（2）変換結果の優先度を決める学習以外の情報（文節数最小，最長一致，2文節最長一致，文法接続検定など）とどちらを優先させるか

文献[3),4)]の調査におけるワードプロセッサの変換性能の違いは，仮名漢字変換があらかじめ用意している辞書（以下，「基本辞書」と呼ぶ）の語彙や同音語の優先順位の問題もあるが，上記の二つの要因をどう設定するかの影響が大きい．例えば，優先する語の範囲を非常に小さく採ると，学習効果が短期間で消えてしまい，実質的に現れないということも起こり得る．したがって，最近使用語優先学習方式を有効に機能させるために，これらの要因と変換精度の問題は，十分検討すべき問題である．しかし，この学習をモデル化したり，変換性能への影響を定量的に評価したりした例はなく，多くのシステムでは経験的に設定しているのが現状である．

本論文では，仮名漢字変換の精度を上げる重要な要因として学習を採り上げ，そのモデル化とモデルの検証を試みたので報告する．まず第2章では，最近使用語優先学習方式のモデルを，計算機システムの仮想記憶管理におけるワーキングセットの概念と対比して提示する．ここでは，最近使用語優先学習方式を評価する重要な尺度として，「学習語生存確率」の概念を導入する．続く第3章では，その実現方式とモデルの関係を議論する．第4章と5章では，実際の文章に対してこのモデルをあてはめて実験を行い，学習語生存確率の推定，実測，また学習語生存確率と変換性能の関係を述べる．

2. 最近使用語優先学習方式

2.1 最近使用語優先学習のモデル

仮名漢字変換における学習とは，一般的に「変換処理にそれ以前の変換結果候補選択の情報を反映させること」である．この定義の中には，意味的な学習，構文的な学習も含まれる．最近使用語優先学習とは，その中でも，「最近使用した語を含む変換結果候補を優先的に出力する」という学習方式であるといえる．

次に，最近使用語優先学習方式を単純にモデル化する（**図1**）．

まず，ある一連の仮名漢字変換入力において，変換結果候補から選択された単語を，その順番に，

$W_1, W_2, W_3, \cdots, W_k$

とする*,**．

このとき，W_k からさかのぼって t 単語の範囲に含まれる語を「最近使用したので優先すべき語」と定義する．この範囲を「学習区間」，t を「学習区間長」，学習区間に含まれる語を「学習語」と呼ぶ．したがって学習語は，

$W_{k-t+1}, W_{k-t+2}, \cdots, W_k$

である．さらに学習語の中で同じものがあれば，最も後に現れたものだけを残した集合を $R(t, k)$ とする．このとき，最近使用語優先学習とは，次の仮名漢字変換での変換結果候補の中で，$R(t, k)$ 内の単語を含むものを，含まないものより優先する方式であるといえる．また，一つの読みに対して，$R(t, k)$ 内の単語 W_x，$W_y(x>y)$ を含む二つの変換結果候補があれば，

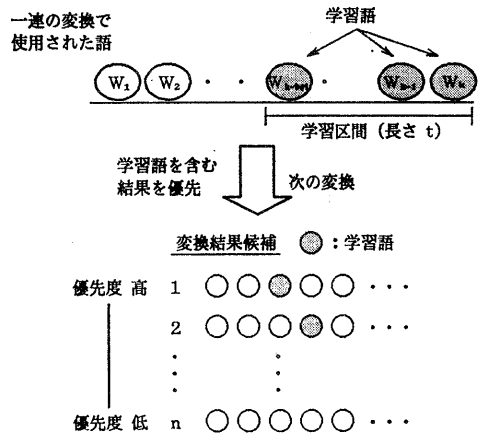


図 1　最近使用語優先学習方式
Fig. 1　Model of Recently Used Word Preference Method in Kana-to-Kanji Translation.

---

* ここでは，仮名漢字変換の辞書に格納されているもの，あるいはそれと同等に扱えるものを単語と考える．したがって，活用語尾や複合語に関して，どの部分を単語と見なすかについては，変換システムでの定義によって異なる．また，助詞や助動詞といった付属語は，単語ではあるが，学習の対象とは考えないのが一般的である．

** このときの添字 1～k は，単語を使用した論理的な時刻を示していると見ることができる．以下，単語に関する「時刻」という表現は，すべてこの論理的な時刻を指す．

より最近使用された語である W_x を含む変換結果を優先するのが，最近使用語優先の考えからは自然である．ただし，これは最近使用語優先学習方式についてだけの議論であり，文法，最長一致法，文節数最小法などのさまざまな情報も使って変換結果候補のもっともらしさを評価するときには，最近使用された語を含む候補が必ず優先されるかどうかはわからない．

2.2 同音語の扱い

最近使用語優先学習は，同音語の選択に大きな効果を発揮することが知られている．例えば，初期状態で，

「かいとう」→「解答」*

という単語を変換結果の第1候補として出力する仮名漢字変換が，「回答」をユーザが選択した後には，

「かいとう」→「回答」

を第1候補に出すことができる．変換結果候補を受け取る側からこの現象を見ると，同音語の間での優先順位の変更（学習）が行われたと理解できる．

前節で示した最近使用語優先学習のモデルは，同音語かどうかの区別なく，学習区間に含まれる語を学習語と定義し，それを変換処理で優先する．しかし，この考え方でも，学習区間にある語に関しては最近使用語優先となるし，それは同音語に関しても例外ではない．したがって，このモデルでも同音語の優先順位の変更を表現できる．

さらに，このモデルでは読みの異なる語に対しても最近使用の優先順位を定義している．例えば，初期状態で，

「つかうがいかはない」→「使うが以下はない」

という変換結果を第1候補としたが，ユーザは「使う外貨はない」を選択したとする．このとき，「以下」と「外貨」は読みが違うので，同音語の優先順位の変更だけでは，以降の変換で，

「つかうがいかはない」→「使う外貨はない」

を第1候補にすることが難しい．しかし，提案したモデルでは，「以下」と「外貨」の間の優先関係も定義しており，その情報を優先すれば，一般にいわれる「文節の区切り方の学習」を行ったのと同様の効果を，簡単に得ることができるという利点がある．

ただし，この方式では，学習区間からはみ出した語に関して最近使用語優先の情報が消えてしまう．したがって，大量に変換を行っていると，ある同音語の優先順位が，読みの異なる語の使用によって消える場合もあり，提示したモデルがどんな場面でも優れているとは断定できない．しかし，本研究では，以降に示すような解析的議論が行えることに着目し，提示したモデルに従って議論する．

2.3 学習語生存確率

以上のモデルから，最近使用語優先学習の効果が現れる必要条件は，変換結果として得たい語が，$R(t, k)$ に含まれていることである．したがって，その確率は最近使用語優先学習方式を評価する一つの尺度である．

この確率の推定は，計算機システムの記憶管理におけるワーキングセットのモデルに非常によく似ている[4),5)]．仮名漢字変換の単語は仮想記憶のページに，学習区間は主記憶の容量の近似値に，基本辞書は2次記憶に対応し，$R(t, k)$ はワーキングセットに対応する．このモデルから類推すると，$R(t, k)$ に次の変換で使用する単語が含まれている確率 $P(t, k)$ を k に関して平均した値 $P(t)$ は，単語の使用間隔の確率分布から求めることができる．すなわち，ある単語を使用してから次にその単語を使用するまでの間に別の単語を $X-1$ 個使用する（単語使用に関する論理的な時間差が X であるときの）確率分布 $F(X)$ とすると，

$$P(t)=\sum_{i=1}^{t} F(X=i) \qquad (1)$$

となる（**図2**）．これは，最近使用語優先学習が効く平均確率を表している．この $P(t)$ を「学習語生存確率」と呼ぶ．（注：学習語生存確率は，仮想記憶でのページフォールト率の逆の考え方である．）

仮想記憶では，高速アクセスが可能な1次記憶に，参照するページができるだけ高い確率で残っていることが重要である．その確率が記憶管理方式の性能を直接示す．一方，提示した最近使用語優先学習のモデルでは，使用する単語が学習区間内に残っていることが重要ではあるが，それは変換性能を向上させる必要条件にすぎない．例えば，二つ以上の同音語が学習語である場合の扱いなどもあり，学習語生存確率が変換性

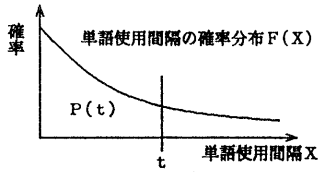


図2 学習語生存確率
Fig. 2 Learned Word Survival Probability.

* 矢印の左が変換前の仮名文字列，右が変換結果の候補の一つを意味する．以後，同様の記述は同じ意味である．

能を直接示すことにはならない．

2.4 学習区間長と変換性能の議論

学習区間長 t を大きく採ると，$P(t)$ が1に近くなることは，(1)式から明らかである．このとき，最近使用語優先学習は非常に高い確率で効く．しかし，学習区間長を大きくすることには，問題点もある．

一つの問題は，学習区間長を長くするに従って学習のために管理する情報が増えるので，辞書検索速度の低下や記憶容量の不足が考えられることである．この問題は比較的よく認識されており，このような現実的な制約から学習区間をあまり長くすべきでないということは確かである．

また，一般には理解されにくいが，仮名漢字変換結果という観点からも，学習区間を極端に長くすることは，必ずしもよい効果をもたらすとは限らない．一般に基本辞書にある単語の優先度は，大量の文章から得た単語の使用頻度に基づいて付けられている．これは，長期的に見た単語の使用確率を予測しており，単語の使用がその頻度に従って完全に均一であると仮定すれば，欲しい結果が平均的によい順位になる．これに対して，最近使用語優先学習方式は，最近使った語はまた使われる，つまり単語の使用には偏りがあるという仮定に基づく予測モデルである（**図3**）．一般に文章を書く場合には，ある内容を記述する．したがって，その内容に関する語はある短期間に何度も使用される．すなわち，単語の使用には偏りがあり，その仮定は妥当であるが，それはあくまでもある短期間を対象とした議論である．

もし，学習区間長 t を極端に大きく採ると，確かに $P(t)$ は1に近くなるが，それとともに単語の優先順位は使用された時刻の新しい順に変わってしまう．すなわち，長期的に見て現れやすい語という情報によって付けられた基本辞書の優先順位が失われ，偶然の単語の出現順番の情報に代わってしまう．この結果，最近使用語優先学習は有効に機能しているものの，欲しい結果を平均的によい優先順位にするための情報が消えてしまい，変換結果を得るための平均的な手間が増大する可能性がある．

以上の二つの点から考えると，単純に学習区間を長くして学習語生存確率を大きくするのではなく，学習区間長を制限することには重要な意味があるといえる．

3. 最近使用語優先学習方式の実現方法

3.1 実現の代表的方式

最近使用語優先学習を行うためには，学習語の情報を格納する必要がある．その方式としては，大きく分けて次の二つが考えられる．

（1）学習語を学習用のバッファ（以下，「学習辞書」と呼ぶ）に格納する

これは，仮想記憶と同じ方式であり，学習辞書が主記憶に，基本辞書が2次記憶に相当する．一般に学習辞書には登録語数の制限があり，辞書が一杯になった場合には，LRU(Least Recently Used) 方式などで古い語を追い出す．登録可能語数の制限は，学習区間の限定と見ることができる．以下ではこの方式を，「学習辞書方式」と呼ぶ（**図4**）．

学習辞書方式では，基本辞書を全く書き換えずに学習を実現できる．すなわち，処理速度の向上のために基本辞書が ROM のような書換え不可の媒体に格納されているときには有効な方法となる．記憶容量の点からは，登録可能語数に比例した容量を必要とするので，登録可能語数が小さい場合には記憶容量を小さくできる．

しかし一方では，LRU の管理が必要であり，データ管理が複雑になるという欠点がある．

（2）最近の変換に使用した学習語であるということを基本辞書に直接書き込む

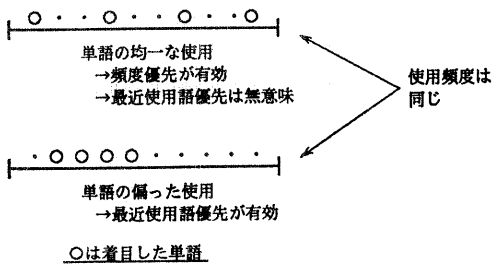


図3 単語使用の偏りと最近使用語優先
Fig. 3 Deviation of Word Used in Text and Recently Used Word Preference.

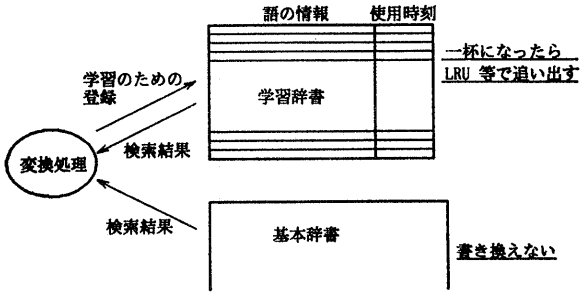


図4 学習辞書方式
Fig. 4 Learned Word Dictionary Method.

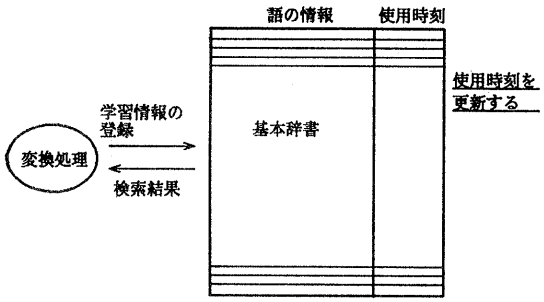


図 5 基本辞書書換え方式
Fig. 5 Basic Dictionary Rewriting Method.

この方法は，学習の情報を基本辞書に直接書き込む．したがって，基本辞書（の学習に関する情報）が書換え可能な場合にだけ採用できる手法である．実現として最も簡単なのは，基本辞書の各単語に使用時刻を記録する領域を用意するという方法である．（注：単語の優先順位をポインタなどのリンクで表現している実現方法の場合も，時刻の前後関係が論理的なリンクを定義するので，この方式で議論できる．）以下，この方式を「基本辞書書換え方式」と呼ぶ（**図5**）．

基本辞書書換え方式は，各単語の使用時刻を記録するだけでよく，LRU 等の複雑な管理が必要ないことが大きな利点である．学習区間の限定は，変換処理時に，現在の時刻からある一定以上古い時刻に使用された単語を，学習語として扱わないことで実現できる．

しかし，基本辞書のすべての語に対して時刻を記録する領域が必要になるので，記憶容量が多く必要になる．例えば，10万語の辞書（市販の実用的日本語ワードプロセッサは，最低でもこの程度の語数を持っている）の場合，時刻の領域を2バイトとしても，200kバイトの領域が学習のために必要になる．

この二つの方式には，それぞれ利点・欠点がある．どちらを採用するかは，主に実現上の制約の問題である．

3.2 学習辞書の登録可能語数と学習区間長

基本辞書書換え方式の場合，2.1 節に示した最近使用語優先学習方式のモデルでの学習区間長を，直接定義することができる．すなわち，現在の時刻からある一定時刻以内の単語だけを学習語と見なせばよい．このときの学習語生存確率は，(1)式に従って，単語使用間隔の分布から知ることができる．

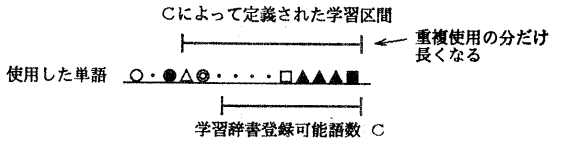


図 6 学習辞書登録可能語数と実際の学習区間の関係
Fig. 6 Relation between the Capacity of Learned Word Dictionary and the Real Learning Period.

学習辞書方式でも，学習辞書に登録されている語に関して，同様の方式で学習区間を設定することはできる．しかし，古い学習語は学習辞書の LRU 管理で自動的に追い出されることを利用して，学習辞書に登録されている語をすべて学習語と見なす方式もある．学習区間内で単語は重複使用されることがあるので，この方が同じ記憶容量で実質的な学習区間を長くすることができ，有利である（**図6**）．以下，学習辞書方式といった場合には，学習辞書内にある語をすべて学習語と見なす方式を指すものとする．このとき，単語の使用状況（重複使用の状況）によって学習区間長が一定しないので，単語使用間隔の分布から学習語生存確率を正確に求めることはできない．ただし，近似的に推定することはできる．

学習辞書の登録可能語数を c とすると，それによって定義される学習区間長 t_c は，単語の使用状況によって一定はしないものの，必ず $t_c \geqq c$ となる．なぜなら，連続した c 語の間で重複使用される語があるので，t_c は異なり語数が c になるまでの単語数だからである（図6）．また，学習語生存確率の性質から，

$$P(t_c) \geqq P(c) \quad (t_c \geqq c) \qquad (2)$$

なので，$t_c = c$ として単語使用確率分布から学習語生存確率を求めると，実際の学習語生存確率の下限値を与えることになる．

4. 単語使用間隔分布の実測と学習語生存確率

4.1 単語使用間隔分布の実測

最近使用語優先学習方式のモデルから，単語の使用間隔の分布がわかれば，基本辞書書換え方式での学習語生存確率，あるいは，学習辞書方式での学習語生存確率の予測値（下限値）を知ることができる．そこで，実際にいくつかの文章から単語使用間隔を測定して，検証を試みた．

実験は，我々が作成した実験用仮名漢字変換システム[6]の上で行った．このシステムの最近使用語優先学習は，記憶容量の節約を考慮し，学習辞書方式を採用している．実験は次の手順で行った．

(1) 学習辞書の登録可能語数を，実用上無限大（LRU による追出しが起こらない状態）にする．

（2）学習を行いながら変換処理を行う．すなわち，変換作業中に選択した語を学習辞書に順次登録していく．

（3）辞書は一杯にならないので，以前に使用された語はすべて学習辞書に残っている．そこで，以前に使用されたときの時刻と今の時刻の差（すなわち，単語使用間隔）を測定し，分布を採る．

変換作業での候補選択処理は，変換精度自動測定ツールによって自動的に行うことができる．このツールは，仮名テキストとそれに対応する変換後の文字列を入力として，変換結果が一致するまでの再変換要求の回数を数え，また，その結果の各単語を学習辞書に登録する．なお，本実験で用いた辞書は登録語数約9万語で，学習は自立語だけに関して行った．辞書中の複数の単語から構成される複合語（名詞の連続など）は，別々に学習辞書に登録し，用言に関しては語幹だけを学習した．

実験では，三つの文章（情報処理学会論文2編：文章A[7]，文章B[8]，卒業論文1編：文章C[9]）に関して単語使用間隔の分布を測定した．それぞれの文章に使われている単語数（自立語）を**表1**に示す．

実験の結果得られた，各文章の単語使用間隔の分布を**図7**に示す．

この結果からまずわかることは，単語使用間隔の短い語が非常に多いということである．どの文章に関しても，複数回使用される語のうちの60%前後が使用間隔100語以内である．また，使用間隔が長くなると急激に割合が減少することもわかった．

使用間隔が短い語には，形式名詞（例：こと，もの），形式動詞（例：なる，する），連体詞（例：この，その），代名詞（例：これ，それ）などが含まれているので，この結果は決して不自然ではない．それ以外の語はもう少し広い間隔で使用されていると思われるが，いずれにしても，文章中の語は短期間に何度も繰り返して使用されることがわかった．最近使用語優先学習方式が機能する前提条件は，「同じ単語は短期間に何度も使われる」ということである．この単語使用間隔の測定結果は，それを裏付けている．

表1　実験用文章の使用単語数
Table 1　Number of Words Used in Text of the Experiments.

文章A，文章B：情報処理学会論文
文章C：卒業論文

	文章A	文章B	文章C
延べ使用語数	2112	2640	7678
異なり使用語数	484	617	1066

注）単語数は自立語だけを対象に数えた．

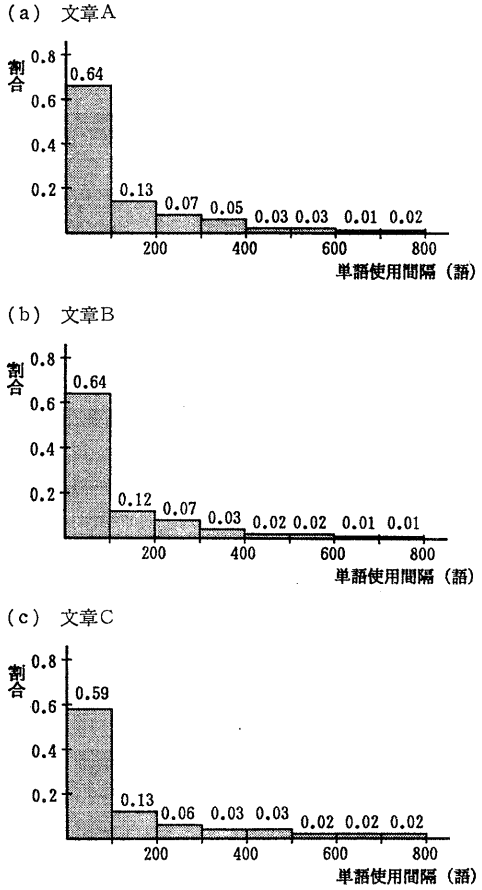


図7　単語使用間隔の分布
Fig. 7　Distribution of the Intervals of Using the Same Word.

4.2　学習区間長と学習語生存確率

次に，図7の三つの分布から(1)式に従って学習区間長ごとに学習語生存確率を計算した．結果を**図8**に示す．

図7からもわかるように，文章によって使用する単語もその使用状況が異なるので，学習語生存確率も，当然文章によってばらつきがある．しかし，どの場合も学習区間長 t が100で0.6前後，300で0.75〜0.8前後，500で0.8〜0.9前後という結果が得られた．文章間の値のばらつきは最大でも0.1以下であった．

この値は，基本辞書書換え方式の場合，学習語生存確率を直接示している．また，さきほども議論したが，学習辞書方式の場合，学習語生存確率の下限値を

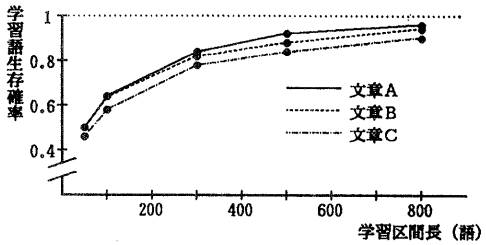


図 8 学習区間長と学習語生存確率の関係
Fig. 8 Relation between the Length of the Learning Period and the Learned Word Survival Probability.

示している．

4.3 学習辞書の登録可能語数と学習語生存確率

学習辞書方式について，登録可能語数を実際に 50, 100, 300, 500 にして，LRU によって辞書を管理しながら学習語生存確率を実測した．この結果をさきほど示した下限値と比較することによって，学習辞書方式の場合に単語使用間隔の分布から，実際の学習語生存確率がどれくらい推測できるのかがわかる．実測値と下限値を，文章ごとに**図 9** に示す．

単語使用間隔の分布（図 7）からも予測できることではあるが，短期間での単語の重複使用が多いので，学習語生存確率は予測値よりもかなり大きくなる．実際には，最大で 0.1〜0.15 程度大きい値になった．

もう少し詳細に結果を観察すると，学習辞書の登録可能語数 c が小さいとき，実測値と予測値の差はあまり大きくないことがわかる．これは，極端に短い期間では単語の重複使用があまりなく，(2)式での t_e と c が近いからである．また，c の増加に従って実測値と予測値の差は大きくなるが，c がある程度以上になると，またこの二つの差は小さくなっている．これは，文章中の使用語彙が有限なので，実測値が飽和するからである．事実，文章Aの場合は文章に使用している異なり単語数が 500 以下なので，学習辞書の登録可能語数を 500 にすると学習語生存確率は 1 になってしまった．

以上の結果から，学習辞書の登録可能語数を学習区間長として学習語生存確率を予測した場合，最大十数パーセントの誤差となることがわかった．また，その誤差は学習辞書の登録可能語数が極端に大きいとき，あるいは極端に小さいときは，小さくなることがわかった．さらに，本実験で使った文章（情報系の論文）に限っていえば，学習辞書の登録可能語数を 500 語程度用意すれば，実際の学習語生存確率は 95% 以上になることがわかった．このとき，学習辞書の各単語の

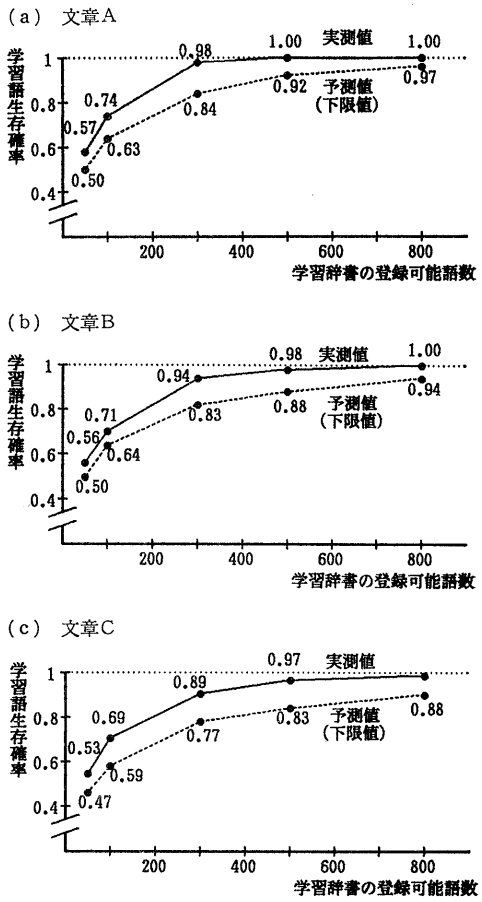


図 9 学習辞書の登録可能語数と学習語生存確率の関係
Fig. 9 Relation between the Capacity of the Learned Word Dictionary and the Learned Word Survival Probability.

情報を 100 バイトとしても，全体で 50 k バイト程度の容量で十分である．したがって，記憶容量の節約をしたい場合，学習辞書方式は有効であるといえる．

5. 学習語生存確率と変換性能の関係

5.1 実験方法

前章までで，学習語生存確率について解析的な議論を行い，さらにそれを検証した．次に，この学習語生存確率が変換性能に与える影響を検討する．

まず，実験にさきだって文節数最小法や 2 文節最長一致法という仮名漢字変換の代表的な基本手法と，最近使用語優先学習をさまざまに組み合わせた場合の変換性能を調べた．その結果，次の二つの方法が，比較的高い変換性能を示した．

(1) 文節数最小法＋最近使用語優先（文節数最小法

をより優先)

(2) 2文節最長一致法＋最近使用語優先（2文節最長一致法をより優先)

次に，この二つの変換手法に関して変換精度を測定し，学習語生存確率との関係を調べた．変換性能は次式に示す変換率で表現した．

$$\text{変換率}=\frac{\text{変換に成功したデータ数}}{\text{全データ数}}\times 100\ (\%)$$

入力データは，単語使用間隔分布や学習語生存確率の測定に用いた3文章を，句読点ごとに区切ったものである．句読点単位に区切ったので，入力データのそれぞれには複数の自立語が含まれることもある．そのすべてが正しく変換されないと変換に成功したとは見なさない．

実験では，最近使用語優先学習方式の効果だけが現れるように，基本辞書の単語の優先順位を，メモリへの単語の物理的配置順（ほぼランダムと考えてよい）とした．また，初期状態の学習辞書は，空にして実験を行った．

5.2 実験結果

結果を**図10**に示す．どの場合も学習語生存確率と変換率はほぼ正の相関関係になっていることがわかる．文節数最小法や2文節最長一致に学習を組み合わせた場合，学習語生存確率は直線的に変換性能に影響を及ぼすことがわかった．二つの手法間の変換性能に若干の差があるが，これは手法の能力の問題であろう．一般的には，文節数最小法の方が広い範囲からの情報を利用して変換を行っている分，変換率が高くなる．

このグラフをさらによく観察すると，学習語生存確率の高い方が，変換率の伸びが大きくなっていることがわかる．これは，入力データが句読点単位であることと変換率の定義式のためである．すなわち，入力単位に二つ以上の単語が入った場合，そのうちどれか一つの単語でも正しく変換できなければ，変換成功とは見なされないからである．

以上の結果から，比較的変換率の高い手法と学習の組合せ方では，学習語生存確率が変換性能の向上に線形的な効果を示すことがわかった．これを逆に考えれば，変換性能の見積りには学習語生存確率の考え方が不可欠であることが示されたといえる．

また，第4章での議論から，学習語生存確率は単語使用間隔の分布から知ることができる（学習辞書方式の場合は下限値の予測値，基本辞書書換え方式の場合

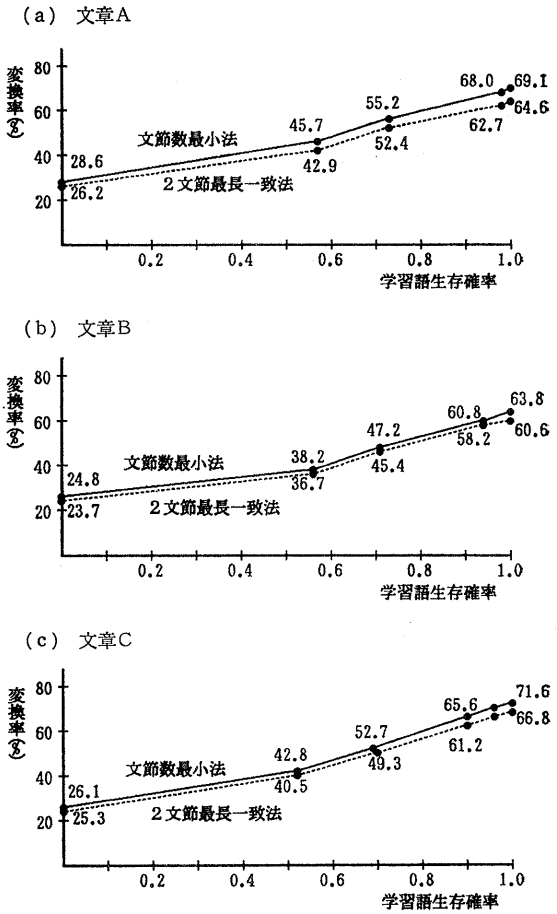


図10 学習語生存確率と変換精度（変換率）の関係
Fig. 10 Relation between the Learned Word Survival Probability and the Translation Accuracy (Success Rate) in Kana-to-Kanji Translation.

は正確な値)．したがって，本論文に示したモデルに基づけば，最近使用語優先学習方式が変換性能へ与える影響をかなり解析的に議論できることがわかった．

5.3 学習の悪影響の議論

これまでの実験からわかるように，最近使用語優先学習方式は，変換性能を大きく向上させる．しかし，一般には学習したことによって変換に失敗するという悪影響もある．実際に本実験でも，「各ツールの検出数は」のように正しく変換できていた例が，「書く」という動詞を学習したことによって，「書くツールの検出数は」のように誤変換した例もあった．このようなことは，学習辞書の登録可能語数に従って多くなると予想される．

今回測定した範囲での平均的な変換率では，その悪

影響を観測することができなかった．すなわち，学習の効果に対して悪影響の起こる頻度は小さいという結果になった．したがって，処理速度と必要とする記憶容量を除いて，本実験の結果を変換の平均変換率という立場で見れば，記憶容量の許す限り，学習区間は長ければ長い方がよいことになる．

しかし，非常に大量な，多くの語彙を使った文章に関しては，最近使用語の範囲をあまり大きく採ると別の悪影響が起こる可能性のあることは，第2章でも議論した．この点は，より大量のデータを使用した実験によって明らかになるであろう．

6. おわりに

本研究では，仮名漢字変換における最近使用語優先学習方式をモデル化し，学習語生存確率という評価尺度を提示した．また，そのモデルの検証を行った．その結果，次の成果を得た．

（1）単語使用間隔分布と学習語生存確率を実測し，提示したモデルを検討した

実際に3文章から単語使用間隔の分布を採り，学習語生存確率を調べた．その結果，単語使用間隔は非常に短いものが多く，最近使用語優先学習方式の有効性を裏付ける「単語使用の局所性」が明らかになった．学習辞書方式では，実測値と最大で十数パーセントの誤差はあるものの，学習辞書の登録可能語数から性能を予測できることがわかった．

（2）学習語生存確率と変換性能の関係を評価した

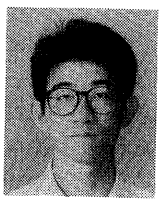
学習語生存確率は，変換性能に対して，線形的な影響を及ぼすことがわかった．したがって，学習語生存確率を知ることは，最近使用語優先方式を採り入れた場合の変換性能を予測するために有効であることがわかった．

本研究で最近使用語優先学習方式をモデル化したことによって，例えばある単語使用間隔分布の文章をある大きさの学習辞書を使って変換した場合の変換性能が予測できる．実際の場面では，同音語の扱い，最近使用語優先以外の情報との関係など，もっと複雑な問題もある．また，今回行った実験はモデルの検証を主眼としているので，大きな文章や学習辞書の長期間にわたる使用に関しては検討の余地がある．これらが今後の課題である．

参考文献

1) 牧野　寛，木澤　誠：べた書き文の分かち書きと仮名漢字変換―二文節最長一致法による分かち書き―，情報処理学会論文誌，Vol. 20, No. 4, pp. 337-345 (1979).
2) 吉村賢治，日高　達，吉田　将：文節数最小法を用いたべた書き日本語文の形態素解析，情報処理学会論文誌，Vol. 24, No. 1, pp. 40-46 (1983).
3) 酒井貴子，本宮志江，下村秀樹，並木美太郎，中川正樹，高橋延匡：日本語ワードプロセッサの仮名漢字変換における変換処理と精度についての考察，情報処理学会ヒューマンインタフェース研究会報告，35-10 (1991).
4) Denning, P. J.: The Working Set Model for Program Behavior, *Comm. ACM*, Vol. 11, No. 5, pp. 323-333 (1968).
5) 高橋延匡，土居範久，益田隆司：オペレーティング・システムの機能と構成，岩波講座　情報科学-16，岩波書店 (1983).
6) 下村秀樹，本宮志江，酒井貴子，並木美太郎，高橋延匡：OS/omicron 仮名漢字変換システム第2版の設計思想，第42回情報処理学会全国大会論文集，5Q-1 (1991).
7) 下村秀樹，並木美太郎，中川正樹，高橋延匡：最小コストパス探索モデルの形態素解析に基づく日本文誤り検出の一方式，情報処理学会論文誌，Vol. 33, No. 4, pp. 457-464 (1992).
8) 下村秀樹，並木美太郎，中川正樹，高橋延匡：人間の文章誤り検出能力と誤り検出機能の効果に関する実験，情報処理学会論文誌，Vol. 33, No. 12, pp. 1607-1617 (1992).
9) 酒井貴子：日本語文章作成環境における入力系の研究，東京農工大学工学部電子情報工学科卒業論文 (1991).

（平成5年2月5日受付）
（平成5年12月9日採録）

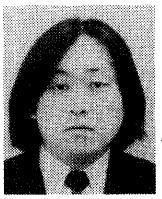


下村　秀樹（正会員）

昭和40年生．平成5年東京農工大学大学院博士後期課程（電子情報工学専攻）修了．同年日本電気(株)入社，現在に至る．日本語情報処理，特に仮名漢字変換，日本語文章作成環境の研究に興味を持つ．工学博士．

酒井　貴子（正会員）

昭和43年生．平成3年東京農工大学工学部数理情報工学科卒業．平成5年同大学院博士前期課程（電子情報工学専攻）修了．同年(株)日立製作所入社．現在，システム開発研究所に勤務．自然言語処理，特に仮名漢字変換に興味を持つ．

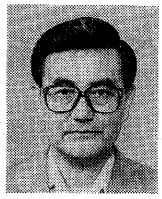


並木美太郎（正会員）

昭和59年東京農工大学工学部数理情報卒業．昭和61年同大学院修士課程修了．同4月(株)日立製作所基礎研究所入社．昭和63年より東京農工大学工学部数理情報助手．平成元年4月より電子情報助手．並列処理，日本語情報処理のソフトウェア/ハードウェアアーキテクチャに興味を持ち，コンパイラ，オペレーティングシステムなどシステムプログラムの研究・開発に従事する．工学博士．

中川　正樹（正会員）

昭和52年東京大学理学部物理卒業．昭和54年同大学院修士課程修了．同在学中，英国 Essex 大学留学 (M. Sc. in Computer Studies)．昭和 54 年東京農工大学工学部数理情報助手，平成元年1月数理情報助教授，同4月電子情報助教授．オンライン手書き文字認識，日本語計算機システム，文書処理の研究に従事．理学博士．



高橋　延匡（正会員）

昭和8年生．昭和32年早稲田大学第一理工学部数学卒業．同年(株)日立製作所中央研究所入社．HITAC 5020 モニタ，TSS の開発に従事．昭和52年より東京農工大学工学部数理情報教授．平成元年電子情報教授．理学博士．オペレーティングシステム，日本語情報処理，パターン認識の研究に従事．電子情報通信学会，ソフトウェア科学会，計量国語学会，ACM 各会員．