北海道大学工学部研究報告
第119号（昭和59年）

Bulletin of the Faculty of Engineering,
Hokkaido University, No. 119 (1984)

かな漢字変換システムのシミュレーションによる性能評価

川合　英夫*　栃内　香次　永田　邦一

（昭和58年9月30日受理）

Simulation Study of the Kana-Kanji Translation System

Hideo KAWAI, Koji TOCHINAI and Kuniichi NAGATA

(Received September 30, 1983)

Abstract

Effects of the size and the control algorithm of the Kanji word dictionary on the performance of the Kana-Kanji translation have been studied using a simulation technique.

In the researcher oriented Japanese word processing system reported previously, the size and the control algorithm of the Kanji word dictionary are considered to be important for the performance of the system.

In this study, the Kanji word dictionary was simulated using the data collected during input experiments on the system.

Simulation results indicate that the size of the dictionary of the current system is suitable, however the performance is expected to be increased by the use of a modified control algorithm.

1．序　　論

日本語のテキスト処理における最大の問題は，極めて多数種の漢字を使用する日本語文章をいかに効率よく入力するかという点にある。この問題に対する決定的な解決法は未だ見出されていず，使用される場に応じてさまざまの方式が出現している。

かな漢字変換方式は，これら種々の入力方式のうちで，おそらくもっとも適用範囲の広い方式と考えられ，現に広く普及しつつある。この方式はその名の示すとおり，日本語文をすべてかな（あるいはローマ字）で入力し，必要な部分を計算機内で漢字に変換して漢字かな混り文を得るものである。したがって本質的に，

1）かな－漢字の変換のための大容量の辞書を用意する必要がある。

2）同音異語（字）の発生が避けられない。

という二つの問題点が存在する。

われわれは先に，作成対象文書をある専門分野の論文，報告等に限定することにより，

1）使用される漢字語が限定されるため小容量の辞書でよく，

2）同音異語の発生も少なく，

3）辞書の内容（収録語）を当該分野，さらに個々の使用者に適応させることにより良好な変

電子工学科　電子機器工学講座
*現在　三菱電機（株）

換性能が得られる，
ことを示し，これにもとずいて実用システムの試作を行った。[1),2)] このシステムは，現在われわれの研究室で学会予稿，その他の文書作成に使用されるとともに，多数の資料による変換性能の測定，使用経験の蓄積を行っている[3)]。

本研究は，このシステムの性能評価の一環として行われたもので，変換辞書の容量，収録語の適応方式等と変換性能との関係をシミュレーションによって求めるものである。これらは，実際にシステムを稼動させて実測する場合は，測定に長時間を要し，またプログラムの一部を修正する必要があり，手間がかかるが，シミュレーションによれば短期間に広範囲にわたるデータを得て，大域的な性質を把握することができる。このような観点から，システムのモデルを構築し，GPSS によりシミュレーションを行った，その結果，

1）試作システムで採用している変換辞書容量，適応アルゴリズムはほぼ妥当であるが，

2）さらに改良が可能と思われる点がある，

ことが判明し，今後のシステム改善に資する指針が得られたのでここに報告する。

2．かな漢字変換システムＫＫＨの概要

本研究において検討の対象とする日本語テキスト処理システムの概要を述べる。このシステムはすでに報告したように，論文，予稿など，ある専門分野の学術文書を作成することを主目的としており，[1),2)] その概要は以下に示すようになっている。

1）システムは北大大型計算機センターの HITAC VOS 3 システム上に作られている。

2）入力は同じシステムに接続された一般の TSS 端末からローマ字表記形式で行われる。

3）出力は当面，大型計算機センター内に設置されているレーザビーム漢字プリンタに行われる。

4）漢字とかなの区分は使用者の指定によるものとし，漢字語の先頭を大文字とし，漢字からひらがなへの境界にスペースを 1 個おく方法を用いる。なお，漢字が連続している場合の語境界の指定も使用者が与える。

5）変換辞書の容量は 2500 語とし，使用者ごとに個別に持つ。

6）変換辞書には初期収録語として，先に行われた情報処理に関する 30 篇の論文に出現する漢字語の調査で得られた 3671 語の異なり語の中から，頻度 2 以上の 2303 語を収容する。

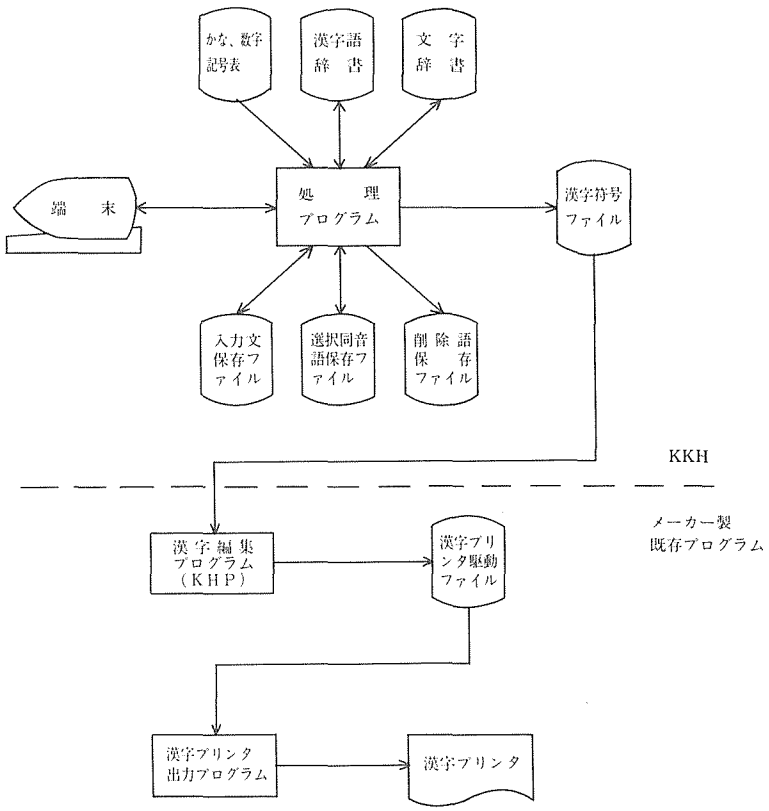
7）本システムにより文書入力を行う場合，辞書に登録されていない語が出現したならば，それを登録する必要がある。このとき，辞書に空きがあればそのまま登録するが，空きがない場合は今後使われないと予想される語を削除し，そのあとに登録する。

8）削除すべき語を決定するために，各語には出現頻度と前回参照されてからの間隔を示す二つのカウンタ（頻度カウンタ，履歴カウンタ）を設け，頻度が小で履歴が大のものを削除するようにしている。

9）未登録語の新規登録を容易にするために，別に漢字辞書を用意して出現した漢字を登録しておき，語を構成する各文字をこの辞書から得るようにしている[3)]。

図 1 に試作システム全体の構成を示す。なお，われわれは本システムを KKH と称しており，簡単のため以下本論文においても KKH という名称を用いる。

上述のように，KKH では変換辞書に未登録の語（新出語）が出現したとき，変換辞書が一杯であれば，いずれか 1 語を削除してそのあとに登録する。したがって，削除語を選出するアルゴリズムは重要である。現在の KKH では，登録されている各語の頻度カウンタの値 n と履歴カウン



図—1　KKH システムの構成

タの値 h を用い，以下に示す手順により削除語を選出している。

1）ある文書を入力するためにシステムを起動したとき，変換辞書中のすべての漢字語について $h+1 \rightarrow h$ とする。

2）辞書に登録されているある語が参照されたとき，その語について

$n+1 \rightarrow n$，$0 \rightarrow h$ とする。

3）新出語を登録するとき，辞書が一杯であれば各語について h/n を求め，この値が最大なものの中から1語を選んで削除する。

1），2）から，稀にしか使用されない語は n が小さく，h が大きくなるので，3）によりそのような語が選出される。ただし，過去に多数回使用された語では n が大きく，したがって h が相当大きくなるまで選出されない。そこで，他に h_0，n_0 というパラメータを設け，$h > h_0$ のもののみを対象とし，さらに $n \leqq n_0$ のものはすべて $n = n_0$ としてから h/n を求めようにして，n の変化に敏感すぎないようにしている。実際のシステムでは，使用者により異なるが，

$h_0 = 10 \sim 20$，$n_0 = 2 \sim 5$ としている。

以上により，かなから漢字語への変換性能を決定する重要な要素として，

1）変換辞書の容量，

2）削除語選出アルゴリズム，

の2項目があげられる。これらは，実際に多数の文書を入力して辞書内容の変化を観測することにより，その適否を検討することが可能であるが，大量の入力実験を必要とし，また辞書の容量や削除語選出アルゴリズムを種々変えて多数の実験をくり返すのは容易でない。本研究は，この

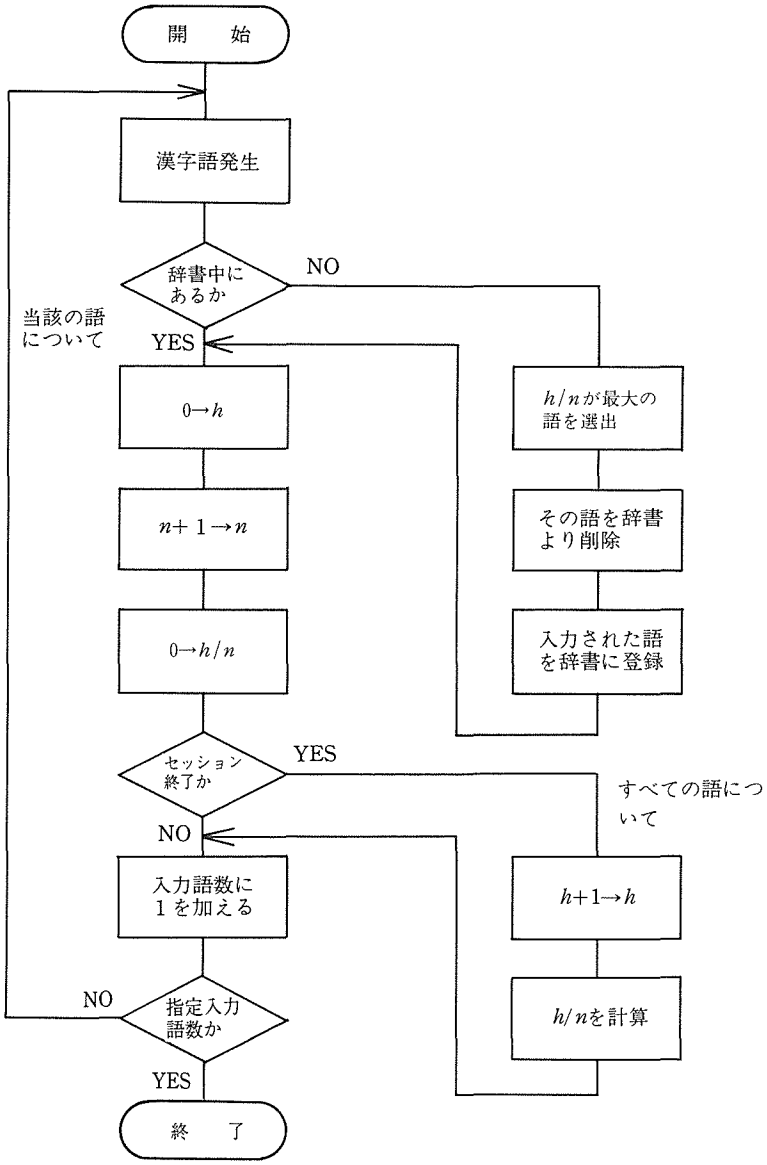
部分をモデル化してシミュレーションを行うものである。

3．シミュレーション・モデル

前述のように，本研究の目的は変換辞書の容量，削除語選出アルゴリズムと変換性能との関係を求めようとするものである。KKH の現実の動作においては，1 回のセッションごとに 300～500 語の漢字語が入力され，先に述べたような処理をうける。そこで，この部分を図 2 に示すようにモデル化した。以下，図 2 にしたがってこのモデルの動作概要を述べる。

1）あらかじめ与えられた分布にもとづき，漢字語を 1 語発生させる。

2）その語が辞書に登録されているか否かを検査する。



図—2　モデルの動作概要

3）当該の語が辞書に登録されているときは，その後の履歴カウンタh，頻度カウンタnについて，$0 \rightarrow h$，$n+1 \rightarrow n$とする。

4）当核の語が未登録のときは，辞書によりh/nが最大の語を選出して削除し，そのあとにこの語を登録する。

5）発生した語ののべ語数をしらべ，セッション終了にあたるか否かを判定する。

6）セッション終了に到達している場合は$h+1 \rightarrow h$とし，辞書中の各語についてh/nを計算し直す。ついで，のべ語数に1を加える。セッション終了に到達していない場合は，単にのべ語数に1を加える。

7）のべ語数があらかじめ指定した総入力語数に達しているか否かを判定する。

8）到達していなければ1)に戻り，次の語の処理を行う。到達している場合はシミュレーションの実行を終了させる。

前述のように，KKHの実際の使用では，1回のセッションで学会の大会講演予稿1篇分程度が入力される。この場合，総入力字数は2000～3000字で，そのうち漢字語はのべ300～500語である。一方，辞書内容の変動を観測する実験では，のべ10000語程度の漢字語を入力すればおよその傾向を知ることができる[4]。そこで，前記総入力語数は10000語に設定し，それを20回前後のセッションに分割するようにした。

以上にもとづき，シミュレーション・プログラムを作成した。プログラムはGPSS[5]で書かれており，上記総入力語数，セッション回数などは簡単に変更できる。なお，削除語を選出する方式についても，プログラムの一部を修正することにより，h/nが最大のものを選ぶ方法以外の種々の方式を組込むことができる。また，漢字語は任意の分布を与え，それにしたがって発生させることができる。具体的には，GPSSの機能を利用して一様乱数を発生させ，それを所望の分布に変換している。本研究では，実際に行った入力実験から得られた分布形を表の形で与え，それにしたがって変換するようにした。なお，シミュレーションの際は，漢字語を実際に漢字符号で表わす必要はないので，この変換結果を4桁の整数に直し，その値により個々の漢字語を識別する。また，同音語の有無についても，実際の入力実験で得られた同音語率に一致するように同音語標識を付け，それにより判別している[6]。

4．シミュレーション結果

4．1　実験条件

いくつかのパラメータについて，変換辞書の動作のシミュレーションを行った結果を以下に述べる。前章で述べたように，このシミュレーション・プログラムでは，入力漢字語の頻度分布，辞書の容量，総入力語数，セッション回数を自由に変えることができ，また，削除語選出アルゴリズムは，その部分のプログラムを多少修正することにより変更できる。そこで，本研究では総入力語数は10000語，セッション回数は20回前後とし，他の3種のパラメータを変えてシミュレーションを行った。

1）入力漢字語の頻度分布　　KKHを使用して入力実験を行って得られた，実際の文献における漢字語の頻度分布を用いた[7]。具体的には表1に示す2種の資料を使用した。また，変換辞書の初期収録語についても，実際のKKHの変換辞書の初期収録語と同一の分布にしている。ただし，シミュレーションの方は辞書容量一杯まで初期収録語を収容し，空きは設けていない。

2）変換辞書の容量　　1500語，2000語，3000語の3種についてシミュレーションを行った。

3）削除語選出アルゴリズム　　前述のように，現行KKHではh/nが最大のものを選出して

表1 実験に使用した資料

	資料 A	資料 B
資料内容の概要	かな漢字変換システムの研究に関する修士論文	情報処理システムの入門的解説書
文字総数	19,631字	71,297字
漢字総数	4,169語	8,544語
新出語総数	100語	386語
平均新出語率	2.4%	4.5%
同音語のべ語数	819語	1,744語
平均同音語率	19.6%	20.4%

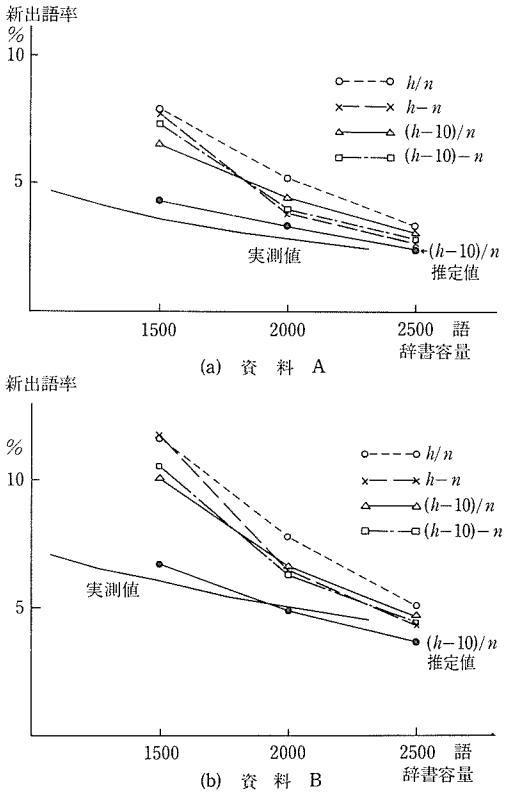
変換辞書の初期収録語はいずれも同一（2,303語）
新出語、同音語は入力過程で出現したものの累計

いるが，n（頻度）の効果が強すぎる感がある。そこで，本研究ではこの他，$h-n$が最大のものを選出するアルゴリズムについてもシミュレーションを行った。さらに，この各々について，$h \geqq 10$ のもののみを削除対象とする場合についてもシミュレーションを行い，h/n, $(h-10)/n$, $h-n$, $(h-10)-n$, の4種について比較を行った。

4．2 新出語率

入力語のうち，辞書に未登録のものを新出語という。新出語は出現した時点で直ちに登録されるので，入力語の累積とともにその割合は減少する。新出語には，最初から辞書になかったものと，当初は存在したが，その語が再度入力されるまでの間に削除されたものとがある。したがって削除語選出アルゴリズムの影響をうける。また，当然ながら辞書の容量に大きく依存する。図3に，のべ10000語の入力に対する新出語率を示す。ここで図3（a）は資料Aに対する結果を，また（b）は資料Bに対する結果を各々示す。

アルゴリズムによる差異はあまり大きくないが，頻度の髪響を小さくした$h-n$系の方が若干よく，またいずれの場合も$h \leqq 10$のものを選出しない場合の方がよい。実測値[7]と比較するといずれも大きくなっているが，これは実測値の方は初期収録語が少なく，辞書に空きがあってはじめのうち削除語が発生しないためと考えられる。これを確めるために，削除語が発生せず，新出語はすべて最初から辞書に存在しなかった語であるとみなして結果を補正すると，図中に「推定値」と記した曲線が得られ，実測



図—3 新出語率

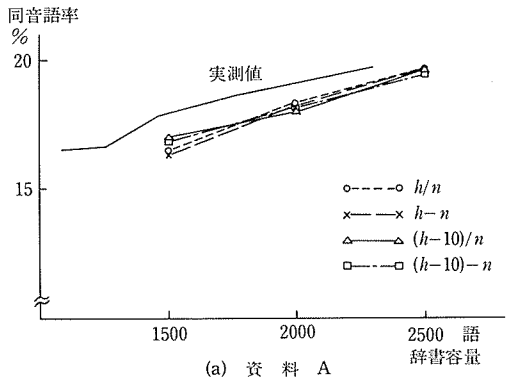
値とほぼ一致する。すなわち，実測値とのちがいはモデルが不適当なためではなく，実験条件のちがいにもとづくと考えられる。

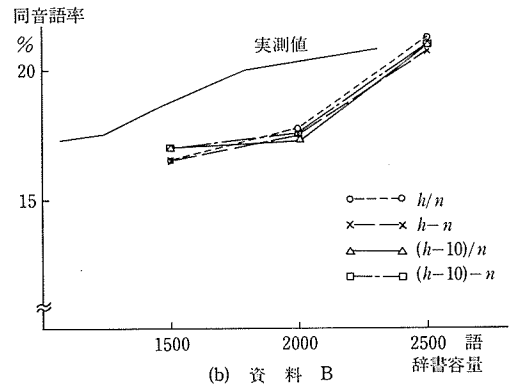
4．3　同音語率

同音語は，語の累績，使用分野等の影響をほとんどうけず，一定の割合で存在すると考えられる。したがって，変換辞書の容量に比例すると考えられる。このシミュレーション結果を図4に示す。前節と同様，図4（a）は資料Aに対するもので，また（b）は資料Bに対するものである。この結果から，上記のとおり辞書容量にほぼ比例し，削除語選出アルゴリズムにはほとんど無関係であることがわかる。

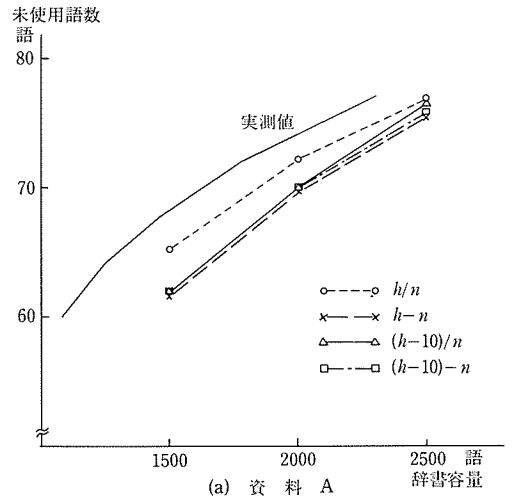
4．4　未使用率

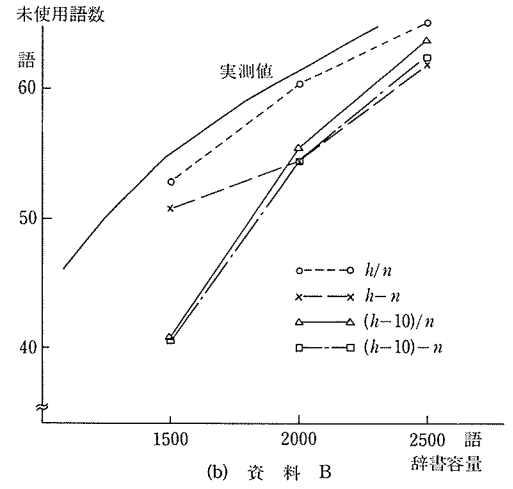
辞書の初期収録語のうち，のべ10000語の入力中1回も出現しなかったものを未使用語という。このシミュレーション結果を図5に示す。これからわかるように，削除語選出アルゴリズムの影





図—4　同音語率





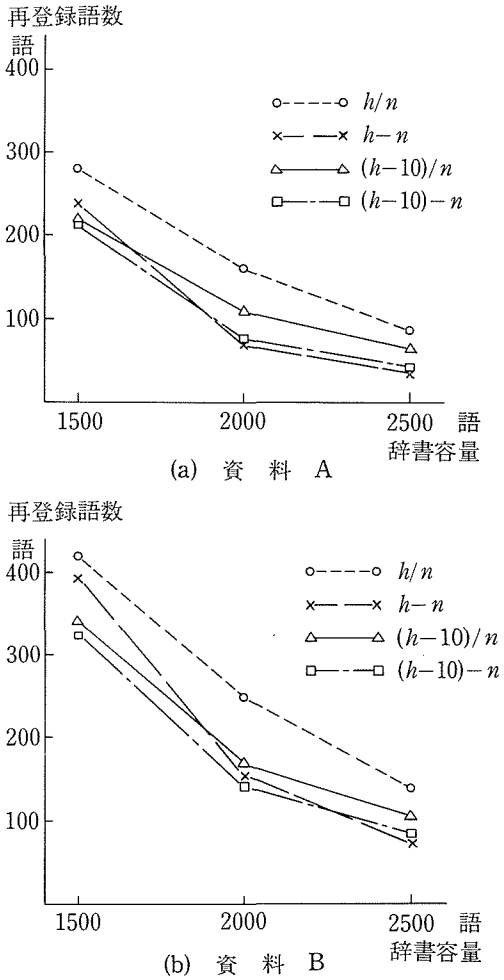
図—5　未使用語数

響は比較的小さく，また，容量が2000語をこえるといずれの場合も50%以上となる。

4．5 再登録率

ある漢字語が辞書から削除された後に入力文に出現し，再度登録される場合があり，これを再登録語という。図6はこのシミュレーション結果を示すものであるが，これからわかるようにこの値は削除語選出アルゴリズムによって変化し，$h-n$および$(h-10)-n$アルゴリズムの場合が比較的よい結果を示す。なお，図6で縦軸は語数であるが，たとえばこれを50語とすると，10000語の入力に対し0.5%にあたり，十分小さい値であるといえる。図6で，辞書の容量を2500語とし，削除語選出アルゴリズムを$h-n$，$(h-10)-n$とした場合，ほぼこの程度になっていることがわかる。

なお，この値は実測値が得られていないので，正確な比較はできないが，実際にKKHを使用している場合，明らかに再登録語であることに気がつくことは数文献に1回程度なので，大きく見積っても上述の結果程度であると思われる。



図−6 再登録語数

5．結　　論

以上のシミュレーション結果，および実測値との比較結果をまとめると，以下のような結論を得ることができる。

(1) 本研究で構築したシミュレーション・モデルにより，現実のKKHシステムにおける変換辞書の挙動をほぼ表わすことができた。

(2) 現行KKHで採用している変換辞書容量，2500語は妥当な値である。

(3) 削除語選出アルゴリズムは，現行KKHで採用しているh/n方式よりも，$h-n$方式の方がよい。今回は実験しなかったが，頻度nを考慮せず，履歴hのみで判定する方式でも差支えないように思われる。

(4) 一旦削除された後，再登録される語はかなり多い。それゆえ，実際のKKHシステムでは，削除語を別なファイルに保存しておき，簡単に登録できるようにすることを考慮する必要がある。

(5) 辞書の容量，削除語選出アルゴリズムを変え，より広い範囲にわたるシミュレーションを行って，さらに適切な値を見出すことも必要であろう。

謝　　辞 本研究の遂行にあたり，実験の一部をお手伝いいただき，また種々有益な御討論をいただいた研究室の皆様に感謝します。

参照文献

1) 栃内香次，斉藤　康：情報処理学会論文誌，Vol.24, No.2, pp.209—213(1983)
2) 斉藤　康，岡沢好高：栃内香次，永田邦一：北大工学部研究報告，No.108, pp.43—52(1982)
3) 岡沢好高，栃内香次，永田邦一：北大工学部研究報告，No.116, pp.79—86(1983)
5) 日立製作所：離散型シミュレーション・システムGPSS機能篇
4) 栃内香次，岡沢好高：情報処理学会第26回全国大会講演論文集，2H—4（1983）
6) 川合英夫：北大大学院工学研究科修士論文（1983）
7) 岡沢好高，斉藤　康，栃内香次，永田邦一：情報処理学会第25回全国大会講演論文集，7J—2（1982）