ソフトウエア工学 34−20
（１９８４．２．１０）

ローマ字漢字変換システムＫＫＨの ユーザインタフェースと使用効率

栃内香次　伊藤太亮　荒木健治　鈴木康広　永田邦一
（北海道大学）

１．はじめに

ＫＫＨは，研究者が自分の専門分野に関する論文，講演予稿その他の学術文書を作成するのに使用することを主目的として筆者らにより開発されたローマ字漢字変換方式日本語ワードプロセッサである[1]．このシステムは１９８１年夏に稼動を開始し，数次のレベルアップを経て１９８３年秋に第２版が完成した．

ワードプロセッサの性能指標として考慮すべき事項は多数あるが，本稿では主として次の２点について検討を行う．

１）文書入力をどれだけ連続して中断されることなく続けられるか．

２）中断されたときに必要な操作はどれだけ複雑か．

文書入力が中断されるのは，

ａ）変換辞書に未登録の語（以下，新出語という）が出現した場合，

ｂ）同音語の選択を必要とする場合，および，

ｃ）鍵盤操作ミスその他の誤まりを犯した場合，

であるが，このうちａ），ｂ）２項が上記１），２）に関係する．この観点から，システムのレベルアップは，実際に多数の学術文書を入力して新出語および同音語の出現状況を分析し，これらの出現率の減少をはかることと，出現した場合に必要な操作手順の簡単化の２点を中心に行われた．以下，本稿ではこれらについて関連する諸事項を述べる．

２．ＫＫＨシステムの概要

ＫＫＨシステムは図１のように構成されており，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システム上に作られている．

２）入力は一般のＴＳＳ端末からローマ字表記で行われる．

３）出力はセンターに設置された漢字プリンタに行われる．

４）漢字とかなの区分は使用者が指定する．すなわち，漢字語の先頭を大文字とし，漢字からかなへの境界には空白を１個おく．また，漢字が連続する場合の語境界も使用者が指定する．

５）漢字語への変換を行うための辞書（漢字語辞書＜Ｋ辞書＞という）は使用者ごとに個別に作られ，容量は２５００語である．

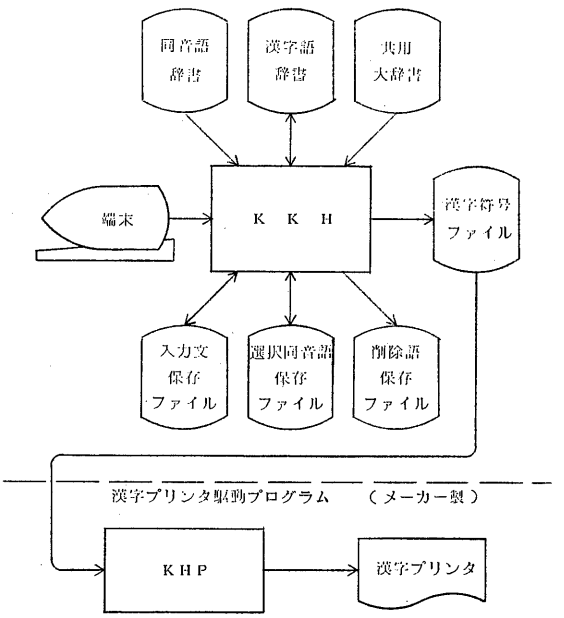

図１　ＫＫＨシステムの構成

6）文書入力中に新出語が出現したとき，変換辞書が一杯ならば頻度が小さく，かつ長期間使われていない1語を削除し，そこに登録する．これにより，一人の使用者が同一分野の文書作成を続けると，変換辞書の収録語は入力漢字語の累積につれて使用者および分野に適応してゆく．

7）新出語を変換辞書に登録する際，複数の使用者が共用する大容量漢字語辞書（L辞書）を検索し，該当する語があれば確認のうえそれを登録する．ない場合は新出語の各漢字について，その音よみ，訓よみの組を与えて文字辞書（M辞書）を検索し，該当する漢字があればそれをとり出して語を組立てる．文字辞書にもない場合は漢字符号表により外部から漢字符号を与える[2]．

8）同音語が出現したときはそのすべてを表示し，使用者が選択する．端末には漢字が表示されないので選択は各漢字語につけられた補助情報により行う．なお，選択結果を固定し，以後システム内で自動選択する機能，および同音語の前後の文字連鎖のパターンにより自動選択する機能がある[3]．

3．新出語の辞書への登録

前述のように，KKHの変換辞書は入力漢字語の累積につれて使用者，分野に適応する．したがって入力のべ語数に対する新出語出現率は次第に減少する．図2はこの様子を示したもので，いくつかの分野の文献を入力し，新出語出現率の推移を求めた結果である．ここで，変換辞書の初期収録語はいずれも同一で，情報処理学会誌の論文30篇に頻度2以上で出現した2303語を用いている．そして，このうち資料Aは初期収録語と同一分野，B，Cは異なる分野，またDは近縁の分野の文献である．

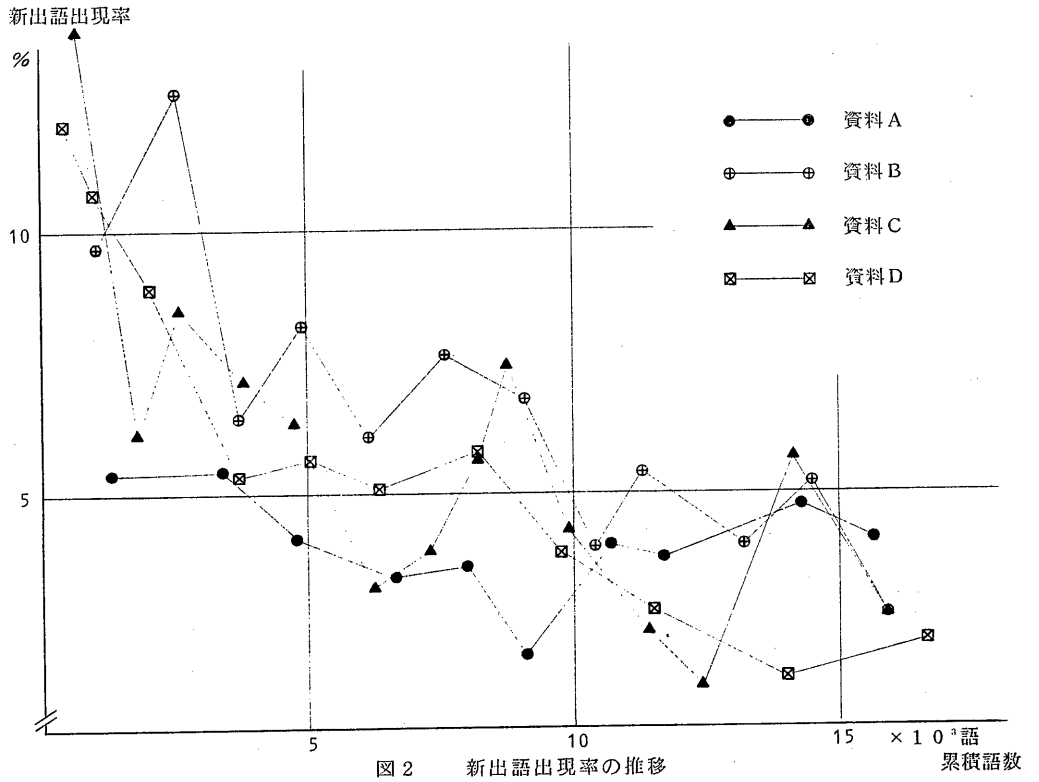

図2　新出語出現率の推移

新出語が出現したとき，この各文字について漢字符号を求め，よみ，補助情報とともに変換辞書に登録する必要がある．与えられた漢字について漢字符号表を検索して漢字符号を求めるのはかなり手間のかかる操作である．

文書中には多種類の漢字語が出現するが，それを構成する漢字の種類はそれほど多くない．たとえば，上の実験で変換辞書の初期収録語とした2303語については，使用されている漢字の字種は814字にすぎない．新出語についてもこの傾向は保存されると予想される．

以上により，初期のレベルアップで文字辞書が導入された．この結果，上記の実験に際して出現した新出語を構成する漢字の90％以上は文字辞書に含まれていることがわかり，文字辞書は極めて有効であると結論された．変換辞書の適応が進んだ段階では，図1に示されるように新出語出現率は4％程度である．したがって，文字辞書にも存在せず，外部から漢字符号を与える必要のある漢字は，入力した漢字の総数に対して0．2～0．3％となり，極めて少数である．

さらに，複数の使用者が存在する場合，ある使用者の新出語が他の使用者については既出である場合がありうる．そこで，各使用者の変換辞書にある語をまとめた各使用者共用の辞書（大辞書＜L辞書＞という）を導入した．この辞書は読出し専用であり，適当な期間ごとに各使用者の変換辞書と比較して内容の更新を行う．これにより，各使用者の変換辞書は小容量のまま多数の語をシステム内に保存することが可能となる．

以上により，語のよみを与えて漢字符号を得る手順は図3に示すように階層的に行われ，手間のかかる操作を必要最少限のとどめることが可能になった．なお，新出語を大辞書からとり出す際は，そのまま自動的に変換辞書に登録するのではなく，一旦表示して使用者の確認を得た後に登録する．

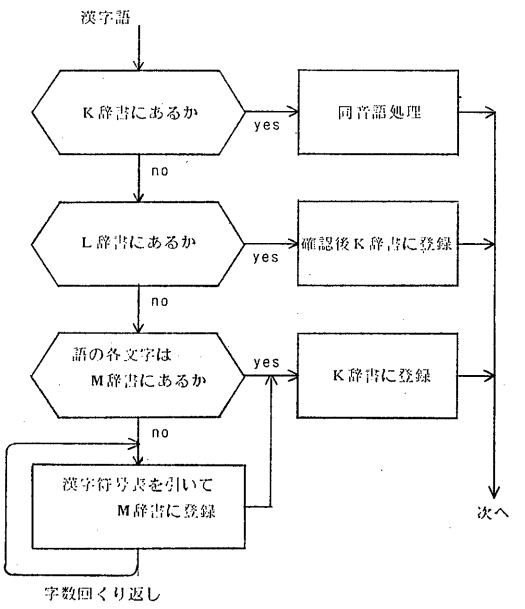

図3　　漢字語変換手順

4．同音語の選択

同音語の出現に関しては，入力の累積による効果はほとんどなく，出現率は入力漢字語のべ語数の20％前後である．しかしながら，一つの文献内では同音語のうちいずれか1種のみが頻出し，他は極めてわずかしか出現しないという性質がある．これを利用して，同音語選択の際にその選択結果を記録し，以後その同音語が出現したときはこの情報を用いてシステム内で自動的に選択する機能を設けた．これを同音語の固定化とよぶ．この機能は任意の同音語について設定できる他，すべての同音語について，n 回（n＝1，2，…）同一の選択が行われたなら自動的にそれに固定することが可能である．ここで n はシステム・パラメータであり，n＝1とすれば最初の選択に固定することになる．この場合，使用者が選択

する回数（手動選択回数）は同音語の語種にほぼ等しい．実測によればこの出現率は入力漢字語のべ語数の約10％であり，同音語全体の出現率の1/2となる．

手動による同音語選択回数をさらに減少させるために，その後同音語とその前後に出現する文字との連鎖関係を利用する自動選択機能を組込んだ．いま，同一のよみWをもつ同音語W_1，W_2が各々，$x_1 W_1 y_1$，$x_2 W_2 y_2$という形で出現しているものとする．ここでx_1，x_2，y_1，y_2は任意の文字である．これらの3つ組をその出現頻度とともに記録しておき，この同音語がさらに入力文中に出現したとき，3つ組xWyの比較によりこれがW_1，W_2のいずれであるかを判定することが可能である．このアルゴリズムの詳細は他の文献[3]にゆずり，結果だけ述べると，上記3つ組の記録（同音語辞書）が150語程度に達したところで，入力同音語のべ語数の約50％が自動選択されるという結果が得られている．

以上により，同音語のうち手動で選択する必要のあるものは入力漢字語のべ語数の10％以下とすることが可能となった．固定化と文字連鎖による自動選択は相互に重なり合っており，両者を組合せてどの程度になるかは今後の検討課題である．

5．誤変換および入力ミス

KKHの出力（漢字かな混り文）に現われる誤まりは，誤変換によるものと入力ミスによるものの2種に大別される．このうち，誤変換は以下に示す機構によって発生する．

1）同音語をもたない漢字語W_1が変換辞書に登録されているとき，この語の同音語W_2が入力されるとこれはW_1に変換される．

2）同音語W_1，W_2が変換辞書に登録されているものとする．ある文書の入力時に，固定化あるいは自動選択機能によりW_1が自動的に選択される状態になっているとき，W_2が入力されるとこれはW_1に変換される．

1）による誤変換の発生率は，最初3％程度から入力漢字語の累積とともに漸減し0.5％以下程度に減少するという実験結果が得られている．また，2）による誤変換の発生率もほぼ0.5％程度である．したがって，総合的な誤変換発生率は1％以下と見積ることができる．

一方，入力ミスは不正確な鍵盤操作によって発生する．どのような操作ミスが多いかの詳細な分析は行っていないが，

a）指使いの難しいキーの誤操作，

b）シフトキーの不正確な操作，

によるものが比較的多い．a）については通常の文字以外の特殊記号によって種々の制御機能を実現している部分で多く発生するので，その改善が必要と考えられる．またb）は本システムの字種指定方式の根幹をなす部分なので，本質的には字種指定を行わない，いわゆる「べた書き」入力方式を採用すべきであると考えられる．ただし，入力全字数に対する入力ミスの発生率は0.3～0.5％であり，それほど大きな問題にはなっていない．

6．総合性能

システムの総合的な性能指標として，次に示す2種の変換率を用いる．

1）語単位正変換率　　入力漢字語のうち完全にシステムのみによって正しく変換されたものの比率．すなわちこの値は入力漢字語から新出語，手動選択同音語，

および誤変換された語を除いたものである．

2）文字単位正変換率　入力した全文字のうち完全にシステムのみによって正しく変換されたものの比率．したがってこの値は入力された全文字から上記各々に該当する語を除き，さらに入力ミスによる誤まり文字を除いたものである．

このうち，語単位正変換率は前3章で述べた諸量から求めることができる．すなわち，変換辞書の適応が相当進んだ後は，

a）新出語出現率≒4％，

b）手動選択同音語出現率≒10％，および，

c）誤変換語発生率≒1％，

と見積ることができる．したがって，語単位正変換率≒85％となる．前記4種の資料による入力実験では各々約15,000漢字語の入力後，85～90％という結果が得られた．

文字単位正変換率は，入力ミスがない限り正しく変換される漢字以外の文字を含めて計算されるので，上記より10％程度高い値を示し，上記の実験において95～98％という結果が得られている．この値は入力文中のかなと漢字の比率によっても変化し，漢字の比率が小さいほど高い値を示す．したがってこれを性能指標にすることには多少問題があるが，一方，入力操作感覚の点からは全文字数を考慮する方がよいと考えられるので，性能指標として両者をともに用いている．

なお，文字単位正変換率として，95％という値が実用下限として報告されており[4]，KKHはこれを満足している．ただし，実際に操作して十分満足できるためにはこれよりもう少し高い，97％位が必要のようである．上記の実験から2種の正変換率間の関係を求めると図4に示すようになり，語単位正変換率の文字単位正変換率への寄与率は0.3～0.4となることがわかる．

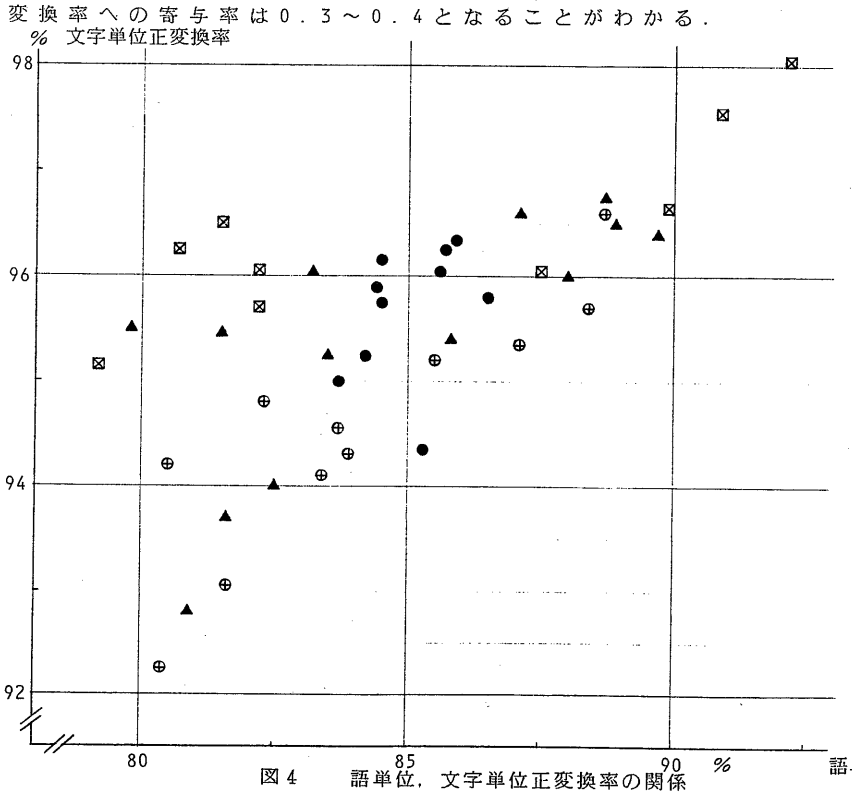

図4　語単位，文字単位正変換率の関係

システムの操作性，あるいは使い勝手を表わすには種々のアプローチがあると思われるが，同一のシステムにおける改善の度合を示す指標として，文書入力に要する所要時間をとり，これと上記文字単位正変換率との関係を求めた．この結果を図５に示す．ここで，縦軸は入力所要時間／入力文字総数，すなわち１字あたり入力所要時間である

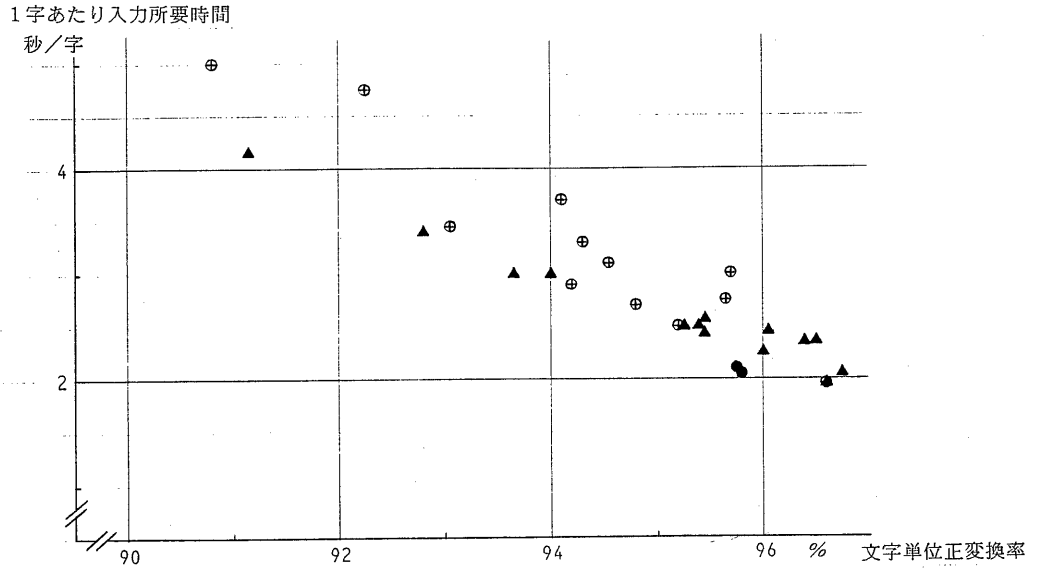

図５　文字単位正変換率と処理時間

縦軸は操作者のタイピング習熟度によって大幅に変るから，この値は相対的な意味しかもたないが，明らかに正変換率と強い相関があることがわかる．したがって，正変換率を指標に改善をはかる方法は有効であるといえる．なお，図５の実験において，操作者は著者（栃内）自身であり，タイピングの正規の訓練は全く受けていない．

７．おわりに

われわれの開発している研究者個人使用向きローマ字漢字変換方式日本語ワードプロセッサ，ＫＫＨについて，実際に使用しつつその結果の分析にもとづいて性能改善を行ってきた経過を述べた．このような手法により，常に最も効果的な部分から改善をすすめることができ，また，現在の仕様に由来する限界を知ることができ，今後の方向づけを行うことができた．おわりに，種々有益な御語討論をいただき，また使用経験をよせていただいた研究室の皆様に感謝します．

文　献

1)栃内，斉藤：情報処理学会論文誌，24，2，pp.209-213(1983.3)
2)岡沢，栃内，永田：北大工学部研究報告，No.116，pp.79-86(1983.10)
3)伊藤，栃内，永田：情報処理学会２７回大会，2H-6(1983.10)
4)森，天野：電子通信学会誌，63，7，pp.729-733(1980.7)