北海道大学工学部研究報告
第116号（昭和58年）

Bulletin of the Faculty of Engineering,
Hokkaido University, No. 116 (1983)

研究者向き日本語処理システムの新出語登録方式と性能測定

岡沢好高* 栃内香次 永田邦一

（昭和58年6月30日受理）

Kanji-Word Registration Method And Performance Measurements In the Researcher Oriented Japanese Word-Processing System

Yoshitaka Okazawa, Koji Tochinai and Kuniichi Nagata

(Received June 30, 1983)

Abstract

In the researcher oriented Japanese word-processing system reported previously, a small and user-adaptive Kanji-word dictionary is in use. It is required to record unregisterd Kanji-words appearing during text processing into the dictionary using a Kanji-code book.

In this report, a method for Kanji-word recording using a Kanji-character dictionary is described. When an unregistered Kanji-word appears during text processing, the Kanji-character dictionary is referred to with each characters of the word. Kanji-codes are taken, and then the total Kanji-code train of the word is synthesized.

The Kanji-character dictionary is also user-adaptive, and more than 80% of unregistered Kanji-words can be synthesized by characters in the Kanji-character dictionary, which has a limited capacity of only about 1,000 characters.

Experiments of text processing were carried out to obtain total system performances. The results show that the Kanji-words in the Kanji-word dictionary are adapted to the user and the field, and 91–95% of input characters are correctly translated into Kanji-codes.

It is concluded that the system is practical for Japanese word-processing.

1. はじめに

われわれは先に，研究者が自分の専門分野に関する論文等を作成するのに適した日本語入力方式として，個々の使用者に適応した比較的小容量の変換辞書を用いてかな漢字変換を行う方式を提案し，入力システムの試作を行った[1)-3)]。この方式において，変換辞書は動的に構成され，使用につれて新たに出現する語を次々と登録するとともに，長期間使用されない語は辞書から削除される。したがって，収録語は個々の使用者に適応し，研究分野の変化などによる使用語の変動にも追従する。この変換辞書を中核とし，入力システムは以下のように構成され，本学大型計算機センターの HITAC M-200 H 上に構築されている[1)]。

電子工学科 電子機器工学講座
* 現在 東芝(株)

1） 入力は漢字表示機能のない一般の TSS 端末から行う。

2） 使用者の慣れを考慮し，ローマ字入力とする。

3） 漢字，かな，カタカナなどの文字種の区別は使用者の指定による。

4） 出力は大型計算機センターに設置されているレーザビーム漢字プリンタに行う。

上述のように，われわれの方式では使用中に新たに出現する語を次々に辞書に登録し，もはや使われないと予想される語とおきかえてゆく。したがって，入力操作中にこれら新たに出現する語を登録する手段が必要であり，これを行うための操作手順は入力速度の低下をきたさないよう十分効率的で，かつ使用者に無用の負担を与えないものでなければならない。

本論文は，個々の漢字を収録した文字辞書を別に設け，語を構成する各漢字をこの辞書からとり出して語を合成することにより，未登録語の新規登録を効率よく行う方式について述べる。また，いくつかの文献を入力し，本システムの総合的な性能を測定した結果について併せて報告する。

2. 文字辞書による新出語登録方式

2.1 方式の概要

かな漢字変換方式では，漢字であらわされる語（以下，漢字語という）の読みをかな，あるいはローマ字で与え，これをキーとして変換辞書を検索する。変換辞書には読みとそれに対応する漢字の漢字符号との組が収録されており，入力した読みに一致するものがあれば，対応する漢字符号列が得られる。一致するものがない場合は入力した読みに対応する語が辞書に登録されていないことを意味する。このような語を新出語，あるいは未登録語ということにする。また，一致するものが2個以上あるときを同音異義語，あるいは簡単に同音語といい，その中から適切な1語を選択する操作が必要である。本節ではこれらのうち新出語の処理をとり扱う。

新出語が出現したときに必要な操作は，その語を構成する各漢字に対応する漢字符号を求め，読みとの組を作って変換辞書中に登録することである。このために，漢字と漢字符号の組を記載した「漢字コード表」が用意されている[4]。しかしながら，漢字の字種が極めて多いこと（漢字符号の定められているものだけでも6000字をこえる），同音漢字が多数存在すること，および1個の漢字に多数の読みが存在する場合が多いこと，等のため，漢字コード表をしらべて漢字符号を見出す操作はかなりの手間を要し，長時間かかることになる。

そこで，漢字コード表の部分をシステム内に組込み，上述の操作を計算機の支援の下で行うことが考えられ，それにより検索の手間が大巾に省かれることが期待できる。さらに，漢字語の場合と同様，語を構成する個々の漢字についても，使用者，使用分野を限定すると必要な字種は少くなり，小容量のコード表でよいことになる。したがって検索方式は単純なものでよく，同音漢字の問題も軽減されることが期待できる。

以上の観点から，変換辞書の場合と同様に，比較的小容量で，かつ使用につれて新たに出現する漢字を登録するように構成された辞書を用意し，この辞書を検索して漢字符号を得る方式を試作入力システム[3]に組込むことにした。これにより，新出語の大部分はこの辞書に収録されている漢字のみによって合成でき，この辞書にも存在しない漢字が出現したときのみ，手作業で漢字コード表を検索することになる。この辞書を文字辞書とよぶ。

2.2 収録文字の選定

本システムは，使用対象をある専門分野に限定し，かつ使用者も個人に限定することによって，

表1　論文で使用される漢字の調査結果

		漢字語辞書に存在する語	文字辞書により合成できる語	文字辞書により合成できない語	誤変換となる語	計
文献1	語数	1224	204	30	10	1468
	%	83.4	13.9	2.0	0.7	100.0
	語種	383	117	21	7	528
文献2	語数	414	76	0	4	494
	%	83.8	15.4	0.0	0.8	100.0
	語種	160	31	0	1	192

比較的小容量の変換辞書で良好なかな漢字変換を行っている。文字辞書に収録される漢字についてもこのことは成立し，小容量で十分であると予想される。

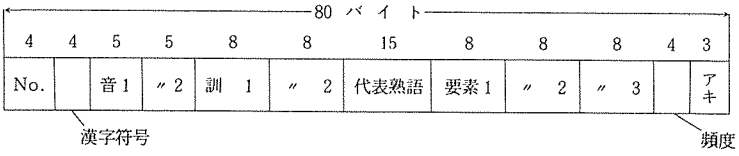
これを確認するために，試作システムで漢字語辞書に初期値として収録した語（以下，初期収録語という），2303語に含まれる814字の漢字によって他の文献にあらわれる漢字がどれ位カバーされるかを調査した。調査対象とした文献は情報処理学会論文誌から任意に選んだ2編の論文で，漢字語辞書に収録されている語を選出した分野と同一の専門分野の文献である。

この調査結果を表1に示す。これからわかるように，2編の論文にあらわれる漢字語のべ数の83～84%が辞書に収録されている語でカバーされ，したがってそのまま変換可能である。さらに，のべ語数の97～99%は，上記814字でカバーされている。それゆえ，上記814字を収録した文字辞書を設けることにより，新出語のほとんどはこの辞書に収録されている文字でカバーされ，漢字符号表を人手によって検索する必要のある新出語はのべ漢字語数の1～3%程度ですむことが予想される。さらに，文字辞書も使用につれて使用者への適応が進み，この割合はさらに低下することが期待できる。以上により，文字辞書は容量1000字とし，初期値として上記814字を収録することとした。以下，これを初期収録文字という。

2.3　文字辞書の構造

漢字語辞書と同様，文字辞書についても字の読みを与えて検索することになるが，一般に1個の漢字が複数の読みをもつので，音，訓それぞれ2種類ずつの読みで検索できるようにした。また，読みが1種類のみ，とくに音読みのみでは同音漢字が多数存在するので，音，訓それぞれ1個ずつの組を与えて検索するものとした。

前述のように，われわれのシステムは漢字表示機能のない端末を用いて入力できるようにしている。そこで，文字辞書についてもこのことを考慮し，各文字についてその構成要素（へん，つくり等）およびその文字を含む代表的な熟語を表示して，検索された漢字の確認と同音漢字からの選択操作を行えるようにしている。以上のことから



(a) 構　造

	文字番号	コード	音読み	訓読み	熟語	要素情報	頻度
数…	444	BFF4	suu su	kazu kazo	Suuzi	kome onna	70
語…	224	B8EC	go	katar	Nihongo	gonben itutu kuti	38
解…	61	B2F2	kai ge	toku tok	Kaisyaku	tuno katana usi	26
流…	782	CEAE	ryuu ru	nagar nagas	Ryuukou	sanzui	5

(b) 内　容　例

図1　文　字　辞　書

文字辞書の構成を図1 (a) のように定めた。また，図1 (b) にその内容例を示す。各文字には頻度情報がつけられ，辞書中では頻度順に配列されている。

2.4 新出語登録手順

文の入力中に新出語が出現したとき，その語を構成する各文字について，以下の手順により文字辞書を検索して該当する文字をとり出し，語を合成して漢字語辞書に登録する。

1) その文字の音読み，訓読みを各1個ずつ入力する。
2) 入力した音，訓の組に一致する漢字があればその情報（音，訓，代表熟語，要素）を表示し，使用者に確認を促す。同音，同訓のものがある場合はそのすべてについて代表熟語と要素を表示し，使用者に選択を促す。
3) 同音，同訓のものからある1字が選択された場合は，選択された文字に関する情報を再度表示し，確認を促す。
4) 一致する文字がない場合は，読みを変えて再度検索することができる。
5) さらに，音，または訓のみ一致するものがあればそれを表示し，同様に選択および確認処理を行う。
6) それでも一致するものがない場合，へん，つくり等文字の要素によっても検索することができる。
7) 最終的に一致するものが見出されない場合は新出文字として，あらためてその字の漢字符号を入力し，さらに音，訓，代表熟語，構成要素を入力して文字辞書への登録を行う。
8) 以上の操作を，語を構成する各文字についてくり返し，語が完成した後に漢字語辞書への登録を行う。

実際にはかなり多数の文字について，音（まれに訓）のみが存在し，一方が存在しないか，または存在してもほとんど知られていず，文字辞書に登録されていない場合がある。このような場合は一方のみを与え，一方はブランクとして検索する。また，語の一部にかなを含む形で漢字語を使用する場合（得ら，書く，など）があり，その場合は該当する文字がかなであることを指定し，文字辞書によらずに登録する。「々」等の記号が含まれる場合も同様である。

新出文字を登録するとき，文字辞書がすでに一杯であれば，頻度情報を用い，頻度最小のものから1字を削除し，そこに登録する。

3. 試作システムの性能

すでに作成され，稼働している入力システム[3]に，上記文字辞書と関連機能を組込み，入力実験を行った。

3.1 入力資料

表2に示す2編の文献を用いた。資料Aは本システムの設計，開発に参加した大学院学生の修士論文[5]，Bは情報処理システムに関する入門的解説書[6]で，ともに1人の著者によるものである。したがって，入力漢字語の累積とともに辞書中の語は使用者に適応し，新出語，新出文字の出現率は減少すると予想される。また，A，Bを比較すると，Aは特定の専門分野に関する文献であり，使用される語の範囲はかなり狭いことが

表2 入力資料一覧

資　料	総字数，C	漢字語数，K	漢字語字数，Kc	Kc/K
A	19,631字	4,169語	7,401字	1.78
B	71,297字	8,544語	14,851字	1.74

予想される。一方，Bは情報処理分野全般にわたっており，Aに比べてより広い範囲の語が使われていると考えられる。

3.2　変換特性

A，B 2編の資料について，資料Aでは1章ごと，Bでは約1/2章ごとに分割して入力し，各区分毎に新出語，同音語，新出漢字，誤変換語を集計し，これから変換率を求めた。

1）新出語　新出語の出現率は図2に示すように資料A，Bとも語の累積とともに減少し，著者と分野への適応が進むことを示す。ことに，分野が狭く限定されている資料Aでは著しい。

2）同音語　同音語出現率も同じく図2に示してあるが，これからわかるように語の累積による影響はほとんど見られず，著者により，あるいは入力区分毎にばらつきが大きい。これは次のように解釈される。

(i) 辞書中には入力中の文献に出現する語が累積され，それに伴って同音語も一定の割合で増加する。

(ii) 一方，辞書が一杯になった後，新出語が出現するごとに不要な語が削除され，それに伴って同音語も一定の割合で減少する。

(iii) 観測される同音語出現率は上記2者の和であり，当該文献で使用される語のあらわれ方による。これは著者個人により，また各文献の部分により差異が大きいと考えられ，したがって一定の傾向は示さない。

いずれにせよ，出現する漢字語のべ語数の20％前後は同音語を有し，しかも新出語と異なり語の累積によっても減少しない。したがって市販の汎用日本語ワードプロセッサと同様，本システムにおいても同音語の選択処理をいかに行うかが，システムの総合的な性能，使い勝手をよくする上で重要である。

3）新出文字　新出語中，文字辞書に登録されていない漢字の字数は資料Aでは12字，Bでは66字で，いずれも極めて少く，文字辞書を設けた効果は大きい。

4）誤変換語　本システムにおいて，誤変換は次のような機構で発生する。

(i) 読み w を有する語が辞書に登録されているものとし，その語を W_1 とする。

(ii) ここで，入力文中に W_1 と同一の読み w を有する語 W_2 が出現したとする。

(iii) W_1 は辞書中に読み w, 同音語数1（その語のみ）として登録されている。したがって新たに出現した語 W_2 は W_1 に変換され，誤変換となる。

したがって，誤変換語についても累積による影響はあまり大きくないと考えられるが，不要な語の削除により若干減少することが期待できる。この結果は図3に示すようにほぼ予想どおりとなっている。

この他，辞書への登録ミス，入力時の打鍵ミス等に起因する誤まりが発生するが，これらは誤変換とは別に扱っている。これらは偶発的に発生し，特定の傾向は示さない。また，量的にも入力総字数の

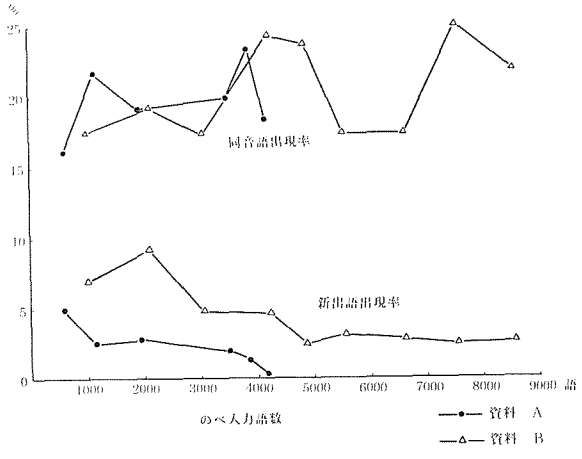


図2　個人の継続使用による変換特性の推移

0.5％以下で極めて小さく，全体の性能に与える影響も小さいのでここでは無視している。

5）正変換率　入力された語のうちから新出語，同音語，誤変換語を除いた残りの語は，何の操作も必要とせず，自動的に正しく漢字に変換される。そこで，この比率を正変換率と称し，これにより変換特性をあらわす。

正変換率には次の2種類の定義が考えられる。

(i) 語単位正変換率　入力文中の漢字語の部分だけについて求めた正変換率で，次式によってあらわされる。

正変換率（語単位）＝

$\{K-(W_n+W_h+W_e)\}/K$

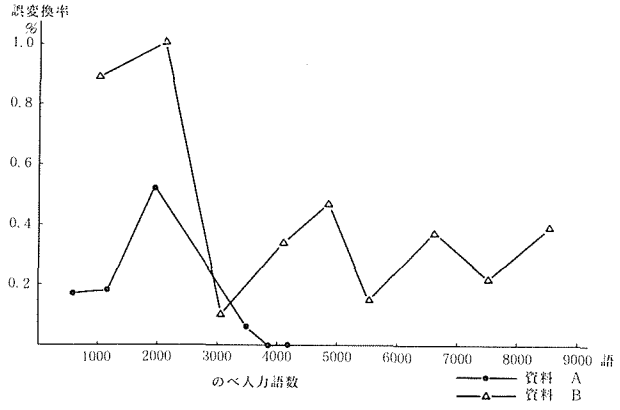


図3　誤変換率の推移

表3　実験結果一覧

(a) 資料 A

章	1	2	3	4	5	6	全体
入力字数，C	2987	2535	3477	7327	1984	1321	19631
漢字語数，K	597	558	774	1549	382	309	4169
漢字字数，K_C	1093	970	1368	2742	652	576	7401
新出語数，W_n	29	14	22	29	5	1	100
W_n/K (%)	4.9	2.5	2.8	1.9	1.3	0.3	2.4
新出文字字数	5	1	1	4	0	1	12
同音語数	96	121	148	308	89	57	819
W_h/K (%)	16.1	21.7	19.1	19.9	23.3	18.4	19.6
誤変換語数，W_e	1	1	4	1	0	0	7
W_e/K (%)	0.17	0.18	0.52	0.06	0.00	0.00	0.17
正変換率$_1$ (%)	78.9	75.6	77.5	78.2	75.4	81.2	77.8
〃 $_2$ (%)	92.3	90.7	91.2	91.8	91.9	91.8	91.6

(b) 資料 B

章	1 A	1 B	2 A	2 B	3 A	3 B	4 A	4 B	4 C	全体
入力字数，C	8106	8184	7581	9825	6733	5367	8897	7982	8622	71297
漢字語数，K	1011	1112	938	1161	642	671	1067	927	1015	8544
漢字字数，K_C	1740	1929	1665	2005	1050	1219	1869	1583	1791	14851
新出語数，W_n	70	102	45	53	16	21	30	23	26	386
W_n/K (%)	6.9	9.2	4.8	4.6	2.5	3.1	2.8	2.5	2.6	4.5
新出文字字数	10	17	9	11	1	3	6	2	7	66
同音語数，W_h	177	214	163	282	152	117	185	232	222	1744
W_h/K (%)	17.5	19.2	17.4	24.3	23.7	17.4	17.3	25.0	21.9	20.4
誤変換語数，W_e	9	12	1	4	3	1	4	2	4	40
W_e/K (%)	0.89	1.08	0.11	0.34	0.47	0.15	0.37	0.22	0.39	0.47
正変換率$_1$ (%)	74.7	70.5	77.7	70.8	73.4	79.3	79.5	72.3	75.2	74.6
〃 $_2$ (%)	94.6	93.0	95.0	94.0	95.8	95.3	95.7	94.5	94.8	94.7

注）正変換率$_1$：語単位正変換率，正変換率$_2$：文字単位正変換率

ここで，K: 漢字語数，W_n: 新出語数，W_h: 同音語数，W_e: 誤変換語数。

(ii) 文字単位正変換率　漢字の他，かなその他を含めた全字数に対する正変換率で，新出語，同音語，誤変換語各々の漢字の字数を求め，次式によって求める。

$$\text{正変換率（文字単位）} = \{C-(C_n+C_h+C_e)\}/C$$

ここで，C: 入力総字数，C_n, C_h, C_e: 新出語，同音語，誤変換語各々の字数。ただし，今回の測定では C_n, C_h, C_e は直接得られていないので，次式により換算した。

$$C_n+C_h+C_e=(W_n+W_h+W_e)\cdot\frac{K_C}{K}$$

ここで，K_C: 漢字語全体についての字数。すなわち K_C/K は漢字語1語あたりの平均字数をあらわす。

変換辞書単独の変換特性をあらわすには前者が適しているが，かな漢字変換システム全体としての性能をあらわす場合は後者が使われることがあり，たとえば実用レベルの目安として95%という値が示されている例がある[7]。

資料A，Bについて，新出語，新出文字，同音語，誤変換語など，すでに述べた諸量を含め，全体をまとめた結果を表3に示す。この結果，比較的漢字以外の文字の多い資料Bについて，ほぼ上記95%に達していることがわかる。

3.3　初期収録語と変換特性

以上の実験では，漢字語辞書の初期収録語を2303語としている。資料A，Bの入力後，辞書の内容を検討したところ，Aでは初期収録語の約3/4が，またBでは約2/3が未使用であった。し

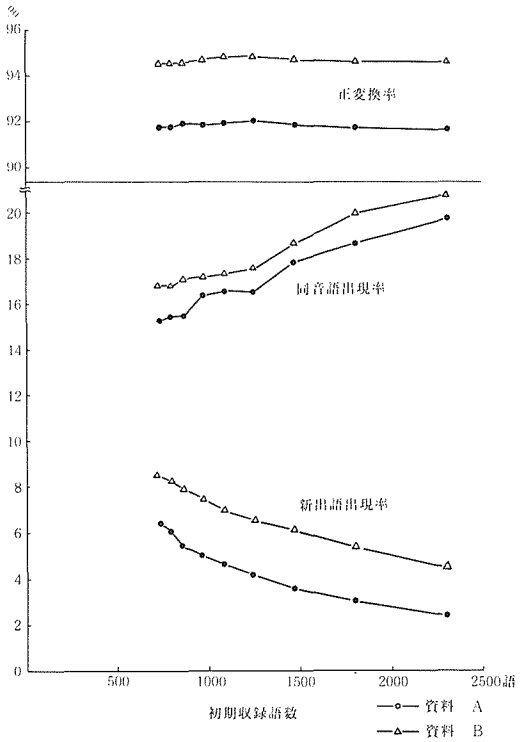


図4　初期収録語と変換特性

たがって，初期収録語をより少数にすることが可能と考えられ，それにより同音語出現率の減少が期待される。そこで，初期収録語を2303語より少くした場合について上記諸量を求めた。図4にその結果を示す。この結果から，同音語出現率は予想どおり減少し，一方新出語率は増加して，正変換率自体はほとんど変化しないことがわかる。ただし，同音語の選択操作は新出語の辞書への登録操作に比べ，はるかに簡単なのでこれから直ちに初期収録語が少数でよいという結論にはならない。

4. おわりに

使用者ならびに使用対象分野を限定した研究者向き日本語文入力システムにおいて，新出語の登録を容易に行うために，独立の文字辞書をおいてそこから必要な文字をとり出して語を合成し登録する方式を開発し，この機能を組込んだシステムの変換性能を求める実験を行った。得られた結果をまとめると以下のようになる。

1) かな漢字変換用漢字語辞書に，対象とする分野でよく使われる語を論文等から抽出し，初期収録語として登録することにより，小容量（2500語）で平均91～95％の文字単位正変換率が得られた。
2) 辞書中の収録語は，使用の継続により使用者，分野に適応し，新出語出現率は1～3％にまで低下した。
3) 文字辞書の組込みにより，新出語登録の際に漢字コード表を人手で検索しなければならない回数は，文字辞書組込み以前に比べ1/6～1/8に減少した。
4) のべ数千語の入力後も，辞書中には未使用の語が1/2以上あり，辞書の容量，収録語の選出，使用による適応過程等に改善の余地があることが示された。
5) 辞書内容の使用者への適応が進むにつれて，同音語の影響が大きくなり，同音語の処理方式がシステム全体の性能を左右する要素として重要になる。

本システムは総合的には十分実用に供し得る性能を有し，現在学会の大会や研究会等の予稿，卒業論文，修士論文，その他各種の資料等の作成に使用されている。今後，上記の結果を考慮し，とくに同音語選択の自動化について検討する予定である。また，他の専門分野の文献について入力実験を行い，初期収録語と異なる分野への適応性についても検討したい。

謝　辞

本研究を行うにあたり，種々有益な御意見，御討論をいただいた研究室の各位に感謝します。

参照文献

1) 栃内香次，斎藤　康：情報処理学会論文誌，Vol. 24, No. 2, pp. 209-213 (1983).
2) 栃内香次，永田邦一：電子通信学会技術研究報告，EC82-20 (1982).
3) 斎藤　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).
4) 日立製作所：HITAC 読み順配列漢字コード表，日立製作所コンピュータ事業本部技術本部 (1980).
5) 斎藤　康：ローマ字漢字変換方式による研究者向き日本語処理システム，北大大学院工学研究科修士論文 (1982).
6) 御牧　義：情報処理システム入門，昭晃堂，東京 (1979).
7) 森　健一，天野真家：電子通信学会誌，Vol. 63, No. 7, pp. 729-733 (1980).