

日本語科学技術文書処理支援環境 IDE : I. ヒューマン・インターフェースについて

宇津宮, 孝一
九州大学大学院総合理工学研究科情報システム学専攻

古保里, 学
九州大学大学院工学研究科情報工学専攻

野美山, 浩
九州大学大学院工学研究科情報工学専攻

<https://doi.org/10.15017/17598>

出版情報 : 九州大学大学院総合理工学報告. 6 (1), pp.69-76, 1984-06-01. 九州大学大学院総合理工学研究科
バージョン :
権利関係 :



日本語科学技術文書処理支援環境 IDE

I. ヒューマン・インターフェースについて

宇津宮孝一*・古保里 学**・野美山 浩**

(昭和59年3月31日 受理)

An Interactive Japanese Document Preparation Environment-IDE

I. Its Human Interfaces

Kouichi UTSUMIYA, Manabu KOBORI
and Hiroshi NOMIYAMA

Taking the properties of scientific and technical Japanese documents, and human factors for Japanese into account, we have developed a well-designed prototype of an interactive Japanese document preparation environment for scientists or engineers. This environment, called IDE, has better human interfaces easy to understand and easy to use. IDE introduces such better interfaces as an improved Kana-to-Kanji conversion algorithm, a hierarchical dictionary system, and a screen editor for Japanese texts in its editing system. These interfaces are discussed in detail and their effectiveness is shown experimentally by using the first version of IDE.

1. はじめに

近年の計算機による日本語処理技術の進歩がもたらした、いわゆる日本語ワード・プロセッサの出現と社会への浸透は、特に日本語文書作成上の諸問題を一気に解決してしまったかの印象を人々に与えている。

ところが実用に供されているシステムを、日本語科学技術文書の作成などに使用してみると、いかに作成者がシステムの性癖に注意し、精神的負担を強いられながら文書作りをしているかがよく分る。清書された文書には、こうした苦心の跡はみじんも現れてこない。このように一般文書とは性質の異なる日本語科学技術文書などの処理に関していえば、例えば、大学の研究者等が快適に使用でき、真に日本人になじむ使いやすいシステムは皆無であるといっても過言ではない。

その原因には、日本語入力を中心とする日本語処理技術に未だ解決すべき問題が山積していること、日本

語文書処理における個々の要素技術が未熟でしかも統合化されていないことなどがもちろんあげられる¹⁾。しかしそれにも増して、上述したシステムに日本人の人間の要因 (human factors) を考慮したヒューマン・インターフェースが欠如しているからではないかと思われる。

我々は、計算機システムにおけるヒューマン・インターフェースの研究の一環として、日本語科学技術文書処理支援環境 IDE (Interactive Japanese Document preparation Environment) の試作研究を進め²⁾、今回、ローマ字 (かな) 漢字変換方式に基づく入力・編集系のプロトタイプを実現した。これは、日本語科学技術文書の特徴を考慮し、新たに導入したローマ字 (かな) 漢字変換アルゴリズムと辞書環境、及び日本語画面エディタなどによって、使いやすい文書作成環境を提供しようとするものである。

本論文では、設計当初から考慮を払ってきた IDE のヒューマン・インターフェースを中心にして、その環境、特色、概要、及び試作版 IDE-0 の使用実験結果などについて述べる。

* 情報システム学専攻

** 工学研究科情報工学専攻修士課程

2. 日本語科学技術文書処理における諸問題

2.1 文書自身に起因する問題

日本語科学技術文書は一般文書とは違って、複合語、派生語、外来語から成る専門用語などが多用される³⁾。そしてこれらの用語の字数は長い場合が多く、また、カタカナ書きの外来語を除き、用語を構成する要素の読みは漢音が多い⁴⁾。

このような文書の作成を、間接日本語入力のための代表的な方式であるかな漢字変換で行おうとすると、

(1) 専門用語を収録する辞書

(2) (1)において、漢音が多いことに起因する同音異字語の出現

(3) カタカナ、記号など種々の字種の選択

などの新たな問題点が生じる。更に、複雑な図式を文書中に伴うことも考慮しておかねばならない。ただ、固有名詞は限定されること、使用される漢字の種類は一般文書に比して比較的少ないといった有利な点も含まれる。

2.2 利用者の問題

日本語科学技術文書処理システムを利用して文書作成を行う者は、例えば大学等の研究者においては次の特徴がある。

(1) 計算機の全くの素人ではないが、使用頻度はそれ程高くない。もちろん、文書処理の非専門家である。

(2) 英数字キーボード以外は不得手である。それも習熟訓練は受けていない。

(3) 各人の専門分野は非常に異なる。

(4) 各人は、計算結果等を科学技術文書の一部としても利用する。

このような特徴は、システムの利用者モデルを設定する際に十分考慮する必要がある。

2.3 システムの問題

一般の会話型システムにおいては、WYSWYG (what you see is what you get.) つまり見えるものがそのまま得られるという概念が重要である。諸外国では、この点を考慮したエディタ等の組織的な研究が進んでいるが、日本語文書処理に関しては非常に立ち遅れている⁵⁾。

従って、こうしたことを容易にするためには、高解像度ビットマップ・ディスプレイ装置とそのソフトウェアを有する画面指向のシステムが、理想的には各個人

人に必要である。更に、高品質の印字装置が利用できなければならない。しかしながら、このような快適なシステム環境は、現時点では非常に困難である。

3. IDE の環境とヒューマン・インターフェース

3.2 使用環境

日本語科学技術文書処理を中心とする IDE の使用環境は、上述した問題点の解決を図ってゆくという観点から、次のように設定した。

(1) 利用者モデル

計算機を自ら使用した経験はあるが、文書作成の非専門家である。使用頻度はそれ程高くない層から高い層までを含む利用者層である。当面、英数字キーボードに慣れた大学の研究者をその対象とする。

(2) システム環境

上記利用者層は、研究活動を含む知的作業を計算機によって支援する一環として、日本語科学技術文書処理を行う。従って、システム環境としては、日本語ディスプレイ装置を用いて、TSS 環境下で文書処理を行い、清書結果を高品質日本語プリンタに出力する環境を設定した。ただし、日本語ディスプレイ装置の制約上、差し当っては日本語文の作成・清書を中心としたシステム環境に限定する。

3.2 入力方式

上述した環境に最適だと考えられる日本語入力方式は、ローマ字 (かな) 漢字変換方式である⁶⁾。これは、直接的に漢字かな混じり文を入力できないので、間接入力方式と呼ばれる。この方式は、Fig. 1 に示すように3つの段階、すなわち、ローマ字 (かな) 入力、漢字列への変換、及び変換結果の確認という過程を経なければならない⁷⁾。

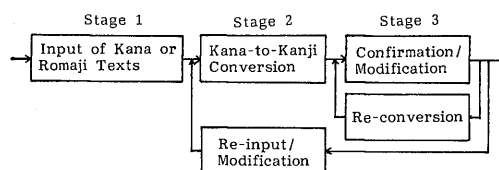


Fig. 1 Preparation process of Japanese documents.

これらの過程に費やす時間、各段階への移行の円滑化、及び全体の理解しやすさや使いやすさがシステムの良し悪しを決定する要因となる。

従って、これらの部分のヒューマン・インターフェースに対する配慮が特に重要である。そのため、IDEの入力方式で考慮した部分は次のとおりである。

(1) 通常の書き言葉に近い形の日本語表現

日本語文章の書き方には、英語のような正書法がない。そのため、様々な書き方が許されるので、こうした許容を認める形のローマ字列の入力が利用者にとって自然である。ところが、文章を書く場合、分ち書きをし過ぎると人間にとって負担となり、逆に、乏し過ぎると機械に受け入れられない。

従って、許容度を広くした(拡張された)文節入力
が、後刻発生する所望の日本語への再変換のことを考
えれば不自然でない最小単位であろう。このような理
由から、拡張された文節分ち書きローマ字入力を基本
とした。

(2) かな漢字変換アルゴリズムの改善

100%正しい変換率を達成するのは不可能である。
誤変換や再変換が多くなれば、それだけ文書作成者の
負担が重くなり、快適な入力環境でなくなる。

従って、辞書を含めた変換アルゴリズムの巧拙が、ヒューマン・インターフェースの重要な要因となる。我々は、大河内らが提案した文節末の付属語から切り出してゆくアルゴリズム^{8),9)}の変形である双方向最長一致法という変換アルゴリズムを考案して、日本語科学技術文書入力の変換率の向上を図った。

(3) 入力済み日本語列の活用

既に入力変換した漢字列については、ディスプレイ画面上で容易に再利用する画面編集機能（後述）を用いて、直接、文書の一部として活用する。その結果、類似文章の再入力の手間を省くことができる。

3.3 編集方式

誤変換、再変換を伴うローマ字（かな）漢字変換方式においては、とりわけ、ディスプレイ画面の構成、画面上での種々の編集操作などが使い勝手に大きな影響を与える。従って、これらの点には細心の注意を払って、そのインターフェースを作成しなければならない。

IDE で重視した部分は次のとおりである.

(1) ディスプレイ画面上では、目線ができる限り一定になるよう、変換後のテキスト、入力ローマ字列、その他のコマンド、メッセージなどの入力・表示域の適正配置を行う。

(2) 次の作業が円滑に行えるよう，カーソル位置

を適正化する.

(3) 画面エディタは任意の文字列を対象とする.

(4) 再変換は、同音異字語メニュー画面により、一括して容易に行う。

4. IDE の入力・編集系

IDE と利用者とのインターフェースの中心である
入力・編集系について詳述する。

4.1 入力・編集系の構成

Fig. 2 に IDE の入力・編集系のシステム構成を示す。図に示すように、これは3つの部分、すなわち、画面入出力管理部、コマンド解析部、及びテキスト入力部から成る。各部の機能は次のとおりである。

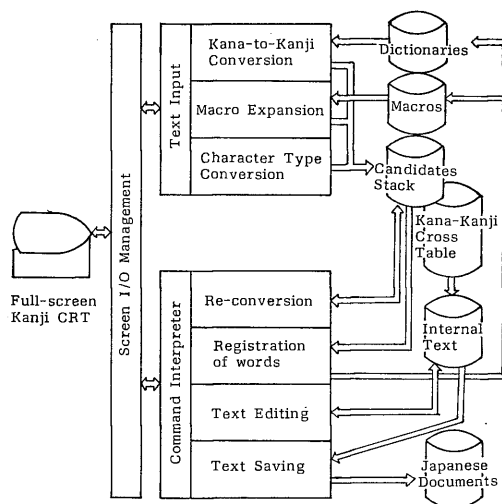


Fig. 2 Organization of IDE input and editing system.

(1) 画面入出力管理部

後述する種々の画面を統一的に管理し、しかもディスプレイ装置使用者にとっても使いやすい画面単位の入出力を実現する。

(2) コマンド解析部

画面のコマンド入力域 (**Fig. 3**) から入力されたコマンドを解析し、画面上に表示されたテキストの編集処理や保存処理、同音異字語等の再変換処理、及び新出語の個人用辞書 (後述) への登録処理などを行う。

(3) テキスト入力部

入力されたローマ字（かな）テキストに対して、指定された字種への字種変換、マクロ名から日本語列へ

の展開、及びローマ字（かな）漢字変換処理等を行う。

4.2 入力・編集における画面と処理

先に述べたように、IDE のヒューマン・インターフェースをよりよくするためには、入力・修正が容易で、理解しやすく、使いやすい日本語画面エディタが必要である。

4.2.1 画面

(1) 画面構成

画面は、主として、次の6つの画面、すなわち、初期設定、入力・編集、同音異字語選択、マクロ表示、個人用辞書登録、及び HELP の各画面から成る。

(2) 入力・編集画面

入力・編集画面の例を Fig. 3 に示す。図中、④は変換後の日本語テキスト表示域、⑤は変換結果表示域、⑥はローマ字（かな）テキスト入力域である。入力、確認など文章作成作業を円滑に行うためには、目線が一定になるよう、④、⑤、⑥の領域を近接した部分に置くことが人間工学的に最適であろう。また、変換後のカーソル位置は、かな漢字の正変換率が95%以上達成されねばならない点から考えると、当然、ローマ字テキストの次入力位置となる。これは、紙と鉛筆を用いた通常の文章入力モデルとは、紙の部分に当たる④の日本語テキスト表示域が、上方向にスクロールされるという点で一致しない。しかし、この方が合理的であり、慣れればそれ程不自然ではないと思われる。

(3) 同音異字語一括選択画面

日本語科学技術文書の入力においては、同音語の出現のために、少なくとも数パーセントの割合で再変換処理が必要となる。この再変換処理が、能率よくしかも分かりやすい形で行われることがヒューマン・インターフェースの観点から重要である。

我々は、変換結果表示域上で同音異字語の候補を次々に表示してゆくコマンド（Nコマンド）のほか、メニュー方式によって、一括して選択できる方法を提供した。この画面の例を Fig. 4 に示す。その結果、1行に複数個の同音異字語が現れる場合の再変換を能率よく容易に行える。

4.2.2 編集コマンド

編集操作においては、紙と鉛筆の文章作成モデルと同様の機能を有するコマンドが必要である。例えば、左側のような語の挿入を、

例) 我々の研究では → 我々の研究では
(室) .IB SITU

右側のように類似の操作で行う。この点に関しては、対象とする日本語ディスプレイ装置の機能不足から、必ずしも最適な編集操作はできないが、概念的には理解しやすいものである。

次の3種類の編集コマンドを用意している。

(1) 一般コマンド

テキストの表示や保存、辞書の割り当て等を行う。

(2) テキスト表示域コマンド

テキスト表示域上の文字列の編集操作

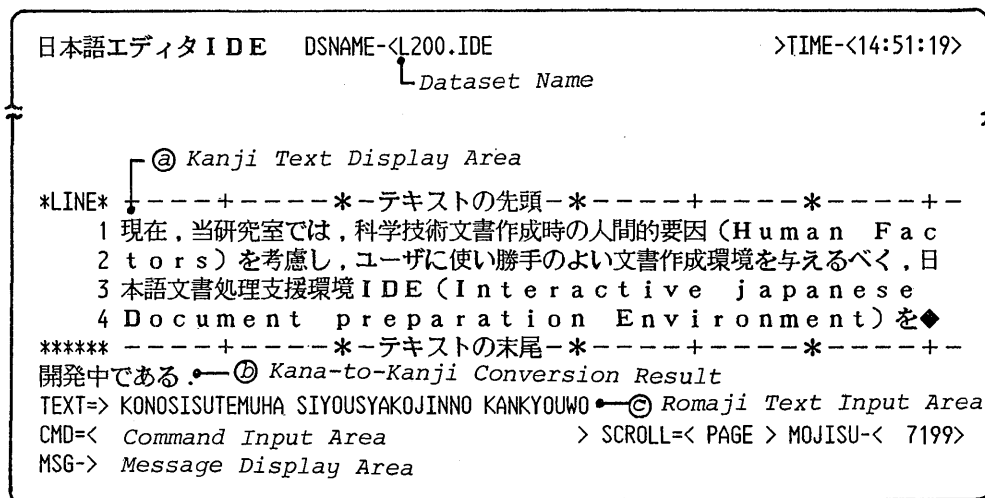


Fig. 3 Screen layout.

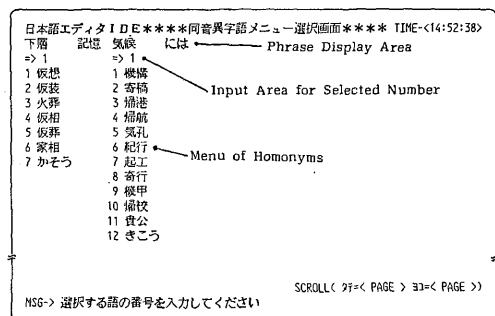


Fig. 4 Homonym selection screen.

(3) 変換結果表示域コマンド

変換結果に対する処理

Table 1 に上記 (2), (3) のコマンド一覧を示す。

4.3 かな漢字変換

4.3.1 基本方式

前述した使用環境に適応させるため、かな漢字変換の実現に次の基本方式を採用した。

(1) 文書作成者の入力負担の軽減に対して

ローマ字表記を読み言葉としてばかりでなく、書き言葉として位置付けし、例えば、小文字かな「っ」などは次のように発音と無関係に表現できる記法を導入した。

「設計」: 'SETU; KEI' 又は 'SEKKEI' と入力。

次に、入力単位に関しては、Fig. 5 に示す拡張文節を基本とする文節分ち書き方式を採用した。これ

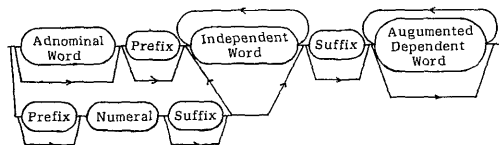


Fig. 5 Augmented phrase structure.

は、研究段階にあるべたがき入力方式を除けば、現時点では最適の方式である。通常、作成者が迷う「～によって」や「～に関して」などを拡張付属語と見なし、拡張文節の一部として許容するので、その負担は非常に軽減できる。

(2) 科学技術文書の特徴に対して

複合語、外来語（カタカナ）、外国語（特に英語）などが多用される科学技術文書の特徴を考慮し、次のようにローマ字（かな）漢字変換に反映させた。

複合語の切り出しは、後述する双方向最長一致法という変換アルゴリズムを用いて、その解決を試みた。

外来語（カタカナ）については、辞書に未登録の語はカタカナで表示することによって、これは、結果的に外来語が出てくるという単純な方法であるが、経験的にある程度うまくゆくものである。

外国語特に英語については、ローマ字列として認識できないものを英字列として扱う方法を採用した。この方法では、誤り入力されたローマ字列と英字列とは、英単語辞書を用いなければ区別できないが、当面は、ローマ字列の修正・再入力を容易にするという手

Table 1 Editing commands.

Command	Function
(I)*	IA, IB DD-DD, DS MM-MM, MS CC-CC, CS RR-RR, RS XX-XX, XS PP-PP, PS QQ-QQ, QS Mark the position after or before which the moved or copied character strings should be inserted Delete character strings Move character strings Copy character strings Replace character strings Re-convert a character string Register a character string in a personal dictionary Register a character string in a macro library
(II)*	.N .T .M .S .F .D .R Select a next homonym Return a word to its first candidate Menu selection of homonyms Convert another phrase structure Return the conversion result of a phrase to its first candidate Delete a phrase Replace a phrase

* Commands used in Kanji text display area

** Commands used in conversion results display area

法で解決を図った。

(3) 辞書に対して

ローマ字（かな）漢字変換の9割以上は、使用する辞書環境に左右される。これは、文書作成者の分野や環境が異なる科学技術文書に関しては、特に著しい。常に大辞書を引くという従来の方式に代えて、個人の使用環境に適応させた個人用辞書を中心にした新しい辞書環境を構築することによって、ローマ字（かな）漢字変換に対応した。

4.3.2 辞書環境

ローマ字（かな）漢字変換に使用する従来の辞書モデルは、大辞書中心であり、机上作業で、通常、研究者等が科学技術文書を作成する場合とは非常に趣を異にする。すなわち、通常の文書作成においては、各個人が有する（頭の中の）辞書で大半のことを済ませ、これで解決できない語を国語辞典等で引く。しかも、一度辞書引きした語は、少なくともその文書作成中は覚えているものである。これらの点を明確にすると、辞書モデルとしては階層構成が自然であり、それによって IDE の辞書環境を Fig. 6 のように設定した。

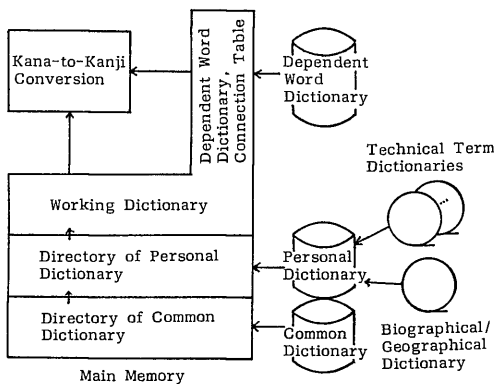


Fig. 6 Dictionary system.

この中で、作業用辞書は、会話中に辞書引きされた単語を一時的に保存しておく短期的な学習機能をもつ主記憶内辞書で、主として同音異字語処理に使用される。

共通辞書は、約8万3千語の自立語を収録した辞書で、利用者が共通的に使用するものである¹⁰⁾。

個人用辞書は、各人の分野で使用する専門用語や人名・地名等の固有名詞、及び各個人がよく使用する語彙などを収録した辞書であり、個人の環境に適応する

長期的学習機能をもつ辞書である。

辞書階層上で個人用辞書を重視すれば、大辞書を引く際の検索効率の問題や科学技術用語などに非力であった従来のシステムの弱点を大幅に改善することができる。個人の環境に適応する辞書のみでも十分高い精度の変換を行えるという報告もある¹¹⁾。

4.3.3 変換アルゴリズム

IDE は、科学技術文書に多用される複合語の処理を重視し、更に、入力者の負担を軽減するために、Fig. 5 に示す構造をもった拡張文節に対して変換処理を行うアルゴリズムを採用した。

この変換アルゴリズムを双方向最長一致法と呼び、Fig. 7 に示すように、次のステップで変換処理を行う。

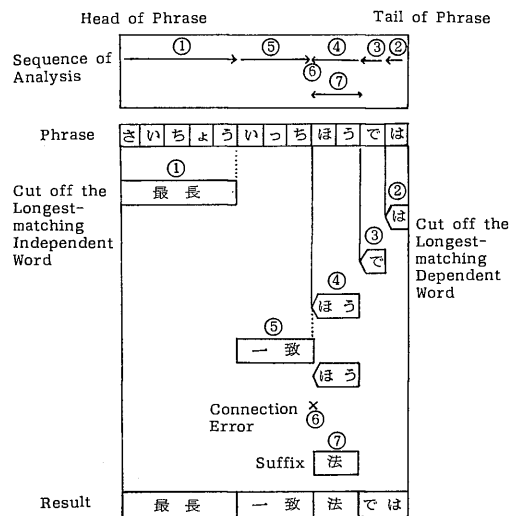


Fig. 7 Leftmost and rightmost longest-matching algorithm⁹⁾.

ステップ1：文節頭から自立語辞書（個人用辞書，共通辞書）を引き，最長一致した語を候補語とする。

ステップ2：逆順付属語辞書を用いて，文節末から（拡張した）付属語列を，単語間の接続テストを行いながら切り出してゆく。

ステップ3：入力文字列が双方で一致しない部分が残れば，ステップ1と同様繰り返し自立語検索を行う。残り部分のかな文字数が3以下のときは，接尾辞の処理を行う。

ステップ4：自立語と付属語が重複するときは，自立語の最長一致を優先して候補語とする。

こうして得られたものを変換結果として表示する。

以上の手順によって、科学技術文書中に多く現れる「単純語の繰り返し+接尾辞」型の複合語は、ほぼ正しく変換される。また、双方向から行うことによって、次のような構造的にあいまいさをもつ文節

「くるまでは」: 車では、来るまでは

に対しても、二通りの変換結果を得ることができる(編集コマンド . S)。

しかしながら、接頭辞を検出しにくい場合や単純語中に更に接辞を含むような場合については、このアルゴリズムの中にこれらを解決する手立てを講じなければならない¹²⁾。

4.3.4 個人用辞書

IDE の個人用辞書を次の方針で設計・試作した。

(1) 見出し数は2〜3千語とし、できる限り主記憶への常駐を試みる。最悪の場合でも1回のアクセスに抑える。

(2) データセットの構成は、更新が容易なようにB氏木構成とする。

(3) 辞書引きは、共通辞書に先んじて引く。その後の処理は共通辞書の場合と同じである。

(4) 辞書内容には、頻度のほか参照情報も格納

し、同音異字語処理に反映させる。

(5) 辞書内容の登録と更新は、読み振り、品詞情報の推定をできる限り自動的に行い、利用者の負担を軽減する。

これらの方針に基づいて、主として情報処理の分野を対象にして個人用辞書を試作してみた。試作個人用辞書は、情報処理学会論文誌の論文約30編(延べ約3万5千語)などから収集した単語¹¹⁾、共通辞書との突き合わせを行って、これに収録されていない語及び優先順位の異なる語を選定したものである。約1,300語を収録しているが、現時点では、関係深い人名・地名などの固有名詞は含まれていない。

5. IDE の使用実験

完成したプロトタイプ IDE-0 (第0版)を用いて、IDE の仕様の確認と予備的な評価を行うために、日本語科学技術文書作成に関して実施した使用実験とその結果について述べる。

5.1 ローマ(かな)漢字変換結果

専門用語を含まない文書の例として、武者小路実篤の「人生論」、科学技術文書の例として、情報処理学会論文誌から選んだ3編の論文及び技術マニュアル1部を対象とし、IDE-0 の漢字への変換率を調査した。

Table 2 Hit ratios for two algorithms.

Samples	Number of Phrases	Algorithms	
		Leftmost Longest-matching	Leftmost and Rightmost Longest-matching
I*	12,407	92.1 %	93.2 %
II*	20,631	83.5	89.1

* JINSEIRON written by Saneatu Mushanokoji

** A computer manual

Table 3 Hit ratios for three dictionary environments.

Dictionary Environments	Samples (Number of Phrases)				
	JINSEIRON (1,735)	Papers			Computer Manual (2,116)
		Operating Systems (2,564)	Speech Recognition (2,417)	Japanese Document Preparation (2,346)	
I*	89.7 %	80.0 %	81.8 %	76.8 %	77.6 %
II**	93.3	88.2	90.0	86.4	84.3
III***	93.3	94.3	90.5	93.1	92.2

* Common dictionary only

** Working dictionary besides I

*** Personal dictionary besides II

(1) 変換アルゴリズム

Table 2 に変換アルゴリズムの相違, すなわち従来の左方向最長一致法と IDE で採用した双方向最長一致法による変換率を示している. 科学技術文書の例に対しては約 5% の変換率の改善があり, このアルゴリズムの有効性が示された.

(2) 辞書環境

辞書環境の違いによる変換率の推移を Table 3 にまとめている. 表は, 作業用辞書が意外に科学技術文書の変換率向上に有効であること, 試作個人用辞書は有効ではあるが, まだ十分にその効果を発揮していないことなどを示唆している. これは, 辞書見出しの選択に問題があり, まだ個人の環境に十分適応していないことが原因だと思われる.

5.2 画面操作

画面操作については, 使い勝手と応答時間が特に重要である. 応答時間については, 現在の版が画面全体を送受信していることが原因か, 多少気になる. 一方, 画面の操作は, カーソルの移動及びファンクション・キーの用法に関して必ずしもまだ使い心地がよくない. 今後, 使い込みを行いながら改善してゆく必要がある.

6. おわりに

以上, 日本語科学技術文書処理環境 IDE のヒューマン・インターフェース部分を中心に述べてきた. 試作版 IDE-0 の使用経験から見るかぎり, ローマ字(かな)漢字変換率, 編集画面, カーソル位置などまだ改良すべき点が多い.

しかしながら, 我々が設計段階で考慮した本質的な部分は, 十分その効果を発揮しているとの感触を得ることができた. 使用実験で明らかになった問題点は, データを積み重ねてゆけば更に大きな改善が期待できる見通しももてた.

IDE-0 の改善を通してのシステムの詳細な評価及び IDE の清書系については改めて報告をしたい.

最後に, 日頃ご指導頂く本専攻駒宮教授並びに辞書の便宜を図って頂いた本学工学部吉田教授及び北大工学部柝内助教授に謝意を表す. また卒業研究で個人用辞書の試作と評価をして頂いた小鶴康浩氏に感謝する.

参 考 文 献

- 1) 日経エレクトロニクス: 日本語処理特集, 8-29, 324, pp. 179-246 (1983).
- 2) 宇津宮ほか: 日本語文書処理システム IDE の入出力部, カナ漢字変換, 個人用辞書, 第36回電気関係学会九支連大論文集, pp. 460-462(1983).
- 3) 林: 図説日本語, p. 597, 角川書店 (1982).
- 4) 渡辺, 大岸: 情報処理の用字と用語, 情報処理学会日本文入力研究会, 10, pp. 1-10(1983).
- 5) J. Nievergelt, G. Coray, J. D. Nicoud, and A. C. Shaw: Document Preparation Systems, p. 274, North-Holland (1982).
- 6) 栗原, 黒崎: 仮名文の漢字混じり文への変換について, 九大工学集報, 39, 4, pp. 659-664 (1967).
- 7) 中山, 黒須: 日本語入力速度予測モデルの検討, 情報処理学会日本文入力研究会, 13, pp. 1-10(1984).
- 8) 藤崎, 大河内, 諸橋: 「ことだま」文書処理システムの文節わかち書き仮名漢字変換, 情報処理学会論文誌, 23, 1, pp. 1-8(1982).
- 9) 大河内: 分かち書き方式仮名漢字変換のためのバックトラックを必要としない文法解析, 情報処理学会論文誌, 24, 4, pp. 389-396(1983).
- 10) 稲永, 吉田: 日本語処理のための機械辞書, 情報処理, 23, 2, pp. 140-146(1982).
- 11) 柝内, 斉藤: 適応型変換辞書を用いるかな漢字変換, 情報処理学会論文誌, 24, 2, pp. 209-213 (1983).
- 12) 野村: 複次結合語の構造, 国立国語研究所報告 49「電子計算機による国語研究 V」, 秀英出版 (1973).