かな漢字変換の機能レベルと適合システム

戸井田 徹　　木村 久正
（日本電信電話公社横須賀電気通信研究所）

1．まえがき

日本におけるオフィスオートメーション（ＯＡ）の普及には漢字を扱う情報処理技術の発達によるところが大きい。漢字の持つ多様な性質から漢字を含む情報処理システムでは入力が重要な問題である。ＯＡの普及には素人にも操作が容易な入力方式が必要であり、これら要求に適する代表的な方式[1)]としては

１）かな漢字変換方式、
２）タブレット方式、

などの手操作による２つの方式があり各種開発されている。

かな漢字変換方式はかな鍵盤を用いるため習熟すればブラインドタッチの入力が行え、疲労の少ない高速な入力が可能となる。また、日本語に関する文法、構文、意味関係などの性質の研究が進んでおり、これらを利用した自然な入力によるかな漢字変換方式の確立が期待されている。しかし、現状では、同音語の同定にインタラクティブな操作が必要な事、自然な入力を可能にするには多くの処理を必要とすること、等の問題がある。かな漢字変換の機能レベルは入力方式により

ａ）漢字指定
ｂ）文節指定
ｃ）べた書き

に分類され、それぞれ処理時間、メモリ量が異なる。したがって、かな漢字変換処理をシステムに導入する場合には入力方式の選択と処理をどこで分担するかが問題となる。

そこで、本報告では

ａ）センタ
ｂ）端末装置
ｃ）キーボード装置

上で具体的に実現した、かな漢字変換処理の特徴と評価から、処理レベルと入力時間、システム規模との関係を求め、システム導入おける変換処理の機能分担、利用分野を明らかにしたので報告する。

2．かな漢字変換処理

かな漢字変換の技術上の主な問題点は単語の抽出と同音語の同定である。表１に入力方式と処理内容を示す。

表１ かな漢字変換方式

入力方式	入力単位	主な処理
表示選択	文字	漢字辞書検索
漢字指定	単語	単語辞書検索,文法処理
文節指定	文節	単語抽出,形態素解析
べた書き	文	構文・意味解析

実用レベルにあるのは、入力単位を文字、および単語とすることにより、単語抽出における誤りを除いた方式である。

今回、各実験システムへ導入したものは実用的な変換率が期待できる漢字指定方式に属するものである。また、システムおよび装置の特徴により、かな漢字変換処理内容を変え、それらの効果を測定した。

2．1　処理概要

実験システムに導入したかな漢字変換処理の概要を図１に示す。

ａ）入力解析

変換部の前後の情報から単語の品詞などを決定する。

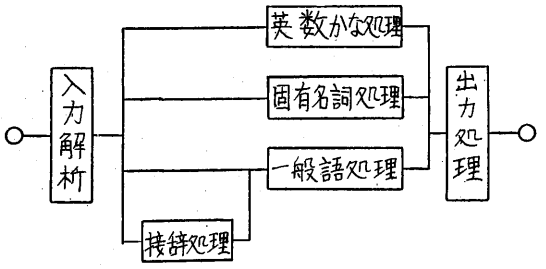

図1 処理の流れ

b）接辞処理

接頭、接尾、数詞を収録したテーブルを利用し、入力解析で接辞と判定された語を変換する。

c）一般語処理

一般語辞書を利用したかな漢字変換処理を行う。

d）固有名詞処理

固有名詞辞書を利用したかな漢字変換処理を行う。

e）出力処理

同音語がある場合、出力順を決定する。

かな漢字変換処理において変換率を向上するため検討した処理内容を以下に示す。

a）辞書指定

かな漢字変換処理を指定された辞書を使用して行う。

［例］固／たろう／固有名詞辞書検索
／かんじ／ 一般語辞書検索

b）頻度処理

同音語がある語は使用頻度が高い語を優先的に出力する。

c）ＩＤＩＯＭ処理[2)]

単語のＩＤＩＯＭ符号が一致する語を優先的に出力する。

［例］□いかん□ １．□移管□
２．□偉観□
３．□遺憾□

e）文法処理

品詞の適合性の判定、活用語尾の適合性判定、付属語の接続性判定を行い整合文字数の多い語を優先的に出力する。

f）学習処理

同音語を有する語において、同音語処理で選択された語を優先的に出力する。

g）同音語処理

出力された語が該当語でない場合は候補語を列挙し選択する。

h）単語登録

辞書に未登録な語を漢字辞書を使用し作成し、私用語として登録する。

２．２ 辞書

辞書の構成を表２に示す。一般語は最も基本的な語彙とし、使用頻度を考慮し収録語数を選択した。固有名詞はカバー率、語彙の共通性を考慮し出現頻度の高い姓、名、主な企業名、ＪＩＳに定められた地名を収録した。漢字辞書は文字単位の修正、および未登録語の作成に用いることから、ＪＩＳＣ６２２６に定められる字種について音、訓、部首、画数を見出しとして収録した。

表2 収録語彙

種類	収録語彙
一般語	25000
固有名詞	姓 10000 名 10000 地名 3000 企業名 3000
漢字	JISC6226 第1,第2水準 (音・訓,部首,画数見出し)

３．かな漢字変換システムの構成

３．１ センタシステム

本システムはセンタおよびクラスタ端末など、かな入力や同音語の処理と、かな漢字変換を分散して行うモデルである。また、本システムはセンタにお

いて、かな漢字変換および文書処理を行うことにより、既存の英数カナ端末を使用した日本語文書処理が行える。[3)]

（１）システム構成

システムの概要を図２に示す。端末はＪＩＳキーボード、サーマルプリンタ（５０文字／行）、音結部を同一筐体に有する携帯形漢字プリンタである。センタは公社内システム（ＤＩＰＳ－１）を使用し、端末とは電話回線（３００ｂｉｔ／ｓｅｃ）で接続する。

キーボードにおいて、ＩＤＩＯＭ記号入力用キー（ＩＤＩＯＭキー）は操作性を向上するため２種類とし、親指で打鍵できるようスペースキーを３分割し、その内の２キーとした。

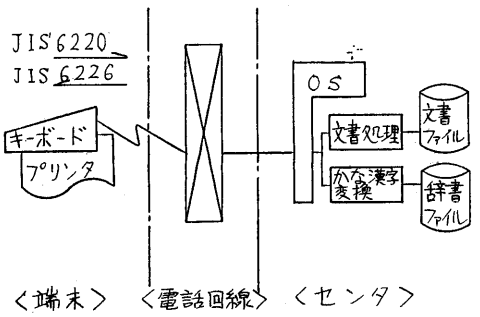

図２ システム構成

（２）入力・変換機能

かな漢字変換を行う方法として

ａ）入力かなデータをファイルに登録し、一括して変換を行うバッチ的処理

ｂ）単語ごとに逐次変換する会話的処理

の２つの方法とした。

（３）評価

本システムでの実験結果の一例を図３に示す。センタにかな漢字変換処理、簡単な文書修正、編集機能を付加することにより、携帯形漢字プリンタにより漢字入力および文書作成ができることが確認された。

コマンドを入れて下さい。
/INP TOIDA1 入力指定
入力文を入れて下さい。
ﾛﾎﾝﾛｺﾝﾋﾟｭｰﾀｼｽﾃﾑﾉﾉﾄｸﾁｮｳﾛﾊﾛｹｲﾀｲ ｶﾞﾀﾛﾒﾉ
ﾛﾌﾟﾘﾝﾀﾛﾃﾞﾛﾌﾞﾝｼｮ ｼｮﾘﾘﾛｶﾞﾛｶﾝｼﾞﾉﾛﾉｺﾄﾃﾞｽ。 入力データ
編集を行って下さい。
01 本システムの特長は形態形 ↙
= 同音語列挙指示
* 2 1形態 2携帯 ／
* 2 同音語選択
01 本システムの特長は携帯形 ↙
= ｗ
02 プリンタで文書処理が可能のことです。
= ==== : ==== 編集位置指定
= ﾛﾍﾝｼｭｳﾛ
= ｶ
= ｶﾝｼﾞ 漢字検索 } 訂正データ入力
* 1S 1天 2典 3店 4点 5展 6転 7添 8殿
* 4 漢字選択
02 プリンタで文書編集が可能な点です。 ↙

注）□で囲まれた部分は端末からの入力である。

図３ 処理例

本システムのかな漢字変換部のステップ数はＣＯＢＯＬで約２Ｋであり、辞書は固定長の索引順編成で１．２Ｍバイトである。

また、かな漢字変換率は、新聞の社説を対象にした正変換率（該当する単語が第１順位で出力される割合）で表わすと９０％である。

本システムの入力速度と入力コスト構成比を図４に示す。ＣＰＵコストは入力かなデータを記録してあるファイルからデータを読み込み、かな漢字変換を行い、出力ファイルに書き込みを始めるまでのコストである。また、かな漢字変換処理の内訳は、新聞社説の用語調査から漢字の単語数を求め、かな漢字変換率、辞書のカバー率（単語が辞書に収録されている割合）から同音語の選択回数、未登録語の入力回数を計算したものである。Ａ４判用紙１頁（１０００文字程度）の文書の変換

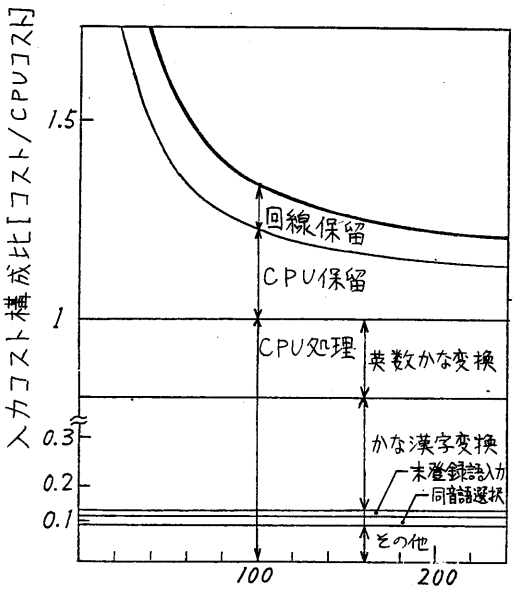

図4 入力コスト構成

処理時間は約１０秒であり、会話形式で入力をおこなう場合、応答速度に問題はない。また、入力コストに占める利用者の入力速度の影響が大きい点や、未登録語の入力をオンラインで行うことはＣＰＵ利用効率の点から問題である。　したがって、入力データのバッファリング、および変換された漢字データの修正は端末で行う方式が適当である。

３．２　文書処理端末

（１）装置構成

端末装置におけるかな漢字変換処理として日本語文書処理端末に導入した例を示す。かな漢字変換処理では、処理内容と変換率の関係、辞書のメモリ量の圧縮方法を検討した。装置の構成を図５に示す。

装置は１ＭバイトのＦＤ２台、ＣＲＴディスプレイ（４０字／２０行）、キーボードからなる。

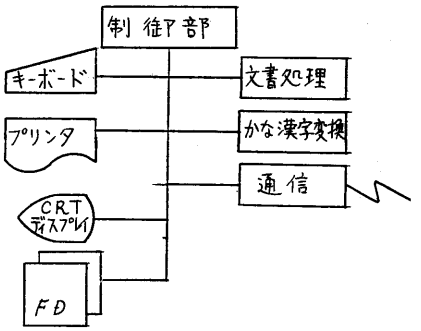

図5 文書処理端末

（２）変換処理と辞書

かな漢字変換処理の各内容を以下に示す。

ａ）学習機能

学習機能は、同音語処理で選択された語をテーブルに記憶し、同音語を有する語の出力順を決定するとき、テーブルに登録されている語を優先的に出力する方法で実現した。新聞の社説を対象とし、その効果を測定した例を表３に示す。学習処理の効果として正変換率（第１順位の語が正しい割合）が５％向上する。

表3 学習処理の効果

単語分類[%]						変換率[%]		
単語数	同音語なし	同音語有り	頻度1位	その他	管理情報	正変換	頻度処理	一意
100	43.4	56.6	37.3	19.8	5.2	85.4	80.3	43.4

この処理に使用するテーブルに必要なエリアを明らかにするため、新聞の社説を対象とし、同音語選択が必要となる語の出現分布について測定した。測定は正変換されない語に着目し、同じ単語が再び現れるまでに、含れる正変換されない単語数（出現距離と呼ぶ）を求めた。測定の結果を図６に示す。

図より、出現距離の平均は１２．６

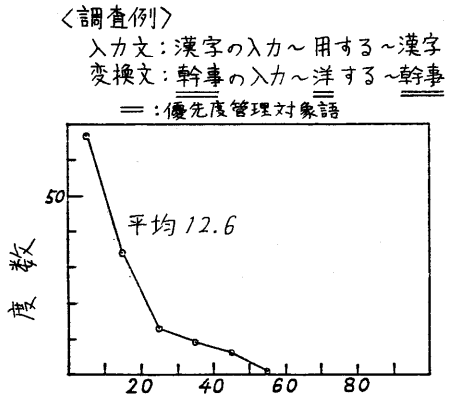

図６　出現距離の分布

語であり、標準偏差（１３．２語）の３倍を基準にすれば優先度管理テーブルのエリアとしては５４語が必要である。優先度管理テーブルは同音語を有する単語の変換ごとにアクセスされるため高速処理を必要とすること、およびエリアも５００バイト程度と少ないことから主メモリに常駐させることにした。

ｂ）辞書構成

一般語辞書の見出し語および漢字文字数の分布を調査した結果、見出し語の平均は３．９３文字、漢字文字数の平均は１．８６文字であり、見出し語、漢字文字数の最大文字数は各々８文字、４文字である。したがって、可変長とすることはデータの圧縮に効果がある。

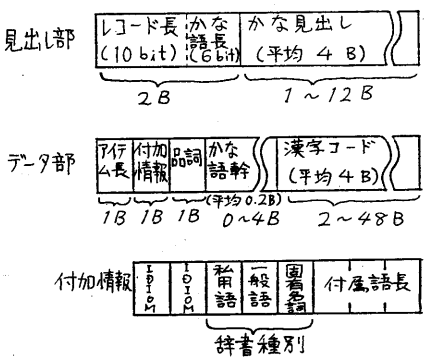

図７　レコード形式

さらに、一般語と固有名詞の識別をフラグにより行い見出し語の重複をなくすことにより、一語平均のメモリ量は約１０バイトとなり固定長に比べ６０％圧縮するできた。レコード形式を図７に示す。

（３）評価

本装置のかな漢字変換部のステップ数はアセンブラで８Ｋ程度であり、辞書は索引部とデータ部および私用語辞書を含め７５０Ｋバイトである。また、かな漢字変換処理時間は平均で５００ｍｓ程度であり、正変換率は９２．６％である。

３．３　キーボード装置

本装置は、かな漢字変換のファーム化を目的として開発したものある。辞書の記憶媒体として大容量のＲＯＭを使用し、また、かな漢字変換処理を単純にし、プログラムの独立性を持たせることによりキーボード部でかな漢字変換を可能とした。[4)]

（１）装置構成

キーボード部と、その評価のため試作した装置の構成を図８に示す。キーボードはＪＩＳおよび五十音のキー配列であり、辞書および変換部は一枚の基盤に実装され、キーボード部に収容されている。

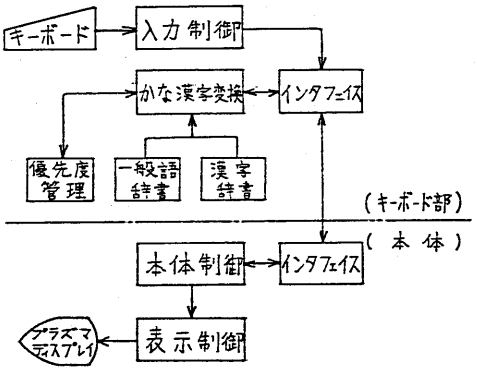

図８　装置概要

（２）変換処理および辞書

本装置のかな漢字変換処理は辞書検索、およびＩＤＩＯＭ記号の適合性判定、学習処理だけを行う簡単なものとした。

辞書の固体メモリ化を図るとき、記憶媒体としては、

ａ）半導体メモリ、

ｂ）磁気バブルメモリ、

が考えられるが消費電力、ディバイスの大きさ、アクセス速度の点から半導体メモリを記憶媒体とした。使用したディバイスは漢字のパターン用として開発された大容量ＲＯＭ（ＭＵＧＲＯＭ）で１Ｍビットのものである。[5] ＭＵＧＲＯＭは漢字コードにより漢字パターン（３６バイト）を読み出す周辺回路を内蔵しているので、本装置では１０文字を１ブロックとし、３６０バイトをアクセス単位とした。辞書構成を表４に示す。[6]

表４ 辞書構成

種別	収録データ	メモリ量
一般語	24000語 インデックス	2Mビット
漢字 一文字	JIS第1,2水準 インデックス	1Mビット
ROM仕様：外寸（53.75×38.1） 40PIN 記憶容量 1,082,880ビット/チップ 消費電力 500mW 5V単一電源		

辞書検索はＭＵＧＲＯＭの使用によりアクセス速度が早いため、辞書検索に使用する装置の主メモリのエリアの圧縮に重点を置き、インデックス部を２段とした。

（３）評価

本装置のかな漢字変換部はアセンブラで２Ｋステップである。　また、かな漢字変換処理は辞書への３回のアクセス時間を含め約６０ｍｓｅｃである。[5]

新聞社説を対象にした場合、正変換率は８５％程度であるが、文法処理を追加することにより９０％以上まで向上が図れるので、かな漢字変換処理のファーム化は実用性を有する。

４．処理レベルと適合システム

（１）処理内容と変換率

かな漢字変換率を向上させる処理内容の効果を新聞社説を対象として測定した。各処理内容と変換率との関係を図９に示す。頻度により同音語の優先度を決定する処理のみでは正変換率は８０．２％であるが、ＩＤＩＯＭ適合性判定、文法処理、学習処理を行うことにより正変換率は９２．６％まで向上する。今回、導入した各処理は用語の統計的処理、形態素解析処理の基本は全て含むものであるので、漢字指定方式によるかな漢字変換処理の変換率は約９３％程度である。

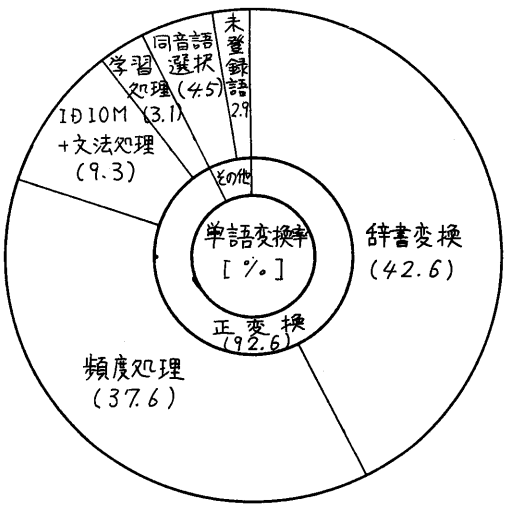

図９ 処理内容と変換率

また、漢字指定方式によるかな漢字変換処理プログラムの規模はアセンブラで１０Ｋステップ以下であり８ビット系のマイクロプロセッサで処理できる。したがって、漢字指定方式による

かな漢字変換処理は端末での処理が可能である。

2）変換率と入力速度

熟練者（１８０タッチ／分）を対象にした場合、入力方式と変換率が入力速度に与える影響について図１０に示す。図において、同音語処理時間、未登録語入力時間は実験から求めた値である。入力の単位を長くすることにより、かなデータの入力時間は短縮され、その効果は変換率が９５％の場合べた書き入力と漢字指定方式では１６％程度である。しかし、初心者においては、かなデータの入力時間が大きくなるため効果は減少する。また、べた書き入力によるかな漢字変換では、意味および格情報を含む辞書を必要とする、などの問題がある。

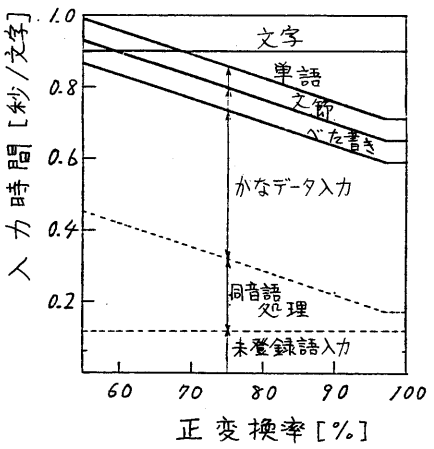

図10 入力時間と正変換率

（３）処理時間と入力速度

かな漢字変換の処理時間と入力速度の関係を図１１に示す。図中の破線は０．２ＭＩＰＳの計算機を使用した場合の処理時間であり[7),8)]、実線はオペレータのキー入力の時間間隔を示したもので、かな漢字変換処理を入力動作に追従させるには、この時間内にかな漢字変換処理を終了する必要がある。　図より、漢字指定方式はマイクロプロセッサでも十分応答性がえられ、文節指定方式においても高速なマイクロプロセッサの使用により応答性が確保できる。

したがって、手操作による漢字入力には漢字指定、文節指定のかな漢字変換を端末装置で行うことにより熟練者に対する応答性が確保できる。

べた書き方式は応答性の点から会話的に漢字を入力するには適せず、プログラム規模の点から端末装置で処理を行うのは困難である。べた書き方式の適用分野としてはインタラクティブな処理の少ない既にコード化されたかなデータを一括してかな漢字変換する処理がある。センタおよびクラスタ端末装置で既にコード化されたかなデータの変換を行い、その結果を端末装置に転送し同音語の同定処理や漢字の修正を行う場合、入力速度は３００字／分以上（図１０で変換率を９５％以上と仮定）となり高速処理が可能となる。

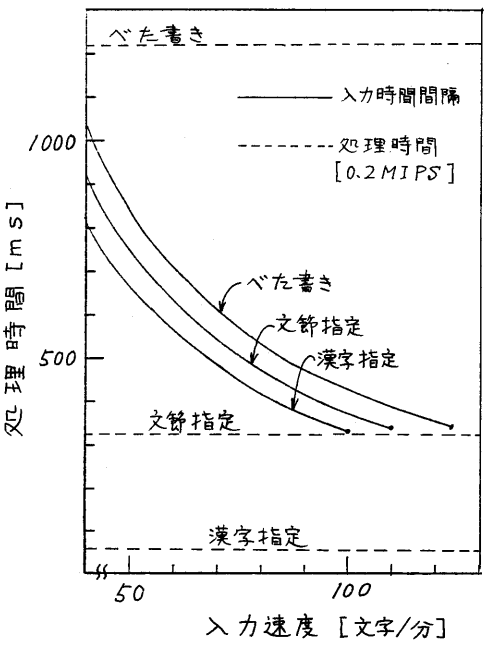

図11 応答性

５．むすび

漢字指定方式によるかな漢字変換処理を端末装置、およびセンタで行った評価から処理レベルと変換率、入力速度、システム規模の関係を明らかにした。また、入力速度から、かな漢字変換処理を装置およびシステムに導入する場合の入力方法の選択と利用分野を検討し以下の結果を得た。

１）手操作による漢字入力では、かな漢字変換は端末装置で会話的に行え、入力方式としては漢字指定、文節指定が可能である。

２）漢字指定方式のかな漢字変換は辞書のＬＳＩ化などによりファーム化できる。

３）既存のかなデータの漢字化はクラスタ端末、およびセンタにおけるべた書きの入力方式によるかな漢字変換と端末装置における同音語処理を組合せ、高速に処理できる。

今回は、入力速度の観点から各かな漢字変換方式の評価を進めたが単語や文節の抽出をオペレータが行うのは思考の中断など精神的な面の問題がある。今後は、これら人間要因も含めた評価を行う。

謝辞

本研究の遂行にあたり御指導をいただいた入力装置研究室小森室長に感謝致します。また、かな漢字変換アルゴリズムについて御助言をいただいたプリンタ研究室小橋補佐、辞書構成につき御助言をいただいた入力装置研究室山階主任に感謝致します。

文献

１）高野他；オフィスオートメーションと標準化、情報処理、Ｖｏｌ２２，Ｎｏ１０、１９８１

２）杉山他；特徴分類形日本語入力方式の検討，信学会研資ＥＣ７９－１３，１９７９

３）戸井田他；プリンタベース形日本語文書処理端末に関する一考察，昭５６信学会総合全国大会，１９８１

４）戸井田；漢字入力用インテリジェントキーボード，昭５６信学会情報．システム部門全国大会，１９８１

５）幸田他；漢字パタン発生用１Ｍビットｒｏｍ，信学技報ＳＳＤ８０－３２

６）山階他；かな漢字変換辞書のＬＳＩ化，情報処理学会第２３回全国大会、１９８１

７）牧野他；べた書き仮名漢字変換システムとその同音語処理，情報処理Ｖｏｌ２２，Ｎｏ１，１９８１

８）木村他；日本語入力用カナ漢字変換システムの試作，情報処理，Ｖｏｌ１７，Ｎｏ１１，１９７６