Vol. 27 No. 11 情報処理学会論文誌 Nov. 1986

連語解析を用いたべた書きかな漢字変換†

本間 茂†† 山階正樹†† 小橋史彦††

べた書き文のかな漢字変換では，文節分かち書きの場合に比較して同音語によるあいまいさがあるばかりでなく，分かち書き処理を自動的に行うことによるあいまいさも生じ，膨大な数の文解釈候補が発生する．このため文節内に閉じた形態素情報を利用した従来の解析法だけでは文解釈候補を十分には絞り込めないという問題があった．本論文では，こうした問題を解決する手段として，単語単体の属性だけでなく単語間の関係を用いる連語解析を提案し，その方法を用いたかな漢字変換法について述べる．本方式ではべた書きかな文から漢字かな混じり文への変換の基本処理に文節数最小法を用い，さらに，単語数最小という条件から候補を絞り込んだ上で連語解析を行い，最終候補の絞り込みを行う．また，連語解析に用いる連語辞書を自動生成する方法を実現し，国語辞典などの用例約30万件から係り受け関係にある単語の組を約30万5千件自動抽出した．これらを適用したべた書き文のかな漢字変換プログラムを作成し，実験を行った結果，新聞社説（約5,500字）を対象に平均文解釈候補数を文節数最小法の約1/7～1/10まで絞り込むことができ，本方式が有効であることがわかった．

1. まえがき

日本語ワードプロセッサの普及に伴い，日本語入力法としてかな漢字変換方式が定着してきている．当初は，漢字を一文字ごとに変換する字単位入力であったため，入力に要するキータッチ数が非常に多く入力速度の低いものであった．その後，入力規則が緩和された漢字指定入力，文節入力が実用化され，キータッチ数から見た入力速度は大幅に向上してきている[1]．しかしながら，文節に区切るための煩雑なキー操作が必要であること，文節に区切るための文法知識が必要であること等の問題点がまだ残されており，オペレータにとって，より自然で，より高速に入力が可能なべた書き入力の実現が要望されている．

べた書き入力では，文節入力の場合と同様に同音語によるあいまいさがあるばかりでなく，分かち書き処理を自動的に行うことによるあいまいさも生ずる．このため，入力文に対して膨大な数の解釈が得られる場合が頻繁に発生する．

こうした問題を解決するために種々の方法が研究されてきた．これらを大別すると，縦型解析方式[2)-6)]と横型解析方式[7)]の二つに分類することができる．前者は，最長一致法に代表されるように入力文の先頭から逐次的に変換処理を行う方式で，処理が簡単になる半面，変換精度が十分でないという問題があった．後者は，文節数最小法に代表されるように総当り的に文全体を評価し，より確からしいものに変換する方式で，処理が多少複雑になる半面，変換精度の面からは有利な特徴を持っていた．しかしながら，変換精度の高いとされている文節数最小法においても文節内に閉じた形態素情報を利用して解析しているため，文解釈候補を十分には絞り込めていないという問題があった．

さらに文解釈候補を絞り込むため，栗原ら[8)]は連語の情報を用い同音語によるあいまいさを低減する方法を提案し，中野[9)]は対象とする世界を限定した中でその効果を確かめている．また牧野ら[10)]は用法辞書，関連語情報を用いて，大島ら[11)]は格文法を用いて同音語を選択する方法を試みている．いずれも同音語の選択に主眼を置いたもので，分かち書きの異なる文解釈候補の選択までには関与していなかった．そこで筆者らは，大規模な連語情報を用いて分かち書きの異なりおよび同音語が原因で複数個存在する文解釈候補を絞り込む方法を検討し，その効果を確かめたので報告する．

2. 連語解析

従来のかな漢字変換処理では，品詞，使用頻度等，単語単体の属性情報を使用する場合がほとんどであった．しかし，処理の高精度化を図るためには，単語単体での属性ばかりでなく，解析対象範囲を拡大して単語間の関係を用いることが必要であると考えられる．この情報の一つに連語情報があげられる．

ここでいう連語とは，通常にいう連語*の定義のうち，特に2個以上の自立語を含むものをいい，他の単語による言い換えが不可能な慣用句に比較してより自由な構成形態をとることができる．例えば「美しい

† Translation of Non-segmented Kana Sentences to Kanji-Kana Sentences Using Collocation Information by SHIGERU HONMA, MASAKI YAMASHINA and FUMIHIKO OBASHI (NTT Electrical Communications Laboratories).

†† NTT 複合通信研究所

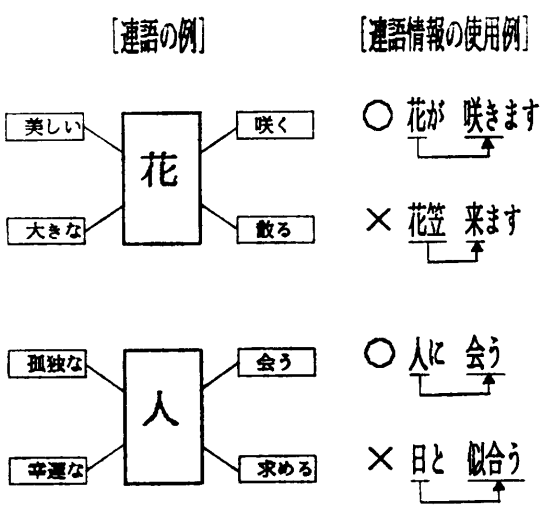


図 1 連語を用いた解析
Fig. 1 An example of sentence analysis.

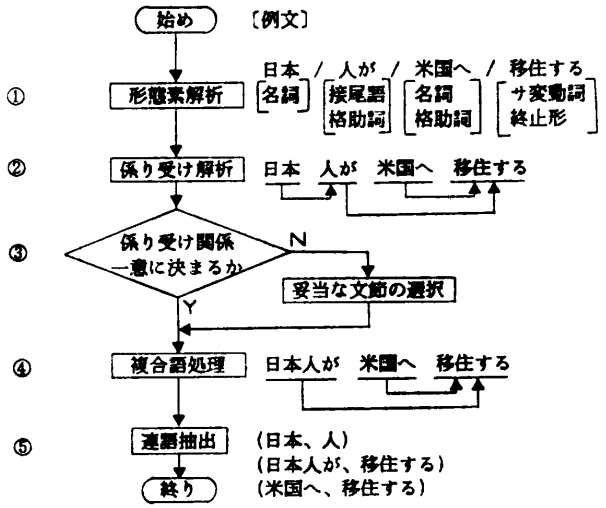


図 2 連語抽出の手順
Fig. 2 Flowchart of collocation extraction from Japanese sentences.

花」,「花が咲く」という表現において「美しい」,「咲く」の代りにそれぞれ,「奇麗な, 大きな, 赤い」,「開く, 散る」等を置くことができる.

このような連語情報をあらかじめ用意しておくことによって, 文に複数の解釈が生じた場合に, 連語情報を用いることによって正しい候補の選択が行える. 例えば図1に示すように「はながさきます」,「ひとにあう」にはそれぞれ2通りの解釈が得られるが, 連語情報を参照することにより,「花が咲きます」,「人に会う」の方を正しく選択することができる. このように, 分かち書きの結果にあいまいさがある場合に文解釈候補を絞り込むことができる.

3. 連語辞書

連語解析を行うためには, 連語を構成する単語の組を収録した辞書(連語辞書)を用意する必要がある. 自立語辞書の収録件数を5万語と仮定すると, 単純計算上では5万の自乗にあたる25億組の単語の組についての連語関係を検査する必要がある. このような数の連語情報を人手で検査しようとすると膨大な工数が必要となり, 実際上不可能である. そこで, ここでは実用的な連語辞書の構築を目指し, 図2に示すように, 日本語文章から連語情報を自動的に抽出することとした[12]. すなわち, 係り受け関係にある単語は緊密な共出現関係を持つことに着目し, 日本語文から係り受け解析により連語情報を抽出した. ただし, 現状では通常の日本語文章から高精度に連語を抽出することは, 構文解析法として残された問題も多く, 困難であるため, 単純な文表現で各単語の典型的な使われ方が示されている国語辞典等[13),14)]の用例に着目し, これらの用例を投入した. 以下に自動抽出の手法を述べる.

手順1:日本語文章を後述する自立語辞書, 接辞辞書, 付属語辞書, 文法辞書を用いて最長一致法により形態素解析し, 各単語の品詞, 活用形を認定する. 特に助詞については係り受け解析の重要な要素となるため, 8種類(格助詞, 接続助詞, 副助詞, 係助詞, 終助詞, 間投助詞, 並立助詞, 準体助詞)に細分化している. 例えば, 図2の解析において,「が」は格助詞, 接続助詞, 終助詞の可能性があるため, 前接する語の品詞を検査し, ここでは前接語が名詞であることから格助詞と認定する.

手順2:文節の形態的な性質と係り受け関係は非交差であるという性質に着目して係り受け解析を行い, 係り受け関係を持ち得る文節を認定する. ここでは, 文節の係り機能は文節末の助詞ある

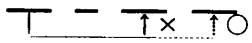
---

* 『2つ以上の単語が結合して1つのまとまった意味を表すが, その結合のしかたが, 1語となるには弱すぎ, また, 文をなすほど大きくはないもの. たとえば,「庭の桜がきれいに咲いた」という文において,「庭の桜が」「きれいに咲いた」とか,「庭の」「桜が」「庭の桜」「咲いた」などはみな連語である. 連語のうち,「咲いた」「行かない」「犬だ」のように, 助動詞が他の語についたものを活用連語といい,「をして」「について」など, 助詞が他の語について助詞相当の語となったものを助詞相当連語という. 連語は, 単語と文の中間的単位であり, 文節や連文節, ときには普通にいう句に相当するものにも当てはめて使う.』松村編: 日本文法大辞典, 明治書院, 東京 (1983)

いは活用語（動詞，形容詞，形容動詞，助動詞）の活用形が定め，受けの機能は文節の自立語の品詞および活用形が定めるものとし，それらの間の係り受け関係の適否を記述したマトリクス表を参照して解析する．例えば，図2の例では，「が」（格助詞）に対して受けの資格を持つのは用言（動詞，形容詞，形容動詞）であることから，「移住する」を抽出する．

手順3：解析の結果，係り受け関係にあいまいさがある場合，文節間の距離関係に着目して隣接文節で係り受け関係を持つものを採用する．例えば，隣接文節以外で係り受け関係にあいまいさが生じた場合，単純に近接する文節間の関係を優先すると次のような誤りが生じる．

《例》 二人は親の許した間柄だ



なお，次の文型では単純に隣接文節を優先することによる誤りが多いため，この文型での連語抽出は行わないこととする．

［文型］ 用語（連体形）＋名詞＋「の」＋名詞

《例》 赤い車の窓　　赤い秋の花

手順4：名詞連続，あるいは，名詞と接辞の連続を複合語と認定する．

手順5：係り受け関係を持つと認定された文節から，

係り単語			助詞	受け単語		
見出し	意味分類	意味分類番号		見出し	意味分類	意味分類番号
藍	染料	917	で	染める	色付	256
哀愁	悲喜	493	を	帯びる	包含	268
愛情	愛憎	481	を	持つ	存廃	285
合図	合図	334	の	口笛	音楽	870
生地	実質	130	を	染める	色付	256
花	花	056	が	咲く	花	056
人	自他	505	に	合う	出会	781
…	…	…	…	…	…	…

図3 連語情報
Fig. 3 Samples of collocations extracted from Japanese sentences.

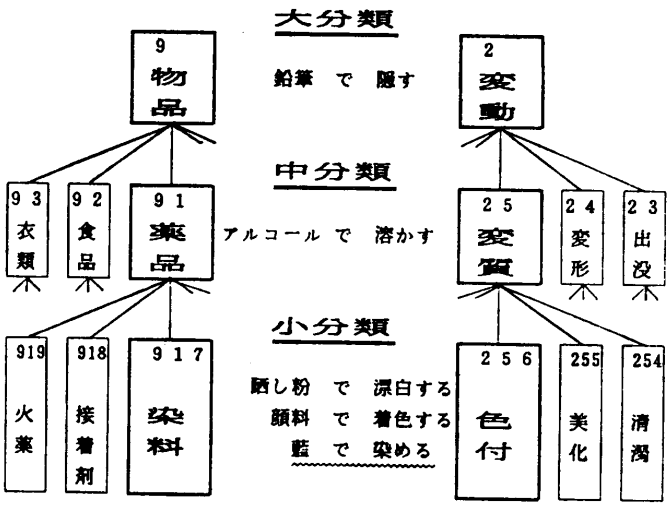
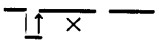


図4 連語の拡張解釈
Fig. 4 An example of sentence analysis using collocation and semantic information.

連語情報を抽出する．ただし，係り文節が名詞連続の複合語の場合，末尾の単語が係りの機能を持つと考える．受け文節が複合語の場合，例えば「情報の／処理／技術」，「新薬の／治療／効果」では，複合語を構成するどの単語が受け機能を持つかを単純に認定することは困難なため，複合語形で連語情報を抽出する．

以上の処理によって得られた連語情報を図3に示す．ここでは，用例約30万件から，98.7%の正解率で約30万5千件の連語を自動抽出した．誤解析例を以下に示す．

《例》 悪事千里を走る



4. 連語の拡張解釈

連語情報はそのまま辞書化して連語処理に用いることも可能であるが，見出しが一致しなければ連語として抽出できない．このため，連語解析の効果をより高めるため，係り単語あるいは受け単語を，意味の類似した単語と置き換えても連語関係が保たれる場合が多いという性質を利用する．すなわち，意味の類似した単語を1カテゴリとして扱うこととし，単語をそれが属する意味カテゴリに付けられている意味分類番号で代表させて表現した．

意味分類，意味分類番号の具体例を図4に示す．意味分類番号は3桁の数字で上位の桁から，大分類，中

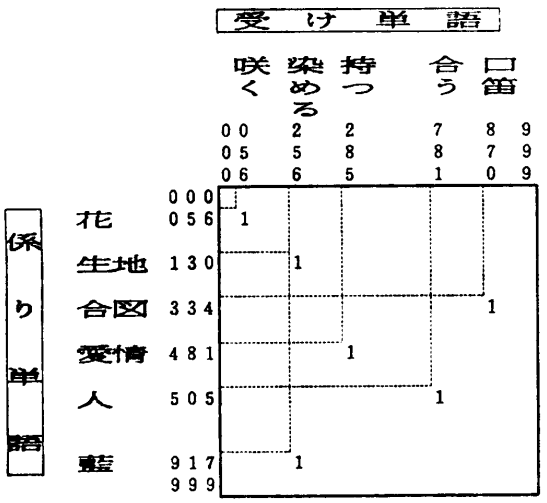


図 5 連語辞書の概念図
Fig. 5 Internal expression of collocation dictionary.

分類，小分類に相当する．したがって，カテゴリ数は1,000である．これにより，前述の連語情報は図5のように1,000×1,000のマトリクス上の'1'，'0'で表現できる．

以上のことから，例えば「藍で染める」がソース・データにあれば，染料（917）と色付（256）の間の連語情報が登録される．すなわち「顔料で着色する」，「晒し粉で漂白する」等は，連語として登録されていなくても染料と色付の関係であるために連語として抽出可能となる．このように，ソース・データを拡張解釈することにより，実効上は膨大な連語が登録されているのと同等の効果が得られる．さらに，意味分類を中分類以上のレベルで見ることにより，拡張解釈できる範囲は広くなる．

しかし，拡張解釈を行うと一つの意味分類が表現する範囲が広くなるため，連語としてふさわしくないものまで連語として認定する場合が発生する．特に，大分類の場合は，ほとんどのカテゴリ間に連語が成立しているので連語解析を行うための情報源とならない．このため，実験により意味分類番号を3桁（小分類）用いた場合，上位2桁（中分類）用いた場合についてその効果を明らかにすることとした．

5. 変換アルゴリズム

変換処理は総当りの辞書検索を基本としている．しかし，それだけでは数多くの文解釈候補が得られるので，文節数，単語数，連語関係による評価，ならびに，いくつかの例外処理を行って文解釈候補の絞り込みを行っている．変換処理の手順を図6の流れに従って以下に述べる．なお，実験に用いた辞書は表1のとおりである．

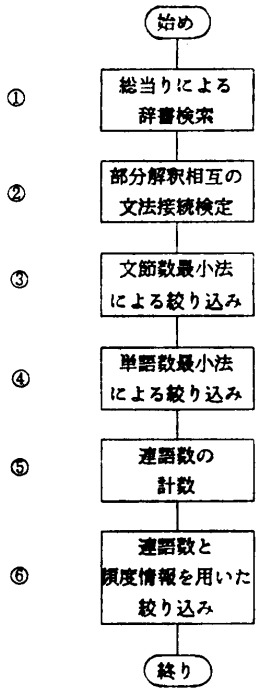


図 6 処理の流れ
Fig. 6 Flowchart of translation of Non-segmented Kana sentences to Kanji-Kana sentences.

手順1：入力文字列に対して左から右へ順次，自立語，付属語，接辞の辞書検索を総当りで行う．

手順2：部分解釈（自立語，付属語，接辞）の組合せの中から文法的に成立するものを選択する．

手順3：文節数最小の候補を選択する．

手順4：単語数最小の候補を選択する．ただし，「自立語一付属語一自立語」を，「自立語一自立語」

表 1 変換辞書
Table 1 Contents of dictionaries.

辞書の種類	内容	容量
自立語辞書	かな見出し，漢字表記，品詞，使用頻度，意味分類番号	約6万5千語
付属語辞書	助詞，助動詞，形式名詞，補助用言など	422語
接辞辞書	接頭語，接尾語，数詞	1,085語
文法辞書	自立語一付属語，付属語一付属語の接続規則を表形式で表現	286×239のマトリクス
連語辞書	3章で述べた辞書内容を持ち，自立語辞書の意味分類番号によって参照する．	1,000×1,000のマトリクス

という複合語に誤解釈してしまうのを防ぐため，複合語を構成する単語の使用頻度が低いものは採用しない．

《例》 自立語—付属語—自立語　　関係の中で
　　自立語—自立語　　関係野中で

また，「自立語—付属語」，「自立語—活用語尾」のいずれにも解釈できるものは，活用語尾も計数に加える．

《例》 自立語—付属語　　課題だ
　　自立語—活用語尾　　過大だ

手順5：連語関係の有無を調べ，各候補文ごとに連語数を計数する．ただし，この段階での文解釈候補数が平均で数十，多いもので数千に及び，処理量が膨大となるため，連語の大半を占める隣接間の連語関係に限定して検査を行う．

《例》 高速で走る電車の窓から

手順6：連語抽出件数が最大の候補を最尤候補として出力する．ただし，連語関係の希薄な単語の組合せを連語として採用することによって生じる誤りを防ぐため，極端に頻度の低い単語が存在している場合は，連語数最大より一つ連語数の少ない候補も調べる．

《例》 文章の校正 ［校正：低頻度］

文章の構成 ［構成：高頻度］

図7に連語解析の具体例を示す．三つの文解釈候補

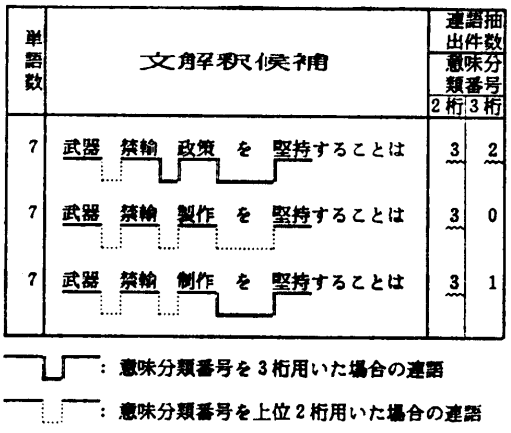


図7 連語解析の具体例
Fig. 7 An example of sentence analysis using collocation information.

表2 実験結果
Table 2 Average number of candidates and error rate.

処理段階		平均文解釈候補数	誤り率(%)(文節単位)
文節数最小		62.6	1.7
単語数最小		16.6	0.8
連語検出数最大(意味分類番号)	(上位2桁)	8.7	1.8
	(3桁)	6.4	1.6

は，文節数最小で単語数最小の条件で選択されたものである．意味分類番号を上位2桁用いた場合は，三つの候補はおのおの連語が3件抽出され，すべて最尤候補として残る．意味分類番号を3桁用いた場合は，上段の候補が連語抽出件数2件で，三つの候補の中で最大であるため，最尤候補となる．

6. 評価実験

入力データは新聞社説10編から抽出した記号等を含まない句読点で区切られた単位の文字列386件で，文節数は1,411（文字数：約5,500字）である．1文あたりの文字列平均長は14文字で，平均3.7文節からなっている．

表2に各処理段階での平均文解釈数と各処理段階で付加される誤り率を示す．総当り的に処理した候補文の中から文節数最小の候補のみを選択すると，平均文解釈候補はまだ62.6通り存在しており，さらに候補数を絞り込む必要があることがわかる．次に，単語数最小の候補を選択すると，平均文解釈候補数は，16.6となり，さらに連語情報を用いた絞り込みでは，平均文解釈候補数は，

8.7 （意味分類番号上位2桁）
6.4 （意味分類番号　　3桁）

まで減少する．この結果，文節数最小法に比較して約1/7～1/10まで候補を絞り込むことができ，本方式の効果を確かめることができた．

一方，候補を絞り込む際の誤り率（正解を漏らす率）は文節数最小の処理段階で全体の1.7%であり，表3に示すように接辞を文節数1としたことによる誤りが大半をしめていることから，出現頻度の高い複合語を自立語辞書に登録することで最先度をあげる方法が有効と考えられる．単語数によって候補を絞り込む際の誤り率は，全体の0.8%であり，「自立語—付属語」を「自立語」と解釈したこと等により生じてい

表 3 誤解析とその原因
Table 3 Classification of errors.

原因	誤解析例	入力文例	件数
文節数最小	教科書の共通か	教科書の共通化	15
単語数最小	体制の元手	体制のもとで	11
未登録連語	例外による不作	冷害による不作	23
未登録語	念書以来	年初以来	26

る．これらについては構文解析を用いてより詳細に文を解釈する必要がある．連語情報を用いた絞り込みでの誤り率は，全体の

1.8% （意味分類番号上位2桁）

1.6% （意味分類番号 3桁）

であり，誤り率の増加を低く抑えた上で平均文解釈候補数を大幅に減少できていることがわかる．この場合の誤り率は，連語情報が辞書になかったために発生している場合が多く，さらに辞書を拡充することで精度向上が期待できる．このほか，未登録語による誤りも全体の1.8%生じている．

以上の結果，誤り率の増加を低く押えた上で平均文解釈候補数を大幅に減少でき，連語解析法がかな漢字変換に十分有効なことがわかった．

7. むすび

本論文では，べた書きかな文字列の解析において，総当り的手法によって抽出した文解釈候補を，文節数最小法，単語数最小数によって絞り込み，さらに，連語情報と使用頻度情報を用いた連語解析法によって，最尤候補文を得る方法について述べた．

その結果，誤り率の増加を低く押えた上で，平均文解釈候補数を，文節数最小法の

約 1/7 （意味分類番号上位2桁）

約 1/10 （意味分類番号 3桁）

に絞り込むことができ，連語解析法の有効性を明らかにした．

今後は，以下の検討を行う予定である．

（1） 現在，隣接した文節のみで行っている連語の検査を，離れた文節においても可能とするための構文解析法の検討．

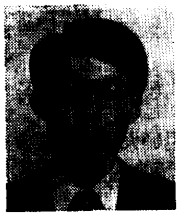
（2） 未登録語を抽出し，その影響を他文節に及ぼさない処理方式の検討．

謝辞 日頃，御指導いただく山崎宅内機器研究部長，小森主席研究員に深謝する．また，有益な御助言をいただいた NTT 基礎研究所の島津主幹研究員に感謝する．

参考文献

1) 本間，山階，小橋：連語解析を用いたべた書きかな漢字変換，情報処理学会日本語文書処理研究会資料，21-2，pp. 1-8 (1985)．
2) 内田，杉山：自由入力形式のカナ漢字変換，情報処理学会自然言語処理研究会資料，27-3，pp. 1-8 (1981)．
3) 平塚，八田，津田：べた書きかな漢字変換の一方式について，情報処理学会日本文入力方式研究会資料，13-4，pp. 1-6 (1984)．
4) 河田，武田，斎藤，中里，楠元：ベター方式かな漢字変換入力システムの試作，情報処理学会日本語文書処理研究会資料，21-1，pp. 1-9 (1985)．
5) 藤田，沼田，山内：かなべた文の逐次単語分割アルゴリズムの一方式，情報処理学会日本語文書処理研究会資料，21-3，pp. 1-7 (1985)．
6) 牧野，木澤：べた書き文の分かち書きと仮名漢字変換一二文節最長一致法による分かち書き一，情報処理学会論文誌，Vol. 20，No. 4，pp. 337-345 (1979)．
7) 吉村，日高，吉田：文節数最小法を用いたべた書き日本語文の形態素解析，情報処理学会論文誌，Vol. 24，No. 1，pp. 40-46 (1983)．
8) 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，Vol. 39，No. 4，pp. 659-664 (1967)．
9) 中野 洋：分野別辞書と同音語，特定研究〈言語の標準化〉林班研究発表会資料，国立国語研究所 (1983)．
10) 牧野，木澤：べた書き文の仮名漢字変換システムとその同音語処理，情報処理学会論文誌，Vol. 22，No. 1，pp. 59-67 (1981)．
11) 大島，阿部，湯浦，武市：格文法による仮名漢字変換の多義解消，情報処理学会論文誌，Vol. 27，No. 7，pp. 679-687 (1986)．
12) 山階，小橋：係り受け解析を用いた連語情報自動抽出法の検討，第 30 回情報処理学会全国大会論文集，5 G-9，pp. 1693-1694 (1985)．
13) 大野，浜西：類語新辞典，角川書店，東京 (1981)．
14) 増田 綱：新和英大辞典，研究社，東京 (1974)．

(昭和61年1月29日受付)
(昭和61年8月27日採録)

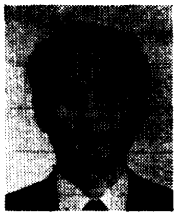


本間　茂（正会員）

昭和33年生．昭和56年東北大学工学部電気工学科卒業．昭和58年同大学院工学研究科情報工学専攻修士課程修了．同年日本電信電話公社入社．以来，べた書きかな漢字変換，音声理解等の日本語情報処理の研究に従事．現在，NTT 複合通信研究所宅内機器研究部音声入出力方式研究室研究主任．電子通信学会，日本音響学会各会員．

山階　正樹（正会員）

昭和26年生．昭和50年山梨大学工学部精密工学科卒業．昭和52年同大学院工学研究科精密工学専攻修士課程修了．同年日本電信電話公社入社．以来，かな漢字変換等の日本語情報処理の研究実用化に従事．現在，NTT 複合通信研究所宅内機器研究部知的通信宅内装置研究室主任研究員．電子通信学会会員．



小橋　史彦（正会員）

昭和22年生．昭和45年徳島大学工学部電気工学科卒業．同年日本電信電話公社入社．以来，かな漢字変換，キーボード配列，オンライン手書き入力，音声入力等の日本語情報処理の研究に従事．現在，NTT 複合通信研究所主幹研究員．電子通信学会，人間工学会各会員．