連語データを利用した仮名漢字変換

小山泰男[†] 安武満佐子[††]
吉村賢治[††] 首藤公昭[††]

仮名漢字変換における変換精度を向上させる現実的な方法として，慣用句などの単語が固定的に共起する表現データを活用して分かち書きと同音語の曖昧性解消処理を行うことが考えられ，商用システムなどでも試みられている．しかし，これらの表現の種別や量と変換精度向上への貢献度との関係についてはこれまで明らかにされていない．筆者らはこれらの表現を連語データとして比較的大規模に収集・整理し，段階的に導入して仮名漢字変換精度を判定する実験を行った．実験は約23,000個の入力仮名文字列とその漢字変換正解および分かち書き正解からなる評価用データを作成して行った．その結果，約72,000個の連語を単語の共起データとして用いる場合，共起情報をまったく用いない仮名漢字変換システムに比べて漢字第1候補正解率で8.9%，分かち書き正解率で4.9%向上すること，市販の平均的なPC用ワープロソフトの漢字第1候補正解率を7.0%向上させることなどが明らかになった．

Kana-to-Kanji Conversion Systems Based on Collocation Data

YASUO KOYAMA,[†] MASAKO YASUTAKE,[††] KENJI YOSHIMURA[††]
and KOSHO SHUDO[††]

Word processors or computers used in Japan employ Japanese input method through keyboard strokes combined with *Kana* (phonetic) character to *Kanji* (ideographic, Chinese) character conversion technology. The key factor of *Kana*-to-*Kanji* conversion technology is how to raise the accuracy of the conversion through the homophone processing since so many homophonic *Kanjis* exist. In this paper, we report the results of our *Kana*-to-*Kanji* conversion experiments for approximately 23,000 input *Kana* strings, which embody the homophone processing based on approximately 72,000 collocation data. It is shown that the collocation data yields 8.9% higher fraction of the conversion accuracy compared with the prototype system which has no word concurrence data, 7.0% raise of fraction of the accuracy of a commercial word processor software, and so forth.

1. はじめに

日本語ワードプロセッサやコンピュータの日本語入力方式として，手書き文字認識，音声認識などの技術を必要としないキーボードによる仮名あるいはローマ字入力が広く用いられている．この方式では，入力されたべた書き入力文字列は形態素解析によって文節分かち書きされ，仮名漢字混じり文に変換されるが，このとき，文節内の同音異字語の候補の中から正しい候補を第1候補として選択する同音語処理が重要な課題である．従来，同音語処理として最終使用語優先方式や最高頻度語優先方式が広く用いられてきたが，これらの方式では語と語の間の意味的な整合性が考慮されていないため，誤った変換を行いがちである．このため，近年，用言の格フレームを用いて単文内の単語の共起の整合性をチェックする方法[1]や隣接単語間の一般的な共起を表現レベルで与えておき，分類語彙表の意味コードを用いてデータの拡張を行う方法[2]，ニューラルネットによって語の共起を記述し利用する方法[3]，特定の話題ごとに共起辞書を作り，キーワードから話題が特定されるとその共起辞書の語を優先する方法[4]，単文内の名詞と動詞の共起を頻度付きであらかじめ調べておき，この情報を推移的に用いて共起の尤度を求める方法[5]など，種々の方法が提案されている．しかし，以上の中で単語の共起を単語の上位概念や意味素性などで制約する方法では語彙レベルの共起を正しく規定できない場合が多い．また，ニューラルネットや分野別共起辞書を用いる方法では数十万語に及ぶ膨大

† 福岡大学大学院工学研究科情報・制御システム工学専攻
Graduate School of Engineering, Fukuoka University
†† 福岡大学工学部電子情報工学科
Faculty of Engineering, Fukuoka University

なデータの取扱いに関する問題が避けられない．さらに，名詞と動詞の共起を調べておく方法においても，網羅性をどう確保するかについては問題が残されている．これらに対して，単語の共起を直接的に連語データとして収集・整理しておき，仮名漢字変換に利用するアプローチが考えられ，商用のシステムなどでも試みられてはいるが，データの種類や量と変換精度との関係はこれまで十分には明らかにされていない．連語データの有用性を定量的に評価することが本論文の主な目的である．そのために筆者らは，約72,000件の連語データを作成し，これを仮名漢字変換に適用して精度の評価を行った．以下，2章で収集・利用した連語データについて述べ，3章で実験システムの概要，4章で評価実験の方法と結果を示し，考察を行う．

2. 連語データ

複数単語からなり，まとまった意味，機能を持つ表現（句）であって，要素となっている単語の通常の意味から表現全体の意味を導くルールを上位概念や意味素性によって規定することが難しいもの，あるいは，単語の組合せが固定的であり，表現中に使われている単語や句が残りの部分の生起を強く示唆する場合，その表現全体をここでは連語と呼ぶ．たとえば，「腹を立てる」という表現は「腹」，「立てる」の単語の通常の意味からは全体の意味が導出しにくい．また，「ぐっすり眠る」，「御多分に洩れず」，「ひとり言を言う」などでは，「ぐっすり」，「御多分に」，「ひとり言を」から「眠る」，「洩れず」，「言う」が後続することがかなり強く示唆されるといってよい．このような表現を大規模コーパスから自動的に抽出する試みがいろいろ行われているが[11)~14)]，筆者らはこれらを人手によって抽出した．統計量に基づく方法ではコーパスにおける表現データの希薄性から収集の十分性の面で問題が避けられないこと，また，自動抽出された表現に対して人間の判断による必要性の判断が不可欠であることなどから，自動抽出のアプローチはとっていない．収集作業にともなう恣意性についてはある程度はやむをえないと考えている．データの収集は新聞，雑誌，小，中学校の教科書，各種辞典類について行った．収集した連語データは文法機能によって付属語性連語と自立語性連語に大別される．

以下の表現例中の記号「/」は表現を通常の文節単位に分かち書きする際の境界を表す記号であり，連語辞書の見出しにも同じ記号が挿入されている．以下で，括弧内に示す数値は特に断らない限り該当する表現の個数を表す．

表1 関係表現（961）
Table 1 Relational expressions.

に/ついて，に/よって，に/おける，に/基づいて， の/ように，でさえも，などと/いった，だけで/なく， にも/かかわらず，とは/言え，に/対して， ばかりで/なく，の/通りに

表2 助述表現（1,438）
Table 2 Auxiliary predicative expressions.

なければ/ならない，かも/しれない，て/いる， たほうが/よい，べきで/ある，に/違いない， おそれが/ある，すら/ない，とは/限らない， ずには/いられない，には/及ばない

表3 強連結の自立語性連語（54,925）
Table 3 Uninterruptible conceptual collocations.

種類	個数	例
名詞性	22,854	幾つか，赤の/他人，目の/毒
サ変名詞性	1,067	大掴み，貰い/泣き，ラッパ/飲み
五段動詞性	9,700	相異なる，汗水/垂らす
一段動詞性	4,789	大人びる，一息/入れる，垢/抜ける
形容詞性	2,076	青臭い，えげつない，途方も/ない
形容動詞性	738	誇らしげ，御機嫌/斜め
副詞性	2,725	案の/定，いつに/なく
連体詞性	353	悪しき，あじな
四文字熟語	2,194	我田引水，不惜身命
～する	284	相手に/する，公に/する
その他	8,145	如何せん，年端も/いかぬ

2.1 付属語性連語（2,399）

付属語性連語は概念間の関係を表す助詞と類似した働きをする関係表現と，話者の判断，態度やテンス，アスペクトなどの情報を提供する助動詞に類似した働きをする助述表現に大別される[7)]．これらの表現の例を**表1**，**表2**に示す．

2.2 自立語性連語（69,968）

自立語性の連語は2種に大別する．ほとんど例外なく連結した形で用いられる強連結連語（uninterruptible collocations）と，ときに分割して用いられることのある緩連結連語（interruptible collocations）である．

2.2.1 強連結の自立語性連語（54,925）

強連結の自立語性連語の種類，個数および例を**表3**に示す．

2.2.2 緩連結の自立語性連語（15,043）

緩連結の自立語性連語の種類，個数および例を**表4**に示す．これらはほとんど例外なく係り受けの関係にある2文節からなる句である．

表4 緩連結の自立語性連語（15,043）
Table 4 Interruptible conceptual collocations.

種類	個数	例
名詞性	505	悪業の/報い，環境の/汚染
サ変名詞性	251	額に/汗，お手数を/おかけ
五段動詞性	7,939	心が/沈む，気を/吐く
一段動詞性	4,202	座を/占める，体を/あける
形容詞性	1,641	態度が/でかい，気が/重い
形容動詞性	63	懐が/暖か，愛情が/細やか
副詞性	14	目を/輝かせて，先を/争って
その他	428	間尺に/合わぬ，心胆/寒からしむ

3. 仮名漢字変換実験システム

3.1 拡張文節

本論文で述べる仮名漢字変換実験システムでは基礎となる言語構造モデルとして拡張文節[6]，優先度決定のためのヒューリスティクスとしてコスト最小法[8]を採用する．

拡張文節は連語を単語と見なして文節の概念を拡張したものであり，その概形は次のとおりである．

< 拡張文節 >::=
< 接頭語 >*< 自立語 | 自立語性強連結連語 >
< 接尾語 >*< 付属語 | 付属語性連語 >*

本実験システムは入力仮名文字列を拡張文節単位に分かち書きする．実際の文節構造規定は上記を大幅に精密化したものであるが本論文では詳細を省く．

以後，拡張文節を単に文節と呼ぶ．また，連語を含まない拡張文節を特に小文節と呼ぶことがある．

3.2 仮名漢字変換過程

本実験システムのアルゴリズムは DP の手法[15]を用いるもので，概略，以下のとおりである．

ステップ 1 辞書探索位置を表すポインタを入力仮名列先頭に初期化する．

ステップ 2 単語，連語（緩連結連語を除く）辞書の探索と文節内の連接チェックにより，ポインタ位置から始まる文節候補を求める．

ステップ 3 各文節候補にその文節で終わる入力先頭からの文節列の部分最小コスト値を与え，その文節列の最終リンクを記録する．

ステップ 4 ポインタ位置を更新する．

ステップ 5 入力末尾までステップ 2〜4 を繰り返す．

ステップ 6 入力末尾で終わる文節候補の部分最小コスト値を比較し，その最小値を持つ文節候補から記録したリンクを入力列先頭まで逆にたどって最小コスト文節列を求める．ただし，係り受けに関与するリンクがあればそれを優先する．

ステップ 7 求めた文節列に対して辞書の内容から漢字変換を行って出力する．

処理結果の評価は第1候補のみについて行うが，最小コスト値を持つ連鎖が複数個存在する場合は，文末側から文節の長さ優先で候補を絞る．また，長さも同じ場合は一般的な使用頻度順となっている辞書記載の順に従う．緩連結の自立語性連語は文節間の係り受けを表層レベルで規定したデータと見なし，3.3 節で述べるコスト計算の際に評価に反映させる．

3.3 コスト計算

コストは大きく3種を考え，それらを合算して優先度設定に用いる．

まず，分かち書きの各文節に対して文節コスト (+2) を割り当て，これらの和をその分かち書きの基本的なコストと考える．このことから連語を単語とする解析に優先度が与えられる．ただし，接辞よりも単語を優先する考え方から，接辞を含む解析結果にはペナルティーとしてのコスト (+1) を加算するなどの調整を行う．

次に，文節の連続の仕方の尤度を考慮したコストを考える．これを文節間コストと呼ぶ．たとえば，用言の連体形に続く名詞や，連体詞に続く名詞，副詞に続く用言の場合，これらの文節境界にボーナスとして負のコスト (−1) を与え，基本のコストに加算する．

さらに，文節間の係り受けチェックを行い，負のコストを受けの文節に対して加算する．これを係り受けコストと呼ぶ．緩連結の連語が検出された場合の係り受けコストは −3 である．3.7 節で述べるシステム C，D，および E には，その他，名詞，動詞の比較的簡単な意味分類を用いた2文節間の係り受けに関する規定が設けられており☆，これらの規定に適う場合の係り受けコストは −1 あるいは −2 とする．以上のコスト値は試行錯誤により経験的に決めている．コストの計算過程は次のとおりである．

ステップ 3-1 文節候補と各左隣接文節との間にリンクを張り，各左隣接文節の部分最小コスト値に当該文節の文節コスト，+2 を加算した値を各リンクの累積コスト値とする．

ステップ 3-2 必要に応じて文節間コスト値を加算し

☆ 深さ4の樹状階層構造のノード約150個を意味カテゴリーとして用いた規定となっている．

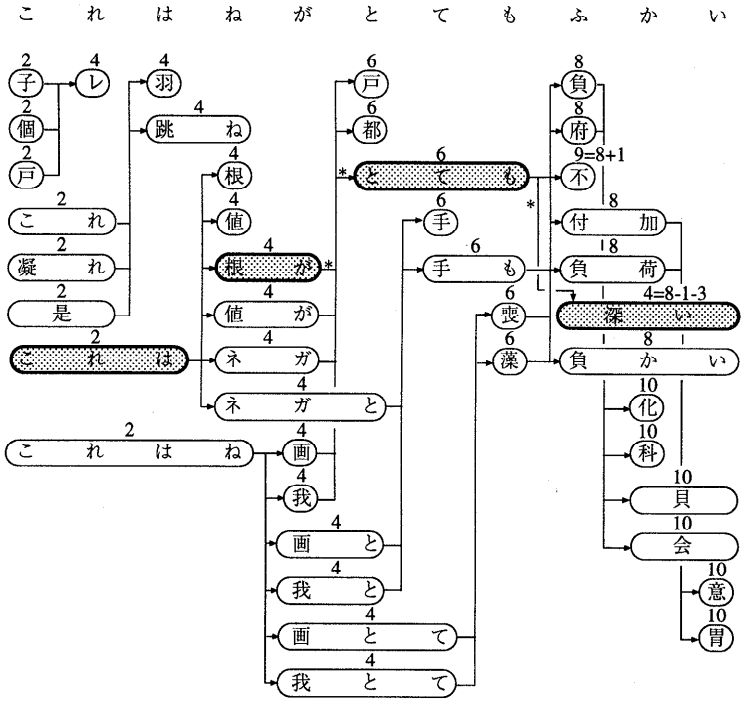


図1 入力仮名列「これはねがとてもふかい」に対して「これは根がとても深い」と出力される例
Fig. 1 An example of processing; output "これは根がとても深い" for input "これはねがとてもふかい".

てリンクの累積コスト値を更新する.

ステップ 3-3 当該文節と文頭側既出文節との係り受けチェックを行い☆, 記録されたリンクでつながった係り文節があれば, そこに至るパスの各リンクにマークを付け, マークの付いた最終リンクの累積コストに係り受けコストを加算してリンクの累積コスト値を更新する.

ステップ 3-4 左隣接文節とのリンクの累積コスト値が最小であるリンクを記録して, その値を当該文節の部分最小コスト値とする.

ステップ 3-5 ポインタ位置から始まる文節候補がある限り, ステップ 3-1 から 3-4 を繰り返す.

図1 に「これはねがとてもふかい」という入力に対するコスト計算過程を示す. 図1では繁雑さを避けるため, 実際より候補文節を減らし, 上記ステップ 3-4 で記録されるリンクのみを束ねて表示している.「深い」という文節候補に対しては, まず, ステップ 3-1 で左隣接の文節候補「とても」,「手も」,「喪」,「藻」とのリンクに累積コスト 8＝6＋2 が与えられ, ステップ 3-2 において,「とても」とのリンクは副詞に続く形容詞に対する文節間コスト, −1 が加算されて7と更新される. 次に, ステップ 3-3 で緩連結連語辞書から「根が」が「深い」に係ることが分かり,「根が」—「とても」—「深い」の各リンクに＊マークが付され, 同時に「深い」と「とても」のリンクの累積コストには係り受けコスト, −3 が加算されて4と変わる. 次に, ステップ 3-4 で「深い」については「とても」とのリンクの累積コスト値4が最小であることから, このリンクを記録し, 文節「深い」の部分最小コスト値を4とする. 次いで, ステップ 6 で, 末尾文節の部分最小コストのうちの最小値4を持つ文節「深い」からリンクに付けられた＊マークを優先しながら記録されたリンクをたどり, 先頭に至るパス「これは」—「根が」—「とても」—「深い」を得る.

3.4 表記のゆれ

日本語には**表5**に示すような表記上の多様性が認められる. 本システムではこれらの自由度を許容した現実的な評価を行うため, 単語, 連語辞書には代表的な表記（代表表記）とその異表記（ゆれ表記）を収録している. 漢字変換に際しては代表表記を優先させるが, 次候補としてはゆれ表記を提示することを想定している. 本実験システムにおいても正解と出力にゆれによ

☆ 緩連結連語データや 3.7 節で述べる係り受けの規定に適う2文節間の係り受けを検出する処理であり, 非交差, 係りの一意性などの条件を用いた文解析は行わない.

表5 表記のゆれ
Table 5 Notational variants of Japanese words.

種類	例
送り仮名の有無	売り上げ ～ 売上げ ～ 売上
長音の有無	メモリー ～ メモリ
拗音のゆれ	バッファ ～ バッファ
ヴァとバ	ヴァイオリン ～ バイオリン
ディとデ	ディジタル ～ デジタル
漢字と記号	電話 ～ TEL
平仮名と漢字	ください ～ 下さい
片仮名と漢字	ウナギ ～ 鰻
異漢字	浜 ～ 濱

表6 単語辞書の構成
Table 6 Composition of word dictionary.

品詞	個数	品詞	個数
名詞	93,776	助詞	300
サ変名詞	14,470	助動詞	323
五段動詞	11,262	形式名詞	170
一段動詞	5,676	補助用言	133
形容詞	3,362	接頭語	26
形容動詞	9,032	接尾語	262
副詞	6,006	冠数詞	23
連体詞	418	助数詞	608
感動詞	585	その他	8,957
接続詞	330	—	—

る表記のずれがあった場合，それを正解として評価する方法をとった．たとえば,「おこなう」に対して正解データが「行う」，辞書中の代表表記が「行なう」となっている場合でも，辞書中のゆれ表記情報から「行う」も許容できる出力と見なすことができる．

3.5 連語の分割した取扱い

本システムでは強連結の連語が一語扱いになるケースが多いが，これらを終始一語扱いにすると操作性の点で問題が生じる場合がある．たとえば，利用者は連語の「腹を立てる」ではなく「(バッターボックスに)原を立てる」を意図していたり,「立てる」を仮名書き「たてる」にしたいと思っている場合に一語扱いでは修復に手間がかかる．そこで，必要に応じて「はらを」と「たてる」を別々に扱えることが望まれる．そのため，連語データの見出しにおける小文節の境界を「/」で明示している．

3.6 基本システムA

以上に述べた処理の枠組みの中で，まず，連語データの効果を測定するための基準として，連語や係り受けなどの単語の共起に関する情報をまったく持たない仮名漢字変換システムを作成した．これをシステムAと呼ぶ．システムAの語彙的資源は約156,000個の通常の単語であり，その内訳は**表6**のとおりである．

表7 システムCの単語共起に関する情報
Table 7 Information about word concurrence, installed additionally in system C.

	種類	数量	例
I.	連語的表現	約3,000件	意志/決定, 家事/手伝い
II.	単語レベルの2文節係り受け情報	約57,000対	雨が/降る, 実績が/上がる, 愛嬌が/良い
III.	意味カテゴリーによる2文節係り受け情報	約10,000対	[人]が/行く, [植物]を/育てる, [食品]を/食べる, [場所]で/開催する

3.7 システムB，C，DおよびE

利用する単語共起のデータを変えてさらに4種の実験システムを作成した．システムAに付属語性連語(2,399)と自立語性連語(69,968)を追加したシステムをシステムBと呼ぶ．

次に，市販のPC用日本語ワープロソフトとしてWXG Ver.2.05☆をとり，そこで装備されている共起情報すべてをシステムAに移植してシステムCを作成した．移植した情報は**表7**のとおりである．システムCにおいては表7のIは強連結連語，IIとIIIは緩連結連語と同様に扱われる．

次に，システムCに付属語性連語を追加したシステムをシステムDとし，最後にシステムDに自立語性連語を加えたシステムをシステムEとした．ここで追加した自立語性連語とシステムCのデータとの重複は，強連結自立語性連語とIとの間で1件，緩連結自立語性連語とIIとの間で3,544件あった．これらは追加データから削除しており，DからEへの実質的な追加の連語データは66,423件である．なお，単語の出現頻度や出現順による影響を除去し，純粋に連語データによる変換精度を測定するため，これらを管理する学習機能は全実験システムで省かれている．

4. 実　　験

4.1 評価用データ

実験に先立ち評価用の入力および正解データを準備した．これは連語の収集に用いたテキストとは独立の新聞記事，雑誌記事，日本語文型辞典類，ビジネス文書などから人手によって抽出，設定したもので，次の例に示すような22,923個の3項組からなる．

例；{入力仮名文字列：にわにばらがさいている，
漢字変換正解：庭に//バラが//咲いて・いる，
文節分かち書き正解：にわに//ばらが//さいて・いる}

☆ エー・アイ・ソフト（株）製．

このデータ中の分かち書きは付属語性連語と強連結の自立語性連語を単語扱いにした文節を単位とする.「//」は明らかな文節境界,「・」は形式名詞,形式動詞などの取扱いの違いからくる,任意性のある境界を意味する.これにより,システムの分かち書き結果を評価する際,現実に即した柔軟性が確保される.上例では,「咲いて」と「いる」を分かち書きした出力も,これらを分かち書きしない出力もともに正しいと見なすことができる.

入力の分かち書き正解には評価用データ全体として平均4.7個の区切り記号「//」が入っている.「//」で区切られた1文節の平均仮名文字数は5.06である.入力データのうち,「//」が3個以下のものは10,964個である.

4.2 評価の方法

実験は,システムA～Eに評価用入力データを読み込んだ後,変換の精度を評価する機能を付与して行った.評価用入力データにおける入力仮名文字列を a,文節分かち書きの正解を b,漢字変換の正解を c とし,実験システムの文節分かち書き出力を b',漢字変換出力を c' とするとき,変換精度として以下の3種類について評価した.

(1) 完全一致;$b=b', c=c'$
(2) ゆれ許容一致;$b=b', c\simeq c'$
(3) 分かち書き一致;$b=b'$

$b=b'$ は分かち書きが一致したことを表すが,出力中の文節間境界を「//」,出力中で採用されている連語の中の小文節単位の境界を「/」と表すとき,各文字間位置における境界の一致不一致の判定は**表8**によって行う.すなわち,正解データと出力とのいずれか一方が「//」で,他方に切れ目の記号がない場合に境界の不一致と見なす.また,出力文字列中に1カ所でも境界の不一致があれば,その出力は分かち書きに失敗したものとする.$c\simeq c'$ は c' が c と一致しているか,不一致の部分が3.4節で述べた表記のゆれとして許容できることを表す.

4.3 実験結果および考察

実験結果の主要部を**表9**,**表10**に示す.表中の数値は4.2節で述べた場合**(1)**～**(3)**に該当する入力件数,括弧内は正解率である☆.表9は,22,923件の入力中4文節以下の比較的短い入力10,964個(入力あたりの平均文節数2.7)に限定した場合の実験結果,表10は全入力に対する結果である.

表8 分かち書き結果の評価
Table 8 Evaluation of segmentation results.

正解データ中の切れ目	出力中の切れ目	判定
なし	なし	一致
なし	//	不一致
なし	/	一致
//	なし	不一致
//	//	一致
//	/	一致
・	なし	一致
・	//	一致
・	/	一致

4.3.1 システムAとBの比較

システムAとBを比較すれば共起情報をまったく持たないシステムに対する連語データ全体の効果が分かる.表9によれば,(1)完全一致で9.4%,(2)ゆれ許容で11.1%,(3)分かち書きで4.0%正解率が向上していることが分かる.表10ではそれぞれ(1)6.3%,(2)8.9%,(3)4.9%の上昇である.(1),(2)の正解率の向上の幅は正解率の高い短い入力の場合が大きくなっているが,改善率☆☆を比較すると短い入力と入力全体とでは(1)で19.5%と18.3%,(2)で18.4%と18.5%と大差がない.しかし,(3)については4.4%と6.1%となって長い入力の方が改善率が高い.分かち書きの精度はシステムAでもかなり高く,入力が長くなって分かち書きが難しくなってくるほど,連語データの効果が現れている.

4.3.2 システムBとCの比較

システムBとCの実験結果を比較すると,入力全体として,(1)完全一致で1.9%,(2)ゆれ許容で2.6%,(3)分かち書きで4.1%,システムBの方がシステムCより正解率が高い.システムCは連語データをあまり使っておらず,主として表7に示す係り受け情報を利用したシステムであり,市販されているPC用日本語ワープロソフトの平均的能力を持つと推定される.したがって,共起情報として本論文の連語データだけを用いても,処理の枠組みの能力を3章で述べたものと同等と仮定すれば,市販ワープロを若干上回る程度の効果が期待できることが分かる.

4.3.3 システムCとDの比較

付属語性連語データの効果を見るため,システムCとDの結果を比較すると,入力全体として正解率が

☆ 正解率 = 正解件数/入力件数.

☆☆ システム β のシステム α に対する改善率 = (システム β の正解件数 − システム α の正解件数)/システム α の正解件数,とする.

表9 比較的短い入力（10,964）に対する実験結果
Table 9 Results of experiments for short inputs.

	システムA	B	C	D	E
(1) 完全一致	5,279 (48.1%)	6,308 (57.5%)	6,116 (55.8%)	6,228 (56.8%)	6,892 (62.9%)
(2) ゆれ許容一致	6,600 (60.2%)	7,816 (71.3%)	7,660 (69.9%)	7,850 (71.6%)	8,525 (77.8%)
(3) 分かち書き一致	9,925 (90.5%)	10,358 (94.5%)	9,995 (91.2%)	10,134 (92.4%)	10,417 (95.0%)

表10 入力全体（22,923）に対する実験結果
Table 10 Results of experiments for all inputs.

	システムA	B	C	D	E
(1) 完全一致	7,844 (34.2%)	9,277 (40.5%)	8,848 (38.6%)	9,148 (39.9%)	10,028 (43.7%)
(2) ゆれ許容一致	10,978 (47.9%)	13,013 (56.8%)	12,424 (54.2%)	13,031 (56.8%)	14,027 (61.2%)
(3) 分かち書き一致	18,424 (80.4%)	19,555 (85.3%)	18,620 (81.2%)	19,113 (83.4%)	19,682 (85.9%)

(1) 完全一致で1.3%，(2) ゆれ許容で2.6%，(3) 分かち書きで2.2%上昇している．改善率は (1) で3.4%，(2) で4.9%，(3) で2.6%である．

4.3.4 システムDとEの比較

自立語性連語データの効果を見るため，システムDとEの比較を行うと，入力全体として，(1) 完全一致で3.8%，(2) ゆれ許容で4.4%，(3) 分かち書きで2.5%の上昇が認められる．改善率は (1) で9.6%，(2) で7.6%，(3) で3.0%である．

4.3.5 システムCとEの比較

2章で示した全連語データを加えることにより，システムCの正解率をどれだけ改善できるかをシステムCとシステムEとで見ると，人力全体として，(1) 完全一致で5.1%，(2) で7.0%，(3) で4.7%となっている．改善率は，(1) 13.3%，(2) 12.9%，(3) 5.7%である．このように連語データが既存のシステムの正解率を向上させる効果を持つことが数量的に確認された．

4.3.6 入力の長さと変換精度との関係

入力の長さと変換精度との関係を細かく見るため，システムEについて，入力の長さ別に結果を集計した．**図2** にその結果を示す．グラフの横軸は入力中の文節の個数，縦軸は正解率である．ただし，横軸上の指示値2は1および2文節入力をまとめて表示している．

図2から入力が長くなるにつれ正解率は比較的なだらかに減少していくことが分かる．一般に，記号列をいくつかの部分記号列に分割する仕方は記号列の長さに関して指数関数的に増加する．たとえば，1文節の長さを最短で1仮名文字，最長で18仮名文字，平均で5仮名文字とすれば，言語データを使わない機械的

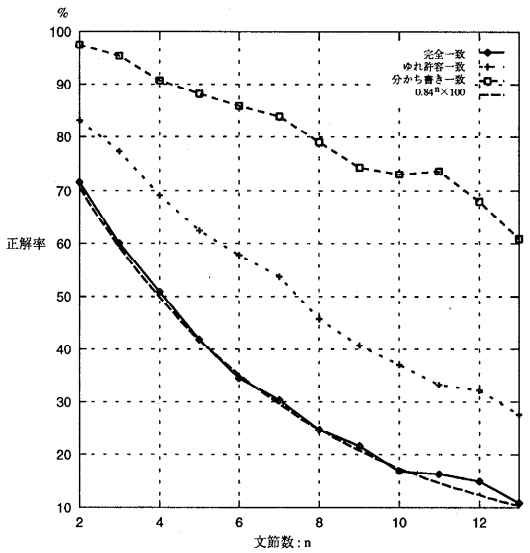


図2 システムEの入力長さ別実験結果
Fig. 2 Results of experiments by system E for each length of the input.

な分かち書きの仕方は4文節（20仮名文字）の入力の場合，524,285通りである☆．一方，図2において4文節入力に対する分かち書き正解率は90.7%である

☆ 長さ m の文字列を最短 a，最長 b 文字のセグメントに分かち書きする仕方の数 $B(m,a,b)$ は，

$$B(m,a,b)=\sum_{k=a}^{b} B(m-k,a,b); \quad b<m \text{ のとき}$$
$$=\sum_{k=a}^{b} B(m-k,a,b)+1 \quad ; \quad a<m\leq b \text{ のとき}$$
$$=1 \quad ; \quad a=m \text{ のとき}$$
$$=0 \quad ; \quad m<a \text{ のとき}.$$

したがって，たとえば，$B(20,1,18)=524,285$.
特に，$b\geq m$ のとき $B(m,1,b)=2^{m-1}$.

表11 システムEのゆれ許容正解の内訳
Table 11 Details of tolerative success of system E.

	4文節以下の入力	入力全体
自立語性連語を含む入力件数：p	3,655	8,651
pのうち正解した件数：q	3,220	5,586
自立語性連語を含まない入力件数：r	7,309	14,272
rのうち正解した件数：s	5,305	8,441
自立語性連語を含む場合の正解率：q/p	0.881	0.646
自立語性連語を含まない場合の正解率：s/r	0.726	0.591
入力件数：p+r	10,964	22,923
正解件数：q+s	8,525	14,027

表12 連語データによる悪影響
Table 12 Undesirable effects of concurrence data.

	4文節以下の入力	入力全体
完全一致だったものがゆれ許容一致になった	130	318
分かち書きの正しかったものが正しくなくなった	133	729
合計	263	1,047

ことを考えると，組合せ的爆発がよくおさえられていることが分かる．また，図2から完全一致の場合の正解率の曲線は，文節数を n としたとき $2 \leq n \leq 10$ の範囲では $0.84^n \times 100$ の曲線とほぼ一致していることが分かる．すなわち，84%の確率で生起する文節n個が前後の相関なしに同時に生起する確率に相当していて興味深い．

4.3.7 連語を含む入力に対する処理結果

入力に含まれる連語が実験結果にどのように影響しているかを調べるため，システムEについてゆれを許容した正解14,027件の内訳を調査した．その結果を**表11**に示す．

表11から自立語性連語を含んだ入力に対する正解率は含まない入力より全体として5.5%高いこと，自立語性連語を含む4文節以内の比較的短い入力に対する正解率は88.1%に及ぶことなどが分かる．

4.3.8 連語データによる悪影響

システムAとシステムEの出力結果を詳細に調べた結果，システムAで正解だったのにシステムEで不正解に変わった入力件数は**表12**のとおりであった．

これらは，たとえば，強連結連語には「御茶を/濁す」が代表表記として登録されていたため，「お茶を/濁す」という変換がゆれ許容の一致と見なされた場合や，「気が/付く」という強連結連語が優先されるため，「捜索機が/着く」という変換ができなかった場合などである．完全一致がゆれ許容一致になった318件については表記の与え方を連語と単語で統一することで解消すると思われる．分かち書きに失敗した729件はシステムEの分かち書き全失敗件数3,241件の22.5%であり，これらを解消することは今後の課題である．

4.3.9 その他の考察

本実験で用いたゆれ表記は約45,000件である．送り仮名，平仮名単語についての網羅性は比較的高いが，「丁ねい」などの漢字仮名混ぜ書き語の収録が不十分であることが分かった．常用漢字以外の漢字を平仮名に置き換えたゆれ表記をさらに充実させる必要がある．

変換精度向上のために拡張文節の考えを採用し，その効果が確かめられたが，実用システムにおいては連語を単語として一括処理すると，その部分的な表記について選択の自由が妨げられる．この点については連語データ中に通常文節に分割する境界を与えることで対処した．操作性におけるこの評価は行っていないが，実用システムでは不可欠な仕組みと考えている．

実験における誤変換の最も顕著なものはカタカナ文字や姓名を中心とする未登録語に起因するものであった．単語辞書の充実も引続き行っていく必要がある．

本稿では，全実験システムで単語の出現頻度や出現順で優先度をつける学習機能を除外したが，これらによる効果は全実験システムでほぼ同様に表れると考えられるため，これらの学習機能を併用した場合も連語データ利用による正解率の向上には大勢としては変化はないだろうと考えている．しかし，詳細については今後，実験を行う必要がある．

今回の実験では第1優先度の変換結果のみを対象として評価したが，第2，第3位まで含めた場合の正解率，同一コスト値の場合の取扱いなどについては検討の余地が残っている．

5. おわりに

本論文では大規模な単語共起表現データの利用が仮名漢字変換の精度に与える効果を測定する実験の結果を示した．本研究の主な特徴は次のとおりである．(1) 分かち書きの単位として連語を単語と見なす拡張文節の考えを採用した．(2) 変換精度向上のため約72,000件の比較的大規模な連語データを利用した．(3) 連語データはすべて人手で採取，整理した．(4) 精度評価のため大量の評価用データ約23,000件を作成し，オープンテストによって評価を行った．(5) 実験用入力は平均約29仮名文字からなる比較的長い仮名文字列である．(6) コスト最小法によって曖昧さの低減を図った．(7) 文節区切りの個人差や送り仮名などの表記のゆれも許容する現実的な評価を行った．(8) 一般的能力

を持つと思われる市販の日本語ワープロソフトWXG Ver2.05との比較，およびその精度改善についても実験を行った．(9)連語に対して単語としての扱いと，通常の文節列としての扱いという二様の扱いをして，操作性の向上を狙った．

実験の結果，連語データの有効性がいくつかの角度から数量的に確認された．当面の課題は代表表記の統一，ゆれ表記の充実，単語辞書の増強，連語データの増強，コスト値設定の手直しなどである．

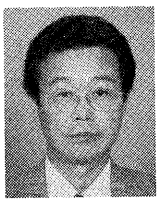
連語は固定的な単語の共起ではあるが，助詞の置き換えや，語順の変更，その他の変化形が許される場合もある．本論文のシステムにおいてはこれらの変化形には十分には対処していない．これらの変化形データの整理は連語データ収集にかかわる今後の重要な課題である[9]．以上の諸点を考慮すればさらに精度を改善できる余地が残っている．本論文の実験では処理の効率を重視しておらず，正確な計測は行わなかったが，実験で得られた印象から連語データによる効率の低下は実用上問題にならないと考えている．

謝辞 本研究の基礎となった連語データの収集・整理作業を永年にわたり行っていただいた福岡大学工学部電子情報工学科知能工学研究室の諸氏，本研究に多方面でご協力をいただいたエー・アイ・ソフト（株）の開発担当諸氏に心から感謝の意を表します．

参考文献

1) 大島義光，阿部正博，湯浦克彦，武市宣之：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).
2) 本間　茂，山段正樹，小橋史彦：連語解析を用いたべた書きかな漢字変換，情報処理学会論文誌，Vol.27, No.11, pp.1062–1067 (1986).
3) 小林　勉，中里茂美，長崎秀紀：ニューロかな漢字変換の実現，東芝レビュー，Vol.47, No.11, pp.868–870 (1992).
4) 山本喜大，久保田淳市：共起グループを用いたかな漢字変換，第44回情報処理学会全国大会論文集，4p-11, pp.189–190 (1992).
5) 高橋雅仁，吉村賢治，首藤公昭：単文内での共起情報を用いた同音語処理，情報処理学会論文誌，Vol.37, No.6, pp.998–1006 (1995).
6) 首藤公昭，楢原斗志子，吉田　将：日本語の機械処理のための文節構造モデル，電子通信学会論文誌，Vol.J62-D, No.12, pp.872–879 (1979).
7) 首藤公昭，吉村賢治：日本語における語の固定的共起，文法的知識と意味的知識の蓄積管理シンポジウム論文集，電子情報通信学会 (1989).
8) 吉村賢治，武内美津乃，津田健蔵，首藤公昭：未登録語を含む日本語文の形態素解析，情報処理学会論文誌，Vol.30, No.3, pp.294–301 (1989).
9) 安武満佐子，小山泰男，吉村賢治，首藤公昭：固定的共起表現とその変化形，言語処理学会第3回年次大会発表論文集，pp.449–452 (1997).
10) 小山泰男，安武満佐子，吉村賢治，首藤公昭：曖昧な文節区切りに対応したかな漢字変換評価用テキストデータ，情報処理学会研究報告，97-NL-122, pp.91–96 (1997).
11) 新納，井佐原：語義の特異性を利用した慣用表現の自動抽出，情報処理学会論文誌，Vol.36, No.8, pp.1845–1853 (1995).
12) Ikehara, S., Shirai, S. and Uchino, H.: Statistical Method for Extracting Uninterrupted and Interrupted Collocations from Very Large Corpora, *Proc. 16th Internat. Conf. on Computational Linguistics* (*COLING-96*), pp.574–579 (1996).
13) Frantzi, K.T. and Ananiadou, S.: Extracting Nested Collocations, *Proc. 16th Internat. Conf. on Computational Linguistics* (*COLING-96*), pp.41–46 (1996).
14) Smadja, F.: Retrieving Collocations from Text: Xtract, *Computational Linguistics*, Vol. 19, No.9, pp.143–177 (1993).
15) Viterbi, A.J.: Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, *IEEE Trans. Information Theory 13*, pp.260–269 (1967).

(平成10年3月11日受付)
(平成10年9月 7 日採録)



小山　泰男（正会員）

昭和26年生．昭和49年山梨大学工学部計算機科学科卒業．昭和49年三協精機製作所入社．昭和53年日本ケミカルコンデンサ入社．昭和55年セイコーエプソン（株）入社．昭和60年エー・アイ・ソフト（株）出向．ワードプロセッサや日本語IMEなどの研究・開発に従事．自然言語処理，情報検索などに興味を持つ．言語処理学会会員．

安武満佐子（正会員）

昭和44年生．平成5年福岡大学理学部応用物理学科卒業．現在同大学工学部電子情報工学科助手．自然言語処理に関する研究に従事．

吉村 賢治（正会員）

昭和30年生．昭和53年九州大学工学部電子工学科卒業．昭和55年同大学院工学研究科電子工学専攻修士課程修了．昭和58年同大学院工学研究科電子工学専攻博士後期課程修了．福岡大学工学部教授．工学博士．自然言語処理に関する研究に従事．電子情報通信学会，人工知能学会，言語処理学会，認知科学会各会員．

首藤 公昭（正会員）

昭和18年生．昭和40年九州大学工学部電子工学科卒業．昭和42年同大学院工学研究科電子工学専攻修士課程修了．昭和45年同大学院工学研究科電子工学専攻博士後期課程単位取得後退学．福岡大学工学部教授．工学博士．自然言語処理に関する研究に従事．電子情報通信学会，人工知能学会，言語処理学会，認知科学会，ACL各会員．