論　文

# 国語辞書の記憶と日本語文の自動分割*

長尾　真**　辻井潤一**
山上　明**　建部周二**

**Abstract**

A dictionary is an important tool for language processing. There have been various forms of dictionaries which were used for mechanical processing of natural language. But most of them were constructed in the course of developing experimental systems and so the sizes of vocabularies were too small. In order to process a large amount of natural language data, we need a dictionary of large vocabularies. We stored such a dictionary and provided convenient access methods for it. Being looked as a data base, a Japanese dictionary has many interesting characteristics. We discuss them and design the most efficient data structure for a Japanese dictionary. Morphological analysis of Japanese sentences is performed by using it.

## 1. はじめに

近年，自然言語処理の研究は人工知能研究の主要テーマとして精力的に研究され，日本語文の解析についてもいくつかの成果が積み重ねられて来ている[1]．しかしながら，現在つくられている日本語文解析システムのほとんどは，入力がローマ字表現あるいは片仮名表記されていることを前提にしており，日本語の通常の表記法である漢字かな混り文をそのまま受付け処理しているシステムはほとんどない．英語のブランクに相当する単語ごとの区切りがない日本語では，漢字の使用が単語の単位を認定する重要な役割を果している．したがって，字種情報を用いないこれらのシステムでは，ほとんどが文節単位の分ち書きを前提にしている．植村はその論文[2]で，文節単位の分ち書きを前提としない漢字かな混り文の処理の基本的な問題点を列挙しているが，今日に致るまで十分に解決されているとは言い難い．

国立国語研究所では，大規模な語い調査を計算機で行う場合に，入力の段階で人間が単語単位に分割して，各種の統計処理を行っている．しかしながら，このような前処理をするためには，ある程度の訓練を受けた人が相当量の作業をする必要があり，漢字入出力装置が発達してきたにもかかわらず，日本語データを計算機処理する際の大きな障害になっている[3]．

ここでは，日本語情報処理の第一段階として，自然な入力文をうけつけ，それを単語単位に自動的に切断する研究を行った．この処理を行うために，大きな語い数をもつ国語辞典を，計算機から柔軟に参照できるようにし，これと字種情報（漢字，かな，カタカナ等の区別）や活用情報を用いることによって，かなり精度のよい単語単位への分割結果を得た．以下では，辞書の記憶構造とその参照手法，文節境界の認定と文節内での接続検定の手法，長い漢字列の単語への分割について，その詳細を述べる[4]．

## 2. 機械辞書

自然言語処理の基本的な道具の１つに辞書がある．計算機処理に使われる辞書には，その処理の目的にしたがって様々な形態がある．ここでは広範囲で大量の日本語データを処理することを目的としているので，登録されている単語数の大きい辞書を持つ必要があっ

* Data-Structure of a Large Japanese Dictionary and Morphological Analysis by Using It by Makoto NAGAO, Jun-ichi TSUJII, Akira YAMAGAMI and Shuji TATEBE (Faculty of Engineering, Kyoto University).

** 京都大学工学部電気工学第二教室

**Table 1** Numbers of lexical entries in the dictionary.

| 品詞 | 語数 | 品詞 | 語数 | 品詞 | 語数 |
|---|---|---|---|---|---|
| 自動詞上一段 | 88 | カ変動詞 | 12 | 連体詞 | 95 |
| 自動詞下一段 | 463 | 感動詞 | 146 | サ変名詞 | 6,760 |
| 自動詞五段 | 1,208 | 接頭語 | 51 | タルト型形容動詞 | 252 |
| 他動詞上一段 | 57 | 接尾語 | 64 | ダ型形容動詞 | 1,103 |
| 他動詞下一段 | 726 | 形容詞 | 626 | 名詞 | 46,279 |
| 他動詞五段 | 1,351 | 副詞 | 1,387 | | |
| サ変動詞 | 271 | 代名詞 | 119 | | |

(*) 助詞，助動詞，接続詞は除く．

た．また，新聞記事や科学論文に現われる一般の単語について，計算機で使用できるような整理された意味記述を与えることは，現在の段階では不可能である．そこで，一応現在出版されている辞書をそのまま使用することにした．使用した辞書は，三省堂出版の「新明解国語辞典」(金田一京助等編) で，電子技術総合研究所推論機構研究室と三省堂とが共同で磁気テープ化されたものを使用させて頂いた．この辞書の概要を **Table 1** に示す．

### 2.1 辞書の記憶方式

辞書の記憶の方式は，これまで機械翻訳の実験との関係で論じられてきた．特に，西村は機械翻訳システム YAMATO の作成報告書において，辞書および文法ルールを木構造で表現することを提案し，それまでの機械翻訳システムで使われた辞書の記憶方式を木構造表現の立場から整理している[5]．**Fig. 1** に木構造(西村の完全形) によって，WALK, WORD, WORDS, ZERO の各単語を記憶した例を示す．この木構造の完全形を基礎にして，これまでいろいろな辞書の形式が提案されてきている．しかし，いずれも実験的な機械翻訳システムで用いられたものであり，項目数も小さく，すべての項目を主記憶上におくことが可能であった．

したがって，ポインタによる項目の探索を行っても，それほど大きな効率の低下をひきおこさないが，我々

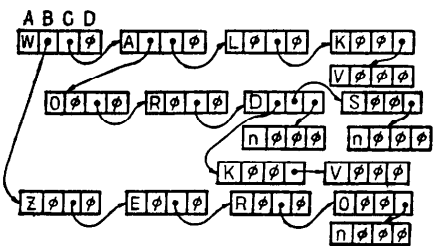



**Fig. 1** Structure of a dictionary (Nishimura's complete form).

の辞書は約7万項目と非常に大きいため，2次記憶装置の使用が不可避となり，ポインタを使った単純なデータ構造においては，ポインタをたどるたびに2次記憶とのデータ転送をひきおこすことになる．このような場合には，主記憶上での項目同志のつき合せに要する時間よりも，主記憶と2次記憶装置間でのデータ転送に要する時間の方が大きくなり，この回数をなるべく少なくする記憶構造が望ましい．

2次記憶上の大量のデータを効率よく探索する方式として，現在よく使われているものに，ハッシング法によるものと，B-tree のような multi-way tree による方式とがある．計算機処理における辞書引きでは，入力文から切り出された文字列を手掛りにして辞書をひくが，この切り出された文字列は，次節で述べるように複合語や活用変化の存在のために，必ずしも辞書中の単語と正確に対応しているとは限らない．ハッシング法では，このような変形したキーからデータ探索をすることはむつかしくなる．そこで，我々は次節で述べるような Multi-way tree を変形した構造で辞書の記憶を行った．

### 2.2 日本語辞書の記憶構造

日本語の辞書をひく際には，基本的には漢字で引く場合と読みのひらがなで引く場合の2通りの探索方法がある．すなわち，各辞書項目には探索用のキーとして，その語の漢字書き文字列と読みのひらがな列の2つがある．これらのキーは次のような性質を持っている．

(1) キーによる項目の指定が一意的でない．漢字で表記される語には，同音異語が非常に多い．また，数は少ないが同字異語（二重――にじゅう，ふたえ）も存在する．

(2) 辞書項目を指定するのに，上記の2つのキー以外に，それを変形した複数個の指定方法がある．これは漢字かな混り文の正書法が確立していないために，同じ単語の表記法が複数個存在するためである．

ここで，(1)の同音異語・同字異語の問題は，情報検索システムで使われる転置ファイルを各キーごとに対して作ることによって解決できる．しかしながら，(2)の表記のゆらぎについては，すべての変形されたキーに対して転置ファイルを作ることは，記憶量の増大につながり，得策ではない．しかも，ほとんどの単語の探索は漢字列かひらがな列で行われ，変形されたキーで探索されることは（通常の日本語文を対象にしているときには），非常に少ないと考えられる．したが

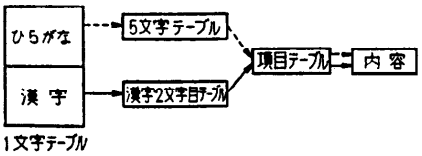



**Fig. 2** Index table structure for retrieval of words.

って，その場合の探索経路を転置ファイルのようなデータ構造で表現しておくことは得策ではないと考えられる．そこで我々は，漢字列，ひらがな列による探索にはできるだけ高速に応答でき，しかも変形されたキーによる探索にも多少時間はかかるが，すばやく応答できるような記憶構造として，**Fig. 2** で示す構造をとることにした．

図中，1文字テーブル・5文字テーブル・漢字2文字目テーブルはそれぞれ項目テーブルを探索するためのインデックス用のテーブルである．太線は漢字から，点線は読みからの探索をするための探索経路を示す．ここで，項目テーブルは読みのひらがなによってソートされており，また漢字2文字目テーブルは漢字の1文字目によってまとめられているので，変形したキーによる探索も，これらのテーブルを参照することによって容易に行われる．すなわち，この記憶方式は国語辞典的な配列と漢和辞典的な配列との両方の性質を備えており，どちらからの探索にも答えることができる．

各テーブルの構成を簡単に示す．

[1文字テーブル]：漢字コードに対応して，4,096 のエントリからできているテーブル．ひらがなのコードに対応するエントリには，そのひらがなから始まる5文字テーブルの最初のエントリへのポインタ，漢字コードに対応するエントリには，漢字2文字目テーブル中のその漢字に対するエントリへのポインタがそれぞれ入っている．

[5文字テーブル]：「項目テーブル」の 40 エントリに対して，この「5文字テーブル」の1エントリが設けられている．その内容は，対応する 40 エントリの辞書項目群中の，最初の辞書項目の読みの先頭から5文字のひらがなをとってきて，それを清音ひらがな書きに変換したものと，40 エントリの項目群へのポインタとの対である．読みからのアクセスの場合には，この 40 項目がディスクから主記憶装置への転送の単位となる．

[漢字2文字目テーブル]：漢字から辞書項目へのアクセスをするためのテーブル．2文字目にくる漢字コードと，対応する辞書項目へのポインタが対になっている．

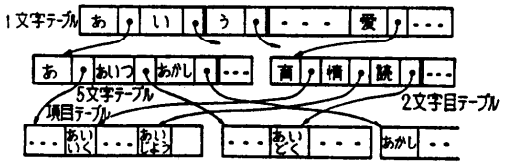



**Fig. 3** Relation between index tables and lexical entries.

[項目テーブル]：見出し語についての統語的な情報を持つ．すなわち，語の読み・漢字表記・品詞・活用型・重要語と非重要語の区別，及び意味記述部分へのポインタが入っている．意味記述部分は不定長であるが，この項目テーブル中のエントリは固定長であり，見出し語を清音ひらがな書きしたものをキーとしてソートされている．

**Fig. 3** に1文字テーブル・5文字テーブル・漢字2文字目テーブル・項目テーブルの関係を示す．5文字テーブルのキーの文字数をひらがな5文字までと制限したのは，測定の結果，字数5までで 92% の辞書項目の見出しを含むことが判明したためである（**Table 2** 参照）．**Table 3** に各テーブルの記憶容量を示す．

### 2.3 辞書の探索方式

基本的な探索機能として，前節に示した内部構造に基づいて，次のようなものを設けた．

（1） ひらがな書きされた見出しをキーとする探索．

（2） 漢字で書かれた見出しをキーとする探索．

（3） (1)で探索された項目の次にくる項目の探索．項目テーブルは清音ひらがな書きにソートされているので，この配列上で後続する項目を次々にとり出す．これは同音異義語探索用である．

（4） (2)のキーと同じキーを持つ別の辞書項目の探索，これは同字異語および派生語（代数→代数学）探索用の機能である．

**Table 2** Frequency of words in different lengths.

| 字数 | 1 | 2 | 3 | 4 | 5 | 6 | 7以上 |
|---|---|---|---|---|---|---|---|
| 累積度数 | 391 | 4,217 | 17,891 | 42,449 | 53,043 | 55,932 | 57,557 |
| 累積相対頻度(%) | 0.6 | 7.3 | 31.0 | 73.7 | 92.1 | 97.1 | 100 |

(派生語は統計から除く)

**Table 3** Storage requirement for the index tables.

| テーブル名 | 容量 | テーブル名 | 容量 |
|---|---|---|---|
| 1文字テーブル | 24 KB | 漢字2文字目テーブル | 700 KB |
| 5文字テーブル | 18.6 KB | 見出し項目テーブル | 4.3 MB |

さらに，3節，4節で述べる漢字かな混り文の処理を行うためには，変形されたキーによる探索を行う必要があり，このために下記に示すような探索機能を設けた．

（5） 省略可能な送りがなを省略した見出しをキーとする探索．

（6） 漢字表記の単語の一部をひらがな書きした見出しをキーとする探索．

（7） 活用語の語幹をキーとする探索．

（1），（2），（3），（4）の探索機能がいずれも **2.2** で述べた5文字テーブルや漢字2文字目テーブル等の内部構造を直接参照するだけで実現できたのに対して，この（5），（6），（7）の各機能は，5文字テーブルや漢字2文字目テーブルを手掛りにして，項目テーブル中の特定の個所へ到達した後は，項目テーブル中の記述をプログラムでチェックしながら探索する．したがって，これらの探索は（1）～（4）の機能に比べて，処理時間を要する．

上記の（1），（2）の機能について，日本科学技術情報センター（JICST と略す）の論文抄録中に含まれる単語をランダムに100語とり出し，そこからランダムに1語ずつとり出して1,000回の辞書探索を行ったところ，一語あたり83 msec の処理時間であった（辞書記憶に使用したディスク装置の平均アクセス時間は38 msec）．

## 3. 日本語文の単語単位への分割と品詞認定

日本語文の基本的な構成単位に文節がある．我々は，与えられた自然な入力文から，まず文節境界の認定を行い，次に文節内での単語の品詞とその活用形の決定を行う．第一段階の文節境界の認定は，句読点と字種の情報だけを使って行うもので，これにはかなりの誤認定が含まれるが，第二段階の活用形と品詞を使った処理を行う過程で，この誤認定のほとんどが回復される．

### 3.1 品詞認定の基本的なアルゴリズム

品詞の認定は，あらゆる言語情報処理の最初に行われる処理である．したがって，ここでの失敗は次に続く処理の致命的な失敗につながるために，この品詞認定の処理はできるかぎり正確に行われる必要がある．したがって，我々の品詞認定プログラムでは，可能な解釈がすべて出力されるようになっている．次にこの品詞認定のアルゴリズムをステップごとに述べる．

［ステップ1］ 句読点，および字種の変化する位置（ひらがな列→漢字又は片仮名列，数字列）を文節の切れ目とする．以下のステップでは，この仮に定められた文節内での処理を行う．

**Table 4** Tables used by the procedure for morphological analysis.

| テーブル名 | 内　　容 | 語数 |
|---|---|---|
| 付属語テーブル | 助詞，助動詞，接続詞 | 152 |
| ひらがな書き自立語 | 普通ひらがなで書く自立語…こと，とき etc. | 269 |
| 活用語尾テーブル | 用言の活用語尾 | 254 |
| ひらがな語幹 | 語幹の一部にひらがなを含む自立語 | 60 |
| 特殊動詞テーブル | 漢字が先頭にくる動詞で，活用がルール化できないもの | 94 |

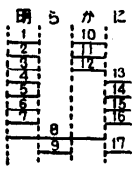
［ステップ2］ 文節中の文字列で，単語（自立語，付属語）とみなすことのできる文字列をすべて検出して，ユニットと呼ぶ単位をつくる．

このユニット作成のために，**Table 4** に示す各種のテーブルが使用される．ただし，一般の自立語は字種（漢字・片仮名）だけで判定され，辞書引きはこの段階では行われない．また **Table 4** のテーブルは2節で述べた西村の木構造表現の完全形に従って作られている．

これは，ここで使われるテーブルの大きさが，辞書に比べてはるかに小さく，すべてを主記憶上に作れること，またこのような木構造でテーブルを作成しておくことによって，1回の探索で入力文のある特定の文字位置から始まる単語とみることのできる文字列をすべてとり出すことができることによる．これによって作成されたユニットの例を **Fig. 4** に示す．

［ステップ3］ ユニットの接続検定と連鎖の作成

このステップでは，ステップ2で作られたユニットを文節の先頭から探索して，接続の可否をチェックし，文節の可能な解釈をすべて求める．ユニットは，各単語の文節内の位置（何字目から何字目までか）と，品詞・活用情報および接続条件の3部分からできている．文節内での単語の相互接続の可否は，連続する2語の品詞と活用情報だけから決定できる．また，ユニットはすべて一様な構造をしており，品詞・活用情報や接続条件はコード化されているので，このステップでの処理は比較的簡単に実行できる．現在，接続条件を 50




| ユニット | 位置 | 品詞 | ユニット | 位置 | 品詞 | ユニット | 位置 | 品詞 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1～1 | 名 | 7 | 1～1 | 副 | 13 | 4～4 | 形動尾用 |
| 2 | 1～1 | 形容詞幹 | 8 | 1～3 | 形動幹 | 14 | 4～4 | 動尾用 |
| 3 | 1～1 | 形動幹 | 9 | 2～2 | 動五尾未 | 15 | 4～4 | 動上尾用 |
| 4 | 1～1 | 動五幹 | 10 | 3～3 | 動五尾未 | 16 | 4～4 | 動上尾未 |
| 5 | 1～1 | 動上幹 | 11 | 3～3 | 副 助 | 17 | 4～4 | 格 助 |
| 6 | 1～1 | 動下幹 | 12 | 3～3 | 終 助 | | | |

**Fig. 4** Example of units generated by the program.

(1) 実験の結果に対するそぼくな疑問……
(2) 動作が安定しはじめた頃に……
(3) 実験をくり返すことによって……
(4) 昨年米国の消費した……

**Fig. 5** Examples of the misrecognition of boundaries.

のタイプに分類しているが，このタイプの数は実験結果によって，より精密化してゆく予定である．このステップで可能な解釈が得られなかった場合には，ステップ1で仮定された文節の境界に誤りがあったと考えられるので，次のステップ4でこの誤りの回復を行う．解釈が求まった場合にはステップへ行く．

［ステップ4］ ステップ1での処理が失敗する原因には，次の3通りの場合が考えられる．

(i) (仮定された)文節中に，ひらがな書きされた自立語が埋もれていたために，実際には2つ以上の文節が入っていた場合(**Fig. 5**, 例(1), (2))．

(ii) 自立語の1部の漢字がひらがな書きされている場合 (**Fig. 5**, 例 (3))．

(iii) 前の文節が漢字表記の自立語で終っているために，次の文節の先頭の漢字列と連続してしまっている場合 (**Fig. 5**, 例 (4))．

ここで，(iii)の誤りは，通常一語とはみなされない漢字連続(その後専門分野……，昨年米国……) が単独の自立語と認定されるが，ステップ3までの処理では，この誤認識は発見されない．この誤認識の回復は次節4で述べる漢字連続の分割によって行われる．このステップ4ではそれ以外の2種類の誤りが回復される．

この2種類の誤りを回復するために，境界検定ルーティンと自立語探索ルーティンが用意されている．上記失敗原因の(i)は自立語探索ルーティンによって，また (ii) は境界検定ルーティンによって，処理される．いずれのルーティンも自立語の辞書引きを行うが，辞書引きに要する時間は非常に大きいので，なるべくその回数が少なくなるように工夫されている．すなわち，まず埋もれた自立語の位置を仮定し，そのひらがな列を除いてできるひらがな列を付属語列として解釈し，解釈が成立すれば，そこに自立語の存在する可能性が強いとして実際の辞書引きが行われる．付属語列の解釈を行うための処理時間の方が辞書引きに要する時間よりもはるかに小さいためである．

［ステップ5］ 文節の先頭にくる自立語の辞書引きを行う．自立語は一応辞書に登録されていることを前提にしているが，科学論文にあらわれる専門用語や長い漢字列の複合語などは，そのままの形では辞書に登録されていない．また，活用語については，原形にもどして辞書引きをする必要がある．そこで，我々の品詞認定プログラムにおいては，まず，ひらがなで書かれた部分の解釈をステップ3までで行った後に，その解釈の正当性を確認するために，自立語の辞書引きを行う．また，ステップ3までで複数個の解釈が存在する場合にも，この段階であいまい性が解消されることが多い．この時点では，自立語部分の品詞の解釈がすでに確定しているので，活用語についても辞書引きは容易である．ただし，長い漢字連続からなる複合語や，2つの漢字表記自立語が連続している場合には，次節で述べる漢字連続の分割処理が行われる．

**Table 5** Score of results and processing time.

処理結果の正答率

| | 単一の正答 | | 複数の結果に正答を含む | | 失敗 | |
|---|---|---|---|---|---|---|
| | 単語 | 文節 | 単語 | 文節 | 単語 | 文節 |
| 辞書引き前 | 73.0 | 61.4 | 20.9 | 33.2 | 6.1 | 5.4 |
| 辞書引き後 | 87.0 | 74.8 | 9.5 | 20.0 | 3.5 | 3.4 |

(単位は%)

処理時間

| | 全処理時間 | 1文当り | 1文節当り | 1単語当り |
|---|---|---|---|---|
| 辞書引き前 | 3分35秒 | 4.6秒 | 0.43秒 | 0.17秒 |
| 辞書引きを含む | 27分40秒 | 35.3秒 | 3.32秒 | 1.30秒 |

磁気テープの入出力を含む．
使用計算機は TOSBAC-40 C, 主記憶は 64 KB.

### 3.3 結果と検討

啓林館・昭和 50 年発行・高校理科用教科書「改訂化学Ⅰ」5ページ以降の 47 文 (500 文節) について処理した結果を **Table 5** に，処理結果の一例を **Fig. 6** (次頁参照) に示す (Fig. 6 の下線は筆者)．

図中，例 (1) は最後の文節の解釈にあいまいさのある場合，例 (2) は埋もれた自立語を検出した例である．

全体の処理結果をみると，品詞接続条件だけからみると誤りではないが，実際の文にはあまり現れないような品詞列が結果として出力され，複数個の結果になっている場合が多い (例：分解 (名) で (格助詞) き (動詞・か変・連用))．しかし，これを防ぐことはこの段階の処理では不可能であり，今後この処理結果を使って，統語処理や意味処理に進む際に解決しなければならない．

## 4. 漢字連続の分割

「磁気記憶装置」という語は，「磁気」，「記憶」，「装置」の3つの基本的な語からできている．このように，よ

例（1）
多くの物質の中には，分子の存在が認められないものもある．
*
* 多く（名　）の（格助　）／(1)／*
*
* 物質（名　）の（格助　）／(1)／*
*
* 中（名　）に（格助　）は（副助　），（読点　）／(1)／
* 中に（名　）は（副助　），（読点　）／(0)／
* 中に（名　）は（副助　），（読点　）／(0)／*
*
* 分子（名　）の（格助　）／(1)／*
*
* 存在（名　）が（格助　）／(0)／*
*
* 認め（動下未）られ（助動用）ない（形容体）もの（形名　）も（副助　）ある（動五終），（句点　）／(1)／
* 認め（動下未）られ（助動未）ない（助動体）もの（形名　）も（副助　）ある（動五終），（句点　）／(1)／*
例（2）
周期表では，電子配列の似た元素が族としてまとめられている．
*
* 周期表（名　）で（格助　）は（副助　），（読点　）／(0)／*
*
* 電子配列（名　）の（格助　）／(0)／*
*
* 似（動上用）た（助動体）／(1)／*
*
* 元素（名　）が（格助　）／(1)／*
*
* 族（名　）と（格助　）し（動サ用）て（接助　）まとめ（動下未）られ（助動用）て（接助　）いる（動上終），（句点　）／(3)／
* 族とし（動五用）て（接助　）まとめ（動下未）られ（助動用）て（接助　）いる（動上終），（句点）／(2)／*
(*) 図中，各文節の末尾につけられた数字は以下のことを示す．
1. ステップ3までで解釈が成立し，ステップ5の辞書引きで正当性が認められたもの
0. 1と同じであるが，辞書引きでは正当性が認められなかったもの
3. ステップ4の失敗回復ルーティンで成功し，ステップ5の辞書引きで正当性が認められたもの
2. 3と同じであるが，辞書引きで正当性が認められなかったもの
(**) (例1)の下線「ない」は複数個の解釈が成立するもの．
(例2)の下線「まとめ」は，ひらがな表記の自立語を失敗回復ルーティンによって検出したものを示す．

**Fig. 6** Examples of morphological analysis.

り基本的な単語の結合した複合語は複次結合語ともいわれる[6]が，前節のステップ4やステップ5の自立語の辞書引きにおいて，このような複次結合語はそのままでは，辞書項目に登録されていない．そこで，このような複次結合語をより基本的な単語に分割して辞書を引く必要がある．日本語の複次結合語は，他の言語では句で表現されるようなものを多く含んでおり，機械翻訳や情報検索に言語処理手法を適用する場合にも，これを基本的な構成要素に分割することは不可欠である．また，前節の品詞認定プログラムにおいても，「その後専門分野を……」，「毎月消費する……」のように連続する漢字列が必ずしも1つの自立語には対応しない場合も多い．

**Table 6** Ratio of different structures of compound nouns.

| 3文字語（例） | 2-1 空芝-域 | 1-2 新-形式 | 1 1 > 1 内-外-径 | 1-1-1 空-対-空 |
|---|---|---|---|---|
| 計 1,001 | 793 (79.2) | 203 (20.3) | 2 (0.2) | 3 (0.3) |
| 4文字語（例） | 2-2 高級-言語 | 1-2-1 非-正規-性 | 2-1-1 分散-系-中 | |
| 計 337 | 298 (83.7) | 26 (7.4) | 13 (3.9) | |
| 5文字語（例） | 2-(2-1) 総合自動化 | (2-1)-2 周波数分析 | (1-2)-2 非直線動作 | (2-2)-1 電圧電流図 |
| 計 240 | 83 (34.6) | 78 (32.5) | 39 (16.3) | 25 (10.4) |
| | 1-(2-2) 各試験装置 | 2-(1-2) 擬似逆行列 | | |
| | 8 (3.3) | 7 (2.9) | | |
| 6文字語（例） | 2-2-2 電信電話装置 | (2-1)-(2-1) 抵抗値許容差 | (1-2)-(2-1) 軸駆動発電機 | (2-(2-1))-1 地球物理学士 |
| 計 159 | 105 | 22 | 5 | 5 |
| | (2-1)-(1-2) 固定子各巻線 | その他 | | |
| | 3 | 8 | | |

## 4.1 漢字連続の性質

漢字連続の性質を調べるために用いた資料は，日本科学技術情報センター（JICST）発行の文献速報用の磁気テープ電気工学編(昭和49年6月～11月，10巻)中の抄録文である．この資料には，異なり数 74,127 の漢字連続が含まれていた．うち約 2% は，「昨年米国」，「毎月消費」のような独立した自立語とは認められないものであった．

また複次結合語の構成には様々なパターンがある[7]．JICSTの資料についてこの結合パターンの割合を調べた結果を **Table 6** に示す．将来はこのような構成要素間の関係も決定する必要があるが，これは意味処理の問題になるので，ここでは構成要素の認定だけを行う．

## 4.2 漢字一文字の性質

複次結合語を基本構成単位の2文字漢語と1文字漢語（接頭語，接尾語，独立語等）に分割する時に，役に立つ手掛りを与えるのは，1文字語の存在である．1文字語には，「非」，「不」，「的」，「中」……のように接頭語的，接尾語的な性質の強いものから，「熱」，「線」……のように独立語的なものまでいろいろある．また，漢字の中でも，構，索，慣，追のように全く1文字語としては使われないものも多い．漢字連続の分割には，

**Table 7** List of prefixes and suffixes in japanese (left is the table for prefix and right is the table for suffix).

| 漢字 | 接頭語% | 総頻度 | 漢字 | 接頭語% | 総頻度 | 漢字 | 接尾語% | 総頻度 | 漢字 | 接尾語% | 総頻度 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 左 | 100 | 6 | 弱 | 72 | 11 | 権 | 100 | 5 | 側 | 85 | 61 |
| 超 | 100 | 20 | 主 | 70 | 55 | 灯 | 100 | 15 | 法 | 84 | 278 |
| 副 | 100 | 9 | 横 | 69 | 13 | 箱 | 100 | 10 | 炉 | 82 | 28 |
| 諸 | 98 | 67 | 縦 | 69 | 12 | 策 | 89 | 19 | 器 | 81 | 181 |
| 非 | 97 | 44 | 全 | 66 | 109 | 者 | 89 | 93 | 鏡 | 81 | 22 |
| 各 | 95 | 186 | 低 | 65 | 108 | 例 | 89 | 96 | 等 | 81 | 133 |
| 他 | 90 | 31 | 無 | 64 | 50 | 棒 | 88 | 18 | 盤 | 81 | 16 |
| 肺 | 85 | 7 | 再 | 63 | 49 | 系 | 87 | 210 | 片 | 81 | 16 |
| 逆 | 78 | 48 | 枝 | 62 | 8 | 型 | 87 | 89 | 員 | 80 | 10 |
| 新 | 76 | 43 | … | … | … | 剤 | 85 | 21 | | … | … |

このような個々の漢字の統計的な性質を手掛りにすることが考えられる.

3文字からなる漢字連続は，もしその前の2文字か後の2文字からなる2文字列のいずれか一方だけが辞書に登録されている場合には，それぞれ 2-1，あるいは 1-2 に分割してほぼ間違いがない(3文字漢字連続のうち，前の2文字列又は後の2文字列の一方だけが辞書に登録されている1,632例に対して，この分割を行った結果，正答率は 98.2% であった). この場合，残った1文字の漢字は，それぞれ接頭語的，あるいは接尾語的に使われたと考えられる. そこで，このことを使って各漢字について，接頭語的，接尾語的に使われた相対頻度を求めると，これがその漢字の性質を示していると考えることができる. 11,963 の3字漢語を使って，この処理を行った結果，**Table 7** を得た. この表中の数字は，ほぼその漢字の性質を反映していると考えられる.

### 4.3 複次結合語の分割アルゴリズム

我々はこの漢字一文字の性質に注目して，これを使ったルールだけで複次結合語の分割を行うアルゴリズム (BUNCUT) も開発し実験を行ったが，**Table 8** に示すようにこのアルゴリズムでは長い漢字列に対してはかなり失敗が多くなる. そこで，ここでは漢字の性質と辞書とを併用したアルゴリズムについて述べる (Table 8 の J & T).

**Table 8** Scores of segmentation of compound words.

| アルゴリズム＼文字数 | 3 | 4 | 5 | 6 | 7 | 8 | 9以上 |
|---|---|---|---|---|---|---|---|
| BUNCUT | 82.6 | 87.2 | 80.0 | 81.2 | 66.4 | 63.4 | 54.0 |
| J & T | 94.9 | 93.6 | 92.7 | 91.5 | 90.0 | 84.9 | 78.0 |

(表中の数字は正答率%を示す)

BUNCUT: 漢字一文字の性質を使ったルールのみで分割

J & T: 辞書引きを行った後にルールで分割

A, B, C, D, E: 各漢字

a, b, c, d, e: 各漢字の接尾語テーブルの値 (%)

a′, b′, c′, d′, e′: 各漢字の接頭語テーブルの値 (%)

(1) 3字の連続の分割……/ABC/……

if max(a, a′)>max(c, c′), then /A/BC/ otherwise /AB/C/

(2) 4字の連続の分割……/ABCD/……

if min (max (a, a′), max (d, d′))>max (b, b′, c, c′, then /AB/CD/ otherwise /A/BC/D/

(3) 5字の連続の分割……/ABCDE/……

if max (a, a′)≥max (c, c′, e, e′), then /A/BC/DE/

otherwise if max (c, c′)≥max (a, a′, e, e′), then/AB/C/DE/

otherwise /AB/CD/E/

(4) 6字以上の連続は，(1)，(2)，(3) の形に還元して行う

**Fig. 7** Rules for segmenting compound nouns.

(a) 成功例

(1) 正／三角／形／格子／標本／化／図形

(2) 一般／産業／用／制御／機器／分野

(3) 大／容量／揚水／発電／電動／機

(4) 集果／性／光学／繊維／素子／単体

(b) 失敗例

(1) 通信／路頭／長

(2) 電話／器用／難燃／材料

(3) 高速／作動／電／磁界

**Fig. 8** Examples of segmentations.

このアルゴリズムでは，まずはじめに，与えられた漢字列から単語として辞書に登録されているものをチェックする. この段階で一意に分割できるものについては処理を終了する. 一意に分割できないものには次の2種類がある.

(1) 辞書が不備で，登録されていない語があった場合(多くの専門用語は辞書にない).

(2) 辞書に登録されている語で2通りの分割が可能であった場合 (不安定……不安，安定のいずれもが辞書にある).

次に，この辞書だけではうまく分割できなかった部分について，**4.2** で決めた漢字一文字の性質 (Table 7) を使って分割する.

**Fig. 7** に Table 7 を使ったルールを示す

**Fig. 8** (a) は分割の例，Fig. 8 (b) は失敗例である. 失敗例(1)は，辞書が一般向きに作られており，専門分野であまり使われない語を含んでいたための失敗例，例 (2) は Table 7 の不備によるものである. また，例 (3) は複次結合語の形成の際に，縮退 (電界磁界→電磁界) があった場合である. 今後このような失敗例に対しても，うまく処理できるアルゴリズムを開発する必要がある.

## 5. おわりに

本論文では，言語処理の第一段階である形態素分析の日本語文に対する適用について考察した. 品詞認定

および漢字列の分割のプログラムは，いずれも一般的に作られており，そこで使われるテーブル類や辞書も特定の分野にあわせては作られていない．したがって，ある専門分野の用語集等の適用分野固有の情報が使用できる場合には，ここで示した結果よりも良い処理結果が得られることが期待できる．情報検索システムにおける自動インデクシング等の問題に適用する場合には，このような配慮も必要であると思われる．また，自然言語データに対して統計処理を行う場合，従来は人手で前処理の施されたデータに対して行うか，あるいは全く字づらだけの統計(例えば一文あたりの文字数)がとられていた．前者の場合には，大量のデータを得るためには大きなコストがかかり，後者では意味のある統計量が得にくいという問題があった．今後，このような問題に対して，本論文で述べた手法を適用してゆく予定である．

なお，本研究の一部は文部省科学研究費補助金によって行った．

## 参考文献

1) 長尾 真，辻井潤一：自然言語処理プログラム，情報処理，Vol. 18, No. 1, 1977.
2) 植村俊亮：漢字かなまじり文 KWIC 索引，情報処理，Vol. 10, No. 5, 1969.
3) 中野 洋：品詞認定の自動化，国立国語研究所報告 39, 1971.
4) 長尾，辻井，山上 明，建部周二：言語情報処理のためのサポートシステムII，信学会研資 AL 77-25, 1977.
5) 西村恕彦：機械翻訳プログラムの作成，電気試験所報告，No. 712, 1970.
6) 野村雅昭：複次結合語の構造，国立国語研究所報告 49, 1973.
7) 野村：三字漢語の構造，同上 51, 1974.

(昭和52年8月9日受付)
(昭和52年12月12日再受付)