

## 漢字かなまじり文 KWIC 索引\*

植村俊亮\*\*

## Abstract

This paper describes an experimental Japanese KWIC indexing system. A *kanzi*-teletypewriter is used as an off-line I/O device. The method of automatic segmentation of Japanese sentences is based upon the classes of characters (*kanzi*, *hiragana*,...) and some other characteristics of *kanzi*. This method is enough simple and of practical use.

The KWIC index list may be printed either laterally (from left to right) or vertically (downward). Examples are shown.

Automatic processing of Japanese sentences causes many troubles. Some of them are discussed and an inverted Japanese word-list is also mentioned.

## 0. まえがき

電子計算機の入出力に日本語とくに漢字かなまじり文を使用することにはいくつかの困難な問題があつてなかなか実現しない。しかし情報の検索や提供のシステムは、利用者として不特定の、計算機についてはなら特別な教育を受けていない人を想定しなければならぬので、入出力言語として日本語を使用したい要求は切実であり、またそれが実現すれば、実用上の効果ははなはだ大きいと考えられる。

本論文は情報の提供用索引としてすでに広く実用に供されている KWIC (Keyword-in-Context) 索引を漢字かなまじり文を対象として電子計算機で作成させる実験の報告である。電子計算機への日本語入出力装置として漢字テレタイプライタを使用した。この索引作成システムは、与えられた入力日本語文から索引の見出し語を自動抽出する方法(1種の自動わかし書き)として辞書をあまり使用しないプログラム方式によつたこと、出力としてよこ書きとたて書きの任意の索引が得られることなどに大きな特徴を持ち、単純ではあるが実用的なものである。なお日本語を自動処理する場合の一般的な問題についてもふれた。

## 1. 日本語 KWIC 索引自動作成の問題点とその解決

## 1.1 KWIC 索引

\* A Japanese KWIC Indexing System by Syunsuke UEMURA (Automatic Control Division, Electrotechnical Laboratory)

\*\* 電気試験所制御部

KWIC (Keyword-in-Context; 文脈付き見出し語) 索引<sup>1)</sup>とは、名の示すとおり、見出し語(keyword)とともにその見出し語の使用された文脈(context)をも表示する索引である。Fig. 4, Fig. 5にある日本語 KWIC 索引の例からも明らかなように、文脈を併記することによって単なる見出し語だけの索引よりもはるかに精密な検索(主として人手による検索であるが)を可能にする。

この索引にはいろいろの方面への応用が考えられている<sup>2)</sup>が、本論文ではその中心である科学技術文献の題名を対象としたいいわゆる題名 KWIC 索引だけを考えるものとする。

KWIC 索引は電子計算機で簡単に作成できるのが大きな特徴で、新語の発生などへの対応策が原則として不必要であり、機動性に富んでいる。たとえば英語の KWIC 索引では、見出し除外語(non-keyword; 冠詞など見出し語にする必要のない語)とよばれる一群の語の辞書を一度作成しておけばよく、この辞書にない語はすべて見出し語にするので、索引作成をはなはだ単純に行なえる。日本語で KWIC 索引を作成する場合にも、この特徴を失なわないように留意する必要がある。

## 1.2 日本語の自動処理

とくに KWIC 索引の場合に限らずごく一般的に、漢字かなまじり文を対象にした日本語の自動処理を行なおうとすれば、解決しなければならない課題がいくつかある。それは

- (1) 日本語入出力装置

(2) わかち書きの処理

(3) 文字の大小順序と見出し項目の並べ方

の3つに大別できよう。

このうち(1)については特別の仕様の漢字テレタイプライター(以下漢テレと略称する)をオフラインで使用し、またこの漢テレのコードにしたがってのみ(3)に対処した。これについては5., 6., 7.に述べる。

わかち書きとは、文章を一定の規則にしたがってくぎって、間隔を置きながら書くことをいう。通常、英語は単語単位にわかち書きされるが、日本語にはわかち書きの習慣がない。これは日本語を自動処理する場合の最大の難関の1つである。

このシステムでは、わかち書きのために独自の手法を用いた。その概略は1.3, 1.4に、詳細は3.にある。

### 1.3 日本語 KWIC 索引とわかち書き

KWIC 索引作成の立場からわかち書きの方法を検討する。

(1) あらかじめわかち書きされた入力データを使用すること。

人間がわかち書きを行なうので、正確な KWIC 索引を作成できる。しかし人手による予備編集(わかち書き)は情報提供用索引としての KWIC に要求される機動性と相反する。作成される KWIC 索引がたいへん見づらいので、入力データはわかち書きしても出力の KWIC 索引は再編集して読みやすいものにする必要がある。

(2) 入力データは予備編集をしないで、なんらかの方法で自動わかち書きを行なわせる。

(i) テーブル・ルックアップ方式

カタカナを対象とした JAKIS<sup>2)</sup>はこの方式である。わかち書きを大規模な辞書を参照して行なう。これはすべての見出し語を辞書に登録しておくもので、辞書にない語は見出し語にならない。実用規模の辞書の編成、それによる索引作成に必要な時間、新語の発生に対応する辞書のメンテナンスなど問題が多く、KWIC 索引本来の手軽さ、機動性を生かしきれない。

(ii) プログラム方式

辞書を最小限にとどめたプログラムによるわかち書きが可能であれば、KWIC 索引本来のよさをそのまま保存できよう。この場合プログラムは特定の字、語に関したものでなくて、ごく一般的な、原則的なも

のであることが望ましい。

プログラム方式で一般的な用途のための完全なわかち書きを行なうことは困難かもしれないが、KWIC 索引作成のために、必要な程度のわかち書きは可能である。

### 1.4 プログラム方式によるわかち書き

わかち書きされていない漢字かなまじり文をわれわれがさして困難なく読み下し、理解できるのは、そこに多くの字種(漢字、ひらがな、カタカナなど)が入りまじっていて字種の変化によっていく分のくぎりの認定ができるからだと考えられる。これは電報のようにカナ文字だけを使ってしかもわかち書きされていない文章を読み下すのはたいへん困難であることから、理解できる。むろん字種の使いわけは厳密なわかち書きとは深いつながりはなく、これによって日本語の単語単位の認定までも行なえるとは考えられないが、KWIC 索引のための見出し語の検出でいどのわかち書き(くぎり)には十分に有効である。

たとえば

**ハザードのない最簡論理回路の設計\***

という題名からは KWIC 索引のための見出し語として、まず

**ハザード**

**最簡論理回路**

**設計**

を選べばよく、また

**国産の COBOL コンパイラ**

からは

**国産**

**COBOL**

**コンパイラ**

を選べばよいが、これらはいずれも字種の変化する部分に着目すると、各語の意味は熟知しなくても容易に選定できる(いいかえればわかち書きができる)。

さらに**最簡論理回路**だけではなくて

**論理回路**

**回路**

をも見出し語にするためには、漢字列を意味のある単位にくぎらなければならぬが、科学技術用語では、この例のように原則として2字ごとにくぎればよい(む

\* 以下本論文中で説明や例のために用いた題名はすべて、次の2つの論文誌から引用したものである。

情報処理 第8巻 1~6号 (1967)

電気試験所集報 第31巻 1~12号 (1967)

実験もこれによって行なった。

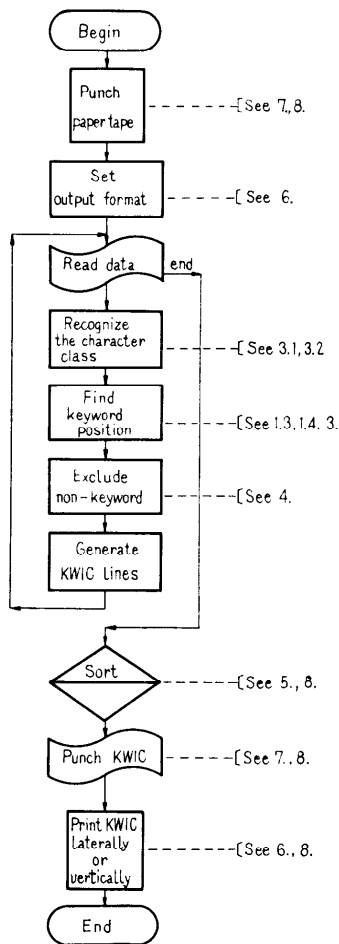


Fig. 1 System Flow

るん漢字列の処理にはいくつかのさらに細かい手続きが必要であり、それは 3. で述べる。

本システムでは以上のような観点から辞書をできるだけ小さくして、プログラムによって見出し語の検出を行なった。ここではわかち書きすべき文字の位置だけを検出するのであって、実際に間隔を置くのは見出し語の最初の文字の前だけである。しかし処理の内容は自動わかち書きにほぼ相当する。

## 2. システムの概要

システム全体の流れを Fig. 1 に示す。各部分について言及している章、節を同図中に注記してある。日本語入出力装置としての漢テレは紙テープを介してオフラインで使用した。

## 3. 漢字かなまじり文中の見出し語位置の検出

### 3.1 概説

見出し語位置とは題名中で見出し語とすべき部分の最初の文字（さきの例ではハザードのハ、最簡論理回路の最、論、回、設計の設など）をいう。この位置を決定できれば、その文字の直前に間隔を置いてKWIC索引の見出し語にすればよいことになる。見出し語位置検出のためには、まず入力データを読み込んだ段階で、各文字の字種を判定する。次に得られた字種のストリングをもとに、時には文字そのものをも参照しながら見出し語位置を検出する。

### 3.2 字種の判定

実験に使用した漢テレ(7. 参照)はこの問題についてよく配慮されたコード系を持ち、文字1字を表現する紙テープ上の2フレーム(情報ビット $6 \times 2 = 12$ ビット)のうち、第1フレーム目によって字種を容易に判定できる。したがってプログラムのにもきわめて楽であった。この処理によって、入力データ中の漢字が続く部分、ひらがなの部分、…を知ることができる。

### 3.3 漢字列の処理<sup>3)</sup>

日本語中で使われる漢字は文字がいくつも連らなって新しい語彙を形成していく。科学技術用語でこの連らなり方を見ると、けっして無原則ではなくてほぼ次の3つの規則にしたがっている。

(1) 2字の漢字が1単位となり、いくつかの単位が連らなっていく。

能力  
記憶能力  
短期記憶能力

(2) 漢字2字で1単位となった語あるいは語の列の前または後に1字の漢字がつく。

計算  
計算機  
容量  
大容量

これは日本文法という接頭(尾)辞にはほぼ相当するが、ここでは漢字のみをさすのでかりに接頭(尾)字と呼ぶことにする。

(3) 上の2つの形の混合。どんなに長い漢字列でも(1)、(2)またはその繰り返しと考えることができる。

## 大容量磁性薄膜記憶装置

## 半導体設計者

以上の3つの規則を書き替え規則の形で表現すると次のようになる。

```
<K-string> ::= <K-string> <K-string>
                | <K-prefix> <K-unit>
                | <K-unit> <K-suffix>
                | <K-unit>
<K-unit> ::= 容量 | 導体 | 計算 | …
<K-prefix> ::= 大 | 半 | …
<K-suffix> ::= 機 | 器 | …
```

以上は漢字列がある意味を持った単位の連らなりで構成されていることを示すもので、むしろその連らなりの意味関係までは示さない。容量が大きい記憶装置と半導体を設計する者の違い、建築と会社が連らなって建築会社になるのに対して出版と会社は連らなって出版社になるなどの問題には、まったく立ち入っていない。

これらを表層構造 (surface structure) と深層構造 (deep structure) とに対応させて研究するのは興味ある問題だと思われる。

また明らかに上の原則に反する例として

## アウエルバッハ氏来日

## 第2回支部評議会

のそれぞれ氏、回などがある。これらを接頭字とすれば原則にはかなうが意味的には明らかにまちがいである。

KWIC 索引作成のためには漢字列の意味的な連らなりまで考慮する必要はなく、みかけ上くぎるだけでよいので、その規則を次のように定めた。

(1) 漢字列は原則として2字ごとにくぎって見出し語とする。

(2) ただし接頭字や接尾字のみはあらかじめ辞書に収めておいて処理する。

(3) (1), (2)によるあいまいさや処理の誤りを防ぐために、途中でくぎりたくない語 (漢字列) も辞書に収めておき, (1), (2)に先立って処理しておく。

(3)は、実験の途中で経験的に加えられた規則である。書き替え規則中でたとえば K-prefix に属する大は K-unit 中でも大会などと使われるから、

## 大会論文募集

という漢字列では、どちらの大か判断できない。そこで大会は (3) の辞書に収められている。

この3つの規則によって漢字列をくぎる位置を検出し、したがって KWIC 索引の見出し語位置とすることができる。現段階では

接頭字 (K-prefix) の辞書—非, 不, 半など 20 字  
接尾字 (K-suffix) の辞書—器, 系, 的など 20 字  
に対して, (3)のための

保護処理用の辞書—機器, 系統など 47 語 (94 字) であるが, これ以前記の2つの論文誌の題名中の漢字列をほぼ完全に処理できた。誤まった例は安河内のような人名, さきのアウエルバッハ氏の氏の処理などであった。なお氏は上記の接頭字の辞書中に収めて応急の処置としたが, 安河内はいまだに正しくくぎれないで, 安河と内になる。

各辞書がどの程度の規模に収まるかは今後の課題であるが, これ以上急速には増加しないものと考えられる。ただしまったく分野の異なる論文誌を対象にするときは, いくつかの語を辞書に加えなければならないであろう。

## 3.4 ひらがな, カタカナその他の処理

科学技術文献中で, ひらがなで書かかれているが見出し語にするべきであるという場合は少ない。前記の2つの論文誌の題名のうち, 見出し語と考えられる部分がひらがなで書き始められていたのはわずかに次の3例であった。

## ため込み式, たたみこみ, あいまいさ

なお次の2例は原文では漢字が使用されていたが, 漢テレにこれらが含まれていなかったために, やむをえずひらがなになった。

## はい(排)気特性, けい(螢)光特性

すなわち KWIC 索引の見出し語を検出する場合, ひらがなの部分はあまり重要ではない。また助詞や助動詞も, ごく限られた種類のもの単純な組み合わせが使われるにすぎない。この代表的なものを Table 1 にあげた(助詞, 助動詞以外に一部の動詞も含む)。本システムではひらがなに関してはこれらを辞書に収めている。ひらがなの列は先頭の部分が促音であるか

Table 1 Hiragana strings frequently used in titles of technical literature

"に"の類:	における, について, によせて, により, による, に
"と"の類:	として, と
"の"の類:	のための, への, の
"する"の類:	する, して, した, された, させる (として)
その他:	が, の, を, から, より, で, て, その, および, まで







順序に並べてはじめて索引になる。漢字かなまじり文についてこの分類 (sorting) を自動的に行ない十分に自然な結果を得ることはきわめて困難である。本実験では使用した漢テレのコードによってのみ分類を行ない、それ以上の処理はなにもしていない。この漢テレのコードにしたがうと、

アルファベット、ひらがな、カタカナ、漢字の順になり、アルファベットはホレリスコードと同じ大小順序、ひらがな、カタカナはほぼ五十音順、漢字は音読みの順になる。音読みは1つの漢字に1つだけ与えられている。ここに示した索引の例 (Fig. 4, Fig. 5) はこのコード順そのままである。

さらに自然な順序を得るために、1つの漢字に1つの音読みと1つの訓読みを与えて使いわける実験を計画したが、送りがなの不統一 (組み合わせ⇔組合せ) まで考慮しなければならず、まだ着手していない。

この問題については 4), 5) などの研究が報告されている。漢字処理の自動化をはかる場合いつかは解決しなければならない問題である。

## 6. 印刷の形態

### 6.1 よこ書きとたて書き

この実験に使用した漢テレはよこ書き用に設計されていた。よこ書きでは1行に40ないし50字印刷できるが、KWIC索引の特徴である文脈の並記のためには不足であると考えられたので、索引の1行の左半分と右半分を別に印刷し、のちにのり付けして1行80字のよこ書き索引を作成することにした (すなわち索引の左半ページ分だけをまず印刷し、つづいて右半ページ分を印刷して、左右をつなぎ合わせる)。この作業には手間がかかる以外には問題はなく、ただ漢テレの行送り機構の精度が低いために左右に多少のズレが生じる点だけが未解決である。Fig. 4 はこうして作成したよこ書き索引を示す。

一方漢字かなまじり文を出力として使用することから自然に、たて書きの索引が着目された<sup>6)</sup>。これによれば (たての) 1行の長さ (字数) を自由に選べてもはり合わせの手間もいらぬなどの利点がある。この場合に生じるいくつかの問題点については次節でふれるが、Fig. 5 はたて書きの索引の例である。

たて書きの索引は、科学技術文献の索引としては時代に逆行するものだという感想が多く、現実の問題としてもアルファベットやローマ数字を題名中に使用することが多いので、不自然になる。しかし、もっと基

礎的な計量言語学への応用としては、このたて書きの KWIC 索引の考え方は広く利用されると思われる<sup>6)</sup>。

### 6.2 よこ書きとたて書きの変換

よこ書きの機械にたて書きの印刷をさせることは、単にプログラム上の行と列の交換の問題であると考えていたが、実際にはいくつかの問題点があって完全なよこ→たて変換は困難である。問題点を列挙する。

(1) ある種の記号、文字が必要であること

よこ書き用の記号 ( , ) , 「 , 」 , — , — などではそれぞれたて書きでは へ , へ , へ , へ , へ , へ などに変換しなければならない。コンマ ( , ) やピリオド ( . ) , 日本語の句読点、促音文字、よう音文字、その他上付きや下付きの数字・記号などはすべてよこ書きとたて書きとで使いわけないと見苦しい。

Fig. 5 のたて書き索引は漢テレに可能であった範囲でこのたぐいの変換を行なっている。

(2) 文字の大きさ、幅、印字間隔、行送り間隔

よこ書きとたて書きを同じ機械で行なわせると、印字間隔と行送り間隔が逆転して、よこ書きの行送り間隔がたて書きの印字間隔になり、活字の大きさ (天地) と幅 (左右) の異なるある種の漢テレではなおのこと、同じ字数のたて書きとよこ書きとでページの大きさがかなりちがってしまう。

(3) 数字 (0, 1, …, 9) やアルファベットの処理

数字を使う場合、よこ書きでは一般にアラビア数字を使うのがすう勢のようだが、これをたて書きにした場合どこまで、どんな表記の漢数字にするかは、いちがいにきめがたい (たとえば昭和 42 年度をたて書きで四二年度にするか四十二年度にするか)。アルファベットはたて書きでははなはだ見にくい、これはたて書きにすること自体不適當なのだと考えられる。

Fig. 5 のたて書き索引はこの点についてはなにも処置していない。

## 7. 入出力装置としての漢字テレタイプライタ

実験に使用した漢字テレタイプライタは当所情報基礎研究室が機械翻訳のための特別仕様<sup>7)</sup> によって発注したもので、とくにコードについて深い考慮がはらわれている。これまでふれてきたように本システムはこの漢テレに負うところが大きい。一般に電子計算機の入出力装置としての漢テレは、低価格で手軽に使えるところが魅力であるが、印刷速度が約 2 字/秒、新し



いものでも3字/秒なのでオンラインではとうてい使用できない。現在は紙テープによるオフラインシステムであるが、紙テープは計算センタにははなはだやっかいなしろものである。インクリメンタルレコーダを介する磁気テープによるオフライン処理が1つの方向だと思われる。印刷品質は Fig. 4, Fig. 5 にあるとおりで、あまりよくない。

## 8. 実験

プログラム言語; COBOL

データディビジョン 約 210 エントリ

手続きディビジョン 約 530 ステートメント

主要機器構成; FACOM 230-50 (32 kW)

漢字テレタイプライタ(谷村新興製,  
商品名漢字テレプリンタ)をオフラインで使用。

計算時間の例としては、情報処理学会誌1年間の全題目(148件)について索引作成に要する時間は3.5分、自家製の分類ルーチンによる分類の時間が4.7分、1031行の索引ができたが、それをオンラインで紙テープにパンチする時間が約40分であった(たて書きよこ書きとも)。これをオフラインで印刷するのに約11時間を必要とした。見出し語処理の精度についてはすでに述べた。

## 9. まとめ

漢字かなまじり文を対象に KWIC 索引を作成するシステムを報告し、あわせて日本語自動処理のいくつかの問題点にふれた。このシステムは比較的単純なプログラム方式によるが、十分実用的な KWIC 索引を作成できる。見出し語を自然な索引の順序に並べるこ

とが残された課題である。

終わりにご指導いただいた応用数学研究室木澤 誠室長に深く感謝する。

また漢テレ使用の便宜をはかっていただいた情報基礎研究室、ご討論いただいた西村 恕彦技官、計算を引き受けていただいた田無計算センタの諸氏、逆配列プログラムを作成された東京理科大学の田尻正滋君に心から感謝する。

## 参考文献

- 1) Luhn, H. P.: Keyword-in-Context index for technical literature (KWIC index), ASDD Tech. Report RC-127, IBM, 1959.
- 2) 橋本, 伊藤, 山崎, 笹森: 日本語キーワード索引システム, 第3回ドキュメンテーション研究集会発表論文集, 167-176, 日本科学技術情報センタ, 1966.
- 3) 現代語の語彙調査 婦人雑誌の用語, 国立国語研究所報告 4, 国立国語研究所, 1953.
- 4) 西村 恕彦: 文字列の配列順序についての問題, 情報処理, 10 (1), 21-25, 1969.
- 5) 田中章夫: 電子計算機によるワードリスト作成上の一問題, 国立国語研究所報告 31, 115-132, 国立国語研究所, 1968.
- 6) 植村俊亮: 日本語 KWIC 索引の試み, 第5回情報科学技術研究集会発表論文集, 143-152, 日本科学技術情報センタ, 1968.
- 7) 西村 恕彦: 電試漢テレ仕様案, 機械翻訳研究委員会資料, 情報処理学会, 1965.
- 8) たとえば  
坂本義行: Automatic Language Processing System—KWIC index file—, CL 研究委員会資料 69-5, 情報処理学会, 1969. など.

(昭和44年3月28日受付)