# 計算機による日本語文書校正のための基礎データの収集と解析

3B−6　その1　KWICによる特定表現の抽出とその結果

鈴木　恵美子　中山　妙子　吉岡　千草　山田圭以　福田恭子
東京家政学院筑波短期大学

## 1. はじめに

コンピュータは数値計算をするためのものだと思われており、実際大量の数値データの処理に威力を発揮しているが、私たちのゼミでは、将来的には計算機（ワードプロセッサ）により　日本語文書の校正処理を行うことを目的とし研究を進めている。それを行うためにはまず、大量の日本語文書の構成に関するデータと起こり得る誤りのデータを収集し、統計を取る必要があると考えている。

今回は、これらを行うための基本的なツールであるKWICのプログラムを開発した。また、これを使用して、特定の表現（論文や報告書によく使われる「〜に関する」、「〜に対する」）の出現状況やその表現の前後の単語関係を調べて、その結果を考察したのでそれについて述べる。

## 2. 大量の文書からの特定表現の抽出

図1 Volごとの総出現数(〜に関する)　図2 Volごとの総出現数(〜に対する)

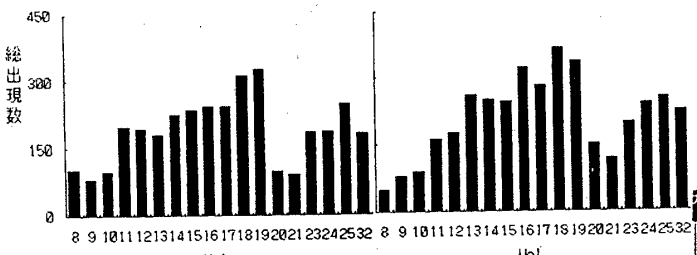


図3 Noに対しての総出現数(〜に関する)　図4 Noに対しての総出現数(〜に対する)

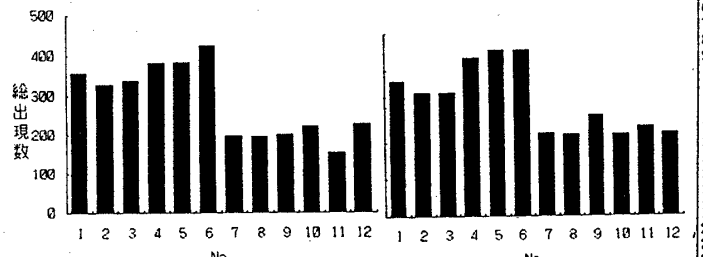


図1よりVol ごとにみた「〜関する…」の出現数は最高値がVol.19の326 、最高値がVol.9 の80で、範囲は246 である。

図2よりVol ごとにみた「〜対する…」の出現数は最高値がVol.18の367 、最高値がVol.8 の49で、範囲は318 である。

図3よりNoごとにみた「〜関する…」の出現数は最高値がNo.6の425 、最高値がNo.11 の152 で、範囲は273 である。

図4よりNoごとにみた「〜関する…」の出現数は最高値がNo.6の455 、最高値がNo.8の221 で、範囲は234 である。

それぞれ200 以上のひらきがあることがわかる。

「〜に関する…」が1つ出現するために必要な頁数は、Vol.8 で2.55ページ、Vol.21で5.62ページと3 ページ以上も差があるのに対し「〜に対する…」が1つ出現するために必要な頁数は、Vol.8 、Vol.10を除いては平均している。当初、この数字をもとに「〜に関する…」「〜に対する…」それぞれ5000例を収集することを目標として読み進めたが、物理的に入手できる論文誌の数も制限があったため、最終的に論文誌158 冊、「〜に関する…」が3398例、「〜に対する…」が3800例で、収集作業は終了した。

全体的に見ると、古い年代に発行された論文誌より、新しい論文誌の方が「〜に関する…」「〜対するの出現率が高いことがわかった。しかし、Vol.20 Vol.21では減少している。これは、雑誌の名前がVol.19までは「情報処理」だったが、Vol.20からは、「情報処理学会論文誌」と変わると共に内容が変化したためだと考えられる。

## 3. 入力ファイルの形態について

人手によって見つけられた、キーワードを含む一文をMS−DOSファイルにキーワードから入力してデータファイルを作成する。

データを入力する際の注意としは、

1. 全て全角で入力する
2. 公式は記号化して別に記録しておく

$ax^2+bx+c$ → （式A）

```
大小区別無効[MS
a:kan15.dat  [前画面] level-1                    [X001:Y00001]
1  スペース行列処理に関する通常のＧａｕｓｓ消去法の消去順序を変更すること↓
2  により、発生する非零要素数に関して、ほぼ消去の全域にわたり、かなり改良↓
3  されたアルゴリズムを提案した。↓
4  ↓
5  ここで、ＡＤＤＲは主記憶上の番地を、ＣＯＵＮＴはアトタイミングが真になる↓
6  るための回数に関する条件、いわゆる通過回数指定を表わしている。↓
7  ↓
8  ３．６　ステップ数に関するデバッグ手法。↓
9  ↓
10 分割の手続きＳｔｅｐ４の実行回数１に関する、次のような帰納法で証明され↓
11 る。↓
12 ↓
13 またもちろん、ｄ＝ｍ（ｘ）についてＱ（Ｇ）−Ｑ（Ｈ）であっても、数式４↓
14 ０５０となることがあるので、Ｑ（Ｇ）−Ｑ（Ｈ）をＧ，Ｈに対するハッシュ↓
15 値と見なおして、Ｇ，Ｈに関するデータレコードは順に並べておくことにする↓
16 。↓
17 ↓
18 領域解析法として、従来より種々の方法が提案され、最近では領域分割の段階↓
19 においても認識対象に関する簡単な事前情報（セマンティックス）を利用する↓
20 ことが考えられている。↓
21 ↓
22 本論文において、事前情報や種々の要因に依存する経験的な閾値等を用いない↓
23 領域解析法を提案し、本方法が領域分割（物体識別）に関する人間の視覚にほ↓
24 ぼ忠実であることを実験によって確めた。↓
25 ↓
26 ＤＦＡＭに新たに追加された機能は、（１）シーク時間、回転待ち時間とデー↓
27 タ転送時間、応答時間および転送データ長に関する統計情報の収集とヒストグ↓
28 ラムの作成、（２）ディスク・スケジューリング方式に基づく入出力要求のス↓
29 ケジュールである。↓
30 ↓
31 ＲＢ（Ｒｅｑｕｅｓｔ　Ｂｌｏｃｋ）は１個の入出力要求に関するすべての情↓
32 報を含む制御表である。↓
33 ↓
34 にもかかわらず、小さなプログラムに関するこれらの実験は、価値あるもので↓
35 あったばかりでなく、当時としては不可欠ですらありました。↓
36 ↓
37 その一つの理由は、この結びつきがかくも緊密である間は、たまたま存在して↓
38 いる計算機の特性が、好ましいものであるかどうかにかかわりなく、↓
39 プログラミングに関する思考を支配してしまう、↓
40 という傾向があったことでありました。↓
41 ↓
42 一方、第二の展開とは、ＮａｕｒやＦｌｏｙｄやＨｏａｒｅの、プログラムの↓
43 正しさを形式的方法で証明する可能性に関する論文によってもたらされた発見↓
44 とその応用に存するものでありました。↓
```

図5　入力ファイル

Colection of data and its analysis to be used for japanese proofreading by using the computer

emiko SUZUKI, taeko NAKAYAMA, chigusa YOSHOKA, kei YAMADA, yasuko FUKUDA
Tokyo Kaseigakuin Tsukuba Junior College

本研究は、財団法人日本科学協会の笹川科学研究助成によって実施したものです。

## 4. 特定表現の抽出方法

ここで、ＫＷＩＣのプログラムの流れについて述べる（図４参照）。

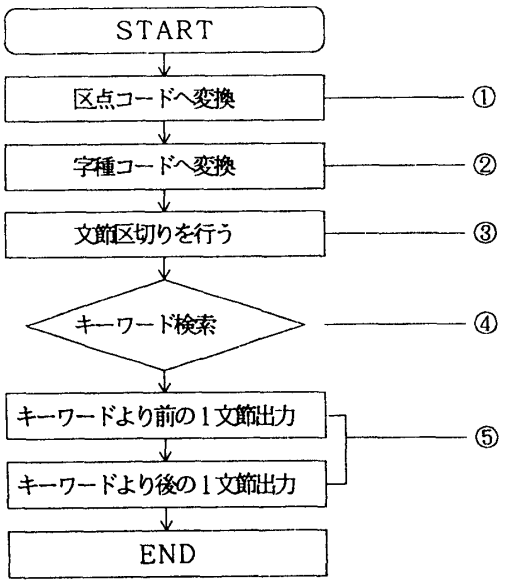


図6　KWICのフローチャート

２バイトコードのままでは、文節区切りの判断が煩雑になるため、①の部分で全ての文字を区点コードに変換する。

次に②の部分で区点コードを見て、字種別にコード化する。

表１　区点コード別字種コード

| | 区点コード | 字種コード |
|---|---|---|
| 記　　号 | 0102～0390 | 6 |
| ひらがな | 0401～0483 | 0 |
| カタカナ | 0501～0586 | 1 |
| 漢　　字 | 1601～9404 | 2 |

ここで、我々が必要としていたのは、キーワードの前後一文節のみであるが、後の処理のことを考えて③の部分で予め、文全体を文節区切りしておく。

文節区切りの基準としては、日本語の文章は普通ひらがなから漢字に変わる所で文節が区切れるので、それが出現したら文節として区切りとした。さらにその他にも文節が区切れると思われる場合（「私のコンピュータ」などは、ひらがなからカタカナに変わる時に文節が区切れる）についても文節区切りの条件とした。

私は犬と一緒に散歩へ行きました。

⇩　②により

２０２０２２０２２０２０００００６

⇩　③により

２０／２０／２２０／２２０／２０００００６

⇩　③時点の出力

私は／犬と／一緒に／散歩へ／行きました。

図7　文節区切りの例

④の部分でキーワード（ここでは「～に関する」と「～に対する」）を本文中から探索する。

⑤の部分で、予め処理③で区切られた文章のうち前後一文節だけを出力する。

## 5. 出力結果

下にＫＷＩＣのプログラムを実行した時の出力結果を揚げる

```
(キーワードは「～に関する…」)

スパース行列処理                ||に関する||通常のＧａｕｓｓ消去法の
回数                            ||に関する||条件、
３．６　ステップ数              ||に関する||デバッグ手法。
実行回数ｉ                      ||に関する||次のようなｎ
Ｇ．Ｈ対象                      ||に関する||データレコードは
認識（物体識別）                ||に関する||簡単な
転送データ長                    ||に関する||人間の統計情報の
入出力要求                      ||に関する||すべての
プログラム                      ||に関する||これらを
プログラミング                  ||に関する||思考
可能性                          ||に関する||論文によってもたらされた
正しさ                          ||に関する||初期のｎの
形式的証明                      ||に関する||初期の
構文                            ||に関する||形式理論の
演算）                          ||に関する||公理がありさえるれ
文法                            ||に関する||何ｎの
証明                            ||に関する||これらの
実行                            ||に関する||仕様とを
仕事                            ||に関する||私たちの
品質                            ||に関する||重要な
現象                            ||に関する||数学的ないし
製造」                          ||に関する||一つの
製造」                          ||に関する||こ
射影幾何学                      ||に関する||所要の
信頼性                          ||に関する||形式的手法を
ｄｅｑｕｅ                      ||に関する||アルゴリズム（９）ムに
置換行列                        ||に関する||情報を
マイクロプログラミング          ||に関する||成仕事は
マイクロプログラム              ||に関する||仕事をする
手法                            ||に関する||も
マイクロプロセッサ              ||に関する||も
分散処理                        ||に関する||も
使用経験等                      ||に関する||ユーザからの
生産技術                        ||に関する||国際会議。
生産技術                        ||に関する||２つ一般論、
（５）製造技術                  ||に関する||一般論
検査                            ||に関する||研究が
直接計測等                      ||に関する||研究
製造技術                        ||に関する||一般論は、
主題                            ||に関する||各種の
パブリック・ヘルス              ||に関する||セッションは、
システム・スクリーニング        ||に関する||システムの
診断                            ||に関する||セッションのう
計算機アーキテクチャ            ||に関する||もの
容量                            ||に関する||検討をｎ
自動解析                        ||に関する||研究がｎ
グ情報処理                      ||に関する||実際的手法について
```

図8　出力ファイル

## 6. おわりに

今後の課題としては、今回のＫＷＩＣのプログラムでは入力形式に制限があったので、今後なくしていくことと、文節区切りの精度を上げることである。

また、次に発表する辞書を使用して辞書引きできるようにしたい。

【参考文献】

[1] 千早　耿一郎　「悪文の構造」，木耳社
[2] 情報処理学会第４５回全国大会講演論文集
[3] 田中　　他　「日本文の自動分かち書き（ひら仮名文字列の分かち書き）」，情報処理学会第２０回全国大会講演論文集
[4] Borland international, Inc.「Turbo Pascal　リファレンスガイド」，（株）マイクロソフトウェア　アンシエイツ