

## テキストデータ処理用プログラムの開発

松 本 康<sup>\*</sup>

### 1. はじめに

「テキストデータ」とは、学習者の語彙や教科書用語のような、「ことば」の形のデータを意味する。本稿は、これらのテキストデータの計量的分析のための一つの道具として、筆者が開発したプログラムについて、その構成と使用手順を述べ、今後の展望を示すものである。

従来、社会科教育においては、教育内容や教材レベルでの議論は盛んに行われてきたが、実際に学習者が使う語彙や資料に使用される「ことば」の文析はあまり行われてこなかった。<sup>1)</sup>「ことば」の分析をおこなう場合、その意味内容、文脈、頻度等の多様な角度からの検討が必要となるが、最も困難な課題は、扱うべきデータ量の多さである。手作業による処理量の限界から、分析できる範囲が限定され、手法上も大きな進展は見られなかった。テキストデータの処理に計算機を使用することは、このような作業上の労力を軽減すると同時に、今までなしえなかった指標による分析への道を開くものである。ここで述べるプログラムは、テキストデータ処理用としては比較的単純なものであるが、一つの開発例として紹介しておきたい。

### 2 プログラム開発の経緯

筆者がテキストデータの分析を企図したのは1984年であるが、当時、既成の統計パッケージでは、日本語を含むテキストデータをローデータのレベルで処理できるものはなく、統計パッケージによる処理の前処理用のプログラムを自作する必要があった。この時は、単語レベルのデータとして入力されたテキストデータを、このプログラムによって一次処理した後に、統計パッケージで本処理にかける、という使い方を意図していたのである。

開発に着手した1984年当時、最初はパーソナルコンピュータ上のBASICによる対話型のシステムを考えていた。ところが、研究室の計算機（富士通FM-11）の記憶容量では大規模データの処理が難しいことと、日本語処理システムのセッティングがやっかいであることから、パーソナルコンピュータのシステムをあきらめ、大型計算機上でプログラムを開発することにした。当時の筑波大学情報処理センターの計算機（FACOM-M380）には、すでに日本語端末と、日本語入

---

\*筑波大学大学院博士課程教育学研究科

力が可能なユーティリティー PFD (Programming Facility for Display users) が備わっていたため、磁気ディスク上に直接データファイルを作成し、そのデータを処理するためのプログラムを書くだけで良かった。言語は FORTRAN<sup>2)</sup> を使用し、バッチジョブで走らせることにした。

その年の内に、連想語データ処理用のプログラム WASORT.FORT (Word Association data SORTing program) と、教科書用語分析用のプログラム TXTSORT.FORT (TeXT data SORTing program) の第1版ができあがった。どちらも基本的には同じプログラムであるが、入力データの形式や、印刷書式の違いなどから、ふたつのプログラムに分けた。

WASORT.FORT の第1版は、一次処理用の簡単なソーティング機能のみを持つものだったが、その後、必要に応じていくつかの分析用サブルーチンを加えてバージョンアップしてゆくうちに、一つのプログラム・パッケージに近い形式を持つようになった。現在は第4版で、1個のメインプログラムと16個のサブルーチンおよびブロックデータ部からなる、約1400行のプログラムである。TXTSORT.FORTは、現在第2版であり、1個のメインプログラムと5個のサブルーチンからなる、約300行のプログラムである。どちらのプログラムも現在は筑波大学情報処理センターの大型計算機 FACOM-M780 の OS の下で稼働するが、FORTRAN77が使える計算機ならば移植は可能である。以下、これらのプログラムの構成と使用手順を述べる。

### 3. WASORT.FORT の構成と使用手順

#### (1) 機能

WASORT.FORT は、学習者の連想語データの分析を目的とするプログラムで、いくつかの指標についての基本統計量と反応語のソーティング結果を算出・印刷する機能を持つ。さらに、これらの分析結果を他のデータファイルに書き出すことによって、SPSS,SAS等の統計パッケージによる二次処理をすることができる。現在分析可能な指標は、複数の刺激語についての、反応語数、反応語種類数、上位反応語、特定用語についての反応パターン、各刺激語間の類似度、孤立語、共通反応語等である。

#### (2) プログラムの構成

プログラムの構成は図1に示す通りである。( ) 内はサブルーチンの名称である。

「コントロール部」のMAINは、作業内容の選択と、作業進行のコントロールをおこなう。必要な作業は各サブルーチンに分割されており、MAINの中のCALL文によってサブルーチンが選択される。ここにはCALL文の中に結果を書き出す任意の媒体(磁気ファイル、磁気テープ、フロッピーディスク)をファイル番号の形で指定できるが、この時にはジョブ文の中にも詳細な指

定が必要である。

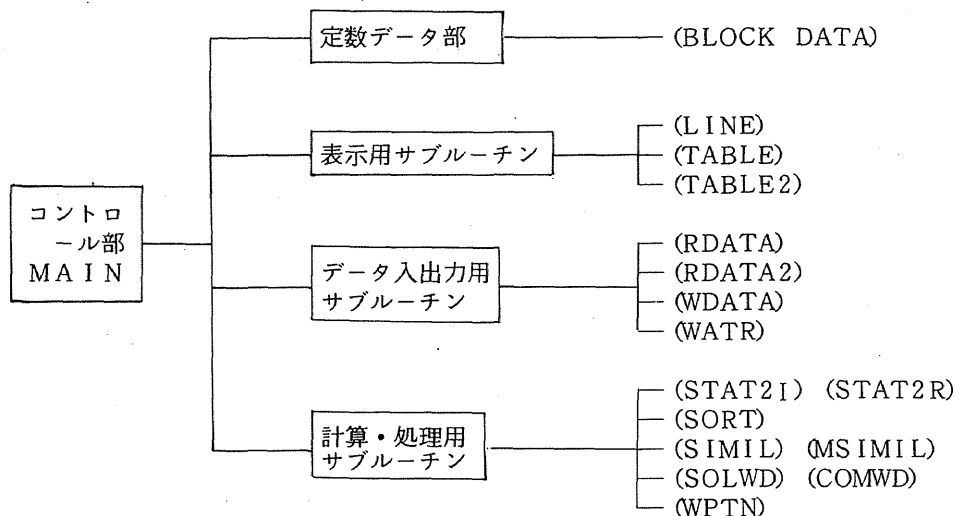


図1. 連想語データ処理用プログラム (WASORT.FORT) の構成

「定数データ部」(BLOCK DATA)では、作業上必要な定数の指定をおこなう。ここでは調査対象者数、刺激語数、最大可能反応語数、各サブグループの人数、刺激語(日本語データとして入力)の指定をおこなう。「表示用サブルーチン」は、印刷用の形式を決めるもので、コマンドに近い。LINEは横線を引くためのサブルーチンで、TABLE、TABLE2は、ソートされた結果を表形式に書き出すためのサブルーチンである。このうち、TABLE2はCOMWDの中でのみ使用される。

「データ入出力用サブルーチン」は、他のファイルを対象として、データの入力と、処理結果の書き出しを行う。RDATA、RDATA2は、共にデータの入力を行うが、RDATAは、刺激語の提示順序が特定の系列に整えられたデータを扱い、RDATA2は、刺激語の提示順序がランダム化されたデータを扱う。WDATAは配列形式に整えられた入力データの書き出しと、任意の処理結果の書き出しを行う。

「計算・処理用サブルーチン」には、現在8つのサブルーチンがある。STAT2I、STAT2Rは、数値型(整数型および実数型)の任意の2次元配列についての基本統計量を算出する。SORTは、刺激語ごとに反応語のソーティングを行い、上位反応語を印刷する。SIMILは、反応語のマッチングにより、刺激語間の類似度行列を個人ごとに算出する。さらにこの類似度行列が

ら、集団の平均類似度を算出するのがMSIMILである。MSIMILでは、全体の平均類似度の他に、サブグループ別（男女別、成績等の属性別）の平均類似度を計算することができる。SOLWDおよびCOMWDは、SIMILによる計算結果をもとに、それぞれ孤立反応語、共通反応語についての上位反応語表を作成する。WPTNは反応パターンの分析用のもので、特定のキーワードについての反応の有無を1-0型の反応パターンデータとして他のファイルに書き出す。このデータは、さらに統計パッケージによる分析にかけられる。以上述べたサブルーチンの他に、利用者は分析目的に従って、任意の新しいサブルーチンを付け加えることができる。

### (3) データ形式と処理手順

WASORT.FORTは単語を処理単位とする。入力データは、1レコードあたり30バイトの固定長ブロック形式で、1レコードが1データに対応する。30バイトのうちの10バイトは、編集用の行番号と区切り用の空白に使用するスペースであるため、実際には、1データの長さは20バイトである。漢字を含む日本語データの場合、2バイトが1文字に対応するため、日本語データとしては最大10文字までの文字列を一つの単語データとすることができる。1レコードに1データというのは無駄の多い形式であるが、これは主として、データファイル作成時のシステムの日本語変換機能の制約によるものである。ワープロで作成したデータを媒体変換すれば、80バイトのレコードに複数のデータを入れることができる。

データファイルには、第1行目に対象者の人数とコメントを書き、第2行以降に、個人ごとに属性データと反応語を入力する。個人データとしては、最初の行にID番号、性別、学年等の属性データを入れ、次の行以降に、反応語を1行1語として入力する。反応語は、調査対象者数×刺激語数×反応語数の3次元配列として読み込まれる。この際、刺激語数は同じであるが、反応語数には個人差があるため、一つの刺激語に対する反応語が終了した時には、終了を意味する区切り記号として、'XX'コマンドをデータ行に入力しておく。このコマンドを読み込むと、入力プログラム(RDATA)は次の刺激語に対する反応語の読み込みへと、制御を移す。データファイル作成から処理までの手順は以下ようになる。

- ① 磁気ファイル上にデータファイルを作成する。
- ② プログラムのMAINに作業内容を指定する。
- ③ ジョブ文にファイル、媒体等を指定し、バッチジョブとして起動する。

## 4. TXTSORT.FORTの構成と使用手順

### (1) 機能と構成

TXTSORT.FORTは、教科書用語の頻度分析用プログラムで、教科書中のテキストデータにつ

いて、WASORT.FORTと同様の分析をおこなうものであるが、WASORT.FORTの第2版に準拠しているため、頻度分析・ソーティング等のごく限られた機能を持つものである。プログラムの構成はWASORT.FORTと同様であり、データ入力用およびコントロール用のMAINと5つのサブルーチンからなる(図2)。そのうち計算・処理用のサブルーチンは、基本統計量の算出用のSTAT2D、ソーティングおよび反応パターン分析用のSORT1、およびソーティング専用のSORT2の3つである。

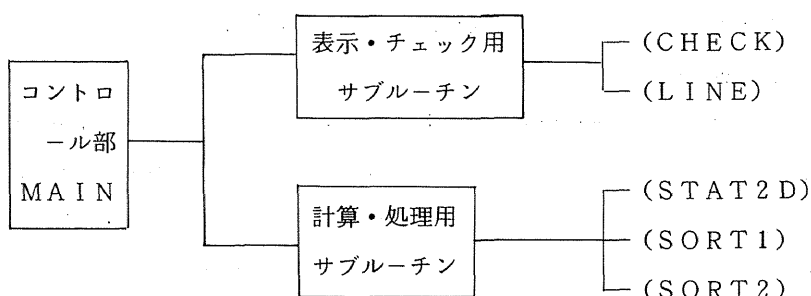


図2. 教科書用語分析用プログラム (TXTSORT.FORT) の構成

## (2) データ形式と使用手順

データ単位はWASORT.FORTと同様に、1データあたり20バイトの単語を単位とする。分析にあたっては、節－文－単語、の3レベルについての処理が可能である。データファイルの形式は、WASORT.FORTとはほぼ同じであり、データとなる用語は、節の数×節中の文の数×文中の単語数、の3次元配列に読み込まれて処理される。データファイル作成からプログラム実行までの手順は、3の①～③と同様であるので、ここでは教科書からのデータの抽出法について述べる。

- ① 分析対象とするテキスト部分に、節番号(通し番号)をつける。
- ② 各節ごとに、すべての文について、文番号をつける。
- ③ 各文ごとに抽出する用語を単語単位で選び出す。活用のある用語は、終止形とする。

こうして抽出された単語データが処理されるが、各用語は、①、②の情報とともに配列に読み込まれ、ソーティング、マッチング等の処理を受ける。普通の頻度分析の他に、用語分析の一方法として、各用語が同一文中、あるいは同一節中にどの程度一緒に使用されているかをカウントすることによって、用語間の類似度を算出することができる。このような指標によって、教科書に使用される用語についての一つの議論を展開することが可能であろう<sup>3)</sup>。TXTSORT.FORTは、WASORT.FORTほどの多くの機能はないが、今後、さらに新しい機能を付け加えてゆくことが可能である。また、両プログラムを統合して、より普遍的なプログラムに仕立ててゆくことも考

えられる。

## 5. おわりに

以上が筆者の開発したテキストデータ処理用プログラムの概要であるが、現在の問題と将来的な展望をまとめると、以下の3点になる。

- (1) もともと科学技術計算用の言語である FORTRAN で書かれたプログラムであるため、テキストデータ処理上の操作性はあまり良くない。また FORTRAN 独特のパラメータ指定をプログラムの各所に行わねばならず、FORTRAN に慣れていないユーザーには使いにくい。これに関しては、文字処理に適した他の言語<sup>4)</sup>への移植が考えられる。
- (2) 大型計算機用のプログラムであるため、センターの計算機使用の登録がしてあるユーザーでないと使用できず、手軽には使いにくい。より機動力のある使い方を可能にするためには、パーソナルコンピュータ上の、MS-DOS 等の OS でも使用できるようにプログラムを修正・移植する必要がある。
- (3) 基本的には単語単位のデータが同一であるか否かという単純な判断に基づく処理を行うプログラムであり、部分文字列処理および品詞の分類や文レベルでの意味分析等の機能はない。現在は単純集計に近い作業をおこなうものであるが、人間の文解釈に近い振る舞いをさせるに越したことはない。言語学の分野では、すでにパソコンのレベルでも日本語の構文解析が可能なプログラムが開発されており<sup>5)</sup>、プログラムの機能を高めることは技術的にはそれほど困難ではない。(1)、(2)の問題の解決ともかかわって、今後の課題である。

## 注

- 1) たとえば、教科書中に現れる用語を分析したものに、米本要三「中学教科書にあらわれる経済用語」社会科教育研究 No. 24, 1967, pp. 30-38, がある。また、学習者の使用する用語を対象としたものに、吉田昇「社会科における用語の使用概念の発達」教育心理学研究 Vol. 6, No. 4, 1959, pp. 238-243, 藤岡信勝「高校生・短大生にみる社会認識—社会科学用語の意味を手掛かりとした実態調査—」北海道大学教育学部紀要 25, 1975, pp. 61-80, 野口眞代「児童・生徒の連想語にみる社会認識の発達」お茶の水女子大学人文科学紀要 29-2, 1976, pp. 89-113, 松本康「『社会概念』の発達に関する基礎的研究」筑波社会科学研究 No. 6, 1987, pp. 24-35, がある。また、教育工学の立場から授業記録のデータベース的な扱いを試みたものに、大谷尚「パーソナルコンピュータによる授業記録分析システム」日本科学教育学会年会論文集 1983 年, pp. 81-82, がある。

- 2) 森口繁一『JIS FORTRAN 入門〔上〕〔下〕第3版』（1984）東京大学出版会。
- 3) このプログラムを使用した教科書分析の試みについては、以下の報告を参照されたい。松本康「教科書分析における計量的方法－用語の分析について－」（菱山謙二・松本康・松浦利隆「＜研究会報告＞コンピュータ利用による研究調査法」筑波社会科学研究№ 5, 1986, pp. 74-81 所収）。
- 4) PL/I で書かれたテキストデータ処理用のプログラム・パッケージとして、広島大学平和科学研究センターの松尾雅嗣が開発した LEX (program package for lexical analysis of a text) がある。これは大型計算機用の対話型プログラムで、汎用パッケージに近い操作性を持っている。松尾雅嗣「テキスト語彙処理プログラム LEX の開発について：概要と論理」広島平和科学 2, 1978/1979, pp.63-90.
- 5) 言語学者の草薙裕は、日本語の文構造を解析できる自然言語処理用の BASIC プログラムを開発している。草薙裕『パーソナルコンピュータによる自然言語処理』（1984）工学図書。