# 第4節　身の上相談研究のための若干の覚書（太郎丸）

## 1 問題

### 1.1 文書データの収集

文書データが社会学にとって価値があることは、かなり以前から広く認められている。しかし文書データをどのように選択、サンプリングすべきかについてはこれまで議論されてこなかったようである。もちろん研究対象となりうる資料をすべて１つ残らず読破し、さらに詳細に分析することが理想的かもしれないけれども、この場合、日本の有権者全員に質問票調査をする場合と同じ問題が生じるのは明らかである。つまり研究対象となる文章が膨大な量になる場合、１つ１つの文書の検討が不十分になってしまう可能性が高まるのである。このような場合、文書データも母集団となる文書群全体からサンプリングする必要が生じる。サンプリングすれば、検討する文書の数を比較的少量に押さえつつ、対象となる文書全体の特徴について推論することが可能となるからだ。本稿の第１の課題は、身の上相談の歴史的変化を調べる場合のサンプリングの方法について、私自身の成功と失敗を通して、若干の”コツ”を書き残しておくことである。

### 1.2 文書データのテキストファイル化

文書データをテキストファイル化することは、文書データの分析にとって必要不可欠な作業となりつつある。キーワードで簡単に検索できるというだけでもそのメリットは計り知れないほど大きい。統計的な解析に懐疑的な論者でも、フィールドノートやインタビューの内容をコンピュータを使って管理することには前向きである場合が多い(Pfaffenberger 1988, 佐藤 1992)。Microsoft Access のような汎用のデータベースソフトを利用すれば、より柔軟で高度なデータベースの構築が可能だし、質的データを管理するためのソフトウェアもすでにいくつか開発されている(Kelle 1995)。この場合文書のテキストファイル化は必要不可欠である。統計解析可能な定型的なデータを作るために、コンピュータコーディングを行う場合はいうまでもない。統計解析への批判は存在しても、テキストファイル化そのものを否定するような議論は、管見の範囲ではまったく見当たらない。つまりテキストファイルをどのように活用するかに関しては、さまざまな立場があるものの、文書データをテキストファイル化すること自体に関しては異論はないと考えていいだろう。それにもかかわらず、どうすれば効率的にテキストファイル化することができるかという技術的な問題は、社会学者によっては書き残されてこなかった。

そこで、本稿の第２の目的は、私が新聞紙上に掲載された身の上相談の内容をテキストファイル化する際に行った試行錯誤の結果を紹介して、テキストファイル化の技術の現状とデータの利用法を書きとどめておきたい。

## 1.3 資料の選択:読売新聞の「人生案内」欄

私の研究目的は、身の上相談に掲載された相談の内容の歴史的変化を調べることを通して、人々の生活世界の変化を明らかにすることであった。そこで読売新聞紙上に掲載されてきた「人生案内」欄をデータとして使用することにした。読売新聞は、日本でも最大手の新聞の一つであるだけでなく、第２次世界大戦前から継続して身の上相談を掲載している稀有の出版物でもある。一般の読者を対象にした身の上相談を分析する場合、おそらくこれ以上の資料はないだろう。

もちろん日本中のすべての人々が読売新聞を読んでいるわけではない。読売新聞を購読する人の中でも、すべての人が「人生案内」欄を読むわけではないだろうし、「人生案内」欄の読者も全員が潜在的投稿者というわけではない。また見田宗介(1965)も認めているように、何らかの事情で新聞を読まない・読めない人々が存在する。一部の「障害」者、乳幼児、その他にもさまざまな理由で新聞を読まない・読めない人々がたくさんいることを忘れてはならない。

このような難点を補うために、「人生案内」欄以外のさまざまな身の上相談も収集・参照すべきなのだろうか。実際そのような意見を私は何度か耳にしたことがある。しかし、そのような研究戦略は、少なくとも現時点においては採用しがたい。それには２つの理由がある。第１の理由は、データの解釈にかかわる問題である。今回の研究の目的は、相談の歴史的な変化を調べることであった。そのためには、資料は長期的に継続して出版されていなければならない。今回は 1934 年から 1994 年という長期間に渡ってその変化を調べたため、資料はその期間、継続的に出版されていることが望ましい。しかし現実にはそのような資料は私の知る限り読売新聞以外には存在しない。もしもアドホックに資料を収集するようなことをすれば、さまざまなデータの偏向を産むことは避けられないだろう。例えば、1934 年の資料として A 新聞と B 新聞を利用したが、1994 年の資料としては、A 新聞と C 新聞を利用したとしよう。しかしこの場合、1934 年の身の上相談の内容と、1994 年の身の上相談の内容に違いが発見されたとしても、それは時代の変化によるものではなく、B 新聞と C 新聞の編集方針の違いかもしれない。このようにアドホックに異なるタイプの資料を補うことは分析を混乱させるだけで、ほとんどメリットはない。

「人生案内」以外の身の上相談を利用しないもう１つの理由は、分析のコストが大きすぎるからである。テキストファイル化やコーディングの自動化が今後進めば、大量の文書データを扱うことは可能になるかもしれないが、現段階では、多大な時間と労力が必要である。「人生案内」を分析するだけでも困難なのに、その他の身の上相談まではとても手が回らないというのが現状である。

このようなことから資料の選択に関して教訓めいたことを言うならば、「手を広げるな！」ということになるだろう。対象を限定し、できるだけ同じ種類の文書を集め、その変化を調べるべきだろう。対象が限定されるため、当然議論できる範囲も限られるけれど

も、その範囲内では、厳密で強力な議論が可能になるだろう。

## 2 文書データのサンプリング

今回は、1934 年、1964 年、1994 年のそれぞれの読売新聞を約 100 強ずつ系統抽出した。サンプリングした日の新聞に「人生案内」が掲載されていない場合もあったが、特に予備サンプルを充当するようなことはしなかった。このような方針は各年ごとに最低 100 ケースはないと、年別の計量分析が難しいと考えたからである。また時間と労力の制約から、合計 300 ケース程度が限界と判断したのである。

しかし、このようなサンプリング方針は誤っていたと現時点では考えられる。むしろ、1934 年、1944 年（この年には人生案内は掲載されていないので実際には回収不能だが）、1954 年、……1994 年の新聞からそれぞれ 40 ケース程度、合計 320 ケース程度をサンプリングした方がより妥当性の高いサンプリングになっていただろう。もっと端的に毎年 5 ケース程度サンプリングし、全部で 60 年だから合計 300 ケースほどサンプリングしたほうがもっとよかっただろう。その理由は２つある。

30 年おきにサンプリングすべきではない第１の理由は、その年に起きた特定の出来事が一時的に相談に影響を及ぼす可能性があるからだ。例えば、1964 年は「妊娠中絶」への言及が目立つのだが、30 年おきのサンプリングでは、1963 年、1965 年に関しても同様のことがいえるのかどうかよくわからないのである。その時代の趨勢なのかもしれないし、1964 年に固有の特徴なのかもしれない。このような問題は、毎年少しずつサンプリングしておけば、すぐに明らかにすることができただろう。１年あたりのケース数は少なくなってしまうけれども、数年分のデータをまとめて１つのカテゴリーを構成するのは容易なことである。また、事後的にさまざまな時代カテゴリーを構成できるメリットも見逃せない。

第２の理由は、編集者の個性を排除するためである。1964 年に「妊娠中絶」への言及が多いのは、たまたまその問題に強い関心を持つ編集者が「人生案内」を担当していたからかもしれない。編集者はある程度、月日が経てば交代する。それでも一貫した変化の趨勢が発見できるのならば、それは編集者の個性の影響ではなく、社会全体の変化の趨勢を反映している可能性が高い。

私は、SSM 調査（社会階層と社会移動に関する全国調査。1955 年以来 10 年おきに行われている）に関わっていたため、長期間インターバルをおいてサンプリングを繰り返すことを当たり前と考えていた。しかし、新聞記事をサンプリングする場合、そのような必要はまったくない。毎年少しずつサンプリングすればよいのである。1 年あたりのケース数が少なすぎるならば、1960 年代の記事を１まとめにして分析すればよい。これは一度テキストファイル化してしまえば、簡単なことである。

# 3 文書データのテキストファイル化

サンプリングした文書をテキストファイル化するためには、２つの方法がある。文書を手でコンピュータに打ち込むか、OCR ソフトを使ってテキストファイルに変換するかである。手で打ち込むには多くの人手と労力が必要であるけれども、OCR ソフトは必ずしも信頼にたるだけの精度を持っているとは限らない。どちらがいいかは、資料として用いる文書の印刷の鮮明さ、文字の大きさ、スキャナや OCR ソフトの性能に依存する。今回、私が行った試行錯誤を紹介することで、現状（1997 年）の技術水準を書きとめておこう。

## 3.1 ネットワークからのダウンロード

最近の新聞記事はすでにテキストファイル化してあり、ネットワークを介して入手することができる。1996 年に電子メールで読売新聞社に問い合わせたところ、以下のような回答をいただいた。

> 現在、読売新聞記事のデータベースであるＹＯＭＩＤＡＳのオンライン・サービスは、下記の商用ネットワーク３社を通じて行われています。サービス内容、利用料金、使えるパソコン、ワープロの機種などは各ネットワークによって差があります。申し込み、問い合わせは直接各社にお願いします。
>
> ①ジーサーチ　営業部　０３－５４４２－４３８１
> ②日本経済新聞　データバンク局　０１２０－２１２－２１２（フリーダイアル）
> ③ＮＥＣ　ＰＣ－ＶＡＮ　０３－３４５４－６９０９
>
> **ＹＯＭＩＤＡＳには１９８６年９月以降の読売新聞東京本社発行の朝夕刊最終版のほとんどの記事の他、大阪、西部本社発行の朝夕刊と中部本社発行の朝刊の各最終版の主要記事、及び英字紙「デイリー・ヨミウリ」が収められています。内容は政治、経済からスポーツまで多岐に渡っており、収容件数は１００万件を越えています。提供データは記事タイトルとテキスト全文で、掲載年月日、朝夕刊の別などを示すアドレスがついています。検索は思い付いた言葉、つまり欲しい記事に関係のありそうな言葉で行えます。ただし記者の名前から記事を見つけることはできません。**

1994 年の記事は、これを使えば簡単に入手できたのだが、1994 年のデータを入力したときには、「人生案内」の記事まではアップロードされていないに決まっていると私が決めつけていたため、結局、ネットワークは利用しなかった。そのかわりに人海戦術でキーボードから入力することにした。後日、ネットワークを使えば簡単に「人生案内」の記事を入手できると、立教大学の村瀬洋一氏からも教えていただいた。

この私の失敗から言えることは、入力をはじめる前に入念なネットワークの渉猟が必要だということだろう。そのことによって飛躍的に作業量が軽減される可能性がある。特に大手の新聞社や出版社、政府の刊行物に関しては、チェックしてみる価値はあるだろう。

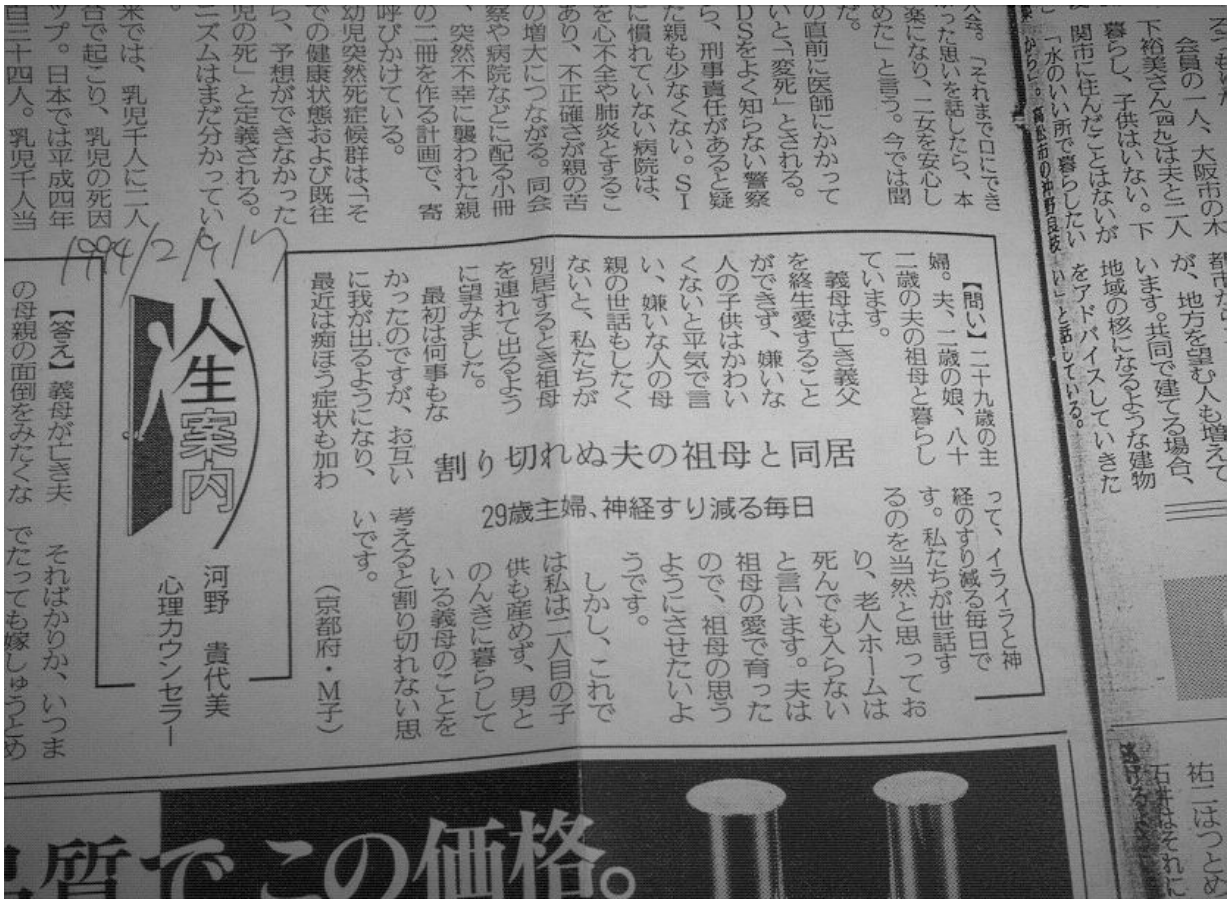
## 3.2 デジタルカメラの利用

1964 年と 1934 年の記事は、ネットワークにはアップロードされていなかったので、OCR

ソフトを使ってテキストファイル化する可能性を検討してみた。OCR ソフトを使ってテキストファイル化する場合、まず紙面を画像ファイルとしてコンピュータに取り込む必要がある。これには普通スキャナが利用される。しかし単に画像としてコンピュータに入力するだけならば、デジタルカメラを使うことも可能だ。つまりデジタルカメラで紙面を撮影し、コンピュータに送ればよい。もしもこれが可能ならば、テキストファイル化は非常に簡単になる。なぜなら、必要な新聞を所蔵している図書館にデジタルカメラを持って赴き、そこで必要な紙面を撮影して帰ってくれば、あとは OCR ソフトにかければテキストファイル化できることになるからだ。コピー代が必要ないし、コピーする時間と労力も必要ない。

そこでデジタルカメラで新聞紙面の撮影を試みた。しかし結論を言えばこの試みは失敗であった。利用したデジタルカメラは Canon Power Shot 600N という機種で、最大 832×608 画素での撮影が可能であった。、1997 年 6 月当時としては、かなり高級な部類に入るデジタルカメラで、白黒(B/W)文字モードという文書撮影専用のモードを持っていることに特徴があった。そこでこの文書撮影専用のモードを利用して 1994 年の「人生案内」欄の一部を撮影してみた。その結果が下の図である。

図 1 新聞紙面をデジタルカメラを使って撮影した結果



人生案内

【問い】二十九歳の主婦。夫、二歳の娘、八十二歳の夫の祖母と暮らしています。

義母は亡き義父を終生愛することができず、嫌いな人の子供はかわいくないと平気で言い、嫌いな人の母親の世話もしたくないと、私たちが別居するとき祖母を連れて出るように望みました。

最初は何事もなかったのですが、お互いに我が出るようになり、最近は痴ほう症状も加わって、イライラと神経のすり減る毎日です。私たちが世話するのを当然と思っており、老人ホームは死んでも入らないと言います。夫は祖母の愛で育ったので、祖母の思うようにさせたいようです。

しかし、これでは私は二人目の子供も産めず、男とのんきに暮らしている義母のことを考えると割り切れない思いです。

（京都府・M子）

## 割り切れぬ夫の祖母と同居

29歳主婦、神経すり減る毎日

【答え】義母が亡き夫の母親の面倒をみたくなるのはそればかりか、いつまでたっても嫁しゅうとめ

河野 貴代美
心理カウンセラー

この原稿が報告書の一部として印刷された場合、図 1 がどの程度鮮明に印刷されるのかは不明だけれども、パソコンの画面上で私が見る限り、何とか判読できる程度である。つまり人間が認識できる最低限の精度は得られるということだ。しかしこれを OCR ソフトでテキストファイル化しようとすると、ほとんど認識できなかった。5 パーセント程度の認識率だった。つまり OCR ソフトは、図 1 をほとんど判読できなかったということだ。

こうしてデジタルカメラでの入力は失敗に終わった。この試みの後 1 年半が経過し、デジタルカメラの性能は飛躍的に向上しているので、今後デジタルカメラで文書が取り込める見込みは、必ずしも暗いものではないだろうけれども、1997 年の時点の技術では不可能だった。

### 3.3 スキャナからの取り込み

次に 1994 年の記事のコピーをスキャナにかけてみた。スキャナを使って取り込んだ画像の OCR ソフトによる認識率は、99 パーセントで実用に耐えうる認識率である。ただし「人生案内」欄は記事のレイアウトが複雑なので多少手間がかかるが、それでも何とかなるといっていいだろう。

しかし 1964 年の記事はうまくいかなかった。周知のように新聞の印刷の鮮明さはここ 30 年の間にすこぶる向上している。1964 年の読売新聞の紙面は 1994 年のそれに比べるとどうしても不鮮明なのである。これが認識率の低下につながった。認識率は大体半分以下でとても実用に耐えるものではなかった。これもスキャナと OCR ソフトの性能が向上すれば、今後可能になるかもしれないので、決して悲観的になる必要はないだろうけれども、1997 年の時点で 30 年以上前の新聞に対してはまったく実用性はなかった。

### 3.4 キーボードからの入力

こうして結局キーボードから入力することになった。これは入力者のタイピングのスピードに依存するが、私の場合、1 時間にだいたい 3 つ程度の記事を入力することができた。したがって 100 個の記事を入力しようとすると、約 33 時間の作業が必要になる計算である。実際にはアルバイトを雇って入力してもらった。1964 年の場合は一般の学生に任せて問題なかったけれども、1934 年の記事は、旧仮名遣いで漢字も旧字体、しかも紙面の汚さはさらにひどく、まったく判読できない文字もあったため、信頼できる大学院生にお願いすることにした。入力された文書はコンピュータ上でもう一度読み直し、誤入力をチェックして、分析に使えるようなテキストファイルとなった。

## 4 おわりに

結局、キーボードから手で入力するという方法でテキストファイル化は行われたわけだ。しかしキーボードから入力することは悪いことばかりではない。元の文書を直接よく見る

ことで、他の記事や広告にも目が止まり、「人生案内」がどのような文脈で掲載されていたのかを知ることができた。また入力の過程で、元の文書を精密に読むことになったので、後のコーディングが若干楽になった。実際、私はコーディングの計画は文書をコンピュータに入力しながら考えていた。したがって、人海戦術を用いる場合は、分析者自身も入力の労をとったほうがいいとも考えられる。

ただし、それでも入力は自動化されたほうがよいし、今後そのような方向に進むことは間違いないだろう。しかし、1997 年の時点で 30 年以上前の新聞を OCR ソフトにかけようとした私の試みはことごとく失敗に終わったのである。

Kelle, U.(Ed). 1995. *Computer-Aided Qualitative Data Analysis: Theory, Methods and Practice*. Sage.
見田宗介. 1965『現代日本の精神構造』弘文堂.
Pfaffenberger, B. 1988. Microcomputer Applications in Qualitative Research. Sage.
佐藤郁哉. 1992. 『フィールドワーク　書を持って街へ出よう』新曜社.