「言語情報処理実習Ⅱ」について

文学部　伊　土　耕　平

1．はじめに

昨年本誌８号（1997年９月）に「『言語情報処理実習Ⅰ』について」なる文章（以下「前稿」と呼ぶ）を寄せたので、今年は続けて「Ⅱ」について書くことにする。平成９（1997）年度にどのような授業を行なったかについて、おもに報告する。

2．授業の概要

「言語情報処理実習Ⅰ」「同Ⅱ」は国文学科の開講科目で、平成８年～10年度は両方とも伊土が担当している。大まかに言えば、Ⅰは前期開講で、使用語彙の度数分布・類似度や相関係数の計算など、言語作品を数量的に分析する際の基本的な方法を身につけるのが目的である。Ⅱは後期開講で、Ⅰで学んだ方法を応用して、自分なりに何かの言語作品（文学作品や新聞の社説など）を分析するのが目的である。

Ⅱの内容について手っ取り早くご理解いただくために、平成９年度のシラバスを引用しよう。

言語情報処理実習Ⅱ　１単位　後期　２，３，４年次　専任　伊土耕平

［テーマ］言語作品からデータを採り、パソコンで処理する。

［概要］受講生が自分で選んだ言語作品（現代小説など）からデータを採り、パソコンに入力して、何かの処理を行ない、その結果にもとづいて何かを主張する。その処理には統合型ソフト＝クラリスワークスを使用する。

［授業内容］

９月　概説

10～11月　データ入力（各自が作業する。講義者が間違いなどを随時チェックする）

12月　データの検定・類似度の計算

１月　リポート作成・発表

［参考文献］石綿敏雄『文科系のためのコンピューター入門』（創拓社1994年）

DB - WEST編『パソコン国語国文学』（啓文社1995年）

［授業の評価方法］平常点と、リポートによる。

［履修上の注意事項］

各自フロッピーディスク（2HD，3.5inch）を２枚以上用意すること。クラリスワークスは、電算実習室のパソコンにインストールされている。

クラリスワークスの操作が出来ることを前提とするので、受講者は初回の授業までにマスターしておくこと。（ちなみに言語情報処理実習Ⅰではクラリスの操作法の指導も行なう）。

講義者のコンピューターの技能は初級程度である。高度なデータ処理を望む受講生の期待には応えられない。しかし、この実習で身につける程度の技能でも、国語学・国文学の論文を書くためのデータは作成できるはずである。

いい加減に作られたデータでは、それにもとづいて何か主張しても、説得力に乏しいことは言うまでもない。その意味では、一つ一つの用例をきちんと理解し正確に入力していくことが、重要である。そのような作業を経験することは、国語学・国文学の研究上、必要なことであると考える。

［テキスト］使用しない。

シラバスは以上である。ここには書かなかったが、９～10月にはデモンストレーションを行なった。つまり、パソコンを利用してどんな研究が出来るか、その具体例を示したのである。学生はそのデモを見るだけではなく、自分でも計算などをした。

３．平成９(1997)年度の授業

では実際の授業はどのようであったか。順を追って述べていこう。

３．１　受講者数

登録した者は５名である。うち２名は最初から最後まで一度も出席しなかった。つまり幽霊である。また、残り３名のうち１名は、最終段階で脱落してしまった。

なぜこんなに少ないのか。もちろん私の人気のなさが第一原因であろうが、この年に限り他学科生の受講を許可制にしたことも、影響しているのかもしれない。他学科生の受講は、結局ゼロであった。

この対策として、今年３月発行の『(国文学科）研究室だより』16号に宣伝文を書いてみた。その成果かどうか、今年度（1998年度）前期の受講生は14名に回復した。

３．２　デモンストレーション

学生に実際にデータ入力をさせる前に、私がこれまでに作成したデータなどを利用して、パソコンでどんな研究が出来るかを示すデモンストレーションを行なった。以下に述べる①～③のことである。

まず最初は、私が以前作成した『古事記』の接続詞のデータを使用して、①接続詞「故」「爾」「於是」の中では、どれとどれが意味・用法が近いか、②「作品の後半にいく

ほど故の使用率が高くなる」などの傾向があるかどうか、の二点を検証した。

①については、前期も行なったことで（つまり復習となる）、かつ前稿にも書いたので、ここでは省略する。

②を具体的に説明すると、まず『古事記』全体を十等分することを考える。私のデータでは、接続詞の一つ一つについて、「02808」などのように、所在情報もついている（この場合はテキスト28頁８行の意。テキストは西宮一民編『古事記　新訂版』桜楓社1992年）。全体が約200頁であるから、20頁ごとに区切れば、全体を十等分することになる。その区間を、１、２、３……10などとし（関数TRUNCなどを使用）、その区間ごとに「故」などの使用率を出し、両者をペアにして相関係数を計算することができるわけである。

途中の計算などは省略すると、結果は、「故」は－0.403、「爾」は0.155、「於是」は0.396であった。「爾」と比較すれば、「故」は作品の後半にいくにつれて使用率が下がる傾向があり、「於是」は、逆に上がる傾向があると言える。「爾」には、そのような傾向は見られず、まんべんなく使用される。

この結果はなかなか興味深い。と言うのは、『古事記』において、「爾」は言わばニュートラルな接続詞と考えられる。とすれば作品全体にまんべんなく現れると考えられるからである。それに対して「故」は、言わば強調のマーカーであるから、力の入った作品前半部に多く現れると考えられる。「於是」については、いろいろ考えられるが、説明は省略する。もっとも、数字のペア（区間と使用率）が10組しかないので、その中で相関係数を出してもあまり意味がないかもしれない。しかし一つのヒントにはなろう。

デモンストレーションの③は、前年度のこの授業で学生が作成したデータを利用して、「名詞の使用率と名詞の平均語長との間に相関がある」という仮説を検証した。どのように考えたのかと言えば、名詞の使用率の高い作家（作品）は名詞にある種のこだわりがあると考えられる。すると、使用する名詞が長めになるのではないか。例えば、単に「イギリス」という語を使うのではなく、「グレートブリテン」という長い語形を使うであろうというように（がさつな考え方であるが）。

前年度、学生が作成したデータは、『羅生門』『裸の王様』など七作品の冒頭から500レコード程度のものである。１レコードが「見出し語（かな表記）・見出し語（漢字表記）・品詞・所在」で構成されている。これらのうち、まず名詞だけを検索し、全体から見た名詞の割合を出す（＝名詞の使用率）。次に、名詞一語ずつの長さ（＝かなで何字か）を出し（関数LENを使用）、名詞全体で平均を出す（＝名詞の長さの平均）。両者をペアにして、相関係数を出すわけである。

結果は0.845で、かなり高い相関係数となった。よって先の仮説は支持される。

と言いたいところであるが、やはりデータが少なすぎる。②の場合は、数字のペアが10組だったが、こちらは７組と、さらに少ない。先の仮説が正しいと強く主張するのは無理

である。それでも、一つの手がかりにはなろう。

以上、三種のデモンストレーションを行なった。要するに、何を明らかにするためにどのようなことをするか、の例を示したわけである。学生は、これらを参考にして、自分なりにテーマを決定しなければならない。

以上の他に、カイ二乗検定なども行なったが、省略する。ちなみに、ここまでで授業四回分である。

3．3　対象作品決定とデータ入力

各自が別々のことをしていたのでは効率が悪いので、共同でデータ入力をすることにした。対象は、ひとりの学生の希望を取り入れて、太宰治の小説とした。太宰の作品の中でもどれにすればよいか？　まず、発表年代によってデータがどう変わるかという観点が考えられる。さらに、これは私の個人的な予想であるが、「私」「自分」などの一人称代名詞が後年にいくほど減っていくのではないか。それを確かめるのも意味があろう。すると、一人称代名詞で書かれた作品でなくてはならない。これら二点を考慮して、「思い出」(1933年発表)、「逆行」(1935年発表)、「富嶽百景」(1939年発表)、「東京八景」(1941年発表)、「人間失格」(1948年発表）の五作品を対象に選んだ。

学生一人が一作品を担当し、作品の冒頭から単語ごとに、見出し語などを入力していく。データの構成は、先述の昨年度のものと同様である。500レコード以上をノルマにした。先述のように学生は三人しかいないのであるから、私が二作品を担当し、データを補った。

この入力作業に授業四回を使った。途中、適宜私がデータをチェックし、最終的には全データに目を通した。間違いや不統一なものを正して、言わば“オフィシャル”なデータを一つ作成し、全員に配るのである（＝ファイルをコピーさせる)。このようにしないと、例えば同じことを計算しても答えが異なる、などの不都合が生じてしまう。

一人が500レコードというのは少なすぎると思われるかもしれないが、学生にやらせると、宿題を含めても、このくらいになってしまうのである。

3．4　リポート

全員にデータを戻して、各自がリポートを作成する。冬休みをはさんで提出されたリポートは、品詞の使用率をグラフにしたりして、いろいろと工夫されていた。

前年度と同じく学生作成のリポート集を作成することにしたが、数が少ないので、私が書いたものも合わせて『97年度　言語情報処理実習Ⅱリポート集』なるものを作成した。学生の執筆部分に対しては、修正すべき点などのコメントが書き込んであるので、学生は勉強になるはずである。

学生の書いたリポートを掲げるのははばかられるので、以下に、私の書いた部分をその

まま引用しよう。統計学的には問題となるところも多々あるのであるが、そのまま掲載する。なお、以下の内容は、先述の『研究室だより』と重複する部分があることをお断りしておく。

太宰５作品のデータから

１．感情形容（動）詞の使用率

形容詞（ここでは形容動詞も含む）には、感情形容詞と属性形容詞があることが知られている。前者の例：嬉しい、悲しい、など。後者の例：白い、広い、など。もちろん、両者をスパッと分けるのは難しい。古代語では、前者はシク活用、後者はク活用、などと多少形態的に区別できる。もっとも、これでも完全には分けられない（例えば「憂し」「つらし」は感情なのにク活である）。

それはともかく、感情形容詞の使用率に変化があるのではないか。年をとるにつれて人間は涙もろくなることがある。つまり、後期にいくほど、感情形容詞の使用率が高くなるのではないか。これが仮説である。

データは次のとおり。ただし、感情形容詞は「私」の感情を直接表すものだけとした。例：○不気味だ。○不安だ。×ましだ。×変だ。（後二者は“判断”）

作品	年	感形容詞	使用率	語数
思い出	1933	6	0.0119	503
逆行	1935	1	0.0019	526
富嶽	1939	5	0.0094	531
東京	1941	11	0.0221	497
人間	1948	18	0.0362	497
平均	1939.20		0.0163	
相関係数	0.884			

感情形容詞の認定に多少不安は残るものの、0.884という高い相関係数を得た。よって、年をとるにつれて、感情形容詞の使用率が高くなるという仮説は支持された。

と考えるのは早計である。上の事実は、たんに形容詞の使用率が高くなることと連動しているだけのことかも知れない。そこで、その点を検証してみる。

感情形容詞と属性形容詞の個数のペアは次のとおりである。

	思	逆	富	東	人	
感情：	6	1	5	11	18	：計　41
属性：	32	29	46	38	46	：計 191
計	38	30	51	49	64	

この分布に有意差はあるか？　カイ二乗値を出してみると12.08となり５％水準で有意である（自由度４）。よって分布に偏りがあると考えてよい。

２．「私」の使用率

これら５作品は、「私」とか「自分」とか、一人称の言葉を使って語り手が語るという構成をとっている。この「私」（以下「自分」も含む）の使用率にも変化があるのではないか。と言うのは、一般に、若い頃は自意識が過剰である。年をとるにつれ、過剰さが中和されて、言わば、落ちついてくる。とすれば、「私」の使用率も少なくなっていくのではないか。これが仮説である。

データは次のとおり。「私」の使用率は、その作品の、「『私』の数／名詞の総数」である。

「私」数	名詞数	発表年	使用率
30	258	1933	0.116
33	286	1935	0.115
25	257	1941	0.097
16	270	1939	0.059
7	218	1948	0.032
	平均	1939.2	0.084
	相関係数	−0.875	0.084

これも、−0.875という高い相関度を示した。仮説は支持された。

３．おわりに

以上、二つの考え方を示した。何かの参考にしてほしい。

もっとも、今回の分析は実験的なものであり、上の仮説が本当に正しいかどうかは、データの量を増やすことはもちろん、太宰という人をもっと調べたり、他の作品を分析したり、さまざまのことをしたうえで判断されるべきである。今回はパイロット（水先案内人）的なものである。

今回提出されたリポートを見て一番言いたいことは、“もっとしつこく考えよ”と

いうことである。そのためには、データを加工してみることも大切である。結局、

データを見る（加工も含む）⇄仮説を考える

という二つのことの相互作用が、研究の基本である。

また、考えのヒントを得るためには、やはり読書が大切である。参考文献のリストなどを見て、いろいろな本を読んでほしい。上に示した感情形容詞／属性形容詞のことなどは、じつは常識的なことなのである。

私の“リポート”は以上である。いろいろと問題があることは承知している。最大の問題は、5作品の間で相関係数を出すのは、やはりほとんど意味がない、ということであろう。統計学に詳しいある教育心理学者のご教示によれば、このデータを無相関検定してみると、5％水準で有意ではないとのことである。ちなみに、最近個人的に、太宰のもう2作品からデータを採って合計7作品にして、「私」の使用率に関して相関係数を出してみた。結果は－0.620で、上述の場合に比べて、だいぶ下がってしまった。もっと調べればもっと下がるかもしれない。

しかし、上述のようなデータしかないのだから仕方ない。このように言うと、日頃厳密な統計学的調査を行なっている先生方からはお叱りを受けるだろう。それは甘んじて受けるしかない。しかし、やはり国文学国語学の研究においては、単語切りから最後の相関係数などの計算まで、一通り経験することが大切であると考える。それを半期でやるとなると、どうしても上述のような程度のことになってしまうのである。

つまるところ、結果よりも考え方が重要である。とくに私の場合、学生に教えられるのはパソコンの使用法や統計学よりも、考え方の方である。（したがって、上記のリポートの中で学生に特に言いたいことは、「3．おわりに」の部分である。）

もっとも、上記の考え（とくに仮説を考えるところ）は、あまりにもがさつであると思われるかもしれない。本当のところを言えば、“考え”以前に“直観”があるのである。今回のデータを眺めて、「私」の使用率が下がるなどと直観したわけである。しかし、直観を人に教えることはできない。

3．5　評価など

リポートの出来と平常点（おもに出席数）とで評価をした。

4．おわりに

以上が平成9年度の「言語情報処理実習Ⅱ」の授業内容である。私のコンピューター技能は原始人レベルで、あまりたいしたことはしていない。恥ずかしいかぎりである。国文学科の新任の滝川氏はコンピューターに強いので、来年度からは交代してもらいたい。

この科目を持っていて一番悩むことは、直観や思いつきを教えるのは不可能だということである。教員がいろいろなことをやって見せるしかないのだろうか。類似度や相関係数などの手法を教えても、学生はそれらを使って何かするということができない。簡単に言えば応用力がない。

それでも、学生はそれなりに興味を持ったらしく、二人ともこのようなことを卒論などでやってみたいと言った。私に対する気配りにすぎないのかもしれないが、嬉しいことではある。

なお、平成10（1998）年度から実習室の機種が変更され、ソフトもクラリスワークスからオフィス97になった。この変更によって授業内容を大きく変えることはない。ただ、作業がだいぶ楽になった。例えば、ある言語作品中の単語の数を数えるのに、クラリスワークスでは、データベースで「小計パート」を作ったりして、大変ややこしかったが（他にもっとよい方法があったかもしれないが)、オフィス97の中のエクセル97では、「ピボットテーブル」を少しひねって使えば、簡単に単語の個数を数えることができる。また、類似度を出すのも、ピボットテーブルを使えばかなり楽である。

以上、情報公開の一環として、「言語情報処理実習Ⅱ」の授業内容について述べた。ご意見などをお寄せいただければ幸いである。