「言語情報処理実習Ⅰ」について

文学部　伊　土　耕　平

1．はじめに

平成8（1996）年度から、新カリキュラムの一部として、「言語情報処理実習Ⅰ」「同II」なる科目が開講された（ともに半期、1単位、2年生以上対象)。これは国文学科の開講科目であり、多少はパソコンがいじれるということで、私（伊土）が担当することになった。このあたりからすでに、怪しい雰囲気が漂っている。

本稿は、これらのうち「Ⅰ」について、平成8年度にどのような授業を行なったかを報告するものである。国文学科という、およそコンピューターとは縁がなさそうな学科の者がどのような授業を行なったか、興味がおありの方もあろう。

以下に、この授業の内容や学生の満足度アンケートなどについて、述べていく。ご意見などお寄せいただければ幸いである。

2．授業の内容

2．0　「言語情報処理」という言葉

本論に入る前に、この授業の名前に使われた「言語情報処理」という言葉について、一言しておきたい。これはおそらく「情報処理」という言葉があって、それに、言語に関係するという意味で「言語」と付けたのであろう。「であろう」と言うのは、私はこの命名に一切関与していないからである。

この「言語情報処理」という言葉は、私はあまり聞いたことがなかったが、それでも先日、難波の古書店で『言語情報処理』という本を見つけた。この本の発行は古くて、1981年である（西田富士夫著、コロナ社、「情報工学講座」の一冊)。もっともこの本は、英題を“Fundamentals on Language and Logic Processing”と言い、コンピューターでデータを実際に処理することより、述語論理などの解説が主である。

また、昭和61（1986）年度の科学研究費の特定研究に「言語情報処理の高度化のための基礎的研究」というのがあり、それに関連して雑誌『日本語学』が、1987年の5月に「言語情報処理の言語学」という特集をしている。この研究には多数の高名な研究者が参加しているから、影響力は大であろう。このあたりから「言語情報処理」という言葉が普及したのであろうか。

以上、「言語情報処理」という言葉は、少なからず使われている。先に「あまり聞いたことがなかった」と書いたが、それは単に私が勉強不足であっただけである。もっとも、

最近よく聞く言葉は「（自然）言語処理」という言葉である。「言語処理学会」という名の学会もあるし、同学会は『自然言語処理』という名の雑誌を出してもいる（本学の図書館にもある)。もっとも、この学会は、かなり工学寄りである。また最近、『自然言語処理』という分厚い本が出た（長尾真編、岩波講座ソフトウェア科学15、1996年)。

とにかく、言語に関するデータをパソコンで処理する、という内容であればよいのだと私は理解した。新カリキュラム作成の国文学科の中心人物であった木村先生に「簡単なことならできます」と答えたことから、この科目の担当者は、私に決ってしまった。

２．１　授業の概要

さて、授業の内容であるが、概要をご理解いただくために、シラバスを掲げよう（一部省略した)。

科目名：言語情報処理実習Ｉ　単位：１　期間：前期　配当年次：２，３，４

テーマ：パソコンによる言語データ処理。

概要：統合型ソフト＝クラリスワークスを使用し、言語データの初歩的な処理を実習する。

授業内容：

４月　クラリスワークスの操作（表計算・検索・ソート・データベース・集計）

５月　統計的処理（度数・累積度数・分布）

６月　類似度の計算（使用率の比較・最小値）

７月　リポートの作成

参考文献：宮島達夫「語いの類似度」（『国語学』82号、1970年）、同「総索引への注文」（同76号、1969年）

履修上の注意：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。その他に用意するものはない。クラリスワークスは、電算実習室の各パソコンにインストールされている。データは、講義者が作成したものを使う。

クラリスの操作方法については、何回かにわたって指導するが、それ以後は各自で練習して身につけること。

講義者のコンピューターの技能はせいぜい中級程度である。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、国文学の論文は十分書けるはずである。

評価は、出席回数と、リポートの出来具合による。

テキスト：使用しない。

上のシラバスには大きな間違いがあった。「講義者のコンピューターの技能はせいぜい中級程度」というところである。「中級」にも及ばないことを、この授業で十分に思い知らされた。

実際の授業も、だいたい上のように進行した。中でも重点を置いたのは、「類似度」である。これは、例えば『万葉集』と『古今和歌集』を比較して、語彙の割合構成がどのくらい近いか、を数値で出すものである[1)]。数値は１から０までの値をとり、「１」であればまったく割合構成が同じ、「０」であれば共通して使われる語がまったくない、という意味になる。この「類似度」は、国文学の研究上いろいろなところに応用できる便利な数値であり、学生にぜひ学んでほしいものである。

さらに、シラバスには書かなかったが、「相関係数」もとりあげた。これは、例えばある一人の作家を取り上げたとして、例えば、「執筆年」と「助詞の使用率」の間に相関関係があるかどうか、つまり、年齢が高くなるにつれて助詞の使用率も高くなる、などの傾向があるかどうか、を調べるものである。この係数は、－１から１までの値をとる[2)]。この「相関係数」も、国文学研究ではなかなか有用である[3)]。

以下に、より具体的に、細部にわたって、授業内容について述べていこう。記述の順序は、時間の流れに従う。

２．２　準備―平成８（1996）年３月―

授業の具体的な内容に入る前に、枕として、“準備”について少し述べよう。じつは、私はクラリスワークス（以下「クリラス」と言う）はそれまで使ったことがなかったのである。それどころか、統合型ソフトなど、さすがに聞いたことはあったが、見たこともさわったこともなかったのである（ここで国文学科がいかに遅れているか知れよう)。クラリスを情報処理センターのパソコンに積むことが決まったのは、いつだったか忘れたが、そのときは「なんとかなるだろう」くらいにしか思わなかった。

ちなみに、私が自宅で使うのは「Ninja4」である（！）。Ninja4はデータベースのソフトで、あまり複雑な計算はできないが、シンプルで、なかなか使いやすい。私はどちらかと言えば、一つ一つの用例を重視するほうで、統計学的な手法を駆使するというタイプではない。だからNinja4でも十分なのである。ただし、カイ二乗値を計算するときなどは、さすがに不便だなと思う。

さらに言えば、Ninja4にできて、クラリスにできないこともあるのである。「ファイル連結」「同除」「一括更新」などである（などと言い切ってよいのだろうか)。

話をもとに戻そう。結局私は、授業開講の１ヵ月前、具体的には平成８年３月に、集中的に練習してクラリスをマスターするという作戦を立てた。その昔、ワープロを習得したときも１ヵ月くらいであったし、ＮＨＫＦＭラジオの「日曜喫茶室」という番組でだいぶ

前に、司会のはかま満緒が「１ヵ月くらい集中してやればパソコンはマスターできる」と言っていたのを思い出しもした。

正直言うと、マウスなるものも、このころ使い始めたのである。このような情けない状態ではあったが、春休みに何日も大学に通った成果が出て、１ヵ月でなんとか一通りはマスターできた（自宅のパソコンはウィンドウズが積めるほどパワーがないのである)。ときどきはクラリスのテクニカルサポート（電話で疑問に答えてくれるサービス）も利用したが、私の疑問の解消率は50%くらいであったろうか。自慢ではないが、テクニカルサポートで“そのような作業は、クラリスではできない”と言われたことでも、自分なりに工夫してできたことが何回かある。あまり他人を頼りにしてはいけない。

２．３　幻の人数制限

いよいよ４月になった。私はひそかに、この授業は受講希望者が多いだろうと思っていた。しかし一つのクラスで面倒みられるのは20人くらいまでである。そこで人数制限をすることにして、講読か何か忘れたが、クラス分けがあったときに、クラス分けの前に、受講希望者に集まってもらうことにした。

当日現れた学生は、たったの一人であった。これにはさすがに拍子抜けがした。宣伝不足と、私の人気不足が原因であろう。この結果、人数制限の必要はなくなった。

あとから考えてみると、この授業は他学科の学生が多く受講したので、国文だけを対象にして人数制限をしても意味がなかったのである。それにしても、この事態をどう考えるべきか？　人数制限（をするほどの人気）というのは幻想であった。

２．４　開講時の様子―４月―

結局この授業に登録した学生は全部で16人であった。その内訳は以下のとおりである。

学　科	２年	３年	４年	計
国　文	2	3	4	9
史　学	1		2	3
地　理			1	1
文　財		1		1
社　会			1	1
産　社	1			1
計	4	4	8	16

いろいろな学科・学年の学生が受講した。ちなみに、女子は２名であった（国文３年と４年に一人ずつ)。

シラバスにもあるように、4月はクラリスの操作方法について簡単に説明した。しかしこれは意外に難しいことである。その最大の理由は、学生の技能のレベルがさまざまであることだ。学生の中には、キーボードに触るのは初めてという者から、私などよりよっぽどレベルの高い者までいた。このような集団に教えなくてはならないのである。ほとんど絶望的であった。無論、私の教え方に問題がある、という面もある。

シラバスにも書いたのだが、クラリスの操作についてあまりたくさんは教えないので、技量が不足している者は自分で自習してほしかった。しかし、実際に自習した者はあまりいなかったようである。何人かの学生は、この段階で脱落していった。

私自身、先に述べたように、自力で習得したのであるから、いまどきの学生に出来ないはずがない。脱落していった学生は努力不足である。パソコンに限らず何でも、最初のうちは失敗したり、ひどい目にあったりする。そこを通り越して技術が身につくのではないか。最近の学生は甘えているように思えてならない。中には、私の『一週間でマスターするクラリスワークス』（このタイトルは違うかもしれない）というマニュアルを借りたまま、来なくなってしまった学生もいる。いまだに返してもらっていない。

クラリスには、ワープロ、表計算、データベース、グラフィックスの四機能がある。それらすべてを熟知するというのは大変である。国語国文学の研究に有用なのは表計算とデータベースであるが、学生がどこまでその有用性を実感できたか、心もとない。

2．5　度数分布―5月―

一応クラリスの操作はマスターしたとして、5月には、度数を数えることと、その分布をグラフにすることを行なった。例えば、新聞の社説の電子化テキストについて、単語ごとに切り、「原発」が何個、「事故」が何個と、単語の個数（＝出現度数）を数えるのである。単語切りはワープロ画面で手作業で行なったが、度数を数えるのは、ワープロ書類をデータベース書類に変換したあと、自動的に集計できる。

単語切りが自動的にできればかっこいいのであるが、これは大変難しいことである。コンピューター自身が、「辞書」とやりとりをしながら、切る範囲を考えなくてはいけないからである。以前、大阪松蔭女子大学でそういうシステムを実際に見たことがあるが、私の力量では到底作ることができない。結局、ワープロ画面で、単語の切れ目ごとに改行していくという、原始的な方法を取らざるをえなかった。それをデータベースに変換し、集計機能を使って、「原発」が何個、「事故」が何個と数えるわけである。

しかし、人間が単語切りをするのも、じつは難しいことなのである。私は昔、国立国語研究所で『国定教科書用語総覧』の編集作業をしていたが、そのときも単語切りには悩まされた。例えば、「十キロメートル四方」という言葉は切るか切らないか？　「奈良大学付属図書館閲覧係」はどうか、等々々。国研のその研究室では実に細かい規則を作って対

応していた。この授業ではそこまでできず、不徹底で曖昧な部分が出てきてしまった。

いっそのこと切れるだけ切る、というのがもっとも簡単で、わかりやすいのであるが、これではまた別の問題が生ずる。例えば、ある文章に「現実的思考」という言葉が多用されていたとする。この言葉はその文章の内容的・文体的なキーワードになりうる。これを「現実的」と「思考」に切ってしまうと、この文章の内容的・文体的な把握は難しくなってしまうのである。

ともかく単語切りはできたとして、次に単語の数を数えなくてはならない。この“数を数える”というのは、数量的研究において、もっとも基本的なことである。この作業をNinja4でしようとすると、大変時間がかかる。その点クラリスのデータベースの集計機能を使えば、あっというまである（もちろんデータの量にもよるが)。これには感動した。

次に度数分布を調べた。度数分布にもいろいろなものが考えられるが、例えば、ある単語の使用度の順位を横軸に、その語の使用度数を縦軸にとって、グラフに書いてみるとどうなるか。（当然、右下がりのグラフになる。）

クラリスには自動的にグラフを書く機能があるが、トラブルが多かった（私はどうもパソコンで図を書くのが苦手である)。それ以前に、“何をするのか”を学生にうまく説明できなかったこともあり、この作業は、全体的によくなかった。

２．６　類似度と相関係数―６月―

６月には、まず、「第６章までのまとめと補足（Ｑ＆Ａ）」なるプリント（Ａ４サイズ、４ページ）を配布した。「第６章」というのは、それまでに配布したマニュアルの章である[4]。マニュアルは懇切丁寧に記述したつもりであるが、操作手順しか書かなかったので、学生の理解はいまひとつであった。そこで考え方を説明したものを作成・配布したのである。データの構造から「ソート」、クラリスで度数を数える仕組みなどと、かなり丁寧に書いたつもりであるが、あるとき、ある学生から「ソートって何ですか？」と聞かれ、力が抜けた。

次に類似度の計算を実習した。具体的には、例えば次のようなものである。『古事記』には「故（かれ）」「尓（しかして）」「於是（ここに）」という接続詞が多用される（合計800回ほど)。これら三者のうち、どれとどれが意味用法が近いか。

じつは、このことに関するデータを私はすでに作成しており、計算結果などは公表している[5]。そのデータを学生にコピーさせて、使った。

もう少し具体的に言うと、まず、このデータでは、接続詞の後件を36種類に分類してあり、それらの後件ごとに「故」「尓」「於是」の使用度数を出してある。後件は、例えば「誰それが誰それと結婚した」とか「○○天皇が天下を治めた」など、内容によって分類したものである。次に、「故」「尓」「於是」を『万葉集』や『古今集』などの作品に見

立て、後件の分類の一つ一つを「あはれ」「いも」などの単語に見立てて、類似度を計算するわけである。もし、例えば、「尓」と「於是」の意味用法が近いのであれば、両者とも同じような文脈（≒後件）で使われるはずであり、36個の後件の割合構成も同じようになるはずである。

ちなみに、この計算には、足算や割算はもちろんのこと、IFやLOOKUPなどの関数も使わなければならない。少し複雑である。

この類似度は便利なので学生にぜひとも知っていてほしいことだったが、あまり理解されなかったようである。私の作成したデータの意味が理解されなかったのが、ネックだったようである。そもそも『古事記』の接続詞のことなど、学生は知らない。類似度計算の実習は、失敗であった。

次に、相関係数の計算を実習した。村上征勝『真贋の科学―計量文献学入門―（行動計量学シリーズ６）』（朝倉書店、1994年）に、例題として、日蓮の著作24編における著作年と助詞の出現率の間の相関係数を計算しているところがある（p.31)。そのデータをそのまま学生に転記させ、計算をさせた。この計算は比較的容易だったようである。計算方法などは省略するが、ちなみに答えは0.68で、両者の相関度は高い。

２．７　締めくくり―７月―

実習も最後に近づいたということで、もっと実のあることをしたかった。ちょうどその頃、史学科の水野柳太郎先生から、『日本書紀』の電子化テキストが出来たので見に来いと言われた。拝見しに伺うと、先生はそのデータのコピーを下さった。そこで、このデータを使って、『日本書紀』の巻ごとに漢字の使用率を出し、各巻の間で類似度を出してみることにした。

学生一人に一つの巻を担当させることにした。各巻の冒頭から1000字の漢字について、ワープロ画面で一字一字切り離し、データベースによって漢字の異なりの一覧と一字ごとの使用度数を出し、表計算で使用率を出したあと、他の巻のデータと比較して、類似度を出した。冒頭から1000字だけという不完全なデータであるから、結果にはあまり意味がないが、一応記しておくと、どの組をとっても（巻一と巻二、巻二と巻三などの組）、だいたい0.45くらいの類似度であった。

ちなみに、『日本書紀』の巻ごとの用字法に関しては、山田英雄「日本書紀各巻の成立と助字法」（『史学雑誌』63の２、1954年）という有名な論文がある。これは、漢字すべてではなく、助字の漢字だけを使って巻ごとの異同を論じているものである。結論としては、「巻四、五は前後と全く異なり、巻六－十五までは（巻八を除く）は略同じく、巻十六は又孤立し、巻十七、十八は略同じく」などと書かれている。この結論と先の類似度による結果とが合致すれば良いのであるが、そう甘くなかった。

最後の締めくくりは、評価である。シラバスにはリポートで評価すると書いたが、「Ｉ」では技術が習得できたかどうかが大切であると考え、リポートはやめにした。この時期には、どの学生がどの程度のことができるか、は大体把握できていたので、それによって評価した。こちらの要求する作業ができればＡという、到達度評価にした。

３．学生の満足度アンケート

教師の一人よがりになってはいけないので、授業の最終回に、受講生に満足度アンケートを行なった。アンケートと言っても、ワラ半紙を配り、満足度は100点満点で何点か、などを書かせるだけである。無論無記名であったが、小人数なので誰が書いたかわかってしまうと考えたのか、あまりひどいことを書いた学生はいなかった。ちなみに、私は他の多くの授業でこのようなアンケートを行なっているが、50人に１人くらいは必ずたちの悪い学生がいて、えげつないことを書くものである。

以下に、点数の高い順に、ほぼ全文を掲げよう（原則として原文どおりであるが、あまりにひどい誤字などは訂正した)。

100点　私は全く初めてパソコンを使わせて頂いて、教えて頂いて大変感謝しています。おかげで少しは使えるようになったか？と思います。

　先生は分かりやすいプリントを毎時間作ってくれて全然知識のない私でも理解することができました。先生本当にありがとうございました。（以下略）

95点　ワープロしか使わないので、データベース、表計算ソフトは少々難しかった。この授業は国文学科の人には難しいようである。この統合ソフトを使う機会がない!

　データベース、表計算ソフトの使い方が少しわかり、勉強になりました。

　授業に関しては、理解しにくい箇所もあった。古事記の接続詞の所など、結果はよくわかったのですが、関数などが？？？であった。

79点　時々、自分が何をやっているのかわからなくなったりした。しかしウインドウズのソフトはやたらとわかりにくい。

78.8点　初めのうちはよく分かったが中ごろはよく分からなかった。後半はよく分かった。クラリスワークスがとても遊べるモノだと理解した。使い方の一部ははあくしたつもりだ。

75点　４回生になり就職活動もあるのだから全く面白くなければ出てこないと思います。「クラリスワークス」の様な統合パッケージを学べばその他の専門ソフトを学ぶときに変な先入観やコンプレックスなく学べるのではと思いました。

70点　ワイワイやれて楽しかったです。授業のすすめかたがちょっと不満。最初の２～３回はクラリスの操作方法や基本的な操作の授業をしたほうが良いかも知れませ

ん。レベルの異なる人を同時に教えるのはたいへんなことですね。

70点　色々と我がままをいいましてすいません。不慮の事故が多かったので何ですが御迷惑をおかけしました。しかしウィンドウズのソフトはどうしてこう不親切なんでしょう。P.S.家のクラリスは復興できませんでした。

65点　理屈をきっちり教えて欲しいです。（別プリントにあったけど）初心者なので、パソコンで何かできたという喜びで大半を占めています。だから、これをどう生かしたらいいのかもひとつ分からない所もありますが、使う機会があればいいなあと思います。有難うございました。

60点　これだけまともにパソコンを使ったのは初めてだったから最初のうちは大変不安だった。とりあえず前期が終わったわけだが、まだまだ不安だ。プリント通りにやれと言われれば出来るが、いざ応用するようになると全く手が出ない。自分一人で使えるという目標はもろくも崩れてしまったようだ。もっともっと素人向けの講義内容並びにプリントを今後はお願いしたい。

59点　先生も初心者っぽくてアットホームな感じでおもしろかったが、パソコン自体の腕が上がった気はしなかった。何のためにってとこがあいまいで、時折知っている人にしかわからんような専門語みたいなんがよく分からなかった。（ペーストetc.）それが何を意味するするのかをまず教えてほしかった。

（点数無記入）　表計算やデータベースを本格的には使ったことはなかったので、基本的な計算は理解できた。しかし類似度を出すとか累積度数を出すなどの目的意識をもって処理を行なうまではまだいっていないと思う。しかし授業の後半になって、何度も同じ操作をするのと、マニュアルがその操作の目的などが書かれるにつれて理解度は増したと思う。

上記の点数について、点数無記入のものを除いて単純に平均を出すと、75.2点となる。最高点と最低点をカットして平均を出してみても、74.1点である。初めてにしてはまずまずだったかなと思う反面、脱落していった学生の多さを思うと（16人中５人が脱落）、やはりダメな授業であったかと思う。しかし私のパソコンの実力を考えれば、こんなものかなとも思う。結局自己評価としては、65点くらいであろうか。とにかく、私自身もよい勉強になった。

４．おわりに

以上が平成８年度「言語情報処理実習Ｉ」の報告である。「もっとよい方法がある」とか「こんなことをすればどうか」などのご意見があれば、ぜひお教え願いたい。

べつに卑下するわけではないが、私のパソコン能力のレベルは実に低い。そして今後も大幅な改善は望めないだろう。この点、学生たちに対して申し訳ない気持ちでいっぱいである。しかし、私自身は結構おもしろかった。パソコンがややこしい計算をさっさと片付ける様子には、快感を感じた（もっとも、時々は、「この操作はウィンドウズの限界を越えています」というメッセージとともに、営々と築き上げた膨大なデータを破棄しなくてはならないこともあったが)。

最後に、学生のうち、まじめに取り組んでくれた人には感謝したい。また、水野柳太郎先生にもお礼申し上げます。

注

1）宮島達夫「語いの類似度」（『国語学』82、1970年）を参照のこと。

2）統計学の参考書、例えば、飯尾晃一『統計学再入門』（中公新書、1986年）を参照のこと。

3）相関係数を使用した国語国文学の論文としては、最近のものでは、橘豊『日本語表現研究』（おうふう、1996年）がある（p.131以降)。

4）「マニュアル」の章立ては次のとおりである。

1．クラリスで度数を数える方法
2．単語の切り方について
3．品詞表示について
4．クラリスの表で区間ごとの度数を出す方法
5．度数別の個数（度数１の語は何語か？など）を表で数える
6．統計値２つ
7．類似度の出し方
8．相関係数
9．「日本書紀」の使用漢字を分析してみよう

なお、このマニュアルは、最初に全章ができていたのではなく、授業の進行に合わせて、継ぎ足していったのである。このマニュアルを作るのは大変であった。

5）拙稿「『古事記』の接続詞について（上）」（『国語国文』64の２、1995年)。ちなみに結果は、「尓」と「於是」の間の類似度が、0.687と高かった。「故」「尓」間、「於是」「故」間はそれぞれ0.384、0.399であった。「尓」と「於是」が同じような文脈に用いられるというのは、本居宣長なども言っていることで、この類似度のデータは、それを裏付けたことになる。