I hereby declare that the conditions of the Ordinance and Regulations for the degree of Master of Science (M.Sc) at the University of St. Andrews have been fulfilled by the candidate, Pamela A. Scott.

Professor A.J. Cole.

I hereby declare that this thesis is a record done by myself, not accepted in any previous applications for a higher degree in the University of St. Andrews or elsewhere.

(Mrs.) Pamela A. Scott.

This work is a computer-assisted analysis of the style of Jane Austen, together with four eighteenth century novelists. The date of my admission as a research student under Ordinance General No. 12 and enrolment as a candidate for this degree under this resolution was October 1969.

A Comparative Study

of Jane Austen and the Eighteenth century novel

by quantitative methods.

by

Pamela A. Scott M.A.

Computing Laboratory

University of St. Andrews.

I Introduction.

II Literary Summary.

III Problems of "style".

IV Past history of stylistic analysis by quantitative methods.

V Aim of this thesis.

VI Method.

VII Cluster Analysis.

VIII Conclusion.

IX Basic Programs.

X Tables.

I

Literary analysis has always belonged to the realms of the abstract and most critics refuse to be tied down to facts, devoting most of their criticism to levels well above the ground. This, however, has not deterred me from attempting to bring my appraisal of Jane Austen strictly down-to-earth and dealing only with such things as can be calculated on a computer.

One is immediately faced with the question of whether in fact this can be done. Though computers are being used more and more widely for non-mathematical applications, I do not think they provide a panacea. I maintain that little of any real importance about an author's work can be identified, far less measured, and what is measurable may turn out to be peripheral or secondary, although, of course, one hopes not. Mechanical methods are simply the means to an end and must be preceded by a sound knowledge of the text and a devotion to its literary art. Moreover, first of all the same critical intellectual process should take place as in the usual literary study.

Measurement is not new to literature. Every statement about the density of Shakespeare's metaphors or the number of references to military matters in his plays makes some kind of numerical estimate, and even the airy medium of poetry depends for much of its effect on the simple manipulation of numbers - for example, a long line of iambic pentametre balanced by a short, sharp spondee.

In my study of Jane Austen and her debt to the eighteenth century, I hope to have the indulgence of the critics who love and admire her work as I do and who feel I am doing her a great disservice by reducing these great novels to a mere string of numbers. I do not pretend to give a full literary study. This work will be deliberately partial and quantitative because it is only within these limits that I can do what I have set out to do. This, then, is my apology for undertaking something which purists will condemn as worthless. I hope to prove them wrong.

II

My object in this thesis is to use a quantitative method to analyse the style of Jane Austen. I will then do the same for the four main eighteenth century novelists, Defoe, Richardson, Fielding and Smollett. (I omit Sterne as he is so unlike any author either of eighteenth century or later). These sets of data will then undergo various tests and the results will be compared.

I am concerned here with explaining my purpose in doing this and the actual techniques I use will be left until later.

The eighteenth century saw the rise of the novel as we know it today. Before that, there had been several prose works which all contributed to its rise but in themselves were not novels. These included Elizabethan short stories; works like Cervante's Don Quixote translated into English in 1612; Bunyan's Pilgrim's Progress, 1678; the Spectator Papers of 1711-1712, by Addison and Steele; and Swift's Gulliver's Travels, 1726. Not until the second decade of the eighteenth century, however, did the novel proper emerge, and when it did, it came from a writer who was not a gentleman but a tradesman without benefit of a university education. The Life and Surprising Adventures of Robinson Crusoe, of York, Mariner, the most important work of Daniel Defoe has several important differences from the other works I have mentioned. For the first time, the hero is an individual. He does not represent mankind at large, nor is he the personification of a virtue. He is not everyman but one particular man, despite the claims of Defoe and the other eighteenth

century novelists that in their works they show humanity in general. The novel, when it appeared, had a realism and immediacy not found previously in prose fiction. We see ourselves in the characters they portray and although they may be eccentric and have extraordinary adventures, yet they are rooted firmly in time and place and there is nothing mythical about them.

The emergence of the novel in the eighteenth century also has a sociological reason. It is a middle-class form of literature and it emerged at a time of social upheaval. The bourgeoisie were gaining power both financially and politically and their tastes are reflected in this new literary genre.

I do not intend to discuss the style of these eighteenth century novelists, fascinating though the period is, but in order to assess the work of Jane Austen, it is important to realise what the situation was when she came to the novel. Her debt to these eighteenth century predecessors is like an iceberg. Only the tip is visible and the rest is submerged in the very syntax of her style.

Though a contemporary of the Romantics, in spirit and in technique Miss Austen belongs clearly to the Augustan age. One of the most forthright moralists in the English language, her values are pure eighteenth century. Sentiment is generally by-passed and when she does use it, it is usually for the purpose of satire. Indeed, there are times in her novels when she is utterly ruthless.

For example, in Persuasion, the most tender of the novels, the death of the Musgrove's son is described in harsh terms. Charlotte Bronte criticised her for ignoring "anything like warmth or enthusiasm, anything energetic, poignant, heartfelt". She is complaining that Miss Austen is not like herself, a Romantic novelist. In fact, though she was about the same age as the Romantics (Wordsworth was five yearsher elder and Coleridge three) she was untouched by the fire of the Romantic movement. This does not mean that she was ignorant of the power of feeling or that she ignored it. As her novels bear witness, she, like Pope and Johnson, thought it ought to be controlled.

All this is inherent in her work, yet some aspects of her style owe a more direct debt to the eighteenth century. Her narrative viewpoint is an accumulation of the styles of Defoe, Richardson and Fielding. She dropped the participating narrator of Defoe and Richardson and told the story in Fielding's manner, as a confessed author, yet she was so discreet in this that she does not intrude upon the reader in any way. She adopts a detached attitude and evaluates the characters from a comic and objective point of view. At the same time, however, she varied the narrative stand-point so as to give us not only editorial comment but much of Defoe and Richardson's psychological closeness to the subjective world of the characters. In her novels, with the exception of Mansfield Park, we are shown one character whose mind is given privileged status and we see the other

characters through her eyes. Jane Austen's own role as narrator shows the character of the heroine and so the reader can balance one against the other and from there we can use the heroine's observations for two purposes - to reveal the character of the person she is observing and by her reaction to reveal her own character.

In theme, too, her roots are eighteenth century. Like Defoe, she faces the social and moral problems raised by economic individualism and the middle-class quest for improved status. From Richardson, she borrowed the minute presentation of everyday life and in general her theme, that of the procuring of marriage, and especially the female role in this. Like Fielding, she paints the social system as it is and should be, although its application to the characters and their situations is in general more serious and discriminating than he makes it.

Her reading also was steeped in these old masters, particularly Richardson and we see the influence of Sir Charles Grandison on many of the male characters such as Sir Thomas Bertram, Mr. Knightly and even Darcy. Most critics agree, however, that Fielding's influence goes even deeper. Walter Allen in The English Novel suggests her novels are a feminisation of Fielding's. Her world is smaller with virtually no references to external events, yet because of this it is much more intense. What she loses in scope is gained in depth. Like Fielding, she relies for a great part of her effect on the skilful use of irony

both of language and of situation. The comment as in Fielding, is not only direct but implied in the turn of the sentence. Her characters, however, are not those of Fielding. Many of his characters appear only in one situation and in a certain light, and are really little more than highly sophisticated caricatures. These include such masterpieces as Parson Trulliber and Beau Didapper, who, though individuals, yet represent an idee fixe. Miss Austen however, gives us the impression that her characters are all personally known to her and furthermore, by the end of the novel she has convinced us that we know them all likewise. Part of her skill lies in the way she makes them discover themselves (and so allows us to discover them). At the beginning the heroine often lacks self-knowledge e.g. Emma, or is over-influenced by a prejudice against seeing things in their true light e.g. Elizabeth Bennet. Throughout the course of the novel, these barriers are broken down and by the end they know themselves as well as we know them. This concept of "Know thyself" is seen in Hamlet. Polonius to Laertes - "Thus above all, to thine own "self be true" (1.3.78)

We see therefore the debt of Miss Austen to the eighteenth century, yet what she took over she transformed with her genius into something totally new, which has made her a permanent point of reference when discussing the novel. If she had merely accumulated the best of the eighteenth century, she would have been a good novelist, but her world-wide reputation rests with her innovations

and though quite unconscious of the fact, she was in her own way a revolutionary. What makes Miss Austen so new is that she wrote the pure novel and this affiliates her with such novelists as Flaubert and Henry James. The "pure" novel is not so much concerned with externals but with the intricacies of form and like pure art of any kind it has its dangers. It may degenerate into an over-preoccupation with techniques at the expense of human content. It tends to be abstract and is in direct opposition to the works of the great novelists such as Tolstoy, Balzac, Fielding. Its interest lies in subjecting the parts to the whole, the whole being the exploration of relationships between the characters or their relationships to a central theme. Its advocates sought to give the novel the intensity of other forms of literature, especially of poetry. James' "cri-de-coeur" as a pure novelist was "Dramatise! Dramatise!" No-one could be more dramatic than Jane Austen - all her novels could be easily adapted to the stage and her short, witty dialogues would be much appreciated there. The "pure" novelist seeks perfection, yet this can only be obtained within the bounds of certain limits. These limits are an integral part of Miss Austen's work and one of the main reasons she is so successful. Though her subject matter is in a sense trivial and her range very limited, she achieves in a degree equalled by very few, a style which perfectly harmonises form and content. Fortunately for posterity, the early drafts of her novels were unsuccessful with the publishers and although late

nineteenth century and other admirers have exclaimed at the stupidity of those men who rejected Susan and First Impressions, probably these works were jejune and unformed; their premature publication might have robbed us forever of the author whom we know, the slow developer, the patient and creative reviser whose writing began in the classroom and continued to her death-bed and whose genius lay for the most part in taking pains.

We see therefore that Jane Austen is a Janus figure looking back to the eighteenth century and forward to the nineteenth. Her style was deeply rooted in the eighteenth century, yet she showed many of the attitudes to the art of the novel that later writers were to adopt. In this thesis, I hope to be able to show some concrete evidence to back up this view but before I can do this, I must clear up some difficulties which arise over the meaning of the word "style". I hope the reader will bear with me through this analysis as it is really fundamental to this theory.

III

With twenty-nine subdivisions of meaning of the word "style", it is necessary for me to show what I am meaning when I use the word in this thesis. Fortunately most of these are non-literary sense and only three are relevent to literature, and it is these three that I will talk about now. The earliest meaning arose in the fourteenth century with the sense of "characteristic manner of expression ..... considered in regard to clearness, effectiveness, beauty and the like" (O.E.D.) as in "Therefore Petrak writeth / This storie which with heigh stile he enditeth". This sense of style usually applies to a group or literary school, genre, nation, period or individual e.g. Euphuistic style, Ancient style, Homeric style. The second meaning was developed during the Renaissance when the word was first of all refined to specify formal rather than substantive aspects of literary composition and then by the regular process of melioration it came to mean good style, just as poetry usually means good poetry. Thus in brief the three meanings are expression, form of expression and good form. We find that these meanings tend to merge and overlap. The word"style" is not like the words "method" or "system" both of which have clearcut meanings. They imply a procedure leading to a goal but "style" describes the way itself. This difference is seen in their adjectival form - Note the difference of meaning between methodical, systematic and stylish.

The ancient view of style was an ornate form, which implies a distinction between content and form. It was classified into

Grand, Middle and Simple style, decided mainly by diction. It was felt strongly that morality and literary work went hand in hand, and that a bad man could not write good literature. Seneca, who connected the degeneracy of morals with the corruptions of style was yet quite modern in rejecting fixed styles and insisting that style is dominated by usage and changes constantly. This view of style as an individual thing has developed with the centuries from Montaigne: "Comme a faire, a dire aussi, je suy tout simplement ma forme naturelle" ..... "J'ay faict ce que j'ay voulu: tout le monde me reconnait en mon livre et mon livre en moy" (Montaigne : Essais). I.A. Richards expresses his view in his admirable book Aesthetic : Principles of Literary Criticism :- "Every true intuition or representation is also expression. That which does not objectify itself in expression is not intuition or representation but sensation and mere natural fact. The spirit only intuites in making, forming and expressing. He who separates intuition from expression never succeeds in re-writing them."

The implications of this view are far-reaching in that no intuition has any reality until it has achieved expression. In turn this denies the possibility of a choice among means of expression for a given intuition. The intuition is unique with the individual and is so to speak identical with the expression which is thus also a delineation of the will of the particular individual. Here then is the coalescence of the two modern views of style - as meaning and as a reflection of the individual, and it

is in this sense that I use it in this thesis. As I will show later, the concept of style as reflecting the personality of the individual as well as conveying his meaning is basic to this study.

IV

The analysis of style by enumerative methods have an early origin. In the early years of history, the standardisation of Homeric texts was being carried out by Alexandrian scholars, who compiled lists of words appearing in the text (and nowhere else) and of "hapax legomena" (words appearing only once). Similarly, the Masoretic text of the Bible was safeguarded by devoted scholars who counted the verses and the words of each book and determined its middle word and middle letter (See The Jewish Encyclopedia).

In recent times we have the work of Professor Lucius A. Sherman of the University of Nebraska. He noticed that the children in chemistry classes improved if they were allowed to do experiments themselves and he applied this to English. They drew a number of conclusions about the English sentence-structure. Despite a gradual decrease in sentence-length (in words) between Thomas More and Macaulay, a remarkable consistency existed in any single writer's average sentence-length. More's Richard III averaged fifty-three while by De Quincey's time, the average was only thirty-two. Macaulay's average of twenty-three is very unusual. He had a constant average of just over twenty-three in all sentences of his Essays and his History of England in any sample larger than five hundred sentences.

One of Sherman's pupils, Gerwig, found that modern writers used fewer than the six verbs in each sentence that their literary ancestors had done. The modern tendency was to suppress the

superfluity of finite verbs by such means as apposition and verbals and to write more simple sentences. Their research, however, is not scientifically accurate. Macaulay's average could be due to the uniformity of the samples, which were all histories and essays. A very wide range of sentence-lengths in individual authors was found by examining works in different genres. Furthermore, it was discovered that the decreases of prediction-average and the increases of simple sentence average were not merely parallel, as Gerwig had thought, but were functionally related. As the use of finite verbs per sentence decreased and the sentences became shorter, the percentage of simple sentences naturally rose. In fact, formulae were written to compute either, if the other was known. P S = C where P is prediction-average, S is % of simple sentences and C a constant.

During the war EdithRichert had been very impressed by the power of code and cipher analysts and later she wrote a book : New Methods for the Study of Literature in which she claimed "to substitute for the impressionistic, hit-or-miss, every-man-for himself, method of approaching literature" some "graphical and statistical methods," by which an author's style "may be understood with a definiteness and certainty impossinle through reading alone." Her efforts, however, were neither very scientific nor original and she had little influence.

The problem with all these methods is that they represented a mere ennumeration of the externals of their subject. Because the

mathematicalaspects were only half-understood they have always been suspect to more traditional scholars who prefer the historical approach. The main problem in this approach is in identifying the styles of particular periods and accounting for any changes which have taken place. Very few critics using this approach have been able to account for any important individual element. It is obvious that in any period, a variety of styles can be found and the historian must simply select some representative illustrations. Even when he is aware of contradictory tendencies, he is bound to under-rate the likelihood that within a highly mannered style, such as the Ciceronian, there were important individual variations, differences so great that they obscured the significance of the common pattern. Therefore to the student of an individual author's style, the historical study is far from conclusive, and a more concrete method, whatever its theoretical justification seems to have more possibility of success. An approach which begins with the stylistic phenomenon and carries it back to some historical sequence at least suggests a firmly-grounded empiricism and to that extent it is persuasive. Such a procedure is the study of modern "stylistics".

The stylistic device is considered as a conscious deviation from the linguistic norm of a writer's period. In fact only that which is beyond the neutral common denominator qualifies as style to stylists. Where there is an alternative available, the possibility of a stylistic choice exists. For example, the French adjective may be placed before or after the noun it modifies,

subject to certain exclusions. Pierre Ullmann summarises this view in his book Style in the French Novel: "At the risk of over-simplification, one might say that everyting which, in language transcends pure communication belongs to the province of style. Whether the choice and the effect it produces are conscious or not is fundamentally irrelevant to a purely stylistic inquiry and it is also most difficult to determine". This applies to the French language which has stricter standards of grammatical propriety than English and therefore the norm is clearer, and deviations more noticeable.

We now come to the question of the application of statistics to literature. This approach is probably the most far removed from the usual literary study. Students of literature often resented statisticians and accused them of fundamental ignorance about the subject. Statisticians, for their part, were either apologetic and tried to disarm criticism by admitting that they were not entirely qualified for the job, or aggressively assertive of the advantages of a scientific approach to literary study. Literary students argue that literature is a matter of the spirit to be deliberated, understood and appreciated only by the subtlest intellectual effort of trained and sensitive minds. Statisticians concede this point but argue that an opportunity for this science still remains because language is a mass phenomenon (It involves a great many small units in various distributions) and statistics

is a method for treating masses of any units with a verifiable amount of accuracy. An individual work though a unique creation, is still a part of the enveloping mass of language among whose users the author is only one. This does not mean we are to reject the scope of individual contribution but it narrows it down in certain places. For example, the difference between Dickens and Virginia Woolf is artistic, ideological and sociological, yet it is still possible that the frequency of letters, phonemes etc. which they use is similar and even predictable. This means only that the English language has certain features which bind it together and distinguish all its users.

The founder of statistical linguistics, George Kingsley Zipf in "The Psycho - Biology of Language", concentrates on this fundamental level of language and discovers the vast number of units in which repetitiveness makes categorisation successful. In investigating "the relationships which exist between the form of the various speech elements and their behaviour in so far as the relationship is revealed statistically," he sought to prepare the way to "the formulation into tentative laws of the underlying forces which impel and direct linguistic expression". He examined words by their size in syllables, and using a count of Ten Million words of German prose made in 1897-8 by F.W. Kaeding, he found that in German language, one-syllable words are used half the time, two-syllables a third of the time, down to a single fifteen-syllable word. These results were backed up by the work of another

investigator whose results tallied very closely with Zipf's. To verify his conclusion that short words are more common than long ones, Zipf showed that Chinese, English and Latin accord very closely to the same proportion, whether the measure is in morphemes, syllables, or phonemes. From this he formulated the Principle of Least Effort (Human Behaviour and the Principle of Least Effort). This supposedly governs word-length, and tends to keep equilibrium among the various components of language.

The application of statistics rather than mere enumeration, to literary problems, especially problems of style or "individual differences" has moved more slowly. This is due largely to the characteristics of statistics itself which is a means of dealing with masses of units grouped in separate categories. A natural language supplies inexhaustible material for counts of anything from letters to paragraphs. But the application of this to a particular study is more difficult. First the size of the material, the sample, is greatly reduced in size and secondly the categories are less simple and larger in size, so shrinking the size of the samples still further. This can be successful, however, as long as the work is done carefully and patiently and the investigators do not claim more than is apparent by the evidence. Yule, in his Statistical Study suggests that a safe minimum would be a sample of 10,000 words.

Too small a sample was the trouble with students of

Equatorie of the Planetis, tentatively attributed to Chaucer on the basis of a minute study of the percentage of nature to Romance words (as well as other items). Studies were carried out by Elderton and Fuchs, which relied on syllable counts as tests of authorship. Each worked separately and the only author common to both was Shakespeare, Elderton examining Henry IV Part I and Fuchs Othello. The percentage for each word-length corresponded very closely and the average number of syllables per word agrees to two decimal places. Such agreement depends on large enough samples, adequately selected, on careful definition of categories (even in the case of syllables) and of course, on careful counting.

Here the computer can play an important role in the literary field. It acts as a clerk who can do our numbering more quickly and more efficiently than its human counterpart. It is merely the tool and the responsibility as always rests with the wielder not the tool, and therefore in order to obtain results which are accurate and trustworthy, we must look not at the computer, but at the methods employed. Therefore, as a first step towards reliable results I will now explain the assumptions on which this thesis is based.

V

I will briefly repeat my views of style.

1. It is the deliberate choice of an effective means of communication (the usual sense of style for most critics).

2. It is the unconscious expression of a writer's personality in his work. That this unconscious expression is consistently diffused through an author's writing is a major assumption of this thesis.

Literary style is a means of expressing one's personality, and, like handwriting, it is guided and directed by the deepest recesses of our mind functioning effortlessly and unconsciously. As our personality matures so our style changes with it. This view is backed by many psychologists. Gordon Allport in his book Personality summarises this into three points.

1. Style reflects personality.

2. It is an unconscious process.

3. In mature writer's this process is consistent.

Two of these lead on to two other theoretical assumptions (2) leads on to the implication that it may be most easily observed in an involuntary or automatic aspect of writing. (3) leads on to the possibility that it may be measured. Granted that it could be measured are we justified in saying that it is desirable? As I have already said, students of the arts have always felt that they were dealing with matters of the spirit. We must ask ourselves whether the things that can be measured are the really significant ones. The student faced with the whole, must analyse it into separate components which can be treated objectively without losing

the spirit of literary inquiry. The ultimate aims of literary study must be kept to the fore in the student's mind while he is devising suitable categories for quantitative treatment of his text. There must be an aim beyond the mere manipulation of figures. The literary student must produce categories in such a way that he is convinced he is measuring something he feels is significant, not something which is merely measurable.

Having clearly set out my assumptions, I now intend to show the methods I employed to carry them out.

VI

The first question to be decided was on what to base our criteria of difference. The first thing that comes to mind is vocabulary, but this is a relatively unstable aspect of writing, because of several reasons. First, vocabulary is very affected by the type and subject-matter of writing. Two sermons by different writers will probably have more in common than a sermon and a romance by the same writer. Vocabulary, therefore, is useless as a criteria for identifying personality. Secondly, vocabulary is also susceptible to an author's conscious choice e.g. whether to use "improvement" or "amelioration". A clever imitator could copy this conscious choice and so render invalid the lexical criteria of style description and identification. What we need is a criterion stable enough to offer predictive power yet safe from imitation, conscious modification and variability of subject-matter.

The answer lies not in the words but in the structure of sentences. The words must be taken as carriers not of lexical but grammatical content. The method is to parse sentences into parts of speech (but not particulars of dependency).

Undoubtedly, such analysis tends to blur certain distinctions. "The boy took the dog to the park". "The minister preached a sermon to the congregation". These two sentences will appear identical by this analysis. Literary quality is obviously sacrificed here, but unfortunately, this cannot be overcome. The choice of word is obviously important but as it cannot be measured definitely, it is out with the bounds of this discussion.

One important point about grammatical criterion is its remoteness from the writer's immediate interest. He is interested in words, subtle shades of meaning, not in its grammatical components. Though he may pause to consider whether he wants cicatrice or scar, he is unlikely to ponder over whether he wants a subordinating conjunction or a relative adverb, a conjunction or a conjunctive adverb. He is aware of these things, but the awareness is not immediate or tangible.

Moreover, the writer's consistency can be better shown by his grammar than by his syntax, because it is beyond his conscious reach and also because it tends even more than the vocabulary to remain static in adult writers. We are all aware how often, especially at the beginning of sentences we slip into certain familiar structure. For example, in letter-writing we have to make a conscious effort not to begin every sentence with "I". That we can modify our grammatical structures by conscious effort does not invalidate the contention that our grammatical choices are mainly unconscious. Rather it suggests that there must be many other aspects of structure which are not open to casual inspection and which can only be ascertained by the use of computerised methods. These have several advantages.

1. The computer in this sense is like a high-powered microscope in which the cellular structure of a body may be examined.

2. It reduces the possibility of any bias on the part of the observer.

3. It is less apt to make mistakes.

The method I found to be most suitable to my needs was similar to that of L.T. Milic, as described in his book A Quantitative Approach to the Style of Jonathan Swift. I analysed the samples manually into word-classes. Each word was individually classified and given a number according to its word-class. Thus each sentence became a series of numbers, the end of sentence period being denoted by a different number. The data was then punched onto cards, each deck making up one sample.

The main difficulty in this method was in determining the word-classes. Most words were straightforward, but as the English language is rich with exceptions to every rule, many words proved difficult to place.

Basically the word-class system is that invented by Charles C. Fries (The Structure of English) who divides words into two sections -

(1) Parts of Speech - words which bear the main lexical burden of the sentence.

(2) Function Words - Words which constitute one of the major means of conveying grammatical information in English.

Milic adopts this system in general but modifies the actual word-classes. I have taken up some of Milic's modifications and added some of my own. The Parts of Speech are divided into eight classes.

01 Nouns - man, apple.

02 Verbs in strict sense of word - ate, walk, took, go.

03 Descriptive adjectives - good, ripe, old.

04 Descriptive adverbs - quickly, brightly, hungrily.

Classes 05, 06 and 07 are all verbal forms which Fries joins to 02. I agree with Milic in subdividing them for greater analytical refinement but I give a separate class to imperatives which he gives to gerunds.

05 Infinitives - Here I group the word "to" which precedes with the infinitive itself, as these are too closely linked in my opinion to be treated separately. Hence "to go" is treated as one word.

06 Imperatives - close the door.

07 Participles and gerunds - I have closed the door, a chance of leaving.

08 This includes proper names, foreign words, quotations, titles and other miscellaneous substantive expressions - Mrs. Smith, a la carte, Her Majesty.

The Function Words are rather more complex. To Fries, they have three things in common.

(a) Each class has only a limited number of members.

(b) The words lay the emphasis on structural rather than lexical meaning.

(c) Speakers of the language must learn these words as individual items because they are not distinguished by formal features as are Parts of Speech.

I prefer Milic's division of these function words to Fries and have followed it closely, only deviating slightly in class 61. He divides the Function Words into sixteen classes which can be grouped into nine types, each characterised by some functional attribute.

11. This contains the pronouns - personal, reflexive, reciprocal, indefinite, demonstrative, but not the relative pronouns, limiting adjectives and interrogatives. Examples are I, we, this, nobody.

21. The second type contains auxiliary verbs and all parts of the verb "to be" e.g. shall, ought, must, is, am, will, have. I have differed from Milic here in grouping two or more auxiliaries together as a single unit e.g. "he would have one" is parsed 11 21 07. Milic gives 11 21 21 07. To me, "would have" here has one function and to all intents and purposes is a single item.

31. Includes all words which modify class 01 - that is the limiting adjectives - a, the, our, such, both, every, my.

The next three classes of this type have adverbial functions.

32. This is based on the modes of German verbs which have detachable prefixes. For example, go out in the street, get up in the morning.

33. This contains the intensifiers - so, very, more, not.

34. This is a miscellaneous class including all the function adverbs not accounted for by classes 32 and 33. They are mainly adverbs of place (here, thence), time (never, always, seldom, now, never) and possibility (perhaps, only).

The fourth type contains the connectives and is divided into 5 classes.

41. The co-ordinating conjunctions - and, but, or.

42. The subordinating conjunctions - since, because, if, although.

43. The relative conjunctions - who, which, what.

44. The interrogative conjunctions - who, when, which, how.

45. The correlative conjunctions - either ... or.

51. The fifth type contain the true preposition; in, of, by etc.

61. Milic keeps this class for three words which he calls pattern markers. These are - "there", "it" and "to" in specialised use. "There" and "it" are expletives or anticipatory words. For example - "there is something burning", "it must be noted". These two words I accept, but Milic also groups with them the "to" which introduces an infinitive. I have already explained that I grouped this "to" with the infinitive group and my reasons for doing so (Milic, himself, is not happy with his grouping but cannot decide what else to do with "to".)

71. This deals with words that appear at the beginning of a sentence but which have little grammatical meaning e.g. Well, Yes, Oh.

81 This class contains all the numerals in any form. Fries treated this without much consistency and I think Milic's reasons for creating a separate group are justified. They are a finite, smallish group of words to which it is impossible to make any addition. They have no semantic content except with reference to themselves. "Fourth" has a meaning only if we are aware of third

and fifth. It has merely a "structural meaning". Again, two or more numerals grouped together to make up one number are treated as a single item e.g. twenty-four thousand is parsed 81, not 81 81 81.

91. The final class contains the sentence connectors - e.g. however, therefore, nevertheless.

This system is well suited to the requirements of literary analysis and the only remaining difficulty is in the definition of "word". In most cases we can take the typographical description of letters surrounded by spaces, but this causes difficulty when we come to phrases like "on the other hand". Idioms like this should be treated as one unit - they are unanalysable into their separate parts and they could be replaced without any change of meaning by a single word. Similarly with correlative conjunctives "either .... or", "not only .... but also". These are cited in pairs in the dictionary as if they were one unit and for our purpose, we shall take them as this.

Much deliberation went on when I encountered phrases such as "no reason in the least was given", "so far from being sorry that", "a great deal better", "could not but have gone". These are classless, and the only consolation one has is that they make up a tiny percentage (less than 1%) and any error is decreased by the fact that only one person is analysing the text and if he errs, he will probably always err in the same way.

The texts, now reduced to a series of two-digit integer

numbers, were then punched onto cards, forty words to a card, thus filling all eighty columns. Each sample contained two thousand words, thus using fifty cards.

I used fourteen samples in all for the main testing programs, six from Jane Austen, and eight comparison samples. (Later when I used cluster analysis, I added another twelve samples, six from Austen and six comparison, making twenty-six samples in all.) The six Austen samples were made up of one from each completed novel she wrote in chronological order. The comparison samples give a comprehensive picture of the eighteenth century style of writing by taking two samples from each of Defoe, Richardson, Fielding and Smollett.

These samples were chosen at random, but it did not seem practicable to pick them page by page as this would be unfair to the stylistic content. I preferred to choose the sample in one block. All the samples belong to the genre of the fictional novel although this takes different forms, such as journal, diary, letter.

Below I have set out the samples with their dates of publication.

Sample	Work (short title)	Author	Date
1.	Northanger Abbey	Jane Austen	Written 1797 not published till 1818
2.	Sense and Sensibility	"	1811
3.	Pride and Prejudice	"	1813
4.	Mansfield Park	"	1814
5.	Emma	"	1816
6.	Persuasion	"	1818
1.	Journal of the Plague Year	Defoe	1722
2.	Robinson Crusoe	"	1719
3.	Pamela	Richardson	1740
4.	Clarissa	"	1748
5.	Joseph Andrews	Fielding	1742
6.	Tom Jones	"	1749
7.	Humphrey Clinker	Smollett	1771
8.	Peregrine Pickle	"	1751

The aim in this study is to try and show that a writer's personality expresses itself in such a way in his work that it is distinguishable from the work of other authors. The first thing to be done, therefore, was to demonstrate that Jane Austen showed consistency within her own works. If this was the case, I could then go on to compare the Austen samples with the others and (hopefully) find some significant difference between them. If she did not differ from them, then her own consistency would have no relevance here and would merely reveal that the English language had a certain common ground of similarity between all users of it. This might be interesting in itself but it would of course invalidate

the hypothesis that a quantitative formula for individual style differences could be found by word-class analysis. If, however, Jane Austen's style did show consistency, and differed meaningfully from the comparisons, then the hypothesis could be considered valid.

The most obvious place to start was with a count of the distribution of the word-class. The results for one sample are shown in Table I.

Table I - Distribution of words in a 2000 word text (Northanger Abbey) into word-classes, as a percentage of total number of words.

Class	Percentage
01	15.85
02	6.30
03	5.95
04	1.63
05	2.03
06	0.05
07	3.81
08	3.30
11	8.99
21	6.91
31	13.21
32	0.91
33	2.18
34	3.76
41	3.71
42	1.07
43	1.42
44	0.66
45	0.81
51	10.77
61	0.76
71	0.20
81	0.97
91	0.56
98	4.17

Table II - Comparison of the percentages of word-class frequency between two samplesof same work of Jane Austen (Pride and Prejudice).

Group	Sample 1	Sample 2	Difference
01	10.22	11.84	-1.62
02	6.05	6.74	-0.69
03	5.84	4.62	1.22
04	2.50	1.15	1.35
05	2.71	1.64	1.07
06	0.52	0.10	0.43
07	5.27	3.56	1.65
08	3.96	4.43	-0.46
11	10.53	10.20	0.33
21	8.65	8.76	-0.10
31	11.78	12.61	-0.83
32	0.83	0.48	0.35
33	3.44	2.69	0.75
34	2.71	3.95	-1.23
41	3.23	3.66	-0.42
42	0.63	1.44	-0.82
43	2.61	1.92	0.68
44	0.52	0.67	-0.15
45	0.63	0.29	0.34
51	8.24	9.05	-0.81
61	2.19	1.64	0.55
71	0.42	1.25	-0.83
81	0.10	1.15	-1.05
91	0.63	0.10	0.53
98	5.84	6.06	-0.22

If we examine Table I immediately a few basic observations can be made. Several classes stand out as much larger than the rest, these being 01, 31 and 51, but this is hardly surprising as these consist of nouns, limiting adjectives (a, the, all etc) and prepositions. We shall see later that these three in different combinations make up the three most frequent three-word pattern-

groups. At the opposite end of the scale, the sample is sparse in imperatives (06), post-verb participles (32), interrogatives (44), correlatives (45), sentence-markers (61), appellatives (71) and sentence-connectors (91). Pronouns (11) occur more than half as many times as nouns; true verbs (02) and auxiliaries (21) occur approximately as often as each other. Descriptive adjectives occur one-third as many times as nouns. Without any comparison with other samples, however, nothing can be said about the stability of these classes. It will be of interest therefore to examine two samples from the same piece of work to see how they compare.

We see from Table II that the sub-samples are in good agreement, considering the size of each sample is 2000 words, the disagreement never being greater than 1.62 and this is in the case of the nouns which is a large group. I found more agreement in the Function Words than in the Parts of Speech. In the 16 classes of the former (11-91) the disagreement was only twice above 1%. In general, therefore, we can say these samples are a good indication that our initial assumption has a sound foundation. Still, however, little can be said about the significance of these results until we determine some way of measuring the permissible limits of divergence.

As we see in Table III, the divergence varies with the size of the sample. The bigger the proportion of the population covered by the samples, the more closely the samples will resemble the population until the population and the sample are identical.

Table III takes sub-samples one-tenth the size of the original samples - that is 200 words. Immediately we see a huge variation, the difference being as great as 8.54% in the case of nouns and more often than not it is over one per cent whereas in the full samples it was the exception for differences to be greater than one per cent. We notice also that in several word-classes there are no differences, but this is often due to the fact that there are no entries for those classes, as can easily happen to rarer word-types in small samples. We can safely presume therefore that the larger the sample, the more consistent the results will be. This brings us back to the question of how we can evaluate a permissible difference between two samples and when do we draw the line above differences that are too great. As no previous research is available for comparison, I will return to my early assumption. The claim is that the characteristics of Miss Austen's style will be revealed by examining the distribution of the word classes. If this is incorrect, then any set of samples from any other author would be just as consistent as the Austen samples. The six samples from her works were then compared with the eight samples of the eighteenth century authors and the results are seen in Table IV.

The Austen samples show remarkable consistency, the Standard Deviation only being greater than one per cent in three classes, 01, 02 and 11, these being among the largest word-classes. In the comparison samples, however, one notices a greater variation

although this is lessened if we look at the samples in pairs, each pair representing one author.

The results of Table V are easier to see. This shows the means and standard deviations of both sets of samples side by side, each set being taken as a separate population. Examining this in detail, the smallness of the standard deviations in the Austen samples show the greater inner consistency of these samples. In most cases (17 out of 24) the standard deviation of the Austen samples is less than that of the comparisons - that is the dispersion is less and the consistency greater, and in the classes where this is not the case, only once is there any great difference (class 81 where it reaches 0.36). These figures help to support my supposition that the Austen samples make up one homogeneous unit, with certain notable characteristics, rather than just a set of random numbers.

After studying these figures we see that the differences between the samples are only within certain limits. Within the boundaries of the language itself, certain restrictions are imposed. For instance, no one could write a recognisable piece of English of any size without the use of verbs. Similarly a piece of writing needs nouns and prepositions in certain proportions. These vary according to the style and content of the writing but certain linguistic requirements must be obeyed. The number of nouns must be greater than the number of prepositions and the number of noun determiners should not be much less than number of nouns. Therefore,

although the author would appear to be quite free to write as he wishes, in fact he is fairly restricted. In normal literary writing in English, nouns are used between 15% and 20%. Although little data is available one source tells us that French is nearer to 25% and Spanish 33%. Similarly, fewer adverbs are used in Spanish than French, and fewer in French than English. Individual variation therefore can only occur within the limits of the language. Where these are too narrow to permit a range of individual expressions, then the language may be considered as a set of linguistic constants.

Let us now leave the distribution of word-classes in general, and proceed to a more detailed study. My method was to compare the six Austen samples with the eight comparison samples as if the latter constituted a population. This procedure was justified as showing that the Austen figures had the cohesion characteristic of a set of related phenomena. However, although the Austen word-class dispersion showed more consistency than the comparisons, this allows nothing more to be inferred. In order to show that Austen's figures in any word-class are peculiar to her and can be used as a criterion of identification, it will be necessary to show that the individual controls vary enough from the Austen figures so that no confusion can exist. Returning to Table IV, we see the Austen figures for class 01 range from 16.00 to 11.06, with a mean of 14.05. Defoe with 17.00 and 15.02 has one higher and one lower than the Austen range, Richardson is within the range, Fielding is

above it with 16.86 and 16.43 and Smollett is well above it with 18.16 and 21.04. In other words, if we relied on class 01 to distinguish works of Jane Austen from those of Richardson, success would not be possible. This is generally true for the other word-classes and therefore individually they will not serve to differentiate between Miss Austen and all the comparison authors.

For this to be possible, Miss Austen and all the comparisons must show different values of any word-class in which they are compared. It is naturally possible for their use of nouns to differ in quality though not in quantity but this cannot be measured by my method. To distinguish between Austen and the comparisons in their use of nouns it would be necessary that individual frequencies were spaced wide enough apart to permit distinction with no overlapping. In our case, although Jane Austen's and Richardson's use of nouns may be vastly different, on the evidence we have here we must assume that in the case of class 01 nouns they are alike.

Other means therefore must be found to differentiate them. The first thing I did was to join the word-classes into groups of greater stability. Here one must be careful, however, that one is not just putting together groups which look as if they would give us the results we want. We must bear in mind what I said earlier about always keeping the literary aim in view. The word-classes must have certain characteristics in common to validate their being grouped together. (Because of the random variation,

the consistency would probably only hold true for one particular set of figures and therefore it does not represent the characteristics of the population). Only in this way would there be any chance of success of finding a group which would give consistent results, no matter what samples were chosen.

The first grouping I made was of the Parts of Speech and the Function Words which my earlier explanation and the basic lay-out make an obvious choice. The total for the Parts of Speech is obtained by summing classes 01 to 08 and the Function Words from classes 11 to 91. The results can be seen in Table VI. (The results are in percentages but do not add up to 100% because the end of sentence marker is computed as a percentage though it belongs to neither of the two groups).

We see the ratio is approximately 40-60 in both sets. In more detail, the Austen values for the Function Words are generally lower than the comparisons, the highest being 58.53. The Parts of Speech total therefore is generally higher, except for Fielding whose results are similar and for Smollett who has the highest Parts of Speech value of all. As with the word-class count, therefore, this test does not distinguish Jane Austen from all the comparison samples. Some interesting observations however can be made. According to linguistic theory a high P/S value is synonymous with a formal style and conversely a high FW suggests a colloquial style. According to this hypothesis, Miss Austen would tend to the formal

rather than the colloquial, just as Fielding would, though neither to such an extent as Smollett. On the other hand, Defoe and to a lesser extent Richardson are, by this theory, exponents of the colloquial style. I think most literary critics would agree with this general assumption, remembering Fielding's claim to write a "literary epic, something hitherto unattempted in our language" and Smollett's formalised mode, while Defoe writes in the form of a journal and claims his works are not novels at all, and Richardson uses the letter-form, both of which are obviously prone to colloquialisms. This of course cannot be taken as absolute proof of a writer's formality or colloquiality but it provides a good pointer on which to base further evidence.

The next text was to group the word-classes, according to the types modifiers (class 31-34) and connectives (41-45). It should be noted that a grouping such as connectives will have a higher consistency than the individual classes of which it is made up because the process of syntactic is a constant necessity, though the means of achieving it may vary. Different stylistic effects are obtained by using subordinating conjunctions, co-ordinating conjunctions etc but the total will be constant i.e. a writer may hesitate between "but" and "although" but he will have to use one of them and so the sum of connectives will be constant. Table VII shows the results of the connectives as individual word-classes and as a total. Classes 44 and 45 are too small to be of significance

in themselves, but grouped with the other connectives, they play a meaningful role. On the whole, the figures for the total connectives are appreciably lower in Jane Austen than in the comparisons and she is fairly consistent overall, the range being only 1.75 as opposed to 3.89 of the other set. The standard deviation was calculated for the totals and found to be 0.73 for Miss Austen and 1.51 for the others. In every case, the range for the Austen samples is lower than that for the comparisons. In the latter, only Defoe and Fielding show much consistency on most classes, the others varying quite considerably in all classes. Smollett's totals are similar but they are made up in different ways. His first sample uses a high percentage of co-ordinating conjunctions (41) while in the second sample, he uses more relatives. We see too that apart from Smollett, the comparison samples use connectives to a greater extent than Miss Austen. It is interesting to note that the two writers of the colloquial style (Defoe and Richardson) have a decided preference for the simplest form of connectives - the co-ordinating conjunction (41).

The results from the grouping of Modifiers in Table VIII tell us very little. Miss Austen shows a greater consistency in individual classes as well as the total than the comparisons although only slightly so in class 34 which is at the best of times a hotch-potch class made up of loosely defined terms. It is interesting to note that in samples 3 and 5 of Jane Austen the noun

determiners (31) have lower results than in the other Austen samples. If we refer back to Table IV we see in these same samples, 3 and 5, the noun count is also lower. Table VIII only shows that Miss Austen is more consistent than the comparisons as the range of the total modifiers is only 1.96 in Austen compared with 6.10 in comparisons.

The groupings up to now have only proved that our language has certain necessities of grammar which must be used to roughly the same extent by writers, or at least those writing in the same genre. It would appear, therefore, that the best groupings do not involve a large choice, and it is better to use those groupings in which the choice is purely optional and the range is rather more limited. With this in mind, the next group I tried was on the authors' use of verb forms. These fall into five classes - finite verbs (02), infinitives (05), imperatives (06), participles and gerunds (07) and auxiliaries (21). Classes 05, 06 and 07 will obviously form one group, all of them being infinite verbs or "verbals". This group of words is used by writers to reduce the frequency of finite verbs and research has shown this to be a peculiarly modern development. It has been said of the infinite part of the verb that no other part of our grammar is developing as vigorously. (This is based on the research of Professor Sherman whom I mentioned earlier. In the four twentieth century samples I use later in the thesis, I have found the direct converse to be the case. They all have much lower values for infinitives

(05) than Jane Austen and comparisons. I do not propose therefore to take up Sherman's theory of infinitives as a sign of modernity.)

I grouped classes 02 and 21 and called the group (VA) and classes 05, 06 and 07 (VB) and the results are shown in Tables IX and X. The grouping of auxiliaries and verbals fails to determine anything of consequence as neither set of results has enough consistency and the comparison samples fall within the range of the Austen results. The results for the verbals 05, 06 and 07 are also indecisive. The imperatives value (06) is always low, reaching its maximum value of 1.06 in Richardson. This is quite in agreement with his style, as imperatives are liable to occur more frequently in conversation rather than in descriptive passages and Richardson's letter form is bound to be conversational. Defoe has a very low count for this class (0.00 and 0.10) and this too is borne out by his style. The journal-diary form, consisting as it does of a recorded history of events leaves little room for conversation (and therefore imperatives) and even reported conversation has little appeal to Defoe. In this text like the previous one, Jane Austen is not firmly differentiated from the comparison authors and in fact the mean for both is only different by 0.01%.

It occurred to me that the authors might show various characteristics in the way they began each sentence. Most of us know how easy it is to fall into certain patterns which keep occurring. I therefore decided to write a program to find the first element in

each sentence, mark which word-class it belonged to and count the frequency of each word-class. Since sentence-lengths vary, there will be more sentences in some samples than in others, a factor which will affect the relationship of the results, yet without taking this into account, the results are so clear-cut as to be unmistakable. These results can be seen in Table XI.

We see immediately the overall general preferences. Some classes are used frequently by all the authors (11, 31 41), some not at all (32 and 45).

Miss Austen is very consistent in her use of Parts of Speech as starting words but the Function Words are less so, particularly with class 81 which does not occur at all in first four samples but the last two have values of 10.61 and 11.11. These values are remarkable; their size is rather magnified by the fact that both these samples contain fewer sentences than the other samples (Only 27 sentences in sample 6 as compared to 114 in sample 3)

The comparison samples also follow this trend, although they are less consistent even in the Parts of Speech. Fielding's two samples (5 and 6) have a relatively high value for classes 06 and 08. It is not really surprising that Defoe has a value of 0.00 for class 08 which is made up to a very large extent of proper names. Defoe is concerned usually with only one person at a time in his novels and as the person is usually the narrator, he uses the first person and has little occasion for using proper names.

Richardson, for the same reason (using first person) has one zero value and one low value. As one could predict as a follow-on to this, their values for class 11, the pronouns, are high.

Little else can be deduced from a table which shows little consistency, even among the individual pairs.

The existence of a nominal rather than a verbal style might prove interesting in this study. It is commonly held that bad writing uses a great many nouns with relatively few verbs, yet despite this, many writers, and some of them very good, consciously or not prefer a nominal style. The procedure I used to decide this was simply to divide those word-classes nominal in character (01, 03, 31, 51) from those verbal in character (VA, VB, 04, 32, 33, 34, 42, 43), add each group separately and find their ratio. Table XII shows this. Smollett headed this group making him the most nominal while Richardson just pipped Defoe at the other end of the scale in being the most verbal.

The hall-marks of a nominal style are supposed to be monotony of pattern, ease of writing, impersonality and formality. I do not think this test is ample proof for expounding the view that Smollett has all these things in abundance but certainly most critics will agree that the general trend is correct. Smollett and Fielding, who have the highest values are both less varied and more formal than the others, Fielding being intentionally so to fit in with his theory of a literary epic and Smollett because his

gifts differed from those of the others (Even Defoe, who was perhaps not a genius, could not be accused of formality or stiffness in his writing - it has all the vigour of a personal interest.) The results of this test are in perfect agreement with the test on the Parts of Speech and Function Words. There, a high P/S total indicated formality, a high FW - colloquiality. The results were exactly the same as here, with Smollett heading the "formality" title, followed by Fielding and Austen, with Defoe and Richardson in the rear.

Another characteristic of the nominal style is ease of writing. If the nominal style is easier to write than the verbal, one might expect that a writer would become "nominal" when he was tired or had grown weary of the task. Such characteristics might coincide with the characteristic state of writers in their later periods. On this theory, a writer's later work would be more nominal than his early work. Miss Austen's figures bear this out with her last work Persuasion giving a much higher value than the others. It is interesting also that her two most lively books beyond any doubt, Pride and Prejudice and Emma - samples 3 and 5, have much lower values than any of her other works.

However revealing this may be, it still does not tell us very much about the semantic process which goes into the relationship between noun and verb. Another view to this question is provided by a German psychiatrist's hypothesis about the relationship between verbs and adjectives. This man, F. Busemann, worked

out the adjective-verb quotient which he claimed assigned psychological values to these two parts of speech. The quotient is easily compiled from the word-class frequency distribution and is worth pausing to examine because it claims to show the emotional stability of the subjects.

Busemann carried out an experiment with some children, recording their oral story-telling and he noticed a great difference in the relationship between active constructions (verbs) and qualifying constructions (adjectives) from child to child. Moreover, over a period of time, an increase in the number of verbs followed an increased emotional instability (as independently observed by the children's teacher) while an increase in the number of adjectives reflected greater emotional adjustment. He assumed these to be translatable to two different types of adult personality, the subjective (active) and objective (qualitative). Later, another man, Boder, carried Busemann's work into the literary field. He used a large source of American writing (adult) for his material. Using samples of 300 to 350 words each, he found a wide range of values, varying greatly from one source to another.

Table XIII reproduces some of his results. The highest and lowest valued are so widely different that we wonder if they can be trusted. The problem lies with the size of his samples. 300-350 words leaves so much room for divergence, but since the quotient for each type of writing has been drawn from a sufficiently large number of samples, the averages may be considered representative.

The results are fairly predictable - the dialogue of plays is expected to be active, the prose of doctoral dissertations circumspect. The value for advertisements makes them rather more objective than one might imagine but as fashions in advertising fluctuate with great alacrity, this could be easily explained as a whim of the time (the research dates from 1927, though the paper was not published till 1940). The types of material he uses are not really comparable to the works I am studying, but on a broad base, the line from Fiction to Mencken could be used as a basis of comparison.

Table XIV shows my results for the AVQ quotient. The highest value in the Austen samples is more than twice the lowest value - a wide range. All her ratios except the lowest are higher than the comparison samples. The lowest values belong to Defoe with the other three authors bunched closely together, the means being within 4 of each other. This test therefore is one in which the work of Miss Austen is very distinctive. If we compare her results with Boder's scale, we find their mean value to be the same as the advertisements! However I have already discussed and explained the irregularity of finding the advertisements placed on the scale between Mencken, a good essayist, and Ph. D. theses.

If we accept Busemann's pyschological tests and since the words we are examining all fall into the category of the early novel, the following ranking emerges in decreasing order of emotional stability. Highest is Jane Austen, followed by Richardson, Fielding,

Smollett and Defoe. The significance of the ranking is found in the division of personality into two types. One is characterised by great intellectuality, more concreteness, higher objectivity, less emotionality and less energy and this type favours adjectives. The other has the opposite traits and favours verbs. This is obviously an over-simplified solution to such a complex subject as personality analysis of writers who in their own way reached genius-level, yet we can see great truth behind it. Miss Austen is noted for her high objectivity and her concrete grasp of facts. She did not write in the heat of the moment but was a patient critic of her own work. Defoe, although very concerned with creating an impression of factual reality, by his continual references to particularities is a very subjective writer and less intellectual than the other four authors.

Having gone as far as I could with these logical groupings by word-class, I decided to try another way of grouping these individual words. Frequency distribution is not the only way in which the words of the texts can be used. The method I now turned to was to try and discover how these words were arranged - that is what kinds of patterns they formed - whether noun, verb, noun; determiner, adjective, noun etc.

This method, though simple enough in theory, proved more difficult practically and the programming techniques were made more complex than before. However, before I could even begin this I had to solve another more basic problem, this being the size of each group. Words taken in groups have an incredibly large number of possible combinations which increase very steeply as the number of units in the group is increased. For example, since there are twenty-four word-classes, the number of possible combinations of two words, each of which may be one of the twenty-four word-classes is 24 x 24 which is 576. If the size of the group is three, we have 24^3 or 13,824 possibilities. Four-word patterns give 331,776 combinations. These are the theoretically possible combinations and in practice many of these would be impossible. Words are not individually independent items like sweets which can be arranged in any possible way. Words are dependant on their neighbours and if the first word is "a" or "the", the possibility of the next word being a "noun" is high. Hardly any word is totally unrestricted and this therefore cuts down the number of practical combinations to possibly between one-tenth and one-twentieth of their theoretical possibilities.

The choice of the pattern-size was based on two requirements. It must be large enough to show some of the syntactical structure of the writing, yet small enough to show within a sample of 2,000 words a reasonable distribution of the most frequent and least frequent patterns. If the number of practical possibilities

exceeds the number of words in the sample, it is unlikely that any kind of distribution will result. This therefore rules out four-word patterns. Even if only 5% of the possible 331,776 ways are practical possibilities, that is still about 15,000 likely four-word patterns more than the total number of words in the sample. The decision between two and three-word patterns was based on the fact that two-word patterns can be inferred from three word patterns and are in fact scarcely any advance over single words. Therefore we are left with three word patterns. Since the practical total lies between 600 and 1200 possibilities (or 5% to 10%) there should be room for a sufficient distribution.

Having made this decision, I set about writing the program. It used the same data-decks of numerical class-designators. Beginning with the first word of the sentence (the period being marked by a special number) successive groups of three words are taken, each word in turn, thus making an overlapping sequence, and these are compared with the previous ones, keeping a record of each different pattern and a count of its occurrences. When the end of the sentence is reached, it skips to the beginning of the next. Thus no patterns are recorded which begin with the last or the next to last word of a sentence. For example, taking 98 as the end of sentence mark, the analysis for the following sentence is as follows:-

"The man sat on the floor."

31 01 02 51 31 01 98

The patterns obtained from this would be 310102, 010251, 025131, 513101, a total of four patterns for the six words in the sentence. Therefore in a sample, the number of total patterns is equal to the number of words less twice the number of sentences.

Once the program was running smoothly, I thought the obvious place to start would be with a list of each author's most frequent patterns, and Table XV shows the results. If I had hoped that this table would be a guide to the individual preferences of each author, I was doomed to disappointment for they are in almost complete agreement about the three most frequently used patterns. Where the most popular pattern overall is not in first place, it is invariably second. In fact the three most frequent three-word patterns for each author are included within the four pattern-groups - these being in descending order:

513101 - 'of the man'

310151 - 'the man of'

015131 - 'man of the'

310301 - 'the good man'

As we can see, the first three are all prepositional phrases in some form, which shows the commanding position of the prepositional phrase in the English language.

As individuality is not seen in the choice of favourite pattern of the authors, it might be found in their use of it. Perhaps in the quantitative distribution, we might find some peculiarities. Table XVI shows these results. As we see, the figures are fairly

indistinguishably bunched in both sets of samples, so even in the frequency distribution, the favourite pattern is of little help. The comparison figures are very scattered although taken in pairs per author, the variation is not quite so great. All the comparison authors bar Richardson have higher values than Miss Austen. The range for her is 1.84 which is over 50% of the mean so we see the lack of consistency here. The range of the comparisons is even worse, it representing 70% of the mean. We see therefore that this test adds nothing positive to our purpose.

If we turn from the most frequent pattern, which is the same for all the authors, to the least frequent, we find a considerable choice as the number of patterns that occur only once is large - too large in fact to pin-point a particular pattern to a particular author. Table XVII shows the relationship between the total number of patterns (P), the number of different patterns (D) and the number of patterns which occur only once (U). The D value represents the number of different three-word arrangements of the word-classes which the author makes use of. If its value is high the author would seem to have a more variety pattern of words than if it is low but of course, variety of style can be obtained by many more ways than the syntactic arrangement of words. A large vocabulary, a wide range of sentence-length, variation of mood by pathos, sarcasm etc can achieve great sanity but possibly syntactical variety is more basic than the others. Applying this to ourtable however, little still can be found to distinguish the author.

Miss Austen uses slightly more different patterns but there is really little difference between them. The table is remarkable in its consistency for both sets of results. In samples of 2000 words, roughly 1800-1900 different patterns occur, over 800 occurring only once, and the total pattern varies between 2462 and 2892. The ratio of U/D is fairly constant throughout. There are a few cases which deviate from the general consistency such as the low value of U in sample 8 and the high U in sample 14 but the ratio of U/D remains roughly the same as these samples have a low and high value of D respectively. This ratio of U/D will be constant only when the samples are the same size. Obviously in a very small sample of 20 words, it is possible for all the patterns to occur only once, but as the size of the sample increases, the number of unique patterns and the number of different patterns will increase, but at a decreasing rate and moreover the rates will be different. Therefore there is only a constant value of U/D if samples are the same size.

We noticed that the writers nearly agree with these three favourite patterns. How do they compare for other popular patterns? If all the samples are combined, the ten most frequent patterns are those shown in Table XVIII which shows too the relative rank of each of those ten patterns in each sample. We see the close agreement in the first four places, and the gradual dissimilarity in the rest. On the whole the Austen samples are no more consistent than the comparisons and in fact, in the first four places, they are less so.

The ten most popular patterns are what we might expect - simple everyday constructions as we see from Table XVIIIA which gives the English equivalents.

If we take the 10 most frequent patterns in each sample (not the 10 applicable to all samples) and combine them, we have a total of 22 patterns. This means that there is such a big overlap that out of a possible 140 patterns, we only require 22. Table XIX a and b show us these 22 patterns and their occurrence in each sample as a percentage of the total patterns.

Having finished the tests involved in three-word patterns, it remains for me to insert here a few miscellaneous tests. The 1st element of each sentence was calculated earlier and here I will show the results of calculating the last element in Table XX. Just as the opening element gave us little information, this table too is disappointing. It is of little use as a means of criteria as all the samples are similar in using nouns most often to close sentences. This is obviously a linguistic standard or norm of the language. Similarly, a large number of word-classes never end a sentence. Certain types of words by their very nature are unsuitable for this, providing a function dependent on another word. Such words include the connectives, classes 41-45, although sample 5 does have a low percentage of subordinators (42). This can occur in speech, both direct and reported. The classes most frequently used are nouns (01), finite verbs (02), pronouns (11) and to a lesser extent function adverbs (34). As I have said

however, the results are disappointing. I had hoped that this test would prove interesting, because unlike the opening sentence elements, this is much more likely to be a subconscious choice on the part of the author.

A more fruitful test in the case of Jane Austen is seen in Table XXI. Throughout Miss Austen's work, one cannot help being struck by the number of times a name occurs, and even more so by the number of times she uses the full name - e.g. Mr. Darcy, Edmund Bertram, Lady Russell. I decided to count the frequency with which she uses these full names and compare her results with those of the comparison authors. As I had expected, the computer verified my observation. Miss Austen's count for 0808 is very much larger than any of the comparison samples, only Smollett's first sample being comparable to her. Such a difference is striking if we consider how small the percentages are and this therefore is an important criterion for distinguishing Miss Austen.

It is interesting to compare this table with Table XXII which shows the percentage of pronouns and name elements as introductory words in the sentence. This table shows the consistently high values of Miss Austen in the 08 count, rather than the 11, and this agrees implicitly with the last test which gave the total count of 0808. It must be remembered that word-class 08 includes such things as quotations, foreign words, names of places, but they must make up less than 1% of total words. For the most part, class 08 consists of the names of characters and Miss Austen has

shown a much greater liking for those than any other author. Her use of pronouns as opening elements in the sentence is more frequent than the comparison authors on the whole, but less strikingly so. Both Defoe and Richardson have as high values for this count as she does, yet taking the introductory pronouns and names together, Miss Austen's total makes up 35% of the total introductory words - a large percentage indeed.

VII

CLUSTER ANALYSIS

It was suggested to me during my research that I might make use of cluster analysis to try and draw some meaningful results from my data. I used a series of Fortran IV programs called CLUSTAN, which was developed in the University of St. Andrews by David Wishart and his advice and assistance has been invaluable in this aspect of my work.

Basically the theory behind these programs assumes a population made up of individuals with different characteristics and it uses various tests to arrange these individuals into groups, or clusters, according to these characteristics. In my case, the individuals were the six Austen and eight comparison samples used throughout this thesis, together with another six Austen and six comparisons samples, making twenty-six in all. Table XXV lists these samples. Four of these comparison samples were twentieth century writers and they were included to provide a point of reference by which the consistency of the earlier works could be judged. The "characteristics" of these samples were the results of the various tests I had carried out previously. Table XXIII shows these tests and the results obtained. These "characteristics" were then punched onto cards one set corresponding to one sample, and this data deck was computed with the various CLUSTAN programs.

After the program file was set up the program CORREL was computed. This obtains the similarity matrix and K-linkage lists

from the data. These results were then stored on a data file as they are necessary for the executions of programs HIERAR, KDEND, MODE, CENTRO AND DNDRIT. The similarity matrix is a triangular array of N * (N - 1)/2 coefficients, such that each element measures the similarity between two individuals (N is the number of individuals). The K-linkage lists are the lists of the nearest neighbours for all N individuals, and the number of lists required can be specified.

Next, I used the program KDEND which uses the Cole - Wishart algorithm to find K-partition clusters, according to the method of Jardine and Sibson. A similarity matrix for a population of N individuals is computed, and a linkage parameter, K, and a similarity threshold H are chosen. Each individual is represented by a node on a graph and all pairs of nodes which correspond to pairs of objects having a similarity of at least H are connected. My samples were linked with first single, then double, treble and finally quadruple links. The first results show the linking of those individuals with the greatest similarity in single links. Most remain single and unlinked and very few clusters are made, then as the conditions of similarity are slackened, more and more individuals form into clusters until the similarity conditions have widened enough to form all the individuals into one cluster.

Several clusters form early, 17 and 18; 2, 4 and 7; and 9 and 10. This is not surprising but soon 1 and 16 join together, 1 being an Austen and 16 a non Austen samples. As they join together

so soon, they must have striking similarities. The most interesting grouping in the single-linking occurs at level 9 where we have 4 clusters. They are all predictable except for 16, and later 17 and 18 join this group. Of the Austen samples, 6, 8 and 11 are slow to cluster, particularly 6 which is always last to join with other groups, but on the whole, the Austen samples cluster more quickly than the comparison samples, thus showing great consistency.

With K = 2 (double-linking) the same pattern is repeated. 17 and 18; 4 and 7; and 4 also joins with 2 (we had 2, 4 and 7 in one cluster above). This is, of course, quite possible in double-linking. 1 again links with 16, but also with 5 and 12, showing its cohesion to the Austen group as well as this irregular similarity to sample 16. One inference from this is that 16 could be similar to the Austen group, rather than 1 being dissimilar to it. It is probably true however that 1 is different from the other Austen novels, and I will try to explain this in literary terms later.

As before, 6, 8 and 11 are late among the Austen samples in clustering, 6 being last once again.

This pattern is repeated for K = 3 and K = 4, the only difference being that the threshold coefficient level gets higher and higher and as expected the samples take longer to finally unite into a single cluster.

I then used the program HIERAR which is a hierarchic fusion using eight different methods. We start with N clusters, each containing a single individual, and being numbered accordingly.

In each of the (N - 1) fusion steps, the two clusters which are most similar are combined and the resulting union cluster is given the number of the lesser of the two clusters. This continues until all the fusions are made and the sequence is summarised in a "Dendogram Table". The corresponding dendogram can then be easily drawn by hand but this program also punches out a deck of cards which can be used as input to the program PLINK which draws the dendogram on the graph plotter. Table A gives these results.

The dendogram shows the results that could be expected after examining the previous results of program KDEND. 1 and 16 join together and 3, 5, 8 and 11 form a nice cluster. Eventually 1 joins this group. 2, 4, 7, 9 and 10 form a compact cluster and finally 6 joins this group. Not until the end are these two separate groups united. Of the other compact group, the 4 twentieth-century authors, 2, 3, 24 and 25 cluster quickly but right at the opposite end and linking only at the end is 26. As we see in the program KDEND, this was the last sample to cluster. These results, therefore, although obtained in completely different ways, support our earlier findings.

Another useful program, employing the graph plotter is the program SCAT, which plots X - Y scatter diagrams with X and Y being in this case, principal components. Basically, principal components analysis is an easy way of representing an N-dimensional space on a 2-dimensional plane. During the analysis, the original set of

variables is transformed into a new set of orthogonal variables. The new variables are all linear combinations of the original variables and these are the same number of each. Thus in my case, the thirty-five variables of the 26 samples will produce thirty-five new variables (the components). These principal components contain successively the maximum possible variance between the samples. The first principal component is designed to contain as much as possible of the variance. The second principal component is then designed to contain as much as possible of the variance which remains after the variance explained by the first principal component has been removed. We can then take a third principal component and continue until all the variance has been accounted for.

My first scatter diagram, Table B_1, shows principal component 1 plotted against principal component 2. This shows a very good clustering of the Austen samples 1-12 and also a cluster of the twentieth century authors - samples 23-26.

I then plotted principal component 1 against principal component 3 (Table B_2). If we take these two graphs together and imagine the second lying in a vertical plane to the first one, we can get a 3-dimensional effect. The Austen samples again fall into a neat cluster, but the twentieth century samples are scattered among the other comparison samples. This suggests (but is by no means conclusive) that this method is better for identifying individual authors than periods and that cluster analysis might be

very useful in stylistic discrimination. It also suggests the consistency of Miss Austen's style for her to have such compact results. As expected, sample 6 is more loosely connected than the other and also sample 1 is nearer to sample 16 than any other non-Austen sample.

RELOC is a program by which the population is regrouped by a method called iterative relocation. It attempts to re-arrange the clusters into the best possible combination for the selected similarity criterion. Having executed this program, I followed it with the program RESULT which uses the information provided by RELOC and computes the diagnostic statistics for the new clusters.

The results are hardly different, however, from the previous ones. The K-linkage lists show the five nearest neighbours for each sample in descending order of similarity and it is these lists which provide the clearest results for my purpose. Table C shows these results.

Sample 1 is still most similar to sample 16 (and likewise 16 is most similar to 1, although of course, this does not necessarily follow), and the other neighbours of sample 1 are split between Austen sample (5 and 12) and non-Austen (17 and 24). The other Austen samples however show a strong preference for their own kind. Of the other samples 2-12, four have all 5 neighbours in the Austen range (2, 4, 5 and 11), six have one outsider (3, 6, 7, 8, 9, 10)

and sample 12 has two outsiders. Of these outsiders, none occupy the nearest neighbour postion and only one in sample 2 is in second position. Sample 1 is nearest neighbour to several samples both Austen and non-Austen, all of them bar 16 being a neighbour of 1 itself. Of the other Austen samples, only 9 and 10 appear as nearest neighbour of a non-Austen sample. We see therefore that the Austen range is very tightly bound together except for sample 1 which without doubt is the black sheep of the family.

As expected the group of moderns - samples 23-26 are fairly closely knit for such a small group though less so than the Austen range. Samples 23 and 25 are very similar being nearest neighbour to each other and samples 24 and 26 have their second nearest neighbour within their own group.

Cluster analysis therefore has proved very interesting for this study of Jane Austen's style. The results clearly indicate the similarities shared by all members of Austen, though sample 1 is obviously less closely knit. It also indicates that an author is easier to recognise by his consistency than a period. The eighteenth century, represented by samples 13-22 showed results which were widely distributed and a central point of similarity was difficult to see. The twentieth century samples 23-26 looked more consistent but it is difficult to say how similar they are to each other. It is more probable that they seem superficially similar

due to the tremendous difference, both sociological and linguistic, from the other samples.

One interesting point to emerge from this analysis is that within the Austen range, the samples from the same novel are no more similar than samples from different novels. This is a clear indication that Miss Austen is consistent throughout the range of her novels, and it is not limited to similarities within the same book. This is seen in the passages from Pride and Prejudice found in samples 3, 7, 8 and 9. In all the tests, these samples are no more similar to each other, than to any other Austen sample.

As we have observed, sample 1 clearly does not fall into the general pattern of the other Austen samples. To form my reasons for this, I looked beyond the actual passage to the literary history of Jane Austen's novels. Sample 1 is a passage from Northanger Abbey, a novel which has always raised great controversy, even its date of composition being doubtful. Most critics agree, however, that it was her first novel, being written between 1797 and 1803, although it was not published until after her death. It is probably the least popular of her novels and certainly the most dated.

Miss Austen's immediate literary predecessors, such as Mrs. Radcliffe and her followers, wrote a new kind of literature, known as Gothic and sentimental novels. These works were escapist novels, far removed from reality, and were probably the origin of the modern horror novels, being full of gloomy castles, hidden passages, missing wills etc. When Miss Austen started writing Northanger Abbey, she

was concerned in part with making a burlesque of these novels. This aim carries her through most of her novel and it is only occasionally that Miss Austen shows the true insight into her characters that is so much a mark of her later novels. This therefore in my opinion, accounts for the distinction between sample 1 and the other samples, all of which are taken from her more mature works.

With this in mind, I thought it might prove interesting to run these cluster programs again, this time omitting sample 1 from the data, thus leaving 25 samples, 1-11 the Austen group and 12-25 non-Austen. The characteristics remain the same as before.

The results show that sample 15 (=sample 16 in previous run) now clusters with a non-Austen sample, 13, although it still shows its affinity to the Austen group by having sample 3 as its second-nearest neighbour. (See Table D). With Sample 1 omitted, however, the results fall neatly into the two categories, Austen and non-Austen, with the four "modern" samples making a tidy sub-group in the latter.

Another experiment I tried was to run the programs with the original 26 samples but to omit certain characteristics. I chose to leave out the 9 variables which showed the least difference between the Austen and non-Austen samples, these being the word-classes 05, 32, 34, 44 and 45, the Parts of Speech, Auxiliaries, Verbals, and the count of the most frequent pattern 513101, in each case the difference between the means being less than 1%.

With this new set of data, I ran the CLUSTAN programs yet again. This time sample 1 does not behave abnormally. In the K-linkage lists, four out of the five nearest neighbours are Austen sample, and only in samples 10 and 11 does sample 1 not appear on the list of nearest neighbours for the Austen group. This clearly shows the affinity between Sample 1 and the rest of the Austen group. Table E shows these results.

In the program KDEND, Sample 1 yet again shows its closeness to the Austen group. Again it is closest to sample 4 and 2 and continually clusters with them. Little else in the results has changed, however. As we might have foreseen from the original run of the program, sample 6 is less closely bound to the Austen group than the other samples. It was always the last to cluster and this is still true.

On the scatter diagrams, we see how much more closely sample 1 is integrated with the Austen samples than previously, both for principal components 1 and 2 and also for 1 and 3. Tables F_1 and F_2 show these results.

This last experiment therefore has shown that Sample 1 is not greatly different from the Austen group, and it suggests that Northanger Abbey, though over-concerned with mocking the Gothic novels, still has all the characteristics that we expect in a work of Jane Austen.

VIII

CONCLUSION

Many tests have been applied throughout this thesis in search of a criterion of distinction between Miss Austen and her predecessors. I have shown both successes and failures, the latter occurring when we meet linguistic constants, stable aspects of the language itself. These, in themselves, can be very valuable, but I do not dwell on them here as they are outside my purpose, which is concerned with details, not generalities.

The details of individual preferences found in the previous tests are summarised and categorised in Table XXIV. This gives a list of the eleven most meaningful tests and shows how each author ranked in these tests and his respective values. This table is helpful not for the values but for the relative ranks of Miss Austen and the other authors.

It is interesting to note from the table that each author, except Fielding is found at the extremes on five occasions, Fielding being central every time and in fact being in the exact centre, 3rd position, on five occasions. Of the four comparison authors, Miss Austen is in contiguous places with Richardson on five occasions, with Fielding on four occasions and with Defoe and Smollett on three. We see from this how well integrated she is with the eighteenth century, yet as the cluster analysis has shown, she differs from them. This table seems to me to show the

true nature of her genius. A lesser critic of Miss Austen's era might have selected one author from whom he drew his inspiration. Miss Austen encompassed the whole of the eighteenth century in her work, rather than relying on a single member of it, and it is in the blending of their styles, joined to her own creative ability, that makes her work so successful.

I have not attempted in this essay to try and write anything like a full critical analysis of Jane Austen. That has been done many times before me by greater authorities than myself. Nor have I attempted a full-proof analytical survey of every component in her style. That too has been left for others to tackle. My aim in this thesis was to steer the analysis of her style away from the vague and woolly and towards greater objectivity and precision, consistent with the nature of the problem. The first chapters were therefore dedicated to revealing the shortcomings of the usual methods and to point out the difficulties facing the critic when dealing with a topic such as style.

I hope to have shown how the mind and personality of an author shows in his writing and that the criteria which determine this can to some extent be identified and measured. Miss Austen has shown herself to be too like her eighteenth century precedessors on many occasions for these criteria to be very different, but I think I have shown a significant difference, large enough for the method I used, to be acceptable. The similarities of the five authors must show the prevalent trend of writing at that time and must also show

certain literary constants of the period. The four modern authors, as we have seen in cluster analysis make quite a distinct group. They show the progress of our language through the centuries, although critics argue whether this is a "royal progress" - that is merely an advance in time and not in worth. It is perhaps surprising that there is not a greater difference between the style of the eighteenth and twentieth centuries when we remember that the former saw the rise of the novel from nothing and the four authors we have studied were all pioneers in their own way. Nothing like these novels had ever been conceived of before. The twentieth century authors, however, came to the novel with two hundred years of experience behind them, and yet nothing strikingly different is seen in these tests. The subject-matter of course has been broadened to include any topic under the sun, the last novelist to keep within Jane Austen's restricted sphere being Henry James. The very heart of the language, however, the syntax and arrangement of word-groups have stayed remarkably constant for two hundred years. I cannot think of anything else which has survived the same space of time with so little change.

Thus I have hoped to show how a quantitative method can be applied to literature and the sort of conclusions which may emerge from the close examination of the texts. For different authors, the categories I have used may need redefining to bring out even more of the subtleties of language and they may vary with the individual authors to bring out their own peculiar idiosyncrasies.

In order to get an overall picture, however, I decided not to over-complicate the picture and I leave it to those who follow to carry this study to their own conclusions.

BIBLIOGRAPHY.

Allen, W.	(1954)	The English novel.
	(1955)	Six great novelists.
Baker, E.A.	(1924)	The history of the English novel.
Forster, E.M.	(1962)	Aspects of the novel.
Fries, C.C.	(1952)	The structure of English.
Lascelles, M.M.	(1939)	Jane Austen and her art.
Milic, L.T.	(1954)	A quantitative approach to the style of Jonathan Swift.
Richards, I.A.	(1925)	Principles of literary criticism.
Watt, I.P.	(1963)	The rise of the novel.

IX

BASIC PROGRAMS.

The basic programs for all these group counts is the same and only slight variations are required to adjust it to the individual requirements, such as the Nominal-Verbal Ratio, or the number of connectives.

I read my data on to disk so that I could use the one set of data for all the programs consecutively, and to save space, the data from the samples was read in half-word integers. Below is the basic program for the distribution count of all the word-classes.

After the dimensions, integers and real numbers are set up, the word-class types 01, 02, 03 ----- 91, and also the end-of sentence mark 98 are read into the first column of an array, GR (1,I) with one space for each word-class. The other columns of this array are used to store the values of each word-classfor each sample. The text is then read in from disk, 40 numbers at a time and the first number is compared with each word-class in turn. When it reaches the word-class to which it belongs, the count for that word-class, GR (L,I), is increased by 1.

The other 39 numbers are treated similarly and when all have been compared, the next card is read and the process continues. This loop is only stopped when we read the end-of-sample mark (99). We then branch out and the percentages of the totals in each word-class for that sample are computed. When all the samples in the

Jane Austen group have been computed, the results are printed out and the whole process is repeated for the comparison set of results.

This is a very simple account of the program and readers with experience of programming will see from the program itself that I have accounted for setting up the disk, incrementing the record-marker on the disk, setting initial values to zero, and keeping a count of the number of the words in each sample (KNT) in order to compute the percentages.

The program for the three-word patterns is similar to the word-class count but obviously more complex. The numbers are still read in from disk 40 at a time. The first three are placed in an array, STAK, making the first three-word pattern. The rest of the numbers, in overlapping groups of three (e.g. 123, 234, 345) are then compared with the contents of STAK to see if they constitute a pattern already in STAK. If so, the count for the pattern STAK (JJ, 4) is increased by 1. If not, we test to see if N and JJ the two pointers are the same. If they are, then that pattern is unique and its frequency count is 1. If JJ does not equal N, it is incremented by 1 and the process is repeated. The process continues until the end-of-sample mark is reached.

The program raised more problems than the previous ones. Since the numbers are compared three at a time in overlapping groups, it was necessary to use 42 columns in the array. The first card was read into columns 3 to 42 and when we reached the end of the card,

the last two numbers became the first two numbers of the next card. Otherwise these numbers would never have begun a new pattern themselves. Another problem was the end of each sentence. I did not want patterns running on from one sentence to the next, therefore the second last and last words of a sentence never begin a new pattern. In order to range the numbers in descending order of frequency, I use a subroutine MAX. The first value in the array is set to LARGE, and J is set to 1. We test if N is equal to 1 and if so we return. If not we continue to next number in array and test if it is less than LARGE. If so we continue until we reach a number which is greater than LARGE. LARGE is then replaced by that value and J is given the value which I is at. We return to the main program and the maximum value, LARGE, is printed, together with the pattern it represents. This is then replaced by the last element in the array. This means that only N-1 elements remain to be considered and they are in the first N-1 positions. We then call MAX again, and the largest value among the N-1 elements is calculated and printed, then replaced by the second last element in the array. This process is continued until the twenty largest numbers and their corresponding patterns have been printed.

SAMPLE PROGRAM FOR GROUP COUNT

```
      C       PAM SCOTT MSC  GROUP COUNT
      C       SAMPLE  1 NORTHANGER ABBEY
      C       SAMPLE  2 SENSE AND SENSIBILITY
      C       SAMPLE  3 PRIDE AND PREJUDICE
      C       SAMPLE  4 MANSFIELD PARK
      C       SAMPLE  5 EMMA
      C       SAMPLE  6 PERSUASION
      C       SAMPLE  7 JOURNAL OF PLAGUE YEAR --- DEFOE
      C       SAMPLE  8 ROBINSON CRUSOE --- DEFOE
      C       SAMPLE  9 PAMELA --- RICHARDSON
      C       SAMPLE 10 CLARISSA --- RICHARDSON
      C       SAMPLE 11 JOSEPH ANDREWS --- FIELDING
      C       SAMPLE 12 TOM JONES --- FIELDING
      C       SAMPLE 15 HUMPHREY CLINKER --- SMOLLETT
      C       SAMPLE 14 PEREGRINE PICKLE --- SMOLLETT
0001          DIMENSION TEXT(40),GR(14,25),GRF(14,25),SUMS(25),
             1 MEAN(25),Z(25),SD(25)
0002          INTEGER*4 GR, TEXT*2
0003          REAL MEAN
0004          DEFINE FILE 2(1000,20,U,ID2)
0005          ID1=1
0006          READ(5,40)(GR(1,I),I=1,25)
0007       40 FORMAT(25I2)
0008          DO 22 KL=1,2
0009          READ(5,51)N
0010       51 FORMAT(I2)
0011          M=N+1
0012          DO 75 L=2,M
0013          KNT=0
0014          DO 100 I=1,25
0015      100 GR(L,I)=0
0016      150 READ(2'ID1)TEXT
0017          DO 120 J=1,40
0018          DO 130 I=1,25
0019          IF (TEXT (J).EQ.99)GO TO 200
0020          IF (TEXT(J).NE.GR(1,I))GO TO 130
0021          GR(L,I)=GR(L,I)+1
0022          KNT=KNT+1
0023          GO TO 120
0024      130 CONTINUE
0025      120 CONTINUE
0026          ID1=ID1+1
0027          GO TO 150
0028      200 X=KNT
0029          GRF(L,I)=GR(L,I)*100/X
0030          ID1=ID1+1
0031       75 CONTINUE
```

```
0032          DO 65 I=1,25
0033          SUMS (I)=0
0034          DO 12 L=2,M
0035       12 SUMS(I)=SUMS(I)+GRF(L,I)
0036          MEAN(I)=SUMS(I)/N
0037          Z(I)=0
0038          DO 13 L=2,M
0039       13 Z(I)=Z(I)+(GRF(L,I)-MEAN(I))**2
0040          SD(I)=SQRT(Z(I)/N)
0041       65 CONTINUE
0042          IF (KL.EQ.2)GO TO 300
0043          WRITE(6,160)
0044      160 FORMAT('1','JANE AUSTEN SAMPLES',//,20X,'S1',
             1 9X,'S2',9X,'S3',9X,'S4',9X,'S5',9X,'S6',9X,
             1 'MEAN',9X,'SD',//)
0045          GO TO 310
0046      300 WRITE (6,210)
0047      210 FORMAT('-','COMPARISON SAMPLES',//,21X,'S1',9X,
             1 'S2',9X,'S3',9X,'S4',9X,'S5',9X,'S6',9X,'S7',
             1 9X,'S8',7X,'MEAN',7X,'SD',//)
0048      310 DO 26 I=1,25
0049          WRITE(6,180)GR(1,I),(GRF(L,I),L=2,M),MEAN(I),SD(I)
0050      180 FORMAT(' GR',I2,10(4X,F6.2))
0051       26 CONTINUE
0052       22 CONTINUE
0053          STOP
0054          END
```

SAMPLE PROGRAM FOR 3-WORD PATTERNS

```
      C       P. SCOTT MSC  GROUPS OF THREE WORDS
0001          DIMENSION TEXT(42),STAK(1500,4),SAVE(1500)
0002          INTEGER*4 STAK,SAVE,TEXT*2
0003          DEFINE FILE 2(1000,20,U,ID2)
0004          ID1=1
0005          DO 22 KL=1,2
0006          READ(5,31)NN
0007       31 FORMAT(I2)
0008          M=NN+1
0009          DO 75 L=2,M
0010          N=1
0011          DO 8 JJ=1,1500
0012        8 STAK(JJ,4)=0
0013          J=3
0014          K=4
0015        2 READ(2'ID1)TEXT(M),M=3,42)
0016          DO 15 I=1,3
0017          STAK(N,I)=0
0018       15 STAK(N,I)=TEXT(J+I-1)
0019          DO 9 J=K,41
0020          DO 7 JJ=1,N
0021          DO 4 I=1,3
0022          IF (TEXT(J+I-1).EQ.99)GO TO 200
0023          IF (TEXT(J+2).EQ.98)J=J+3
0024        4 IF (STAK(JJ,I).NE.TEXT(J+I-1))GO TO 3
0025          STAK(JJ,4)=STAK(JJ,4)+1
0026          GO TO 9
0027        3 IF(JJ.NE.N)GO TO 7
0028          N=N+1
0029          DO 25 I=1,3
0030       25 STAK(N,I)=TEXT(J+I-1)
0031          STAK(JJ,4)=1
0032        7 CONTINUE
0033        9 CONTINUE
0034          J=1
0035          K=1
0036          TEXT(1)=TEXT(41)
0037          TEXT(2)=TEXT(42)
0038          ID1=ID1+1
0039          GO TO 2
0040      200 DO 24 JJ=1,N
0041          SAVE(JJ)=0
0042          SAVE(JJ)=STAK(JJ,4)
0043       24 CONTINUE
0044          DO 17 JJ=1,20
0045          CALL MAX(N-JJ+1,SAVE,JK, LARGE)
0046          WRITE(6,180)LARGE,(STAK(JK,I),I=1,3)
```

```
0047    180  FORMAT(' ',I3,10X,I2,I2,I2,/)
0048         SAVE(JK)=SAVE(N-JJ+1)
0049     17  CONTINUE
0050         ID1=ID1+1
0051     75  CONTINUE
0052     22  CONTINUE
0053         STOP
0054         END

0001         SUBROUTINE MAX(N,X,J,LARGE)
0002         INTEGER X(1500)
0003         LARGE=X(1)
0004         J=1
0005         IF(N.EQ.1)GO TO 7
0006         DO 6 I=2,N
0007         IF (X(I).LE.LARGE)GO TO 6
0008         LARGE=X(I)
0009         J=I
0010      6  CONTINUE
0011      7  RETURN
0012         END
```

Table III - Comparison of word-class frequency distributions between two sets of one-tenth samples (200 words) of Pride and Prejudice and Mansfield Park.

Group	1A	1B	Diff.	2A	2B	Diff.
01	7.54	16.08	-8.54	19.70	13.07	6.63
02	8.04	7.04	1.01	2.02	4.02	-2.00
03	4.02	6.03	-2.01	7.58	4.02	3.56
04	4.52	1.01	3.52	1.01	1.01	0.01
05	2.51	3.52	-1.01	2.02	1.51	0.51
06	1.01	1.01	0.00	0.00	0.00	0.00
07	8.54	3.52	5.03	5.05	8.04	-2.99
08	5.53	5.03	0.50	5.05	11.06	-6.00
11	13.07	8.04	5.03	5.05	9.05	-3.99
21	9.05	4.02	5.03	6.57	7.04	-0.47
31	6.53	15.58	-9.05	16.16	10.05	6.11
32	1.01	0.00	1.01	0.51	0.00	0.51
33	3.52	1.57	2.01	2.53	1.51	1.02
34	1.51	0.50	1.01	0.51	5.03	-4.52
41	4.52	2.01	2.51	7.07	5.03	2.05
42	0.00	1.51	-1.51	1.52	0.50	1.01
43	5.03	5.03	0.00	2.53	3.52	-0.99
44	0.50	0.00	0.50	0.00	0.00	0.00
45	0.00	0.50	-0.50	0.00	0.00	0.00
51	7.04	14.07	-7.04	9.09	11.56	-2.47
61	0.00	2.51	-2.51	0.00	0.00	0.00
71	1.01	0.00	1.01	0.00	0.00	0.00
81	0.00	0.00	0.00	3.54	1.01	2.53
91	0.50	0.50	0.00	0.00	0.00	0.00
98	5.03	1.01	4.02	2.53	3.02	-0.49

TABLE IV - Word-class frequency distribution of all the whole samples showing Mean and Standard Deviation.

JANE AUSTEN SAMPLES

	S1	S2	S3	S4	S5	S6	Mean	SD
GR 1	15.85	14.91	11.06	13.76	12.71	16.00	14.05	1.76
GR 2	6.30	5.87	6.40	5.90	7.30	2.93	5.79	1.36
GR 3	5.95	7.31	5.20	5.80	6.50	6.34	6.18	0.65
GR 4	1.63	1.88	1.80	2.25	2.35	1.80	1.95	0.26
GR 5	2.03	1.97	2.15	2.45	1.80	2.27	2.11	0.21
GR 6	0.05	0.29	0.35	0.00	0.40	0.19	0.21	0.15
GR 7	3.81	4.09	4.35	4.85	3.65	6.34	4.52	0.90
GR 8	3.30	5.92	4.20	5.15	4.35	5.06	4.66	0.83
GR11	8.99	7.50	10.36	8.60	11.91	6.25	8.93	1.84
GR21	6.91	6.49	8.70	7.00	8.05	7.43	7.43	0.75
GR31	13.21	13.13	12.21	13.71	11.11	13.87	12.87	0.95
GR32	0.91	0.63	0.65	1.05	0.75	0.90	0.82	0.15
GR33	2.18	3.03	3.05	2.90	2.95	2.84	2.83	0.30
GR34	3.76	1.97	3.35	1.95	3.30	2.41	2.79	0.71
GR41	3.71	4.52	3.45	4.75	3.60	5.30	4.22	0.68
GR42	1.07	1.39	1.05	1.20	1.20	1.14	1.17	0.11
GR43	1.42	2.21	2.25	2.25	1.35	2.18	1.94	0.40
GR44	0.66	0.29	0.60	0.05	0.35	0.09	0.34	0.23
GR45	0.81	0.29	0.45	0.55	0.85	0.71	0.61	0.20
GR51	10.77	9.96	8.65	9.85	9.35	9.65	9.71	0.64
GR61	0.76	1.15	1.90	1.45	0.85	0.38	1.08	0.49
GR71	0.20	0.10	0.85	0.00	0.40	0.09	0.27	0.29
GR81	0.97	0.82	0.65	0.75	0.50	2.32	1.00	0.61
GR91	0.56	0.58	0.35	0.35	0.15	0.38	0.39	0.14
GR98	4.17	3.70	5.95	3.40	4.25	3.12	4.10	0.92

TABLE IV cont'd

COMPARISON SAMPLES

	S1	S2	S3	S4	S5	S6	S7	S8	Mean	SD
GR 1	17.00	15.02	12.56	15.17	16.86	16.43	18.16	21.04	16.53	2.33
GR 2	5.97	8.31	8.61	6.08	6.40	5.59	6.30	5.59	6.61	1.11
GR 3	4.65	4.80	6.54	5.82	4.75	4.86	4.55	4.64	5.08	0.67
GR 4	0.82	1.35	1.01	0.66	1.70	1.91	0.55	0.85	1.11	0.46
GR 5	2.40	2.20	2.07	2.30	2.20	2.16	2.50	1.49	2.17	0.29
GR 6	0.00	0.10	1.06	0.51	0.20	0.00	0.75	0.00	0.33	0.38
GR 7	4.49	3.55	3.27	3.47	4.70	5.10	3.90	6.23	4.34	0.94
GR 8	0.77	1.30	1.83	3.06	3.65	2.26	5.05	2.50	2.55	1.28
GR11	7.20	11.16	12.95	10.01	7.90	7.80	8.65	4.85	8.81	2.35
GR21	5.87	5.86	6.64	6.23	5.80	6.38	5.75	4.42	5.87	0.62
GR31	16.49	14.56	12.03	15.02	14.91	15.94	14.96	18.33	15.28	1.68
GR32	1.43	0.80	0.63	0.87	0.55	0.29	1.20	0.43	0.77	0.36
GR33	2.25	2.00	2.12	2.25	2.10	2.06	1.10	1.60	1.93	0.37
GR34	3.01	3.50	2.31	3.32	3.25	3.38	1.60	2.13	2.81	0.66
GR41	6.53	6.16	6.40	4.44	3.30	4.02	5.35	3.41	4.95	1.25
GR42	1.79	2.45	2.60	1.84	1.40	1.57	1.00	0.91	1.69	0.57
GR43	3.27	2.30	2.31	1.94	3.00	3.53	1.25	3.68	2.66	0.79
GR44	0.10	0.30	0.43	0.46	0.15	0.15	0.45	0.00	0.26	0.17
GR45	0.51	0.40	0.29	0.51	1.15	0.93	0.40	0.32	0.56	0.29
GR51	11.13	10.21	8.81	9.04	11.81	11.23	10.96	14.28	10.93	1.61
GR61	1.12	0.50	0.63	1.28	0.75	0.15	0.65	0.21	0.66	0.37
GR71	0.10	0.25	0.58	0.31	0.00	0.10	0.05	0.00	0.17	0.18
GR81	1.23	0.60	0.58	0.87	0.90	1.28	0.70	0.80	0.87	0.25
GR91	0.41	0.50	0.87	0.72	0.65	0.54	0.15	0.48	0.54	0.20
GR98	1.48	1.80	2.89	3.83	1.90	2.35	4.00	1.81	2.51	0.91

TABLE V - Means and Standard Deviations of Austen and comparison samples, each taken as a separate population.

	Jane Austen		Comparisons	
Group	Mean	SD	Mean	SD
GR 1	14.05	1.76	16.53	2.33
GR 2	5.79	1.36	6.61	1.11
GR 3	6.18	0.65	5.08	0.67
GR 4	1.95	0.26	1.11	0.46
GR 5	2.11	0.21	2.17	0.29
GR 6	0.21	0.15	0.33	0.38
GR 7	4.52	0.90	4.34	0.94
GR 8	4.66	0.83	2.55	1.28
GR11	8.93	1.84	8.81	2.35
GR21	7.43	0.75	5.87	0.62
GR31	12.87	0.95	15.28	1.68
GR32	0.82	0.15	0.77	0.36
GR33	2.83	0.30	1.93	0.37
GR34	2.79	0.71	2.81	0.66
GR41	4.22	0.68	4.95	1.25
GR42	1.17	0.11	1.69	0.57
GR43	1.94	0.40	2.66	0.79
GR44	0.34	0.23	0.26	0.17
GR45	0.61	0.20	0.56	0.29
GR51	9.71	0.64	10.93	1.61
GR61	1.08	0.49	0.66	0.37
GR71	0.27	0.29	0.17	0.18
GR81	1.00	0.61	0.87	0.25
GR91	0.39	0.14	0.54	0.20
GR98	4.10	0.92	2.51	0.91

Table VI - Grouping of word-classes into Parts of Speech (P/S) and Function Words (FW) for both sets of samples, showing mean and standard deviation.

JANE AUSTEN

	1	2	3	4	5	6	Mean	S.D.
P/S	38.92	42.23	35.52	40.17	39.07	40.94	39.47	2.10
FW	56.91	54.06	58.53	56.43	56.68	55.94	56.42	1.33

COMPARISONS

	1	2	3	4	5	6	7	8	Mean	S.D.
P/S	36.09	36.64	36.96	37.08	40.47	38.30	41.77	42.35	38.71	2.31
FW	62.43	61.56	60.15	59.09	57.63	59.34	54.23	55.83	58.78	2.60

Table VII - Connectives as individual word classes and grouped (CONN), in percentages showing mean and range.

AUSTEN

Class	1	2	3	4	5	6	Mean	Range
41	3.71	4.52	3.45	4.75	3.60	5.30	4.22	1.85
42	1.07	1.39	1.05	1.20	1.20	1.14	1.18	0.34
43	1.42	2.21	2.25	2.25	1.35	2.18	1.94	0.90
44	0.66	0.29	0.60	0.05	0.35	0.09	0.34	0.61
45	0.81	0.29	0.45	0.55	0.85	0.71	0.61	0.56
CONN	7.67	8.71	7.80	8.80	7.35	9.42	8.29	2.07

COMPARISONS

Class	1	2	3	4	5	6	7	8	Mean	Range
41	6.53	6.16	6.40	4.44	3.30	4.02	5.35	3.41	4.95	3.23
42	1.79	2.45	2.60	1.84	1.40	1.57	1.00	0.91	1.70	1.69
43	3.27	2.30	2.31	1.94	3.00	3.53	1.25	3.68	2.66	1.43
44	0.10	0.30	0.43	0.46	0.15	0.15	0.45	0.00	0.26	0.36
45	0.51	0.40	0.29	0.51	1.15	0.93	0.40	0.32	0.48	0.86
CONN	12.20	11.61	12.03	9.19	9.00	10.20	8.45	8.31	10.13	3.89

Table VIII - Modifiers as individual word-classes and grouped (MOD) in percentages, showing mean and range.

AUSTEN

Class	1	2	3	4	5	6	Mean	Range
31	13.21	13.13	12.21	13.71	11.11	13.87	12.87	2.76
32	0.91	0.63	0.65	1.05	0.75	0.90	0.82	0.42
33	2.18	3.03	3.05	2.90	2.95	2.84	2.83	0.87
34	3.76	1.97	3.35	1.95	3.30	2.41	2.79	1.81
MOD	20.07	18.76	19.26	19.61	18.11	20.02	19.30	1.96

COMPARISONS

Class	1	2	3	4	5	6	7	8	Mean	Range
31	16.49	14.56	12.03	15.02	14.91	15.94	14.96	18.33	15.28	6.30
32	1.43	0.80	0.63	0.87	0.55	0.29	1.20	0.43	0.78	1.14
33	2.25	2.00	2.12	2.25	2.10	2.06	1.10	1.60	1.94	1.15
34	3.01	3.50	2.31	3.32	3.25	3.38	1.60	2.13	2.81	1.90
MOD	23.18	20.87	17.08	21.45	20.81	21.68	18.86	22.48	20.80	6.10

Table IX - Finite verbs and auxiliaries as individual word-classes and grouped (VA) in percentages, showing means and range.

AUSTEN

	1	2	3	4	5	6	Mean	Range
02	6.30	5.87	6.40	5.90	7.30	2.93	5.79	4.37
21	6.91	6.49	8.70	7.00	8.05	7.43	7.43	2.21
VA	13.21	12.36	15.11	12.91	15.36	10.36	13.22	5.00

COMPARISONS

	1	2	3	4	5	6	7	8	Mean	Range
02	5.97	8.31	8.61	6.08	6.40	5.59	6.30	5.59	6.61	3.02
21	5.87	5.86	6.64	6.23	5.80	6.38	5.75	4.42	5.87	2.22
VA	11.84	14.16	15.26	12.31	12.21	11.97	12.06	10.02	12.48	5.24

Table X - Verbals as individual word-classes and grouped (VB) in percentages, showing mean and range.

AUSTEN

	1	2	3	4	5	6	Mean	Range
05	2.03	1.97	2.15	2.45	1.80	2.27	2.16	0.65
06	0.05	1.29	1.35	1.00	0.40	0.19	0.21	0.40
07	3.81	4.09	4.35	4.85	3.65	6.34	4.52	2.69
VB	5.89	6.35	6.85	7.30	5.85	8.80	6.84	2.95

COMPARISONS

	1	2	3	4	5	6	7	8	Mean	Range
05	2.40	2.20	2.07	2.30	2.20	2.16	2.50	1.49	2.16	1.01
06	0.00	0.10	1.06	0.51	0.20	0.00	0.75	0.00	0.33	1.06
07	4.49	3.55	3.27	3.47	4.70	5.10	3.90	6.23	4.34	2.76
VB	6.89	5.86	6.40	6.28	7.10	7.26	7.15	7.73	6.83	1.87

TABLE XIa - Austen Samples.

Frequency of 1st elements in sentence, in word-classes as a percentage of total words.

GR 1	2.44	0.00	0.88	0.68	1.52	0.00
GR 2	2.44	0.00	0.00	0.00	0.00	0.00
GR 3	0.00	2.63	1.75	0.68	0.00	0.00
GR 4	0.00	1.32	0.88	0.00	0.00	0.00
GR 5	0.00	0.00	0.00	0.68	0.00	0.00
GR 6	1.22	1.32	1.75	2.70	0.00	0.00
GR 7	0.00	0.00	0.00	0.00	0.00	3.70
GR 8	8.54	7.89	7.89	11.49	12.12	0.00
GR11	26.83	31.58	21.93	33.78	19.70	29.63
GR21	2.44	1.32	0.88	1.35	4.55	0.00
GR31	12.20	13.16	14.91	18.24	24.24	14.81
GR32	0.00	0.00	0.00	0.00	0.00	0.00
GR33	6.10	0.00	0.88	1.35	0.00	3.70
GR34	4.88	2.63	1.75	2.03	3.03	0.00
GR41	8.54	18.42	12.28	7.43	4.55	14.81
GR42	2.44	1.32	2.63	3.38	6.06	0.00
GR43	0.00	0.00	1.75	2.03	3.03	3.70
GR44	4.88	1.32	4.39	0.68	0.00	0.00
GR45	0.00	0.00	0.00	0.00	0.00	0.00
GR51	8.54	7.89	2.63	2.03	4.55	0.00
GR61	3.66	5.26	9.65	6.08	3.03	18.52
GR71	3.66	1.32	11.40	4.05	1.52	0.00
GR81	0.00	0.00	0.00	0.00	10.61	11.11
GR91	1.22	1.32	1.75	1.35	1.52	0.00
Total No. of sentences	82	76	114	148	66	27

TABLE XIb - Comparisons

GR 1	0.00	2.70	0.00	0.00	2.70	2.44	5.71	2.44
GR 2	0.00	0.00	0.00	0.00	0.00	0.81	0.00	2.44
GR 3	0.00	2.70	4.92	4.05	2.70	0.00	0.00	0.00
GR 4	0.00	0.00	1.64	0.00	0.00	0.00	0.00	0.00
GR 5	0.00	2.70	1.64	0.00	2.70	0.00	0.00	0.00
GR 6	0.00	0.00	1.64	4.05	5.41	5.69	0.00	2.44
GR 7	3.70	2.70	0.00	0.00	0.00	0.81	0.00	0.00
GR 8	0.00	0.00	3.28	0.00	10.81	10.57	11.43	8.54
GR11	29.63	27.03	29.51	31.08	18.92	25.20	8.57	26.83
GR21	0.00	0.00	0.00	0.00	0.00	1.63	2.86	2.44
GR31	14.81	21.62	4.92	17.57	10.81	17.07	28.57	12.20
GR32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GR33	3.70	0.00	0.00	0.00	2.70	1.63	2.86	6.10
GR34	0.00	5.41	0.00	1.35	2.70	3.25	11.43	4.88
GR41	14.81	16.22	26.23	13.51	16.22	4.88	2.86	8.54
GR42	0.00	5.41	3.28	5.41	5.41	5.69	8.57	2.44
GR43	3.70	0.00	1.64	0.00	2.70	0.81	0.00	0.00
GR44	0.00	0.00	3.28	5.41	0.00	1.63	0.00	4.88
GR45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GR51	0.00	0.00	1.64	4.05	8.11	10.57	11.43	7.32
GR61	18.52	2.70	0.00	4.05	2.70	0.81	0.00	3.66
GR71	0.00	2.70	8.20	5.41	0.00	0.81	0.00	3.66
GR81	11.11	0.00	0.00	0.00	2.70	1.63	0.00	0.00
GR91	0.00	8.11	8.20	4.05	2.70	4.07	5.71	1.22
Total No. of sentences	27	37	61	74	37	123	35	82

Table XII - Ratio of nominals to verbals.

Austen Samples	N	V	Ratio
1	45.78	24.19	1.89
2	45.31	23.47	1.93
3	37.12	27.26	1.36
4	43.12	24.51	1.76
5	39.67	27.26	1.46
6	45.86	21.63	2.12

Comparison Samples	N	V	Ratio
1	49.26	24.40	2.02
2	44.59	26.58	1.68
3	39.94	26.23	1.52
4	45.05	23.19	1.94
5	48.32	24.21	2.00
6	48.46	24.72	1.96
7	48.62	18.76	2.59
8	64.10	30.77	2.08

Table XIII - Boder's Adjective-Verb Quotient (AVQ) for various types of writing.

$$AVQ = 100 \times \frac{03}{02 + 05 + 06}$$

This gives a number, usually of 2 digits expressing the number of adjectives for each 100 verbs.

Source	Mean	AVQ Low	AVQ High	No. of samples
Plays	11	0	69	226
Legal	16	0	31	13
Business	19	0	33	7
Fiction	35	10	80	20
Poetry	36	0	75	18
H. Brisbane ("Today")	42	28	72	15
Letters	43	13	100	39
Emerson ("Journals")	47	11	138	132
Scientific	64	37	90	20
M.A. Essays	64	10	130	30
H.L. Mencken ("Mercury")	72	31	160	36
Advertisements	78	33	167	18
PH. D. Theses	88	50	200	30

From : Boder - The Adjective - Verb quotient - Psychological Record III, (March 1940) - Table I P.318.

Table XIV - The Adjective-Verb Quotient for Austen and comparisons.

Austen samples	03	VB	SUM	AVQ
1	5.95	8.38	14.33	70
2	7.31	8.13	15.44	89
3	5.20	8.90	14.11	58
4	5.80	8.35	14.16	69
5	6.50	9.50	16.01	68
6	6.34	5.40	11.74	117
MEAN	6.18	8.11	14.30	78

Comparison samples	03	VB	SUM	AVQ
1	4.65	8.37	13.02	55
2	4.80	10.61	15.42	45
3	6.54	11.74	18.29	55
4	5.82	8.89	14.71	65
5	4.75	8.80	13.56	53
6	4.86	7.75	12.60	62
7	4.55	9.55	14.11	47
8	4.64	7.09	11.72	65
MEAN	5.08	9.10	14.18	55

Table XV - The three most frequent word-patterns in Swift and the controls.

Samples-Austen	1	2	3
1	513101	015131	310301
2	513101	015131	310151
3	310151	513101	015131
4	513101	310151	310301
5	513101	310151	310301
6	310151	513101	310301
Samples-Comparison			
1	513101	310151	015131
2	513101	310151	310301
3	513101	310301	015131
4	310301	513101	310151
5	513101	310151	015131
6	513101	310151	015131
7	513101	310151	015131
8	513101	310151	015131

Table XVI - Frequency distribution of the most frequent three-word pattern (513101) as a percentage of total patterns.

Austen samples	Percentage	Comparison samples	Percentage
1	3.85	1	4.35
2	3.75	2	3.38
3	2.51	3	2.88
4	3.08	4	2.70
5	2.44	5	4.66
6	4.35	6	4.58
		7	5.68
		8	3.77
Mean	3.33	Mean	4.00
Range	1.84	Range	2.98

Table XVII - Frequency distribution of different patterns (D) patterns occurring once (U) and total patterns (P)

Austen Samples	D	U	P
1	1923	838	2761
2	1940	812	2752
3	1947	850	2797
4	1785	832	2617
5	1936	878	2814
6	1948	783	2731
Comparison Samples			
1	1813	722	2535
2	1822	640	2462
3	1805	815	2620
4	1904	851	2755
5	1711	860	2571
6	1848	854	2702
7	1798	845	2643
8	1976	916	2892

Table XVIIIA - The 10 most frequent patterns in all the samples combined, ranked in descending order, with English equivalents.

Rank	Pattern	English Equivalent.
1	51 31 01	of the man
2	31 01 51	the man of
3	01 51 31	man of the
4	31 03 01	the good man
5	31 01 41	the man and
6	01 51 01	man of strength
7	02 31 01	took the man
8	03 01 51	good man of
9	51 31 03	of the good
10	31 01 21	the man should

Table XVIII - Relative rank of the ten most frequent patterns.

Austen Samples	1	2	3	4	5	6	7	8	9	10
	513101	310151	015131	310301	310141	015101	023101	030151	513103	310121
1	1	4	2	3		8		5	6	9
2	1	3	2	4			8	6	10	5
3	2	1	3	4		10		6		5
4	1	2	4	3	5			9	8	
5	1	2	4	3		6	10			
6	2	1	4	3	5			6	9	
Comparison Samples										
1	1	2	3	4	5			10	9	8
2	1	2	4	3	5	7	10			8
3	1	4	3	2	8				9	
4	2	3	4	1	7	6	8		5	
5	1	2	3	4	6	5	10	7	9	
6	1	2	3	4	5	8	7		9	6
7	1	2	3	4	5	6	9	8		
8	1	2	3	4	10	5	6	7		8

TABLE XIX

Austen

1	2	3	4	5	6	Mean
2.61	2.65	1.57	2.29	1.95	1.76	2.14
1.63	1.34	1.64	1.57	1.49	1.90	1.60
1.70	1.74	1.00	1.18	1.00	1.39	1.34
1.70	1.27	0.86	1.18	1.03	1.46	1.25
0.00	0.00	0.00	0.61	0.00	1.17	0.30
0.91	0.73	0.61	0.57	0.00	0.88	0.62
0.65	0.00	0.50	0.00	0.57	0.00	0.29
0.69	0.51	0.00	0.57	0.00	0.70	0.41
0.00	0.58	0.00	0.00	0.53	0.00	0.19
0.51	0.65	0.54	0.57	0.53	0.66	0.58
0.00	0.00	0.54	0.00	0.53	0.00	0.18
0.00	0.00	0.50	0.61	1.00	0.00	0.35
0.69	0.00	0.00	0.00	0.53	0.00	0.20
0.00	0.55	0.00	0.00	0.00	0.88	0.24
0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.77	0.13
0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.57	0.00	0.00	0.10

Comparison

1	2	3	4	5	6	7	8	Mean
3.43	2.68	2.18	1.71	3.58	2.96	3.59	3.42	2.94
2.17	1.79	1.07	1.60	2.06	2.44	2.01	3.18	2.04
1.93	1.30	1.11	1.02	1.91	1.85	1.82	2.59	1.69
1.42	1.34	1.83	1.74	1.17	1.37	1.44	1.52	1.48
1.26	0.93	0.69	0.51	0.74	0.89	1.32	0.76	0.89
0.63	0.00	0.00	0.51	0.74	0.00	0.76	0.86	0.44
0.00	0.85	0.61	0.58	0.78	0.70	0.91	1.04	0.68
0.63	0.00	0.69	0.62	0.62	0.70	0.00	0.00	0.41
0.00	0.69	0.00	0.00	0.62	0.78	0.72	0.90	0.46
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.89	0.88	0.00	0.00	0.00	0.76	0.00	0.32
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.83	0.10
0.00	0.69	0.73	0.44	0.00	0.00	0.00	0.00	0.23
0.00	0.00	0.00	0.00	0.00	0.63	0.00	0.00	0.08
0.71	0.00	0.00	0.00	0.00	0.63	0.00	0.00	0.17
0.00	0.00	0.80	0.00	0.00	0.00	0.00	0.00	0.10
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09
0.00	0.00	0.65	0.00	0.00	0.00	0.00	0.00	0.08
0.00	0.00	0.00	0.00	0.00	0.00	0.57	0.00	0.07
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.05

TABLE XX - Frequency of last element in the sentence, in word-classes, as a percentage of total words.

Austen

GR 1	39.24	37.50	22.88	28.79	39.29	42.42
GR 2	6.33	11.11	10.17	13.64	2.38	6.06
GR 3	5.06	9.72	9.32	6.06	8.33	1.52
GR 4	5.06	1.39	2.54	4.55	5.95	3.03
GR 5	1.27	2.78	1.69	1.52	0.00	4.55
GR 6	0.00	0.00	0.00	0.00	0.00	0.00
GR 7	2.53	6.94	4.24	4.55	2.38	12.12
GR 8	3.80	9.72	10.17	4.55	5.95	3.03
GR11	13.92	2.78	17.80	18.18	23.81	7.58
GR 21	8.86	5.56	1.69	0.00	1.19	3.03
GR31	1.27	6.94	4.24	6.06	2.38	4.55
GR32	1.27	1.39	0.00	6.06	0.00	0.00
GR33	0.00	0.00	1.69	1.52	0.00	1.52
GR34	10.13	2.78	5.08	3.03	4.76	6.06
GR41	0.00	0.00	0.00	0.00	0.00	0.00
GR42	0.00	0.00	0.00	0.00	1.19	0.00
GR43	0.00	0.00	0.00	0.00	0.00	0.00
GR44	0.00	0.00	0.00	0.00	0.00	0.00
GR45	0.00	0.00	0.00	0.00	0.00	0.00
GR51	0.00	0.00	0.00	0.00	2.38	1.52
GR61	0.00	0.00	5.93	1.52	0.00	1.52
GR71	0.00	0.00	1.69	0.00	0.00	0.00
GR81	0.00	0.00	0.85	0.00	0.00	1.52
GR91	1.27	0.00	0.00	0.00	0.00	0.00
GR98	0.00	1.39	0.00	0.00	0.00	0.00

TABLE XX Cont'd

Comparisons

GR 1	37.04	41.67	40.35	24.00	51.35	47.92	54.43	79.41
GR 2	11.11	5.56	5.26	12.00	0.00	6.25	2.53	5.88
GR 3	0.00	2.78	3.51	8.00	2.70	0.00	8.86	2.94
GR 4	0.00	0.00	0.00	1.33	2.70	2.08	0.00	0.00
GR 5	0.00	0.00	0.00	2.67	5.41	8.33	1.27	0.00
GR 6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GR 7	7.41	2.78	5.26	6.67	2.70	12.50	7.59	0.00
GR 8	7.41	5.56	7.02	10.67	2.70	4.17	11.39	5.88
GR11	11.11	13.89	21.05	13.33	10.81	4.17	1.27	0.00
GR21	0.00	0.00	1.75	0.00	2.70	2.08	0.00	0.00
GR31	14.81	8.33	8.77	5.33	8.11	2.08	2.53	2.94
GR32	0.00	5.56	3.51	4.00	0.00	4.17	5.06	0.00
GR33	3.70	2.78	0.00	1.33	0.00	4.17	0.00	0.00
GR34	7.41	8.33	1.75	6.67	10.81	0.00	1.27	2.94
GR41	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GR42	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GR43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GR44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GR45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GR51	0.00	0.00	0.00	1.33	0.00	0.00	0.00	0.00
GR61	0.00	2.78	0.00	2.67	0.00	0.00	1.27	0.00
GR71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GR81	0.00	0.00	1.75	0.00	0.00	2.08	2.53	0.00
GR91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GR98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table XXI - Frequency distribution of pattern 0808, as a percentage of total patterns, and total occurrences.

Austen	1	2	3	4	5	6		
Samples	1.01	1.87	0.81	1.55	1.28	1.90		
	21	42	17	33	27	43		
Comparison	1	2	3	4	5	6	7	8
Samples	0.10	0.05	0.24	0.74	0.72	0.76	1.59	0.36
	2	1	5	15	15	16	34	7

Table XXII - Frequency distribution of classes 08 and 11, occurring as first word in sentence, as a percentage of total introductory words, giving mean.

Austen Samples	1	2	3	4	5	6			Mean
08	8.54	7.89	7.89	11.49	12.12	0.00			7.99
11	26.63	31.58	21.93	33.78	19.70	29.63			27.24
Total	35.37	39.47	29.82	45.27	31.82	29.63			35.23
Comparison Samples	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Mean**
08	0.00	0.00	3.28	0.00	10.81	10.57	11.43	8.54	5.58
11	29.63	27.03	29.51	31.08	18.92	25.20	8.57	26.83	24.53
Total	29.63	27.03	32.79	31.08	29.73	35.77	20.00	35.37	30.11

TABLE XXIII - Criteria of 35 test for 26 individuals, to be used as characteristics in CLUSTAN programs.

	1	2	3	4	5	6	7	8	9	10	11	12
1	15.85	14.91	11.06	13.76	12.71	16.00	14.76	12.96	15.82	17.18	10.47	15.07
2	6.30	5.87	6.40	5.90	7.30	2.93	6.26	7.70	6.01	5.66	6.81	4.27
3	5.95	7.31	5.20	5.80	6.50	6.34	5.46	5.35	5.06	6.61	4.61	6.73
4	1.63	1.88	1.80	2.25	2.35	1.80	2.00	1.30	1.40	1.30	0.95	1.08
5	2.03	1.97	2.15	2.45	1.80	2.27	2.85	1.70	2.40	2.25	2.40	2.11
6	0.05	0.29	0.35	0.00	0.40	0.19	0.25	0.25	0.00	0.00	0.05	0.05
7	3.81	4.09	4.35	4.85	3.65	6.34	4.20	2.40	4.60	5.31	4.46	4.27
8	3.30	5.92	4.20	5.15	4.35	5.06	5.31	5.50	6.91	5.51	9.37	5.01
11	8.99	7.50	10.36	8.60	11.91	6.25	8.76	11.36	8.01	6.71	10.47	8.59
21	6.91	6.49	8.70	7.00	8.05	7.43	6.01	8.70	5.86	4.56	8.22	7.36
31	13.21	13.13	12.21	13.71	11.11	13.87	12.56	11.31	12.96	13.98	9.77	12.13
32	0.91	0.63	0.65	1.05	0.75	0.90	0.40	0.20	0.95	0.25	0.40	0.49
33	2.18	3.03	3.05	2.90	2.95	2.84	2.80	2.65	1.85	2.10	3.21	2.26
34	3.76	1.97	3.35	1.95	3.30	2.41	2.05	2.80	3.50	3.21	5.26	4.96
41	3.71	4.52	3.45	4.75	3.60	5.30	3.60	3.80	5.01	5.56	3.26	5.20
42	1.07	1.39	1.05	1.20	1.20	1.14	1.35	1.15	1.05	1.00	1.25	0.64
43	1.42	2.21	2.25	2.25	1.35	2.18	3.10	2.80	2.10	2.51	1.80	1.72
44	0.66	0.29	0.60	0.05	0.35	0.09	0.00	0.45	0.15	0.05	0.70	0.64
45	0.81	0.29	0.45	0.55	0.85	0.71	0.65	0.60	0.30	0.60	0.80	0.54
51	10.77	9.96	8.65	9.85	9.35	9.65	11.71	7.95	11.61	12.27	8.02	10.11
61	0.76	1.15	1.90	1.45	0.85	0.38	1.20	0.85	0.60	0.45	1.00	1.23
71	0.20	0.10	0.85	0.00	0.40	0.09	0.15	0.25	0.05	0.20	0.20	0.15
81	0.97	0.82	0.65	0.75	0.50	2.32	0.60	1.45	0.40	0.25	0.25	0.79
91	0.56	0.58	0.35	0.35	0.15	0.38	0.45	0.60	0.40	0.10	0.65	0.10
98	4.17	3.70	5.95	3.40	4.25	3.12	3.50	5.90	3.00	2.35	5.61	4.52

continued overleaf

NVRAT	1.89	1.93	1.36	1.76	1.46	2.12	1.86	1.38	2.00	2.43	1.18	1.93
AVQ	70.00	89.00	58.00	69.00	68.00	117.00	58.00	55.00	60.00	83.00	49.00	104.00
P/S	38.92	42.23	35.52	40.17	39.07	40.94	41.09	37.17	42.19	43.84	39.13	38.59
FW	56.91	54.06	58.53	56.43	56.68	55.94	55.41	56.93	54.80	53.81	55.26	56.90
CONN	7.67	8.71	7.80	8.80	7.35	9.42	8.71	8.80	8.61	9.72	7.82	8.74
MOD	20.07	18.76	19.26	19.61	18.11	20.02	17.82	16.96	19.27	19.54	18.64	19.83
AUX	13.21	12.36	15.11	12.91	15.36	10.36	12.26	16.41	11.86	10.22	15.03	11.63
VERB	5.89	6.35	6.85	7.30	5.85	8.80	7.31	4.35	7.01	7.57	6.91	6.43
0808	1.01	1.87	0.81	1.55	1.28	1.90	1.73	1.82	2.29	2.04	2.50	1.52
513101	69.11	67.34	42.02	56.03	54.03	44.49	70.07	43.02	63.06	66.13	30.06	51.06

TABLE XXIII cont'd

	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	17.00	15.02	12.56	15.17	16.86	16.43	18.16	21.04	14.96	18.68	15.23	17.06	16.40	20.25
2	5.97	8.31	8.61	6.08	6.40	5.59	6.30	5.59	9.11	4.61	7.86	6.55	7.22	4.91
3	4.65	4.80	6.54	5.82	4.75	4.86	4.55	4.64	4.20	3.91	6.34	6.80	5.94	8.40
4	0.82	1.35	1.01	0.66	1.70	1.91	0.55	0.85	0.95	0.40	0.83	1.00	1.57	1.18
5	2.40	2.20	2.07	2.30	2.20	2.16	2.50	1.49	1.55	1.00	0.93	1.60	0.93	1.47
6	0.00	0.10	1.06	0.51	0.20	0.00	0.75	0.00	0.10	1.40	0.64	0.25	0.15	0.29
7	4.49	3.55	3.27	3.47	4.70	5.10	3.90	6.23	4.00	4.51	2.85	3.40	3.39	3.88
8	0.77	1.30	1.83	3.06	3.65	2.26	5.05	2.50	2.05	4.36	5.06	1.15	7.61	1.87
11	7.20	11.16	12.95	10.01	7.90	7.80	8.65	4.85	11.86	6.51	8.45	7.95	8.10	5.60
21	5.87	5.86	6.64	6.23	5.80	6.38	5.75	4.42	8.11	5.01	6.58	7.65	5.65	3.69
31	16.49	14.56	12.03	15.02	14.91	15.94	14.96	18.33	12.96	16.57	12.23	13.76	12.48	15.53
32	1.43	0.80	0.63	0.87	0.55	0.29	1.20	0.43	0.50	1.20	1.08	1.25	1.42	0.79
33	2.25	2.00	2.12	2.25	2.10	2.06	1.10	1.60	2.80	1.25	2.01	1.35	1.03	0.49
34	3.01	3.50	2.31	3.32	3.25	3.38	1.60	2.13	2.35	2.75	4.13	3.45	3.63	3.34
41	6.53	6.16	6.40	4.44	3.30	4.02	5.35	3.41	3.80	4.96	4.62	3.55	4.27	5.26
42	1.79	2.45	2.60	1.84	1.40	1.57	1.00	0.91	2.10	1.05	0.54	0.75	1.47	1.08
43	3.27	2.30	2.31	1.94	3.00	3.53	1.25	3.68	2.25	1.30	2.26	1.40	1.23	1.28
44	0.10	0.30	0.43	0.46	0.15	0.15	0.45	0.00	0.30	0.65	0.20	0.65	0.00	0.20
45	0.51	0.40	0.29	0.51	1.15	0.93	0.40	0.32	0.55	0.40	0.29	0.45	0.25	0.34
51	11.13	10.21	8.81	9.04	11.81	11.23	10.96	14.28	7.81	13.17	8.30	10.86	9.33	12.43
61	1.12	0.50	0.63	1.28	0.75	0.15	0.65	0.21	0.55	0.75	1.18	1.55	0.93	0.84
71	0.10	0.25	0.58	0.31	0.00	0.10	0.05	0.00	1.25	0.90	0.83	0.30	0.54	0.84
81	1.23	0.60	0.58	0.87	0.90	1.28	0.70	0.80	0.65	0.80	0.44	0.30	0.69	0.84
91	0.41	0.50	0.87	0.72	0.65	0.54	0.15	0.48	0.95	0.25	0.05	0.05	0.10	0.15
98	1.48	1.80	2.89	3.83	1.90	2.35	4.00	1.81	4.25	3.61	7.07	6.90	5.65	5.06

continued overleaf

NVRAT	2.02	1.68	1.52	1.94	2.00	1.96	2.59	2.08	1.42	2.98	1.66	2.07	1.90	3.38
AVQ	55.00	45.00	55.00	65.00	53.00	62.00	47.00	65.00	39.00	55.00	67.00	80.00	71.00	125.00
P/S	36.09	36.64	36.96	37.08	40.47	38.30	41.77	42.35	36.94	38.86	39.73	37.82	43.22	42.26
FW	62.43	61.56	60.15	59.09	57.63	59.34	54.23	55.83	58.81	57.54	53.19	55.28	51.13	52.68
CONN	12.20	11.61	12.03	9.19	9.00	10.20	8.45	8.31	9.01	8.36	7.91	6.80	7.22	8.16
MOD	23.18	20.87	17.08	21.45	20.81	21.18	18.86	22.48	18.62	21.78	19.45	19.81	18.57	20.15
AUX	11.84	14.16	15.26	12.31	12.21	11.97	12.06	10.02	17.22	9.61	14.44	14.21	12.87	8.60
VERB	6.89	5.86	6.40	6.28	7.10	7.26	7.15	7.73	5.66	6.91	4.42	5.25	4.47	5.65
0808	0.10	0.05	0.24	0.74	0.72	0.76	1.59	0.36	0.05	0.67	1.15	0.10	1.79	0.48
513101	83.72	64.06	52.94	48.01	88.04	75.53	91.05	107.62	59.06	105.16	59.92	67.03	80.55	87.47

TABLE C

5 K-LINKAGE LISTS - (NEAREST NEIGHBOURS)

S	1	0.637	16	0.665	5	0.699	12	0.776	17	0.808	24
S	2	0.456	4	0.576	7	0.672	9	0.868	1	0.982	10
S	3	0.807	5	1.098	16	1.114	1	1.125	4	1.127	8
S	4	0.425	7	0.456	2	0.640	9	0.874	1	0.943	5
S	5	0.665	1	0.807	3	0.943	4	0.954	8	1.080	2
S	6	1.264	4	1.392	2	1.492	18	1.565	10	1.630	12
S	7	0.425	4	0.576	2	0.668	9	0.745	17	0.959	10
S	8	0.954	5	1.127	3	1.241	21	1.249	1	1.348	11
S	9	0.546	10	0.640	4	0.668	7	0.672	2	0.820	19
S	10	0.546	9	0.959	7	0.982	2	1.167	4	1.224	17
S	11	1.207	3	1.268	5	1.348	8	1.502	1	1.540	12
S	12	0.699	1	1.011	9	1.028	2	1.054	16	1.120	24
S	13	0.903	14	1.185	18	1.215	16	1.472	17	1.742	4
S	14	0.837	16	0.881	15	0.903	13	1.240	18	1.278	21
S	15	0.881	14	1.218	21	1.266	16	1.819	8	2.012	5
S	16	0.637	1	0.837	14	1.037	4	1.054	12	1.083	2
S	17	0.319	18	0.745	7	0.776	1	1.024	4	1.095	9
S	18	0.319	17	1.060	1	1.098	7	1.185	13	1.237	4
S	19	0.820	9	1.154	22	1.288	1	1.356	10	1.373	2
S	20	1.342	18	1.344	17	1.535	10	1.879	9	2.106	22
S	21	1.218	15	1.241	8	1.278	14	1.344	16	1.391	3
S	22	1.154	19	1.881	26	1.914	16	2.070	1	2.106	20
S	23	0.745	25	0.890	24	1.296	5	1.394	1	1.427	12
S	24	0.808	1	0.890	23	1.120	12	1.243	16	1.346	5
S	25	0.745	23	1.144	9	1.299	2	1.419	24	1.464	1
S	26	1.642	10	1.862	24	1.878	25	1.881	22	2.001	19

TABLE D

5 K-LINKAGE LISTS - (NEAREST NEIGHBOURS)

S	1	0.446	3	0.567	6	0.652	8	0.959	9	1.009	11
S	2	0.789	4	1.064	15	1.094	7	1.105	3	1.175	10
S	3	0.414	6	0.446	1	0.622	8	0.923	4	1.005	16
S	4	0.789	2	0.923	3	0.933	7	1.058	1	1.102	6
S	5	1.222	3	1.354	1	1.449	17	1.515	9	1.598	11
S	6	0.414	3	0.567	1	0.656	8	0.729	16	0.932	9
S	7	0.933	4	1.094	2	1.203	20	1.316	10	1.424	15
S	8	0.532	9	0.622	3	0.652	1	0.656	6	0.803	18
S	9	0.532	8	0.932	6	0.959	1	1.130	3	1.198	16
S	10	1.175	2	1.237	4	1.316	7	1.492	11	1.703	8
S	11	0.991	8	1.009	1	1.024	15	1.087	23	1.134	3
S	12	0.877	13	1.158	17	1.192	15	1.449	16	1.694	3
S	13	0.815	15	0.856	14	0.877	12	1.217	17	1.243	20
S	14	0.856	13	1.192	20	1.231	15	1.774	7	1.978	4
S	15	0.815	13	1.016	3	1.024	11	1.051	1	1.064	2
S	16	0.311	17	0.729	6	1.005	3	1.085	8	1.127	15
S	17	0.311	16	1.066	6	1.158	12	1.206	3	1.217	13
S	18	0.803	8	1.118	21	1.327	9	1.332	1	1.336	15
S	19	1.312	17	1.324	16	1.493	9	1.823	8	2.073	12
S	20	1.192	14	1.203	7	1.243	13	1.300	15	1.352	2
S	21	1.118	18	1.829	25	1.853	15	2.078	19	2.079	23
S	22	0.726	24	0.874	23	1.268	4	1.391	11	1.437	7
S	23	0.874	22	1.087	11	1.209	15	1.310	4	1.339	2
S	24	0.726	22	1.110	8	1.264	1	1.393	23	1.499	18
S	25	1.593	9	1.809	23	1.814	24	1.829	21	1.936	18

This Table shows the results with sample 1 omitted. There are therefore 25 samples left, the original sample 2 becoming sample 1, 3 becoming 2 etc.

TABLE E

5 K-LINKAGE LISTS - (Nearest Neighbours)

S	1	0.556	4	0.581	2	0.596	17	0.649	9	0.652	7
S	2	0.351	4	0.434	7	0.581	1	0.651	12	0.689	9
S	3	0.755	5	0.901	4	1.111	7	1.121	8	1.142	1
S	4	0.351	2	0.374	7	0.556	1	0.632	9	0.652	12
S	5	0.733	1	0.755	3	0.812	4	0.888	7	0.935	2
S	6	1.310	12	1.351	2	1.504	4	1.544	18	1.687	1
S	7	0.374	4	0.434	2	0.536	9	0.609	17	0.652	1
S	8	0.993	5	1.049	11	1.121	3	1.122	7	1.128	1
S	9	0.442	10	0.536	7	0.632	4	0.649	1	0.689	2
S	10	0.442	9	0.960	2	0.985	12	1.060	7	1.108	4
S	11	1.049	8	1.199	3	1.223	5	1.331	7	1.395	4
S	12	0.651	2	0.652	4	0.687	1	0.835	9	0.985	10
S	13	0.907	14	0.986	18	1.272	17	1.280	16	1.955	20
S	14	0.907	13	1.029	16	1.045	15	1.212	18	1.280	17
S	15	1.045	14	1.472	21	1.509	16	2.114	8	2.354	5
S	16	0.654	1	0.847	17	1.029	14	1.042	2	1.050	4
S	17	0.318	18	0.596	1	0.609	7	0.771	9	0.805	4
S	18	0.318	17	0.986	13	1.013	1	1.080	20	1.172	7
S	19	0.739	9	0.966	22	1.004	25	1.096	1	1.121	10
S	20	0.995	17	1.080	18	1.643	10	1.941	9	1.955	13
S	21	1.375	16	1.454	14	1.472	15	1.517	8	1.552	3
S	22	0.966	19	1.776	16	1.959	17	2.094	9	2.098	26
S	23	0.770	25	0.873	24	1.142	12	1.202	3	1.208	1
S	24	0.873	23	0.886	1	1.024	12	1.230	25	1.411	16
S	25	0.770	23	0.837	9	0.926	1	1.004	19	1.027	2
S	26	1.736	10	1.820	24	2.074	12	2.086	19	2.098	22

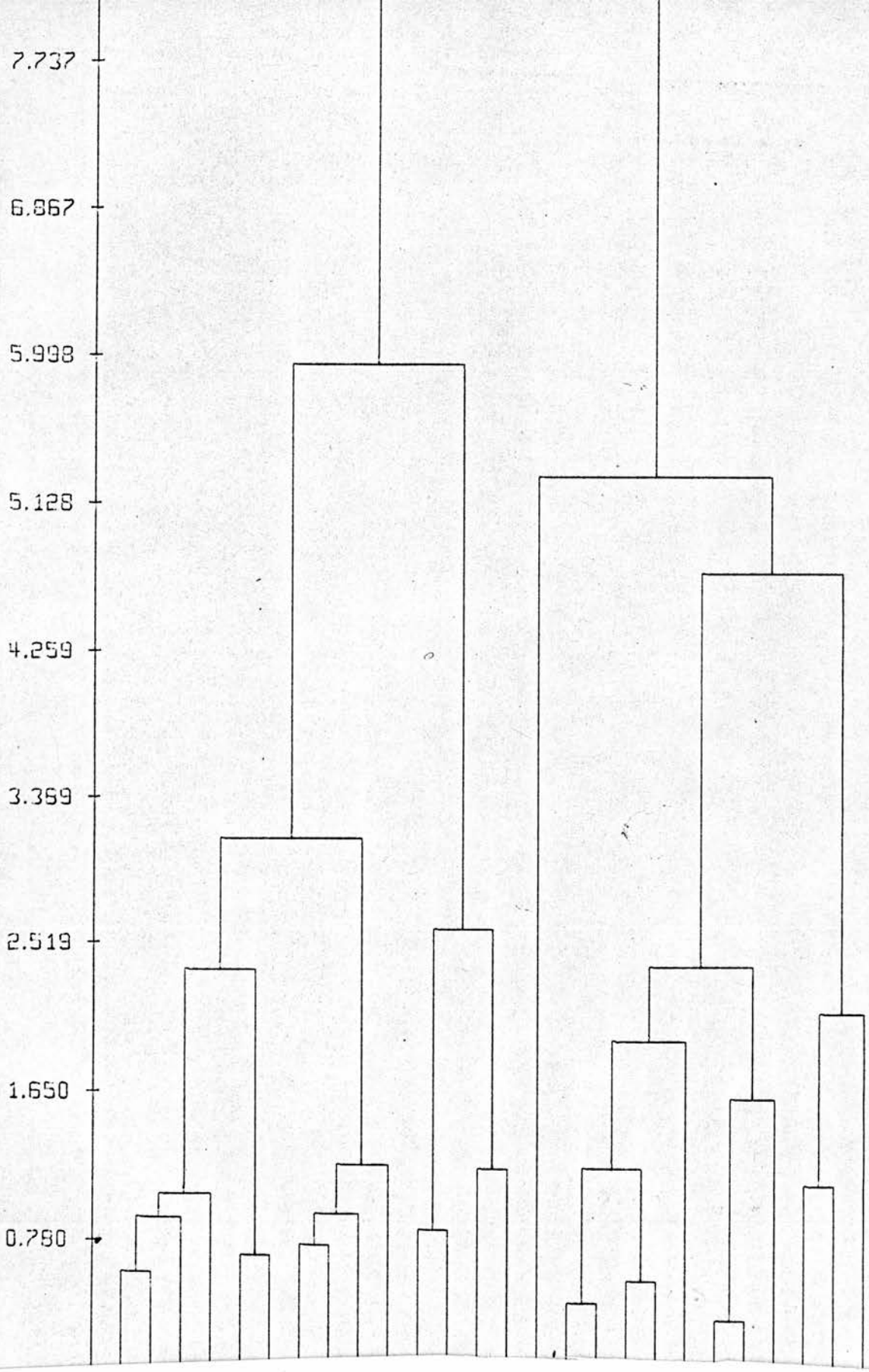
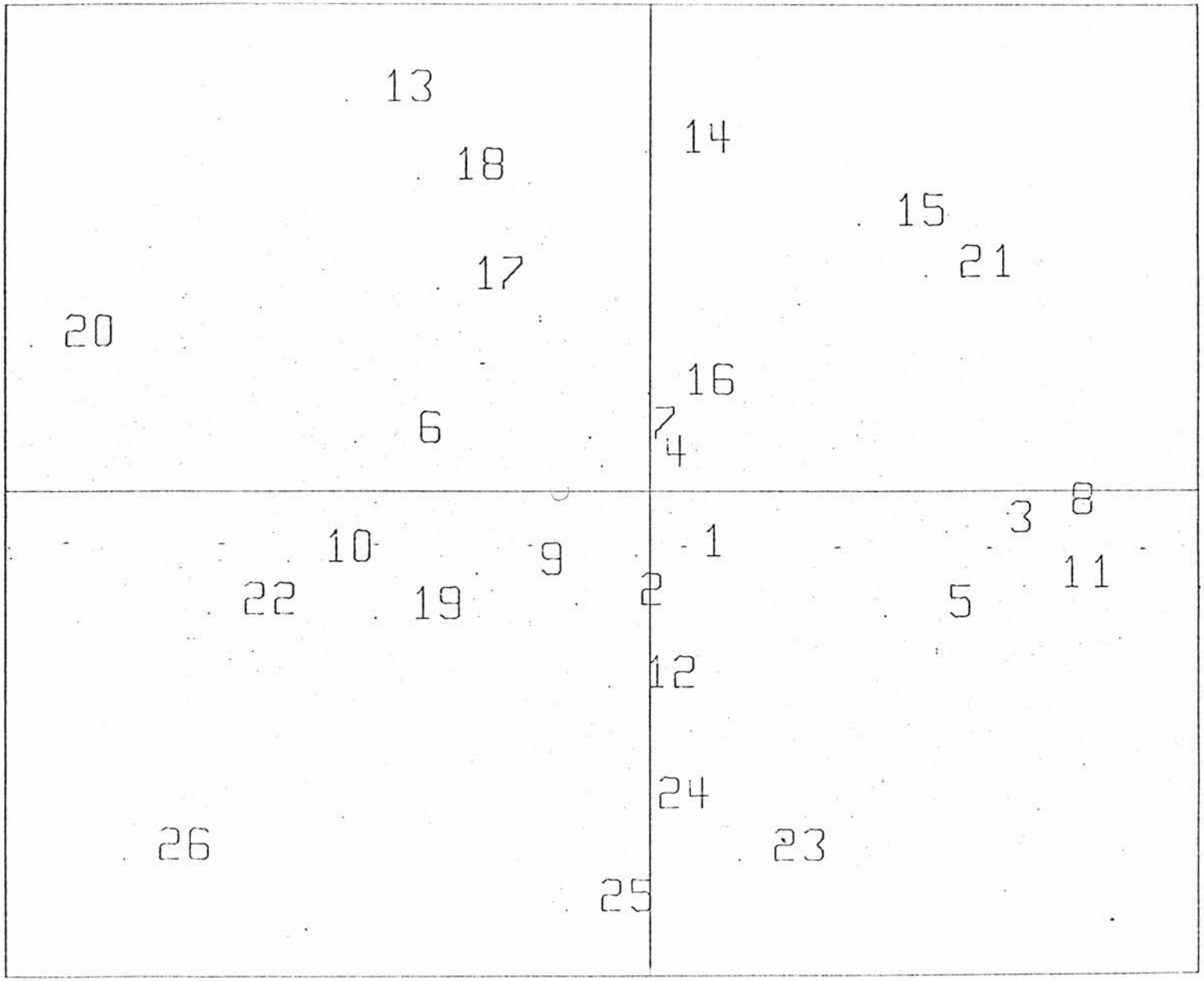
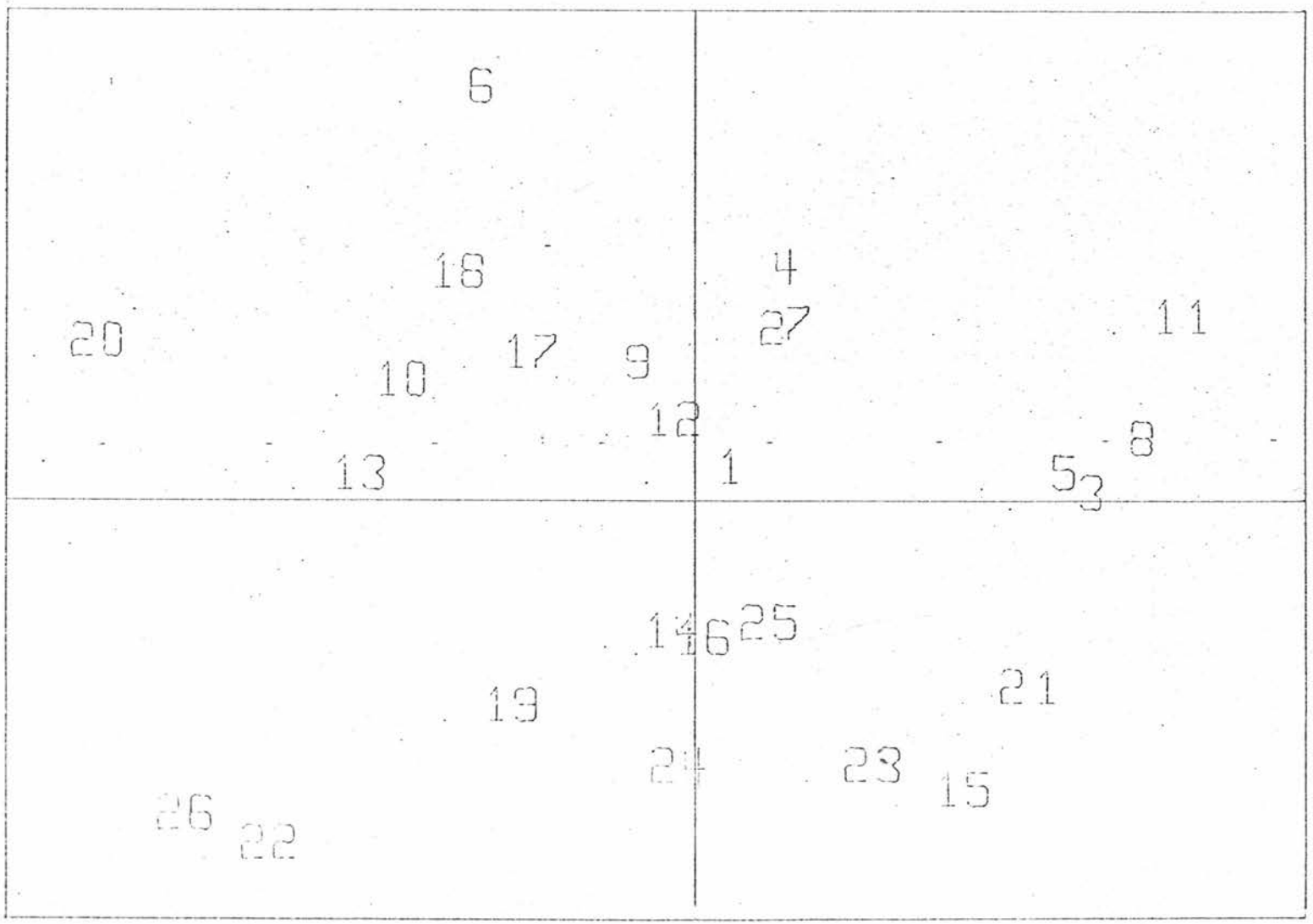


TABLE A

LINK PROGRAM TO GIVE DIAGRAM OF DENDOGRAM DECK

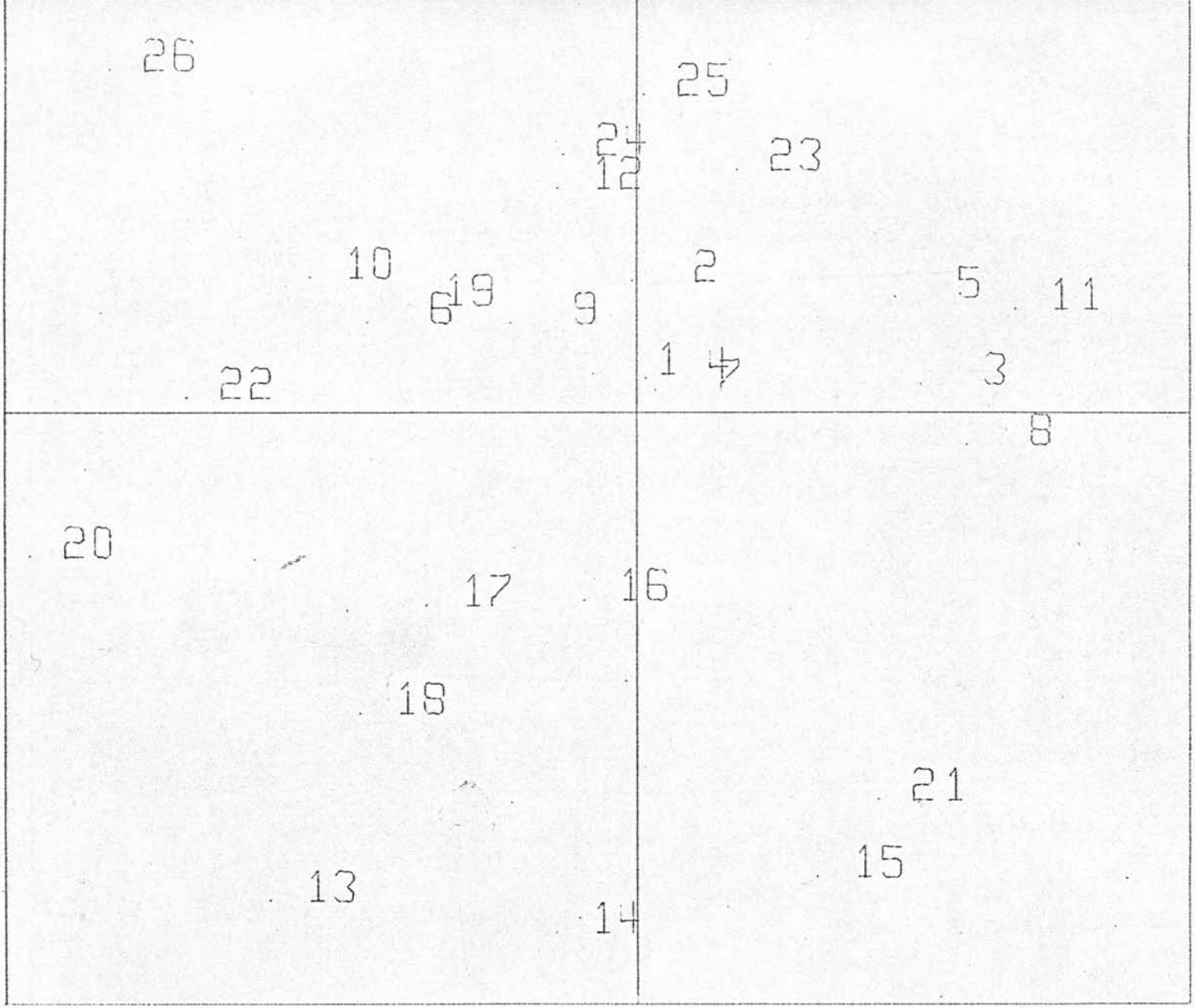

PRINCIPAL COMPONENT 1
PRINCIPAL COMPONENT 2
13
18
17
20
6
14
15
21
16
7
4
3
8
1
10
9
11
22
19
2
5
12
24
26
23
25

MAP NUMBER 1



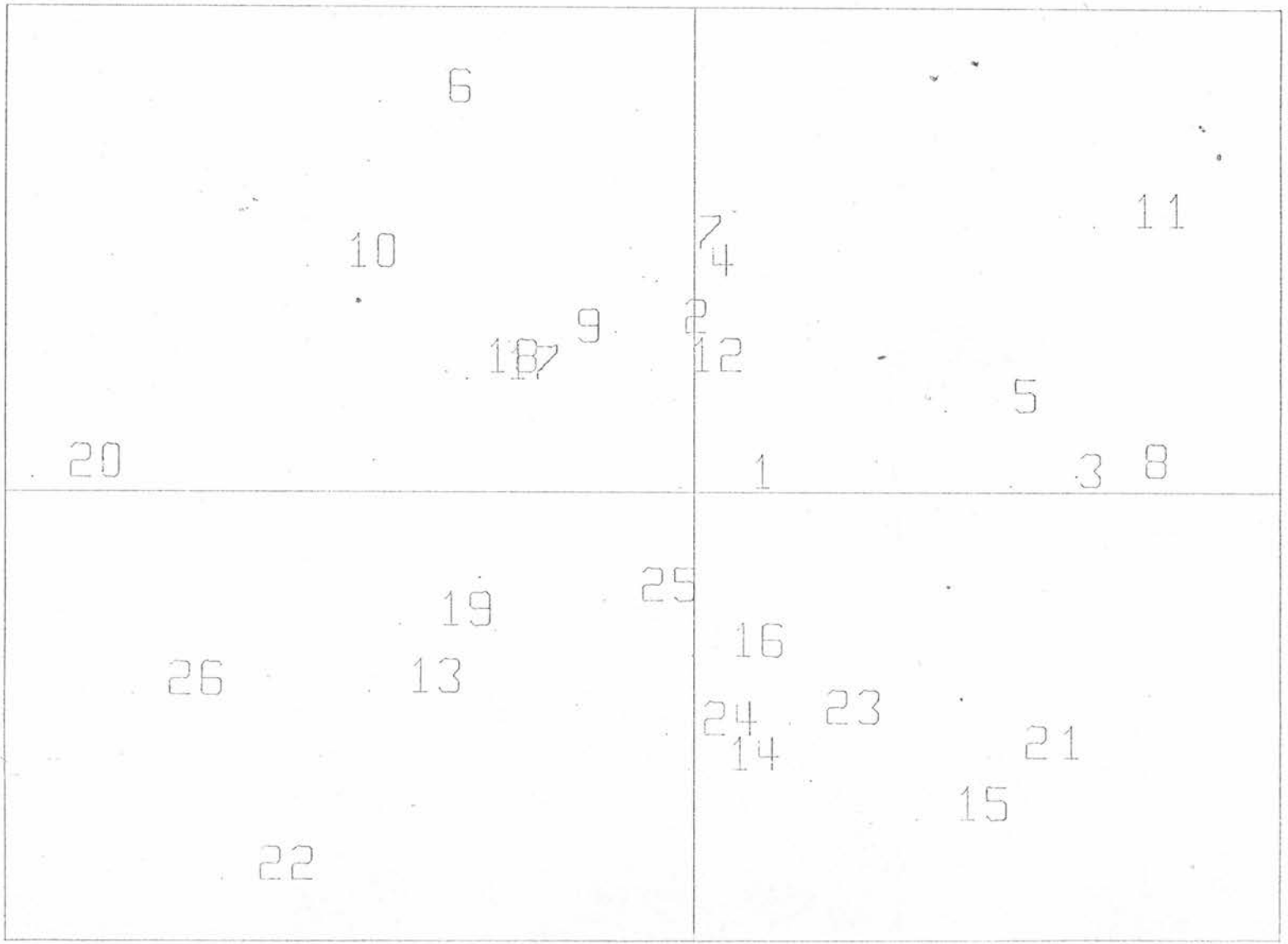
MAP NUMBER 2

TABLE 8.



MAP NUMBER 1

TABLE F



MAP NUMBER 2

TABLE F_2

Table XXIV - Summary table of identification criteria for all authors showing relative rank for each criterion.

Description of criteria	Austen		Defoe		Richardson		Fielding		Smollett	
	Rank	Value	Rank	Value	Rank	Value	Rank	Value	Rank	Value
F.W.	4	56.42	1	62.00	2	59.62	3	58.47	5	55.03
CONN.	5	8.29	1	11.91	2	10.61	3	9.60	4	8.38
MOD.	5	19.30	1	22.03	4	19.27	2	21.25	3	20.67
VA.	2	13.22	3	13.00	1	13.79	4	12.09	5	11.04
VB.	3	6.84	4	6.38	5	6.34	2	7.18	1	7.44
N.V. RAT.	4	1.75	1	2.35	5	1.73	3	1.98	2	2.34
AVQ	1	78	5	50	2	60	3	58	4	56
513101	4	55.50	3	73.89	5	50.48	2	81.76	1	99.34
0808	1	1.47	4	0.75	5	0.50	3	0.76	2	1.02
Introd. 08	1	35.19	4	28.33	3	31.94	2	32.75	5	27.69

TABLE XXV

Sample used in Cluster Analysis.

Sample	Work (short title)	Author	Date
1.	Northanger Abbey	Jane Austen	Written 1797 not published till 1818
2.	Sense and Sensibility	"	1811
3.	Pride and Prejudice 1.	"	1813
4.	Mansfield Park	"	1814
5.	Emma 1.	"	1816
6.	Persuasion 1.	"	1818
7.	Pride and Prejudice 2.	"	1813
8.	Pride and Prejudice 3.	"	1813
9.	Pride and Prejudice 4.	"	1813
10.	Persuasion 2.	"	1818
11.	Persuasion 3.	"	1818
12.	Emma 2.	"	1816
13.	Journal of the Plague Year	Defoe	1722
14.	Robinson Crusoe	"	1719
15.	Pamela	Richardson	1740
16.	Clarissa	"	1748
17.	Joseph Andrews	Fielding	1742
18.	Tom Jones 1.	"	1749
19.	Humphrey Clinker	Smollett	1771
20.	Peregrine Pickle	"	1751
21.	Tom Jones 2.	Fielding	1749
22.	Tristam Shandy	Sterne	1760
23.	Women in Love	Lawrence	1911
24.	Portrait of an Artist	Joyce	1916
25.	The Dead (The Dubliners)	"	1914
26.	Passage to India	Forster	1924