It Ain't my Style: a study in authorship exclusion

Richard S. Forsyth, [then at] University of Luton.

[Cite as:
Forsyth, R.S. (2002). It ain't my style: a study in authorship exclusion. *The 2002 Annual Meeting of the Classification Society of North America*, Madison Wisconsin, 13-15 June 2002.
http://www.classification-society.org/csna02/x31.html
]

E-MAIL: forsyth_rich@yahoo.co.uk
(former) ADDRESS: Department of Computing & Information Systems
University of Luton
Park Square
Luton LU2 3JU U.K.
(+44) (0)1582 - 743288

Abstract: Having fallen under suspicion of writing an anonymous satire, the present author embarked on a study of the feasibility of authorship exclusion, i.e. **disproving** authorship by stylometric methods. A sample of 42 texts written by myself was collected, as well as a multi-author, multi-register control sample of 52 comparison texts. The "Naïve" Bayesian Classifier, which has been reported by Mitchell (1997) and others to give good results in content-based text-categorization, was adapted to deal with the Self-versus-Other classification task, but gave poor results. However, a linear classifier based on frequent function words and punctuation marks achieved 79.8% cross-validated success rate in discriminating Self from Other, and assigned the disputed documents to Other. A step towards authorship exclusion on stylistic grounds has thus been taken, but further work is needed before it can be regarded as reliable. In addition, while no sign of linguistic drift was found over a period of half a century in the multi-author sample, the single-author sample showed clear evidence of chronological change, which, remarkably, exhibited the same trend in five different registers.

Keywords: Authorship attribution, Bayesian methods, machine learning, stylochronometry, text classification.

> "In short, every secret of a writer's soul, every experience of his life, every quality of his mind is written large in his works," -- Virginia Woolf, *Orlando*, 1928. (Woolf, 1993: 145.)

1. Background

Very few stylometrists have ever been suspects in a case of disputed authorship. Indeed, to the best of my knowledge, I joined a singleton set when, in October 2000, an anonymous "parable", highly critical of our departmental leadership, appeared in the

mail-room of my department, and I found that many of my colleagues believed that I had written it. I soon discovered, by informal extra-linguistic enquiries, who did write it (a member of staff who has subsequently left our institution). However, when a second anonymous article, with the same critical tone, was circulated a few weeks later, and various senior colleagues clearly still regarded me as the likely author, I decided that a fortuitous opportunity for a reflexive case study had arisen which was too good to pass up. The problem then became: whether it is feasible by stylometry – the statistical analysis of literary style (Holmes, 1994) – to show that a particular author has **not** written a particular text.

Most of the classic studies of disputed authorship have been framed in a positive context, e.g. Did Hamilton or Madison write the disputed *Federalist* papers (Mosteller & Wallace, 1984)? Did Fletcher or Shakespeare write *Two Noble Kinsmen* (Lowe & Matthews, 1995)? Did Cicero or Sigonio write the "rediscovered" *Consolatio* of 1583 (Forsyth et al., 1999)? Much less attention has been devoted to the obverse problem: how to show that someone did not write something, although arguably Foster's study of *The Funeral Elegy* (Foster, 1989) and Tweedie's investigation of whether Milton wrote (all of) *De Doctrina Christiana* (Tweedie et al., 1998) can be seen as examples of this kind of approach. The situation is analogous to providing a DNA sample with the aim of eliminating oneself from further enquiries in a criminal case. In practice, this need does sometimes arise in forensic linguistics -- in connection with poison-pen letters or emails.

2. Research Questions

The present study must be seen as work in progress. I do not claim to have found a knock-down method of authorial authentication; rather I am treating myself as a guinea pig (a practice with a long and honorable history in medical research) in the hope of making progress towards the answers to a number of interesting research questions.

As a starting-point, the Naïve Bayesian Classifier algorithm is used here, since this "improper" Bayesian method has been reported (Forsyth, 1995; Mitchell, 1997) to give surprisingly good results in text-categorization applications. Thus this study does not focus on evaluating rival algorithms or techniques. Instead it addresses three relatively narrow technical questions and two broader methodological questions.

Technical questions:

(1) Which of 2 alternative methods is better for deriving probability estimates from word frequencies in a group of documents?
(2) In this kind of application, is it better to use the full vocabulary of a group of documents or just the high-frequency words?
(3) Is it necessary to stick to words only, or does it help to use other tokens, such as punctuation, as well?

Methodological questions:

(1) Is authorship exclusion feasible using current stylometric methods?
(2) Is it essential to have a *foil* (i.e. a multi-author sample or corpus representing the language as a whole) to compare with the single author under suspicion in a case like this?

3. Text Sampling

Altogether 110 texts were collected for the present study and recorded electronically, under three different headings. This section summarizes the contents and some of the attributes of these three groupings, as they stood on 19th May 2002. (Further extensions and enhancements are anticipated.)

3.1 CATS – Control Authors' Text Sample

CATS (Control Authors' Text Sample) consists of 52 texts, one text per year from 1950 to 2001. It contains works by many different authors and is intended as a baseline. Four items are anonymous, two are co-authored, and one author appears twice. Because of the anonymous pieces, the exact number of authors involved is not known with certainty, but it cannot be less than 50, and is probably 54.

The main idea behind CATS is to establish a linguistic norm, as background against which to assess an individual author's idiolect. In effect it is a small-scale diachronic corpus, which reflects two major sources of linguistic variation: (1) temporal change (over my lifetime); and (2) variation among different registers.

"Deciding on the range of registers to include in a diachronic corpus can also be difficult." (Biber et al., 1998: 252.) For this investigation five registers were recognized, with a sixth text type (Misc) added to cover other varieties. Within each decade, a minimum and maximum number of each type had to be selected, and a program was written to do this randomly.

Table 1 – Text Types, per Decade

No.	Type Code	Description	Allowable Range
1.	Tech	Scientific or technical prose	2 .. 3
2.	Info	Non-technical non-fiction (e.g. arts & humanities)	2 .. 3
3.	Lett	Personal letters	1 .. 2
4.	Fict	Fiction (extracts from novels, or short stories)	1 .. 2
5.	Poem	Poems	1 .. 2
6.	Misc	Everything else (e.g. adverts, drama)	0 .. 1

Thus, in compiling CATS, it was necessary for each year to find a text of a particular type, first published in that year. The prime consideration was that this choice should not be dictated by personal preference. To avoid this I wandered around Luton University's Library and Luton Borough's Central Library on a number of occasions until I had

several (usually 6) candidates, then made the final selection randomly. Originally I had thought of using a subset of the British National Corpus (BNC) for this purpose, but unfortunately the BNC is not truly in the public domain. It is also rather limited in time-span. Another option considered was to make a random selection from the British Library catalogue. The British Library holds over 150 million items, going back more than 250 years, so it should be an ideal resource from which to draw. However, its electronic catalogues are incomplete before 1975 and are not designed with random selection in mind, so this approach was abandoned as impractical for the time being[1].

Further details of these texts can be found in Appendix 1.

3.2 RATS – Richard's Annualized Text Sample

There are some advantages to working with samples of one's own work, not least of which is that permission is easy to obtain! Another advantage is that both authorship and dating can be unusually secure. For this investigation a sample of 42 texts written by myself was selected, one text per year from 1960 to 2001, called RATS (Richard's Annualized Text Sample). This entailed many hours of searching through boxes of old family papers, school magazines, letters and so forth. It also exercised my seldom-used collection of non-6-sided dice, as well as Python's (pseudo-)random number generator. Then followed a good deal of scanning, typing and editing, since few of the texts were already on disk.

Even with my personal interest in the case, this was an arduous process, and many of the issues of corpus design presented themselves. My main concern was that the texts included should not be chosen merely on the basis of my personal taste. With this in mind, I always ensured that for every year (except 1960) there was always more than one candidate text to choose from, and in years where there was an abundance did not stop seeking texts till I had found at least six. The register constraints were the same as for CATS, as shown in Table 1, which was a great preventive against the temptation to include just personal favorites. Nevertheless, because in many years during the period covered I have not written fiction or verse, and because there were several years from which I could find no letters, it was necessary to shuffle these three categories to match availability in the 1960s, 1970s and 1980s. Another difference between CATS and RATS was that in RATS year of composition rather than first publication was used. (Many RATS texts have never been published.)

The distribution of text types in both CATS and RATS is shown in Table 2.

Table 2 – Distribution of Text Types.

	Tech	Info	Lett	Fict	Poem	Misc	N
CATS (1950-2001)	11	15	7	8	8	3	52
RATS (1960-2001)	12	10	4	5	6	5	42

Further details can be found in Appendix 2.

3.3 FLEA – Fear & Loathing Expressed Academically

A third sample called FLEA (Fear & Loathing Expressed Academically) comprises 16 texts. It includes both the anonymous pieces which I was suspected of writing and a number of internal Lutonian memoranda and messages, including an additional text by myself. Of these 16 texts 11 are anonymous. Some of these 11 are highly vituperative.

The file sizes in CATS, RATS and FLEA vary by more than two orders of magnitude. This ensures that any classification procedure that can deal with them must have faced, and coped with, the problem of unequal text lengths.

Table 3 shows some basic statistics concerning text length in the three samples.

Table 3 – Some Stats on CATS & RATS.

Word Counts:	Mean per text	Lower Quartile	Median	Upper Quartile	Total Words	Total Tokens
CATS	1650.33	281	725	1865	85817	99509
RATS	1366.17	174	851	1917	57379	68159
FLEA	993.25	386	932	1382	15892	17994

Further details of FLEA can be found in Appendix 3.

3.4 Limitations

Undoubtedly CATS has a Lutonian, and indeed Forsythian, bias and can therefore only be regarded as a first approximation. However, for the present purpose (comparison with my own style) this is not a major disadvantage: if a significant difference between self and non-self emerges here, it should be replicated, a fortiori, when using a more broadly-based sample.

4. Classification Procedures

Two programs, NBC1 and NBC2, were written, in Python, for this study. NBC2 is the more conventional of the two: it is essentially a re-implementation of the Naïve Bayesian Classifier described by Mitchell (1997). During training, it constructs a probabilistic model of word or token usage for each of two or more categories, from a number of training texts. When applied to new texts, it computes the posterior probability of each document's word usage $P(C=c_i \mid \text{data})$ for each category c_i in turn and assigns the text being considered to the category with the highest probability. For convenience, NBC2 uses entropies (negated log probabilities) instead of probabilities, and picks the smallest. It also starts with equal priors for each class. Otherwise it is the same as Mitchell's algorithm.

This procedure is called "naïve" or "improper" because it treats each feature as statistically independent.

NBC1 is an adaptation of Mitchell's algorithm to deal with only a single class. During training it forms a probabilistic model from the training texts of that class. During testing, it measures the fit of a number of test documents to that model in terms of mean entropy and *perplexity*. Perplexity is quite often used by corpus linguists (Oakes, 1998) to measure the degree of match between a linguistic model and a text. It is monotonically related to mean entropy according to the formula

$$PP = 2^{AE}$$

where PP is perplexity and AE is average entropy per symbol. It can be interpreted as the mean number of equally probable symbols at each choice point. The higher the perplexity (or entropy) the worse the fit between model and text. (In this paper, all entropies are quoted in bits, i.e. expressed to the base 2.)

Both NBC1 and NBC2 estimate P(t | C), the probability of token t given class C, as follows

$$P(t \mid C) = (f + 1) / (N + V)$$

where f is the frequency of token t in category C, N is the total number of tokens in category C, and V is the vocabulary size of category C – i.e. the number of distinct types. This implies that the "escape probability" (Witten & Bell, 1991) of a token never encountered before is 1 / (N + V). This adjustment fits into a Bayesian inferential framework, but its actual form has only pragmatic justification. It does have the considerable advantage that the problem of zero probabilities is avoided.

In addition, both programs have an alternative method of estimating the probability of a word or token, by taking the mean rates of usage of that token in each text, summing them, and dividing by the number of texts in the training sample. This gives equal weight to each text, whether long or short. This mode will be referred to as Moms (mean of means) to distinguish it from the standard mode (referred to as Mitch). Moms tends to downgrade the probability estimate of words that appear frequently in a long document but not often or at all in the other documents of the sample. For example, the name "Stephens" is ranked 48th most common word in CATS using total frequency (185) divided by total word-count (85817), with a probability of 0.002156, but in Moms mode it is ranked 161st with a probability of 0.000449. Even this is surely an over-estimate, but closer to its expected probability in a larger corpus. (It occurs in 2 out of 52 texts in CATS.)

It should be noted that when either NBC1 or NBC2 is run in self-test mode the leave-1-out method of cross-validation is used. Thus, for example, when using NBC1 to build a model on the 42 text units of RATS and also test it on RATS, the program actually

constructs a model 42 times, using 41 texts to construct the model and applying it to the remaining item each time.

5. Experimental Results

The results of four experiments, using NBC1, NBC2 and various other items of software, are described in this section.

5.1 Experiment 1 – Matching against a Single Model

Ideally, we would like to be able to construct a linguistic model from a sample of texts by a single author and find that it gives a significantly better degree of match to other texts written by that author than to texts by other authors. If we could state, without reference to other authors, that the degree of match between a particular text T and the model derived from author A was within or outside a 95% or 99% confidence level, so much the better. That would fulfill the requirements of an exclusion test.

The first experiment used NBC1 to explore this possibility, using word-based or token-based models. Probabilistic models were constructed on the RATS sample, then tested on both CATS and RATS. Six different conditions were compared, varying two factors – vocabulary source and estimation mode. There were three vocabulary sources:

Words	All letter-initial tokens in the training sample;
Tokens	All tokens (including numbers, punctuation etc) in the training sample;
BN50	Only the commonest 50 words in the British National Corpus.

The listing of the fifty most frequent words in the British National Corpus (written part only) was taken from Stamatatos et al. (2000). The first two vocabulary conditions use the full vocabulary (from RATS), whereas the third uses only high-frequency items. This latter condition was included because many previous stylometric studies (e.g. Burrows, 1992; Craig, 1992) have found that it is primarily the high-frequency words that discriminate between authors.

The two probability estimation modes (Mitch & Moms) were as described above, in section 4.

Table 4 shows the results using the Mitchell method. Three output values have been recorded. The first is AE, the average entropy per symbol. This is an overall measure of fit to the model. High scores indicate worse fits. Because the *escape probability*, i.e. the probability of words or tokens not in the vocabulary, can have a dramatic effect, two other parameters were also recorded: SENT (seen entropy), the entropy derived only from symbols found in the vocabulary; and PUNK, the proportion of unknown words (not found in the vocabulary) for which the escape probability is used.

Table 4 – Comparisons with RATS model, Mitchell Method.

	SELF	OTHER
Words		
AE	10.60	10.63
SENT *	9.72	9.57
PUNK *	0.1413	0.1657
Tokens		
AE	10.24	10.23
SENT **	9.38	9.19
PUNK *	0.1257	0.1468
BN50		
AE	2.86	2.91
SENT	6.42	6.33
PUNK	0.6204	0.6058

Table 5 shows the figures derived using the Moms method. In both tables an asterisk indicates a significant difference ($p < 0.05$) arising from a 2-tailed t-test between the values on the SELF and OTHER texts (RATS versus CATS). A double asterisk indicates a highly significant difference ($p < 0.01$).

Table 5 – Comparisons with RATS model, Moms Method.

	SELF	OTHER
Words		
AE	10.61	10.58
SENT *	9.71	9.51
PUNK *	0.1413	0.1657
Tokens		
AE	10.24	10.17
SENT **	9.37	9.13
PUNK *	0.1257	0.1468
BN50		
AE	2.85	2.91
SENT	6.40	6.30
PUNK	0.6204	0.6058

The first thing to note is that both modes (Mitch and Moms) give rather similar results.

More importantly, in none of the 6 conditions is there a significant difference in average entropy per symbol (AE). In other words, the texts in CATS match the vocabulary model derived from RATS about as well, or badly, as those from RATS itself (employing cross-validation). This was unexpected, and is rather puzzling.

Less surprising is that fact that – in the full vocabulary modes -- the proportion of unfound items is higher in the CATS texts than the RATS texts. In other words, texts written by me contain a higher proportion of words/tokens found elsewhere in my writings than do documents written by other people.

In the full-vocabulary modes, there are also significant differences in SENT. The entropy derived from seen symbols is actually higher for texts by me matched against my own probability model than for texts by other authors. This finding is illustrated by a boxplot, Figure 1.

Unfound items carry a relatively high entropy (low probability) so the fact that average entropies do not differ significantly implies the finding that the entropies derived from seen symbols must be higher in texts by myself (matched against a model from my own writings) than in texts by others. Still, it is rather strange.

A possible explanation is that words towards the tail of my own frequency distribution will more often be found in other writings by myself than in writings by others. In works by other authors, the items that are found in my vocabulary model will tend to be high-frequency items that form the core of the language.

In any case, this experiment shows that a simple-minded approach to vocabulary modelling does not lead to directly to effective authorial identification. Figure 2 illustrates that, while there is a significant difference between texts by me and by others on two parameters (SENT and PUNK), the overlap is still too great to form the basis for reliable classification.

5.2 Experiment 2 – Comparison between Two Models

In experiment 1 texts from both CATS and RATS were compared against a single vocabulary model, derived from RATS. In experiment 2, texts from CATS and RATS are compared to two vocabulary models, one derived from CATS, the other from RATS (with cross-validation). This is a more conventional way of using the Naïve Bayesian Classifier – to compare the best-fitting of two or more models – except that one of the "authors" is a composite.

Mitchell (1997) describes a content-based text-classification task, in which best results were obtained by dropping both very high-frequency items (e.g. the most frequent 50) and very low-frequency items (e.g. those occurring less than 3 times) before forming the models (Joachims, 1996). Accordingly, NBC2 has two user-selectable parameters (droptop and minfreq) which allows either or both ends of the frequency distribution to be excluded.

In content-based discriminations, it is reasonable to suppose that high-frequency words, such as "the", "of", "and" and "to" contribute little, but in authorship, previous studies have suggested that they are crucial.

Table 6 shows the results of running NBC2 on CATS and RATS, in three different vocabulary-trimming modes, using words only. The figures are the numbers of correct classifications, out of 94.

Table 6 – Classification, Self versus Other (Naïve Bayesian Classifier, words only).

	Mitch	Moms
Full vocabulary (minfreq 4 droptop 0) 4202 items	54 57.45%	54 57.45%
High-frequency vocabulary (minfreq 288 droptop 0) 52 items	51 54.25%	53 56.38%
Mid-range vocabulary (minfreq 4 droptop 50) 4152 items	60 63.83%	50 53.19%

Table 7 gives the same information using all tokens, including punctuation and numbers.

Table 7 – Classification, Self versus Other (Naïve Bayesian Classifier, all tokens).

	Mitch	Moms
Full vocabulary (minfreq 4 droptop 0) 4202 items	55 58.51%	58 61.70%
High-frequency vocabulary (minfreq 288 droptop 0) 52 items	59 62.77%	57 60.64%
Mid-range vocabulary (minfreq 4 droptop 50) 4152 items	59 62.77%	48 51.06%

Interestingly enough, the best success rate was obtained in the condition that most closely matched Joachim's – using words only and dropping words that occurred 3 times or less as well as dropping the high-frequency vocabulary items.

Using the settings that gave the highest success rate, the program was also run on the FLEA texts, with the following results: 12 texts classified as from RATS (i.e. by me), and 4 as from CATS. The 12 assigned to me included the single text actually by me, but also four by other named authors. The four assigned to CATS (the Foil) were: fishtale.txt,

footpar1.txt, footpar2.txt and vote.txt. Although clearly biased in one direction (towards RATS) the system does not allocate either of the suspect texts to me. It would seem that they are relatively dissimilar to my own writings as far as vocabulary is concerned.

However, a success rate of 63.83% (on known instances, cross-validated) is hardly impressive in a 2-class discrimination problem; hence little weight can be placed on this result. Indeed, these results are thoroughly disappointing, compared to those reported by Joachim (1996) who achieved 89% in a 20-class problem.

It begins to look as though the success of the Naïve Bayesian Classifier in text-categorization tasks does not generalize from content-based problems with many thousands of training texts to authorship problems with only several dozen training texts; and since even 42 texts amounting to 57379 words is quite a luxury in authorship studies that must cast some doubt on its usefulness in this area.

5.3 Experiment 3 – Discrimination using Frequent Words (& Tokens)

For comparison, and also to avoid ending on an inconclusive note, a third experiment was conducted, using the approach pioneered by Burrows (Burrows, 1992; Craig, 1992; Binongo, 1994; Holmes & Forsyth, 1995; Forsyth et al., 1999) – i.e. using frequent function words.

A program was written to read each text and from it produce a vector of numbers, where each number is the occurrence rate of a vocabulary item. In this experiment the vector length was 50. Three vocabulary sources were compared: CATS, using words only; CATS, using all tokens including numbers and punctuation marks; the top 50 words in the British National Corpus.

Having transformed the texts into numeric feature-vectors, the SPSS package was used to construct a linear discriminant function in stepwise fashion from the 50 variables. Success rates in the CATS/RATS discrimination task were recorded using leave-1-out cross-validation. Results are summarized in Table 8.

Table 8 – Results of Linear Discriminant Analysis (Self versus Other).

Source of Vocabulary	Percentage Success Rate (Cross-validated)	Canonical Correlation
CATS, top 50 words	70.2%	0.611
CATS, top 50 tokens [*]	79.8%	0.660
British National Corpus, top 50 words	74.5%	0.557

The condition which gave the highest success rate [*] also had the highest canonical correlation. The tokens actually used can be seen in Appendix 4. The discriminant formula produced in this condition is shown below.

```
Discfunc = the * 0.372 + comma * 0.508 + and * 0.410 – it * 0.581
      – hyphen * 0.593 - by * 1.513 – but * 1.102 – 3.065
```

This gives positive values for texts from CATS and negative for texts from RATS (i.e. by me).

So now I know something about my own style: I have a relatively low rate of "the", "and", and commas compared to the norm, but a relatively high rate of hyphens[2], "it", "by" and "but".

Figure 3 shows a boxplot of the discfunc scores for both groups. The separation is quite clear: the lower quartile for CATS texts is above zero while the upper quartile for RATS texts is below zero. This separation is confirmed in Table 9, which shows the cross-validated classification results, using the above formula.

Table 9 – Classification Results, cross-validated.

	CATS	RATS
Discfunc > 0	42	9
Discfunc <= 0	10	33

The success rate of almost 80% (19 errors out of 94) is quite respectable. Moreover, if a one-way Analysis of Variance on this measure is performed, the computed F ratio is 70.843, with 1 & 92 degrees of freedom. This between-groups variation represents 43.5% of the total variation. It shows that the idea of attempting a self-versus-other categorization is not a lost cause. Once again the "Burrows approach" has proved its worth.

This function was then applied to the FLEA texts, with the following results. The text acknowledged as mine was assigned to me, just, with a score of –0.0003. Of the 11 anonymous texts, 3 were assigned to me. The two disputed cases, however, were not: the first football parable obtained a score of 0.88 and the second a score of 1.84. Of the four texts signed by other named authors, three were assigned to me – including two by our head of department, the person criticized in the football parables! Accepting that I wrote only 1 of these texts (the one appearing under my name), that makes 6 mistakes out of 16, giving a crude success rate of 62.5%.

Figure 4 displays these results graphically.

Have I proved my innocence? Not conclusively, though the stylometric evidence does point away from my authorship of the disputed documents: of the 16 FLEA texts, the football parables are two of the three least similar to my work, as measured by the discriminant function which best separates CATS from RATS.

5.4 Experiment 4 – Effects of Type & Time

Two further studies were also conducted, to gain some insight into the effects of variation between registers and variation over time. From the perspective of text categorization, these are "contaminating" factors, but from a linguistic point of view they deserve investigation in their own right.

5.4.1 Register Variation

A Principal Components Analysis was carried out on CATS, using the 50 most frequent tokens (not just words) as variables. This produced 17 components with eigenvalues greater than 1, of which the first 7 components were needed to account for over 50% of variance (53.97%). The first 2 components together account for 21.71% of the total variance.

How might we interpret these two dimensions? Figure 5 aids this interpretation process by showing the individual texts of CATS plotted in the space of the first two dimensions, with text type marked. A register signal comes across loud and clear.

To aid in interpreting this plot, Table 10 shows the loadings of the ten variables with the highest loadings on the first component, positive or negative.

Table 10 – Highest-Magnitude Loadings on Factor 1.

Variable	Loading on Factor 1
IS	-0.756
HE	0.706
WAS	0.680
HIS	0.580
HIM	0.558
ARE	-0.537
OF	-0.526
FULLSTOP	0.522
HAD	0.518
SAID	0.510

This dimension contrasts "is" and "are" (negative loadings) with "was" (positive). Past tense verb-forms "had" and "said" also load positively on this dimension, so it has a temporal association. In addition, high scores on this dimension are correlated with high frequencies of (masculine) third-person pronouns. The positive loading of FULLSTOP suggests that texts scoring highly on this dimension have shorter sentences.

The ten variables with the highest loadings on the second principle component are shown in Table 11.

Table 11 – Highest-Magnitude Loadings on Factor 2.

Variable	Loading on Factor 2
THIS	0.592
THEY	-0.590
ON	-0.571
BE	0.547
IF	0.528
A	-0.519
I	0.516
HAVE	0.510
YOU	0.475
THAT	0.441

Here we find a contrast between "I" and "you" on the one hand and "they" on the other. The unmarked form of the verb "to be" also has a high positive loading.

A loading plot, showing how all 50 individual variables contribute to the first two principal components, appears as Figure 6.

The first dimension is primarily a fact-fiction polarity, with fictional texts scoring highly and technical texts getting lower or negative scores. This accords with the variable loadings shown in Figure 5 and Table 10. High loadings on Factor 1 for "said", "had" and "was" are associated with (past-tense) narrative, as are the third-person pronoun forms. By contrast, negative loadings for "is" and "are" link with the "timeless present" of technical exposition. We could label this dimension "fictionality". It has much in common with Biber's Factor 2 (Biber, 1988), which he describes as "Narrative versus Non-Narrative Concerns".

The prime contrast exhibited on the second dimension is between letters (high-scoring) and poems (low-scoring). A single descriptive label for the second dimension is harder to find, but the fact that it tends to contrast letters (high) with poems (low), as well as the high loadings for "I" and "you", suggests that it is associated with the authors' stance towards their intended readers. Letters (at least the ones collected in CATS) are addressed to a specific individual; poems are addressed to an unspecified audience. It could be termed "individuality of addressee".

Interestingly enough, these two contrasts (fict versus tech and lett versus poem) leave informative prose (info) in the middle as a kind of neutral standard. This text-type is the commonest single register in the sample, with 15 exemplars.

5.4.2 Chronometric Analysis

A stepwise regression with year as dependent variable and all 50 features as independents was performed using the CATS texts. This selected a single variable, the word "this". This was the only variable with a significant correlation with date ($r = +0.286$, $p = 0.04$).

One would expect two correlations to be significant at this level by chance, so this is actually an indication that no clear linear trend is detectable with these 50 variables.

On the RATS texts, however, a stepwise regression of year against the top 50 tokens picked four variables, with an adjusted R-squared of 0.458. The regression formula for Y (predicted year) is:

```
Y = 1994.927 – 3.838 * and – 10.825 * all – 4.721 * was – 4.375 * he
```

This indicates a falling-off over time in my usage of these four function words.

Figure 7 displays this predicated value for all RATS texts against actual year of composition, with text type marked. It exhibits quite a strong linear trend. What is striking is that all text types appear to participate in the same trend.

Irritatingly, SPSS does not provide facilities for cross-validating a multiple regression (though it does for discriminant analysis), so no genuine out-of-sample testing has yet been done on this finding. However, this finding probably is worth following up: if reliable prediction of composition date, across several registers, is possible for one author, it raises the possibility of effective stylochronometry.

The most important single variable in the temporal regression equation was “and”. Figure 8 shows that the rate of “and” in my writings declines over time. The fact that “and” was also a variable selected in the discriminant function raises an important question (which cannot be settled here): whether authorship classification techniques need to acknowledge temporal variation. In section 5.3 we obtained reasonably good classification results ignoring the time dimension – as is common practice in stylometry – but to achieve greater accuracy it may be necessary to take account of chronological variation.

6. Discussion & Conclusions

6.1 Technical Questions

It is possible to suggest tentative answers to the questions posed in section 2.

1) The mean of means (Moms) method of probability estimation was not clearly superior to Mitchell’s method in this task, though it is a viable alternative.

2) With NBC2 it was better to use mid-frequency vocabulary items rather than high-frequency items only or the full vocabulary (though a linear classifier using high-frequency tokens proved better than NBC2 in any of its vocabulary modes).

3) There would seem to be a slight advantage in using tokens other than words, such as punctuation symbols, rather than words only.

4) Is some sort of multi-author *foil* necessary in studies of this type? Indispensable!

6.2 Issues

Did I write the two anonymous pieces that, indirectly, provoked this work? It hardly looks like it, but I would have to admit that the evidence presented here would not convince a skeptic.

Is authorship exclusion feasible? Not yet, at least not reliably, using the techniques tried here. On the other hand, experiment 3 showed that a linear classifier using frequent tokens can give respectable performance in a Self-versus-Other classification task. This shows that Self-versus-Other classification is not a forlorn hope, and establishes a baseline performance level.

6.3. Further Findings

The principal-components analysis of the multi-author, multi-register sample (CATS) using high-frequency tokens does pick up clear register signals. This is perhaps a supplement to the work of Biber (1988), who studied register variation in the same manner but using syntactic variables.

Although it was not the main focus of the investigation, the chronometric results are interesting. While no clear evidence of linguistic drift was found in the multi-author sample, there was strong evidence of a trend in the single-author sample. Martindale (1990) found almost the opposite – clear evidence of a long-term pattern in the language as a whole but little sign of trends in individual authors (though he used semantic features and covered a much longer time-span). Remarkably, all text types in the RATS sample appear to participate in the same trend (subsection 5.4.2). This has implications for single-author stylochronometry.

It also has implications for authorship attribution. There is no logical inconsistency between seeking consistent trends in an author's style (for dating purposes) and seeking unchanging stylistic habits (for authorial identification); but in the present case one variable (the word "and") appeared both in the function for discriminating self from others and in the regression formula for assigning dates to my texts. At the very least this suggests that stylometrists seeking linguistic "fingerprints" will have to find ways of taking temporal variation into account. If a fingerprint is going to change shape over time, we need to know how fast.

6.4 Remarks

Though CATS far from perfect, I am now firmly convinced of the value of a multi-author, multi-register, diachronic text sample. This particular selection has several shortcomings – most notably in being rather too Lutonian, indeed Forsythian, for general purposes – but it has already demonstrated beyond doubt the usefulness of a temporally organized multi-author, multi-register text sample. I regard it as a prototype, and hope to find time and resources to compile an improved successor in due course.

The relative failure of the Naïve Bayesian Classifier is somewhat puzzling. Joachims (1996) obtained 89% success rate with this method on a 20-class problem. Here the same method achieved less than 64% success rate on a 2-class problem. The task itself is not intractable, since a simple linear classifier achieved as cross-validated success rate of 79.8%. It could be that methods which work well in content-based classification don't work so well in authorship problems. It might also be due to the fact that Joachims used much larger training samples than were available here. In any case, this matter needs further investigation. Possibly another classification technique, such as the Markovian method used successfully by Khmelev and Tweedie (2001) should also be brought into the comparison process. It is intended that the framework put in place for this study will allow investigation of the efficacy of alternative techniques (such as Markov models) and of attributes other than words (such as syntactic tags) at a later date.

It is perhaps appropriate to conclude such an egocentric investigation with a few personal remarks on the issue of self-consciousness. Authors very seldom investigate their own style systematically. I now have more objective evidence concerning my own linguistic habits than most authors. But how should I take account of this self-knowledge? For example, my lack of fondness the word "and" was only a mild surprise. I feel "the" and "and" to be boring words; but whereas "the" is practically unavoidable, "and" has an alternative, the ampersand, which I tend to use whenever I can get away with it. I didn't realize however, that this relative aversion to the most common conjunction in my native tongue has grown over the years. Should I follow the trend and attempt to exclude that word altogether from my vocabulary (like Georges Perec who wrote a novel in French without using the letter "e", which was later translated into English by Gilbert Adair under the same constraint)? I am tempted to do so, in which case "andlessness" will become a hallmark – albeit a highly artificial one – of my own style. But would it be worth the trouble? And would it be reliable? Texts written with enough time for and-removal would possess the hallmark, but those written in a rush probably would not.

And (or should I say "furthermore"?) paragraphs like these, which discuss "and" (or should I say "the most common conjunction in English"?), would become painfully prolix.

Complications of this nature may seem to be mere froth on the surface of stylometry, but if (as we hope) authorial identification becomes more effective, and as the knowledge of this spreads, we can expect writers who are motivated to disguise their style to become more sophisticated. It is analogous to the situation in cryptography – as decryption techniques become more effective, code-makers seek better methods of encryption. So students of style will have to ponder such questions.

Acknowledgements

I thank David Holmes of The College of New Jersey for inviting me to contribute this paper, and for many interesting discussions of stylometric principles & techniques over the years. I also thank Ellie Maclaren, of Middlesex University, for helpful suggestions, including an alternative acronym for FLEA (Former Lutonians' Emotive Allegations). In

addition, I want to thank my home institution, University of Luton, for agreeing to fund (most of) the cost of travel to present this paper at CSNA-2002 in Madison, Wisconsin. I am also grateful for being able to draw on the resources of Luton University's library and Luton town Central Library, without which this research could not have happened.

Postscript

Although I am willing to swear on oath that I wrote none of the anonymous texts in the FLEA sample, I share most of their negative sentiments and endorse their critical attitude towards the senior management of Luton University. (I would like it to be noted that I do so under my own name.)

Notes

[1] I have experimented with various procedures for randomized selection from the BL catalogue, though none is entirely satisfactory and all are laborious. Nevertheless, I would hope to use such an approach seriously in the future, if time and resources permit.

[2] The tokenizer used in this study does distinguish hyphens, joining words, from dashes, separating phrases, so hyphen means either a hyphen or a minus sign. Adventitious hyphens, however, present in some original texts to preserve right-justification, were removed during textual preparation.

References

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press, Cambridge.

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus linguistics: investigating language structure and use*. Cambridge University Press, Cambridge.

Binongo, J.N.G. (1994). Joaquin's Joaquinesquerie, Joaquinesquerie's Joaquin: a statistical investigation of a Filipino writer's style. *Literary & Linguistic Computing*, 9(4), 267-279.

Burrows, J.F. (1992). Not unless you ask nicely: the interpretive nexus between analysis and information. *Literary & Linguistic Computing*, 7(2), 91-109.

Craig, D.H. (1992). Authorial styles and frequencies of very common words: Jonson, Shakespeare and the additions to "The Spanish Tragedy". *Style*, 26, 199-220.

Forsyth, R.S. (1995). *Stylistic Structures: a computational approach to text classification*. Unpublished doctoral thesis, University of Nottingham, UK.

Forsyth, R.S. (1999). Stylochronometry with substrings. *Literary & Linguistic Computing*, 14(4), 467-477.

Forsyth, R.S. & Holmes, D.I. (1996). Feature-finding for text classification. *Literary & Linguistic Computing*, 11(4), 163-174.

Forsyth, R.S., Holmes, D.I. & Tse, E.K. (1999). Cicero, Sigonio and Burrows: investigating the authenticity of the "Consolatio". *Literary & Linguistic Computing*, 14(3), 375-400.

Foster, D.W. (1989). *Elegy by WS: a study in attribution*. University of Delaware Press.

Holmes, D.I. (1994). Authorship attribution. *Computers & the Humanities*, 28, 1-20.

Holmes, D.I. & Forsyth, R.S. (1995). The "Federalist" revisited: new directions in authorship attribution. *Literary & Linguistic Computing*, 10(2), 111-124.

Joachims, T. (1996). A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. Tech. Report CMU-CS-96-118, Carnegie-Mellon University.

Khmelev, D.V. & Tweedie, F.J. (2001). Using Markov chains for identification of writers. *Literary & Linguistic Computing*, 16(3), 299-307.

Lowe, D. & Matthews, R. (1995). Shakespeare vs. Fletcher: a stylometric analysis by radial basis functions. *Computers & the Humanities*, 29, 449-461.

Martindale, C. (1990). *The clockwork muse*. Basic Books, N.Y.

Mitchell, T. (1997). *Machine learning*. McGraw-Hill, NY.

Mosteller, F. & Wallace, D.L. (1984). *Applied Bayesian & classical inference: the case of the Federalist papers*. Springer-Verlag, NY. [First edition, 1964.]

Oakes, M.P. (1998). *Statistics for corpus linguistics*. Edinburgh University Press.

Stamatatos, E., Fakotakis, N. & Kokkinakis, G. (2000). Text genre detection using common word frequencies. *Proc. 18th Internat. Conf. On Computational Linguistics*, COLING-2000.

Tweedie, F.J., Holmes, D.I. & Corns, T.N. (1998). The provenance of *De Doctrina Christiana*: a statistical investigation. *Literary & Linguistic Computing*, 13(2), 77-87.

Witten, I.H. & Bell, T.C. (1991). The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. On Inforrmation Theory*, 37(4), 1085-1094.

Woolf, A.V. (1993). *Orlando*. Penguin Books, Harmondsworth. [First edition, 1928.]

Appendix 1 – CATS (Control Authors' Text Sample)

```
c:\rich\cats  Fri May 24 10:34:16 2002
  1 1950mose.txt  2814 info 1950  I.G.R. Moses : In Re 38, 39 And 40 Windmill Street, St.
  2 1951fain.txt  8070 fict 1951  William Fain : Harmony
  3 1952steb.txt  1122 tech 1952  Susan Stebbing : A Modern Elementary Logic, Chapter Vi
  4 1953wats.txt  1077 tech 1953  James Watson & Franc : Molecular Structure Of Nucleic A
  5 1954anon.txt   668 misc 1954  Anon : Institute Notes
  6 1955bann.txt  1620 info 1955  Roger Bannister : First Four Minutes, European Games
  7 1956murd.txt  2136 fict 1956  Iris Murdoch : The Flight From The Enchanter, Chap. 18
  8 1957gunn.txt   303 poem 1957  Thom Gunn : On The Move
  9 1958lark.txt   497 lett 1958  Philip Larkin : Letter To Judy Egerton
 10 1959brad.txt  1865 info 1959  Ernle Bradford : Cruising Or Racing
 11 1960shap.txt   259 info 1960  Fern R. Shapley : Andrea Del Castagno : The Youthful Da
 12 1961john.txt    90 lett 1961  Linda Johnson : Letter To Elizabethan Magazine
 13 1962tuch.txt  9112 info 1962  Barbara Tuchman : The Guns Of August, Chapter 9
 14 1963garv.txt  9457 tech 1963  Paul L. Garvin : The Definitional Model Of Language
 15 1964trew.txt  1364 info 1964  John C. Trewin : The World Of The Theatre: Places Of Pl
 16 1965zimm.txt   665 poem 1965  Robert A. Zimmerman : Desolation Row
 17 1966boli.txt   122 lett 1966  Hector Bolitho : Thank-you Note
 18 1967lenn.txt   153 poem 1967  John Lennon : Lucy In The Sky With Diamonds
 19 1968fill.txt  1147 tech 1968  Charles Fillmore : Lexical Entries For Verbs, Part V
 20 1969able.txt   783 fict 1969  Paul Ableman : The Twilight Of The Vilp, Chapter 4
 21 1970anon.txt   273 info 1970  Anon : Negro Leader Killed In Gunfight
 22 1971suss.txt    59 poem 1971  Aaron Sussaman : Death Of A Nun
 23 1972hida.txt  2520 tech 1972  K. Hidaka : Oceanography In Japan
 24 1973lark.txt   585 poem 1973  Philip Larkin : Show Saturday
 25 1974edin.txt  1043 info 1974  Harry G. Edinger : De Officiis, Book Iii, [1]-[8]
 26 1975amis.txt   118 lett 1975  Kingsley Amis : Letter To John Amis
 27 1976jone.txt  3689 tech 1976  Margaret G. Jones : Arthropods From Fallow Land In A Wi
 28 1977osul.txt   281 misc 1977  Peter O'Sullevan : Red Rum's 3rd National, Aintree 1977
 29 1978farr.txt  1489 fict 1978  J.G. Farrell : The Singapore Grip, Chapter 24
 30 1979pym.txt   4412 fict 1979  Barbara Pym : Across A Crowded Room
 31 1980winn.txt   371 lett 1980  Robert Winnett : Letter From Canon Winnett
 32 1981crai.txt   705 info 1981  George Craig : English Instabilities
 33 1982fors.txt   104 poem 1982  Helen Forsyth : Samuel Johnson / Died 1784
 34 1983rose.txt   203 poem 1983  Wendy Rose : Loo-wit
 35 1984jude.txt   622 info 1984  Anon : Epistle Of Jude
 36 1985fox.txt    159 lett 1985  Levi Fox : Letter From Levi Fox
 37 1986cars.txt   371 info 1986  James P. Carse : Finite And Infinite Games, Chapter 41
 38 1987rive.txt   633 tech 1987  Ronald Rivest : Learning Decision Lists
 39 1988rose.txt  2824 tech 1988  J. Rose, J.J. Lowe & : A Radiocarbon Date On Plant Detr
 40 1989conn.txt  5140 fict 1989  Shane Connaughton : Ojus
 41 1990call.txt   358 poem 1990  Philip Callow : Convalescence
 42 1991fors.txt   221 lett 1991  Angus A. Forsyth : Clan Forsyth Society, Letter To R.F.
 43 1992benn.txt   988 info 1992  Tony Benn : The End Of An Era : Foreword
 44 1993hack.txt  2455 fict 1993  Malcolm Hacksley : Ancestral Voices, Chapter 31
 45 1994anon.txt  1214 misc 1994  Anon : Nafta Treaty: Extract From Part Ii (trade In Goo
 46 1995ashw.txt   355 info 1995  Sue Ashworth : Grilled Cypriot Cheese With Tomato & Red
 47 1996sayo.txt  7206 tech 1996  Khalid Sayood : Introduction To Data Compression, Ch. 6
 48 1997cunn.txt   377 info 1997  John Cunningham : Obituary : Lady Tryon
 49 1998stal.txt  1231 tech 1998  Richard Stallman : Free Software Foundation (fsf) & GNU
 50 1999cole.txt   745 tech 1999  Alice Coleman : Notes On The Beginning Of The Third Mil
 51 2000thom.txt   252 info 2000  Roger Thomas : Ceredigion -- Cardigan Bay
 52 2001ripl.txt  1490 fict 2001  Ann Ripley : Harvest Of Murder, Chapter 20

Tokens =  99507; words =  85817.
```

As far as I can ascertain, 17 texts from this sample have non-British (first) authors, and 12 are by female authors. Three of the 52 texts are translations, all from Indo-European languages, as it happens (Latin, Greek and Afrikaans). Translations have been dated according to the year of the English translation, not of the foreign-language original. The target size-range for all text units was 100 to 10000 words, so for documents over 10000 words in length, a segment was chosen (at random) to be close to 1000 words in length,

e.g. a chapter in a novel. Currently, two texts are less than 100 words in length. These may be replaced in later version of CATS.

This particular selection has several shortcomings – most notably in being rather too Lutonian, indeed Forsythian, for general purposes – but it has already demonstrated beyond doubt the usefulness of a temporally organized multi-author, multi-register text sample. I regard it as a prototype, and hope to find time and resources to compile an improved successor in due course.

Ideally that should be freely available to anyone, e.g. for downloading over the internet. The fact that most of these texts are still under copyright, as are the great majority of texts written after 1918, makes that a difficult ideal to live up to. Acquiring permissions is tricky and time-consuming, even for non-profit-making uses. Nevertheless it is hoped to do just that eventually, so that the successor to CATS can be a shared resource. (Anyone who has texts to donate, such as letters, please contact me!)

Appendix 2 – RATS (Richard's Annualized Text Sample)

```
c:\rich\rats  Fri May 24 10:35:32 2002
  1 1960this.rf     69 poem 1960  Richard Forsyth : Thistledown
  2 1961abby.rf    179 info 1961  Richard Forsyth : Abbey Sports, No. 4, Sept. 30th
  3 1962bike.rf    171 lett 1962  Richard Forsyth : Letter From Bryanston About Bike
  4 1963clox.rf    315 tech 1963  Richard Forsyth : Sundials And Clocks
  5 1964rail.rf    317 info 1964  Richard Forsyth : Design Of Things To Come
  6 1965john.rf   1111 fict 1965  Richard Forsyth : Short Story : This Is John
  7 1966pigg.rf   1122 misc 1966  Richard Forsyth : This Little Piggy, Scene 1
  8 1967hove.rf    120 poem 1967  Richard Forsyth : Tim Hovey Died Fucking
  9 1968brum.rf    661 info 1968  Richard Forsyth : Way Out -- Symbolism And Sex
 10 1969memo.rf   1918 tech 1969  Richard Forsyth : The Biochemical Bases For Memory
 11 1970quic.rf    127 info 1970  Richard Forsyth : Translations From The Quicksilver
 12 1971thre.rf    172 poem 1971  Richard Forsyth : Three Minus One
 13 1972sept.rf    175 lett 1972  Richard Forsyth : Letter From Iffley Road
 14 1973soci.rf   3423 tech 1973  Richard Forsyth : Introduction To Social Cybernetics (s
 15 1974ezra.rf   1197 tech 1974  Richard Forsyth : Conversation Model Mark 1.1
 16 1975biog.rf    474 misc 1975  Richard Forsyth : Curriculum Vitae (spoof)
 17 1976basi.rf   2458 tech 1976  Richard Forsyth : The Basic Idea (chapter 6)
 18 1977perf.rf   1268 info 1977  Richard Forsyth : The Perfect Run
 19 1978hex2.rf    927 fict 1978  Richard Forsyth : Ascii Through The Logic Gate
 20 1979sili.rf   1916 fict 1979  Richard Forsyth : Silicon Valley Of The Dolls
 21 1980beag.rf   4050 tech 1980  Richard Forsyth : Beagle -- A Darwinian Approach To Pat
 22 1981pasc.rf   3560 tech 1981  Richard Forsyth : Pascal At Work & Play, Chapter 2
 23 1982ecai.rf   1641 info 1982  Richard Forsyth : Report On 1982 European Conference On
 24 1983arch.rf   2851 tech 1983  Richard Forsyth : The Architecture Of Expert Systems
 25 1984para.rf   1448 info 1984  Richard Forsyth : Paralox (a Game Of Strategy)
 26 1985vaxi.rf   2204 fict 1985  Richard Forsyth : The Christmas Tree Forest, Chapter 2
 27 1986rain.rf    132 poem 1986  Richard Forsyth : Rain Will Come Soon
 28 1987vrs.rf     583 misc 1987  Richard Forsyth : Memorandum Of Agreement
 29 1988cult.rf    923 info 1988  Richard Forsyth : The Cult Of Information
 30 1989birk.rf    163 lett 1989  Richard Forsyth : Letter To Brian Birkhead
 31 1990icon.rf   2098 tech 1990  Richard Forsyth : The Icon Programming Language
 32 1991wism.rf   6001 info 1991  Richard Forsyth : Towards A Grounded Morality
 33 1992hack.rf    777 fict 1992  Richard Forsyth : Obituary & Appreciation Of Fr. Hacker
 34 1993anns.rf   1439 tech 1993  Richard Forsyth : Neural Networks : A Very Brief Introd
 35 1994chor.rf    779 info 1994  Richard Forsyth : Foreword To "chaos Theory In The Fina
 36 1995pens.rf    112 lett 1995  Richard Forsyth : Letter To Provident Mutual
 37 1996feat.rf   8135 tech 1996  Richard Forsyth : Feature-finding For Text Classificati
 38 1997magd.rf    298 poem 1997  Richard Forsyth : Magda The Magyar
 39 1998bocc.rf    120 poem 1998  Richard Forsyth : La Bocca Della Verita`
 40 1999bash.rf    601 misc 1999  Richard Forsyth : High Life In High Town
 41 2000uppr.rf    329 misc 2000  Richard Forsyth : Upperthorpe Reunion Circular
 42 2001read.rf   1015 tech 2001  Richard Forsyth : In Search Of Research
Tokens =  68159; words =  57379.
```

Appendix 3 – FLEA (Fulminations from Luton's Embittered Academics)

```
c:\rich\flea  Fri May 24 10:35:53 2002
   1 2000spom.rf    855 misc 2000  Richard Forsyth : Spom (staff Perception Of Management
   2 fakedai.txt    356 lett 2001  Anon : Hello Colleagues
   3 fishtale.txt  1009 fict 2001  Anon : A Fishy Tale
   4 footpar1.txt   785 fict 2000  Anon : A Football Parable
   5 footpar2.txt   818 fict 2000  Anon : Football Parable Ii
   6 history.txt   3080 info 2001  Anon : A Thoroughly Mismanaged Affair
   7 labs.am        206 lett 2000  Annette Marshall & B : Cis Lab Complaints
   8 midyear.mk    1524 misc 2002  Malcolm Keech : Mid-year Review (cis)
   9 news.mk        386 lett 2002  Malcolm Keech : News On Staff Recruitment
  10 resigned.txt  1040 lett 2001  Anon : Vc & Gang Have Resigned
  11 teamat1.txt   1059 misc 2001  Anon : Team Matters, Mid-april 2001
  12 teamat2.txt   1034 misc 2001  Anon : Team Matters, Mid-may 2001
  13 teamjan.txt   1382 info 2002  Stella Cottrell et a : Team Matters, January 2002
  14 vote.txt       525 info 2001  Anon : No Confidence Ballot
  15 windbag1.txt  1654 misc 2001  Anon : A Welsh Windbag Waffles
  16 windbag2.txt   179 lett 2002  Anon : Website & Windbag
Tokens =  17994; words =  15892.
```

Appendix 4 – High-Frequency Tokens from CATS Texts.

```
    c:\rich\cats  Sun May 26 10:57:48 2002
the               5955    1  5.9845    5.9845  5.5751    5.5751  2.6764  5.3918  9.3548
,                 5268    2  5.2941   11.2786  5.3241   10.8992  2.1771  5.1243 10.7438
.                 4614    3  4.6369   15.9155  4.4362   15.3354  0.4354  4.1379  7.4604
of                3050    4  3.0651   18.9806  2.8825   18.2178  0.0000  2.8453  5.8133
to                2167    5  2.1777   21.1583  2.1376   20.3554  0.8657  2.0478  4.6693
and               2153    6  2.1637   23.3220  2.2707   22.6261  0.6711  2.0882  5.3613
a                 2062    7  2.0722   25.3942  2.0438   24.6698  0.0000  1.9248  9.3750
in                1845    8  1.8541   27.2483  1.6819   26.3517  0.0000  1.4514  5.1429
that               955    9  0.9597   28.2081  0.9610   27.3127  0.0000  0.8264  4.2735
was                948   10  0.9527   29.1608  0.6169   27.9296  0.0000  0.4866  2.0756
is                 888   11  0.8924   30.0532  1.0309   28.9605  0.0000  1.1050  2.5039
he                 830   12  0.8341   30.8873  0.4614   29.4219  0.0000  0.1368  2.7886
it                 704   13  0.7075   31.5948  0.7043   30.1262  0.0000  0.6873  2.6846
for                681   14  0.6844   32.2791  0.7590   30.8853  0.0000  0.6645  2.4331
as                 678   15  0.6814   32.9605  0.5829   31.4681  0.0000  0.5537  1.5625
-                  618   16  0.6211   33.5816  0.6212   32.0893  0.0000  0.5032  3.0096
on                 589   17  0.5919   34.1735  0.6108   32.7001  0.0000  0.5525  1.7143
with               587   18  0.5899   34.7634  0.8480   33.5480  0.0000  0.6058  6.2857
his                555   19  0.5577   35.3211  0.3478   33.8959  0.0000  0.1036  2.6356
be                 547   20  0.5497   35.8708  0.4833   34.3791  0.0000  0.4662  1.9455
not                489   21  0.4914   36.3623  0.4571   34.8362  0.0000  0.3771  1.5625
i                  469   22  0.4713   36.8336  0.8403   35.6766  0.0000  0.3529  6.7114
by                 468   23  0.4703   37.3039  0.4278   36.1043  0.0000  0.3226  2.0344
had                463   24  0.4653   37.7692  0.3126   36.4169  0.0000  0.0095  3.0162
from               455   25  0.4573   38.2265  0.4780   36.8949  0.0000  0.4149  1.4457
this               446   26  0.4482   38.6747  0.4584   37.3533  0.0000  0.3747  1.4599
at                 427   27  0.4291   39.1038  0.4623   37.8156  0.0000  0.4411  1.3423
are                382   28  0.3839   39.4877  0.4713   38.2869  0.0000  0.2755  3.1250
you                377   29  0.3789   39.8665  0.5957   38.8826  0.0000  0.1618  3.8911
or                 376   30  0.3779   40.2444  0.3621   39.2447  0.0000  0.2331  3.0166
she                350   31  0.3517   40.5961  0.3450   39.5897  0.0000  0.0000  3.9130
which              339   32  0.3407   40.9368  0.2537   39.8434  0.0000  0.2331  0.8877
have               335   33  0.3367   41.2735  0.4717   40.3151  0.0000  0.3322  2.0202
but                329   34  0.3306   41.6041  0.3967   40.7118  0.0000  0.3217  1.7391
one                315   35  0.3166   41.9207  0.4174   41.1292  0.0000  0.3436  2.2039
her                311   36  0.3125   42.2332  0.3846   41.5138  0.0000  0.0000  6.0870
were               304   37  0.3055   42.5387  0.1710   41.6848  0.0000  0.0000  1.5905
him                297   38  0.2985   42.8372  0.1607   41.8456  0.0000  0.0000  1.2019
an                 296   39  0.2975   43.1347  0.2692   42.1148  0.0000  0.2353  0.8296
they               287   40  0.2884   43.4231  0.3724   42.4872  0.0000  0.2037  3.1250
--                 284   41  0.2854   43.7085  0.3332   42.8204  0.0000  0.2472  1.5544
;                  274   42  0.2754   43.9838  0.3502   43.1707  0.0000  0.1958  2.1888
we                 271   43  0.2723   44.2562  0.2581   43.4288  0.0000  0.0201  1.7094
said               257   44  0.2583   44.5145  0.1477   43.5765  0.0000  0.0000  1.4664
?                  255   45  0.2563   44.7707  0.2308   43.8073  0.0000  0.0000  2.7778
all                246   46  0.2472   45.0179  0.3087   44.1160  0.0000  0.2320  1.5625
would              241   47  0.2422   45.2601  0.1950   44.3111  0.0000  0.1236  1.1673
if                 224   48  0.2251   45.4852  0.2360   44.5470  0.0000  0.1222  1.9455
there              201   49  0.2020   45.6872  0.1978   44.7449  0.0000  0.1724  0.7126
when               200   50  0.2010   45.8882  0.1693   44.9142  0.0000  0.1236  0.8547
  Total symbols =  99507, total types = 11105, TT-Ratio =  0.1116
```

After each vocabulary item, the nine numbers listed are, in order:

Total frequency in the sample;
Rank in descending order of frequency;
Overall occurrence rate, percent;
Cumulative occurrence rate;
Mean rate of occurrence (mean of means);
Cumulative percentage occurrence rate (Moms);
Minimum occurrence rate; Median rate of occurrence; Maximum rate of occurrence.

The last 7 columns are all expressed as percentages.

Appendix 5 – Specimen Text Unit

There follows the contents of the smallest file in the CATS sample, to show the format used.

```
<head>
   DEATH OF A NUN:
         Aaron Sussaman;
         Pullman, Washington?, 1971.
         Published in Dryad, No. 7/8, 1971.
</head>

<body>
         DEATH OF A NUN

She is not one
anymore,
but three,
like three sisters,
and they are all running
down a steep backyard
into a death
clean as the mouth
of a cat.
The father they are afraid
to look back and see
kneels on a high porch
to leave a saucer of warm milk
for the approaching storm.
</body>

<tail>
by=Aaron Sussaman
year=1971
from=magazine
refline=Sussaman, A. (1971). Death of a nun. Dryad, No. 7/8, Washington D.C. p. 10.
brit=n
poem=y
textype=poem
</tail>
```

All texts are held in plain ASCII form, divided into three sections by minimal mark-up:

<head> </head>
<body> </body>
<tail> </tail>

The <head> contains title, author and publication details in a consistent format, for human consumption. The <body> is the only part that is actually analyzed as text. The <tail> contains attribute data which is meant to be interpreted by software.

Figures & Diagrams

SENT

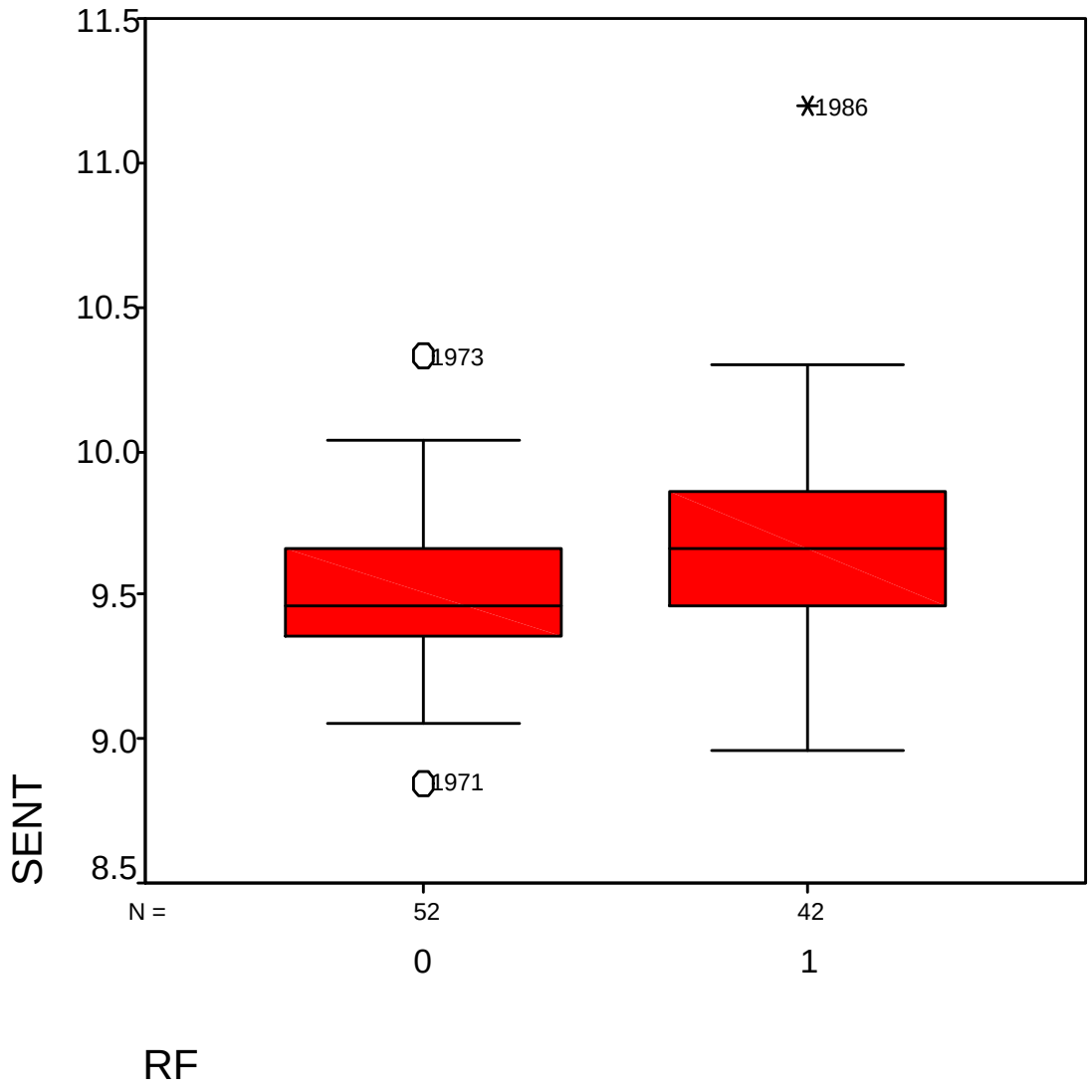

Figure 1 – Mean Entropy of Seen Symbols (SENT) using RATS model on CATS (RF=0) and RATS (RF=1); Probability Estimation Mode = Moms, Vocabulary source = Words.

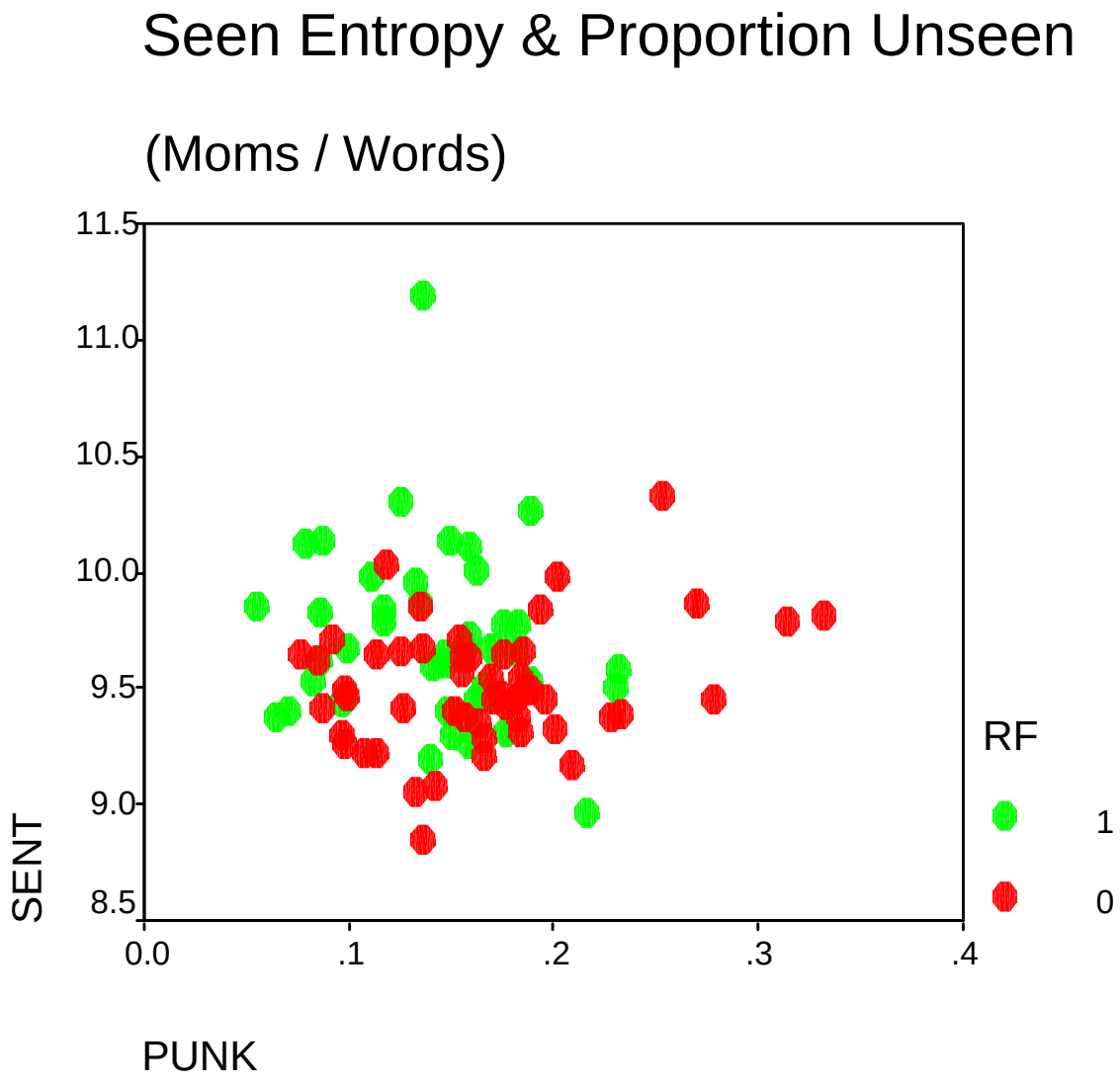

Figure 2 – Plot of Seen Entropy & Proportion Unfound (Self versus Other), Estimate Mode = Moms, Vocabulary Source = Words.

Discriminant Scores from Function 1 for Analysis 1

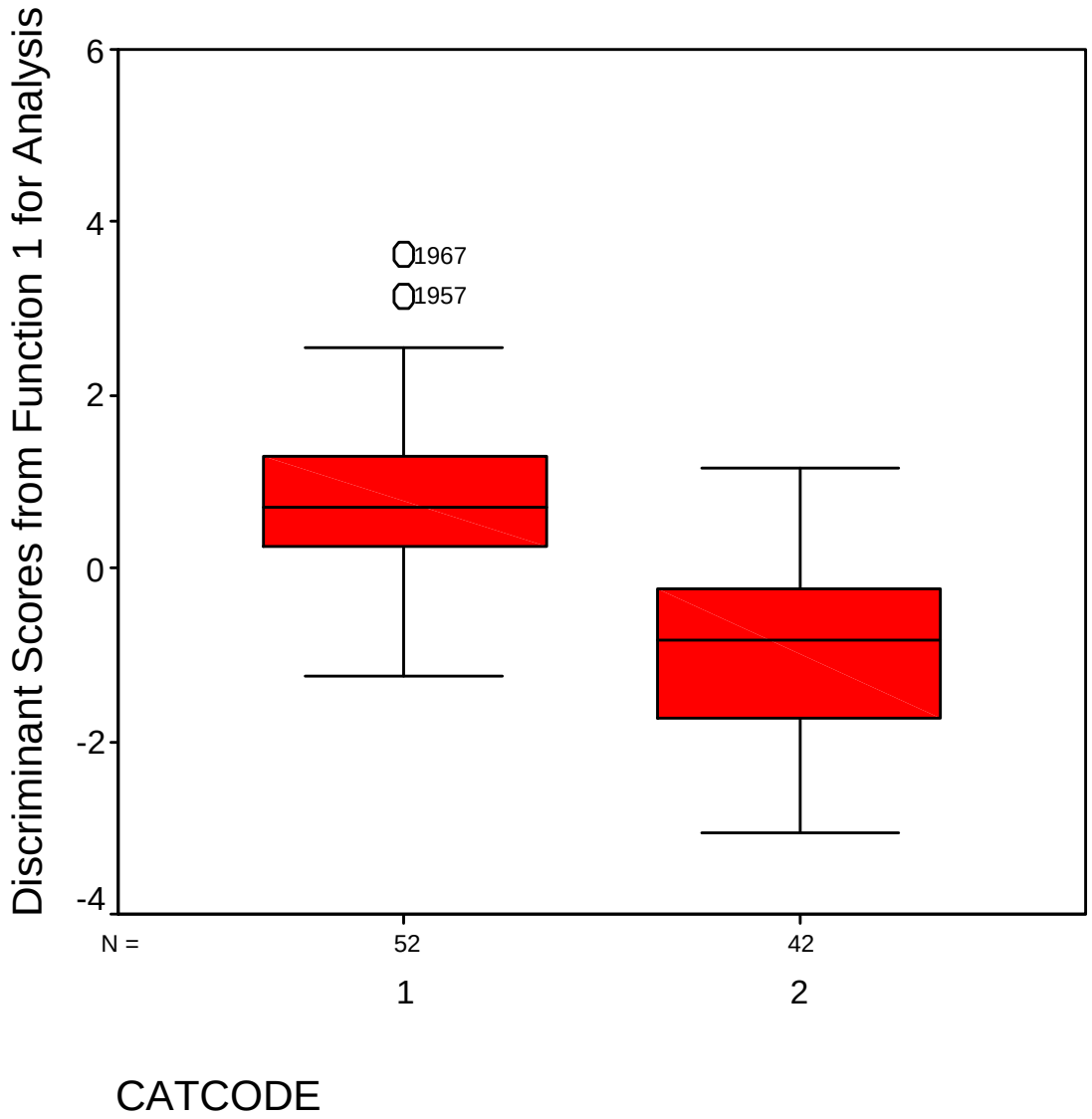

Figure 3 – Scores on best linear discriminant function for CATS (catcode=1) and RATS (catcode=2) texts. The stepwise method selected 7 variables from most frequent 50 tokens in CATS, as shown in the formula below.

```
Discfunc = the*0.372 + comma*0.508 + and*0.410 – it*0.581
     – hyphen*0.593 - by*1.513 – but*1.102 – 3.065
```

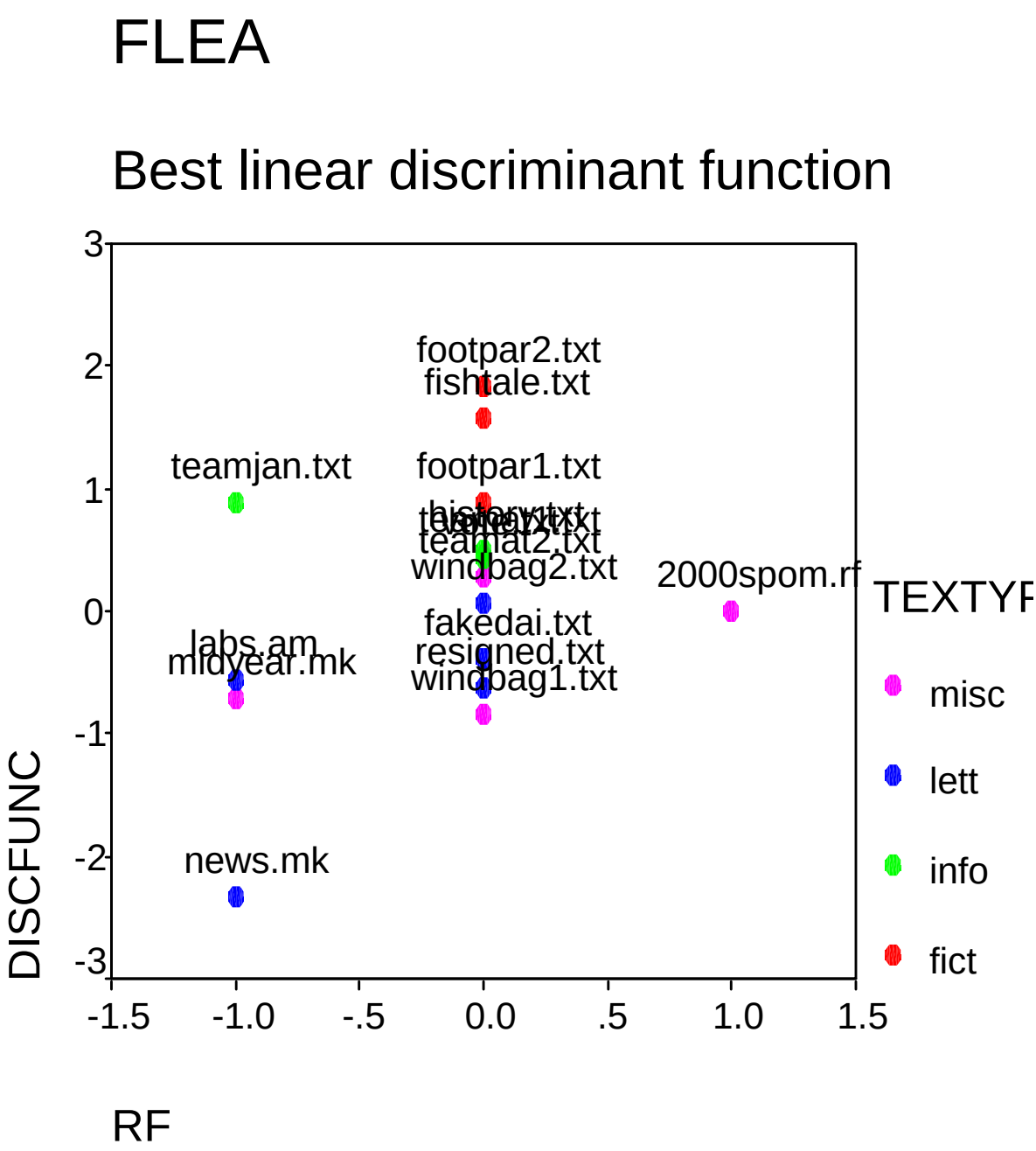

Figure 4 – Results of Applying Best Linear Discriminant Function to FLEA texts (RF=1 for text by myself; RF=0 for Anonymous texts; RF= -1 for texts signed by other named authors). Positive values of discfunc indicate **dissimilarity** to myself; negative values indicate similarity.

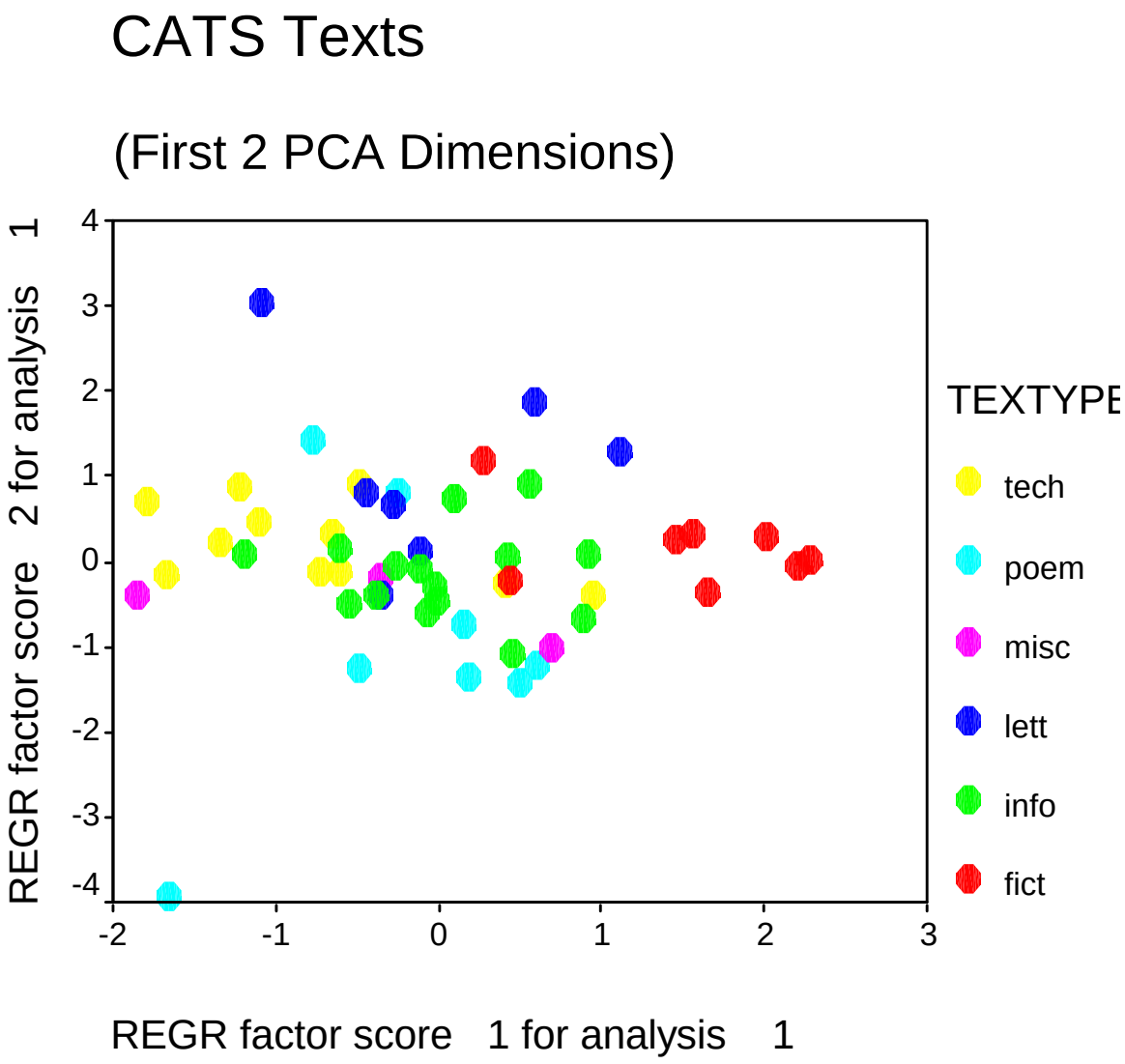

Figure 5 – CATS Texts plotted on First 2 PCA Dimensions, showing text-type.

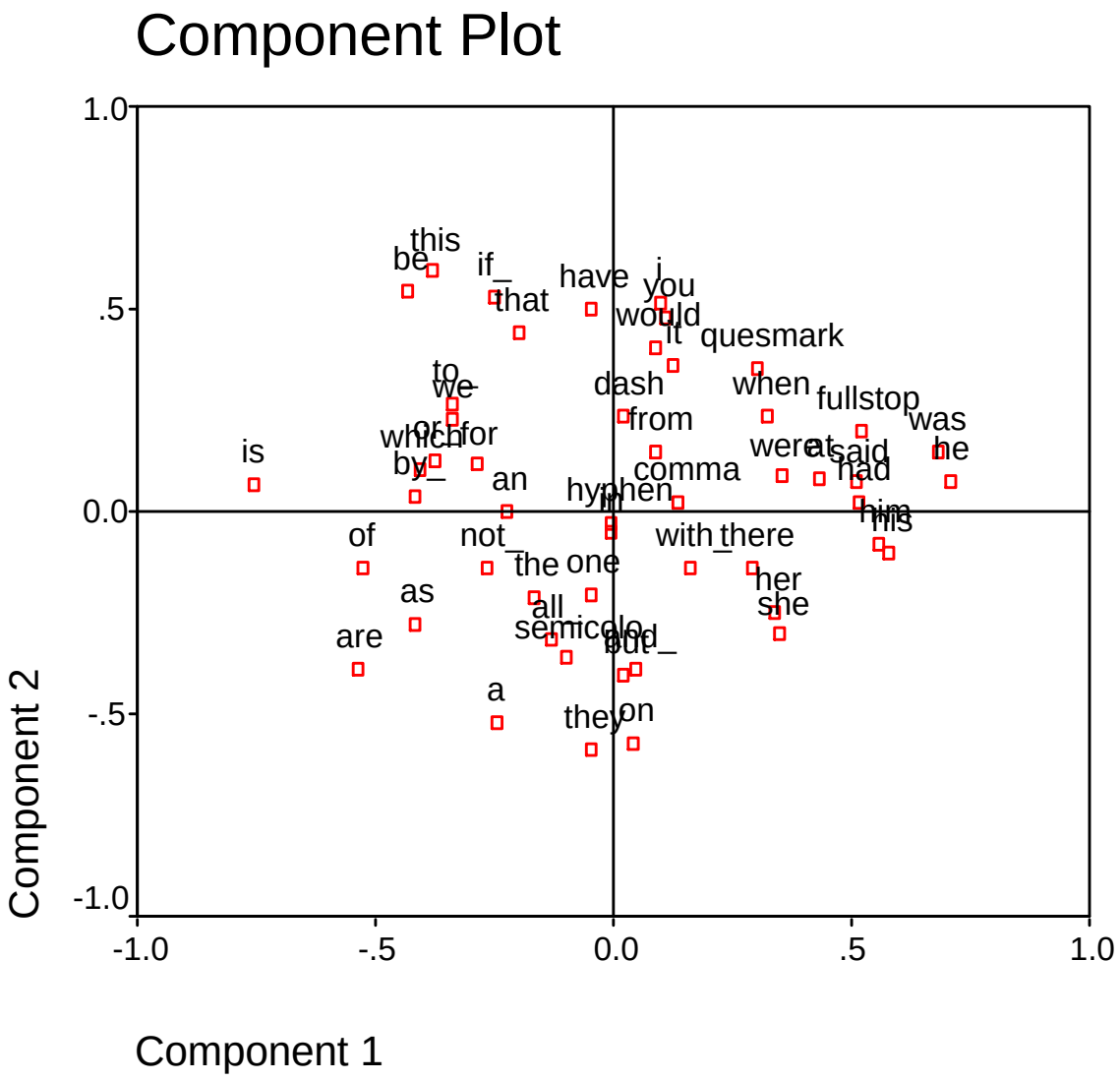

Figure 6 – Factor Loading Plot from CATS (commonest 50 tokens).

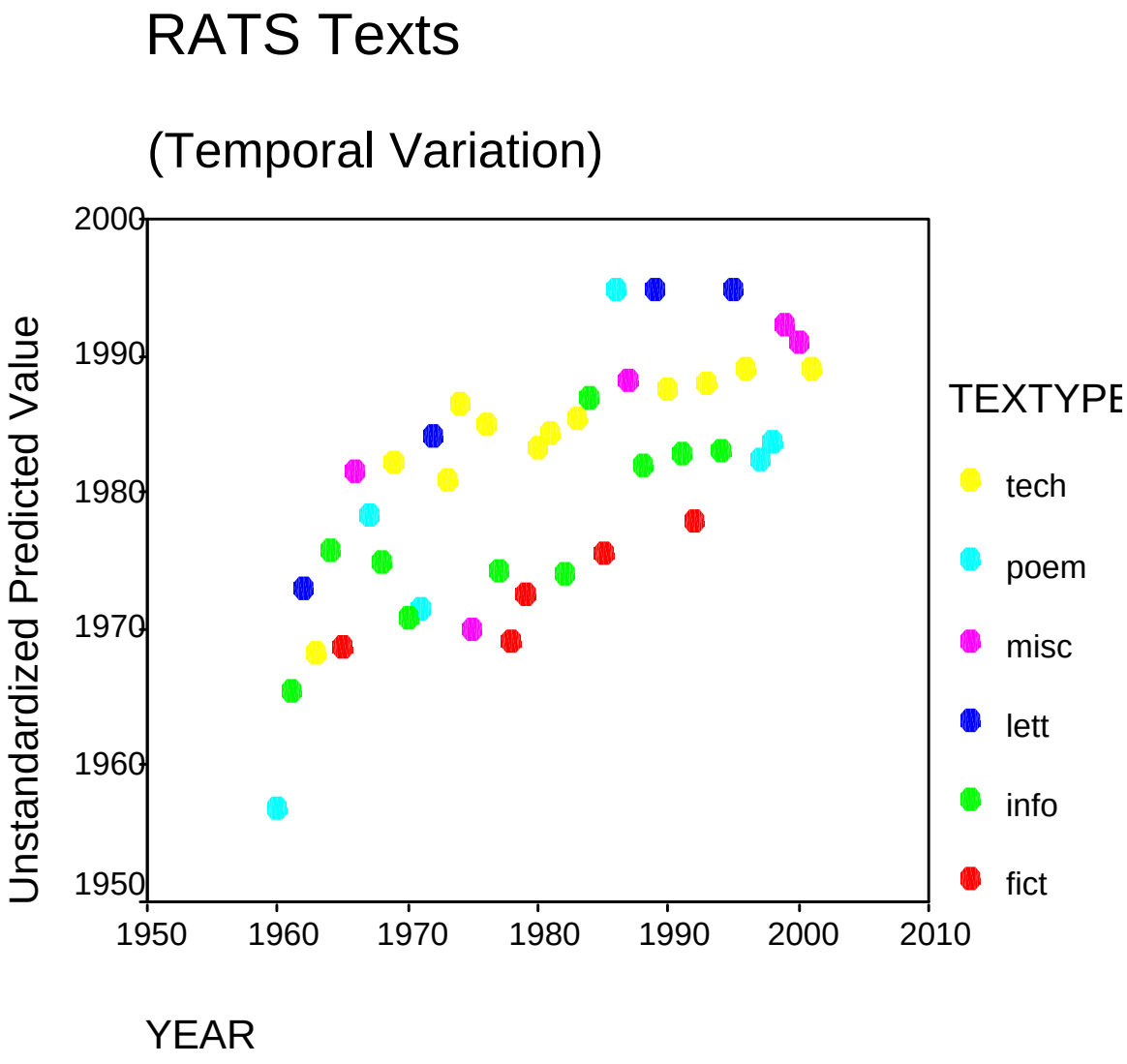

Figure 7 – Plot of Chronological Regression Formula (four variables selected) against Time for RATS texts. Remarkably, all genres appear to participate in the same trend.

On the RATS texts a stepwise regression of year against the top 50 tokens picked four variables, with an adjusted R-squared of 0.458. The regression formula for Y (predicted year) is:

```
Y = 1994.927 – 3.838 * and – 10.825 * all – 4.721 * was – 4.375 * he
```

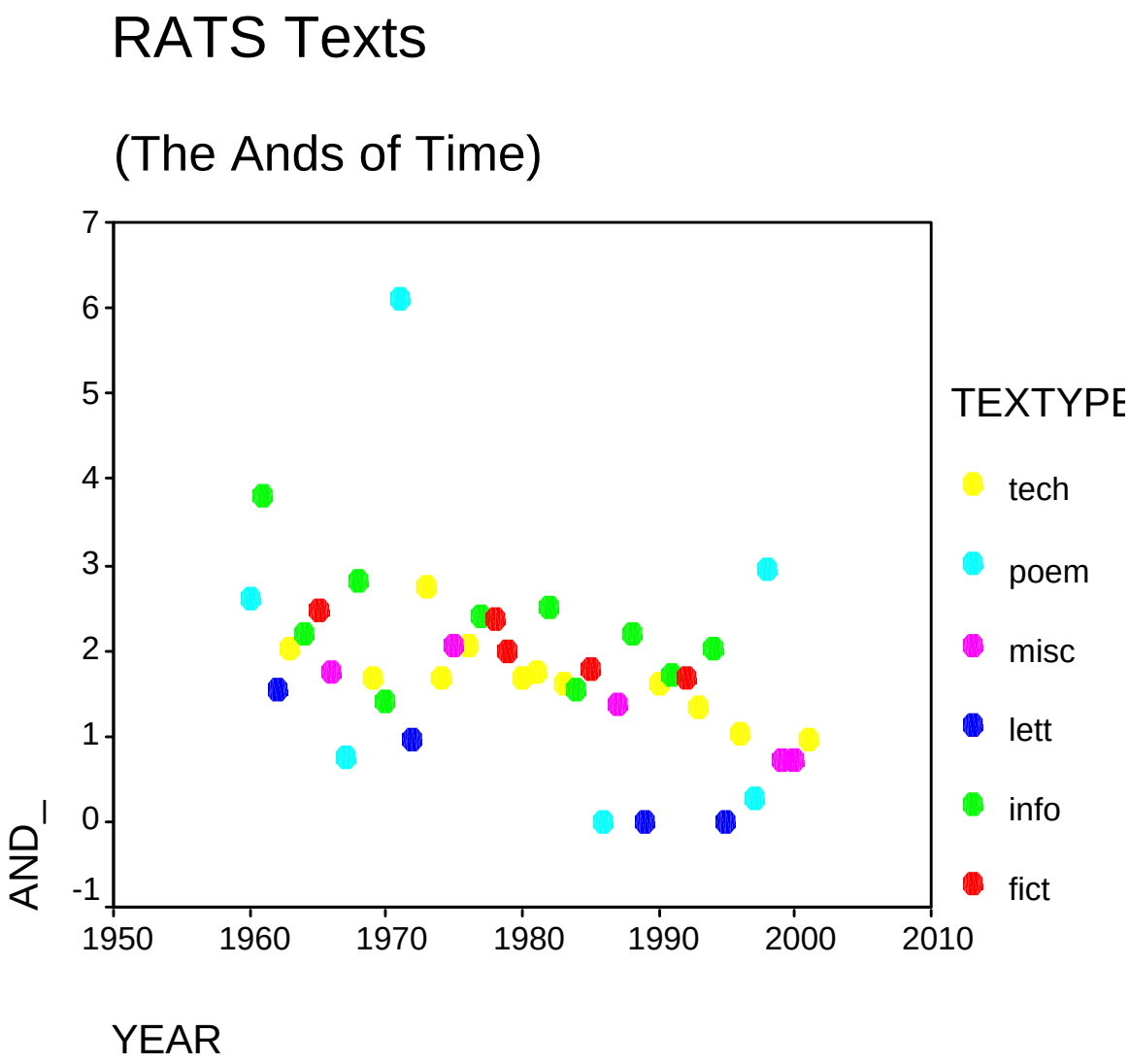

Figure 8 – Declining Frequency per 100 tokens of “and” in texts by RF from 1960 to 2001.