

The search for verbal fingerprints

George K. Barr, D.A.(Glas), ARIBA, BD, PhD.,

7 Tay Avenue, Cornhill, Perthshire, PH6 2PF

Abstract: Many scholars have attempted to detect patterns in speech or in writing which might identify the author as surely as a fingerprint does. The article considers a problem in taking statements which is becoming common at the present time. It reviews some of the features which scholars have used in attempting to discriminate between the works of different authors and finds that they fail to produce a verbal fingerprint. Finally, scale-related patterns are introduced which link different works in a way which has only recently been detected. These patterns may be significant in some instances of authorship determination and take the matter one step further in the search for the elusive verbal fingerprint.

Introduction

A lawyer recently approached me concerning a statement which had been made by a client who had a long history of repeated criminal offences. The statement was considered to be suspect as it contained some police jargon. Could I determine whether the statement had been doctored by the police? There are two things which may be said about this. Firstly, a person who has been questioned repeatedly by the police may become adept at using police jargon, and the person concerned in this case was an old hand at the game. Secondly, a police officer has problems in recording a statement made by someone who is semi-literate, whose every third word is "effing", who seldom completes a sentence grammatically and who supplements his words with "Ums" and "Ahs", nods and winks, shrugs and gestures, rude or otherwise. The police officer's job is to convey to the court the meaning of what the witness is trying to say, but shorn of the nods and winks, the fragmented verbatim text may not make much sense. The police officer is tempted to give some order to the text and may possibly supply some vocabulary. The resulting text may therefore owe something of its structure to the police officer, but in fact expresses what the witness was trying to say more adequately than a verbatim text would without its supporting gestures. For this reason, video recordings may be made when important statements are taken.

The cusum episode

A.Q. Morton believed that he had found a solution when he designed a system in which graphs representing grammatical features were superimposed on graphs derived from the sentence length pattern (the Cusum technique). When the whole text was by one person the graphs followed each other closely but the insertion of foreign material into the text caused the two graphs to separate. Evidence of this kind was presented as expert evidence in the appeal in London of Crossan (1991), the trial of Beck in Leicester (1991) and the trial of Connell in Dublin (1991), also in cases concerning allegedly forged confessions (Morgan, 1991 and Campbell, 1992) and in several prominent appeal cases since. This system has been

severely criticised by De Haan and Schils (1993), Hardcastle (1993), Canter (1992), Hilton and Holmes (1993), Sanford, Aked, Moxey and Mullin, (1994). and others, and now has no place in the preparation of expert evidence. As far as it worked, it depended on the fact that authors may have different habits in the use of grammatical features, and it was these differences that caused the separation of the graphs. It is now apparent, however, that different authors do not have habits which are so distinct that they may be compared to fingerprints which are considered to be unique in each person. Comparison of the occurrences of most grammatical features shows that occasionally the characteristics of different authors are quite distinct and separate, but that usually there are large areas of overlap between authors in their use of these features. Morton (1993) had shown that the epistle to Philemon had the same author as the epistle to Galatians as the insertion of the one into the other did not disrupt the matching of the graphs too much. In fact it can easily be shown that a far better match can be made by inserting Philemon into one of Prof. James S. Stewart's sermons, or into a short story by O. Henry. The characteristics of these texts are in places very similar to those of Philemon, but it would be absurd to suggest that the epistle to Philemon had been written by either or both of these modern writers.

Variable characteristics

The search for verbal fingerprints which might be used as expert evidence goes back to 1851 when Augustus de Morgan suggested that word length might prove to be the distinctive trait of a writer's style. This was soon shown to be inadequate and attention fastened on sentence length as a measure of individual style. As early as 1889 Parker showed that for detective purposes this method was valueless, yet as late as 1939 Prof. Udny Yule was persevering in the examination of sentence length as a discriminator. Others used sentence length to discriminate between the writings of certain Greek authors. As Fig. 1 shows, average sentence length does indeed provide complete separation of the works of some authors, but in the cases of other authors there is much overlap. Average

sentence length does not therefore provide a verbal fingerprint which is unique in each author. It can, on occasion, provide a balance of probability in favour of one author or another where the group of possible authors is very limited. But it can never be assumed that an author's characteristics will always fall within the ranges found in his or her known works. A clear example is the case of Isocrates. Over the greater part of a long writing career, the average sentence lengths of Isocrates' works fell between 31 and 44 words per sentence. In his last work, Panathenaicus, published when he was well over 90 years of age, the average shot up to 54 words, yet the authenticity of that work is not in dispute.

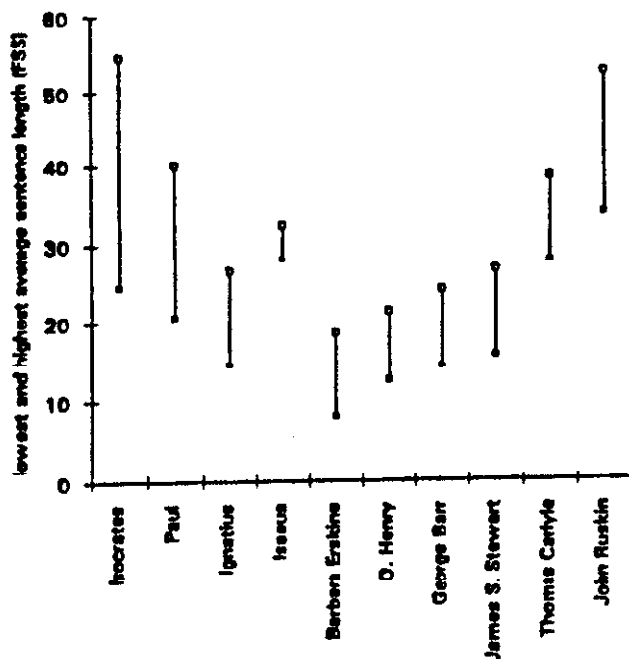


Figure 1

Morton held that the use of the Greek *KAI* or the English *AND* could provide characteristic patterns which would identify individual authors. These words, however, have little discriminating power. A study of the use of *KAI* in the works of St. Paul and the use of *AND* in the sermons and lectures of Prof. James S. Stewart shows that they not only cover virtually the same range (generally from 2.5% to 5.5% of the text) but also show similar rogue passages where the rate of occurrence drops to below 2%.

The rates of occurrence of two- and three-letter words, initial vowel words and nouns all show a great amount of overlap in their use by different authors and the combination of these features has proved to be problematical. None of these provides a reliable fingerprint.

Multivariate Analysis

Some statisticians have combined many features in studies employing multivariate analysis in the hope that the

combination of many elements would throw up distinctive patterns. Anthony Kenny (1986) used 96 features in his study of the Pauline epistles. Ledger (1989) employed 37 variables in his study of Plato's style. Neumann (1988) selected 617 indices in his study of the Pauline epistles. These substantial studies, however, all contained one fundamental flaw which may prove fatal. These authors did not appreciate that almost all parts of speech are *scale-sensitive*. The study of scale in the written and spoken word is a new field which has not yet been fully explored. All creative work in engineering, art, photography, architecture etc. takes place at a particular scale level. An engineer, for example, may be working on large scale oil rigs or on microscopic silicon chips. But while scale has been recognised and used as a working tool in disciplines like architecture, it is only recently that scale has been recognised in literature. One of its manifestations concerns sentence length; higher scale works tend to have longer average sentence lengths. A close study of samples of text exceeding fifty thousand words shows that most parts of speech are either high scale sensitive (occurring more frequently in works with high average sentence length) or low scale sensitive (occurring more frequently in works with low average sentence length). This is readily understandable if one considers that verbs in the imperative, for example, tend to occur in short, snappy sentences. Prepositions, on the other hand, are often related to the clauses of long sentences and occur more frequently in that context.

The important principle which affects multivariate studies is that data derived from material conceived at different scale levels cannot be simply added together. Measurements which relate to different scales are not additive. To illustrate, a journey may be planned using both a small scale road map and a large scale street map. The two sections of the journey, outside and inside the city, may be measured in inches on the respective maps. These dimensions, however, cannot be added together until they have been converted with regard to the respective scales of the maps. Likewise, data drawn from literary material cannot be combined in multivariate studies without appropriate conversion with regard to scale. To date, multivariate studies have failed to produce anything like a 'literary fingerprint' which might be used as expert evidence.

Scale-related Patterns

The recognition of scale differences does, however, open up new possibilities, because scale related patterns can be found which form links between works which have not been recognised by those using traditional statistical methods. These scale-related patterns are found in sentence sequence patterns, in sentence length distribution patterns, and in the way in which sentence expansion can be achieved. A simple example is shown in the sentence sequence patterns in Fig. 2.

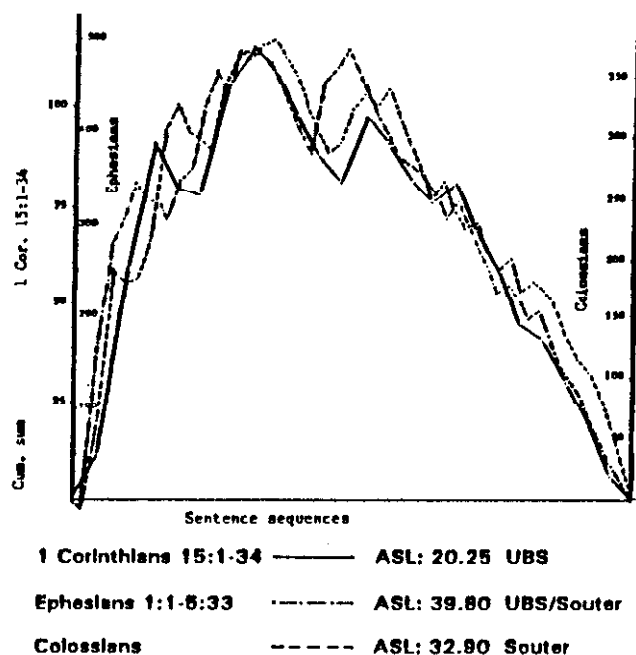
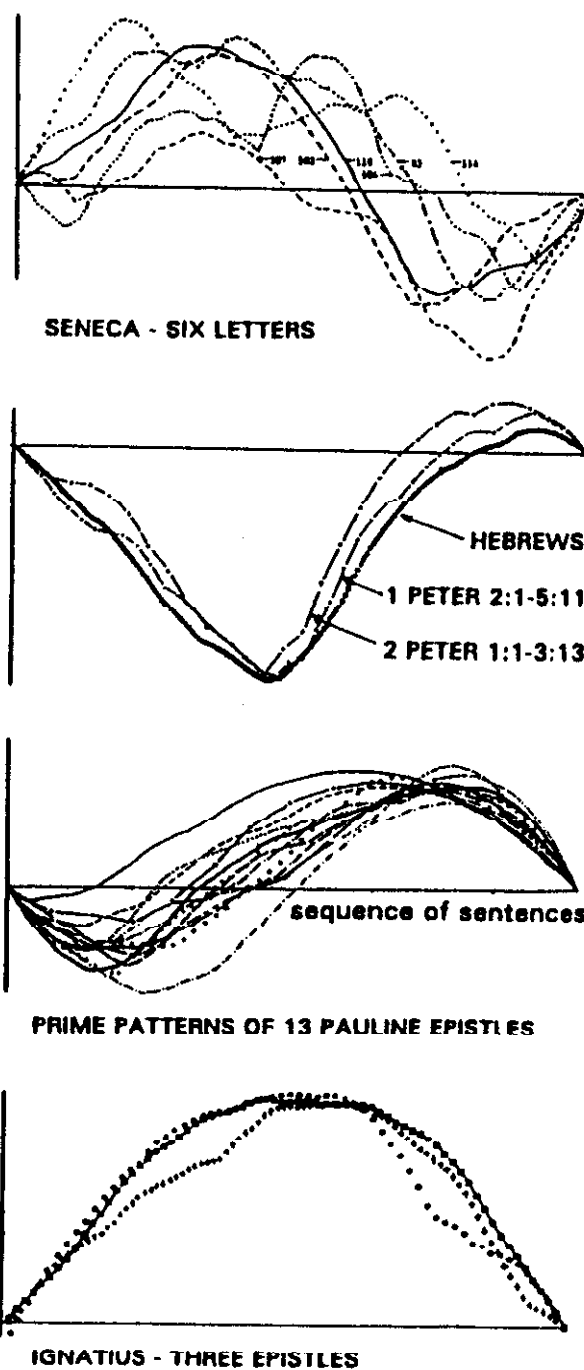


Figure 2.

These are of two whole epistles from the Pauline corpus (Ephesians and Colossians) and a small section of text from 1 Corinthians. The lengths of these texts are very different, ranging from 483 words in the case of the portion from 1 Corinthians to 1582 in Colossians and 2422 words in Ephesians. When these cumulative sum graphs of the variations in successive sentence lengths are drawn out giving careful consideration to the respective scales, the graphs are found to match closely. In other words, the structure of one is a small scale version of the structure of the others.

Where there is a great difference in the lengths of works being compared a problem arises in that the longer works have scope for a great amount of detail to develop which appears as a confusing saw-tooth pattern. The solution to this is to employ a 'SuperQsum' which is a development of the normal cumulative sum graph. In the SuperQsum, the individual sentence lengths are replaced by the respective cumulative sum values. This has the effect of ironing out the variations in the graphs and providing a comparatively smooth curve which expresses the substructure of the text. Fig. 3 compares SuperQsums of selections of several authors' works. It is evident that quite distinctive patterns are obtained for each author which look deceptively like 'verbal fingerprints.'

A close study of the graphs of over half a million words in the works of authors both ancient and modern has not identified any patterns in other authors which might be mistaken for those found in Paul or in the Hebrews/Peter group. This provides hard new evidence regarding authorship which may be of service in Biblical studies. It must be said, however, that while an author may provide



SUPERQSUMS OF SEVERAL AUTHORS

Figure 3.

such clear characteristic patterns in some of his or her works, not all of which work may conform to that pattern. These scale-related patterns, where they occur, may therefore be useful in academic study, but it cannot be expected that they will meet the criteria which expert evidence is required to meet. Nevertheless, they do take one stage further the search for a verbal fingerprint which

may one day identify individual authors with sufficient consistency to be used as expert evidence.

References

- Barr, G.K., 1995, Scale and the Pauline Epistles, *Irish Biblical Studies*, 17: 22-41.
- Barr, G.K. (in press), The use of cumulative sum graphs in literary scalometry, *Literary and Linguistic Computing*, 12.
- Barr, G.K., (under review), Scale sensitive parts of speech in multivariate analysis.
- Barr, G.K., 1997, The structure of Hebrews and of 1st and 2nd Peter, *Irish Biblical Studies*, 19: 1-15.
- Canter, D., 1992, An Evaluation of the "Cusum" Stylistic Analysis of Confessions, *Expert Evidence*, 1 (2): 93-99.
- De Haan, P. and Schils, E., 1993, The Qsum plot exposed, (Papers from the Fourteenth International Conference on English Language Research on Computerised Corpora, Zürich).
- Hardcastle, R.A., 1993, Forensic linguistics: A'n assessment of the CUSUM method for the determination of authorship, *Journal of the Forensic Science Society*, 33(2): 95-106.
- Hilton, M. and Holmes, D., 1993, An assessment of cumulative sum charts for authorship attribution, *Literary and Linguistic Computing*, 8(2) 73-80.
- Kenny, A., 1986, *A Stylometric Study of the New Testament*, Oxford: Clarendon Press.
- Ledger, G.R., 1989, *Re-counting Plato: A Computer Analysis of Plato's Style*, Oxford: Clarendon Press.
- Morton, A.Q., 1991, *Proper Words in Proper Places*, (Computing Science Research Report), University of Glasgow.
- Morton, A.Q., 1993, *The Authorship of the New Testament Epistles*, (Computing Science Research Report), University of Glasgow.
- Neumann, K.J., 1988, *The Authenticity of the Pauline Epistles in the Light of Stylostatistical Analysis*. University of Toronto. SRI. Dissertation Series.
- Sanford A.J., Aked, J.P., Moxey L.M. and Mullin, J., 1994, A critical examination of assumptions underlying the cusum technique of forensic linguistics, *Forensic Linguistics*, 1(2):