New Test. Stud. 54, pp. 139–159. Printed in the United Kingdom © 2008 Cambridge University Press
DOI:10.1017/S0028688508000088

Recent Advances in Computational Linguistics and their Application to Biblical Studies*

J. JOSÉ ALVIAR
Faculty of Theology, University of Navarra, 31080 Pamplona, Spain

This article focuses on novel computer-based techniques for style characterization which have been tested on NT texts. The techniques are derived from the fields of information theory, communications engineering, and bioinformatics, and treat text as linearly sequenced information. They employ computer algorithms capable of detecting patterns in character strings, thereby permitting characterization of a given text or comparison of various texts. The application of the techniques to NT books has so far yielded results generally concordant with other methods, and suggest that the new techniques, if further refined, could complement other approaches to Biblical-textual questions.

Keywords: Stylistics, Biblical analysis, Computer analysis, Linguistics

1 Introduction

The 'feel' of a text – the impression of uniqueness conveyed upon the reader by a cumulus of details such as peculiar words, turns of phrase, or lines of thought – that is what 'stylometry' attempts to quantify. As early as 1851, the English mathematician Augustus de Morgan suggested that measuring the average length of words of the NT books could provide an insight on the authorship of the letter to the Hebrews. The field of stylometry has advanced rapidly since then, and in recent times has been helped by the advent of the computer age. The computer has made possible the storage of large textual corpora in manageable digital form, and given rise to automated procedures for language analysis (computational linguistics). The present article focuses on a novel class of computer-based

* The author wishes to express his gratitude to the following for providing invaluable assistance in the preparation of this article: Derek Abbott, Matthew Berryman, Vittorio Loreto, and William Teahan, for supplying information on experimental results and expert advice; Bernardo Estrada, Robert Kraft, and Réginald Ferdinand Poswick, for providing additional information on publications and institutions; Dither Ruiz and Vincent Hernandez for collaborating in the compression-technique experiments and the preparation of the figures.

techniques, which have been used recently to analyse the style of NT books. The results obtained, and the implied presence of a new conceptual approach to textual questions, are, we feel, sufficiently intriguing to warrant a review.

To situate the new techniques in their proper historical context, we first draw a brief summary of the evolution of stylometry and computational linguistics. It is not our intention to do an exhaustive review (this has already been done by others, as will be seen in the citations below), but rather to open a narrow window, focusing on approaches to style analysis derived from unusual fields such as data compression and bioinformatics. In the central part of the article, we describe the methods and their results. Finally, we remark on the limitations and potential of the new techniques, in the hope that biblical scholars and linguists will consider giving attention to advances in 'foreign' scientific areas.

2 The Evolution of Stylometry and Computational Linguistics[1]

a) *A Brief History*

The first attempts in the paradoxical enterprise of 'measuring' literary style appeared in the late nineteenth and early twentieth centuries. They involved simple criteria for characterising textual pieces, such as the average word length (Thomas Corwin Mendenhall), average sentence length (Udny Yule), or frequencies of different words (George Kingsley Zipf, Yule). In the latter half of the twentieth century more sophisticated techniques came into being. Frederick Mosteller and David L. Wallace, in an influential 1964 study, showed how the frequency of commonly occurring 'function words' (such as prepositions, conjunctions, and articles) could help in discriminating styles. In the 1970s, Barron Brainerd tested the indicative value of article and pronoun frequencies, whereas Andrew Q. Morton tried criteria such as word position within a sentence or in relation to another word. The decades of the 1980s and 1990s saw more progress – in the first place, the appearance of more refined techniques based on word frequency. Herbert Simon Sichel and Morton focused on one extreme of the word frequency distribution, studying the potential of *hapax legomena* and of *dislegomena*,

[1] For more comprehensive reviews, see D. I. Holmes, 'Authorship Attribution', *Computers and the Humanities* 28 (1994) 87–106; idem, 'The Evolution of Stylometry in Humanities Scholarship', *Literary and Linguistic Computing* 13 (1998) 111–17; H. Craig, 'Stylistic Analysis and Authorship Studies', *A Companion to Digital Humanities* (ed. S. Schreibman, R. Siemens and J. Unsworth; Oxford: Blackwell, 2004) 273–88; J. Burrows, 'Questions of Authorship: Attribution and Beyond', *Computers and the Humanities* 37 (2003) 5–32; A. Dean Forbes, 'Statistical Research on the Bible', *ABD* 6.185–206; H. van Dyke Parunak, 'Computers and Biblical Studies', *ABD* 1.1112–24; R. F. Poswick, 'Si la Bible m'était comptée', *Actes JADT 1998* (Nice: CNRS, 1998) 517–27; 'Statistiques et Bible', *Dictionnaire Encyclopédique de la Bible* (Turnhout: Brepols, 2002) 1232–3.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

respectively, to obtain an idea of texts' vocabulary richness, while M. W. A. Smith and John F. Burrows concentrated on the other end of the distribution, reasoning that the statistics of very frequent words could be useful for characterising style. Burrows developed a powerful method that took large sets of frequently occurring function words, then mathematically combined the statistical information of the sets to uncover an author's distinctive use of language. The philosophy underlying his method – that of using not just one stylistic criterion, but rather a mathematical combination of a large set of different criteria (multivariate analysis) – has become predominant in stylometry. It lies at the heart of the various methods currently being used by researchers like Hugh Craig, Peter Dixon, David I. Holmes, David L. Hoover, Gerard R. Ledger, David Mannion, Thomas V. N. Merriam, and David L. Mealand.

The dawn of the computer era had a deep impact on linguistic studies, as reflected in the birth and growth of numerous interdisciplinary institutions and journals.[2] The application of computers to linguistic questions, however, is still in its early stages. Holmes, in a 1998 survey of stylometric techniques, included a section on 'Stylometry and Artificial Intelligence' and referred to artificial intelligence (AI) techniques as 'exciting tools for the future'.[3] Some investigators are in fact currently testing AI approaches to style analysis, such as artificial neural networks (Bradley Kjell, Robert A. J. Matthews, Fiona J. Tweedie, Sam Waugh), genetic algorithms (Richard S. Forsyth, Holmes), and comparison of character sequence probabilities (Patrick Juola). These methods work on digitalized texts, and employ computer programs capable of detecting or comparing patterns in the digital representations of texts. Their basic assumption is that a machine with suitably designed software can detect features that are not evident to the human eye.

b) *Applications to Bible Studies*

Not surprisingly, during this time a number of stylometric techniques have been tested in application to biblical texts. The latter half of the twentieth century, especially, saw a growing number of stylometric methods being brought to bear on Bible texts. The question of authorship of books in the Pauline corpus was met by Kenneth Grayston and Gustav Herdan using a technique based on vocabulary counts, and by Morton with a method relying on features such as sentence length,

[2] For more details, see S. Hockey, 'The History of Humanities Computing', *A Companion to Digital Humanities*, 3–19. Among the institutions that work more specifically with Bible and computers are: the Association Internationale Bible et Informatique (AIBI, founded in 1982 through the initiative of Réginald Ferdinand Poswick, and organizer of the colloquium series 'Bible and Computer'); the Computer Assisted Research Group (CARG, begun in 1979 under the auspices of the Society of Biblical Literature); and the Center for Computer Analysis of Texts (CCAT) of the University of Pennsylvania, under the direction of Robert A. Kraft.

[3] Holmes, 'The Evolution of Stylometry in Humanities Scholarship', 114–15.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

function word frequency and function word positioning within sentences. Other researchers attempted to compare the styles of OT texts – Ronald E. Bee using verb frequencies, Yehuda T. Radday a combination of statistical markers. Among Radday's studies, the one on Genesis is perhaps best known.[4] Relying heavily on computers, it applied a large number of statistical techniques: univariate and multivariate analyses, analysis of pattern vector distribution, cluster analysis, smallest space analysis, reliability analysis, factor analysis, and estimation of vocabulary richness through the Sichel distribution. It led to a conclusion contrary to the venerable Wellhausen hypothesis: Genesis appeared to be more of a unity as far as authorship was concerned, though stylistically measurable differences were noticeable among the 'narrative' passages, the 'God-spoken' ones, and the 'human-spoken' ones.

Anthony Kenny, in another study that applied simpler statistical tools to NT texts, also achieved interesting results.[5] His analysis, based principally on different parts of speech (conjunctions and particles, prepositions, articles, nouns and pronouns, verbs, adjectives, adverbs) and their statistical distribution in the NT books, as well as on the commonest words, last words, sentence lengths, and preposition positions, led to conclusions such as the following: the likeness between Luke and Acts points to one same author; the Apocalypse is dissimilar to John's Gospel [and therefore could have a different author]; the Pauline epistles constitute a stylistic 'cluster', with the distance of individual epistles from the nuclear group varying in this outward order: Rom, Phil, 2 Tim, 2 Cor, Gal, 2 Thess, 1 Thess, Col, Eph, 1 Tim, Phlm, 1 Cor, Titus.

In the 1990s, following the general trend in stylometry, researchers like Étienne Brunet and Mealand applied multivariate techniques (factor analysis or correspondence analysis) to the stylistic study of NT texts. Further methods have been reported in more recent times, with the same underlying philosophy as the AI approaches mentioned above – that of viewing text as linearly sequenced information, characterizable with pattern-detecting computer programs. These techniques are unusual, though, for two reasons: (1) they use procedural tools developed in areas quite removed from linguistics or Bible studies – e.g. the fields of information retrieval (data mining), electronic and communications engineering (signal processing and data compression) and biology (DNA sequencing and phylogenetic tree mapping); (2) the methods, relying heavily as they do upon enhanced computing power, could hardly have been envisioned for use just a few decades ago – indeed, they are strictly 'computer-based', not merely 'computer-

[4] Y. T. Radday, et al., *Genesis: An Authorship Study in Computer-Assisted Statistical Linguistics* (Rome: Biblical Institute, 1985). For a summary of the critique, see Forbes, 'Statistical Research', 199–201.

[5] A. Kenny, *A Stylometric Study of the New Testament* (Oxford: Clarendon, 1986). For a summary of the critique, see Forbes, 'Statistical Research', 193.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

assisted', techniques.[6] In the remainder of this article we focus on these novel techniques – not because they are definitely superior to older methods, but because of the perspectives they suggest for biblical and linguistic studies.

3 Some Recent Methods

a) *Techniques Based on Data Compression*

(1) *Overview*

At the turn of the millennium, several research teams (in the United Kingdom, Russia, and Italy) working on the problem of discriminating linear character sequences (e.g. literary texts, DNA sequences, transmitted signals, etc.), turned their attention to ideas first conceived in the field of information theory. The techniques described by William J. Teahan and David J. Harper,[7] Olga V. Kukushkina, Anatolij A. Polikarpov and Dmitry V. Khmelev,[8] and Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto,[9] were all based on a 'stochastic' model of written language, which conceived the characters of a text as having been 'emitted' one at a time by a source (= the writer), with a probability function governing the likeliness of any given character being emitted next. Such a probability function (an indication of the 'richness' of the information contained in the text – the 'entropy'[10] or 'complexity',[11] in the more technical wording of the authors) might be sufficiently distinctive, the researchers argued, as to permit identification not merely of the idiom, but even of the author, of a given text.

[6] This would seem to bear out H. van Dyke Parunak's scepticism over attempts at writing a history of 'the application of computers to the Bible': 'the field', he stated, 'is progressing so rapidly that such a history would immediately be out of date' ('Computers and Biblical Studies', 1112).

[7] W. J. Teahan, 'Text Classification and Segmentation Using Minimum Cross-Entropy', *Proceedings of the RIAO 2000 Conference* (Paris: Centre de Hautes Études Internationale D'Informatique, 2000) 943–61; W. J. Teahan and D. J. Harper, 'Using Compression-Based Language Models for Text Categorisation', *Language Modeling for Information Retrieval* (eds. W. Croft and J. Laferty; Boston: Kluwer Academic, 2003) 141–66.

[8] O. V. Kukushkina, A. A. Polikarpov, and D. V. Khmelev, 'Using Literal and Grammatical Statistics for Authorship Attribution', *Problems of Information Transmission* 37 (2001) 172–84.

[9] D. Benedetto, E. Caglioti, and V. Loreto, 'Language Trees and Zipping', *Physical Review Letters* 88 (2002) 48702(1–4).

[10] C. E. Shannon: 'A Mathematical Theory of Communication', *Bell System Technical Journal* 27 (1948) 379–423, 623–56.

[11] R. J. Solomonov, 'A Formal Theory of Inductive Inference', *Information and Control* 7 (1964) 1–22, 224–54; A. N. Kolmogorov, 'Three Approaches to the Quantitative Definition of Information', *Problems of Information Transmission* 1 (1965) 1–17; G. J. Chaitin: 'On the Length of Programs for Computing Finite Binary Sequences', *Journal of the Association for Computer Machinery* 13 (1966) 547–69; idem, *Information, Randomness and Incompleteness: Papers on Algorithmic Information Theory* (Singapore: World Scientific, 1987).

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

In practice, it is not feasible to arrive exactly at such a probability function, so the research teams looked for ways of approximating it. Following earlier leads, they turned to existing text-compression algorithms (PPM = Prediction by Partial Match, in the case of Teahan and Harper; LZ77 = Lempel-Ziv, in the case of Benedetto, Caglioti, and Loreto; 16 different algorithms, in the case of Kukushkina, Polikarpov, and Khmelev). Previous research had already shown such compression algorithms to be capable of yielding an estimate of the 'richness' of information in a text.[12] (This may be grasped intuitively as follows: one expects a text with numerous repetitions to be easier to summarize; and a text with hardly any repeated elements to be harder to condense). The estimate of 'richness' of a text could then be contrasted, further studies had suggested, with estimates of 'richness' of other texts. In this way, an approximate procedure for textual comparison could be derived. (In the researchers' more technical terms, the problem was one of measuring the 'relative entropy', 'cross-entropy', or 'relative complexity' between any given pair of texts).[13] It is the application of data compression algorithms (packaged in readily available computer programs) to an 'old' problem –

[12] J. Ziv and A. Lempel, 'Compression of Individual Sequences via Variable-Rate Coding', *IEEE Transactions on Information Theory* 24 (1978) 530–6; A. D. Wyner and J. Ziv, 'Some Asymptotic Properties of the Entropy of a Stationary Ergodic Data Source with Applications to Data Compression', *IEEE Transactions on Information Theory* 35 (1989) 1250–8; T. M. Cover and J. A. Thomas, *Elements of Information Theory* (New York: John Wiley & Sons, 1991) 78–124; P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. C. Lai, and R. L. Mercer, 'An Estimate of an Upper Bound for the Entropy of English', *Computational Linguistics* 18 (1992) 31–40; D. S. Ornstein and B. Weiss, 'Entropy and Data Compression Schemes', *IEEE Transactions on Information Theory* 39 (1993) 78–83; A. D. Wyner, 'Typical Sequences and All That: Entropy, Pattern Matching and Data Compression', *IEEE Information Theory Society Newsletter* (June 1995) 8–14; M. Farach, M. Noordewier, S. Savari, L. Shepp, A. Wyner, and J. Ziv, 'On the Entropy of DNA: Algorithms and Measurements Based on Memory and Rapid Convergence', *Proceedings of the 6th Annual ACM-SIAM Symposium on Discrete Algorithms* (Philadelphia: SIAM, 1995) 48–57; W. J. Teahan and J. G. Cleary, 'The Entropy of English Using PPM-based Models', *Proceedings of the 1996 Data Compression Conference* (Los Alamitos: IEEE Computer Society, 1996) 53–62; M. Li and P. M. B. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications* (2nd ed.; New York: Springer, 1997) 325–78, 476–88; A. D. Wyner, J. Ziv and A. J. Wyner, 'On the Role of Pattern Matching in Information Theory', *IEEE Transactions on Information Theory* 44 (1998) 2045–56; I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner, 'Nonparametric Entropy Estimation for Stationary Processes and Random Fields, with Applications to English Text', *IEEE Transactions on Information Theory* 44 (1998) 1319–27.

[13] J. Ziv and N. Merhav, 'A Measure of Relative Entropy between Individual Sequences with Applications to Universal Classification', *IEEE Transactions on Information Theory* 39 (1993) 1280–92; Li and Vitanyi, *An Introduction to Kolmogorov Complexity*, 537–54; P. Juola, 'What Can We Do with Small Corpora? Document Categorisation via Cross-Entropy', *Proceedings of the Interdisciplinary Workshop on Similarity and Categorisation* (Edinburgh: University of Edinburgh, 1997) 137–42; idem, 'Cross-Entropy and Linguistic Typology', *Proceedings of the Joint Conference on New Methods in Language Processing and Computational Language Learning* (Somerset: ACL, 1998) 141–9.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

textual comparison – which constitutes the novelty of the approach. We now offer a summary of two specifications of this approach.

(2) *The PPM (Prediction by Partial Match) approach*

The PPM algorithm for text compression was proposed by John G. Cleary and Ian H. Witten in 1984,[14] and remains as one the most efficient compression schemes in use today. In simplified terms, this algorithm encodes characters in a text one by one. First it predicts the upcoming character, based on the few characters (usually not more than 5) immediately preceding, relying on frequency counts done for earlier character sequences. (The frequency counts are continually updated as compression proceeds.) The actual upcoming character is then encoded in terms of its mathematical probability of appearing after the preceding letters. In real applications, the PPM algorithm permits highly efficient compression of linearly sequenced information.

In articles published between 2000 and 2003, Teahan and Harper described how the PPM algorithm could be used to classify or compare various texts.[15] First, they applied the algorithm to a set of 'known' texts, arriving in each case at a 'trained' PPM compression procedure. Next, they applied the 'learned' compression procedures to 'test' texts of 'unknown' provenance. Since a learned compression scheme derived from a training text reflects the specific probability of the successive appearance of different characters, such a compression scheme, when applied to test texts, performs efficiently whenever the test texts have character sequences similar to those of the training text; and less efficiently when the character sequences in the test texts are unlike those in the training text. Thus the ease or difficulty in compressing a test text using different 'learned' compression schemes gives a measure of the 'closeness' of the test text relative to the various training texts. Teahan and Harper reported high rates of success in matching texts (in modern Western languages) of 'unknown' authorship with 'reference' texts of known authorship (e.g. the attribution of segments of Reuters news articles to different Reuters reporters, or the attribution of 12 disputed 'Federalist Papers' to one of three suspected authors).

A further idea was set forth in 2001 by the Russian team comprised of Kukushkina, Polikarpov, and Khmelev, in their article on authorship attribution using Markov chain models.[16] As a contrast to the Markov chain method, they described in an Appendix an alternative technique employing 16 different text

[14] J. G. Cleary and I. H. Witten, 'Data Compression Using Adaptive Coding and Partial String Matching', *IEEE Transactions on Communications* 32 (1984) 396–402.

[15] See note 7. Teahan had earlier worked with PPM compression in his PhD thesis: 'Modelling English Text' (Department of Computer Science, University of Waikato, Hamilton, 1998), obtaining peculiar results (pp. 135–7) for the book of Revelation.

[16] See note 8.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

compression algorithms. The authors suggested that an approximate measure of the 'closeness' between any two documents (their 'relative complexity') could be obtained through a simple procedure, summarized thus: For two 'documents' to be compared, say α and β, extracts may be taken – A and B, respectively. A and B may then be concatenated and compressed (using a compression algorithm); additionally, B alone may be compressed. The authors reasoned that the *difference* between the length of the compressed AB text and the length of the compressed B text (in shorthand, $L_{AB} - L_{B}$) should be *smaller* the more similar extract B is to extract A. If A and B are truly representative extracts of sources α and β, then we have a simple way of approximately measuring the 'affinity' between documents α and β. The authors tried out this method in matching 'unidentified' texts with texts of 'known' authorship (all texts in Russian), and achieved high rates of success in assigning 'unknown' texts to their real authors (though the success rates were generally inferior to the rate obtained via the Markov chain method).

A further development was reported in 2005, when an Australian research team comprised of Madeleine Sabordo, Shong Y. Chai, Matthew J. Berryman, and Derek Abbott[17] combined the ideas described above with a procedural suggestion set forth by the Italian research group of Benedetto, Caglioti, and Loreto.[18] In order to estimate the stylistic 'distances' among NT books, the Australian team proceeded as follows. First they took short extracts (e) from different books of the NT. They then appended the 'e' extracts to whole books of the NT (E), to generate all possible E+e combinations. Afterwards, they applied the PPM compressor to the E+e texts. Again, theory predicts that where E and e have similar character sequences, the model generated by the PPM algorithm in compressing the first section E should continue to work efficiently as it starts work on the appended e section, and the resulting file ought to be relatively small-sized. And vice-versa: a bigger, less efficiently compressed E+e file may be expected in cases where the appended e fragment has character sequences quite different from the E file.

The Australian research group took note of the numerical sizes (in bits) of the resulting compressed E+e files and, using a formula proposed by the Italian research team, calculated the 'distance metric'[19] between pairs of NT books.

[17] M. Sabordo, S. Y. Chai, M. J. Berryman and D. Abbott, 'Who Wrote the Letter to the Hebrews? Data Mining for Detection of Text Authorship', *Smart Structures, Devices, and Systems II. Proceedings of the SPIE* 5649 (2005) 513–24.

[18] See note 9.

[19] This formula may be described in the following general way. Let α and β be two different 'books' to be compared. From these 'books' are taken long extracts (let us call them A and B, respectively) and shorter ones (a and b respectively). The extracts are cross-combined to form the text files A+b, B+b, B+a, and A+a. These files are all encoded with a compression algorithm. Extract A and extract B are separately compressed. Afterwards, the length L (in bits or bytes) of each resulting compressed file is measured. The delta (Δ) values are then cal-

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

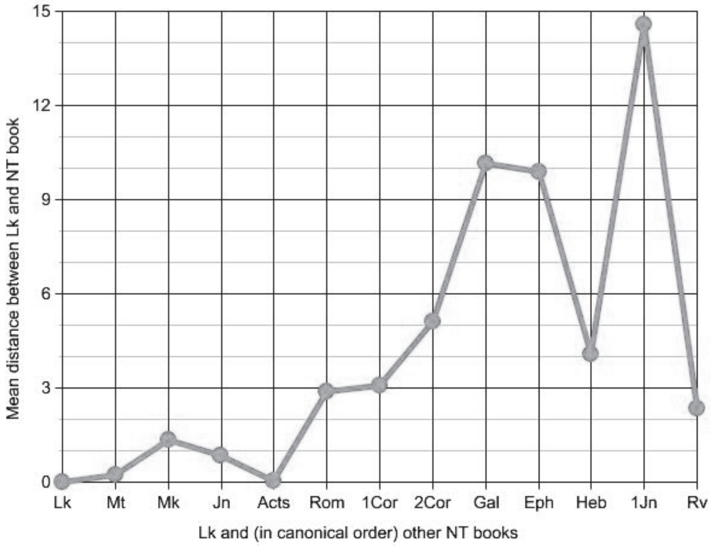

Figure 1. Plot of distances between Lk and other NT books

Figure 1 shows the 'distances' calculated by the Australian team between Lk and other NT books using the PPM technique.

A 'closeness' to the reference book (in this case, Lk) may be observed for the three other Gospels, as well as for Acts (seen in this graph as especially close to Lk); and a greater 'distance' relative to Lk may be noted for the Pauline epistles. Among the Pauline works, the results for the longer epistles Rom, 1 and 2 Cor are similar; in turn, these are rather removed from the results for the shorter epistles Gal and Eph. (Evidently, the method as described here only permits direct comparison between one 'reference' book and other NT books; however, by comparing the different 'distances' from the common reference book one indirectly obtains an idea of the 'closeness' of the other NT books with respect to each other.)

The Australian team concluded that the PPM algorithm yielded sufficiently precise results and was preferable to the LZ77 algorithm propounded in 2002 by the Italian group, whose method we now proceed to discuss.

culated as follows: $\Delta_{Ab} = L_{A+b} - L_A$; $\Delta_{Bb} = L_{B+b} - L_B$; $\Delta_{Ba} = L_{B+a} - L_B$; $\Delta_{Aa} = L_{A+a} - L_{A.}$ Finally, from the delta values the so-called 'distance metric' ($S_{\alpha\beta}$) between 'books' α and β may be computed thus: $S_{\alpha\beta} = (\Delta_{Ab} - \Delta_{Bb}) / \Delta_{Bb} + (\Delta_{Ba} - \Delta_{Aa}) / \Delta_{Aa}$.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

(3) *The LZ77 (Lempel and Ziv) approach*

In 1977 Jacob Ziv and Abraham Lempel proposed one of the earliest compression algorithms, which came to be called LZ77.[20] Though the original version has subsequently been refined, its basic concept continues to lie at the heart of such widely used compression programs as zip and gzip. Simply put, the LZ77 algorithm proceeds in linear fashion: it 'remembers' previously encountered character sequences and thus detects any subsequent recurrence of a specific sequence. It then encodes the information of the repeated sequence economically, employing just two numerical references – to the *length* of the repeated string, and to its *distance* from the previous identical string.

The Italian team of Benedetto, Caglioti, and Loreto attracted attention in 2002 when they published an article in *Physical Review Letters* suggesting the use of this algorithm for text categorization.[21] As briefly mentioned above, their method included an interesting twist – that of appending short samples 'e', extracted from different source texts, to longer samples 'E' extracted from the same corpus of texts, then using the gzip program to compress all the E+e files. Theory predicts that compression should be most efficient when an appended file e is most 'like' the file E to which it is attached. (In this sense, the gzip algorithm behaves like the PPM algorithm described above.) This is so because as the program proceeds with the compression of E, it 'memorizes' character sequences typical to E, so that by the time it gets to compressing e it is already equipped with a set of rules for summarizing e's character sequences. When e is similar to E, many of the character sequence rules learned for E apply, encoding is economically done, and the resulting compressed file E+e turns out to be small; on the other hand, when the appended e text is dissimilar to E, unfamiliar sequences of characters are encountered in e, impossible to resume with the rules previously learned for E, and encoding file E+e becomes more cumbersome (and the resulting compressed E+e file is consequently bigger).

In one experiment, Benedetto, Caglioti, and Loreto extracted long (E) and short (e) samples from different Italian literary works, generated all possible E+e combinations, applied gzip compression, and measured the size (L_{E+e}) of each compressed file. They also compressed the E extracts taken from the different books and measured the resulting file size in each case (L_E). The lowest values for $L_{E+e}-L_E$ were predicted by theory to correspond to those cases where the documentary source of sample e was most 'similar' to the source of sample E. Using this principle, the Italian team attempted to match 'unknown' texts with works of 'known' authorship, obtaining a success rate higher than 90 per cent.

[20] J. Ziv and A. Lempel, 'A Universal Algorithm for Sequential Data Compression', *IEEE Transactions on Information Theory* 23 (1977) 337–43.

[21] See note 9.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

Going a step further, the Italian researchers suggested that further quantification of the 'remoteness' between document pairs could be attempted. They proposed a formula for computing what they called the 'distance metric' ($S_{\alpha\beta}$) between any two documents α and β, calculable from the data arising from the procedure just described.[22] The 'distance' $S_{\alpha\beta}$, they reasoned, reflected the differences in patterns between documents α and β, and therefore signified an estimate of 'literary distance'.

Given a corpus of documents, $S_{\alpha\beta}$ values may be calculated for every document pair, then inscribed in a triangular matrix compactly summarizing the 'closeness' or 'remoteness' of the documents in the corpus. The authors suggested that, by feeding the matrix values into a program used in phylogenetics (the Fitch–Margoliash algorithm, included in the public-dominion 'Phylip' package), a tree diagram could be generated to provide an overall picture of the relative affinities among documents. (In subsequent reports on improved versions of their technique, the Italian team included tree diagrams of Italian literary works, in which works by the same writers tended to cluster together.)

The Italian team's 2002 article provoked mixed responses – some favorable, others critical. Teahan and Khmelev pointed out the limitations inherent in the method.[23] Other researchers tested the method in different applications, introducing procedural modifications of their own.[24] The members of the Italian group themselves have continued to refine and develop their original idea since 2002.[25]

In 2005 the Australian research group of Sabordo, Chai, Berryman, and Abbott (earlier cited) used the technique to measure the relative affinities of NT texts (their principal intention being to shed additional light on the question of the authorship of Heb).[26] A partial presentation of their published data may be seen in Figure 2. The graph shows the calculated 'distances' of various NT books relative to Luke. (The graph has been simplified to show only the calculated mean

[22] See note 19.

[23] D. Khmelev and W. J. Teahan, 'On an Application of Relative Entropy', *Physical Review Letters* 90 (2003) 089803; D. Benedetto, E. Caglioti and V. Loreto, 'Reply', *Physical Review Letters* 90 (2003) 089804.

[24] Apart from the article by Sabordo, Chai, Berryman, and Abbott (note 16), see S. C. Sahinalp, M. Tasan, J. Macker, and Z. M. Ozsoyoglu, 'Distance Based Indexing for String Proximity Search', *Proceedings of the 19th International Conference on Data Engineering* (New York: IEEE Computer Society, 2003) 125–36; T. Agata, 'Authorship Attribution by Data Compression Program', *Library and Information Science* 54 (2005) 1–18; R. Cilibrasi and P. M. B. Vitanyi, 'Clustering by Compression', *IEEE Transactions on Information Theory* 51 (2005) 1523–45; T. Roos, T. Heikkilä and P. Myllymäki, 'A Compression-Based Method for Stemmatic Analysis', *Proceedings of the 2006 European Conference on Artificial Intelligence* (Amsterdam: IOS, 2006) 805–6.

[25] See below, section (c).

[26] See note 17.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

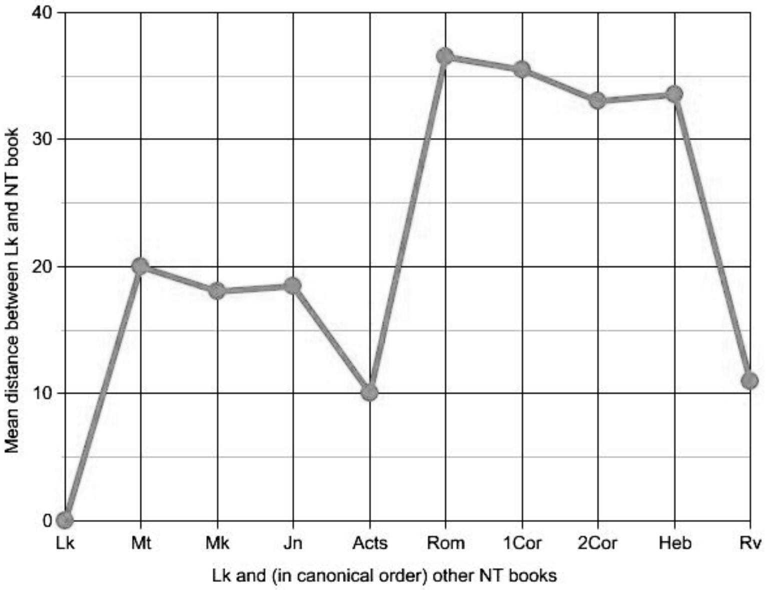

Figure 2. Plot of distances between Lk and other NT books

values, not the uncertainty ranges of the values, which the authors hold to be significant.)

Though some of the results were intriguing (closeness of Acts to Lk; relative affinity of the Synoptics, and even Jn, among themselves; closeness of some Pauline epistles among themselves), the authors noted that the method did not yield results sufficiently precise (i.e. with an acceptably small 'scattering' of experimental values) as to permit firm conclusions. They concluded that the LZ77 algorithm was less useful than PPM as a tool for authorship attribution.

We have performed an experiment along the same line, to illustrate how the original idea of the Italian and Australian research groups might be pursued to the end, to arrive, as the Italian group did in their experiment with Italian texts, at a pairwise distance matrix for NT books and a tree diagram representing the closeness or remoteness of these books among themselves. Ours, however, must be considered a simple proof-of-principle experiment, since significant results may be obtained only through repeated trials with multiple extracts randomly drawn from each NT book.[27] In one trial, long and short samples were taken from the 17 largest books of the NT, cross-combined, then compressed with the LZ77 algorithm. The resulting file size values were used to calculate the 'distances' between

[27] For more details of this proof-of-principle experiment, see www.unav.es/tdogmatica/profesores/josealviar/experiment.html.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

Mt																					
Mc	-0.11																				
Lk	-0.01	0.02																			
Jn	0.09	0.13	0.13																		
Act	0.27	0.22	0.24	0.34																	
Rm	0.17	0.18	0.20	0.30	0.35																
1Co	0.10	0.22	0.12	0.22	0.32	0.15															
2Co	0.25	0.28	0.27	0.34	0.31	0.09	0.20														
Ga	0.15	0.11	0.08	0.24	0.32	0.04	0.06	0.13													
Eph	0.30	0.30	0.20	0.32	0.37	0.08	0.15	0.08	0.16												
Ph	0.21	0.23	0.20	0.26	0.30	0.12	0.03	0.21	0.08	0.08											
Co	0.18	0.17	0.11	0.31	0.36	0.10	0.16	0.14	0.11	0.02	0.06										
1Th	0.29	0.30	0.28	0.46	0.42	0.16	0.14	0.02	0.26	0.20	0.17	0.15									
1Tm	0.20	0.18	0.17	0.27	0.34	0.04	0.17	0.13	0.14	0.08	0.11	0.04	0.22								
2Tm	0.09	0.11	0.07	0.17	0.22	-0.05	0.11	0.10	0.12	0.04	0.08	0.10	0.24	-0.03							
Hb	0.23	0.22	0.21	0.30	0.32	0.20	0.29	0.29	0.14	0.23	0.18	0.27	0.32	0.25	0.16						
Jm	0.11	0.12	-0.05	0.18	0.32	0.05	0.07	0.10	0.09	0.09	0.01	0.12	0.23	0.12	0.04	0.20					
1P	0.20	0.17	0.15	0.28	0.33	0.09	0.13	0.15	0.15	0.01	0.08	0.04	0.23	0.06	-0.01	0.18	0.07				
2P	0.34	0.26	0.25	0.47	0.43	0.30	0.33	0.27	0.33	0.28	0.32	0.29	0.40	0.25	0.13	0.31	0.30	0.17			
1Jn	0.33	0.36	0.32	0.41	0.63	0.23	0.39	0.36	0.30	0.37	0.28	0.38	0.34	0.40	0.29	0.46	0.18	0.31	0.42		
Rv	0.30	-0.02	0.26	0.42	0.51	0.42	0.41	0.43	0.41	0.51	0.47	0.34	0.52	0.38	0.24	0.40	0.29	0.31	0.55	0.55	
	Mt	Mc	Lk	Jn	Act	Rm	1Co	2Co	Ga	Eph	Ph	Co	1Th	1Tm	2Tm	Hb	Jm	1P	2P	1Jn	Rv

Figure 3. Pairwise 'distance' matrix for NT books

different pairs of NT books using the formula proposed by the Italian research group. Figure 3 summarizes the results in the form of a triangular matrix.

To visualize the results better, we followed the suggestion of the Italian team and fed the matrix values to the Phylip tree-drawing algorithm.[28] In simplified terms, this program (employed by biologists to visualize affinities among different DNA sequences) yields a tree diagram that most closely agrees with the numerical distances calculated for different pairs of sequences. The result is shown in Figure 4.

b) *The WRI (Word Recurrence Interval) Method*

We now turn our attention to another recent method, quite different from the compression-based ones. The concept of WRI originally arose from the field of 'data mining' – more specifically, from research on automated keyword extraction. It was proposed in 2002 by a Spanish research group comprised of Miguel Ortuño, Pedro Carpena, Pedro Bernaola-Galván, Esther Muñoz, and Andrés M. Somoza.[29] Its key idea is 'inter-word spacing', i.e. the 'distance' between two successive occurrences of the same word within a given text. Such 'spacing' or 'distance' is formally defined as the number of other words intervening between one

[28] It will be noted that in the distance matrix a few 'distance' values are negative (as Teahan and Khmelev predicted might occur: see 'On an Application of Relative Entropy', 089803). As Benedetto, Caglioti, and Loreto suggested (see 'Reply', 089804), we have rounded these values to 0 to avoid feeding unacceptable negative values to the Phylip tree-drawing program.

[29] M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz, and A. M. Somoza, 'Keyword Detection in Natural Languages and DNA', *Europhysics Letters* 57 (2002) 759–64.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

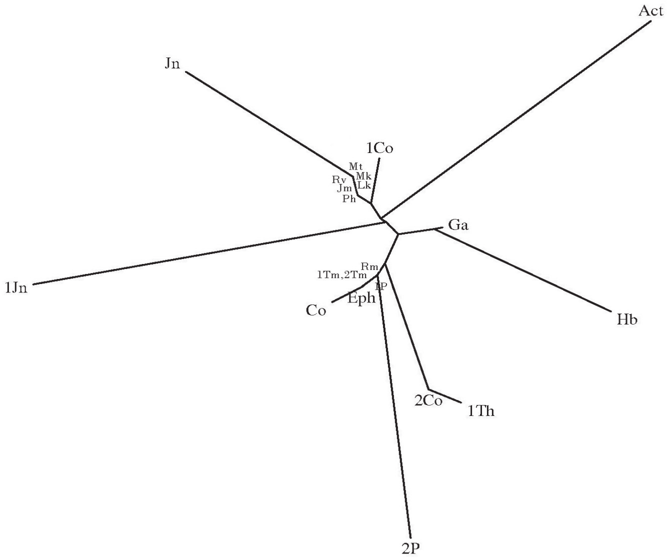

Figure 4. Affinity tree of NT books, generated by the Phylip algorithm from the pairwise distance matrix data

occurrence of a given word and the next occurrence of that same word. For any given word in a document, all its inter-word distances may be calculated (*hapax legomena* are of course excluded); afterwards, the *standard deviation* of that word's inter-word distance values may be computed. The procedure may thereafter be repeated for every distinct word in the text, to produce a table of standard deviation values (of inter-word distances), corresponding to each and every word in the document. Finally, the words may be ranked according to their computed standard deviation values, from highest to lowest.[30]

In several articles, the Australian research group headed by Abbott and Berryman (earlier mentioned) reported experiments using WRI as a criterion for

[30] According to the Spanish research team, a large standard deviation value is indicative of a non-uniform or 'clustered' occurrence of the word in the document body, whereas a small value signifies a more regular distribution of the word within the document. Experimentally, they discovered that the high-standard-deviation words (which they called 'self-attracting' words) were apt for use as keywords. For example, working with the King James version of the Bible, the researchers found that the words with the highest WRI standard deviation values were (in descending order): Jesus, Christ, Paul, disciples, Peter, Joab, faith, Saul, Absalom, John, David, king, Pharisees, Jeremiah, Gospel, Solomon, Mordecai, Esther, Joshua, Elisha. They concluded that these words are indeed representative of the KJV's contents, and that therefore the WRI method holds promise for automated selection of keywords.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

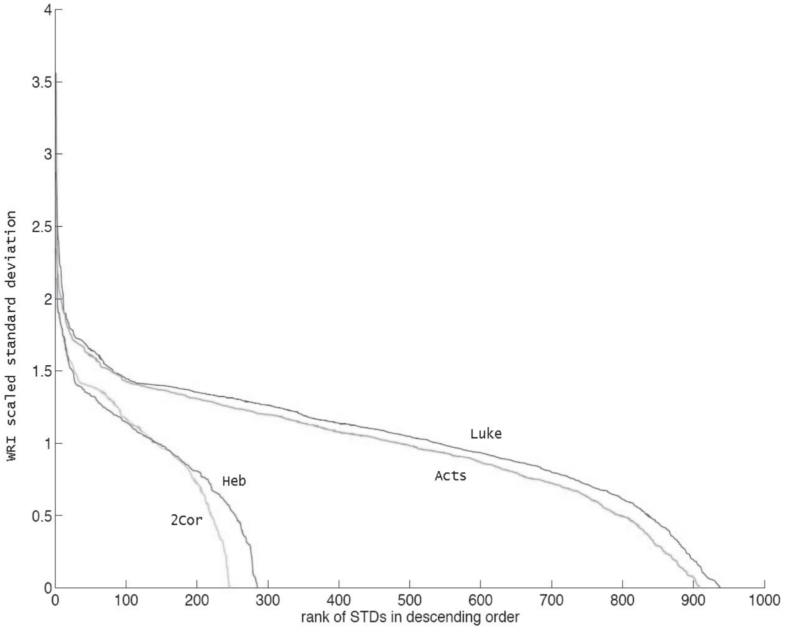

Figure 5. Graph of standard deviation vs. rank

discriminating English literary works.[31] When they drew graphs of standard deviation values against word rank, they observed that the resulting curves for texts by different authors were distinct. They then tested the method on NT books. In Figure 5 we offer a simplified version of their published results for four NT books. The authors themselves have pointed to a similarity between the curves of Lk and Acts, and the dissimilarity of their curves with those of 2 Cor and Heb.

To quantify further the differences or similarities among the curves, the Australian team tried both chi-squared testing and curve fitting via linear regression. The latter method allowed them to calculate the slope value for each book, which they suggested could vary from author to author; Figure 6 shows a simplified view of their results. Once more, a similarity between Lk and Acts may be observed – this time expressed as identical slope values (−0.0015), as well as the difference of these slope values from those of 2 Cor and Heb.

Further experiments by the Australian group on English fiction texts, however, yielded ambiguous results and led them to adopt a more cautious attitude regard-

[31] M. J. Berryman, A. Allison, and D. Abbott, 'Signal Processing and Statistical Methods in Analysis of Text and DNA', *Biomedical Applications of Micro- and Nanoengineering. Proceedings of SPIE* 4937 (2002) 231–40; 'Statistical Techniques for Text Classification Based on Word Recurrence Intervals', *Fluctuations and Noise Letters* 3 (2003) L1–L10; M. Sabordo, S. Y. Chai, M. J. Berryman, and D. Abbott, 'Who Wrote the Letter to the Hebrews?', 513–24.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

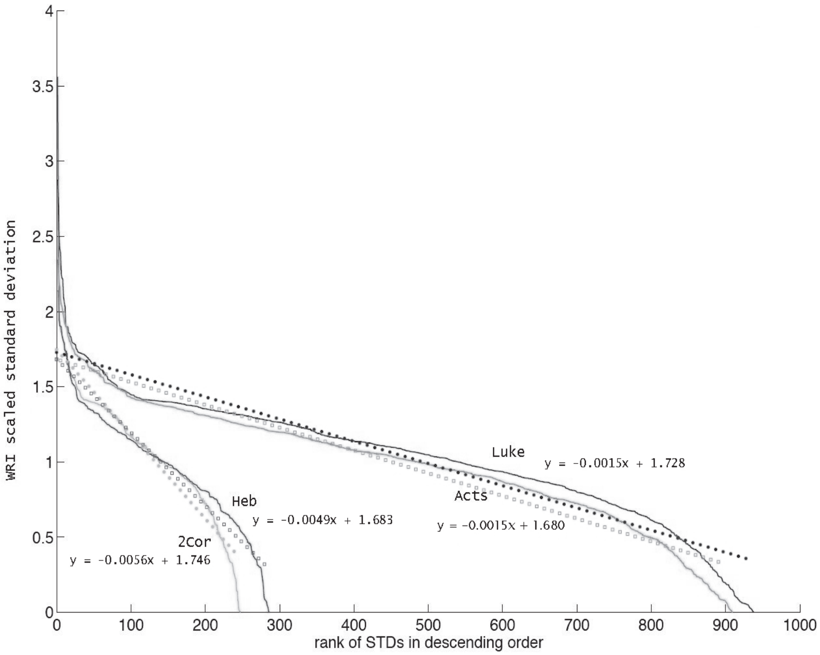

Figure 6. Linear fitting of curves of standard deviation vs. rank

ing the capability of the method (at least 'in its current form') to characterize an author's style.[32]

c) *Other Techniques: A Résumé*

Our survey has focused on two approaches utilized by several research teams around the world. For completeness' sake we now make a brief mention of other computer-assisted methods that have been used recently in textual comparison experiments. The techniques are not yet as widely employed as the ones earlier cited, but appear to hold promise also.

Several of the research groups mentioned above as having worked with the PPM or LZ77 algorithm have tested other text-compression algorithms, particularly bzip2 and improved variants of PPM. Some have gone further, devising even more refined pattern-detecting schemes. The tendency now seems to be towards schemes based on the detection and comparison of 'typical' character strings in

[32] T. J. Putnins, D. J. Signoriello, S. Jain, M. J. Berryman, and D. Abbott, 'Advanced Text Authorship Detection Methods and Their Application to Biblical Texts', *Complex Systems. Proceedings of SPIE* 6039 (2005) 163–75. It seems that the plots vary somewhat with document length, so the issue must be investigated before the WRI technique can be reliably employed for authorship attribution.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

documents as suggested, for instance, by Forsyth and Holmes. In this line, Khmelev and Teahan have proposed the R-measure. The value is derived from simple counts taken of character strings that are common to any two given documents.[33] Teahan, after Khmelev's death in 2004, went on (with the assistance of graduate students) to develop a related value, called the C-measure, for documentary comparison. This value is derived from counts of character strings of a *fixed length* that are common to the documents being compared.[34] The method achieves improved results over the R-measure technique.

For their part, the Italian researchers have been working on the creation of intermediate, 'artificial texts', based on typical sequences detected in the original sources by the LZ77 compression algorithm.[35] Such 'artificial texts' are in principle representative of their sources, and are the ones that may be compared to obtain an estimate of the 'distances' of the original source documents relative to each other. The procedure circumvents several limitations inherent in the LZ77 compression technique, and has been reported by the authors to yield better experimental results.

4 A First Appraisal

a) *Commentary on the Results*

The results obtained thus far with the new methods are hardly startling. Abbott and Berryman's experiments – perhaps the most comprehensive to date done on NT texts – show that compression algorithms which detect typical character sequences situate the four Gospels near each other (with Acts and Rv close by), and books belonging to the Pauline corpus farther away. (The measured closeness of Rv to the Gospels could be due to the presence of apocalyptic passages in the Gospels; this suggests a need to pay careful attention in future trials to the presence of various genres within a NT book). The detected affinity between Acts and Lk concords with findings of other stylometrists like Kenny and with the traditional attribution of Acts to the author of the Lucan gospel. The closeness of the great Pauline letters Rom and 1–2 Cor, as well as their distance from Eph, is in

[33] D. V. Khmelev and W. J. Teahan, 'A Repetition Based Measure for Verification of Text Collections and for Text Categorisation', *Proceedings of the 26th Annual ACM-SIGIR Conference on Research and Development in Information Retrieval* (New York: ACM, 2003) 104–10.

[34] D. S. Hunnisett and W. J. Teahan, 'Context-based Methods for Text Categorisation', *Proceedings of the 27th Annual ACM-SIGIR Conference on Research and Development in Information Retrieval* (New York: ACM, 2004) 578–9.

[35] A. Baronchelli, E. Caglioti, V. Loreto, and E. Pizzi, 'Dictionary-Based Methods for Information Extraction', *Physica A* 342 (2004) 294–300; A. Baronchelli, D. Benedetto, E. Caglioti, and V. Loreto, 'Artificial Sequences and Complexity Measures', *Journal of Statistical Mechanics* (2005) P04002.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

agreement with the consensus among Biblical scholars; not so their distance from Gal. This latter result is ambiguous – it may be taken as an indication of the compression technique's limitations, or as a suggestion that Eph is not really too distant from other epistles of indubitable Pauline authorship. As for the WRI trials, based on an algorithm capable of measuring the clumping or scattering of determined words in a text, the results lend further credence to the idea of a common author for Lk and Acts.

Such results, though suggestive, must be considered as provisional and still subject to discussion. Although mathematical support for the techniques is available, further experimentation needs to be done to gain a more complete idea of the methods' effectiveness in distinguishing literary 'styles'. (In fact, as already mentioned, their proponents are currently working on further refinements.) The task that these researchers face is no mean one. In a critical article regarding the current state of stylometry[36] Joseph Rudman cites, among the challenges ('problems', he calls them) confronting current-day researchers, the need for serious experimental design[37] – e.g. obtaining the best available version of the text to be studied, mastering the history of difficulties of the different versions of a document, having a firm grasp of the mathematical and statistical techniques to be applied (including the knowledge of their assumptions and their limitations). This point of Rudman's critique has some applicability to the procedures reported above: the choice of the particular version of the NT used in the trials is a matter for discussion, as are the number of trials required and the specific conditions that need to be met for reliable conclusions to be reached.[38]

The results obtained thus far with the new methods are, as we have said, far from revolutionary; but not devoid of interest. The mere fact that they generally point in the same direction as conclusions from other studies may be taken as an

[36] J. Rudman, 'The State of Authorship Attribution Studies: Some Problems and Solutions', *Computers and the Humanities* 31 (1998) 351–65. Similar concerns are expressed by A. Dean Forbes, 'Statistical Research on the Bible', 204, and É. Évrard, 'Sur quelques précautions en statistique littéraire', *Bible and Computer: Proceedings of the AIBI-6 Conference* (Leiden/Boston: Brill, 2002) 583–91.

[37] Rudman, 'The State of Authorship Attribution Studies', 359.

[38] As a concrete example, A. Puglisi, D. Benedetto, E. Caglioti, V. Loreto, and A. Vulpiani, 'Data Compression and Learning in Time Sequences Analysis', *Physica D* 180 (2003) 92–107, showed that if one wishes to obtain reliable results using the original LZ77 technique described by Benedetto, Loreto, and Caglioti in their 2002 article (see note 9), it is necessary to control carefully the size of the short extracts 'e' relative to the long extracts 'E'. The measurement of the so-called 'cross-over length', necessary for a sound experimental design, calls for additional trials. A further idea of the abundant work required may be drawn from the report by Kukushkina, Polikarpov, and Khmelev (see note 8) – they estimated that comprehensive testing of 16 different compression algorithms for author attribution work required 'about three weeks of non-stop computing'.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

indication of the potential of such methods for text characterization. It is true that their approach to literary text is simplistic, as they view text as a mere string of characters; however, it is precisely this reductionist approach which renders the text tractable to powerful pattern-detecting computer programs. In this sense, the techniques could be a valuable complement to the other methods of style analysis.

The novel techniques offer some advantages. First, they do not require much 'preprocessing' of texts, and so are more straightforward to use compared to other stylometric methods (such as the discriminant methods).[39] On a deeper level, the 'blindness' of the new methods renders them immune to subjective or ideological biases on issues like authorship, canonicity, etc.[40] Such a 'neutral' starting-point could be useful for objectively counterchecking proposals like the old multiple-source hypothesis of Wellhausen for the Pentateuch, or the Q-source hypothesis for the Gospels. This same trait, admittedly, may also be counted as a weakness, since the new methods cannot possibly take into account the historical, cultural, and social context of biblical texts. They are clearly non-holistic methods.

One further interesting characteristic of the new techniques is that they are applicable to any linear sequence of characters, and may thus be employed for Hebrew and Greek texts. Conceivably, these techniques could be of use in measuring differences among sections within one same book (as Radday and Shore attempted, using other methods, with the book of Genesis[41]): e.g. Gen 1.1–2.4a vs. Gen 2.4b–3.24; Gen 1–11 vs. Gen 12–50; Jn 7.53–8.11 or Jn 21 vs. the rest of Jn. They might also be helpful in comparing presumed Q-originated sections of the Synoptics with other sections; or, for that matter, canonical with contemporaneous non-canonical works (apocalypses, gospels, etc.: e.g. Daniel vs. apocryphal Danielic literature).

b) *Future Perspectives*

But then again, how useful are computer-based techniques for analysing a text so complex as the Bible? In the early 1960s, when computer power was

[39] 'Compression-based text classification methods are easy to apply requiring virtually no pre-processing of the data'. Y. Marton, N. Wu, and L. Hellerstein, 'On Compression-based Text Classification', *Advances in Information Retrieval: Proceedings of the 27th European Conference on Information Retrieval Research* (Lecture Notes in Computer Science 3408; Berlin: Springer, 2005) 300–14, esp. 300.

[40] 'Information theoretic measures have the advantage of making very few assumptions on the models which are considered to have generated the sequences'. D. P. Coutinho and M. A. T. Figueiredo, 'Information Theoretic Text Classification Using the Ziv-Merhav Method', *Pattern Recognition and Image Analysis: Proceedings of the 2nd Iberian Conference on Pattern Recognition and Image Analysis* (Lecture Notes in Computer Science 3523; Berlin: Springer, 2005) 355–62, esp. 355.

[41] See n. 4.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

increasingly being brought to bear upon biblical research, T. M. Knox famously remarked: 'The spirit moveth where it listeth and is not to be reduced to the numerical terms with which alone a computer can cope'.[42] Can a 'blind program on a mindless machine' really detect the subtle patterns that a human author consciously or unconsciously imprints on his or her work? Is it not, indeed, a contradiction to seek to measure that elusive trait that we call 'style'; to attempt to *quantify* what is essentially a *quality*?

The results obtained to date in the field of computational stylistics,[43] and those reported in this article, would seem to contradict Knox's dismissal of the usefulness of machine-based approaches to Biblical texts. While it is true that computational approaches grasp only a limited number of textual peculiarities, they do add richness to the perceived texture of human language. Combined with more traditional methods, non-traditional techniques contribute to a more holistic picture of the text. Modern scholars, in fact, tend to use more than one approach to literary texts, aware that every approach has its limitations. Thus, Holmes assures: 'Stylometry presents no threat to traditional scholarship. In the context of authorship attribution, stylometric evidence must be weighed in the balance along with that provided by more conventional studies'.[44] In the same vein, a 1994 Vatican document on Biblical exegesis acknowledges: 'No scientific method for the study of the Bible is fully adequate to comprehend the biblical texts in all their richness . . . It is not surprising, then, that at the present time other methods and approaches are proposed . . . which serve to explore more profoundly other aspects worthy of attention'.[45]

It will be noted that the ideas and tools here reviewed have a 'foreign' provenance – fields like information theory, data compression and mining, and bioinformatics. This is a sign of the increasing 'globalization' of scientific research: methods are continually being elaborated in disparate areas of human knowledge, with potential application to many other fields. In the concrete filed of textual analysis, as we have seen, instruments are now being proffered by information theorists, communications engineers, and DNA researchers. This trend towards interdisciplinarity suggests that linguists and biblists may have to pay greater attention to advances in research areas ostensibly distant from their own.

Therein lies a challenge – for who can keep attuned to fields so numerous and diverse? Rudman, in his article on the current state of stylometry, complains of a

[42] Cited by Kenny, *A Stylometric Study of the New Testament*, 116.

[43] See Craig's defence of this field in his article 'Stylistic Analysis and Authorship Studies', esp. 287–8.

[44] Holmes, 'Authorship Attribution', 104.

[45] Pontifical Biblical Commission, *The Interpretation of the Bible in the Church* (15 April 1993), I. B.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press

general 'lack of competent and complete bibliographical research' and 'little investigative memory' in current stylometric investigation;[46] in other words, researchers are not always fully aware of what others have already done or are currently doing. This problem is only becoming deeper, for – as we have pointed out – concepts and tools are now being brought to bear upon textual problems from fields farther and farther removed from traditional sectors of stylometry, and the results reported in specialized publications little accessed by biblical or linguistics scholars.[47] (It will be observed that the concepts for the techniques described in this article were first reported in communications and electrical engineering journals, rather than in biblical, linguistic, or computer journals.)

Kenny, in his comparative analysis of Aristotle's Eudemian and Nicomachean Ethics published in 1978 (which partially relied on stylometric techniques), affirmed that 'to be fully qualified to undertake such a task a man must be a professional philosopher, classicist, and statistician'.[48] We might today add: '. . . and an electrical engineer, or perhaps a geneticist'. Perhaps the most realistic solution would consist in more assiduous collaborative work among experts from different fields, in order to pool know-how and resources.[49] Will linguists, biblists, computer scientists, mathematicians, communications engineers, and biologists eventually come to form a tightly knit 'expertise network'? Or will, rather, the trend towards specialization hinder attempts at concerted effort? Only time will tell.

[46] Rudman, 'The State of Authorship Attribution Studies', 354.

[47] For instance, Rudman ('The State of Authorship Attribution Studies', p. 353) reports that 'a quick scan of my working bibliography shows that non-traditional authorship attribution studies have been published in well over 76 journals representing 11 major fields'. Another veteran researcher in Bible and computers, R. F. Poswick, alludes in an online article to the 'overwhelming' bibliography on the Bible and information technology, as well as the need to classify and critically evaluate all this bibliographical information: 'The Bible in the Civilization of the Electronic Writing: An Evaluation (1985–2004)', http:/www.cibmaredsous.be/cib5015J.htm, 1.2.1 (23 July 2004).

[48] Kenny, *The Aristotelian Ethics. A Study of the Relationship between the Eudemian and Nichomachean Ethics of Aristotle* (Oxford: Clarendon, 1978) v.

[49] Craig hints at this possibility in his article 'Stylistic Analysis and Authorship Studies', 281–2. Observing the difficulty for any one scholar to be expert in two or more fields, he points to a trend towards collaborative work among scholars from the humanities, computing, and statistical areas.

https://doi.org/10.1017/S0028688508000088 Published online by Cambridge University Press