Ecdotica

5
(2008)

Alma Mater Studiorum. Università di Bologna
Dipartimento di Italianistica

Centro para la Edición
de los Clásicos Españoles

Carocci editore

Alma Mater Studiorum. Università di Bologna,
Dipartimento di Italianistica,
Via Zamboni 32, 40126 Bologna
ecdotica.dipital@unibo.it

Centro para la Edición de los Clásicos Españoles
cece@cece.edu.es
www.cece.edu.es

Con il contributo straordinario dell'Ateneo di Bologna
e con il contributo della Fondazione Cassa di Risparmio in Bologna

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

CEE
CENTRO PARA LA EDICIÓN DE LOS
CLÁSICOS ESPAÑOLES

FONDAZIONE
CASSA DI RISPARMIO
IN BOLOGNA

Carocci editore,
Via Sardegna 50, 00187 Roma
tel. 06.42818417, fax 06.42747931

INDICE

Saggi

Foro

Questioni

Rassegne

Cronaca

A SURVEY OF NON-TRADITIONAL AUTHORSHIP ATTRIBUTION STUDIES

ANTONIO MIRANDA-GARCÍA
AND JAVIER CALLE-MARTÍN

1. Introduction

In everyday life one often encounters insurmountable difficulties to tell twin brothers apart, especially when no salient physical features lead to their immediate identification. Notwithstanding their likely identical resemblance at first sight, human beings characterize for their uniqueness and individuality, features which are more consistently observed in the inner organisation of knowledge and, more importantly, in the actual use of the speaker's language, written production in particular. It is this writing singularity what constitutes an author's fingerprint, a topic which has traditionally been the object of authorship attribution studies. Assuming «that every author has a verifiably unique style» (Rudman 2000: 170), the most immediate aim of these approaches is to find the stylistic differences allowing to relate works and authors with accuracy.

Authorship attribution is taken to be as old as the hills since it runs parallel with the production of the first literary texts in ancient times as an attempt to find a hand behind some classical pieces, which were more often than not anonymous. Later, the 20th century witnessed a proliferation of this kind of studies which, using a traditional approach, analyse the internal/external dimension of a text within the fields of Stylistics and Literary Criticism. The advent of computers in the Humanities and the increasing availability of machine-readable texts have largely influenced the development of non-traditional authorship attribution studies. Nowadays, this approach is found to have a wide vari-

The present research has been funded by the Spanish Ministry of Science and Innovation (grant number FFI2008-02336). This grant is hereby gratefully acknowledged.

ety of applications, not only as a means to seek the likely author of a disputed piece, but also to ascertain at the court-rooms of many countries the authorship of menacing notices or electronic messages vindicating a terrorist action.

This paper therefore surveys non-traditional attribution studies in the last decades to provide a state-of-the-art which sheds light on the origin, development and main goals of the discipline. The present survey stems from the assumption that authorship attribution is a method rather than an end and, accordingly, the objective of this paper is twofold: *a*) to review the most important techniques used in the field; and *b*) to evaluate their assets and shortcomings, if any. In light of this, our paper has been organised into five different sections: the first deals with the origins and development of Stylometry; the second characterises this same discipline to highlight its features; the third describes lexical authorship attribution studies, Yule's *K*, Zipf *Z* and *principal component analysis* among them. The fourth, in turn, briefly discusses the contribution of other recent approaches. Finally, our conclusions close the paper.

2. Origins and evolution of Stylometry

The term *authorship* is defined in the *OED* as (1) "the occupation or career as a writer of books; (2) "the dignity or personality of an author; and (3) "the literary origin or origination (of a writing)" (Simpson and Weiner 1989). Based on (1) and (3), different types of authorship can be distinguished (i.e. collaborative, individual, precursory, executive, declarative, revisionary, etc.) though, for simplicity, that of individual agency is taken «as a form of human work» (Love 2002: 32-50).

The responsibility of establishing authorship has been changing hands with time. In the ancient world this task was undertaken by editors such as the Alexandrian Zenodotus and Aristarchus, by scholar librarians like Aristophanes of Byzantium, and by authors such as Plutarch and Marcus Terentius Varro, who were guided «by an intuitive recognition of the characteristic features of his manner and diction» (Love 2002: 16). Likewise, authorial attribution was practised by the compilers of the Jewish and Christian Bibles, many of which books were named after a putative author, although some are still the object of authorship investigation (i.e. the *Pauline Epistles*), and especially by the venerated scholars Saint Jerome, one of the Fathers of the Church, and the Car-

thaginian Saint Augustine. Both investigated the philosophical nature of authorship and to the former are due a set of valid criteria to solve authorial problems as stated by Foucault (1969: 204).

For its influence on the external/internal evidence, a few lines must be devoted to the consequences of the introduction of the printing press in the 15th century. There is no doubt that this innovation largely contributed to the standardisation of writing. However, it is not less certain that the printed versions lost some valuable *external* characteristics of the holograph, being therefore of great help to date or to ascertain the provenance of a given manuscript (i.e. script, bookbinding, paper watermarks, etc.). From a linguistic perspective, manuscripts usually provide with some helpful information which is often lost as a result of the modern editorial conventions. Among others, the palaeography of the text may be taken as a reliable clue for authorship attribution not only in terms of the particular script used by the scribe, but also in terms of the inventory of marks of punctuation along with other practices such as word separation, word division, etc.

After a nearly complete absence of a new authorship tradition in the Middle Ages comes a rich period characterized by humanist scholarship, Erasmus of Rotterdam (1466?-1536) being the most salient figure along with Lorenzo Valla (1406-1457), Spinoza (1632-1677) and Richard Simon (1638-1712). The argument for claiming the authorship of some works ranged from the scholar's intuition or simple notes about their dating, to the orthodoxy or unorthodoxy of the doctrine being dealt with, the writer's style also included.

This approach was used with secular texts, not only classical but also more contemporaneous pieces like John Donne's or Andrew Marvel's poetry, Shakespeare's plays,[1] or Milton's *De Doctrina Christiana.*[2] The disputed authorship of some of these items remains still alive in specialised journals, as is the case of Addison's and Steele's periodical essays published under an eponymous *Mr Spectator* (c. 1700), the pseudonymous journalism of the letters signed "Junius" in the *Public Advertiser* (1769-1772),[3] or the well-known 12 disputed papers which were penned as *Publius* in various American newspapers (1787-

[1] Keller has recently provided new evidence on the authorship of *Titus Andronicus* from a historical perspective (2003: 105-118).

[2] Its provenance has also been investigated from a non-traditional perspective (Tweedie, Holmes and Corns 1998: 77-87).

[3] For a comprehensive analysis of these letters from the point of view of authorship attribution, see Ellegård (1962).

1788). The authorship assignment with this traditional methodology has proven to be accurate in many a case as, besides the practitioners' vast erudition, the style of the author was conclusive for a reliable attribution.

Therefore, the study of authorship attribution fell within the scope of Stylistics until the end of the 19th century when the first tentative steps in the use of quantitative data are observed as a way to accept or refute the authorial attribution of doubtful works by means of the traditional approach. The steady employment of Statistics for this purpose resulted in the advent of Stylometry/Computational Stylistics. Statisticians were henceforth enrolled in research groups that devoted their time to find «quantifiable features used as authorial discriminators» (Holmes 1998: 111) as well as to design theoretical proposals and/or experimental tests to solve the authorial problems with a computer-based scientific methodology.[4]

The scope of authorship attribution has been widened with the advent of Forensic Stylometry, aiming at the analysis of the linguistic evidence of a case come to trial, from anonymous letters and guilty pleadings to the claiming of responsibility for a terrorist attack. In light of all this, it is a fact that the burst of electronic messaging for criminal purposes (i.e. emails, sms, etc.) will undoubtedly contribute to the development of authorship attribution as the new challenge of the discipline in the 21st century, always pursuing new techniques and procedures granting a more accurate relation between text and author.

3. Stylometric features

Stylometry seeks how to relate a work and its anonymous or disputed author accurately, the reasons for the anonymity ranging from the author's intentions of remaining unknown in hostile political or religious environments to other spurious reasons such as gaining outstanding notoriety when his/her style resembles that of a reputed author. Theoretically speaking, it is elsewhere assumed that every author signals his/her works with an authorial wordprint which can distinguish him/her from other authors' works like the fingerprints, the iris or the

[4] «[The] growing power of the computer and the ready availability of machine-readable texts» (Holmes 1998: 111) largely contributed to the development of the discipline insofar as the speed and the accuracy of the calculations improved faster than it could be hardly imagined before.

ADN serve to identify a person successfully. Contradictory as the rationale may seem, authorial assignment stems from the assumption that the unconscious features of an author's style are somewhat permanent, whilst the chronological clustering of works is founded on the hypothesis that the author's stylistic features develop rectilinearly in the course of the author's lifetime (Can and Patton 2004: 61-82).[5]

Still, no agreement has been reached among specialists as to the appropriate methodology, the technique and the reliability of the results. Accordingly, Lancashire considers that authorship attribution cannot be established without reliable authorial parameters, which must be «habitual, difficult for the authors to observe, to edit, and to cut, and unambiguous», particularly, those «of which the author is not conscious» (Lancashire 1998: 299).

In traditional approaches, literary critics pinpoint the stylistic features of a piece to relate it with its author by considering both micro- and macro-textual markers. These are subsequently compared with those occurring in the works of the same author or of different authors to assess their likeliness. At a micro-textual level, the word has been the most recurrent marker, particularly on account of its easy handling, whereas at macro-textual level perhaps the punctuation and/or the text organisation, either from a syntactic, pragmatic or prosodic perspective, have been the most recurring factors. In the particular case of the word, its length and letter composition and arrangement, doublets, synonyms, antonyms, rare words, *hapax legomena*, etc. have been used time and again for attribution purposes.

Likewise, the practitioners of non-traditional approaches have done their best to characterise the most reliable stylistic feature which may be safely considered as the corner-stone for attribution. It is therefore great the variety of quantifiable stylistic features claimed as the most reliable *authorial discriminators* to arrive at the most conclusive attribution. These features can be grossly classified into «lexical, syntactic and semantic» (Holmes 1998: 111). It is beyond any doubt that the word has been the most recurrent feature in stylometric studies because,

[5] The functions of authorship attribution are, among others, the following: *a*) to select the most plausible author of a piece from a set a candidates in view of their whole or partial work; *b*) to distinguish which texts are written by the same author and which ones are written by different hands; *c*) to refute the accepted authorship in view of the stylistic evidences found; *d*) to sort out any wrongly attributed work from the canon of an author; *e*) to rank the works of an author chronologically; *f*) to distinguish an authentic work from a pastiche; etc.

according to Tallentire, «no potential parameter of style below or above that of the word is equally effective in establishing objective comparison between authors and their common linguistic heritage» (cited in Holmes 1998: 111).

Taking for granted the suitability of the word for these purposes, there is no agreement as to the number and type of words to employ: i.e. all the words or *tokens* (Labbé 2007; Miranda-García, Calle-Martín and Marqués-Aguado 2008: 210-225), only the different ones or *word-types*, content words, function words (Mosteller and Wallace 1984), the most common words (Burrows 2002: 267-287), the least common words, the *hapax legomena* (Honoré 1979) or *dislegomena* (Sichel 1975), etc., an endless list not deprived of controversy as every analyst claims that the use of one or the other yields the highest accuracy.

In addition to lexical approaches, the syntactic parsing or the semantic tagging of a text can also be used as the input for authorial attribution (Stamatos, Fakotakis and Kokkinakis 2001: 193-214), though they require the manipulation by the analyst, a fact seriously criticized by Rudman (1998: 351-365) insofar as a certain subjectivity is introduced artificially into the texts. Add to them the adoption of other stylistic features such as the metric pattern, the most common vowel, etc.

The variety of statistical techniques and methodologies used is also great insofar as a mere glance at the literature reveals an evolution in search for the most accurate results and an optimization of the procedures (for example, the continuous caveat to avoid text-length dependency). A brief survey of the most important landmarks in Stylometry is accordingly provided in the next section.

4. Three landmarks in lexical authorship attribution studies

Three important stylometric hits must be highlighted as contributing to authorship attribution studies, which are chronologically as follows. The first is associated with the assessment of lexical richness by Yule and Zipf. The second has to do with Mosteller and Wallace's masterpiece on the *Federalist Papers*. The third stems from the meritorious contribution of Burrows with the *Delta* methodology.

4.1. Lexical richness: Yule and Zipf

The evaluation of the *lexical richness* (*LR*) (or *vocabulary richness*, *VR*) of texts has been a common topic in the field of Quantitative Linguistics and in authorship attribution studies, scholars assuming that *LR* constitutes a salient authorial feature by which texts and their authors can be related successfully. The magnitude of this feature, however, is a moot point from a scholarly perspective, as shown in Hoover's experiment carried out by Hoover (2003: 153).

LR can be grossly associated to the vocabulary size of a text (the number of different word types, *V*), which is expected to vary with text length (the number of word tokens, *N*). Accordingly, it is evident that the grammar-context-related sentence «*That "that" that that man said was wrong*», $N = 8$ (word tokens) and $V = 5$ (word types), is less rich than Goneril's words «*By day and night, he wrongs me, every hour*» (Shak, *Lr*, I, 3), $N = 9$ and $V = 9$, on account of the greater value of *V* and the similar value of *N*. This evidence can be confirmed using these same data in *N/V*, or its inverse *V/N*, which yields the results of 625 in the former, and one in the latter.

Unfortunately, the results obtained from Mendenhall's *type token ratio* (*V/N*) and from Baker's *mean word frequency* (*N/V*) are not reliable for attribution purposes in terms of their text-length dependency insofar as *V* increases with *N*. In other words, «the longer the text, the more slowly the vocabulary grows, and hence the less rich the vocabulary becomes» (Hoover 2003: 157). In view of this shortcoming, Mendenhall's original rate has been successively redefined into new formulae to characterise *LR* irrespective of N, even though none has been eventually proven as being text-length independent in itself (Tweedie and Baayen 1998: 330).

Other scholars propose an approach in terms of the elements of the *frequency spectrum* or *lexical profile* of the text, which is accomplished by registering in each row of the leftmost column of an array the number of times that one or more word-types occur in a text. Likewise, the right column holds the number of word-types occurring so many times whilst other columns would contain an accumulative study of tokens, word-types, as well as their percentages with respect to *N*. Accordingly, Table 1 below shows that there is one word-type occurring 528 times (*the most common word*), two words occurring 315 times, 489 words occurring exactly twice (*hapax dislegomena*) and 654 words occurring just once (*hapax legomena).*

TABLE 1
Lexical profile of a text

TIMES	WORD-TYPES	
528	1	← *the most common word (MCW)*
492	1	
315	2	
...	...	
4	187	
3	234	
2	489	← *hapax dislegomena*
1	654	← *hapax legomena*

In light of the distribution of word-types within the lexical profile, Yule (1944) presented the first *Characteristic Constant* (henceforth *K*) in lexical statistics assuming that «the occurrence of a given word is based on chance and can be modelled by a Poisson distribution» (Holmes 1998: 112; see also Tweedie and Baayen 1998: 330; Miranda-García and Calle-Martín 2005b: 287-294). Yule's *K*, an inverse measure of *LR* given that a high *K* value implies a small vocabulary, actually measures the rate at which words are repeated. Accordingly, «vocabulary concentration (a small, focused vocabulary) rather than vocabulary richness (a large, varied vocabulary) is deemed a mark of high quality» (Yule 1944: 122, 131), though «for fiction, a richer vocabulary is likely to be more highly valued» (Hoover 2003: 174).

Similarly, by studying some specific elements of the lexical profile, Sichel (1975: 542-547) noticed that the ratio of *dislegomena* to *N* is roughly constant across a wide range of sample sizes, and Honoré (1979: 172-177) discovered that the ratio of *hapax legomena* to *N* is constant with respect to the logarithm of the text size.

A new research line was built to evaluate *LR* with a limited number of formal parameters of probabilistic models for word frequency distributions. From the different models available, perhaps the most efficient is Orlov's generalized Zipf model (1983), where *V* is a function of one free parameter *Z*, which expresses the text length at which Zipf's law holds. Zipf *Z* can be considered a measure of *LR* inasmuch as an increase of *Z* leads to an increase of *V* (Tweedie and Baayen 1998: 331).

The statistics for lexical richness are classified into two classes. The first comprises those, Yule's K included, which are appropriate to measure the rate of repetition, thereby constituting inverse measures of vocabulary richness. The second agglutinates those which measure vocabulary richness more directly, Zipf's *Z* among them (Hoover 2003: 156). In this fashion, Tweedie and Baayen emphasise that the employment of Yule's *K* and Zipf's *Z* will bring forth a surprising amount of authorial style as shown in the cluster of some works, though a careful use is recommended on account of the textual variability (Tweedie and Baayen 1998: 350). It is, however, a fact that *VR* varies greatly within a single text or in the texts by the same author even when equal-sized texts are surveyed.[6]

In opposition to Tweedie and Baayen's considerations, Hoover (2003: 158) state that *VR* «is a much less useful and a much more dangerous indicator of authorship and marker of style» after replicating their experiment (first with the same texts used by Tweedie and Baayen, then with excerpts of the first 24,000 words of each of their texts, and with other texts), and by applying a set of 17 constants: Yule's *K* and Zipf's *Z* among them. The results lead Hoover to conclude that the grouping of texts becomes more accurate only when *K* and *Z* are used instead of the 17 constants, though admitting that *K* and *Z* are not so conclusive as to grant universal reliability in the following terms: «these measures of vocabulary richness capture some aspects of authorial style, but just as clearly, they fail to separate large numbers of texts by different authors or to cluster all sections of single texts together» (Hoover 2003: 167).[7]

4.2. The *Federalist Papers*: Mosteller and Wallace

The case-study of the *Federalist Papers* has become ground-breaking in literary detection, being subsequently used as the test tube for new stylometric techniques. Mosteller and Wallace, two American statisticians, decided to use statistical methods (in particular, a 200-year-old mathe-

[6] In this vein, Hoover argues that «If the vocabularies of sections of different texts by a single author can vary by more than 1500 words while the vocabularies of sections of texts by eleven different authors can vary by fewer than 70 words, there seems little hope that vocabulary richness alone can be safely used to determine authorship, or to illuminate an author's style» (2003: 168).

[7] Yule's *K* and Zipf's *Z* have been extensively used in a number of experiments of authorship attribution (i.e. Mosteller and Wallace 1984; Smith and Kelly 2002: 411-430; Somers and Tweedie 2003: 407-429; Miranda-García and Calle-Martín 2007: 49-66; Miranda-García, Calle-Martín and Marqués-Aguado 2008: 210-225, etc.).

matical theorem) to solve the problem of authorship of the disputed *Federalist* Papers, which is for them a 175-year-old historical problem «more to advance statistics than history» (Mosteller and Wallace 1984: IX). The *Federalist Papers* were written by Alexander Hamilton, John Jay and James Maddison, and published under the pseudonym of Publius in 1787-1788 persuading the citizens of the State of New York to ratify the Constitution. The authorship of 12 of them is attributed either to Hamilton or to Maddison, thereby termed the *Disputed Papers*. As a matter of fact, their attribution has been controversial among History scholars because the internal evidence obtained from «some positions expressed» in the propagandistic papers «were not held at later times» (Mosteller and Wallace 1984: 3) and, in addition, their oratorical style, which is formal and complex, does not greatly differ as to tell them safely apart.

The investigation was initiated by Williams and Mosteller who, influenced by Yule's (1938) and Williams' (1939) work on word counts and sentence length, counted all the words in the sentences of *The Federalist* to obtain a non-discriminating result, both in the *average sentence length* and in the *average standard deviation* (Mosteller and Wallace 1984: 6-7).[8]

In view of these results, Williams and Mosteller proceeded to the counting of four other variables (percentage of nouns, adjectives, one- and two-letter words, and *the*), whose results were subsequently the input for a statistic which yielded high scores to Hamilton's papers and low to Madison's. They also calculated the *linear discriminant function* with the weighted sum of the rates of the four variables in Hamilton, Madison and the *Disputed Papers*. The evidence pointed to Madison as the author of the 12 papers, but the odds were not valid to produce a reliable assignment for each paper.

Mosteller and Wallace continued this same line of research by rating the occurrence of some marker words (*while*/*whilst*, *enough* and *upon*), which served to distinguish between Hamilton and Madison. The results showed that these four words were highly discriminatory, *upon* being the most, thus leading to the fact that the *Disputed Papers* were clearly Madisonian. Later, the rates of *by*, *of* and *to* were tested, finding that they were not so conclusive as to distinguish individual papers from one another. Therefore, they (1984: 16) decided to base their study on individual words and their distributions to obtain a likely author, a word being a composite of all the words of the same spelling, irrespective of

[8] Mannion and Dixon also investigated the case of *Oliver Goldsmith* by using sentence length (2004: 497-508).

capitalization, lemma and/or word-class for the difficulties involved. They acknowledged, however, that such distinctions would certainly increase the effectiveness of marker words and, subsequently, the accuracy of the study in this way, «to put our worst foot forward at once, we do not distinguish [...]» (Mosteller and Wallace 1984: 16).

Mosteller and Wallace opted for *function words* as the basis for their scrutiny because their rates do not vary significantly, particularly if compared with content words. In addition, they followed the same criterion as Ellegård (1962), who had already used them as the markers in the attribution assignment of the *Junius Letters*. In a preliminary stage, a list and an index of functions words were compiled, which required their selection from standard lists as well the study of their rate variations over time, inter – and intra – textual origin, Poisson distribution, etc.

As for the statistical methodology, Mosteller and Wallace were determined to solve the problem of attribution by applying Thomas Bayes's (1702-1761) theorem on probability. Accordingly, the rates of the function words (prepositions, conjunctions, and articles) used as discriminators were calculated as numerical probabilities to express degrees of belief, and Bayes's theorem was then used to adjust these probabilities for the evidence in hand (Holmes 1998: 112).

The attribution deriving from Mosteller and Wallace's study is in line with that of historians, in the sense that «Madison is extremely likely [...] to have written all the disputed Federalists: Nos. 49 through 58 and 62 and 63, with the possible exception of No. 55.», and, conversely, discards the likelihood of the *Disputer Papers* to Madison's style on the assumption that Hamilton wrote and Madison edited all the papers. On methodological grounds, Mosteller and Wallace claim that «the weight-rate, the robust Bayes, and the three-categories studies give good support for the main study from the point of view of reasonableness of results» (Mosteller and Wallace 1984: 263-264).

New experiments in the attribution of the *Federalist Papers* have been made by Holmes and Forsyth (1995), Martindale and McKenzie (1995), and Tweedie, Singh and Holmes (1996). Likewise, Bayes's theorem has been used by Girón, Ginebra and Riba (2005: 19-30) to determine whether the style of *Tirant lo Blanc* is homogeneous to confirm or refute the position of medievalists as to the existence of two authors.[9]

[9] *Tirant lo Blanc* is a 15th century chivalry book written in Catalonian, cited by Cervantes as the best book of its genre in the world and considered by Vargas Llosa as the first European modern novel.

4.3. Principal component analysis and Burrows's Delta

Principal component analysis (henceforth *pca*) and *cluster analysis* have been widely used for *multivariate statistical analysis* in the sciences and social sciences (i.e. Meteorology, Allometry, Psychology, Stylometry, etc.). On stylometric grounds, *multivariate statistical analyses* essentially consist in selecting the «N most common words in the corpus under investigation and computing the occurrence rate of these N words in each text or text-unit, thus converting each text into a N-dimensional array of numbers» (Binongo and Smith 1999: 445; Holmes, Gordon and Wilson 2001: 406).

The pca technique breaks up the dependence on the original or observed variables (word occurrence rates), which are transformed into a new set of uncorrelated variables (known as *principal components*), and arranged in decreasing order of importance. The *principal components* are the linear combinations of the original variables, and the decreasing arrangement of the former is taken to allow the first few components to account for most of the variation in the latter, thereby reducing the dimensionality of the problem. A *cluster analysis* of the *MCWs* of a text, on the other hand, «provides an independent and objective view of any groupings amongst the textual samples by means of a tree diagram or dendogram», where the joint of two branches indicates that two texts have a small dissimilarity (or a large similarity) in the values of their *N*, the dissimilarity being commonly measured as the Euclidean distance between two texts (Holmes, Gordon and Wilson 2001: 408). Notwithstanding its claimed usefulness, the clustering shortcomings seem to point that the «the problems with the technique are general rather than local», evincing thus that «cluster analysis may be less effective than has been thought», possibly due to the variation found in the frequencies of the MCWs within a single author's works. (Hoover 2001: 438).

As stated above, the *pca* methodology seeks to explain «as much of the total variation in the data as possible with as few variables as possible [...] and to represent each text in some lower dimensional space so that the texts that are similar to one another in the original variable-space are best represented by points that are close to each other in this lower dimensional representation» (Binongo and Smith 1999: 463-464). Although *pca* was not conceived as a discriminating technique, it suggests «the degree of affinity of an otherwise anonymous text with the writings of known authorship» (Binongo and Smith 1999: 464). This technique, which was

first employed in Stylometry by Burrows and Hassall (1988: 427-453), was termed as *eigenanalysis of function words*, and subsequently used by the same Smith (1990; 1991; 1992; 1993) as well as others like Binongo (1993), Holmes and Forsyth (1995); Baayen, van Halteren and Tweedie (1996); Tweedie, Holmes and Corns (1998); Binongo and Smith (1999); Burrows and Love (1999); Craig (1999); Burrows and Craig (2001).

Some years after his pioneering use of *pca*, Burrows replaced *principal components weights* with *principal components scores* as indicators of authorship for a new measure in attribution studies (Burrows 2002: 267-287; 2003: 10). In this vein, the term *Delta* was coined to represent «D for difference and also as a gesture of respect for Udney Yule and those other pioneers in our field who tried to derive simple expressions of stylistic difference. Udney Yule's Characteristic *K* remains one of the most remarkable of these attempts» (Burrows 2003: 10).

As for the aetiology of this new measure, Burrows (2002: 267) explains that it stems from the observation that «methods of comparison and authorial attribution currently employed in computational stylistics are better fitted for *closed games* (two or three claimant authors as in *The Federalist*) than for more open ones (an anonymous text but little or no outside evidence to identify the most likely candidates)». In particular, *pca* must be considered a «test of comparative resemblance», but never «a test of authorship» in the same way as «artificial neural networks (Waugh, Adams and Tweedie 2000: 187-198) or discriminant analysis (Craig 1999), are [only] at their best in closed inquires» (Burrows 2003: 8). Therefore, Delta (represented as Δ, whenever possible), is not in any way conceived «to displace *pca*» but, on the contrary, «to remedy its chief limitation, to complement it and consolidate it in the role for which it is best fitted, in the middle stages of the game» (Burrows 2003: 8). In other words, Delta is destined to be the best fitted measure in open games.

As in most *pca-based multivariate statistical analysis*, the MCWs (particularly function words) constitute the unit of measure as their handling involves the least manipulation, and the results derived from their counting and subsequent calculations are more intelligible to any observer. The procedure to apply Delta is complex to be summarised in a few lines and difficult to follow if a table with data is not provided. Perhaps, the easiest explanation is by Burrows, who states that the simplest way to represent a large set of numerical differences in word frequencies «[is] to add them up and average them out», given that the common objective of many methods in computational stylistics is that «they all amount to an assessment of numerical differences in word-frequencies

and similar phenomena» (Burrows 2003: 11). In brief, Delta score can be defined as «the mean of the absolute differences between the *z*-scores for a set of word variables in a given text-group and the z-scores for the same set of word-variables in a target text» (Burrows 2002: 271).

The results indicate that the measure is more successful than expected in the *open games* even with short texts as it selects the most likely candidates from a large group, and it is accurate «in singling out the true authors of texts of more than 1,500 words in length» (Burrows 2002: 282).

In analogy to Burrows' respectful reference to Smith as *a stalwart gatekeeper* in the field of authorship attribution studies for his role in the Smith vs. Morton controversy, Hoover and Holmes, among others, also deserve a similar appellative for their defence and for their enthusiasm to test the accuracy of innovative techniques or new measures, thus proposing reliable adaptations to improve the results. To the latter is due the assessment of QSUM, and to the former the test on Tweedie and Baayen's study (1998: 323-352) and on Burrows's Delta (2004a: 453-475). The first of Hoover's study seeks to test the effectiveness of the measure with prose texts – as originally applied to poetry by Burrows – and to demonstrate that, by enlarging the original number of frequent words (from 150 to 800), the accuracy of the measure also increases. With the automation of the process Hoover also analyses the different texts using the complete set of *MCWs* or smaller sets resulting from the removal of contractions, personal pronouns, and contractions and personal pronouns as a whole, or from culling at 70%, i.e. the removal of the words for which a single text supplies more than 70% of the occurrences, etc. (Hoover 2004a: 456). The results come to confirm those obtained by Burrows (Hoover 2004a: 470-471). In the second paper, Hoover proposes two modified methods of calculating Delta and three alternatives or transformations to produce results that are more accurate in four out of the five proposals (Hoover 2004b: 477-475).

5. Stylometric controversies

Along with the three main lexical approaches described above, other lexical and non-lexical methodologies were also employed in authorship attribution, not deprived of controversy in themselves. As for word-based approaches, QSUM is possibly one of the most controversial

inasmuch as their results were admitted at court-rooms as exculpatory evidence in cases of allegedly forged confessions. The QSUM authorship test was originally proposed by Morton and Michaelson (in Hilton and Holmes 1993: 73; Holmes 1998: 114), and it stems from cumulative sum (CUSUM) charts, i.e. graphs which represent the variation between a series of values (in a text, number of words per sentence, words of two or three letters, word-classes, verb frequencies, etc.) and its average. As in other statistical techniques, CUSUM charts were also imported from industrial environments and adapted for authorship attribution studies as in Michaelson, Morton and Wake (1978).

The theoretical basis of QSUM rests upon the uniqueness and permanency of each person's communicative habits, to the extent that quantification of their different linguistic habits can serve to tell them apart (Holmes 1998: 114; Hilton and Holmes 1993: 74). Prototypical examples of linguistic habits could be, among others, the distribution of sentences in term of their word-length, the number of adjectives per sentence, or the distribution of words according to class (nouns, adjectives, pronouns, etc.). In practice, a QSUM test consists of, at least, two CUSUMS: one to represent the data of sentence word-length and the other to plot the frequency of the feature under scrutiny, which are then superimposed to observe their similitude or dissimilitude. The sameness of both plots would then point to a consistent correspondence. The application of the QSUM test to a different author's text is expected to produce a different plotting, which will serve to discriminate from one another. Similarly, when texts by different authors are linked, a significant discrepancy is expected at the point where the texts were concatenated, though the sense of significant discrepancy constitutes by itself a serious motive of controversy.

The accuracy of the QSUM method has been seriously criticised on account of the lack of a solid statistical base and the need for a rigorous validation. The methodology was eventually assessed by Hilton and Holmes, who also tested the validity of the weighted Cusum test to conclude that «it performs marginally better than the QSUM test, but the cumulative sum techniques do not give consistently reliable results» (Hilton and Holmes 1993: 74-80). Likewise, the method was found to be unreliable in Holmes and Tweedie's investigation (1995: 19-47).

In addition to the controversy generated by the employment of QSUM, particularly at law courts, two other approaches are worth mentioning for the same reason. One is Morton's method, conceived to identify authors of works written in English by using tests to com-

pare the number of occurrences, both in the authentic and the disputed texts, of: 1) a word in a given position within the sentence; 2) the relative position of a word with respect to others; and 3) the relative position of synonyms and antonyms. Morton's method, used by Merriam to deal with the authorship of several Shakespeare's texts (1987: 57-58), was subsequently tested by Smith (1985a; 1985b), who condemned it for the lack of rigour and the small number of samples used (Holmes 1998: 113), hence giving rise to the Smith vs. Norton controversy.

The other refers to Thisted and Efron's application of Fisher's technique to a Shakespeare poem (1987: 445-455). The attribution was then questioned by Valenza (1990: 1-20), who demonstrated that Thisted and Efron's attribution tests were not accurate when applied to Marlow's and Shakespeare's works.

6. Other approaches: old and new

In addition to the afore-mentioned techniques, there are other valid approaches for the authorship attribution of written material. Some of them can be deemed transformations of previous proposals, which are accordingly based on a particular technique (i.e. *pca*), as happens with the multivariate statistical analysis, or by proposing an *innovative* measure, such as that of *intertextual distance* (Labbé 2007: 33), although it can be traced back to Merriam's use of Morton's method (Merriam 1987: 57-58).

Others can be taken as particular developments deriving from new trends in scientific fields, and which are adapted to solve authorship problems. This is the case of artificial neural networks, used to simulate the performance of human intelligence in discrimination and classification tasks (Waugh, Adams and Tweedie 2000: 187-198; Tweedie, Singh and Holmes 1996: 1-10), or else the replication of the genetic code or DNA structure to entrap an author's salient features as used in keyword detection (Ortuño *et al.* 2002: 759-764).

Finally, some others are mostly founded on morphologically or syntactic parsed corpora, in the assumption that the more discriminant the salient stylistic features are, the more accurate and reliable the technique will be, and authorship attribution will eventually become more successful.

7. Conclusions

The present paper is a state-of-the-art of non-traditional authorship attribution studies to shed light on the assets and shortcomings of the different approaches to the discipline. In light of the previous examination, the following conclusions have been accordingly drawn:

1. The use of non-traditional methodologies for authorship attribution does not aim, in any case, at replacing the traditional approaches. Instead, they must be taken as a complementary tool to discriminate between two works (*closed games*), or among several works (*open games*) by means of a quantitative assessment (Love 2002: 100-101; Rudman 1998: 351-365; Holmes 1998: 111). As for their sequence, they should be applied in succession, non-traditional approaches following traditional ones.

2. There is no consensus as to the existence of a universal non-traditional methodology or technique, regardless of language, genre, etc. In this same line, Mannion and Dixon argue that «it is axiomatic that no single test can be successfully applied to every authorship problem» insofar as more often than not the validity of an approach is dependent upon different variables, i.e *open* versus *closed* games (Mannion and Dixon 2004: 497-508). It is a fact that authorial attribution gains reliability when more than one technique is applied, as in Hoover (2001: 421-444), who replicates Tweedie and Baayen's experiment (1998: 323-352), or further develops Burrows's Delta (Hoover 2003: 151-178).

A call is made here, therefore, for the search of a standard methodology by means of which the results of any given test may be reliably verified. On methodological grounds, it would be desirable a standardization of the texts under scrutiny, a harmonization of the concept of word (especially proper and compound nouns, whether hyphenated or not) as well as the establishment of rules about «the treatment of quotations, numbers and the other special usages» (Mosteller and Wallace 1984: 249). Moreover, the availability of the raw data would be an additional asset for the sake of increasing the scientific rigour of the tests, as they could be checked by other scholars.

3. There is a general agreement as to the inconvenience of comparing works of a different genre or chronology. As a matter of fact, genre can severely distort the task of authorial attribution, even if written by the same pen. In this same fashion, there is also grounded evidence to affirm that an author's lexical richness can vary chronologically when the whole canon is examined (Smith and Kelly 2002: 412). Therefore,

the texts must be chosen with the utmost care not to analyse different genres and, more importantly, chronologically distant pieces.

4. Most of the original approaches were shown to be dependent upon text length, that is why both Yule's *K* and Zipf's *Z* came to minimize the inaccuracy of the previous approaches in the evaluation of an author's lexical richness by simple statistics. Methodologically speaking, therefore, the safest decision to treat works of different length is to divide the material into equal-sized pieces (Smith and Kelly 2002: 412) as this measure will certainly ease both formal and informal statistical analysis of the data (Mosteller and Wallace 1984: 249). The excerpts may range from 1000 words to 3000 words in the sense that larger samples may become less informative for attribution purposes. The randomization of the samples to distort the original ordering of the text as well as the accumulation of blocks to obtain larger samples also contribute to detect any likely variation within a large piece, if any (Miranda-García and Calle-Martín 2005a: 115-130).

5. In lexical terms, the word has been widely accepted as the input for statistical authorship attribution studies, not only in itself but also in terms of lemmas, collocations, etc. Irrespective of whether tokens or types are used, scholars (Hoover 2001: 422; Burrows 2003: 10) usually recommend to have homographs disambiguated (in terms of their class or syntactic function) as they show different rates of frequency. Mosteller and Wallace (1984) were aware of the discriminating value of the marker words *of, by, to*, *by*, in the sense that their rates clearly allowed to assign the *Disputed Papers* to Madison as a block. The distinction becomes even sharper when considering the occurrences of *to* as a preposition or as an infinitive marker, or the different functions of *that* (i.e. determiner, demonstrative pronoun, conjunction, adverb, relative), or the proportional distribution of the occurrences of *wh*-relatives with respect to *that*-relatives, etc., among others. This technique releases the analyst from the bindings of text-dependency, as the preference for one option can be considered a linguistic habit which constitutes a salient stylistic feature irrespective of text-length involved (since it is not the ratio with respect to *N, V*, or the number of sentences, but the proportional distribution of the homographs in terms of their function). Unless automatically accomplished, disambiguation tasks are time-consuming but the added value is worth the effort.

Many benefits derive from the use of a tagged corpus including the lemma, word class, accidence, etc. The advantages are many when the texts of a highly-inflected language are involved (Miranda-García and Calle-Martín 2007: 49-66; Miranda-García, Calle-Martín and Mar-

qués-Aguado 2008: 210-225). In case of a tagged corpus, a lemma-based approach (*ANSWER*, *noun* as opposed to *ANSWER*, *verb*) is liable to offer a more distinctive account than if word-based, as all the occurrences (irrespective of accidence) are associated with the corresponding lemma. Otherwise, the analyst cannot see the wood for the trees as to whether the actual occurrence of a common word like *answers* points to a higher usage as a noun or as a verb, such information being helpful to distinguish between A and B. Moreover, a step forward would be to account for the following group of words: *a*) hyphenated vs. non-hyphenated compounds; *b*) adverbial, prepositional or conjunctive phrases (i.e. *as a result*, *because of*, *on condition that, point out*) as they would require to count them as a whole rather than independently. In view of all this, the employment of tagged works can help to increase the accuracy of the results notwithstanding Rudman's advice of dealing with raw texts free of human manipulation.

6. A greater effort is needed to cope with the text brevity, which characterizes notices, emails and sms, used in Forensic Stylometry. This fact constitutes a serious disadvantage for the stylometrician, which becomes even greater on account of the limited number of specific corpora available. Therefore, the compilation of such corpora is a need for stylometricians and forensic linguists (Santana Lario 2007: 108-110).

There does not seem to be a better colophon to this paper than Burrows' *desideratum* that «we need to match a natural desire to work on celebrated cases like *Henry VIII* and *The Revenger's Tragedy* with a more sober, though less immediately rewarding, concern for testing our methods thoroughly on cases where the true answers are not in any doubt» (cited in Hoover 2001: 422).

References

Baayen, H., H. van Halteren and F.J. Tweedie. 1996. «Outside the Cave of Shadows. Using Syntactic Annotation to Enhance Authorship Attribution». *Literary and Linguistic Computing* 11, pp. 121-131.

Binongo, J.N.G. 1993. «Incongruity, Mathematics and Humor in Joaquinesquerie». *Philippine Studies* 41, pp. 477-511.

Binongo, J.N.G. and M.W.A. Smith. 1999. «The Application of Principal Component Analysis to Stylometry». *Literary and Linguistic Computing* 14.4, pp. 445-465.

Burrows, J. 2002. «'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship». *Literary and Linguistic Computing* 17.3, pp. 267-287.

—. 2003. «Questions of Authorship: Attribution and Beyond». *Computers and the Humanities* 37, pp. 5-32.

Burrows, J. and A.J. Hassall. 1988. «*Ann Boleyn* and the Authenticity of Fielding's Feminine Narratives». *Eighteenth-century Studies* 21, pp. 427-453.

Burrows, J. and H. Love. 1999. «Attribution Tests and the Editing of Seventeenth-century Poetry». *Yearbook of English Studies* 29, pp. 151-175.

Burrows, J. and H. Craig. 2001. «Lucy Hutchinson and the Authorship of Two Seventeenth-century Poems. A Computational Approach» *The Seventeenth Century* 16, pp. 259-282.

Can, F. and J. Patton. 2004. «Change of Writing Style with Time». *Computers and the Humanities* 38, pp. 61-82.

Craig, H. 1999. «Authorial Attribution and Computational Stylistics: If You Can Tell Authors Apart, Have You Learned Anything About Them?». *Literary and Linguistic Computing* 14, pp. 103-113.

Ellegård, A. 1962. *A Statistical Method for Determining Authorship: The Junius Letters, 1769-1772*. Göteborg, Acta Universitatis Gothenburgensis.

Foucault, M. 1969. «What is an author?». In David Lodge (ed.), *Modern Criticism and Theory: A Reader*. London, Longman.

Girón, J., J. Ginebra and A. Riba. 2005. «Bayesian Analysis of a Multinomial Sequence and Homogeneity of Literary Style». *The American Statistician* 59.1, pp. 19-30.

Hilton, M.L. and D.I. Holmes. 1993. «An Assessment of Cumulative Sum Charts for Authorship Attribution». *Literary and Linguistic Computing* 8, pp. 73-80.

Holmes D.I. 1998. «The Evolution of Stylometry in Humanities Scholarship». *Literary and Linguistic Computing* 13.3, pp. 111-117.

Holmes, D.I. and R.S. Forsyth. 1995. «The *Federalist* Revisited: New Directions in Authorship Attribution». *Literary and Linguistic Computing* 10, pp. 111-127.

Holmes, D.I. and F.J. Tweedie. 1995. «Forensic Stylometry: a Review of the Cusum Controversy». *Revue Informatique et Statistique dans les Sciences Humaines* 31, pp. 19-47.

Holmes, D.I., L.J. Gordon and C. Wilson. 2001. «A Widow and her Soldier: Stylometry and the American Civil War». *Literary and Linguistic Computing* 16, pp. 403-420.

Honoré, A. 1979. «Some Simple Measures of Richness of Vocabulary». *Association for Literary and Linguistic Computing Bulletin* 7.2, pp. 172-177.

Hoover, D.L. 2001. «Statistical Stylistics and Authorship Attribution: an Empirical Investigation». *Literary and Linguistic Computing* 16.4, pp. 421-444.

—. 2003. «Another Perspective on Vocabulary Richness». *Computer and the Humanities* 37, pp. 151-178.

—. 2004a. «Testing Burrows's Delta». *Literary and Linguistic Computing* 19.4, pp. 453-475.

—. 2004b. «Delta Prime?». *Literary and Linguistic Computing* 19.4, pp. 477-495.

Keller, S.D. 2003. «Shakespeare's Rhetorical Fingerprint. New Evidence on the Authorship of *Titus Andronicus*». *English Studies* 2, pp. 105-118.

Labbé, D. 2007. «Experiments on the Authorship Attribution by Intertextual Distance in English». *Journal of Quantitative Linguistics* 14.1, pp. 33-80.
Lancashire, I. 1998. «Paradigms of Authorship». *Shakespeare Studies* 26, pp. 296-301.
Love, H. 2002. *Attributing Authorship: An Introduction*. Cambridge, Cambridge University Press.
Mannion, D. and P. Dixon. 2004. «Sentence-length and Authorship Attribution: the Case of Oliver Goldsmith». *Literary and Linguistic Computing* 19.4, pp. 497-508.
Martindale, C. and D. McKenzie. 1995. «On the Utility of Content Analysis in Author Attribution: The *Federalist*». *Computer and the Humanities* 29, pp. 259-270.
Merriam, T. 1987. «An Investigation of Morton's Method: a Reply». *Computers and the Humanities* 21, pp. 57-58.
Michaelson, S., A.Q. Morton and W.C. Wake. 1978. «Sentence Length Distribution in Homer and Hexameter Verse». *ALLC Bulletin* 6.3, pp. 254-267.
Miranda-García, A. and J. Calle-Martín. 2005a. «The Validity of Lemma-based Lexical Richness in Authorship Attribution: a Proposal for the Old English Gospels». *ICAME Journal* 29, pp. 115-130.
—. 2005b. «Yule's Characteristic K Revisited». *Language Resources and Evaluation* 39.4, pp. 287-294.
—. 2007. «Function Words in Authorship Attribution Studies». *Literary and Linguistic Computing* 22.1, pp. 49-66.
Miranda-García, A., J. Calle-Martín and T. Marqués-Aguado. 2008. «Morphological Features in the Translatorship Attribution of the West Saxon Gospels». *English Studies* 89.2, pp. 210-225.
Mosteller, F. and Wallace, D.L. 1984. *Applied Bayesian and Classical Inference. The Case of the Federalist Papers*. New York, Springer-Verlag.
Orlov, J.K. 1983. «Ein Modell der *Häufigkeitsstruktur* des Vocabulars». In H. Guiter and M. Arapov (eds.), *Studies on Zipf's Law*. Brockmeyer, Bochum, pp. 154-223.
Ortuño, M., P. Carpena, P. Bernaola, E. Muñoz and A.M. Somoza. 2002. «Keyword Detection in Natural Languages and DNA». *Europhysics Letters* 57.5, pp. 759-764.
Rudman, J. 1998. «The State of Authorship Attribution Studies: Some Problems and Solutions». *Computers and the Humanities* 31, pp. 351-365.
—. 2000. «Non-traditional Authorship Attribution Studies: Ignis Fatuus or Rosetta Stone?» *BSANZ Bulletin* 24, pp. 163-176.
Santana Lario, J. 2007. «Corpus Delicti: la lingüística de corpus al encuentro de la lingüística forense». In M.A. Martínez-Cabeza, N. McLaren and L. Quereda Rodríguez-Navarro (eds.), *Estudios en Honor de Rafael Fente Gómez*. Granada, Editorial Universidad de Granada, pp. 101-114.
Sichel, H.S. 1975. «On a Distribution Law for Word Frequencies». *Journal of the American Statistical Association* 70, pp. 542-547.
Simpson, J.A. and E.S.C. Weiner (eds.). 1989. *The Oxford English Dictionary*. CD-Rom version 3.1. Oxford, Oxford University Press.
Smith, M.W.A. 1985a. «An Investigation of Morton's Method to Distinguish Elizabethan Playwrights». *Computer and the Humanities* 19, pp. 3-21.

—. 1985b. «An Investigation of the Basis of Morton's Method for the Determination of Authorship». *Style* 19, pp. 341-359.
—. 1990. «Attribution by Statistics: a Critique of Four Recent Studies». *Revue informatique et statistique dans les sciences humaines*, 26, pp. 233-51.
—. 1991. «The authorship of *The Revenger's Tragedy*». *Notes and Queries* 236, pp. 508-511.
—. 1992. «The problems of Acts I-II of Pericles». *Notes and Queries* 237, pp. 346-355.
—. 1993. «*Edmund Ironside*». *Notes and Queries* 238, pp. 202-205.
Smith, J.A. and C. Kelly. 2002. «Stylistic Constancy and Change across Literary Corpora: Using Measures of Lexical Richness to Date Works». *Computer and the Humanities* 36, pp. 411-430.
Somers, H. and F.J. Tweedie. 2003. «Authorship Attribution and Pastiche». *Computers and the Humanities* 37, pp. 407-429.
Stamatos, E., N. Fakotakis and G. Kokkinakis. 2001. «Computer-based Authorship Attribution without Lexical Measures». *Computers and the Humanities* 35, pp. 193-214.
Thisted, R. and B. Efron. 1987. «Did Shakespeare Write a Newly Discovered Poem?». *Biometrika* 74, pp. 445-455.
Tweedie, F.J., S. Singh and D.I. Holmes. 1996. «Neural Network Applications in Stylometry: the *Federalist Papers*». *Computers and the Humanities* 30, pp. 1-10.
Tweedie, F.J. and R.H. Baayen. 1998. «How Variable May a Constant Be? Measures of Lexical Richness in Perspective». *Computers and the Humanities* 32, pp. 323-352.
Tweedie, F.J., D.I. Holmes and T.N. Corns. 1998. «The Provenance of *De Doctrina Christina*, attributed to John Milton: A Statistical Investigation». *Literary and Linguistic Computing* 13.2, pp. 77-87.
Valenza, R.J. 1990. «Are Thisted-Efron Authorship Tests Valid?». *Computer and the Humanities* 30, pp. 1-20.
Waugh, S., A. Adams and F.J. Tweedie. 2000. «Computational Stylistics Using Artificial Neural Networks». *Literary and Linguistic Computing* 15, pp. 187-198.
Williams, C.B. 1939. «A Note on the Statistical Analysis of Sentence-length as a Criterion of Literary Style». *Biometrika* 62, pp. 207-212.
Yule, G.U. 1938. «On Sentence-length as a Statistical Characteristic of Style in Prose: with Applications to Two Cases of Disputed Authorship». *Biometrika* 30, pp. 363-390.
—. 1944. *The Statistical Study of Literary Vocabulary*. Cambridge, Cambridge University Press.

Diseño y preimpresión: Carolina Valcárcel

1ª edizione, febbraio 2009

Carocci editore S.p.A., Roma

Finito di stampare nel febbraio 2009
dalla Litografia Varo (Pisa)

ISBN 978-88-430-5091-8