UNIVERSITY OF MURCIA

International Journal of English Studies

IJES

www.um.es/ijes

Exploring State-of-the-Art Software for Forensic Authorship Identification

VICTORIA GUILLÉN-NIETO, CHELO VARGAS-SIERRA, MARÍA PARDIÑO-JUAN, PATRICIO MARTÍNEZ-BARCO, ARMANDO SUÁREZ-CUETO*
Universidad de Alicante

ABSTRACT

Back in the 1990s Malcolm Coulthard announced the beginnings of an emerging discipline, *forensic linguistics*, resulting from the interface of language, crime and the law. Today the courts are more than ever calling on language experts to help in certain types of cases, such as authorship identification, plagiarism, legal interpreting and translation, statement analysis, and voice identification. The application of new technologies to the analysis of questioned texts has greatly facilitated the work of the language scientist as expert witness in the legal setting, and contributed to the successful analysis and interpretation of style providing statistical and measurable data. This article aims at presenting linguists and researchers in forensic linguistics with an exploration of the strengths, limitations and challenges of state-of-the-art software for forensic authorship identification.

KEYWORDS: Forensic linguistics, language, crime, law, software for forensic authorship identification.

Address for correspondence: Victoria Guillen-Nieto. Departamento de Filología Inglesa. Campus de San Vicente del Raspeig. Ap. 99 E-03080 Alicante, Spain.. Tel.: 34 965909318. Fax: 34 965903800. E-mail: victoria.guillen@ua.es

I. THE AIM OF THIS DISCUSSION

Over the last decade it has become evident that linguists can be of service to the law, and courts, especially in Common Law countries, are calling on language experts more and more to help in certain types of cases, such as authorship identification, voice identification, plagiarism, legal interpreting and translation, statement analysis, etc.

Undoubtedly, the application of new technologies to the analysis of questioned texts has significantly facilitated the work of the language scientist as expert witness in the legal setting, by enhancing the scientific reliability of descriptive linguistic analysis with measurable data, and reducing the time-consuming task involved in the observation, description, analysis, and counting of the data.

The aim of this discussion is to present language experts and researchers with an exploration of the strengths, limitations, and challenges of state-of-the-art software for forensic authorship identification. For the purpose of analysis, this article will be divided into two main parts:

Part one will be devoted to forensic linguistics as an up-and-coming discipline within the field of applied linguistics. Our discussion in this first part will provide the reader with essential background information to understand forensic language researchers' recent, healthy interest in new techniques and methods that may help the language expert explain linguistic findings in statistical terms, and be consistent with the current scientific reliability standard that is demanded for linguistic evidence by the judiciary, especially in Common Law countries.

Part one will be further divided into three sections. Firstly, we will offer the reader a brief overview of authorship identification and the birth of forensic linguistics. Secondly, we will look at stylistic analysis as an approach to forensic authorship identification. And lastly, we will consider the problems faced by the language scientist as expert witness in the legal setting, after the Federal Rules of Evidence in the USA providing the new standard for admitting expert scientific testimony in a federal trial came into force (Daubert v. Merrell Dow Pharmaceuticals 92-102, 509 U.S., 579, 1993). A consideration of a major challenge to forensic linguistics, as seen in latest developments of the discipline, will bring part one to an end.

Part two will concentrate on new advances in software for quantitative data analysis used in forensic authorship identification by examining a selected sample of state-of-the-art tools.

Finally, the concluding remarks section will bring together the most relevant conclusions as to the role played by software for quantitative analysis in forensic authorship identification, and suggestions for further development will be given as to the main challenges in this field.

PART ONE
II. FORENSIC AUTHORSHIP IDENTIFICATION AND THE BIRTH OF FORENSIC LINGUISTICS

The emergence of forensic linguistics as a discipline is closely related to two prominent cases of disputed authorship in police statements in the UK. For many the discipline of forensic linguistics came into being with Svartvik's (1968) publication of his groundbreaking study into the altered police statements in the Timothy John Evans case. On analysing textual modification, Svartvik demonstrated the presence of two different registers in Evan's statements by using a pioneering technique combining language description and statistical analysis. The linguistic evidence provided by Svartvik in his expert testimony was considered to be crucial in the posthumous pardon of Evans, who had been hanged in the 1950s.

Some years later, in April 1988, Rieber and Stewart (1990: 1-4), acting under the sponsorship of the New York Academy of Sciences, organized a workshop on the role of the language scientist as expert in the legal setting, in which they reached the conclusion that the general trend toward the increased use of language experts was running parallel to the specific development of linguistic knowledge and methodology within those areas that were formerly dealt with by judges and lawyers. "The traditional language training and intuitive abilities of law practicioners (…)", claimed Rieber and Stewart (1990: 3), "(…) are no longer a match for the theoretical analytical advances that are from such linguistic subfields as syntax, semantics, and discourse analysis". Until then the use made of language scientists by the legal profession had been largely limited to the area of substance, namely comparisons of samples of handwriting and of tape-recorded voices for authorship identification, a process requiring highly technical instrumentation and skilled knowledge.

Shortly after, in the 1990s, Coulthard introduced a pioneering analytical approach, for which he coined the term *forensic discourse analysis*, to the alleged statement of Derek Bentley who, like the ill-fated Evans, had also been hanged in the 1950s. In his approach to disputed authorship, Coulthard combined insights from different linguistic fields, namely speech act theory, pragmatics, discourse analysis, psycholinguistics, and statistical analyses (Coulthard, 1992: 242-258). The linguistic evidence provided by him in his expert testimony was also decisive in Derek Bentley's posthumous pardon. By that time, criminal justice professionals had begun to realise that linguists could be of service to the law by helping those who had been treated unjustly in the past.

Over the last decade, many linguists have indeed explored the interface between language, crime and the law at the investigation level, and forensic linguistics is ripe for debate and argument, as shown in the discipline's major journal *The International Journal of Speech, Language and the Law*, which is the official journal of the *International Association of Forensic Linguistics* and the *International Association for Forensic Phonetics and Acoustics*, and in the proliferation of publications in the field (Alcaraz-Varó, 2005: 49-66; Coulthard, 1992, 1993: 86-97, 1994, 2005: 249-274; Gibbons, 2003; McMenamin, 2002, 2004; Olsson, 2004; Shuy, 2005: 43-64; Tanner & Tanner, 2004; Turell [ed.], 2005 etc).

Similarly, courts all over the world, especially in Common Law countries—and more than ever in Roman Law countries, are calling on language experts to help in forensic cases of authorship identification, plagiarism, mode identification, legal interpreting and translation, transcribing verbal statements, the language and discourse of courtrooms, language rights, statement analysis, forensic phonetics, textual status, etc.

All in all, forensic linguistics as a science is young and so "(...) nothing is yet cast in stone" (Olsson, 2004: 7). Therefore, a lot of research is still to be done on the part played by universities in this field as regards the development of new techniques and methods.

Having presented this short overview of the birth of forensic linguistics, we will now move on to consider stylistic analysis as a well-established methodological approach to forensic authorship identification.

III. STYLISTIC ANALYSIS AS AN APPROACH TO FORENSIC AUTHORSHIP IDENTIFICATION

Forensic linguistics has benefited to a large extent from the application of descriptive linguistics to the analysis of forensic texts. (Coulthard, 2005: 249-274; Shuy, 2005: 19-48; Turell, 2006: 43-64).

Stylistic analysis as an approach to authorship identification in literary contexts is based on the assumption that it is possible to *identify*, *describe* and *measure* a writer's individual style or idiolect by careful linguistic observation and analysis of his/her unique set of linguistic choices.

Forensic text analysis makes use of stylistics to reach a conclusion and opinion related to the authorship of a questioned writing in the context of litigation (McMenamin, 2002: 163-164). For example, a typical case of disputed authorship involves comparing or contrasting the questioned writing with a set of known writings of one or more candidate authors. Such an analysis is accomplished by analysing the writing style of the two sets of texts, the questioned and the known writings. Results of this analysis may lead the language expert to any of the following options: (a) authorship attribution, (b) authorship identification, (c) determination of the resemblance of questioned writings to known writings, (d) elimination of one or more suspect authors, and (e) neither elimination nor identification because the investigation is inconclusive as not enough linguistic evidence has been found to support either hypothesis.

IV. FORENSIC STYLISTICS AND SCIENTIFIC EXPERT TESTIMONY

Recent changes in the criteria for reliability and evidence, especially in the USA (Daubert v. Merrell Dow Pharmaceuticals (92-102), 509 U.S. 579, 1993), have laid emphasis on the heuristic requirements of the scientific method traditionally used in the natural sciences, and highlighted the need to provide *quantitative* or *measurable probability* as regards the

admission of linguistic evidence at court. The underlying reason for this is that for many non-quantitative results do not constitute scientific knowledge.

This poses a major threat to the language scientist as expert witness in the legal setting, since the methods of inquiry in the humanities and social sciences are unavoidably more relative than those in the natural sciences. On top of that, to account for quantitative or measurable probability is not always possible in forensic linguistic reports for a variety of reasons:

Firstly, some linguistic features may be difficult to identify as discrete units, such is the case of style markers like the overall design of pages, socio-pragmatic aspects, interferences from other languages, etc. Secondly, the linguistic significance of an identified variable may not always be captured by counting. Thirdly, a variable may be linguistically significant because it rarely occurs in the language, but it does not occur frequently enough in the data to be meaningfully counted. Fourthly, there may not be sufficient data for valid quantification. Last but not least, in a given case of disputed authorship there may be no set of known texts for comparison to a set of questioned texts.

Considering the fact that there are disciplines like forensic stylistics in which it is not always possible to count or measure the evidence, concepts such as "inductive probability" (Cohen, 1977) or "nonmathematical but structured sense of probability" (McMenamin, 2002: 129) have been suggested. In both proposals, probability is based on a comparative or ordinal gradation rather than on a quantitative or measurable one.

In connexion with the problem posed by the new standard for scientific expert testimony, other problems facing language experts today are: a lack of training in statistical analysis, and the fact that counting, if not done automatically, is time-consuming, labour intensive, expensive, and too slow for the urgency required in most law enforcement and criminal justice investigations.

In spite of the above-mentioned problems and limitations, many linguists share the opinion that the present emphasis on quantification is important for two reasons: (a) it meets current methodological requirements for the study of linguistic variation (hypothesis testing and verification), and (b) satisfies external requirements for expert evidence as imposed by the judiciary (McMenamin, 2002: 174).

From the above short discussion on the new requirement for scientific expert testimony, two important ideas emerge as to the need to develop an interdisciplinary stylistic approach to have the best of both worlds, qualitative and quantitative analyses, so as to be able to measure linguistic variation and quantify the writing and language data of cases.

We will now move on to the second part of our discussion, the aim of which is to present linguists and researchers in forensic linguistics with an exploration of the strengths, limitations and challenges of software for their work and research in forensic authorship identification.

PART TWO
V. STATE-OF-THE-ART SOFTWARE FOR RESEARCH IN FORENSIC AUTHORSHIP IDENTIFICATION

For the purpose of analysis, ten state-of-the-art tools have been selected. These may be broadly classified into two main groups: (a) software for plagiarism detection and historical authorship investigations, and (b) software for general purpose text analysis. Whereas the former covers software such as *JVocalyse* v 2.05, *CopyCatch Gold* v 2, and *Signature Stylometric System* v 1.0, the latter includes *WordSmith Tools* (WST) v 4.0, *Simple Concordance Program* v 4.09, *Textanz* v 2.4.1.0, *AntCone* v3.2.1, *Yoshikoder* v.0.6.3-preview.1 Build 13, *Lexico* v 3, and *T-LAB* Pro 5.4.

The evaluation results for each of these tools are shown in the Appendix. However, due to space restrictions, only four of the ten selected tools will be fully discussed in the following pages. These are: *JVocalyse*, *CopyCatch Gold*, *Signature Stylometric System*, and *WordSmith Tools* (WST).

As shown in the Appendix, the *Result Template* was adapted from EAGLES (1995) and expanded to include the identification and quantification of style markers at all linguistic levels found in eighty authorship cases, as well as some relevant statistics (McMenamin, 2002: 216-231).

The *Result Template* consists of eleven well-defined parts: (1) User interface, (2) Product documentation and user help, (3) User interface elements, (4) Tool PageRank, (5) Text selection and result presentation, (6) Qualitative analysis (Identification of style markers at all linguistic levels), (7) Quantitative analysis of style markers, (8) Statistical tests for significance of variables, and (9) Measures of authorship discrimination.

Once we have applied this *Result Template* to the selected sample of state-of-the-art software, we will present the reader with the most important findings in the next subsections. (See Appendix for a detailed evaluation of the whole sample).

We will begin our discussion by reviewing software for plagiarism detection and historical authorship investigations, namely *JVocalyse*, *CopyCatch Gold*, and *Signature Stylometric System*. After that, we will examine software for general purpose text analysis, that is, *WordSmith Tools* (WST).

V.1. JVOCALYSE V 2.05

JVocalyse v 2.05 is a product developed by David Woolls (2003: 102-112) of *CFL Software Development*, in association with members of the *Corpus Forensic Linguist group* at the University of Birmingham, to which the renowned forensic linguist Coulthard belongs.

As shown in Figure 1 below, the initial screen of *JVocalyse* enables the language researcher to carry out a number of basic operations by clicking on the following buttons: *Select Files*, *Read Files*, *Save the data*, *Clear*, and *Language* (this allows access to a built-in

list of 450 English function words, although it is possible to work with lists in other languages).

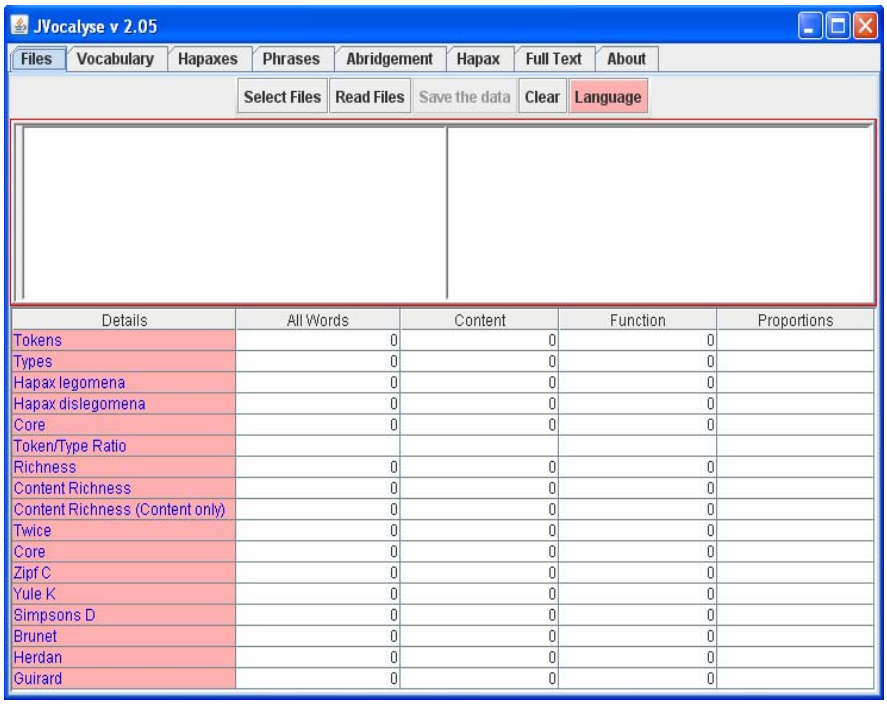


Figure 1: Initial screen of *JVocalyse* v 2.05

JVocalyse makes use of lexically based measures. Table 1 summarises the style markers and statistics available in the tool for measuring the richness of the vocabulary used by the author.

Tokens	The total number of words in each text (Full text).
Types	The total number of different words used (Vocabulary).
Hapax Legomena	Once-occurring words.
Hapax dislegomena	Twice-occurring words.
Core	Number of content words which are included on the Core list. The words on this list are the most common content words which have appeared in writing for children over the last century. They are included because empirical testing has shown that they appear with different total frequencies in a wide range of writing, and that the different usage is frequently associated with authorship.
Token/Type Ratio	Standard division of all the words in the texts (Tokens) by the vocabulary used (Types), to give an average word usage value.
Vocabulary Richness	Three measures of Richness are calculated. Content Richness and Content Only are designed for use with short forensic texts by discounting the effect of the function words.

Twice Ratio	This is the vocabulary used twice as a percentage of the full vocabulary.
Core percentage	This is the percentage of all the content words represented by the total occurrences of the words on the Core vocabulary list.
Abridgement	This function produces an abridgement, *i.e*. a set of sentences in text order. A selected sentence contains links with at least two other sentences at word level.
Hapax distribution	This page allows the selection of a sample or the examination of the number of content words which occur just once in the whole text for successive Word Frames in the texts.
Full text mark up	This page shows the full text with words coloured as follows: Bold red (Content word which occurs just once), Light red (Function word which occurs just once), Light Blue (Content word which occurs twice), Italic light blue (function word which occurs twice), and Black italic (Content words used more than twice).

Table 1: Summary of the actions *JVocalyse* v 2.05 provides

V.1.1. Strengths

On looking at the most remarkable strengths of *JVocalyse* as software for forensic authorship identification, we may highlight the following:

Although the tool is not designed to undertake statistical authorship identification in itself, it does supply measurable data with which the language expert may provide a quantitative analysis of an individual's writing style.

It allows the rapid analysis of suspect documents, producing statistical, vocabulary and phrasal information about the texts.

It lets the user see how different texts have different ratios of content to function words both at the full text (Token) level and at the vocabulary (Type) level.

It facilitates the identification of word strings, either lexical phrases or function phrases, which might reveal linguistic patterns of word use. These listings may be particularly useful when looking for authorial habits or unexpected repetitions.

It allows regularity of patterns to be examined in a full text or in a sample in long texts.

The *Full text mark up* page shows the patterns of word use in different colours, to give a visual representation of the distribution of the vocabulary frequency through the text.

V.1.2. Limitations

JVocalyse only involves lexically based measures (content words, function words, the hapax legomena, the hapax dislegomena, the type-token ratio, etc.). This means that other linguistic style markers that have proved to be relevant in the clarification of some cases of disputed authorship are necessarily overlooked. This is is the case, for example, of markers such as text format, spelling, errors and omissions, punctuation, capitalization, numbers and symbols, abbreviations, word couplets, functional variation of language use, variation in syntax, variation in discourse, and interference features from other languages.

Consequently, *JVocalyse*, despite its unquestionable value for forensic text analysis, seems to be a more appropriate tool for authorship attribution of anonymous or doubtful literary texts, considering the objective lexical style markers it uses for identifying the potential author of questioned writings.

V.2. COPYCATCH GOLD V 2

CopyCatch Gold v 2 is also software for forensic text analysis developed by David Woolls (2003: 102-112) of *CFL Software Development* in association with members of the *Corpus Forensic Linguist group* at the University of Birmingham. Since 2007 *JVocalyse and CopyCatch Gold* have been assembled into a single package: *CopyCatch Suite – 2007*.

The main function of *CopyCatch* is to detect plagiarism and collusion between students by comparing submitted documents and calculating the proportion of words and phrases held in common.

As shown in Figure 2 below, the initial screen is divided into two main boxes. In each one, the language researcher is able to carry out a number of basic working operations, such as *Select Work Files*, *Select Comparison Files*, *CopyCatch*, *Compare with Work Files*, *Clear Work files*, and *Clear Comparison Files*.

The initial screen also shows other relevant buttons like *Language*, which allows the user to load a list of functional words, together with specific technical words as functional for particular subjects; *Help*, which provides a user-friendly manual in English; and *Threshold*, which restricts the number of pairs on show by establishing a previously-defined similarity threshold between two sets of texts.

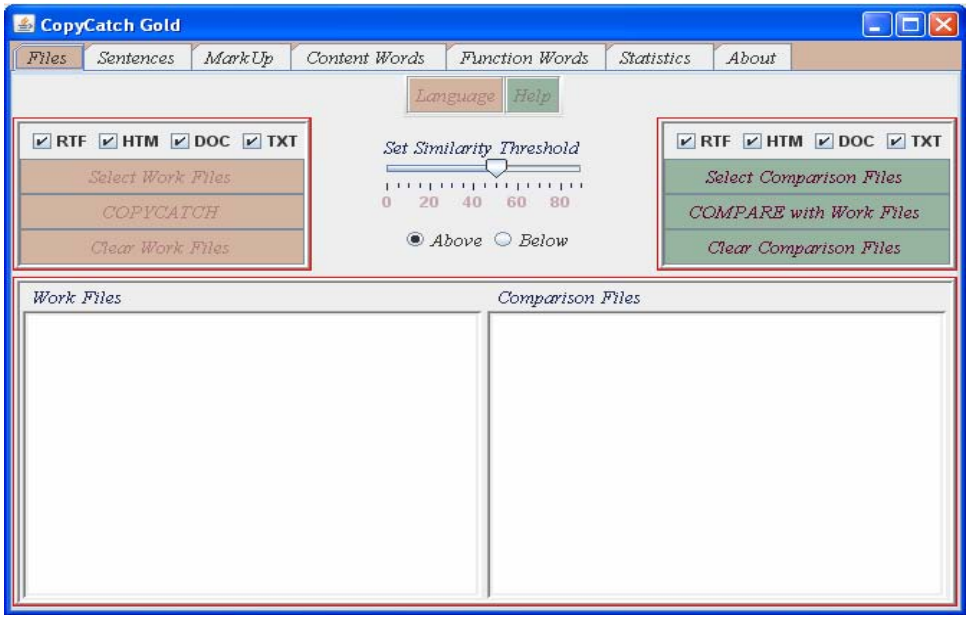


Figure 2: Initial screen of *CopyCatch Gold* v

The approach of *CopyCatch Gold* to plagiarism detection also involves lexically based measures, and allows the identification and quantification of content words, function words, phrases and sentences that are found in common between two sets of texts in a fully contextualized way. More accurately, by clicking on the top buttons, the user will be able to perform the following actions:

Sentences	**Top left listing**. It shows the results in ascending order of percentage of similarity. Percentages are calculated on the total similarity between the texts. **Top right listing**. It shows the sentences which contain three or more content words in common. Sentences are shown in related pairs, *i.e.* Informant 1 [P3 S2], Informant 2 {P5 S4} **Sentence colouring**. Black text shows the words that are different. Red text shows the words that are the same in each sentence.
MarkUp by sentence	Sentences which have been found to have phrasal elements in common are shown.
MarkUp by vocabulary used	Vocabulary which has been found in common is shown. Red text indicates that words are shared once. Blue text means that the words have used more than once across the two texts.
Saving	The files generated can be saved as html or rtf by clicking on the appropriate radio button.
Statistics	The statistics show the breakdown of the texts as occurrences of each type of text. Each file is shown in total

	numbers of content and function words, together with the number of words which are shared between the two files and the number of words shared once only. Adding the two percentages together for each file gives the total amount of similarity.

Table 2: Summary of the actions *CopyCatch Gold* v 2 provides

V.2.1. Strengths

Copycatch Gold is specifically designed for use with forensic texts, the length of which may range from short writings to long documents, particularly in the area of historical investigation of anonymous texts.

The *Language* button on the *Files* tab allows the user to not only modify language but also change to a longer or shorter function word list in any language.

CopyCatch Gold shows all the matching sentences between two sets of texts, cross-referenced by the paragraph/sentence indicator at the end of the sentence. In addition, it allows the user to see words, phrases and sentences where groups of sentences from either text are found in both, and whether they have been kept together in the other text or moved around. The user may also see the work file in full but only the matching sentences from the comparison document.

Forensic linguists may find the *Statistics* included particularly interesting for their research purposes, namely the statistical analysis of the data in terms of the quantity and frequency of words, phrases, and sentences shared between two sets of texts.

V.2.2. Limitations

The main limitation of *CopyCatch Gold* as software for forensic authorship identification is inherent in the different purpose for which it was primarily designed. "The tools were built", explains Woolls (2003: 108), "to allow examination of the structure of the texts, and to give pointers to further phrasal and vocabulary analysis, not to give a swift answer to the question 'Did X write it?' That is not to say that it is the only way they can be used, but that was the intent behind them".

Consequently, relevant style markers for authorship identification that may appear at other language levels are simply not considered for purposes of analysis (text format, spelling, errors and omissions, punctuation, capitalization, interference features from other languages, etc.).

In both authorship identification and plagiarism, the language researcher looks for matching style markers that will serve to demonstrate that two sets of texts were not produced independently. However, there are some fundamental differences between

detecting plagiarism and identifying the author of disputed forensic texts in criminal cases that are often missed.

The first difference relates to the genre and register of writing texts. Whilst in plagiarism the disputed text may be parallel to the plagiarised text in terms of genre and register, in authorship identification, not including authorship attribution of anonymous literary texts, the disputed text, which may well be a suicide letter, a ransom note, an anonymous letter, etc, may exhibit a completely different style to that–or those—of the texts that are known to have been produced by an individual. Moreover, the anonymous author involved in a criminal case will do his best to "mask" or "disguise" his writing style as much as possible in order to hide his true identity.

The second difference relates to text length. Whereas in plagiarism detection—and in historical authorship investigations—the language researcher may have to analyse texts of similar length, in authorship identification, text length may vary considerably from text to text, and be significantly reduced in a judicial case. Hence, the analysis of the richness of the vocabulary used by the questioned author may sometimes be of little use.

CopyCatch Gold may report similarity but makes no judgements on the reasons for the similarity. In this case, quantitative similarity may be used as supportive evidence for forensic authorship identification. But it may also be the case that the software does not find enough evidence to support similarity between two sets of texts. In this case, the findings may not be reliable enough to rule out possible candidates for the authorship of disputed texts, since, as said above, an individual's writing style and text length may vary considerably from genre to genre, and from register to register.

V.3. SIGNATURE STYLOMETRIC SYSTEM V 1.0

Signature Stylometric System v 1.0 is freeware for educational use designed by Peter Millican (University of Leeds). The aim of this tool is to facilitate stylometric analysis, with special emphasis on authorship identification. More specifically, the tool enables the language researcher to compare the styles of different writers, as well as to analyse disputed literary texts and explore authorship identification.

Signature Stylometric System makes it possible to load different files at the same time and build up a single corpus. Additionally, single texts can be divided into halves.

As the reader can see from Figure 3 below, *Signature Stylometric System* projects two-dimensional or three-dimensional graphs with the results (percentages or absolute frequencies) corresponding to the measurement of style markers such as word length, sentence length, paragraph length, the number of letters, and the number of punctuation marks for the selected files.

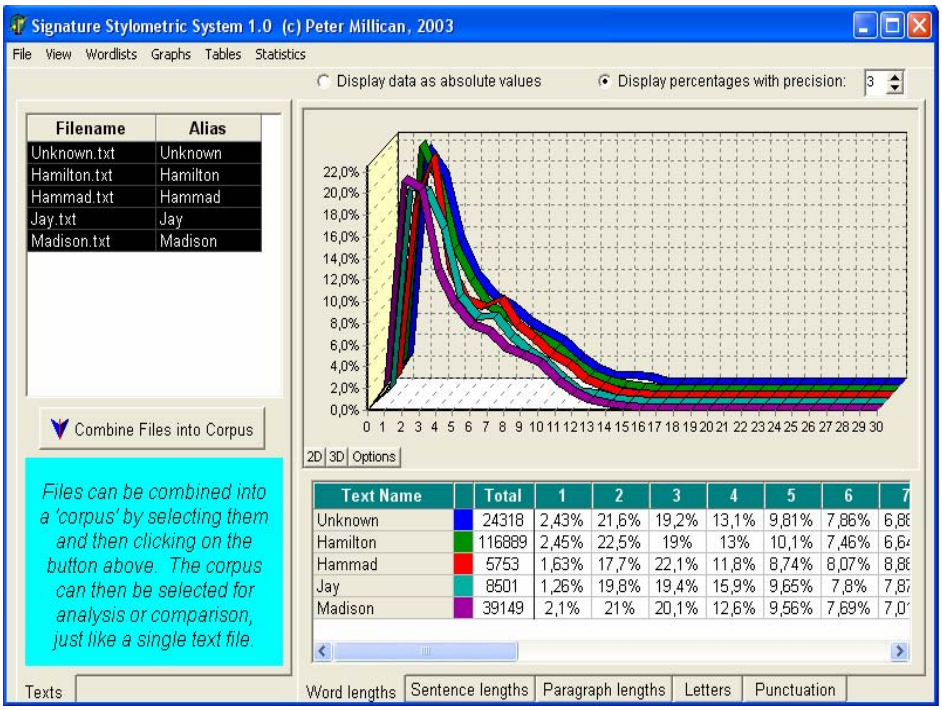


Figure 3: Sample graph displaying percentages in *Signature Stylometric System* v 1.0

Furthermore, *Signature Stylometric System* has a *Statistics* option that performs a Chi-square significance test. This test is used to evaluate relative homogeneity of multiple variables expressed as actual frequencies in various questioned writings.

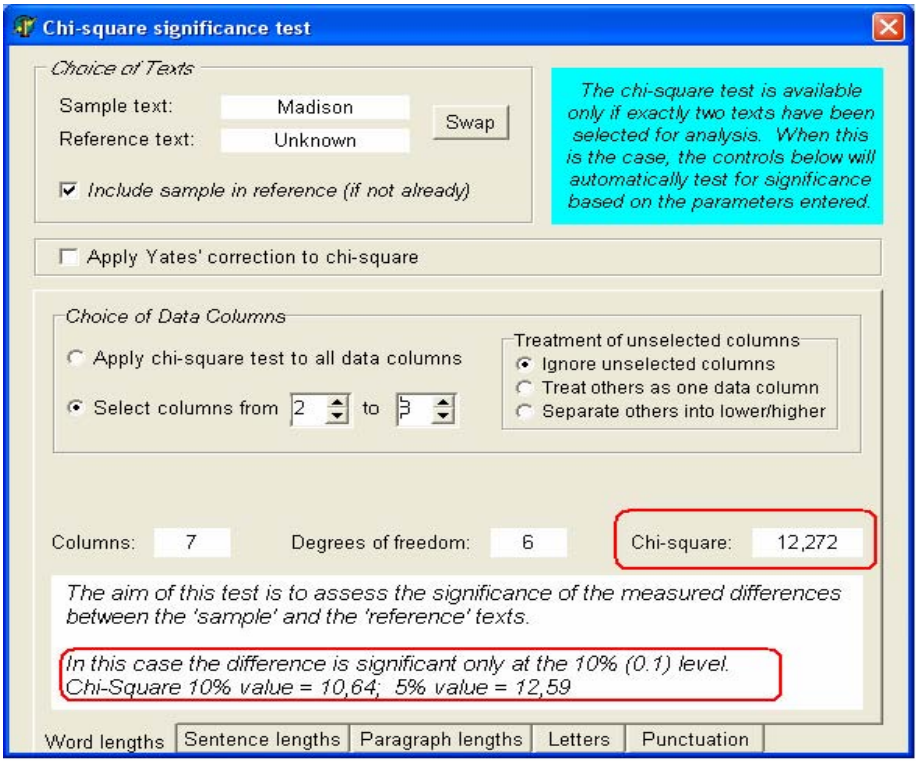


Figure 4: Sample of Chi-Square Significance Test in *Signature Stylometric System* v 1.0

As Figure 4 portrays, the actual Chi-square value is shown in the upper red circle, while the sentence in the lower red circle gives the user an idea about the significance of the difference found between two texts, and displays the standard value against which the actual value can be compared. Since the frequency of language patterns in texts may have more variation than the scientific phenomena for which statistical tests are commonly used, linguists must exercise caution in dealing with the results obtained.

Signature Stylometric System allows the user to introduce a new list apart from the available *Wordlist* or indicate which words (keywords) are the most useful for a given case of authorship identification under the *Wordlists* menu.

V.3.1. Strengths

On considering the strengths of *Signature Stylometric System* for forensic authorship identification, the following features emerge:

Firstly, files can be made into a single text. This facility is particularly useful when dealing with short texts, which is the usual case in forensic authorship identification.

Secondly, this software does not make use of lexically based measures exclusively but rather examines other relevant style markers such as letter and punctuation frequencies.

Thirdly, the analytical approach provides full comparison of all the results obtained and graphic output.

Fourthly, the software package includes a Chi-square significance test to evaluate relative homogeneity of multiple variables expressed as actual frequencies in various questioned writings.

Fifthly, the tool provides the forensic language researcher with a remarkable sample of disputed texts for forensic linguistic analysis.

Lastly, *Signature Stylometric System* seems to be quite suitable for historical authorship investigations.

V.3.2. Limitations

Although the current version of *Signature Stylometric System* was released in May 2003, the tool seems to be in a stage of development; however, no new version has become available since then. Concerning the limitations, the forensic linguist may find at the present stage of development, we might suggest the following: (a) Statistics should be expanded so as to include linguistic correlation and appropriate graphic output, (b) word concordance and phrase recognition should be added to the already existing *Word search* facility, (c) text filtering mechanisms should be developed for removing unwanted textual artefacts, (d) the tool should be adapted to non-standard alphabets and punctuation, (e) Unicode could be included to enable texts to be processed and displayed appropriately in a wide variety of languages, (f) a user manual should be available in different languages—presently there is

only a PowerPoint slide show in English, and (g) the *Text display* facility could be improved.

V.4. WORDSMITH TOOLS V 4.0

WordSmith Tools (WST) v 4.0 —the new version 5.0 will come out soon—is a suite of computer programs developed by Mike Scott (University of Liverpool) and may be defined as a quantitative analysis program for exploring the way grammatical and lexical features behave in their natural setting, namely the text.

The current version of WST (Figure 5) consists of three core tools (*Concord, KeyWords*, and *Wordlist*) and eight utilities. Within each of these tools there are different instruments and functions for analyzing texts and getting statistical support.

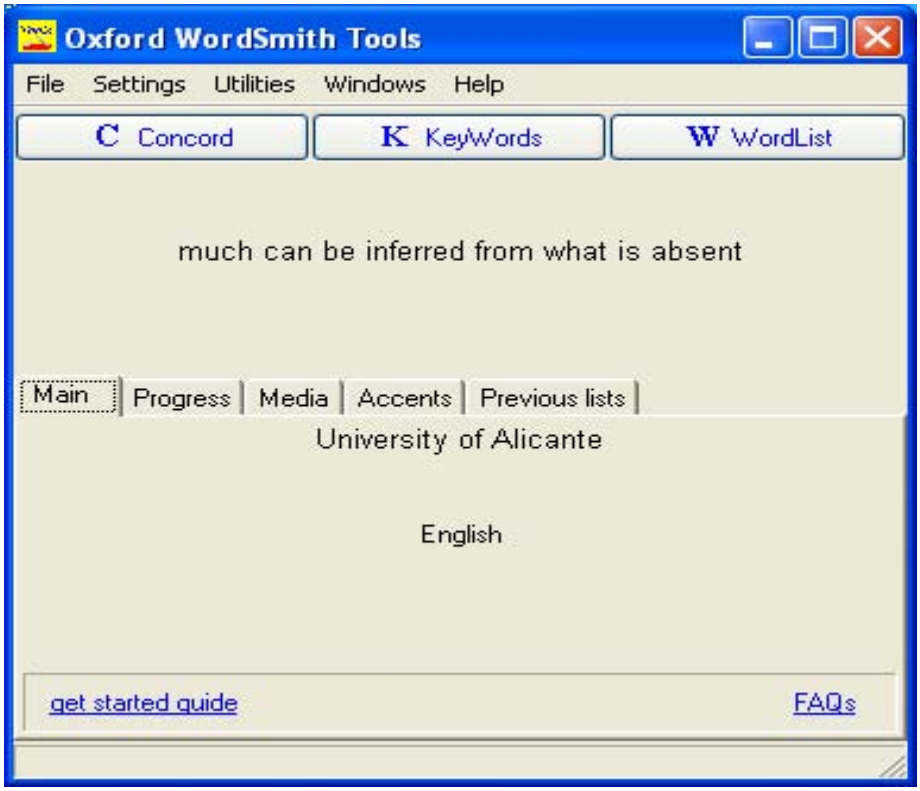


Figure 5: Initial screen of *WordSmith Tools* v 4.0

The utilities complement the core tools mentioned. There is one intended for re-formatting the texts by doing search-and-replace operations (*Text Converter*); one to convert data from WST old formats to the current version (*Data Converter*); one designed to split one text into smaller ones (*Splitter*); one to build up a corpus of texts downloading them directly from the Internet (*Webgetter*); there is one to view a text with words of interest highlighted and do sentence or paragraph alignment of two texts, which is extremely useful for comparing two versions of the same text (*Viewer and Aligner*); one intended for finding pairs of words which are minimally different from each other (*Minimal Pairs*); one for

selecting/changing the language of texts to be processed (*Languages Chooser*); and one to determine the frequencies of individual characters in text files (*Character Analyser*).

Next, we would like to present briefly the most outstanding characteristics and functions of *Wordlist, Concord*, and *Keywords*.

As its name suggests, *WordList* (Figure 6) generates word listings in alphabetical and frequency order, enabling the linguist to compare texts at a lexical level.

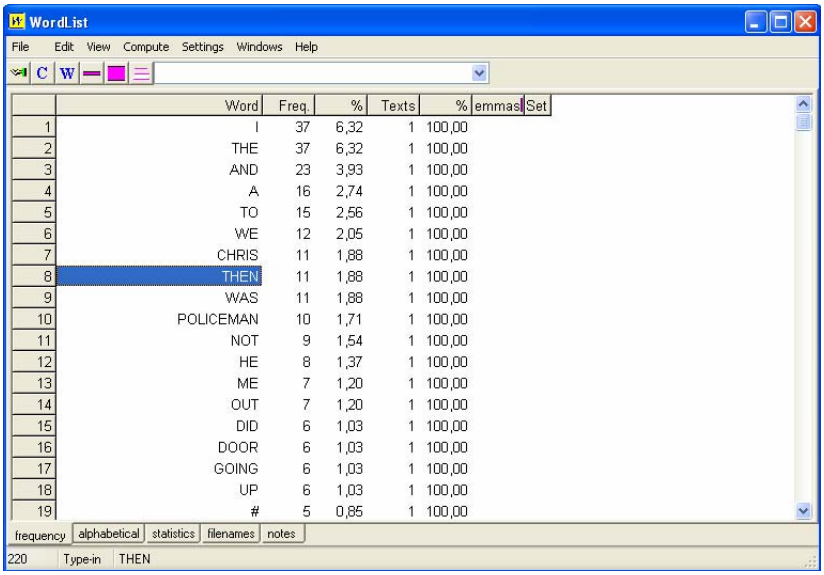


Figure 6: Screen of *WordList* showing the frequency list

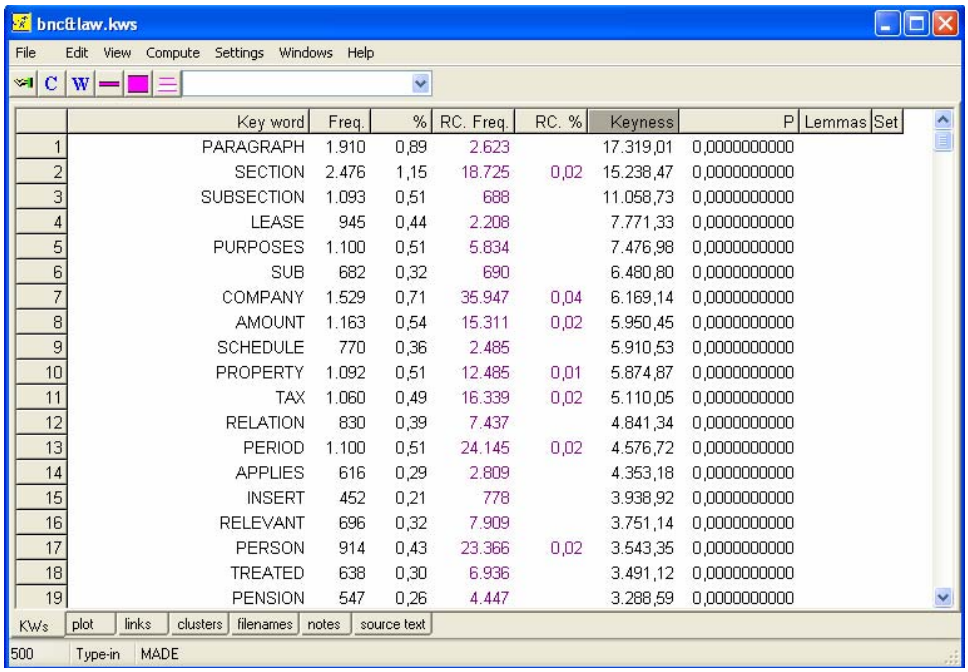
The *Statistics* tab provides information about the text(s)—size, number of running words (tokens), number of types (distinct words), type/token ratio (TTR), standardized TTR (STTR), length of words, number of sentences/paragraphs, among others. Needless to say, this information provides valuable help to measure a writer's individual style and to quantitatively contrast a set of texts. Regarding TTR, it is expressed as a percentage and obtained by dividing the total number of types (distinct words) by the total number of tokens (running words) in text(s). A high value means that the texts under study contain a high number of different tokens. In contrast, a low value implies a high number of word repetitions, which can be interpreted for authorship identification purposes by indicating that the text is less rich or varied in terms of lexical density. However, TTR is sensitive to the length of the text, and this is the reason why it is not the best method to be used when contrasting texts of different lengths, since a longer text may contain more word repetition and then its value could be lower. In contrast, the STTR calculates the TTR at regular intervals and is used to neutralize the influence the length of the text exerts when calculating the TTR; longer texts usually have more word repetition and, consequently, lower values are obtained. The STTR results in a higher average value as it does not count the repetition of words occurring in other parts of the text. This measure can be extremely useful when comparing lexical density across texts, since it can be used as an indicator of an individual's writing style.

The tool *Concord* (Figure 7) has been designed with a twofold purpose in mind. On the one hand, it generates concordances, namely a list which shows every instance of a given search word (also called *key word*, *base* or *query word*) along with its linguistic context (co-text). And on the other hand, it gives collocation information by implementing four association scores (MI, Z score, MI3, and Log-likelihood). The query word can be a single unit, part of it, or several units.

Figure 7: Screen of *Concord*

In forensic linguistics, particularly in authorship investigations (Collins, Kaufer, Vlachos, Butler & Ishizaki, 2004; Coulthard, 1993, 1994), concordances have proved their usefulness. By observing lexical concordances, the language researcher may analyze a word, part of a word, a group of words, a phrase, or expression, etc. in their linguistic context and consequently, discover such features as recurring lexical patterns, idiosyncratic usages of a word or expression, and word meaning.

The remaining tool to be described is *KeyWords* (Figure 8), whose function is to compare two word lists. One of these lists is considered to be the *reference corpus*, which is the baseline for comparison during analysis. The other has to be created with the text the linguist wants to investigate, that is the *study corpus*. The comparison between the two wordlists results in a new listing of keywords, namely words whose frequencies are significantly different in the study corpus in comparison with the reference corpus.



	Key word	Freq.	%	RC. Freq.	RC. %	Keyness	P	Lemmas	Set
1	PARAGRAPH	1.910	0,89	2.623		17.319,01	0,0000000000		
2	SECTION	2.476	1,15	18.725	0,02	15.238,47	0,0000000000		
3	SUBSECTION	1.093	0,51	688		11.058,73	0,0000000000		
4	LEASE	945	0,44	2.208		7.771,33	0,0000000000		
5	PURPOSES	1.100	0,51	5.834		7.476,98	0,0000000000		
6	SUB	682	0,32	690		6.480,80	0,0000000000		
7	COMPANY	1.529	0,71	35.947	0,04	6.169,14	0,0000000000		
8	AMOUNT	1.163	0,54	15.311	0,02	5.950,45	0,0000000000		
9	SCHEDULE	770	0,36	2.485		5.910,53	0,0000000000		
10	PROPERTY	1.092	0,51	12.485	0,01	5.874,87	0,0000000000		
11	TAX	1.060	0,49	16.339	0,02	5.110,05	0,0000000000		
12	RELATION	830	0,39	7.437		4.841,34	0,0000000000		
13	PERIOD	1.100	0,51	24.145	0,02	4.576,72	0,0000000000		
14	APPLIES	616	0,29	2.809		4.353,18	0,0000000000		
15	INSERT	452	0,21	778		3.938,92	0,0000000000		
16	RELEVANT	696	0,32	7.909		3.751,14	0,0000000000		
17	PERSON	914	0,43	23.366	0,02	3.543,35	0,0000000000		
18	TREATED	638	0,30	6.936		3.491,12	0,0000000000		
19	PENSION	547	0,26	4.447		3.288,59	0,0000000000		

Figure 8: Screen of *KeyWords*

To calculate the frequency of the keyword, the user can choose between two statistical measures: Chi-square and Log-likelihood. The latter, according to Scott (2003: 75), "gives a better estimate of keyness, especially when contrasting long texts or a whole genre against your reference corpus".

A keyword list has clear application for work in forensic linguistics since a general-purpose corpus (like BNC, Bank of English, etc.) may be used to establish norms of frequency and usage against which individual texts can be measured and contrasted, allowing the language researcher to accomplish a number of actions such as making a cross-register comparison, recognizing the lexical similarities or differences between two sets of texts, and discovering an author's stylistic preferences.

V.4.1. Strengths

WST is a powerful, reasonably priced tool for exploring language uses, and excels in the number of research features and functions it offers. Among its numerous strengths, we find its capacity to generate useful statistical data, to support different input formats, languages and character sets, to handle multiple tagged or untagged texts at the same time, and to export the results to different file formats, excel spreadsheet included. Moreover, pre-indexing of the texts is not necessary and different kinds of settings can be adjusted to fit a particular research interest.

The variety of ways the package provides for searching a corpus is also one of its strengths. Users may search for words, parts of them, strings, patterns and tags, as well as expand the context or call up the whole text at their convenience.

It should also be noted that albeit WST was not originally designed for forensic linguistics issues, it contains a wealth of features and functions that offer valuable, empirical output for statement analysis, historical authorship investigation, authorship identification, and plagiarism detection. As already mentioned, in forensic linguistics the descripion of written language is the first means of discovering, analyzing, and interpreting style. In this respect, WST provides the language expert with automatic identification of what forms—words, concordances, collocations and clusters are used by a writer, and how and why they are used. The second important issue in forensic linguistics is the measurement of style. WST allows automatic quantification of how much and how often forms are used by a writer and other relevant quantitative measures such as vocabulary richness.

V.4.2. Limitations

Bearing in mind that WST was not developed to work specifically in forensic authorship identification, it has few shortcomings that once overcome will render this versatile software extremely useful not only for forensic linguistics but also for other applied linguistics fields.

On the quantitative side, the lack of some measures for authorship identification is justified by the fact that it was not primarily designed for forensic purposes. On the qualitative side, WST's main weakness is found in its exclusive use of lexically based measures. The lexical orientation toward text analysis of WST may pose a problem to the forensic linguist, since, as mentioned above when evaluating *JVocalise* and *CopyCatch*, significant style markers appearing at other language levels may not be captured by the system.

In our view, WST would certainly be improved as a tool for forensic authorship identification, if it were able to preprocess the text(s) to be analyzed. By *preprocessing* we mean tagging. Along the lines of other dictionary-based tagging programs and parsers, tagging may enable the language expert to advance forensic text analysis, particularly the evaluation of an individual's set of unique linguistic choices and patterns. Conclusively, considering its powerful search and visualization possibilities, this tagging function could make WST a more influential tool in the field of forensic linguistics.

VI. CONCLUDING REMARKS

To the best of our knowledge, the vast majority of the tools that are available on the market today for forensic text analysts were not primarily designed to solve the types of cases involved in forensic authorship identification, the techniques and methods of which have been developing *ad casum* so far.

With the exception of *Signature Stylometric System*, none of the sampled tools reviewed in this article were originally designed for authorship identification purposes. However, they are extensively used by researchers in forensic linguistics, since they provide

language experts with the possibility of measuring linguistic variation and quantifying the writing and language data of cases, which is an essential requirement for providing scientific evidence at court today, especially in Common Law countries.

Despite considerable interest in forensic linguistics, we share the feeling that universities need to invest more resources in the design of specific tools to facilitate automatic authorship identification.

Hitherto most research has been focused on designing suites of programs for plagiarism detection, text genre detection, and attribution of disputed authorship of literary texts (Stamatatos, Fakotakis & Kokkinakis, 1999: 158-164).

The most important approaches to authorship attribution involve lexically based measures, and so a number of style markers have been proposed for measuring the richness of vocabulary used by the disputed author, namely content words, function words, the hapax legomena, the hapax dislegomena, the type-token ratio, etc.

However, as mentioned in part two of our discussion, an approach to automatic authorship identification should not only be based on lexical style markers, which have nevertheless proved to be quite reliable for plagiarism detection and authorship attribution of anonymous literary texts. After all, an individual's idiolect is made up of the unique set of linguistic choices that s/he makes at all linguistic levels.

Needless to say, no tool in the world, at least at present, would be able to provide an answer to the question: "Who is the author of this set of questioned texts?" But today, as has already been discussed in part one, recent changes in the criteria for admitting scientific evidence at court demand more than ever the quantification of an individual's style and measurable probability, especially in the USA.

The design of the "ideal tool" for forensic authorship identification would be possible provided that language researchers and software engineers put their heads together, to combine the best of both worlds, namely language description (qualitative analysis) and the quantification of an individual's style (quantitative analysis).

This ambitious enterprise would involve an interdisciplinary approach to the design of a suite of user-friendly programs. These should have interactive interfaces, available in different languages, that would enable linguists to conduct their linguistic research, and decide on the style markers that are not only recurrent in a set of questioned texts but also *prominent features* of an individual's writing style. These prominent features would ultimately serve to draw a *linguistic profile*.

The "ideal tool", as an extension of the language researcher (Hall, 1976: 25-40), should be able to carry out automatic search for style markers at all linguistic levels (text format, phonological, morpho-syntactical, lexico-semantical and discourse), assist the language expert in marking up prominent features in the text(s), and quantify the data and their distribution across sets of texts–the questioned texts and the known texts.

The "ideal tool" should facilitate statistical tests for the significance of variables without having to import the data produced by the software to a spreadsheet or other statistical packages, namely evaluation of potential relationship among variables expressed as means in comparison writings, standard error of difference (sD), t-Test (t), analysis of

variance (F), evaluation of potential relationship of variables expressed as percentages in comparison writings, and so on.

Finally, the "ideal tool" should be able to provide the measure of authorship discrimination (intra-author vs. interauthor variation in stylometry, principal component analysis, correspondence analysis, discriminant analysis, and multivariate analysis).

This article thus ends in an open-ended way, since it is not our purpose to overcome the difficulties found in forensic authorship identification, but rather to highlight the many problems which exist and which need to be addressed.

In an age in which interdisciplinary enquiry is becoming ever more necessary, forensic linguistics does indeed entail crossing many disciplinary borders to make real advances, and automatically involves us in interdisciplinary research. With this idea in mind, the authors of this article, a group of linguists and computer engineers at the University of Alicante, have embarked on the design of advanced software for forensic authorship attribution and identification. Hopefully our research in progress will be discussed in detail in a further article.

REFERENCES

Alcaraz-Varó, E. (2005). La lingüística legal: el uso, el abuso y la manipulación del lenguaje jurídico. In Mª T. Turell (Ed.), *Lingüística forense, lengua y derecho. Conceptos, métodos y aplicaciones*. Barcelona: IULA, pp. 49-66.

Cohen, L. J. (1977). *The Probable and the Provabable*. Oxford: Clarendon.

Collins, J., Kaufer, D., Vlachos, P., Butler, B. & Ishizaki, S. (2004). Detecting collaborations in text: Comparing the authors' rhetorical language choices in the Federalist Papers. *Computers and the Humanities*, vol. 38:1, 15-36.

Coulthard, M. (2005). Algunas aplicaciones forenses de la lingüística descriptiva. In Mª T. Turell (Ed.), *Lingüística forense, lengua y derecho. Conceptos, métodos y aplicaciones*. Barcelona: IULA, pp. 249-274.

Coulthard M. (1994): On the use of corpora in the analysis of forensic texts. *The International Journal of Speech Language and the Law*, vol 1:1, 27-43.

Coulthard, M. (1993). On beginning the study of forensic texts: corpus concordance collocation. In M. Hoey. (Ed.), *Data, Description and Discourse*: *Papers on the English Language in Honour of John McH Sinclair*. London: Harper Collins, pp. 86-97.

Coulthard, M. (Ed.) (1992). *Advances in Spoken Discourse Analysis*. London and New York: Routledge.

Coulthard, M. (1992). Forensic discourse analysis. In M. Coulthard (Ed.), *Advances in Spoken Discourse Analysis*. London and New York: Routledge, pp. 242-258.

EAGLES (1995). *Evaluation of Natural Language Processing Systems*. EAGLES document EAG-EWG-PR.2. Version of September, 1995. Available on line: http://www.issco.unige.ch/ewg95/ewg95.html

Gibbons, J. (2003). *Forensic Linguistics. An Introduction to Language in the Justice System*. Oxford: Blackwell.

Hall, E. T. (1976). *Beyond Culture*. USA: Doubleday Anchor Book.

McMenamin, G. R. (2004). Disputed authorship in US Law. *Speech, Language and the Law* , vol.11:1, 73-82.

McMenamin, G. R. (2002). *Forensic Linguistics. Advances in Forensic Stylistics*. Florida: CRC Press.

Olsson, J. (2004). *Forensic Linguistics. An Introduction to Language, Crime and the Law*. London and New York: Continuum.

Rieber, R. W. & Stewart, W. A. (Eds.) (1990). *The Language Scientist as Expert in the Legal Setting. Annals of the New York Academy of Sciences*. New Cork: The New York Academy of Sciences, vol. 606,1-135.

Rieber, R. W. & Stewart, W. A. (1990). The interaction of the language sciences and the Law. In R. W. Rieber & W. A. Stewart (Eds.), *The Language Scientist as Expert in the Legal Setting. Annals of the New York Academy of Sciences*. New Cork: The New York Academy of Sciences, pp. 1-4.

Scott, M. (2003). *WordSmith Tools version 4.0*. Oxford: Oxford University Press.

Shuy, R. W. (2005). La aportación de la lingüística al estudio de la intencionalidad criminal. In Mª T. Turell (Ed.), *Lingüística forense, lengua y derecho. Conceptos, métodos y aplicaciones*. Barcelona: IULA, pp. 19-48.

Stamatatos, E., Fakotakis, N. & Kokkinakis, G. (1999). Automatic authorship attribution. In EACL, 158-164. Available on line: http://acl.ldc.upenn.edu//E/E99-1021.pdf

Svartvik, J. (1968). *The Evans Statements*: *A Case for Forensic Linguistics*. Göteborg: University of Gothenburg Press.

Tanner, D. C. & Tanner, M. E. (2004): *Forensic Aspects of Speech Patterns: Voice Prints, Speaker Profiling, Lie and Intoxication Detection*. Arizona: Lawyers & Judges.

Turell, Mª T. (2006). Aplicaciones forenses de la lingüística descriptiva y de corpus. In M. Juan, M. Amengual, & J. Salazar (Eds.), *Lingüística aplicada en la sociedad de la información y la comunicación*. Universidad de las Islas Baleares: Servicio de Publicaciones y de Intercambio Científico, pp. 43-64.

Turell, Mª T. (Ed.) (2005). *Lingüística forense, lengua y derecho. Conceptos, métodos y aplicaciones*. Barcelona: IULA.

Woolls, D. (2003). Better tools for the trade and how to use them. *The International Journal of Speech Language and the Law: Forensic Linguistics*, vol.10:1,102-112.

APPENDIX: RESULT TEMPLATE (EVALUATION OF SOFTWARE FOR FORENSIC AUTHORSHIP IDENTIFICATION)

	Tool	JVocalyse	Copy Catch Gold	Signature Stylometric System	WordSmith Tools	Simple Concordance Program	Textanz	AntConc	Yoshikoder	Lexico3	T-LAB
	Version	2.05	2	1.0.	4.0.0.308	4.09 1997-2007 Alan Reed Build 10 (July 2007)	2.4.1.0	3.2.1 March 2007	v.0.6.3-preview.1 Build 13 October, 2006 Will Lowe	V3.45.0.1 2000-2002	T-LAB Pro 5.4 August 2007
	Type of software	commercial	commercial	freeware for educational use	Shareware. Full version: Single user: 75 € + VAT 10-users: 376€ + VAT ...	freeware	Shareware Full version: 17.42€ $22.95 US	freeware	freeware	Shareware Full version, licence simple: 600 €	Shareware Only full version program : 930€ Standard, 560 € Education
1. User interface	1.1 Installation routine (Wizard install shield?)	executable file	executable file	executable file or wizard	manual installation	Wizard	Wizard	executable file	executable file	Wizard	Wizard
	1.2 Type of user interface? (Menu-driven, GUI)	GUI	GUI	GUI	GUI	GUI	GUI	GUI	GUI	GUI	GUI
	1.3 Interface languages	English	English	English	English	English	English	English	English	French	English, Spanish, French, Italian
2. Product documentation and user help	2.1 Types of help resources available to users	No help resources available in the evaluated version	Help file, e-mail address	ppt file	Manual, blog, forum, e-mail address, webpage	Help dialog included	Help dialog included	txt file	Html	Html and pdf	Help dialog included
	2.2 Are help resources available in several languages?	x	English	English	English, French, German and Spanish	English	English	English	English	French, English	English, Spanish, French, Italian

	Tool	JVocalyse	Copy Catch Gold	Signature Stylometric System	WordSmith Tools	Simple Concordance Program	Textanz	AntConc	Yoshikoder	Lexico3	T-LAB
	2.3 Is appropriate documentation provided? (user manual, tutorials, demos, sample files, online help, etc.)	✓	✓	×	User manual, getting started manual, demo version, on line help, sample files	✓	✓	×	✓	✓	✓
3. User interface elements	3.1 Typed commands	×	×	×	×	×	×	×	×	×	×
	3.2 Clickable buttons	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	3.3 Function keys	×	×	✓	✓	✓	✓	✓	✓	✓	✓
	3.4 Traditional menus	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	3.5 Pull-down pop-up menus	×	×	✓	✓	✓	✓	✓	✓	✓	✓
	3.6 Dialog boxes	×	×	×	✓	×	×	×	×	×	×
	3.7 Icons	×	×	✓	✓	✓	✓	✓	✓	✓	✓
	3.8 Are there keyboard shortcuts available?	×	×	✓	✓	✓	✓	✓	✓	✓	✓
	3.9 Is mouse trackball required?	✓	✓	✓	✓	✓	✓	✓	✓	✓	
	4. Tool PageRank	6/10	6/10	4/10	5/10	3/10	4/10	5/10	3/10	5/10	4/10
5. Text selection and result presentation	5.1 The texts are necessarily taken one by one	×	×	×	×	×	✓	✓	×	✓	×
	5.2 The researcher may control the selection of texts	✓	✓	×	✓	×	×	×	×	✓	×
	5.3 The results are presented in a contextualized way	Not always	✓	×	✓	✓	×	✓	×	✓	×
	5.4 The results may be compared among sets of texts in the same screen	×	✓	×	×	×	×	×	×	×	×
	5.5 The results may be saved as rtf, htm, doc, txt	Some of them. The file format is assigned by the user. The resulting exported data needs to be processed.	Some of them as html and or rtf	×	Yes, as txt, xml, xls and native format	Yes, as rtf, htm and txt	(not available in trial version)	✓ (txt)	html o excel files	×	dat, html
Textual features on of style	6.1 Text format	html, rtf, txt, doc	html, rtf, txt, doc	txt	txt, html, sgml, xml	ans, asc, txt	txt	txt, htm, html, xml, ant	txt	txt	txt
	6.2 Numbers and symbols	×	×	×	✓	×	×	×	✓	×	×
	6.3 Abbreviations	×	×	×	✓	×	×	×	×	×	×

	Tool	JVocalyse	Copy Catch Gold	Signature Stylometric System	WordSmith Tools	Simple Concordance Program	Textanz	AntConc	Yoshikoder	Lexico3	T-LAB
	6.4 Punctuation	x	x	✓	✓	✓	x	✓	x	✓	x
	6.5 Capitalization	x	x	✓	✓	x	✓	✓	x	x	x
	6.6 Spelling (all the various kinds of patterned variants and mistakes)	x	x	x	x	x	✓	x	x	✓	✓
	6.7 Errors and corrections	x	x	x	x	x	x	x	x	x	x
	6.8 Word formation	x	x	x	x	x	x	x	x	x	x
	6.9 Content words	✓	x	x	✓	x	x	x	x	x	x
	6.10 Core words	✓	x	x	x	x	x	x	x	x	x
	6.11 Function words	✓	x	x	✓	x	x	x	x	x	x
	6.12 Sentences	x	✓	✓	✓	x	✓	x	x	x	x
	6.13 High-frequency words and phrases	✓	✓	x	✓	x	✓	x	x	✓	x
	6.14 Hoax hegomena (words only used once in the text)	✓	✓	x	✓	✓	✓	✓	✓	✓	✓
	6.15 Hapax dislegomena (words only used twice in the text)	✓	x	x	✓	✓	✓	✓	✓	✓	x
	6.16 Word strings	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	6.17 Word couplets	x	x	x	✓	✓	✓	✓	x	✓	✓
	6.18 Lexical variation (choices of words and phrases)	✓	x	x	✓	✓	✓	x	x	✓	✓
	6.19 Functional variation of language use (match between structure and function)	x	x	x	x	x	x	x	x	x	x
	6.20 Variation in syntax (sentence structure, coordination, subordination)	x	x	x	x	x	x	x	x	x	x
	6.21 Semantic variation (semantic features of words, phrases, and sentences)	✓	x	x	x	x	x	x	x	x	x
	6.22 Variation in discourse (quoted direct speech vs. indirect discourse, TV forms, reference to time and place, ordering of ideas, etc.	x	x	x	x	x	x	x	x	x	x
	6. 23 Interference features from other languages present in the language subject to analysis	x	x	x	x	x	x	x	x	x	x
	7.1 Tokens (total number of words in each text)	✓	x	✓	✓	✓	✓	x	x	✓	✓

Tool	JVocalyse	Copy Catch Gold	Signature Stylometric System	WordSmith Tools	Simple Concordance Program	Textanz	AntConc	Yoshikoder	Lexico3	T-LAB
7.2 Types (total number of different words used)	✓	×	×	✓	✓	✓	×	×	✓	✓
7.3 Core (number of content words included on the Core list)	✓	×	×	×	×	×	×	×	×	×
7.4 Token Type ratio (total number of words in the text divided by the vocabulary items used)	✓	×	×	✓	✓	✓	×	×	×	×
7.5 Vocabulary richness	✓	×	×	×	✓	✓	×	×	×	×
7.6 Mean sentence length	×	×	×	✓	×	✓	×	×	×	×
7.7 Mean paragraph length	×	×	×	✓	×	✓	×	×	×	×
7.8 Mean of 2-3 leter words	×	×	✓	✓	×	×	×	×	×	×
7.9 Mean of vowel-starting words	×	×	✓	×	×	×	×	×	×	×
8. Statistical tests for significance of variables										
8.1 Frequency distributions (function words, content words, etc.)	✓	✓	×	✓	✓	✓	✓	✓	✓	✓
8.2 Standard error of difference (sD)	×	×	×	×	×	×	×	×	×	×
8.3 t-Test (t)	×	×	×	×	×	×	×	×	×	×
8.4 Analysis of variance (F)	×	×	×	×	×	×	×	×	×	×
8.5 Proportion test (z)	×	×	×	×	×	×	×	×	×	×
8.6 Chi square (x^2)	×	×	✓	✓	×	×	×	×	×	✓
8.7 Coefficient of correlation (r)	×	×	×	×	×	×	×	×	×	×
8.9 Frequency estimates (P)	×	×	×	×	×	×	×	×	×	×
8.10 Olkin's likelihood ratio (lambda)	×	×	×	×	×	×	×	×	×	×
9. Measures of authorship discrimination	×	×	×	×	×	×	×	×	×	×

	Tool	JVocalyse	Copy Catch Gold	Signature Stylometric System	WordSmith Tools	Simple Concordance Program	Textanz	AntConc	Yoshikoder	Lexico3	T-LAB
	7.2 Types (total number of different words used)	✓	×	×	✓	✓	✓	×	×	✓	✓
	7.3 Core (number of content words included on the Core list)	✓	×	×	×	×	×	×	×	×	×
	7.4 Token/Type ratio (total number of words in the text divided by the vocabulary items used)	✓	×	×	✓	✓	✓	×	×	×	×
	7.5 Vocabulary richness	✓	×	×	×	✓	✓	×	×	×	×
	7.6 Mean sentence length	×	×	×	✓	×	✓	×	×	×	×
	7.7 Mean paragraph length	×	×	×	✓	×	✓	×	×	×	×
	7.8 Mean of 1-3 letter words	×	×	✓	✓	×	×	×	×	×	×
	7.9 Mean of vowel-starting words	×	×	✓	×	×	×	×	×	×	×
8. Statistical tests for significance of variables	8.1 Frequency distributions (function words, content words, etc.)	✓	✓	×	✓	✓	✓	✓	✓	✓	✓
	8.2 Standard error of difference (sD)	×	×	×	×	×	×	×	×	×	×
	8.3 t-Test (t)	×	×	×	×	×	×	×	×	×	×
	8.4 Analysis of variance (F)	×	×	×	×	×	×	×	×	×	×
	8.5 Proportion test (z)	×	×	×	×	×	×	×	×	×	×
	8.6 Chi square (x2)	×	×	✓	✓	×	×	×	×	×	✓
	8.7 Coefficient of correlation (r)	×	×	×	×	×	×	×	×	×	×
	8.9 Frequency estimates (P)	×	×	×	×	×	×	×	×	×	×
	8.10 Olkin's likelihood ratio (lambda)	×	×	×	×	×	×	×	×	×	×
9. Measures of authorship discrimination		×	×	×	×	×	×	×	×	×	×

