Applied Linguistics 25/4: 431–447

Author Identification, Idiolect, and Linguistic Uniqueness

MALCOLM COULTHARD

University of Birmingham, UK

For forty years linguists have talked about idiolect and the uniqueness of individual utterances. This article explores how far these two concepts can be used to answer certain questions about the authorship of written documents—for instance how similar can two student essays be before one begins to suspect plagiarism? The article examines two ways of measuring similarity: the proportion of shared vocabulary and the number and length of shared phrases, and illustrates with examples drawn from both actual criminal court cases and incidents of student plagiarism. The article ends by engaging with Solan and Tiersma's contribution to this volume and considering whether such forensic linguistic evidence would be acceptable in American courts as well as how it might successfully be presented to a lay audience.

INTRODUCTION

It is now over thirty-five years since Jan Svartvik published *The Evans Statements: A Case for Forensic Linguistics* (1968), in which he demonstrated that incriminating parts of a set of four linked statements—purportedly dictated to police officers by Timothy Evans and incriminating him in the killing of his wife and baby daughter—had a grammatical style measurably different from that of uncontested parts of the statements. It was later discovered that both victims had actually been murdered by Evans' landlord, John Christie. This marked the birth of a new discipline—the linguistic investigation of authorship for forensic purposes. Little more happened for a quarter of a century, with the notable exception of work by Roger Shuy in the USA (1993, 1998), but, during the past ten years, there has been a rapid growth in the frequency with which lawyers and courts in a number of countries have called upon the expertise of linguists in cases of disputed authorship. The texts examined range from questioned suicide notes, through anonymous letters, mobile phone text messages and contemporaneous police records of both interviews and confession statements, to the essays of students suspected of plagiarism.

IDIOLECT AND UNIQUENESS OF ENCODING

The linguist approaches the problem of questioned authorship from the theoretical position that every native speaker has their own distinct and individual version of the language they speak and write, their own *idiolect*, and the assumption that this *idiolect* will manifest itself through distinctive and idiosyncratic choices in texts (see Halliday *et al.* 1964: 75; Abercrombie 1969). Every speaker has a very large active vocabulary built up over many years, which will differ from the vocabularies others have similarly built up, not only in terms of actual items but also in preferences for selecting certain items rather than others. Thus, whereas in principle any speaker/writer can use any word at any time, speakers in fact tend to make typical and individuating co-selections of preferred words. This implies that it should be possible to devise a method of *linguistic fingerprinting*—in other words that the linguistic 'impressions' created by a given speaker/writer should be usable, just like a signature, to identify them. So far, however, practice is a long way behind theory and no one has even begun to speculate about how much and what kind of data would be needed to uniquely characterize an *idiolect*, nor how the data, once collected, would be analysed and stored; indeed work on the very much simpler task of identifying the linguistic characteristics or 'fingerprints' of whole *genres* is still in its infancy (Biber 1988, 1995; Stubbs 1996).

In reality, the concept of the linguistic fingerprint is an unhelpful, if not actually misleading metaphor, at least when used in the context of forensic investigations of authorship, because it leads us to imagine the creation of massive databanks consisting of representative linguistic samples (or summary analyses) of millions of idiolects, against which a given text could be matched and tested. In fact such an enterprise is, and for the foreseeable future will continue to be, impractical if not impossible. The value of the physical fingerprint is that every sample is both identical and exhaustive, that is, it contains all the necessary information for identification of an individual, whereas, by contrast, any linguistic sample, even a very large one, provides only very partial information about its creator's idiolect. This situation is compounded by the fact that many of the texts which the forensic linguist is asked to examine are very short indeed—most suicide notes and threatening letters, for example, are well under 200 words long and many consist of fewer than 100 words.

Nevertheless, the situation is not as bad as it might at first seem, because such texts are usually accompanied by information or clues which massively restrict the number of possible authors. Thus, the task of the linguistic detective is never one of identifying an author from millions of candidates on the basis of the linguistic evidence alone, but rather of selecting (or of course *de*selecting) one author from a very small number of candidates, usually fewer than a dozen and in many cases only two (Coulthard 1992, 1993, 1994a, 1994b, 1995, 1997; Eagleson, 1994).

An early and persuasive example of the forensic significance of idiolectal co-selection was the Unabomber case. Between 1978 and 1995, someone living in the USA, who referred to himself as FC, sent a series of bombs, on average once a year, through the post. At first there seemed to be no pattern, but after several years the FBI noticed that the victims seemed to be people working in

Universities and Airlines and so named the unknown individual the Unabomber. In 1995 six national publications received a 35,000-word manuscript, entitled *Industrial Society and its Future*, from someone claiming to be the Unabomber, along with an offer to stop sending bombs if the manuscript were published.[1]

In August 1995, the *Washington Post* published the manuscript as a supplement and three months later a man contacted the FBI with the observation that the document sounded as if it had been written by his brother, whom he had not seen for some ten years. He cited in particular the use of the phrase 'cool-headed logician' as being his brother's terminology, or in our terms an idiolectal preference, which he had noticed and remembered. The FBI traced and arrested the brother, who was living in a wooden cabin in Montana. They found a series of documents there and performed a linguistic analysis—one of the documents was a 300-word newspaper article on the same topic written a decade earlier. The FBI analysts claimed major linguistic similarities between the 35,000 and the 300 word documents: they shared a series of lexical and grammatical words and fixed phrases, which, the FBI argued, provided linguistic evidence of common authorship.

The defence contracted a linguist, who counter-argued that one could attach no significance to these shared items on the grounds that anyone can use any word at any time and therefore shared vocabulary can have no diagnostic significance. The linguist singled out twelve words and phrases for particular criticism, on the grounds that they were items that could be expected to occur in any text that was arguing a case—*at any rate*; *clearly*; *gotten*; *in practice*; *moreover*; *more or less*; *on the other hand*; *presumably*; *propaganda*; *thereabouts*; and words derived from the roots _argu* and propos*_. The FBI searched the internet, which in those days was a fraction of the size it is today, but even so they discovered some 3 million documents which included one or more of the twelve items. However, when they narrowed the search to documents which included instances of all the twelve items they found a mere 69 and, on closer inspection, every single one of these documents proved to be an internet version of the 35,000-word manifesto. This was a massive rejection of the defence expert's view of text creation as purely open choice, as well as a powerful example of the idiolectal habit of co-selection and an illustration of the consequent forensic possibilities that idiolectal co-selection affords for authorship attribution. The first example will be taken from the area of plagiarism detection.

ON DEFINING AND DETECTING PLAGIARISM

At its simplest, plagiarism, or more accurately the type of plagiarism we as linguists are competent to deal with, is the theft, or unacknowledged use, of text created by another. As my own university's website expresses it:

> Plagiarism is a form of cheating in which the student tries to pass off someone else's work as his or her own. . . . Typically, substantial

> passages are 'lifted' verbatim from a particular source without proper attribution having been made.
> *http://artsweb.bham.ac.uk/arthistory/declaration_of_aship.htm*

Any investigation of plagiarism is based consciously or unconsciously on a notion of *idiolect*. In other words it is expected that any two writers writing on the same topic, even if intending to express very similar meanings, will choose an overlapping, but by no means identical, set of lexico-grammatical items to do so. Indeed, and more importantly for some cases I will treat below, linguists from all persuasions subscribe to some version of the 'uniqueness of utterance' principle (Chomsky 1965; Halliday 1975) and so would expect that even the same person speaking/writing on the same topic on different occasions would make a different set of lexico-grammatical choices. It follows from this that, in any comparison of two texts, the more similar the set of items chosen, the greater the likelihood that one of the texts was derived, at least in part, from the other (or, of course, that both were derived from a third text), rather than being composed independently.

In most cases involving students there is little doubt about guilt, as these two examples of essay openings from Johnson (1997: 214) demonstrate—all items which student B 'shares' with student A are highlighted in bold:

> **A.** It is essential for all teachers to understand the history of Britain as a multi-racial, multi-cultural nation. Teachers, like anyone else, can be influenced by age old myths and beliefs. However, it is only by having an understanding of the past that we can begin to comprehend the present.

> **B.** In order for **teachers** to competently acknowledge the ethnic minority, **it is essential to understand the history of Britain as a multi-racial, multi-cultural nation. Teachers** are prone to believe popular **myths and beliefs; however, it is only by understanding** and appreciating **past** theories **that we can begin** to anticipate **the present**.

Even these short extracts provide enough evidence of shared items to question the originality of at least one of the essays, or both, of course, if a third text later proves to be the common source. When this level of sharing is also instanced in other parts of the same texts there is no room for doubt or dispute. The case of essay C, however, is not as clear-cut (items which C shares with one or both of essays A and B are highlighted):

> **C.** It is very important for us as educators to realise that **Britain as a nation** has become both **multi-racial** and **multi-cultural.** Clearly it is vital for **teachers** and associate teachers to ensure that **popular myths and** stereotypes held by the wider community do not **influence** their teaching. By examining British history this will assist our **understanding** and in that way be better equipped to deal with **the present** and the future.

Even though there is still quite a lot of shared lexical material here, it is evident that the largest identical sequences are a mere three running words long. Even so, one would still want to categorize this degree of lexical overlap, if instanced in other parts of the text, as unacknowledged, through more sophisticated borrowing and therefore as plagiarism, even if it does not fit easily within the Birmingham observation that 'Typically, *substantial* passages are "lifted"'. I will not discuss here the important question of whether a significant proportion of student written texts, which technically fall within the textual definition of plagiarism, are not the results of deliberate attempts to deceive at all, but rather a consequence of what is coming to be known as 'patchwriting', that is genuine but flawed attempts by students, who have somehow failed to acquire the academic rules for acknowledging textual borrowing, to incorporate the work of others into their own texts (see Pecorari 2002; Howard 1999).

Johnson's (1997) solution to the detection of this kind of student plagiarism or *collusion*, was to move away from using strings or sequences of items as diagnostic features and to focus instead on the percentage of shared individual lexical types and tokens as a better measure of derivativeness.[2] Intensive testing has shown that this measure of lexical overlap successfully separates those essays which share common vocabulary simply because they are writing on the same topic, from those which share much more vocabulary because one or more of them is derivative (see Woolls and Coulthard 1998). For example, in Johnson's study, whereas essays A, B, and C shared 72 different lexical types in their first 500 words, a set of three other essays from the same batch, whose authors had not colluded, shared only 13 lexical types, most of which were central to the topic under discussion. Further work (Woolls 2003) has shown that the most significant evidence is not the mere quantity of shared lexis, but rather the fact that, in the case of some shared items, both texts have both selected them and then only used them once. As such, 'once-only' items are, by definition, not central to the main concern of the text, otherwise they would have been used more frequently. The chances of two writers independently choosing several of the same words for single use are so remote as to be discountable.

If proof were needed of the distinctiveness and diagnostic power of words used once-only—*hapaxes* as they are technically labelled—it comes from successful internet searches in cases of suspected plagiarism. Experience confirms that the most economical method to use when checking the internet for suspected plagiarized text is to search using distinctive collocates whose individual items occur only once in the text in question. I will exemplify with the opening of a story written by an 11-year old girl:

The Soldiers (*all spelling as in the original*)

> Down in the country side an old couple husband and wife Brooklyn and Susan. When in one afternoon they were having tea they heard

> a drumming sound that was coming from down the lane. Brooklyn asks,
> 'What is that glorious sound which so thrills the ear?' when Susan replied in her o sweat voice
> 'Only the scarlet soldiers, dear,'
> The soldiers are coming, The soldiers are coming. Brooklyn is confused he doesn't no what is happening.
> Mr and Mrs Waters were still having their afternoon tea when suddenly a bright light was shinning trough the window.
> 'What is that bright light I see flashing so clear over the distance so brightly?' said Brooklyn sounding so amazed but Susan soon reassured him when she replied . . .

The first paragraph is unremarkable, but the second shifts dramatically, '*What is that glorious sound which so thrills the ear?*'. The story then moves back to the opening style, before shifting again to '*What is that bright light I see flashing so clear over the distance so brightly?*'. It is hard to believe that the same author could write in both styles and raises the question of whether the other borrowed text(s) might be available on the internet.

If one takes as search terms three pairs of collocated *hapaxes* 'thrills—ear', 'flashing—clear' and 'distance—brightly' one again sees the distinctiveness of idiolectal co-selection; the single pairing 'flashing—clear' yields over half a million hits on Google, but the three pairings together a mere 360 hits, of which the first thirteen are all from W. H. Auden's poem 'O What is that sound'. The poem's first line reads 'O what is that sound which so **thrills** the **ear**' while the beginning of the second verse is 'O what is that light I see **flashing** so **clear** Over the **distance brightly**, brightly?'. If one adds a seventh word and looks for the phrase 'flashing so clear', all of the hits return Auden's poem.

DO PEOPLE REPEAT THEMSELVES?

Whereas (occasional) identical strings in two texts which are supposed to have different authors can be indicative of 'borrowing' or theft, it is harder to argue the case when the texts are (supposedly) produced by the same author but on different occasions—even when there is no suggestion that the author had sight of the first text when s/he was producing the second. The example I want to use is from a famous English murder case, dating from 1978, where one piece of strongly contested evidence was a record of a police interview with a suspect.

In this case, four men were accused, and subsequently convicted, of killing a 13-year old newspaper delivery boy, Carl Bridgewater, solely on the basis of the confession of one of them, Patrick Molloy—there was no corroborating forensic evidence and Molloy retracted his confession, but to no avail. He admitted that he did actually say (most of) the words recorded in his confession, but insisted that he was being told what to say, while he was dictating the confession, by a policeman who was standing behind him. He also claimed that he had only made the confession after being physically and verbally abused for some considerable time.

The police, however, as support for the reliability of the confession, produced a contemporaneous handwritten written record of an interview which they claimed had taken place immediately before the confession and which contained substantially the same information expressed in the same language as the confession statement. Molloy denied that the interview had ever taken place—in his version of events he was being subjected to abuse at that time—and counter-claimed that the interview record had been made up later on the basis of the then pre-existing confession. As is evident from a cursory glance at the two extracts below taken, respectively, from the statement which Molloy admitted making and the interview record which he claimed was falsified, the similarities are enormous; I have highlighted them in bold. Most linguists would agree, on the basis of such similarities, that either one of the two documents was derived from the other or that both had been derived from a third. However, at the time of the original trial, no linguist was called to give evidence—in fact there were no forensic linguists practising in Britain at the time—so it was left to the lawyers to evaluate the linguistic significance of the interview and confession. As a result, the same phenomenon, massive identity in phrasing and lexical choice, was argued by the defence to be evidence of falsification, and by the prosecution to be evidence of the authenticity and reliability of both texts, on the grounds that here was an example of the accused recounting the same events, in essentially the same linguistic encoding, on two separate occasions.

Extract from Molloy's statement

> (17) **I had been drinking and cannot remember the exact time I was there but whilst I was upstairs I heard someone downstairs say be careful someone is coming.** (18) **I hid for a while and** *after a while* **I heard** *a* **bang** *come from downstairs.* (19) **I knew that it was a gun being fired.** (20) I went downstairs and **the three of them were still in the room.** (21) **They all looked shocked and were shouting at each other.** (22) **I heard Jimmy say, 'It went off by accident'.** (23) I looked and **on the settee** I saw the *body of the boy.* (24) **He had been shot in the head.** (25) **I was appalled and felt sick.**

Extract from disputed interview with Molloy

P. How long were you in there Pat?

(18) **I had been drinking and cannot remember the exact time that I was there, but whilst I was upstairs I heard someone downstairs say 'be careful someone is coming'.**

P. Did you hide?

(19) Yes **I hid for a while** and then **I heard** the **bang** I have told you about.

P. Carry on Pat?
(19a) I ran out.
P. What were the others doing?
(20) **The three of them were still in the room.**
P. What were they doing?
(21) **They all looked shocked and were shouting at each other.**
P. Who said what?
(22) **I heard Jimmy say 'it went off by accident'.**
P. Pat, I know this is upsetting but you appreciate that we must get to the bottom of this. Did you *see the boy's body*?
(Molloy hesitated, looked at me intently, and after a pause said,)
(23) Yes sir, he was **on the settee.**
P. Did you see any injury to him?
(Molloy stared at me again and said)
(24) Yes sir, **he had been shot in the head.**
P. What happened then?
(25) **I was appalled and felt sick.**

Both the prosecution assertion that identity of formulation in two separate texts is indicative of reliability and the apparent willingness of the lay jury to accept this assertion depend on two commonly held mistaken beliefs: first, that people can and do say the same thing in the same words on different occasions and secondly, that people can remember and reproduce verbatim what they and others have said on some earlier occasion. The former belief can be demonstrated to be false either by recording a person attempting to recount the same set of events on two separate occasions, or by simply asking a witness to repeat word for word what s/he has just said. The second belief used to have some empirical support, at least for short stretches of speech (see Keenan *et al.* 1977 and Bates *et al.* 1980), but was seriously questioned by Hjelmquist (1984), who demonstrated that, even after only a short delay, people could remember at best 25 per cent of the gist and 5 per cent of the actual wording of what had been said in a five minute two-party conversation in which they had participated. Confirmatory evidence about the inability to remember even quite short single utterances verbatim was specially commissioned from Professor Brian Clifford and presented at the 2003 'Glasgow Ice Cream Wars' Appeal. This was used to challenge successfully the claim of police officers that they had independently remembered, some of them for over an hour, verbatim and identically, utterances made by the accused at the time of arrest. Clifford's experiment tested the ability to remember a short, 24-word utterance and found that most people were able to recall verbatim no more than 30 to 40 percent of what they had heard.[3]

By the time of the Bridgewater Appeal in 1997, it was possible to provide extra supporting evidence of two kinds. First, as a direct result of Johnson's (1997) work on plagiarism discussed above, which demonstrated the significance of vocabulary overlap, an analysis was done of the shared vocabulary in the two Molloy texts. It became evident that the highlighting in the two Molloy extracts presented above actually understates the similarities between the two texts—a closer examination revealed that there was in fact not one single word in Molloy’s statement, neither lexical nor grammatical, which did not also occur in the interview record. I have only seen that degree of overlap on one other occasion, when two students had in fact submitted identical essays for assessment. Ironically, the computer analysis showed the degree of similarity to be only 97 percent—the 3 percent of different words made up of spelling errors produced by one of the two students.

In the Bridgewater case there also was secondary, supporting linguistic evidence, of a different kind, to support the claim that the interview record was both falsified and based on the statement. If we assume that the police officers had indeed, as Molloy claimed, set out to create a dialogue based on the monologue statement, they would have faced the major problem of what questions to invent in order to link forward and apparently elicit the pre-existing candidate answers which they had derived from the statement. In this scenario one would expect there to be occasions when a question did not fit successfully into the text into which it had been embedded—and indeed there are.

In a developing interview, a police question usually links backwards lexically, repeating word(s) from the previous answer. However, in creating a question to fit a pre-existing answer, there is always the danger that the question will only link forward. I will give two examples. The original statement has a two-sentence sequence—(21) ‘They all looked shocked and were shouting at each other.’ (22) ‘I heard Jimmy say, “it went off by accident”’—which appears word for word in the interview record, except that the two sentences are separated by the inserted question ‘Who said what?’. However, in this context the word ‘said’, although it is cataphorically unremarkable—*said* links with *say*—is anaphorically odd because the men have just been described as ‘shouting’. One would therefore have expected an anaphorically cohesive follow-up question to be either ‘What/Why were they *shouting*?’ or ‘Who was *shouting* (what)?’; one would certainly not predict ‘who *said* what?’. The choice of ‘said’ is a most unexpected choice—except of course for someone who knows that the next utterance will be ‘I heard Jimmy *say* . . .’—then it has an evident logic.

An example of a *grammatical* misfit is where the statement version ‘on the settee I saw the **body** of the **boy**. **He** had . . .’ is transformed into ‘Q. Did you see **the boy’s body**? Yes sir, **he** was on the settee’. The statement version correctly uses the pronoun ‘He’ because the referent is the ‘boy’ in ‘the body of the boy’, but the reformulated version in the police interview, ‘the boy’s **body**’, would be more likely to have elicited ‘**it**’ as a referent. We also find examples of process misfit: in the exchange reproduced below, the question ‘what happened’ requires a report of an action or an event, but in fact the response is a description of two states:

P. What **happened** then?
M. I **was appalled** and **felt sick**.

Had the reply been 'I vomited', it would, of course, have been cohesive. Similar process misfits are:

P. What were the others **doing**?
M. The three of them **were** still in the room.
P. What were they **doing**?
M. They all **looked shocked**

It is possible to continue in this vein, but I think these examples are sufficient to show that textual oddities like these support the claim, which was based on the identity of expression, that the interview record was created from the pre-existing statement. Sadly, it was not possible to test the acceptability and persuasiveness of these arguments in court, as the Crown conceded the appeal shortly before the due date, when compelling new evidence from document and handwriting analysts emerged to convince the judges of the unsafeness of the conviction.

THE EVIDENTIAL VALUE OF SINGLE IDENTICAL STRINGS

In the Bridgewater Four case there was a whole series of identical strings of words to support the claim that the interview record was derived from the statement, but for anyone unconvinced by the assertion that the identities were due to borrowing rather than identical encoding on two separate occasions, the claim of fabrication was supported by other linguistic evidence of a different and independent kind. The final questions I will address in this article are how much weight can one place on a single identical string and how significant is the length of a string when assessing its evidential significance? These questions go to the heart of current thinking about uniqueness in language production.

As Sinclair (1991) pointed out, there are two complementary assembly principles in the creation of utterances/sentences; one is the long accepted principle that sequences are generated word by word on an 'open choice' basis. When strings are created in this way, there is for each successive syntagmatic slot a large number of possible, grammatically acceptable, paradigmatic fillers and thus one can easily, if not effortlessly, generate memorable but meaningless sequences like 'colorless green ideas sleep furiously'. The other assembly principle proposed much more recently as a result of corpus work (Sinclair 1991), is the 'idiom principle', according to which pre-assembled chunks made up of frequent collocations and colligations are linked together to create larger units. In practice, both principles work side by side, which means that any given short string might be produced by either principle and therefore might be either an idiosyncratic combination or a frequently occurring fixed phrase. Nevertheless, the longer a sequence is, the more likely it is that at least some of its components have been created by the open choice principle and consequently, the less likely that the occurrence of this identical sequence in two different texts is a consequence of two speakers/writers coincidentally selecting the same chunk(s) by chance.

The data I will use for exemplificatory purposes come from the Appeal of Robert Brown in 2003. As in the Bridgewater Four case, here too there was a disputed statement and a disputed interview record; the difference was that Brown claimed that the statement was in reality a dialogue which had been represented as a monologue. He claimed that a police officer had asked questions to which Brown said he simply replied 'Yes' (Judge's Summing-up, p. 95 section E), and that, although the interview did occur, the record of it was made up afterwards—'no police officer took any notes' (Judge's Summing-up, p. 93 section E).

Below are two sentences from the statement matched with items occurring in the (invented) interview record:

(i)	Statement	I asked her if I could carry her bags she said 'Yes'
	Interview	I asked her if I could carry her bags and she said 'yes'
(ii)	Statement	I picked something up like an ornament
	Interview	I picked something up like an ornament

In what follows I have used examples from Google, rather than from a corpus such as the Bank of English or the British National Corpus, on the grounds that Google is accessible to the layperson for whom the argument is designed. While the above utterances/sentences may not seem remarkable in themselves, neither of them occurs even once in the hundreds of millions of texts that Google searches and even the component sequences quickly become rare occurrences:

String	Instances
I picked	1,060,000
I picked something	780
I picked something up	362
I picked something up like	1
I picked something up like an	0
an ornament	73,700
like an ornament	896
something like an ornament	2
I asked	2,170,000
I asked her	284,000
I asked her if	86,000
I asked her if I	10,400
I asked her if I could	7,770
I asked her if I could carry	7
I asked her if I could carry her	4
I asked her if I could carry her bags	0
if I could	2,370,000
if I could carry	1,600

It is evident that 'if I could' and perhaps 'I asked her' have the characteristics of pre-assembled idioms, but even then their co-selection in sequence is rare, at 7,770 occurrences. The moment one adds a 7th word, 'carry', the odds against it occurring become enormous, with a Google search yielding only 7 instances. Indeed rarity scores like these begin to look like the probability scores DNA experts proudly present in court. However, unlike the DNA expert, the linguist/expert has the disadvantage that everyone in the courtroom considers themselves to be a language expert and, among other things, 'knows' that they can remember what they and indeed others said in past conversations and feels confident that they can 'repeat' what was said verbatim at a later date. It will never be enough for the linguist to simply assert the uniqueness of encoding, it will need to be demonstrated in an accessible way.

AN ATTEMPT AT COURT PERSUASION

When faced with the problem of having to convince the Appeal Court judges in open court in the *R vs. Robert Brown* Appeal, I prepared the following presentation, which I hoped would both interest and persuade the three judges of the evidentiary strength of the identical formulations discussed above. I should say that, as in the Bridgewater case, this was not the only linguistic evidence to support Brown's claims about the unreliability of the police records.

As a first step I used Google to find out something about other cases involving Lord Justice Rose, who was to preside. The first three citations for the words 'Lord Justice Rose Appeal' were about an appeal against conviction for perjury by a famous British politician, Lord Archer. The first hit was:

> Guardian Unlimited | Special reports | Archer loses **appeal** bid
> . . . was not present at today's hearing, had his application for permission to **appeal** against the conviction rejected within hours. **Lord Justice Rose**, sitting with . . .
> (*www.guardian.co.uk/archer/article 0,2763,759829,00.html—30k*)

I accessed this citation, part of which is reproduced below as 'Guardian Extract', and from it selected the first phrase quoted from Lord Rose—'For reasons we will give later in the day'—which I have highlighted in bold. Given the nature of Appeal Court judgments, which are often released after the decision has been announced, this seemed an unremarkable phrase and yet a Google search returned only 7 occurrences of the phrase—all of which, on closer examination, proved not only to be attributed to Lord Rose, but were all in fact different reportings of the same uttering at the end of the Archer appeal. Even reducing the phrase to the apparently less specific 6-word utterance 'For reasons we will give later' only produced two more examples, this time not uttered by Lord Rose. Thus, here was an example of the uniqueness of an apparently ordinary utterance by Lord Rose himself.

Guardian extract

> **Archer loses appeal bid** Lord Justice Rose, sitting with Mr Justice Colman and Mr Justice Stanley Burnton in London, told Archer's QC Nicholas Purnell: '**For reasons we will give later in the day** we are against you in relation to conviction.'
>
> At the start of the hearing Nicholas Purnell QC, outlining the grounds of appeal, said: 'The submission that we make on behalf of Lord Archer is that **the first and fundamental ground** which interconnects with all the other grounds of appeal was that the learned trial judge wrongly exercised his discretion not to sever the trial of Edward Francis.'
>
> Mr Purnell said the decision of the judge, Mr Justice Potts, not to sever the trial of Francis had an '**unbalancing effect on the equilibrium**' of the trial.
>
> Counsel argued that Mr Francis was 'in a position effectively as a **substitute prosecution witness** and a substitute prosecutor'.

I then took three shorter phrases quoted in the same article, this time from Nicholas Purnell, Lord Archer's QC, each of them seeming, at least to this lay reader, to be equally unremarkable phrases for a lawyer to utter, *the first and fundamental ground, unbalancing effect on the equilibrium,* and *a substitute prosecution witness*. For these phrases I found 7, 10 and 4 instances respectively, but again for every phrase all the instances were reports of the same single occasion of uttering by Mr Purnell.

Armed with these examples, taken from an audience-relevant text, I hoped to convince the judges that uniqueness of utterance was a demonstrable fact. Sadly, at a pre-trial case conference, the defence barrister who had chosen to call me as an expert witness smiled indulgently and described my intended presentation as *whimsical* and decided not to use it. Fortunately, my other linguistic evidence was presented to the judges and accepted and the appeal was granted.

AUTHOR IDENTIFICATION IN AMERICAN COURTS

While the analytical techniques and arguments and the derived opinions reported above would be and in some cases already have been accepted in British courts, the situation is less clear in the USA. In a recent article entitled 'The linguist on the witness stand: forensic linguistics in American courts', Tiersma and Solan (2002) noted that:

> although the [American] legal system has often welcomed linguistic expertise, there are a number of areas in which they are more hesitant to do so. One example is the use of linguistics to identify authors. (Tiersma and Solan 2002: 229)

Tiersma and Solan cited the rigorous demands of the American legal system's Daubert criteria, which, in their opinion, many authorship identification methods fail to meet. The Daubert criteria were created in a Supreme Court ruling at the end of an appeal in the case of *Daubert vs. Merrell Dow Pharmaceuticals, Inc.* (509 U.S. 579 (1993)). Essentially the argument was over whether expert evidence could be rejected on the grounds that the experts involved had not published their work. In their ruling the Supreme Court observed that 'the adjective "scientific" implies a grounding in the methods and procedures of science' and then went on to propose four criteria with which to evaluate the 'scientific-ness' of a method:

1 whether the theory offered has been tested;
2 whether it has been subjected to peer review and publication;
3 the known rate of error; and
4 whether the theory is generally accepted in the scientific community. (509 U.S. 593) quoted in Tiersma and Solan (2002).

There is an extensive and lucid discussion of the Daubert criteria as applied to linguistic evidence in the article by Solan and Tiersma in this issue (pp. 448–65).

In one sense, Tiersma and Solan are raising purely American problems, because in the British and Australian legal systems it is the expert rather than the method that is recognized, so these courts can and do allow opinion evidence from anyone considered to have:

> specialised knowledge based on . . . training, study or experience [provided that the opinion is] wholly or substantially based on that knowledge. (Evidence Act 1995 Sec 79)

However, knowing that their evidence would also satisfy the Daubert criteria gives extra confidence to British linguist/experts and it is quite conceivable that similar criteria might be introduced into British Courts at some point in the future, even if only piecemeal, as the result of individual judgments. A court in Northern Ireland, for example, has recently ruled that forensic phonetic evidence based solely on auditory analysis, that is with no acoustic support, is no longer permitted.

Although I await the views of lawyer-linguists Solan and Tiersma with interest if not with some trepidation, I would argue that the methods of author attribution discussed above do meet the four Daubert criteria:

1. Whether the theory offered has been tested

Work by many people on a large number of cases has shown that there is no longer any dispute that the occurrence of shared identical items is conclusive evidence that two texts have not been independently created; what remains to be agreed is how few shared identical items are necessary to support a decision.

2. Whether [the theory] has been subjected to peer review and publication

Publications like this one and those by Johnson (1997), Woolls and Coulthard (1998) and Woolls (2003) have been subjected to peer review, there have been many presentations on this kind of authorship assignment at international conferences and to peer audiences in universities worldwide. In addition, the Copycatch Gold collusion detection program is in use internationally in over 50 universities, including the British Open University.

3. The known rate of error

This is perhaps the most difficult criterion—in cases of plagiarism it is traditional to err on the side of caution, and so I know of no cases of error. However, in this article, in the final section, I have taken the extreme position that a single and relatively short string can be conclusive evidence; this in itself is a challenge to the academic community to test the error rate and at the same time to fix an acceptable statistical equivalent for 'beyond reasonable doubt'.

4. Whether the theory is generally accepted in the scientific community

There is no doubt that the basic tenets of idiolectal variation and the uniqueness of utterance are generally accepted across the whole linguistic community; the disagreements are over how far certainty of assignment depends on the amount and kind of shared vocabulary and on the length of individual sequences and their composition in terms of idiomatic and open choice items.

CONCLUSION

The evidence discussed above suggests that the concepts of idiolect and uniqueness of utterance are robust and provide a basis for answering certain questions about authorship with a high degree of confidence. As demonstrated we can use the concepts to help us search when we suspect plagiarism and to categorize and classify when we already have texts of various kinds whose authorship is suspect or disputed. There are still many author identification problems where the methodology is less developed and reliable and where Solan and Tiersma's cautions are well-heeded, but the future for author identification is encouraging.

Final version received July 2004

NOTES

[1] For an accessible version of events, from someone who wrote a report on the language of the manuscript, see Foster (2001). The full text of the Unabomber manuscript is available at:
http://www.panix.com/~clays/Una/.

[2] An automated version of this analytic method, produced by Woolls (2002), is now available as the computer program *Copycatch Gold*.

[3] *http://news.bbc.co.uk/1/hi/scotland/3494401.stm*

REFERENCES

Abercrombie, D. 1969. 'Voice qualities' in N. N. Markel (ed.): *Psycholinguistics: An Introduction to the Study of Speech and Personality*. London: The Dorsey Press.

Bates, E., W. Kintsch, C. R. Fletcher, and **V. Giulani.** 1980. 'The role of pronominalisation and ellipsis in texts: Some memorisation experiments,' *Journal of Experimental Psychology: Human Learning and Memory* 6: 676–91.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. 1995. *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Chomsky N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MIT Press.

Clemit, P. and **D. Woolls.** 2001. 'Two new pamphlets by William Godwin: A case of computer-assisted authorship attribution,' *Studies in Bibliography*, 54: 265–84.

Coulthard, R. M. 1992. 'Forensic discourse analysis' in R. M. Coulthard (ed.): *Advances in Spoken Discourse Analysis*. London: Routledge, pp. 242–57.

Coulthard, R. M. 1993. 'Beginning the study of forensic texts: Corpus, concordance, collocation' in M. P. Hoey (ed.): *Data Description Discourse*. London: Harper Collins. pp. 86–97.

Coulthard, R. M. 1994a. '*Power*ful evidence for the defence: An exercise in forensic discourse analysis' in J. Gibbons (ed.): *Language and the Law*. London: Longman, pp. 414–42.

Coulthard, R. M. 1994b. 'On the use of corpora in the analysis of forensic texts,' *Forensic Linguistics: the International Journal of Speech, Language and the Law* 1/1: 27–43.

Coulthard, R. M. 1995. Questioning statements: Forensic applications of linguistics, text of inaugural lecture. Birmingham: English Language Research.

Coulthard, R. M. 1997. 'A failed appeal,' *Forensic Linguistics: the International Journal of Speech, Language and the Law* 4/ii: 287–302.

Eagleson, R. 1994. 'Forensic analysis of personal written text: A case study' in J. Gibbons (ed.): *Language and the Law*. London: Longman, pp. 362–73.

Foster, D. 2001. *Author Unknown: on the Trail of* Anonymous. London: Macmillan.

Halliday, M. A. K. 1975. *Learning how to Mean*. London: Edward Arnold.

Halliday, M. A. K., A. McIntosh, and **P. Strevens.** 1964. *The Linguistic Sciences and Language Teaching*. London: Longman.

Hjelmquist, E. 1984. 'Memory for conversations,' *Discourse Processes* 7: 321–36.

Howard, R. M. 1999. 'The new abolitionism comes to plagiarism' in L. Buranen d A. M. Roy (eds): *Perspectives on Plagiarism and Intellectual Property in a Postmodern World*. Albany: State University of New York Press, pp. 87–95.

Johnson, A. 1997. 'Textual kidnapping—a case of plagiarism among three student texts,' *Forensic Linguistics: The International Journal of Speech, Language and Law* 4/ii: 210–25.

Keenan, J. M., B. MacWhinney, and **D. Mayhew**, 1977. 'Pragmatics in memory: A study of natural conversation,' *Journal of Verbal Learning and Verbal Behavior* 16: 549–60.

McMenamin, G. 2004. 'Disputed authorship in US Law,' *International Journal of Speech Language and the Law* (formerly *Forensic Linguistics*) 11/1: 73–82.

Pecorari, D. E. 2002. Original Reproductions: An Investigation of the Source Use of Postgraduate Second Language Writers.

Unpublished Ph.D. thesis, University of Birmingham.

Shuy, R. 1993. *Language Crimes: The Use and Abuse of Language Evidence in the Courtroom*. Cambridge MA: Blackwell.

Shuy, R. 1998. *The Language of Confession, Interrogation and Deception*. London: Sage.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Solan, L. and **P. Tiersma.** 2004. 'Author identification in American courts,' *Applied Linguistics*, 25: 448–65.

Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.

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

Tiersma, P. and **Solan, L.** 2002. 'The linguist on the witness stand: Forensic linguistics in American courts,' *Language* 78: 221–39.

Woolls, D. 2002. *Copycatch Gold* a computerised plagiarism detection program.

Woolls, D. 2003. 'Better tools for the trade and how to use them,' *Forensic Linguistics. The International Journal of Speech, Language and Law* 10/i: 102–12.

Woolls, D. and **Coulthard, R. M.** 1998. Tools for the trade,' *Forensic Linguistics. The International Journal of Speech, Language and Law* 5/i: 33–57.