… and then …

Language Description and Author Attribution

> At the highest stratum of all there is the inter-penetration of minds. Each individual constructs his private linguistic universe, and through his utterances gives hints as to its nature.
> *Towards an Analysis of Discourse page 130*

Introduction

Since the mid-1980s I have been involved with forensic applications of authorship attribution, working to develop and refine a methodology. In what follows I will give examples from cases where it was possible to exploit techniques for the description of interaction, grammar and lexis that were developed in Birmingham by John Sinclair, his colleagues and his students.

Over the years many people have asked me which bits of *Towards an Analysis of Discourse* were written by me and which by John. I would often reply flippantly that mine were the bits that were easy to understand. When I came to prepare this lecture, I thought it would be interesting to end it with a comparison of John's style with my own – the results surprised me. In the Appendix are two longish extracts from the book; you might like to read them before proceeding any further and try to decide which is by John and which by me.

1. 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 Bloch 1948, Halliday et al 1964, 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, in fact they 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. No one has

even begun to speculate about how much and what kind of data would be needed to uniquely characterise an *idiolect*, nor how the data, once collected, would be stored and analysed. 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 under 200 words long and many contain 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 which massively restricts 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 (and, of course, sometimes *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, b, 1995, 1997, Eagleson, 1994).

An early and persuasive example of the forensic significance of idiolectal co-selection was the case of the Unabomber. Between 1978 and 1995, someone living in the United States, 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 for **Un**iversities and **A**irlines and so named the unknown individual the **Una**bomber. In 1995 six national publications received a 35,000 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. (For an accessible account written by someone involved in the case see Foster 2001.)

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, an ex-Berkeley University lecturer in mathematics, 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 log cabin in Montana. They impounded a series of documents and performed a linguistic analysis – one of the documents was a 300-word newspaper article on the same topic as the manuscript, written a decade earlier. The FBI analysts claimed there were major linguistic similarities between the 35,000 and the 300 word documents: they noted that they shared a series of lexical and grammatical words and fixed phrases, which, they argued, provided linguistic evidence of common authorship.

The defence contracted a distinguished linguist, who counter-argued that one could attach no significance to these shared items because anyone can use any word at any time and therefore shared vocabulary can have no diagnostic significance. She singled out twelve words and phrases for particular criticism, on the grounds that they were items likely 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**.

In response the FBI analysts searched the web, which in those days was only a fraction of its current size, 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 not one but all twelve of the items, they found a mere 69 and, on closer inspection, every single one of these proved to be a version of the 35,000 word manifesto. This was a massive rejection of the defence expert's view of text creation as purely free and open choice and a powerful illustration of the idiolectal habit of repeating co-selections. I will return to this topic in Section 4 below.

2. Hidden voices in monologue

In November 1952 two teenagers, Derek Bentley aged 19 and Chris Craig aged 16, were seen climbing up onto the roof of a London warehouse. The police surrounded the building and. three unarmed officers went up onto the roof to arrest them. Bentley immediately surrendered; Craig started shooting, wounding one policeman and killing a second. Bentley was jointly charged with his murder, even though he had been under arrest for some time when the officer was killed. The trial, which lasted only two days, took place five weeks later and both were found guilty. Craig, because he was legally a minor, was sentenced to life imprisonment; Bentley was sentenced to death and executed shortly afterwards. Bentley's family fought tenaciously to overturn the guilty verdict and were eventually successful 46 years later, in the summer of 1998. (The feature film *Let Him Have It, Chris,* released in 1991, gives a mainly accurate account.) The evidence which was the basis for both Bentley's conviction and the subsequent successful appeal was in large part linguistic.

In the original trial the problem for the Prosecution, in making the case against Bentley, was to demonstrate that he could indeed be guilty of murder despite being under arrest when the murder was committed. At this point it would be useful to read the statement which, it was claimed, Bentley dictated shortly after his arrest. It is presented in full below; the only changes I have introduced are the numbering of sentences for ease of reference and the highlighting, by underlining and bold, of items to which I will later refer.

Derek Bentley's Statement

(1) I have known Craig since I went to school. (2) We were stopped by our parents going out together, but we still continued going out with each other - I mean **we have not gone out** together until tonight. (3) I was watching television tonight (2 November 1952) and between 8 p.m. and 9 p.m. Craig called for me. (4) My mother answered the door and I heard her say that I was out. (5) I had been out earlier to the pictures and got home just after 7 p.m. (6) A little later Norman Parsley and Frank Fasey called. (7) **I did not answer the door or speak to them**. (8) My mother told me that they had called and I then ran out after them. (9) I walked up the road with them to the paper shop where I saw Craig standing. (10) We all talked together and then Norman Parsley and Frank Fazey left. (11) Chris Craig and I then caught a bus to Croydon. (12) We got off at West Croydon and then walked down the road where the toilets are - I think it is Tamworth Road.

(13) When we came to the place where you found me, Chris looked in the window. (14) There was a little iron gate at the side. (15) Chris then jumped over and I followed. (16) Chris then climbed up the drainpipe to the roof and I followed. (17) Up to then **Chris had not said anything.** (18) We both got out on to the flat roof at the top. (19) Then someone in a garden on the opposite side shone a torch up towards us. (20) Chris said: 'It's a copper, hide behind here.' (21) We hid behind a shelter arrangement on the roof. (22) We were there waiting for about ten minutes. (23) **I did not know** he was going to use the gun. (24) A plain clothes man climbed up the drainpipe and on to the roof. (25) The man said: 'I am a police officer - the place is surrounded.' (26) He caught hold of me and as we walked away Chris fired. (27) **There was nobody else** there at the time. (28) The policeman and I then went round a corner by a door. (29) A little later the door opened and a policeman in uniform came out. (30) Chris fired again then and this policeman fell down. (31) I could see that he was hurt as a lot of blood came from his forehead just above his nose. (32) The policeman dragged him round the corner behind the brickwork entrance to the door. (33) I remember I shouted something but I forgot what it was. (34**) I could not see** Chris when I shouted to him - he was behind a wall. (35) I heard some more policemen behind the door and the policeman with me said: '**I don't think** he has many more bullets left.' (36) Chris shouted 'Oh yes I have' and he fired again. (37) I think I heard him fire three times altogether. (38) The policeman then pushed me down the stairs and **I did not see** any more. (39) I knew we were going to break into the place. (40) **I did not know** what we were going to get - just anything that was going. (41**) I did not have** a gun and **I did not know** Chris had one until he shot. (42) I now know that the policeman in uniform that was shot is dead. (43) I should have mentioned that after the plain clothes policeman got up the drainpipe and arrested me, another policeman in uniform followed and I heard someone call him 'Mac'. (44) He was with us when the other policeman was killed.

Bentley's barrister spelled out for the jury the two necessary pre-conditions for them to convict: they must be "satisfied and sure",

> i) that [Bentley] knew Craig had a gun and
> ii) that he instigated or incited Craig to use it." (Trow p179)

The evidence adduced by the Prosecution to satisfy the jury on both points was linguistic. For point i) it was observed that in his statement, which purported to give his unaided account of the night's events, Bentley had said "I did not know he was going to use the gun", (sentence 23). In his summing up, the judge who, because of the importance of the case was the Lord Chief Justice, made great play with this sentence, telling the jury that its positioning in the narrative of events, before the time when there was a single policeman on the roof, combined with the choice of "*the* gun" (as opposed to "a gun") must imply that Bentley knew that Craig had a gun well before it was used. In other words "the gun", given its position in the statement, must be taken to mean "the gun I already knew that Craig had".

The evidence used to support point ii), that Bentley had instigated Craig to shoot, was from the police officers. In their written statements and in their verbal evidence in court, they asserted that Bentley had uttered the words "Let him have it, Chris" immediately before Craig had shot and killed the policeman. As the judge emphasised, the strength of the linguistic evidence depended essentially on the credibility of the police officers who had remembered it recorded it, written it down later and then sworn to its accuracy. When the case came to Appeal in 1998, one of the defence strategies was to challenge the reliability of Bentley's statement. If they could throw doubt on the veracity of the police, they could mitigate the incriminating force of both the statement and the phrase "Let him have it", which Bentley, supported by Craig, had vehemently denied uttering.

At the time of Bentley's arrest the police were allowed to collect verbal evidence from those accused of a crime in two ways: either *by interview*, when they were supposed to record contemporaneously, verbatim and in longhand, both their own questions and the replies they elicited, or *by statement*, when the accused was invited to write down, or, if s/he preferred, to dictate to a police officer, their version of events. During statement-taking the police officers were supposed not to ask substantive questions.

At trial three police officers swore on oath that Bentley's statement was the product of unaided monologue dictation, whereas Bentley asserted that it was, in part at least, the product of dialogue, and that police questions and his replies to them had been reported as monologue. There is no doubt that this procedure was sometimes used for producing statements. A senior police officer, involved in another murder case a year later, explained to the Court how he had himself elicited a statement from another accused in exactly this way:

I would say "Do you say on that Sunday you wore your shoes?" and he would say "Yes" and it would go down as "On that Sunday I wore my shoes" (Hannam 1953: 156)

There are many linguistic features which suggest that Bentley's statement is not, as claimed by the police, a verbatim record, see Coulthard (1993) for a detailed discussion; here we will focus only on evidence that the statement was indeed, at least in part, dialogue converted into monologue. Firstly, the final four sentences of the statement

> (39) I knew we were going to break into the place. (40) I did not know what we were going to get - just anything that was going. (41) I did not have a gun and I did not know Chris had one until he shot. (42) I now know that the policeman in uniform that was shot is dead.

form some kind of meta-narrative whose presence and form are most easily explained as the result of a series of clarificatory questions about Bentley's knowledge at particular points in the narrative. In searching for evidence of multiple voices elsewhere in the statement we must realise that there will always be some transformations of Q-A which will be indistinguishable from authentic dictated monologue. In the Hannam example quoted above, had we not been told that "On that Sunday I wore my shoes" was a reduction from a Q-A, we would have had some difficulty in deducing it, although the preposed adverbial 'On that Sunday' is certainly a little odd.

We can begin our search for clues with the initial observation that narratives, particularly narratives of murder, are essentially accounts of what happened and to a lesser extent what was known or perceived by the narrator and thus reports of what did **not** happen or was **not** known are rare and special. There is, after all, an infinite number of things that did not happen and thus the teller needs to have some special justification for reporting any of them to the listener, in other words there must be some evident or stated reason for them being newsworthy. It is interesting to remember in this context Halliday's work on the statistics of markedness, done while he was based at Cobuild in the early 90's, when he found that positive finite clauses were 8 times more likely to occur than negative clauses.

We can see typical examples of 'normal' usage of negative reports in the sentences below which are taken from a crucial confession statement in another famous case, that of the Bridgewater Four, which is discussed in more detail below:

> i) Micky dumped the property but **I didn't know where**.
> ii) Micky Hickey drove the van away, **I don't know where he went** to
> iii) **We didn't all go together**, me and Vinny walked down first.
> (Molloy's Statement)

In examples, i) and ii) the second negative clause functions as a *denial* of an inference which the listener could have reasonably derived from the first clause. Example iii) is similar, but this time it is a denial of an inference which the narrator guesses the listener might have made, as there is no textual basis for the inference. In other words

such negatives are an integral part of the ongoing narrative. We find examples of negatives being used in a similar way in Bentley's statement

(6) A little later Norman Parsley and Frank Fasey called.
(7) **I did not answer the door or speak to them**

When Bentley reported that his friends had called, the listener would reasonably expect him to have at least talked to them and therefore this is a very natural denial of a reasonable expectation.

However, there are some negatives in Bentley's statement which have no such narrative justification, like sentence (17) below:

(16) Chris then climbed up the drainpipe to the roof and I followed.
(17) Up to then **Chris had not said anything.**
(18) We both got out on to the flat roof at the top.

Chris is not reported as beginning to talk once they have got out onto the roof, nor is his silence contrasted with anyone else's talking, nor is it made significant in any other way later in the narrative. A similarly unwarranted negative is:

(26) He caught hold of me and as we walked away Chris fired**.**
(27) **There was nobody else** there at the time.
(28) The policeman and I then went round a corner by a door.

None of the possible inferences from this denial seem to make narrative sense here - i.e. that as a result of there being no one else there a) it must be the policeman that Craig was firing at, or b) that it must be Craig who was doing the firing, or c) that immediately afterwards there would be more people on the roof. So, the most reasonable explanation for the negatives in these two examples is that, at this point in the statement-taking process, a policeman asked a clarificatory question to which the answer was negative and the whole sequence was then recoded and recorded as a negative statement by Bentley. The fact that some of the statement may have been elicited in this way is of crucial importance in sentence (23):

(23) **I did not know** he was going to use the gun

This is the one singled out by the judge as incriminating. This sentence would only make narrative sense if it were linked backwards or forwards to the use of a gun - in other words if it has been placed immediately preceding or following the report of a shot. However, the actual context is:

(22) We were there waiting for about ten minutes.
(23) **I did not know** he was going to use the gun.
(24) A plain clothes man climbed up the drainpipe and on to the roof.

If it is accepted that there were question/answer sequences underlying Bentley's statement, it follows that the logic and the sequencing of the information were not under his direct control. Thus the placing of the reporting of some of the events must depend on a decision by the police questioner to ask his question at that point, rather than on Bentley's unaided reconstruction of the narrative sequence. Therefore, and crucially, this means that the inference drawn by the judge in his summing up about Bentley's prior knowledge of Craig's gun was totally unjustified - if the sentence is the product of a response to a question, with its placing determined by the interrogating police officers, there is no longer any conflict with Bentley's later denial "I did not know Chris had one [a gun] until he shot". Nor is there any significance either to be attached to Bentley saying "the gun". All interaction uses language loosely and co-operatively and so, if the policeman had asked Bentley about "the gun", Bentley would have assumed they both knew which gun they were talking about. In that context the sensible interpretation would be 'the gun that had been used earlier that evening' and not 'the gun that was going to be used later' in the sequence of events that made up Bentley's own narrative of the evening.

3. Using corpus evidence

One of the marked features of Derek Bentley's confession is the frequent use of the word "then" in its temporal meaning - 11 occurrences in 588 words. This may not, at first, seem at all remarkable given that Bentley is reporting a series of sequential events and that one of the obvious requirements of a witness statement is accuracy about time. However, a cursory glance at a series of other witness statements showed that Bentley's usage of "then" was at the very least atypical, and thus a potential intrusion of a specific feature of policeman register deriving from a professional concern with the accurate recording of temporal sequence.

Two small corpora were used to test this hypothesis, the first composed of three ordinary witness statements, one from a woman involved in the Bentley case itself and two from men involved in another unrelated case, totalling some 930 words of text, the second composed of statements by three police officers, two of whom were involved in the Bentley case, the third in another unrelated case, totalling some 2270 words. The comparative results were startling: whereas in the ordinary witness statements there is only one occurrence, "then" occurs 29 times in the police officers' statements, that is an average of once every 78 words. Thus, Bentley's usage of temporal "then", once every 53 words, groups his statement firmly with those produced by the police officers. In this case it was possible to check the findings from the 'ordinary witness' data against a reference corpus, the Corpus of Spoken English, a subset of the COBUILD Bank of English, which, at that time, consisted of some 1.5 million words. "Then" in all its meanings proved to occur a mere 3,164 times, that is only once every 500 words, which supported the representativeness of the witness data and the claimed specialness of the data from the police and Bentley, (cf Fox 1993).

What was perhaps even more striking about the Bentley statement was the frequent post-positioning of the “then”s, as can be seen in the two sample sentences below, selected from a total of 7:

> Chris **then** jumped over and I followed.
> Chris **then** climbed up the drainpipe to the roof and I followed.

The opening phrases have an odd feel, because not only do ordinary speakers use "then" much less frequently than policemen, they also use it in a structurally different way. For instance, in the COBUILD spoken data "then I" occurred ten times more frequently than "I then"; indeed the structure "I then" occurred a mere 9 times, in other words only once every 165,000 words. By contrast the phrase occurs 3 times in Bentley's short statement, once every 194 words, a frequency almost a thousand times greater. In addition, while the “I then” structure, as one might predict from the corpus data, did not occur at all in any of the three witness statements, there were 9 occurrences in one single 980 word police statement, as many as in the entire 1.5 million word spoken corpus. Thus, the structure "I then" does appear to be a feature of policeman's (written) register.

When we turn to look at yet another corpus, the shorthand verbatim record of the oral evidence given in court during the trial of Bentley and Craig, and choose one of the police officers at random, we find him using the structure twice in successive sentences, "shot him *then* between the eyes" and "he was *then* charged". In Bentley's oral evidence there are also two occurrences of “then”, but this time the "then"s occur in the normal preposed position: "and *then* the other people moved off", "and *then* we came back up". Even Mr. Cassels, one of the defence barristers, who might conceivably have been influenced by police reporting style, says "*Then* you". Thus these examples, embedded in Bentley’s statement, of the language of the police officers who had recorded it, added support to Bentley’s claim that it was a jointly authored document and so both removed the incriminating significance of the phrase “I didn’t know he was going to use the gun” and undermined the credibility of the police officers on whose word depended the evidential value of the claimed-to-be remembered utterance “Let him have it Chris”.

In August 1998, 46 years after the event, the then Lord Chief Justice, sitting with two senior colleagues, criticised his predecessor’s summing-up and allowed the Appeal against conviction.

4. Uniqueness of encoding, again

In 1979 four men were 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 subsequently retracted his confession, but to no avail. He admitted that he did actually say the words recorded in the confession, but insisted that he was being told what to say, 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, immediately beforehand.

The police, however, as support for the reliability of Molloy's confession, produced a handwritten contemporaneous record of an interview which, they claimed, had occurred immediately before the confession. It contained substantially the same information, expressed in the same language, as the confession statement. Molloy denied that this interview had ever taken place – in his version of events he was being subjected to abuse at that time. He counter-claimed that the interview record had been made up later on the basis of the by-then pre-existing confession. As is evident from a cursory glance at the two extracts below, the first from the statement which Molloy admitted making and the second from the interview record which he claimed was falsified, the similarities are striking; I have added sentence numbers and highlighted identical shared items in **bold** and close paraphrases in *italic*.

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**.

Linguists of all persuasions subscribe to some version of the 'uniqueness of utterance' principle and so would expect that even the same person speaking/writing on the same topic on different occasions would make an overlapping but different set of lexico-grammatical choices. Most linguists would also agree, on the basis of the number and length of the identical shared strings, 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 in Britain at the time – so it was left to the lawyers to evaluate the linguistic significance of the evident similarities between the interview and the 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.

The prosecution assertion that identity of formulation in two separate texts is to be expected and indicative of reliability depends on two commonly held mistaken beliefs: firstly, 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 indeed others have said on some earlier occasion. The former belief can be demonstrated to be false simply by recording someone attempting to recount the same set of events on two separate occasions. 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) and Hjelmquist and Gidlung (1985), who demonstrated that, even after only a short delay, people could remember at best 25 percent of the gist and 5 percent of the actual wording of what had been said in a five minute two-party conversation in which they had just participated.

Confirmatory evidence of 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, even when such a small stretch of language was involved, most people were able to recall verbatim no more than 30 to 40 percent of what they had heard. (for details see **http://news.bbc.co.uk/1/hi/scotland/3494401.stm)**
This confirmed that the only way in which these two Molloy extracts could have come to share so much vocabulary and phrasing would be if one had been derived from the other or both from a third text. Sadly, it was not possible for me 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.

5. Coherence and cohesion in discourse

In the same Bridgewater Four case there was secondary, supporting linguistic evidence of a different kind to reinforce the opinion that the interview record was falsified and to demonstrate that it was derived from 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 actually pre-existing answers, which they had extracted from the statement. In this scenario one would expect there to be occasions when a question did not fit successfully, coherently and/or cohesively, into the text into which it had been embedded – and indeed there are.

In a developing interview, a question usually links backwards lexically, often 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 cohesive with the next utterance – "said" links with "say" in "I heard Jimmy say" – is odd in terms of coherence. The men have just been described as "shouting", so one would have expected a coherent follow-up question to be either 'What/Why were they *shouting*?" or "Who was *shouting* (what)?"; one would certainly not anticipate "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 "said" 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 "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 in the reformulated version in the police interview, "the boy's **body",** would be likely 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 these examples are sufficient to show that certain oddities of cohesion and coherence support the opinion that the interview record was falsified on the basis of the pre-existing statement.

6. Uniqueness of encoding yet again - the evidential value of single identical strings

At the time of this lecture, the University of Birmingham website carried the following observation on plagiarism:

> 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_ship.htm*

As is evident from the two extracts below, plagiarism may not be detectable if one is looking only for ‘substantial passages’. The sophisticated plagiarist may not reproduce even a single sentence word for word, but no one would dispute that the extract from the Mackay biography is derived from the Wall biography. As before **bold** is used to indicate identical words, *italic* to indicate close paraphrases.

Two Biographies of Andrew Carnegie

a. With all of these problems it was little short of a miracle that the “stichting” board *was ready to lay* the *corner***stone** for the building *in* **the summer of 1907** at **the opening of the Second Hague International Conference**. It then **took six** *more* **years** before **the Palace** *was completed* during which time there *continued to be squabbles* **over details**, *modifications of architectural plans and lengthy discussions about furnishings… For ten years the Temple of* **Peace** was *a storm of* **controversy**, **but** *at last***, on 28 August 1913, the** Grand **Opening ceremoni***es were held* . (J F Wall, *Andrew Carnegie)*

b. The *foundation* **stone** *was not laid until* **the summer of 1907**, *in nice time for* **the opening of the Second Hague International Conference**. Actual construction of the **palace took** *a further* **six years**, delayed and exacerbated by constant *bickering* **over details**, *specifications and materials. For an entire decade* the **Peace** *Palace* was *bedevilled by* **controversy, but** *finally,* **on 28 August 1913, the opening ceremony** *was performed***.**
(J Mackay, *Little Boss: A Life of Andrew Carnegie)*

Plagiarism detection raises the question of how unique is encoding and how little identical text does one need to claim that it was copied and not created independently. 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, and then

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. We must now ask 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 op cit), 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 have been 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 the same or two different speaker/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 only difference was that Brown claimed that, although the interview itself 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 set beside sentences occurring in the (?invented) interview record:

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"
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 an academic corpus like the Bank of English or the British National Corpus, on the grounds that Google is easily accessible to the laypeople, like judges and jury members, for whom

the argument was designed. While the above utterances/sentences may not seem remarkable in themselves, neither of them occurred even once in the billions 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
if I could	2,370,000
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

Focussing on the second pair of sentences, 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 the same sequence is rare, at 7,770 occurrences. The moment one adds a 7th word, "carry", the odds against these 7 running words occurring become enormous, with the 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 expert linguist has the disadvantage that everyone in the courtroom considers themself to be a language expert. It will never be enough for the linguist to simply assert the uniqueness of encoding, it will always need to be demonstrated in an accessible and persuasive way.

When I came, in April 2006, to produce this written version of my 2005 lecture, I decided it would be prudent to check my claim about the uniqueness of "I asked her if I could carry her bags" and rest assured that there were indeed no instances. To my horror this time Google found two examples.

However, as we are often told, it is the exception that proves the rule. Since Robert Brown's successful appeal a website devoted to his case has been set up, (http://www.eamonnoneill.net/Candp.html), where text of the confession now appears. But what about the second embarrassing citation? It is in an article I myself wrote about the case and made available to my students on a website. So the 9-word string is still unique, it's just that it has now been reproduced twice.

7. Uniqueness and internet plagiarism

If proof were still needed of the diagnostic power of idiolect, we can show it through focussing on distinctive collocations and can demonstrate their importance in successful internet searches for suspected plagiarism. Experience confirms that the most economical method to use, when checking via the internet, is to search by using 3 pairs of collocates whose individual items occur only once in the text in question. I will exemplify with the opening of a story written by a 12-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 style shifts dramatically in the second: *"What is that glorious sound which so thrills the ear?"*. The story then moves back to the style of the opening, before shifting again to *"What is that bright light I see flashing so clear over the distance so brightly.* The reader feels it is very unlikely that the same author could write in both styles and this 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 immediately sees the forensic power of idiolectal co-selection. The single pairing 'flashing-clear' yields over half a million hits on Google, but the three pairings together yield a mere 360 hits, of which the first thirteen, when I first searched, were 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 are from Auden's poem.

8. And so, my dear Watson, which text was written by Sinclair?[1]

The Holy Grail in authorship studies is to find valid and reliable markers which consistently distinguish authors. So far there has been more searching than finding. Grant (2005) tested over 170 markers of authorship proposed by others and found the vast majority of them wanting. One marker which does seem to work, however, is sentence length, as Winter and Woolls found in a pilot investigation in 1996. They were responding to a challenge made by the then Head of the School of English, Kelsey Thornton, to distinguish between the individual styles of two authors, who had jointly written a late-Victorian novel. Winter and Woolls were provided with the first 1,000 running words from each of the first five, and the last six chapters, (28-33), of the novel and also, for comparative purposes, 2,500 words from the beginning of a single author novel, written by one of two authors.

Winter had suggested that the frequency of lexical items which were used only once (often called *hapaxes*) – in other words the degree of lexical novelty and variety - might provide a discriminatory measure. Research by Holmes (1991) appeared to support this view, although his findings were based on significantly longer text samples. The question in this case was whether such a measure, labelled *lexical richness*, could provide results when applied to much shorter text extracts. The lexical richness score is derived from the relative frequency of *hapaxes* expressed as a function of the length of the text. Thus a greater proportion of once-only usage, results in a higher lexical richness score. In their investigation Winter and Woolls focused on both lexical richness and average sentence length.

Using the two measures together allowed Winter and Woolls to locate each text in a two dimensional space, with sentence length plotted against lexical richness. The results for the 1000 word extracts showed a marked difference between the lexical richness scores and an evident, though lesser, difference between the average sentence lengths of the odd numbered chapters 1, 3 and 5 and the even numbered chapters 2 and 4. This suggested that the two measures were identifying a real stylistic difference between the two authors. When the results for the final six chapters were added, chapter 32 was found to have lexical richness and sentence length scores comparable with those of chapters 2 and 4, while the scores for chapter 33 placed it close to those for chapters 1, 3 and 5 – see Chart 1 on the next page.

The scores for the remaining chapters, 28-31, fell in between the two groupings and this led Winter and Woolls to suggest that the two authors may have collaborated on these chapters, a hypothesis later confirmed by Kelsey Thornton, after consulting a diary written by one of the authors. Winter and Woolls then divided the 2,500 word extract from the single-author novel into three consecutive 835-word samples and labelled them CTLA, CTLB and CTLC. The scores for all three extracts were remarkably similar to each other and to those for chapter 32. It was therefore, correctly as it turned out, argued that all had been written by the same author.

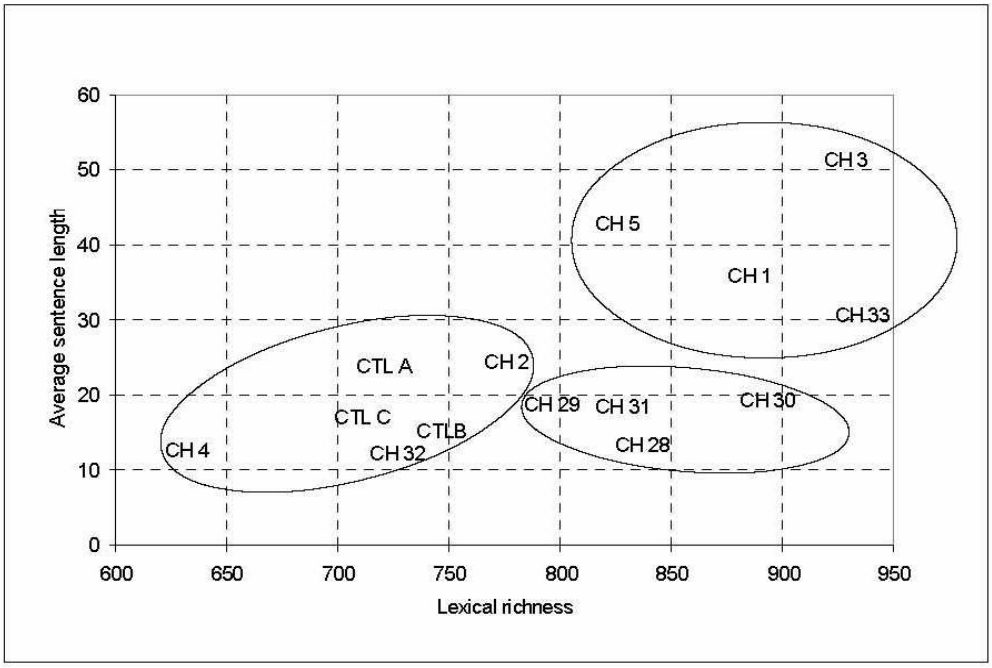

Chart 1 Sentence Length and Lexical Richness Scores. For Novel Extracts

We decided to test whether these two measures would correctly assign Texts A and B. We first we needed baseline scores for both authors on the two measures and to calculate them we used two publications from each author that had been written in the early 90's and then added for comparison two jointly authored texts written by Woolls, giving six texts in all. As is evident in Chart 2 below, Sinclair and Coulthard differed significantly in terms of sentence length, but not in terms of lexical richness:

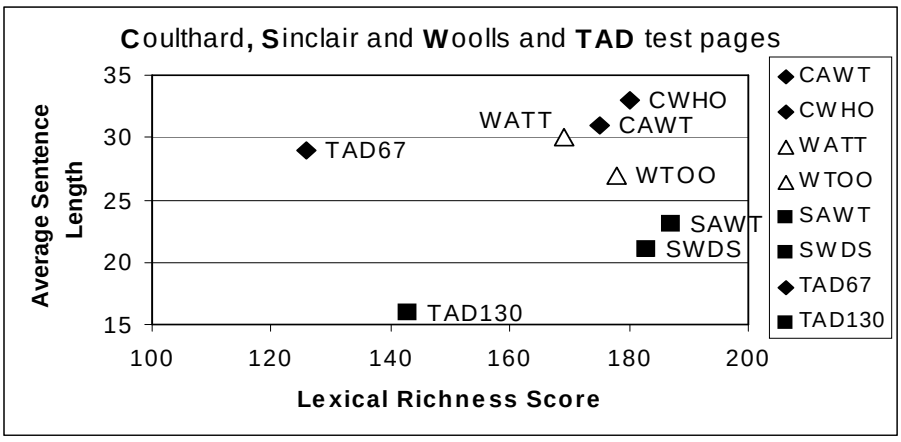

Chart 2 Sentence Length and Lexical Richness Scores for Sinclair and Coulthard

Key to labels in chart
CAWT is Coulthard 1994c; CWHO is Coulthard 2000a; WATT is Clemit and Woolls 2001; WTOO is Woolls and Coulthard 1998; SAWT is Sinclair 1992/1994; SWDS is Sinclair 1993

When the scores for Texts A and B were compared with those for the known texts, sentence length correctly grouped Text A with Sinclair and Text B with Coulthard, although the lexical richness score associated neither text with either author.

Although this analysis was undertaken in a light-hearted manner, it does raise serious and interesting questions. If Texts A and B are indeed representative of the styles of Sinclair and Coulthard in the early 70's, then, over a 20-year period, both have moved to writing longer and lexically richer sentences – although, of course, this trend may not continue. As I was writing and revising this text I became painfully aware of my own diagnostically significant long sentences and I must admit that I have consciously divided up any particularly long ones that I noticed at the revision stage.

In this spirit of adventure and academic curiosity we decided to apply a Woolls program called 'Citereader', (Woolls n.d.), to the two texts. This program was initially devised to identify acknowledged and unacknowledged citations in students' essays. It works on the assumption that there will normally be style differences between the embedded citation and the embedding text, as we saw in the Soldiers story discussed above, and that citations will have a less direct connection with the rest of the student's own writing and are likely to be more concise. Essentially the program allocates a score to each sentence based on the relationship of that sentence to the rest of the text and on the rarity and semantic complexity of the component words. This means that the same sentence, occurring in a different context, for instance as a citation, would almost certainly get a different rating. The score for any given sentence is a sum of the scores for each individual word. Grammatical words score lowest, then core lexical words, then lexical words which occur frequently in the text; infrequent and longer words are given a high rating, with the highest of all reserved for hapaxes. The scoring system is designed so that long sentences do not inevitably achieve higher scores, but only do so if they contain significant quantities of higher rated individual words. Short sentences with infrequent and complex vocabulary can also achieve a high score.

An unexpected finding from applying the Citereader program to a large number of texts was that the Citereader scores for different individual authors proved to be quite consistent across a range of texts. On reflection this is not so surprising because, following the Winter and Woolls findings for the jointly authored novel, one would expect authors to display individual style features fairly consistently, So, when a series of texts is put through the program, they tend to be grouped by author. Some authors have consistently more sentences with low scores, some have a significant grouping in the mid range, while yet others have a greater proportion of high scoring sentences. For that reason we decided to see what, if anything, Citereader would say about the Sinclair and Coulthard extracts, when compared with an analysis of the known texts.

The program is designed to assign all analysed sentences to one of 8 levels of complexity and, as we can see in Table 1 below, the styles of the three authors under consideration are clearly separated, particularly by the proportion of sentences falling

into the three lowest categories, 1-3, which have been grouped together, for exemplificatory purposes, in Table 1 below:

File	Words	% Levels 1 - 3	L1	L2	L3	L4	L5	L6	L7	L8
CAWT	4085	24%	33	6	7	17	14	5	13	32
CWHO	5827	25%	42	8	10	16	20	17	15	46
WATT	4700	34%	49	6	8	18	15	18	9	31
WTOO	6443	38%	77	12	12	32	25	21	18	36
SAWT	7184	57%	141	27	25	38	22	18	19	21
SWDS	7585	63%	197	19	27	27	25	19	10	21

Table 1 Citereader Analysis of the Coulthard, Sinclair and Woolls Texts

The major difficulty for any authorship analyst in a forensic context, as we noted above, is usually the shortness of the texts provided for analysis and it was for this reason that we chose shortish extracts from *TAD* for comparison purposes. We then took similar length short extracts from the complete Sinclair, Woolls and Coulthard texts already analysed above and compared them first. Even though there were now only a few sentences on which to base the comparison, the texts were still clearly separated. When we added in the Sinclair and Coulthard texts, as you can clearly see from Table 2 below, Text A was placed with Sinclair's SAWT extract and Text B with Coulthard's CAWT extract.

File	Words	% Levels 1 - 3	L1	L2	L3	L4	L5	L6	L7	L8
Text B	322	11%	0	1	0	2	2	0	0	6
CAWTp3	347	15%	0	1	0	1	1	0	1	5
WATT	405	29%	3	2	1	2	1	0	1	5
WTOO	334	37%	1	1	1	4	2	1	2	0
Text B	353	50%	6	2	3	5	2	2	0	1
SAWTp15	337	55%	5	0	2	4	3	1	1	1

Table 2 Citereader Analysis of the Coulthard, Sinclair and Woolls Extracts

Many forensic cases will not yield such clear categorisations as this, but this party trick, designed to entertain a group of John Sinclair's friends, may have produced results of great significance for forensic authorship analysis. Only further work will tell, but for the moment it may be wise not to place too much trust in the text.

NOTE

For this Section I have had a great deal of assistance from David Woolls of CFL Software Development, both with the original analyses and with their verbal and visual presentation.

Selected 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, Kintsch, W, Fletcher, C R, and Giulani, V. 1980 'The role of pronominalisation and ellipsis in texts: some memorisation experiments', *Journal of Experimental Psychology: Human Learning and Memory*, 6, 676-691.

Biber, D 1988 *Variation across Speech and Writing*, Cambridge, CUP.

Biber, D 1995 *Dimensions of Register Variation: a Cross-linguistic Comparison*, Cambridge, CUP.

Bloch, B 1948 'A set of Postulates for phonemic analysis', *Language*, 24, 3-46.

Clemit, P and Woolls, D 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, 242-57.

Coulthard, R M 1993 'Beginning the study of forensic texts: corpus, concordance, collocation', in M P Hoey (ed.), *Data Description Discourse*, London, HarperCollins, 86-97.

Coulthard, R M 1994a On the use of corpora in the analysis of forensic texts' *Forensic Linguistics*, 1, i, 27-43.

Coulthard, R M 1994b '*Power*ful evidence for the defence: an exercise in forensic discourse analysis', in J. Gibbons (ed), *Language and the Law*, London, Longman, 414-27.

Coulthard, R M 1994c 'On analysing and evaluating written text', in R.M. Coulthard (ed), *Advances in Written Text Analysis*, London, Routledge 1994, 1-11.

Coulthard, R M 1995 *Questioning Statements: Forensic Applications of Linguistics*, text of inaugural lecture, Birmingham, English Language Research.

Coulthard, R M 1996 'The official version: audience manipulation in police reports of interviews with suspects', in C R Caldas-Coulthard and R M Coulthard (eds) *Texts and Practices: Readings in Critical Discourse Analysis*, London, Routledge, 164-176.

Coulthard, R M 1997 'A failed appeal', *Forensic Linguistics*, 4 ii 287-302.

Coulthard, R M 2000a 'Whose text is it? On the linguistic investigation of authorship', in S. Sarangi and R. M. Coulthard (eds) *Discourse and Social Life*, London, Longman, 271-87.

Coulthard, R M 2000b 'Suppressed dialogue in a confession statement' in Coulthard R M, Cotterill J and Rock F (eds.) *Working with Dialogue*, Tubingen, Niemeyer, 417-424.

Coulthard R M 2000c 'Patterns of lexis on the surface of texts' in M Scott and G Thompson (eds) *Patterns of Text: in Honour of Michael Hoey*, Amsterdam, John Benjamins, 239-254.

Coulthard, R M 2001 'Forensic linguistics', new entry in *The Linguistics Encyclopedia*, 2nd edition, (ed Malmkjaer, K) Routledge.

Coulthard, R M 2002 'Whose voice is it? Invented and concealed dialogue in written records of verbal evidence produced by the police', in J Cotterill (ed) *Language in the Legal Process*, Palgrave, 19-34.

Coulthard, R M 2003 'What did they actually say? A forensic linguist's evaluation of police evidence', *Dialogue Analysis 2000*, Tubingen, Niemeyer, 93-106.

Coulthard, R M 2004a 'Text and authorship – on forensic applications of linguistics', *Structures e Discours: Melanges offerts à Eddy Roulet*, A. Auchlin, M Burger, et al (eds) Editions Nota Bene, 99-113.

Coulthard, R M 2004b 'Author identification, idiolect and linguistic uniqueness' *Applied Linguistics* 25, 4, 431-447.

Eagleson, R 1994 ‘Forensic analysis of personal written texts: a case study’, in J Gibbons (ed.) *Language and the Law*, London, Longman, 362-373.

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

Fox, G 1993 ‘A comparison of ‘policespeak’ and ‘normalspeak’: a preliminary study’ in: Sinclair, J M, Hoey M P, and Fox G *Techniques of Description: Spoken and Written Discourse*, Routledge, London, 183-195.

Grant, T 2005 *Authorship Attribution in a Forensic Context*, unpublished PhD thesis, Department of English, University of Birmingham.

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

Hjelmquist, E 1984 ‘Memory for conversations’, *Discourse Processes*, 7, 321-36.

Hjelmquist, E and Gidlung, A 1985 'Free recall of conversations', *Text*, 3:169-186.

Holmes, D I 1991, ‘Vocabulary Richness and the Prophetic Voice’, *Literary and Linguistic Computing*, 6 (4), 259-268.

Keenan, J M, MacWhinney, B, and Mayhew, D 1977 ‘Pragmatics in Memory: a study of natural conversation’, *Journal of Verbal Learning and Verbal Behavior*, 16, 549-560.

O’Neill, E ‘Crime and punishment: time won’t heal: an investigation’, a report on the case of Robert Brown, http://www.eamonnoneill.net/Candp.html

Sinclair, J McH 1966 'Beginning the Study of Lexis', in Bazell C E, Catford J C, Halliday M A K and Robins R (eds.) *In Memory of J R Firth*, London, Longman, 410-30.

Sinclair, J McH and R M Coulthard 1975 *Towards an Analysis of Discourse: the English Used by Teachers and Pupils*, London, Oxford University Press.

Sinclair, J McH 1987 'Collocation: a progress report', in Steele R & Threadgold T (eds.) *Language Topics: Essays in Honour of Michael Halliday*, Amsterdam, John Benjamins, 319-31

Sinclair, J McH 1991 *Corpus Concordance Collocation*, Oxford, Oxford University Press

Sinclair, J McH 1992 'Trust the text' in M Davies & L Ravelli (eds.) *Advances in Systemic Linguistics: Recent Theory and Practice*, London, Pinter, 5-19; reprinted in R M Coulthard (ed.) *Advances in Written Text Analysis*, London, Routledge, 1994, 12-26.

Sinclair, J McH 1993 'Written discourse structure', in Sinclair J M, Hoey M P and Fox G (eds.) *Techniques of Description: Spoken and Written Discourse, A Festschrift for Malcolm Coulthard*, London, Routledge, 6-31.

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.

Trow, M J 1992 *“Let him have it Chris”*, London, HarperCollins.

Winter, E O and Woolls, D 1996, ‘Identifying authorship in a co-written novel’, Internal report for University of Birmingham

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 102-112.

Woolls, D (n.d.) www.copycatchgold.com\citereader.htm

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.

Appendix - Extracts from *Towards an Analysis of Discourse,* Sinclair and Coulthard, London: OUP 1975

TEXT A At the highest stratum of all there is the interpenetration of minds. Each individual constructs his private linguistic universe, and through his utterances gives hints as to its nature. A problem which has always been with linguistics is the relation between subjective and objective ways of understanding the nature of language. Firth tried to exorcise this dichotomy along with the others but did not succeed. But through the concept of orientation we are able to build both subjective and objective aspects into a coherent model of verbal communication.

One possibility is that participants can maintain a consistent orientation towards each other throughout an interaction. Another is that they can converge on or diverge from each other. Or their orientation may be sensitive to smaller units of the discourse and may vary considerably. Or one participant may adopt an idiosyncratic mode. Because orientation is signalled through a complex network of choices, there are many configurations.

In classroom discourse we have mainly examples of consistency. The teacher's orientation is rarely challenged. The process of education is seen as the pupils accepting the teacher's conceptual world, since he is the mouthpiece of the culture. In some lessons the quality of acceptance seems to superficial – literally making the same noises as the teachers; as when the teacher indicates clearly the answer required and then demands a choral response of the target word or phrase.

The domination of the teacher's language is fully displayed in earlier chapters of this book. The basic IRF structure, giving the teacher the last word, allows him to recast in his own terms any pupil response. Pupils acknowledge the domination by choosing elliptical responses, and by avoiding initiating. Programmed instruction texts often take this sort of interaction to embarrassing extremes.

In an interview between doctor and patient, there is an attempt to construct a conceptual frame compounded of what each brings to the interaction. The doctor brings his expertise in classification and diagnosis and the patient brings his symptoms. The doctor is able to dominate, but the patient retains many subtle ways of insisting on his own view of things. P 130

TEXT B In our effort to make things as simple as possible initially, we chose classroom situations in which the teacher was at the front of the class 'teaching', and therefore likely to be exerting the maximum amount of control over the structure of the discourse. While it was basic to our theory that the verbal and non-verbal context would affect the discourse, we had no theoretical basis for distinguishing between important and unimportant features and therefore set out to control as many potential variables as possible – age, ability, class size, teacher/pupil familiarity, topic of lessons.

Our initial sample consisted of the tapes of six lessons, all based on the hieroglyph materials reproduced in Appendix I, all taught to groups of up to eight 10-11 year-old children by their own class teacher. The system of analysis outlined in Chapter 3 was devised for and based on these lessons. However, once we felt able to handle the controlled sample, we collected a wide variety of tapes covering children of different age groups, in different schools, being taught different subjects by teachers with differing degrees of formality. The system required some, but not major, revision and is now able to cope with most teacher/pupil interaction inside the classroom. What it cannot handle, and of course was not designed to handle, is pupil/pupil interaction in project work, discussion groups, or the playground.

Armed with the results of this research, we are currently attempting to specify a descriptive apparatus capable of greater generality. We have selected a small number of situations which contrast with the classroom along various dimensions but which all have clearly recognizable roles, objectives, and conventions. Chapter 6 gives a brief account of work in progress and indicates the main lines of a developing theory of language interaction. Publication of this volume is designed to promote the generalization of the descriptive apparatus by making it readily available to critics and fellow practitioners. P 67