Some forensic applications of descriptive linguistics

Malcolm Coulthard
(The University of Birmingham)

Abstract

In recent years the descriptive linguist has been invited more and more frequently to give evidence in court in cases where the authorship of texts is in dispute. This paper describes some of the tools and techniques used by the linguist and illustrates with real data drawn from English cases, most of them ones in which the author himself participated. The paper ends with a discussion of linguistic approaches to the detection of plagiarism.
Keywords: forensic application; descriptive linguistics; authorship.

Introduction

Thirty seven years ago Jan Svartvik published The Evans Statements: A Case For Forensic Linguistics in which he demonstrated that disputed parts of a series of four statements dictated to police officers by a young man called Timothy

VEREDAS - Rev. Est. Ling., Juiz de Fora, v.9, n.1 e n.2, p.9-28, jan./dez. 2005

Evans, which incriminated him in the murder of his wife and daughter, had a grammatical style measurably different from the style of the uncontested parts of the statements and a new discipline was born. Initially its growth was slow. In unexpected places there appeared isolated articles in which an author, often a distinguished linguist, analysed disputed confessions or commented on the likely authenticity of purported verbatim records of interaction or identified and evaluated inconsistencies in the language which had been attributed to immigrants or aboriginals in police records of depositions, (for details of early cases see LEVI, 1994a).

There was, however, in those early days no attempt to establish a discipline or even a methodology for forensic linguistics - the work was usually undertaken as an intellectual challenge and almost always required the creation, rather than simply the application, of a method of analysis. By contrast, in the past five years, there has been a rapid growth in the frequency with which Courts in a series of countries have called on the expertise of linguists and, in consequence, there is now a developing methodology and a growing number of linguists who act as expert witnesses, a few even on a full time basis, (see LEVI, 1994b; EADES, 1994; MCMENAMIN, 2002; ROSE, 2002).

1. What do forensic linguists do?

Forensic linguists are most frequently called in to help a court answer one or both of two questions: what does a given text 'say' and who is its author? In answering these questions linguists draw on knowledge and techniques derived from one or more of the sub-areas of descriptive linguistics: phonetics and phonology, lexis, syntax, semantics, pragmatics, discourse and text analysis. For this reason, just as some of those within the general field of linguistics often prefer to distinguish themselves as phoneticians, lexicographers, grammarians or discourse analysts, so within forensic language analysis there are two distinct sub-classes of expert, *forensic phoneticians* and *forensic linguists*.

2. What does a text say?

For the phonetician this is a question of decoding words and phrases from tape-recordings - when a recording is of poor quality the non-expert may hear one thing, while the expert with a trained ear and the help of sophisticated equipment and software may perceive something entirely different. In one case in which I was involved an indistinct word in a clandestine recording of a man later accused of manufacturing the designer drug Ecstasy, was crucially mis-heard by a police transcriber as 'hallucinogenic':

> but if it's as you say its *hallucinogenic*, it's in the Sigma catalogue
> whereas what he actually said was
> but if it's as you say its *German*, it's in a Sigma catalogue.

In another case, a man with a strong West Indian accent was transcribed as saying that he got onto a train and then "shot a man to kill" whereas the phonetician was able to demonstrate that what he had actually said was the innocuous and contextually much more plausible "showed a man ticket".

The forensic linguist is concerned not with deciphering words, but rather with their interpretation. The meaning of phrases or even individual words can be of crucial importance in some trials. Perhaps the most famous British example comes from the 1950s, the case of Derek Bentley and Chris Craig. Bentley, already under arrest at the time, was said to have shouted to Craig, who had a revolver in his hand, "let him have it, Chris"; shortly afterwards Craig fired several times and killed a policeman. There was a long debate in court over the interpretation of Bentley's ambiguous utterance, which was resolved in favour of the prosecution's incriminating interpretation, "shoot him" rather than the defence's mitigating "give him the gun"; this made Bentley an accessory to murder, for which he was convicted and later hanged.

Although lawyers and judges usually see themselves as the guardians of the meaning of legal texts, and indeed one judge in the United States refused to admit the linguist Ellen Prince as an expert on the grounds that it is the function of the Court to decide on meaning, linguists are occasionally allowed to express a professional opinion. Kaplan et al (1995) report on a case which went to the Supreme Court in 1994. The facts are as follows: a certain Mr Granderson pleaded guilty to a charge of destroying mail, for which the maximum sentence was 6 months imprisonment. In fact the judge decided not to imprison him but instead to fine him and also put him on probation for 5 years. Subsequently Mr Granderson violated his probation by being caught in possession of cocaine. In such cases the law instructed the Court to 'revoke the sentence of probation and sentence the defendant to not less than one third of the original sentence'. This presented the Court with a problem because, if it took 'original sentence' to refer to 'probation', imposing a sentence of 'not less than one third' would in fact reduce the penalty as it had more than 20 months still to run, so it was decided that the correct interpretation required the court to sentence him to *20 months in jail*, even though that was a sentence more than three times greater than the original maximum.

Kaplan et al (1995) argued on linguistic grounds that this interpretation was inadmissible, on the grounds that the Court had treated the ambiguous phrase 'original sentence' as if it could simultaneously have two different meanings: they had interpreted it as referring to 'imprisonment' for the purpose of determining the *type* of punishment, but then to the imposition of 5 years (of probation) for determining the *length* of the sentence. To clarify the confusion they observed that this is the linguistic equivalent of a Frenchman taking the phrase *Pierre a fait tomber l'avocat* to mean, Pierre did something to a lawyer (avocat meaning 1) and caused an avocado (avocat meaning 2) to fall. The Supreme Court accepted the argument and changed the sentence.

More often, the dispute is not over what the original professional producer of a message intended an item to mean, but rather what a non-expert, the ordinary man-in-the-street, might reasonably have interpreted that expression to mean. For example, there are currently several cases coming to the Courts about the meaning and clarity of warnings on cigarette packets (see DUMAS, 1990) and about the explicitness and honesty of cautionary advice given to women contemplating breast implants.

To illustrate I will focus on a case where a 58 year old cement worker sued an insurance company which was refusing to pay his disability pension on the grounds that he had lied when he responded to four of the questions on the original proposal form. One question read as follows

> Have you any impairments?.... Loss of sight or hearing?.... Loss of arm or leg?...Are you crippled or deformed?... If so explain....

The insurance company argued that the man had lied when he replied to this question in the negative, since "he was overweight, had a high cholesterol level and occasional backaches", despite the fact that none of these conditions had ever caused him to take time off work, (PRINCE, 1981, p.2). In her evidence Prince focused on the vagueness of the word *impairment*, and argued, apparently successfully, that any 'co-operative reader' would reasonably infer that, given the phrases that followed the word *impairment,* the word was being used in that question to mean a relatively severe and incapacitating physical condition and that therefore the man had indeed answered "no" 'appropriately and in good conscience', even though it was an untrue answer to the question (claimed to have been) intended by the insurance company (ibid: 4).

When what is at issue is the meaning of single words, the problem facing the linguist is how to discover and then how to demonstrate to a lay audience what the 'ordinary' meaning of an item actually is. One fairly new method is to use evidence of actual usage derived from a *corpus*, that is a collection of millions of words of spoken and/or written texts stored in computer-readable form. The University of Birmingham's COBUILD team has built up a 400 million word corpus, whimsically labelled the Bank of English, which is an invaluable resource for the forensic linguist, as is the 100 million word British National Corpus.

3. Maite please can you put in here a ref to relevant Spanish corpora?

Sinclair (ms) used the Bank of English corpus when he was asked to give an opinion on the ordinary man's understanding of the word *visa.* Apparently in British law a visa is not in fact an 'entry permit', but rather 'a permit to request "leave to enter"'. Sinclair was asked to provide evidence that this is not the commonly understood use and meaning of the word. He based his evidence mainly on a 5 million word corpus of *The Times* newspaper, although he supplemented this data by reference to the whole of the Bank of English. *The Times* sub-corpus included 74 instances of "visa" and "visas" in the sense under consideration, of which over 50 co-occurred or *collocated* with common verbs like "grant", "issue", "refuse", "apply for", "need" and "require". Sinclair noted that, although the commonest modifier of "visa(s)" is "exit" the word also co-occurs with "entry" and "re-entry":

> You cannot *enter* an Arab country with an Israeli visa stamped in your passport... British passport holders *do not require* visas...
> Non-Commonwealth students who *require* an *entry* visa will *need* a *re-entry* visa, even if you only *leave* the country for a couple of days...

so, he concluded that

the average visitor, encountering everyday English of the type recorded in the corpus, would deduce that a visa was a kind of permit to enter a country... There is nothing... in these examples to suggest that a person who is in possession of a valid visa, or who does not require a visa, will be refused entry. The implication is very strong that a visa either ensures entry, or is not needed for entry. The circumstances of someone requiring "leave to enter" in addition to having correct visa provision does not arise in any of the examples, and the word "leave" does not occur in proximity to "visa(s)" except in the meaning "depart", (Sinclair, ms)

This is one example of what can be achieved with a fairly common word and a reasonably small corpus and demonstrates very clearly the usefulness of the method. However, it also shows that it is essential to have a substantial number of instances of the word in question and is therefore in itself a justification for the collection of very large corpora - if for instance one were interested in a word which occurs on average once every 2 million words, one would ideally need to consult the whole of the 400 million word corpus.

4. Who is the author?

In a significant number of cases what is in question is authorship. The phonetician is typically asked whether the voice on sample tape-recordings is the same as the voice committing the crime. Often there have been attempts at disguise, sometimes simply to protect the speaker, but sometimes more interestingly in order to pretend to be someone else, as in the case of spoof royal telephone calls or telephone banking fraud. The phonetician has ways of identifying disguise - an assumed accent may slip - but there are certainly cases when the disguise succeeds. (ROSE, 2002; SCHLICHTING, F.; SULLIVAN, K. 1997)

Authorship of course is an ambiguous concept - the physical production of a text may be separated from the creation of its content, as anyone who has dictated a letter and then had to correct the spelling mistakes knows only too well. Thus there are cases where there is no dispute that a given text is written in a particular hand or spoken by a particular voice, but there may still be real doubt about who was the *author* of the *message*. The most obvious examples are suspect suicide notes, but there are instances where letters were written under duress, written confessions dictated by police officers to the accused and one interesting case where a woman claimed that her taped and apparently spontaneous confession was in fact a recording of her reading aloud a statement prepared in advance by a police officer.

5. The linguistic investigation of authorship

The linguist can also approach 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, p.75). Every speaker

has a very large active vocabulary built up over many years, which will differ from the vocabularies which 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 characterise 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, b, 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 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 in **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.

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, which he had 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 because 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 would 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 or 'lemmas' *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 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. (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/).

6. Mistakes and errors

It is a basic tenet of linguistics that not only is language rule-governed, but that so also is its production in written or spoken form, although, of course, any spoken or written text may display items which break the rules of the *standard* language. We can divide such rule-breaking into two categories: 'performance' *mistakes*, where the speaker/writer knows that s/he has broken a rule, which, of course, doesn't prevent her/him from breaking it again, (and again, and again, as learners of foreign languages know to their chagrin) and 'competence' *errors,* where the speaker/writer is working with a set of rules which are non-standard, but rules which s/he nevertheless follows consistently, even though s/he may still make performance mistakes which break their own rules. In the short texts that are typically the focus of the forensic linguist, it is

usually only possible to focus on the grammatical and orthographic rule-breaking, because in order to examine characteristic vocabulary choice one would need much more textual data than is typically available.

The most difficult author-identification cases are those involving anonymous letters, because there is usually a fairly large number of potential authors and only a small amount of written text to analyse. For this reason, success is in the main limited to those cases which involve semi-literate authors, who necessarily provide a comparatively large number of idiolectal mistakes and errors in a comparatively small amount of text. (Obviously all intending anonymous letter writers should use a word-processor spell-checker and also the style-improver options now available in the latest programs, in order to homogenise the style and thereby disguise their idiolect.)

Below I reproduce a few short extracts from a typed anonymous letter, which the addressee-company suspected was written by one of its own employees. I have highlighted the words which contain non-standard features by italic; (there are many more instances in the rest of the letter of the particular phenomena I have chosen to focus on):

> .... I hope you appreciate that *i* am *enable* to give my true *idenitity* as
> _ this *would* ultimately jeopardize my position.....
> .... *I* would like to *high light* my greatest concern....
> .... have so far *deened* it unnecessary to *investegate* these *issus*.....

There are several interesting non-standard features immediately apparent, although one of the problems of dealing with typewritten text is that errors and mistakes may be confused and compounded - one may not know, for any given item, particularly if it only occurs once, whether the 'wrong' form is the product of a mis-typing or a non-standard rule - for instance if a (British English) text includes the word 'color' is this a typing mistake or a spelling error, or even worse the result of the computer user being unable to change the spell-checker to British English?

In examining the non-standard items in the extracts above we note firstly, the writer is an inexperienced typist, the first person pronoun "I" appears also as "i", and the very unusual "l"; secondly, some of the words have metathesized (reversed) letters and others additional or omitted letters; thirdly, the writer has serious problems when spelling words containing unstressed vowels - thus we have the following spellings "enable" = "unable", "investegate" = "investigate" and elsewhere "except" = "accept"; fourthly, the writer is unsure about when to write certain sequences of morphemes as a single word and when as two separate words - thus "high light" and "with out". In addition, but not exemplified here, there are homonym problems, "weather" appears for "whether" and "there" for "their". Finally, the writer has some grammatical problems: the frequent omission of markers of past tense and of the 3rd person singular present tense and even of articles - "have now (a) firm intention". Collectively these mistakes and errors are idiosyncratic and idiolectally distinctive and proved to be instanced in the authenticated letters of only one of the eight employees who had access to the information contained in the threatening letter. He turned out to be the employee already suspected by the company.

7. Fabricated texts

There are occasions when someone claims that a text is in part or completely falsified – i.e. that the real author is different from the purported author. In this context the fabricator, whether s/he is creating an interview record, a confession statement or a suicide note is acting as an amateur dramatist or novelist, imagining what the purported speaker/author would have produced in the same circumstances. As with any fabrication, be it bank notes or written texts, the quality of the finished product will depend on the degree of understanding that the falsifier has of the nature of what s/he is falsifying. Depending on the nature of the text being examined different linguistic approaches are suitable.

8. Spoken and written language

This case concerns a disputed statement, in which the accused had confessed to involvement in a terrorist murder. He claimed that some of what was contained in the statement had been fairly accurately recorded, but denied having dictated a substantial proportion of the statement.

It is now well established within linguistics (HALLIDAY, 1989) that spoken and written language have different principles of organisation and can usually be distinguished both grammatically and lexically. As a generalisation, spoken language tends to have short clauses, a low ratio of lexical to grammatical words and represents what happened as *process* by the use of verbs, whereas written language tends to have longer clauses, a higher lexical density and represents what happened as *product* by the use of nominalisations. For example, the following sentence, which the accused admitted to having said, displays short co-ordinated clauses and very low lexical density that are typical of spoken narrative:

> I drove down to the flats & I saw him up on the roof & I shouted to him & he said that he would be down in a couple of minutes.

We notice that the sentence contains thirty two words, only seven of them lexical, divided into five clauses, giving an average of 6.4 words per clause and a lexical density of 1.4 lexical items per clause. The disputed first sentence, presented below, is in marked contrast to the above undisputed sentence consisting, as it does, of a mere three clauses which contain forty seven words, (I have conservatively treated '1987' and 'ABC' as single words), 25 of which are lexical, giving an average clause length of 15.7 and a lexical density of 8.3:

> I wish to make a further statement explaining my complete involvement in the hijacking of the Ford Escort van from John Smith on Monday 28 May 1987 on behalf of the A.B.C. which was later used in the murder of three person (sic) in Newtown that night.

In other words, this sentence has the high lexical density, massive subordination and frequent use of nominalisations - for example *statement,*

involvement, hijacking and *murder* - typical of written texts. On cross examination the police officer/scribe conceded that the statement may not after all have been recorded verbatim, but continued to maintain that all the words were indeed spoken by the accused, though "perhaps not in that exact order"!

9. Conversational rules about explicitness and detail

Some cases require reference to the socio-linguistic rules which govern the production of speech. Grice (1975) in his seminal article 'Logic and conversation' observed that one of the controls on speakers' contributions is the *quantity maxim,* which he summarised as

> a) make your contribution as informative as is required (for the current purposes of the exchange),
> b) do not make your contribution more informative than is required.

What Grice is concerned with here is the fact that all utterances are shaped for a specific addressee on the basis of the speaker's assumptions about shared knowledge and opinions and in the light of what has already been said, not only in the ongoing interaction but also in relevant previous interactions. This appeal to what Brazil (1985) called 'common ground', makes conversations frequently opaque and at times incomprehensible to an overhearer.

It is for this reason that it would be impossible to present truly 'authentic' conversation on the stage, because the real addressee of any stage utterance is in fact the audience who needs supplementary background information. Thus, there has arisen the dramatic convention of over-explicitness, which allows characters to break the quantity maxim and to say to each other things they already 'know', even things that are strictly irrelevant, in order to transmit essential information economically to the audience. This is a convention which the dramatist Tom Stoppard parodies at the beginning of *The Real Inspector Hound*:

> Mrs Drudge (into phone) Hello, the drawing room of Lady Muldoon's country residence one morning in early spring... Hello! - the draw- Who? Who did you wish to speak to? I'm afraid there is no one of that name here, this is all very mysterious and I'm sure its leading up to something, I hope nothing is amiss for we, that is Lady Muldoon and her houseguests, are here cut off from the world, including Magnus, the wheelchair-ridden half-brother of her ladyship's husband, Lord Albert Muldoon, who ten years ago went out for a walk on the cliffs and was never seen again - and all alone for they had no children.

When we come to consider the fabricator of texts, we can see that he is in a situation directly analogous to that of the dramatist - he is creating his text with the overhearer, in this case the judge (and jury) in the trial, in mind, and for this reason is anxious to make the incriminating information that is being transmitted by the text as unambiguous as possible. Thus, at times the fabricator, just like the dramatist, will break the maxim of quantity, though rarely as extremely as in the extract below, which is taken from the beginning of a fabricated

telephone conversation, in which a convicted defendant, Mr B, is trying to discredit, Mr A, who had given evidence against him at his trial; note particularly utterances B2, B3 and A4:

A1: Hello.
B1: Hello, can I speak to Mr A please?
A2: Speaking.
B2: Are you surprised I've phoned you instead of coming down and seeing you as you asked in your message over the phone yesterday?
A3: No, I'm not surprised. Why are you phoning me here for? Why don't you come in to see me if you want to see outside?
B3: Well you've dragged me through a nightmare and I don't intend to give you an opportunity to set me up again for something else or beat me up again and abandon me miles away as you did outside Newtown prison with the two detectives; and for your information, as you may know, I've filed an official complaint against you and the two C.I.D. detectives.
A4: The detectives and I beat you up and the C.I.D. they denied, they didn't beat you up but you cant do anything because you got no proof.

Over-explicitness can be realised in the choice of nominal groups as well. In the disputed confession attributed to William Power, one of the so-called Birmingham Six, in a famous British case dating from the mid 1970's, (see COULTHARD, 1994a) there was frequent reference to "white plastic (carrier) bags":

Walker was carrying *two white plastic carrier bags....*
Hunter was carrying *three white plastic carrier bags....*
Richard was carrying *one white plastic carrier bag....*
Walker gave me *one of the white plastic bags....*
Hughie gave J Walker his *white plastic bag....*

Our knowledge of the rules of conversational composition tells us that it is unlikely that Power would have used the combination, 'numeral + white + plastic + carrier + bags' even once. Firstly, it is a noted feature of speech that speakers do not normally produce long noun phrases of this kind; rather they assemble complex information in two or three bits or bites. Secondly, it represents a degree of detail we do not see in the rest of his statement. Finally, the detail does not seem to have any importance in the story as *he* tells it and it is very unusual for narrators to provide detail which has no relevance to *their* story. Let us compare the way similar information came out in Power's interview with the police, which has a ring of authenticity:

Power: He'd got a holdall and *two bags*
Police: What *kind of bags*?
Power: They were *white*, I think they were *carrier bags.*

and even then nothing was said about 'plastic'. The extract below taken from cross-examination during the trial shows clearly that, once a full form of a referring expression has been used, a speaker's normal habit is to employ a shortened version on subsequent occasions.

Mr Field-Evans: And did you say '*two white plastic carrier bags?*'
Power: Yes sir
Mr Field-Evans: Whose idea was it that Walker was carrying *two white carrier bags?* Were those your words or the Police Officers' words?
Power: They were the Police Officers'. They kept insisting that I had told them that they carried *plastic bags* into the station.
Mr Field-Evans: Does *the same* apply to what Hunter was carrying?
Power: I don't know what you mean sir.
Mr Field-Evans: I am sorry. Whose idea was it that you should say that Hunter was carrying *three white plastic bags?*
Power: Well, sir, I said that.
Mr Field-Evans: But was it your idea?
Power: No. They kept saying that I had already told them that they were carrying *plastic bags* into the station. When I said that, they said "who was carrying *them*? who was carrying *them*?" They threatened me. I said "They were all carrying *them*." They asked me how *many* were they carrying and I just said *one, two, three, one and one*.

10. Register features

Linguists have long recognised that the language that any given individual uses varies according to the contexts in which and the topics for which s/he is using it - this is called *register* variation. Thus, at its simplest a policeman at work will have a series of linguistic options which mark him as a policeman, as indeed will a doctor, an economist, a linguist, etc. When a text is being falsified there is always the possibility that the real author will allow features of his own usage to enter into the text; these features may be idiolectal, as we saw above, but may also be due to register.

To illustrate this I will focus on the statement of Derek Bentley, who has already been referred to above, which was made some three hours after his arrest. At his trial Bentley claimed that this statement was in fact a composite document, not simply transcribed, but also in part written, by the police. I will focus on one small linguistic feature simply to illustrate how a register analysis works; obviously a full analysis would focus on a whole series of features. (I reproduce the whole of Bentley's statement below.)

11. Derek Bentley's Statement

I have known Craig since I went to school. 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. I was watching television tonight (2 November 1952) and between 8 p.m. and 9 p.m. Craig called for me. My mother answered the door and I heard her say I was out. I had been out earlier to the pictures and got home just after 7 p.m. A little later Norman Parsley and Frank Fasey called. I did not answer the door or speak to them. My mother told me that they had called and I then ran after them. I walked up the road with them to the paper shop where I saw Craig standing. We all talked

together and then Norman Parsley and Frank Fazey left. Chris Craig and I then caught a bus to Croydon. We got off at West Croydon and then walked down the road where the toilets are - I think it is Tamworth Road.

When we came to the place where you found me, Chris looked in the window. There was a little iron gate at the side. Chris then jumped over and I followed. Chris then climbed up the drainpipe to the roof and I followed. Up to then Chris had not said anything. We both got out on to the flat roof at the top. Then someone in a garden on the opposite side shone a torch up towards us. Chris said: 'It's a copper, hide behind here.' We hid behind a shelter arrangement on the roof. We were there waiting for about ten minutes. I did not know he was going to use the gun. A plain clothes man climbed up the drainpipe and on to the roof. The man said: 'I am a police officer - the place is surrounded.' He caught hold of me and as we walked away Chris fired. There was nobody else there at the time. The policeman and I then went round a corner by a door. A little later the door opened and a policeman in uniform came out. Chris fired again then and this policeman fell down. I could see he was hurt as a lot of blood came from his forehead just above his nose. The policeman dragged him round the corner behind the brickwork entrance to the door. I remember I shouted something but I forget what it was. I could not see Chris when I shouted to him - he was behind a wall. I heard some more policemen behind the door and the policeman with me said: 'I don't think he has many more bullets left.' Chris shouted 'Oh yes I have' and he fired again. I think I heard him fire three times altogether. The policeman then pushed me down the stairs and I did not see any more. I knew we were going to break into the place. I did not know what we were going to get - just anything that was going. I did not have a gun and I did not know Chris had one until he shot. I now know that the policeman in uniform is dead. 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'. He was with us when the other policeman was killed.

12. *"then"*

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 582 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 suggested to me that Bentley's usage 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.

To test this hypothesis I created two small corpora, 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, which totalled some 930 words of text, the second composed of statements by three police

officers, two of whom were involved in the Bentley case and the third in another unrelated case, which totalled some 2270 words. The results were startling: whereas in the three ordinary witness statements there is only one occurrence of "then" in 930 words, "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 58 words, groups his statement firmly with those produced by the police officers. In this case I was fortunate in being able to check the representativeness of my '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 running words collected from very many different types of naturally occurring speech. "Then" in all its meanings proved to occur a mere 3,164 times, that is, on average, 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 occurrences in the 582 word text

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

This sounds odd 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 they use "then I" ten times more frequently than "I then"; indeed the structure "I then" occurred a mere 9 times in the whole of the spoken sample, 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, not only does this "I then" structure, as one might predict from the corpus data, not occur at all in any of the three witness statements, also there are 9 occurrences in one single 980 word police statement, as many occurrences as in the entire 1.5 million word spoken corpus. Taken together the average occurrence in the three police statements is once every 119 words. Thus, the structure "I then" does appear to be a feature of policeman's (written) register.

More generally, it is in fact the structure Subject (+Verb) followed by "then" which is typical of policeman's register; it occurs 26 times in the statements of the three officers and 7 times in Bentley's own statement. 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 one might expect to be have been influenced by police reporting style, says "*Then* you". Recently other linguists, working on a whole series of British police and witness statements, have confirmed the use of postposed "then" to be a marked feature of police register.

13. Idiolect and the detection of plagiarism

One major authorship detection problem that is part of the day-to-day life of academics is plagiarism. At its simplest, plagiarism, or more accurately the type of plagiarism that 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, 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, 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 composed independently.

In most plagiarism cases involving students, there is little doubt about guilt, as these two examples of essay openings from Johnson (1997, p. 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. 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 longest identical sequences are a mere three running words. Even so, one would still want to categorise this degree of lexical overlap, if instanced in other parts of the text, as unacknowledged, though more sophisticated, borrowing and therefore as plagiarism, even if it doesn't 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; (An automated version of this analytic method, produced by Woolls (2002), is now available as the computer program *Copycatch Gold*). 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; 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. (For an application of this methodology to Spanish texts see TURRELL, 2004)

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 plagiarised 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.

Afterword

Forensic linguists still have a long way to go in certain areas to convince courts of the reliability and validity of their methodology, particularly in the United States, (see TIERSMA; SOLAN, 2002), but I hope I have said enough to convince you that the discipline has firm foundations and can already offer interesting and well-founded opinions in certain areas.

Resumo

Nos últimos anos, a Lingüística Descritiva tem sido cada vez mais convocada em tribunais para fornecer subsídios em casos em que a autoria de textos está em disputa. Este artigo descreve algumas das ferramentas e técnicas usadas pelo lingüista, ilustrando-as com casos reais em língua inglesa, na maioria dos quais o autor do artigo esteve envolvido na qualidade de perito real. O artigo finaliza com uma discussão a respeito de abordagens lingüísticas para a detecção do plágio.
Palavras-chave: aplicação forense; lingüística descritiva; autoria.

References

BALDWIN, J.; FRENCH, J. *Forensic Phonetics.* London: Pinter, 1990.

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

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

BRAZIL, D. C. *The Communicative Value of Intonation.* Birmingham: English Language Research, 1985.

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

COLLINS, H. Modal profiling in oral presentations. In: BARBARA, L.; SCOTT, M. (Eds.). *Reflections on Language Learning.* Clevedon: Multilingual Matters, 1994. p. 214-29.

COULTHARD, R. M. (ed.). *Talking about Text.* Birmingham: ELR, 1986.

_____. Forensic discourse analysis. In: COULTAHARD, R. M. (Ed.). *Advances in Spoken Discourse Analysis.* London: Routledge, 1992. p. 242-57.

_____. Beginning the study of forensic texts: corpus, concordance, collocation. In: HOEY, M. P. (Ed.). *Data Description Discourse.* London: Harper Collins, 1993. p. 86-97.

_____. *Powerful evidence for the defence: an exercise in forensic discourse analysis.* In: GIBBONS, J. (Ed.). 1994a. p. 414-42.

_____. On the use of corpora in the analysis of forensic texts. *Forensic Linguistics: the International Journal of Speech, Language and the Law,* 1, p. 27-43. 1994b.

_____. A failed appeal. In: *Forensic Linguistics: the International Journal of Speech, Language and the Law,* 4, ii, p. 287-302. 1997.

DAVIS, T. *Forensic handwriting analysis.* In: COULTHARD, R. M. (Ed.), 1986. p. 189-207.

_____. ESDA and the analysis of contested contemporaneous notes of police interviews. *Forensic Linguistics: the International Journal of Speech, Language and the Law,* 1, p. 71-89. 1994.

_____. Clues and opinions: ways of looking at evidence. In: KNIFFKA, H.; COULTHARD, R. M.; BLACKWELL, S. (Eds.). *Papers from the First International Conference of Forensic Linguists,* Bonn July 14-16 [1993] 1995.

DUMAS, B. Adequacy of cigarette package warnings. In: LEVI, J.; GRAFFAM, Walker A. (Eds.). *Language in the Judicial Proces.* New York, Plenum,1990. p. 309-52

EADES, D. 1994 Forensic linguistics in Australia: an overview. *Forensic Linguistics: the International Journal of Speech, Language and the Law,*1, p. 113-32. 1994.

EAGLESON, R. Forensic analysis of personal written text: a case study. In: GIBBONS, J. (Ed.). *Language and the Law*, London, Longman, 1994. p. 362-373.

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

FOX, G. A comparison of "policespeak" and "normalspeak" a preliminary study. In: SINCALIR, J. M.; HOEY, M. P.; FOX, G. (Eds.). *Techniques of Description: Spoken and Written Discourse, A Festschrift for Malcolm Coulthard.* London: Routledge, 1993. p. 183-95.

FRENCH, J. P. An overview of forensic phonetics. *Forensic Linguistics: the International Journal of Speech, Language and the Law,* 1, p. 169-181. 1994.

GIBBONS, J. (Ed.). *Language and the Law*, London: Longman, 1994.

GRICE, H. P. Logic and conversation. In: COLE, P.; MORGAN, J. (Eds.). *Syntax and Semantics III: Speech Act.* New York: Academic Press, 1975. p. 41-58.

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

_____. *Spoken and Written Language.* 2nd edition. Oxford: OUP, 1989.

_____.; MCINTOSH, A.; STREVENS, P. *The Linguistic Sciences and Language Teaching.* 1964.

HOWARD, R. M. The new abolitionism comes to plagiarism. In: BURANEN, L.; ROY, A. M. (Eds.). *Perspectives on plagiarism and intellectual property in a postmodern world.* Albany: State University of New York Press, 1999. p. 87-95.

JOHNSON, A. Textual kidnapping – a case of plagiarism among three student texts. *Forensic Linguistics: The International Journal of Speech, Language and Law,* 4, p. 210-25. 1997.

KAPLAN, J. P.; GREEN, G. M.; CUNNINGHAM, C. D.; LEVI, J. N. Bringing linguistics into judicial decision making: semantic analysis submitted to the US Supreme Court. *Forensic Linguistics: the International Journal of Speech, Language and the Law,* 2, i. 1995.

LEVI, J. N. *Language and the Law: A Bibliographical Guide to Social Science Research in the USA.* Chicago: American Bar Association, 1994a.

_____. Language as evidence: the linguist as expert witness in North American Courts. *Forensic Linguistics: the International Journal of Speech, Language and the Law,* 1, p. 1-26. 1994b.

MCMENAMIN, G. *Forensic Linguistics: Advances in Forensic Stylistics.* London: CRC Press, 2002.

PECORARI, D. E. *Original Reproductions: an Investigation of the Source Use of Postgraduate Second Language Writers.* 2002. (Unpublished PhD thesis), University of Birmingham.

PRINCE, E. Language and the law: a case for linguistic pragmatics. *Working papers in Sociolinguistics.* Austin: Southwest Educational Development Laboratory, 112-160. 1981.

ROSE, P. *Forensic Speaker Identification.* London, Taylor and Francis. 2002.

SCHLICHTING, F.; SULLIVAN, K. he imitated voice – a problem for voice line-ups? *Forensic Linguistics. The International Journal of Speech, Language and Law,* 4, p. 65-148. 1997.

SCOTT, M.; JOHNS, T. *Microconcord.* Oxford: OUP, 1993.

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

SINCLAIR, J. McH ms Unpublished expert opinion on the ordinary man's understanding of the word "visa". 1993.

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

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

TIERSMA, P.; SOLAN, L. The linguist on the witness stand: forensic linguistics in American courts. *Language* 78. p. 39-221. 2002.

TURREL, T. Textual kidnapping revisited: the case of plagiarism in literary translation. *International Journal of Speech, Language and the Law,* 11, page nos missing. 2004.

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

_____. Better Tools for the Trade and how to Use them. *Forensic Linguistics. The International Journal of Speech, Language and Law,* 10, p. 102-112. 2003.

_____.; COULTHARD, R. M. Tools for the Trade. *Forensic Linguistics. The International Journal of Speech, Language and Law,* 5, p. 33-57. 1998.