Journal of Law and Policy

Volume 21
Issue 2
SYMPOSIUM:
Authorship Attribution Workshop

Article 8

2013

On Admissible Linguistic Evidence

Malcolm Coulthard, Ph.D.

ON ADMISSIBLE LINGUISTIC EVIDENCE

*Malcolm Coulthard**

PREAMBLE

This is a very unconventional journal article, the likes of which I have never before written. It is based on a paper that was conceived of and written for the Authorship Attribution Workshop ("Workshop") hosted at Brooklyn Law School in October 2012 with the intention of exploring the boundaries of admissibility of linguistic evidence in U.S. courts. This paper focuses on admissible linguistic evidence in an English court and explores whether some or all of it would be accepted in a U.S. court, where the *Daubert* acceptability criteria,[1] particularly information about known rates of error, are more rigorous than the criteria currently in force in the U.K. Interestingly, it is likely that *Daubert*-like criteria will be introduced into the U.K. in the not too distant future, so it was not just academic curiosity that that led me to inquire whether my evidence would be admissible. Specifically, I wondered if in the U.S. I would be permitted to express an opinion on the evidence or only to act as a "tour guide,"[2] simply presenting the linguistic evidence to the court without evaluation. The general consensus of the Workshop's evidence experts was that most of my evidence would indeed be allowable in a U.S. court.

Comments made during the Workshop about my presentation and analytic advances outlined by Dr. Tim Grant during his presentation have led me to revise and add to my analysis. As a consequence, what you will read below is, I hope, a more

* Federal University of Santa Catarina, Brazil.

[1] *See* Daubert v. Merrell Dow Pharm., 509 U.S. 579 (1993).

[2] *See generally* Lawrence Solan, *Linguistic Experts as Semantic Tour Guides*, 5 FORENSIC LINGUISTICS 87 (1998).

convincing and more soundly based analysis of the evidence, and one that would comply better with the *Daubert* criteria. I leave it to you, the reader, to reach your own decision on admissibility. Interestingly, and again uniquely in my own experience, I will be able to present my evidence in court for a second time later this year because the first trial ended with a hung jury.[3]

INTRODUCTION

Professors Peter Tiersma and Larry Solan note that although "[U.S.] courts have allowed linguists to testify on such issues as the probable origin of a speaker, the comprehensibility of a text, whether a particular defendant understood the *Miranda* warning, and the phonetic similarity of two competing trademarks," in other linguistic areas the situation is more problematic, as the system now requires evidence to conform to the *Daubert* principles.[4] Solan notes,

> it must be conceded that, in cases where conclusions depend on observations about the frequency or rarity of particular linguistic features in the texts under examination, many linguists would have considerable difficulty in stating a "known rate of error" for their results, even if this phrase is interpreted as a likelihood ratio. It is for this reason that some linguists will be forced to change their way of reaching and presenting their opinions, while others may choose to see their role more as that of "tour guides" than opinion givers.[5]

Solan goes on to address the problem that is unique to experts in linguistics—the fact that the judges of fact, whether they be actual judges or jury members, are seen for most

---

[3] As the case is still ongoing, I have changed the names of all of the participants.

[4] Peter Tiersma & Lawrence M. Solan, *The Linguist on the Witness Stand: Forensic Linguistics in American Courts*, 78 LANGUAGE 221, 221 (2002).

[5] MALCOLM COULTHARD & ALISON JOHNSON, AN INTRODUCTION TO FORENSIC LINGUISTICS: LANGUAGE IN EVIDENCE 210 (2007) (citing Solan, *supra* note 2).

purposes to be their own experts in the area of language use and interpretation. The law is, much of the time, concerned with the meaning(s) that ordinary speakers attach to words and expressions.[6] Even so, Solan argues that there is still a role for the linguist, which is to explain and elucidate facts *about* language and usage as a result of which judge and jury will then be in the same position as the linguist and so can make linguistically informed decisions.[7] He explains that his linguistic training has made him "more sensitive to possible interpretations that others might not notice" and as a consequence he can point these out to the jury. However, he adds, "[O]nce I point these out and illustrate them clearly, we should start on an equal footing."[8]

One of Solan's points that is crucially relevant to what follows is that, although juries and judges may well be able to process words, phrases, and even sentences as well as any professional linguist, they may have problems with long documents or with a series of related documents because they may not be able to make the necessary links: "Of course a jury can read the document[s]. . . . But not all jurors, without help, can focus on a phrase in paragraph 24 of a contract that may have an impact on how another word should be interpreted in paragraph 55."[9]

To facilitate a discussion of Solan's points, I present below an edited version of an expert report I wrote where there was one questioned email and tens of thousands of emails available for searching written by many authors whose authorship was unchallenged. As a Coda, I add a new analysis produced as a consequence of the stimulating discussion at the Workshop.

I. EXPRESSING OPINIONS

The lawyers in the case I discuss below wanted me to express my opinions using degrees of likelihood: "it is

---

[6] Solan, *supra* note 2, at 91.

[7] *Id.* at 92.

[8] *Id.*

[9] *Id.* at 94.

(quite/very) (un)likely that *X* is the author of the email." However, as Philip Rose argues convincingly, expressing an opinion in this form is tantamount to expressing an opinion on the likelihood of the accused being guilty, which is the exclusive role of the judges of fact.[10] All that a linguist can comment on is the degree of similarity or difference between linguistic choices in the questioned and the known texts. Rose supports his argument by pointing out that no expert can make an estimate of the likelihood of guilt or innocence on the basis of the linguistic evidence alone; only those with access to all the available evidence can assess the value of each piece of it.[11] For this reason, I prefer to approach questions of authorship attribution as a two-stage process, asking first if the choices in the questioned document are *compatible* with choices made by the potential authors in their known documents. If the choices are not compatible, no further analysis is undertaken. Then, as a second stage for those candidate author(s) for whom the choices are indeed compatible, one comments on how *distinctive* the particular linguistic choices are, on a five-point scale from *not distinctive* to *exceptionally distinctive*.

II. THE BRIEF

I was asked to express an opinion on the likely authorship of a questioned email sent from the email account of a Mr. Stephen Goggin to a Mr. Denis Juola at 16.30 on July 23, 2004. I was briefed that, given the timing and content of the email, in particular the knowledge of and explicit reference to an earlier phone call to Mr. Juola timed at 14.50, only a small number of people—Mr. Goggin; Mr. Tim Widdowson, the CEO; Mr. John Shuy, the Finance Director of MaxiSoft; and possibly their PA, Ms. Janet Gavalda—could have been in a position to author and type the email. I was asked to proceed on the assumption that, although the email was sent from Mr. Goggin's e-account, it may not have been physically typed on his computer, because

---

[10] PHILIP ROSE, FORENSIC SPEAKER IDENTIFICATION 76 (James Robinson ed., 2002).

[11] *Id.* at 68.

Ms. Gavalda had authorized access, which included the facility to send emails in his name from her own machine.

III. TEXTS

A. Emails

I was given electronic access to a large, though selective, database of some 190,000 emails and other texts, including all those authored by Goggin, Widdowson, and Shuy. During my analysis, it became evident that it would have been useful to be able to search in addition a corpus of emails written by Ms. Janet Gavalda in her own voice. However, there was no separate collection of her output available, and so it was only possible to examine those occasional emails authored by her which happened to have been reproduced in other emails sent or received by Goggin, Widdowson, and Shuy, or by other authors included in the database.

My initial analysis focused on three emails: the questioned email sent at 16.30 on July 23rd, and two undisputed emails, one sent by Goggin to Juola at 17.02 and another sent by Widdowson on August 18th to Shuy and Gavalda titled "Chief Exec's Update."

B. Minutes

In addition, I examined eight sets of contemporaneous committee meeting minutes that had been produced by Ms. Gavalda over a fourteen-month period from April 2003 until June 2004.

C. Handwritten Notes

I was also provided with both scanned and transcribed versions of two handwritten entries for July 23rd in a notebook belonging to Mr. Goggin:

> an untimed entry headed "Audit committee report" and consisting of brief notes of a telephone conversation with Widdowson and possibly also Shuy, concerning both an

> "Audit committee report" that had been leaked to the *Guardian* newspaper and an article that was anticipated to appear shortly in another newspaper the *Sunday Times*. This conversation preceded the 14.50 phone call;
>
> a later entry in the notebook headed "D Juola 14.50, 23/07/04" consisting of notes of the topics covered during the 14.50 telephone call.

At a later date, I was provided with notes made by a financial analyst, Caldas, of a telephone conversation with Widdowson two days earlier, on July 21st, also discussing the leak to the *Guardian*.

IV. LINGUISTIC UNDERPINNING

My analysis will focus on linguistic choices and is based on the premise that all language production is rule governed. The underlying linguistic theory is that all speaker/writers of a given language have their own personal form of that language, technically labeled an *idiolect*. A speaker/writer's *idiolect* will manifest itself in distinctive and cumulatively unique rule-governed choices for encoding meaning linguistically in the written and spoken communications they produce. For example, in the case of vocabulary, every speaker/writer has a very large learned and stored set of words built up over many years. Such sets may differ slightly or considerably from the word sets that all other speaker/writers have similarly built up, in terms both of stored individual items in their passive vocabulary and, more importantly, in terms of their preferences for selecting and then combining these individual items in the production of texts.[12]

Thus, whereas any speaker/writer can use any word at any time, what in fact happens is that they make typical and repeated selections and coselections of preferred words, which

---

[12] *See, e.g.*, COULTHARD & JOHNSON, *supra* note 5, at 161; Malcolm Coulthard, *Author Identification, Idiolect and Linguistic Uniqueness*, 25 APPLIED LINGUISTICS 431 (2004); Timothy D. Grant, *Test Messaging Forensics: TXT 4N6: Idiolect Free Authorship Analysis?*, *in* THE ROUTLEDGE HANDBOOK OF FORENSIC LINGUISTICS 508, 508–09 (Malcolm Coulthard & Alison Johnson eds., 2010).

collectively constitute a kind of linguistic fingerprint. Admittedly, this analogy is not precise since a single fingerprint sample has all the necessary information, whereas a single piece of language data has only a minute fraction of the total.

Linguists divide all words into two groups, which they call content, or *lexical*, and formal, or *grammatical*. *Lexical* words are nouns, verbs, adverbs, and adjectives, and it is these words that carry almost all of the message or content of a text, as well as the features of the idiolectal distinctiveness of the author. The *grammatical* words are rather like cement or glue and bind the lexical words together. There are very large numbers of lexical words but only a few hundred grammatical words—thus, a speaker has a very wide choice of content words but a very limited choice of grammatical words. For this reason, linguistic authorship attribution, particularly when the texts involved are short, tends to focus on variation in the selection of the lexical words and on how much overlap there is between authorial choices in known and questioned texts.[13]

Complicating and partly determining the selection of individual lexical words is topic. Given the same basic topic, different speakers/writers will still choose to mention and/or omit different aspects and choose differing lexis to encode any given topic item. Thus, while the occurrence of individual lexical items shared between topically related texts is significant

---

[13] There is, of course, another tradition of authorship attribution represented in this volume by the papers written by Argamon, Juola, Koppel, and Stamatatos. Those works analyze almost exclusively high frequency items, which tend to be word fragments and short grammatical words. *See* Shlomo Argamon & Moshe Koppel, *A Systemic Functional Approach to Automated Authorship Analysis*, 21 J.L. & POL'Y 299 (2013); Patrick Juola, *Stylometry and Immigration: A Case Study*, 21 J.L & POL'Y 287 (2013); Moshe Koppel et al., *Authorship Attribution: What's Easy and What's Hard?*, 21 J.L. & POL'Y 317 (2013); Efstathios Stamatatos, *On the Robustness of Authorship Attribution Based on Character* N-*Gram Features*, 21 J.L & POL'Y 421 (2013). This type of analysis works well with long texts and large collections of texts, as a reading of the articles will confirm, but is unable to cope with very short texts like the questioned email in this case. *See, e.g.*, Argamon & Koppel, *supra*. Both methods have strengths and weaknesses, but I have no doubt that in the future a much more successful method that combines the two will emerge.

in authorship attribution, much more significant is the shared occurrence of coselected items or what linguists call *collocates*, as for instance when *employee* is coselected or collocated with *disgruntled* and/or with *former*.

For example, the questioned email, which is presented in full below (and with the original typos), sets out a situation in which MaxiSoft is ***under attack*** by means of ***rumours*** that are being ***peddled*** by either ***disgruntled employees*** or ***competitors***, these ***rumours*** being concerned with ***revenue*** which, it is claimed, should not have been ***recognised*** and costs which have not been ***fully expensed***.

> As we discussed on the telephone, it would appear that MaxiSoft is currenltly **under attack** from some quarter. There are various rhumours flying around that we anticipate will receive some press coverage over the comming days. We do not know the source of these rhumours, which may be from **disgruntled** (current/former) **employees** or unsuccessful **competitors**.
>
> One of the **rhumours** being **peddled** is that becase of the delay in the finalisation of the HIS contract, we may have **recognised** some **revenue** assocaited with that work. However, I reassure you that such allegations are completely false and that we will refute and defend any such allegations. In addition, all the cost of supporting the HIS bid to date have been **fully expensed**. This issue may not be raised in the press, but I thought I would let you know just in case.

Text 1: Questioned email sent on July 23, 2004 at 16.30

As I noted above, any speaker/writer can use any word at any time and thus for the vast majority of words we can find many instances of their use by large numbers of authors. For simplicity's sake, I will use the Google search engine to illustrate this observation. If we take the eleven word forms I have bolded in the questioned email above and use the Google search engine, we find that all of them are common, some extremely so—there are many millions of hits for most of the items, and even the least used of the group, *peddled*, occurs some 1.5 million times. In other words, none of these word

forms is in any sense rare. See Table 2 below for rounded occurrence figures:

Word	**Google Occurrences**
Under	5 billion
Attack	823 million
Disgruntled	13 million
Employees	727 million
Competitors	185 million
Rumours	50 million
Peddled	1.5 million
Recognised	85 million
Revenue	454 million
Fully	1.2 billion
Expensed	1.8 million

Table 2: Google Word Frequency Searches on Feb. 29, 2012

However, as noted above, what distinguishes speakers/writers and the texts they produce is their coselections. Thus, when we look at some of the coselections in the production of word sequences, we note how quickly the frequency of occurrence decreases as a given phrase lengthens. Here are two examples chosen from the end of the first paragraph of the questioned email:

Words and Phrases	**Google Occurrences**
competitors	185,000,000
unsuccessful competitors	16,100
or unsuccessful competitors	639
employees or unsuccessful competitors	0
* * * * * * * * *	
Disgruntled	12,800,000
disgruntled current	16,800
disgruntled current former	2,570
disgruntled current former employees	55
disgruntled current former employees or	1
disgruntled current former employees or unsuccessful	0

Table 3: Google Word and Phrase Searches, on Feb. 29, 2012

We find this same phenomenon of rapidly reducing numbers of occurrences when we examine the co-occurrence of individual words and short phrases which, although they have not been coselected in a strict linear sequence like those above, still co-occur in the same text. Again, as one would expect, the number of texts sharing a given set of co-occurring items decreases, often dramatically, each time one more item is added. Below, as exemplification, are the cumulative occurrence figures for the first three pairs of collocates pairs that I highlighted in the questioned email. I have presented the search figures in the sequence in which the collocate pairs occur in the email—note an "*" has been used to indicate that I am also including instances where other words occur between the chosen pair of collocates.

Words and Phrases	**Google Cumulative Occurrences**
Under attack	18,000,000
+ Disgruntled * employees	5,500,000
+ Rumours * peddled	0

Table 4: Google Cumulative Searches on Feb. 29, 2012

It is very clear, without needing to include in the search any of the further narrowing coselections of *competitors*, *recognise* + *revenue* and *fully* + *expensed*, that the questioned email has a unique set of lexical coselections—they did not occur together in any of the billions of texts that Google searched.

Thus, we can see clearly that, although in theory anyone can use any word at any time, the topics they choose, the aspects of the topic they decide to focus on, and their preferred linguistic realizations ensure that texts quickly become linguistically unique. This raises the question of who in the software company conceptualized and then linguistically encoded the press problems in ways similar to those used by the author of the questioned email.

A search in the database yielded examples of Widdowson using most of the distinctive vocabulary items in a series of emails written over the period July 16 to August 19, 2004. All of these emails are concerned with the problems raised by the *Guardian* journalist.

In the case of the questioned email, we must also deal with features of typing and copyediting. Some typists are more accurate than others and, because typing is a semiautomated, learned activity, it is possible to characterize less competent typists by the kinds of fingering mistakes they make; I myself frequently missequence, or *metathesize*, letters, and *teh* in particular is a very common mistake in my typing. In addition to typing *mistakes*, i.e. misfingerings, which the typist will recognize as incorrect if s/he rereads what s/he has typed, texts also include what linguists distinguish as *errors*. Errors are nonstandard spellings and grammatical and punctuation choices which the typist does not recognize as such, of which *rhumours*, misspelled identically three times in the questioned email, is an example.

Potentially masking all this idiolectal evidence about a typist is the word-processor's spell-checker, which can save even a poor typist who doesn't proofread and makes not simply typing mistakes but also errors from betraying her/his incompetence. For instance, my spell-checking program automatically corrected the *teh* example above, not once but twice and also warned me that *rhumours* is a nonstandard spelling. Of course, another personal variable is if, when, and to what extent an individual typist actually bothers to use the spell-checker.

V. ANALYSES

A. Stephen Goggin as a Candidate Author

1. Orthography

For its length, the questioned email has a comparatively large number of typing mistakes—four—and one repeated spelling error. There are several categories of mistake and some words have been categorized twice in the listing below because there are alternative possible explanations for the form which has been typed. The first four categories are typing *mistakes*, and the fifth is a spelling *error*:

1. metathesis of letters: *assoc**ai**ted*, *curren**l**tly*
2. omission of letter: *bec**a**se*
3. double keying: *co**mm**ing*
4. additional letter: *curren**l**tly*
5. spelling error: ***rh**umours*, *co**mm**ing*

I have highlighted these items in bold in the email reproduced below:

> As we discussed on the telephone, it would appear that MaxiSoft is **currenltly** under attack from some quarter. There are various **rhumours** flying around that we anticipate will receive some press coverage over the **comming** days. We do not know the source of these **rhumours**, which may be from disgruntled (current/former) employees or unsuccessful competitors.

> One of the **rhumours** being peddled is that **becase** of the delay in the finalisation of the HIS contract, we may have recognised some revenue **assocaited** with that work. However, I reassure you that such allegations are completely false and that we will refute and defend any such allegations. In addition, all the cost of supporting the HIS bid to date have been fully expensed. This issue may not be raised in the press, but I thought I would let you know just in case.

An examination of emails which Goggin affirmed that he had sent from his computer around the period of the questioned email shows that they are completely error free. In particular, the 17.02 email, sent a mere thirty minutes after the questioned email, has no spelling or keying mistakes. In other words, Goggin did not send mistake- or error-filled emails from his computer.

2. Opening and Closing

The questioned email has an in-text opening heading of "Strictly Private and Confidential" in bold. There are no examples of this heading in any Goggin emails. The message closes with "Best Regards," yet the message sent to the same recipient, Juola, only half an hour later at 17.02 and accepted as authentic by Goggin ends simply with "Regards." Indeed, an analysis of all the emails sent by Goggin to Juola in the preceding six months shows that some eighty percent of them end simply with "Steve," and in the twenty percent of emails where there is a closing, it is, as in the 17.02 authentic email, invariably an unmodified "Regards." There are no examples of "Best Regards." In other words, neither the opening nor the closing of the questioned email were choices that Goggin made in his emails at the time.

3. Lexical Choices

Three distinctive lexical choices in the questioned email are *disgruntled*, *peddled*, and *under attack;* none of them occur in any emails Goggin accepts as authentic. Neither does Goggin,

who was a salesman, not an accountant, send any emails with the phrases *recognising revenue* or *fully expensed*.

4. Finding Regarding Goggin as a Candidate Author

The linguistic choices made by the author of the email are not consistent with those instanced in Goggin's other emails.

B. Others as Candidate Authors

1. Content and Expression

The language of the questioned email has significant lexical links with that of the person(s) who briefed Goggin in the earlier telephone call already mentioned above, which was recorded in his notebook as "audit committee report." This person must have been Widdowson or Shuy because Goggin says they were the only other participants. Relevant words and phrases in Goggin's notes on this briefing are highlighted in bold in the extract below and can be compared with the same items occurring in the immediately following extracts taken from the questioned email:

> Someone trying to suggest that **we have recognised revenue**
>
> Take so long – **delay**
>
> **Under attack** **competitor/disgruntled employee**

> As we discussed on the telephone, it would appear that MaxiSoft is currenltly **under attack** from some quarter. There are various rhumours flying around that we anticipate will receive some press coverage over the comming days. We do not know the source of these rhumours, which may be from **disgruntled** (current/former) **employee**s or unsuccessful **competitor**s.
>
> One of the rhumours being peddled is that becase of the **delay** in the finalisation of the HIS contract, **we**

> may **have recognised** some **revenue** assocaited with that work.

We can see, highlighted in the text of the questioned email, all the important lexical items from the briefing notes not simply recurring but recurring in the same collocational groupings. In other words, the author(s) of these two messages which are closely related in time, the one spoken and the other written or dictated, is/are choosing to present the company's problem with the press within the same conceptual framework: that is, not as a legitimate, although admittedly annoying and distracting, investigation by a journalist but as a motivated "attack" either by aggrieved insiders or by those competing for contracts. Not only is the conceptualization of the problem in the email the same as in the telephone briefing but so also is its lexical encoding: "under attack," "disgruntled employees/competitors," "delay," and "we have recognised revenue."

These linguistic facts strongly suggest the possibility of single authorship; in other words, whoever briefed Goggin earlier in the day also authored the questioned email. A search of Shuy's emails did not produce examples of him using any of the central lexis used in the questioned email. Widdowson, however, does use much of this vocabulary.

Two days before the telephone briefing of Goggin, Widdowson briefed company analyst Caldas. In this briefing, the company is also presented as *under attack*, an attack which is characterized as *malicious* and which involves someone who is *feeding* to the *press* claims about *revenue* having been *recognised* before a contract has been signed. Caldas's notes include the following items

> **disgruntled employee** dismissed False letter to **GRD** [*Guardian*]
>
> *feeding* to *jornos*
>
> why **rev recognised** before signed?
>
> subject direct *malicious* **attack**
>
> also signed & **RR'd** [revenue recognised]
>
> co **under attack**

In an email sent to a market analyst on August 13th, Widdowson again refers to the problems with *The Guardian* and

again characterizes the encounter as *malicious* and as an **attack**: "[t]he last few weeks have really been quite extreme and we appreciate the quality of the advice provided and your dogged determination to see off this *malicious* **attack**."

Five days later on August 18th, Widdowson circulated a text entitled "CEO Statement" in which he referred again to the problems with the *Guardian* journalist and used six of the lexical items that occurred in the questioned email, including the same collocations in the same close proximity:

> Having had the initial *malicious* **rumour** planted . . . .
>
> Our response to this direct **attack** was however measured. . . .
>
> [T]here is little evidence that the *malicious* **rumours peddled** by the Guardian journalist have had any material effect on the perception of MaxiSoft in the healthcare IT supply market with either existing or prospective customers. It is an interesting contrast to note that most in the supply market see straight through the recent newspaper 'noise', speculating that it emanates from a **disgruntled former employee** seeking to further a particular selfish personal agenda.

We can compare this lexical encoding with the questioned email:

> We do not know the source of these **rhumours,** which may be from **disgruntled** (current/**former**) **employee**s or unsuccessful competitors.
>
> One of the **rhumours** being **peddled** is that becase of the delay in the finalisation of the HIS contract, we may have recognised some revenue assocaited with that work.

These particular lexical items do not co-occur in any other company emails, let alone in such close proximity to each other.

Widdowson also uses *peddle* on other occasions to disparage communications: in an October 1st email he refers to information having "been *peddled* around already" and on October 12th he characterises a Mr. Steer as "*peddling*."

In addition, Widdowson, an accountant, unlike Goggin, does write frequently about *recognising revenue* and uses the expression "*fully expensed*." In an email sent to Goggin on August 6th and titled "Message re *Guardian* Update," Widdowson writes, "The balance of the SPfiN-related **revenue recognised** in 04 was in respect of earlier deliverables of existing product and services," and on July 16th, a week before the questioned email was sent, in an email entitled "draft script for our friend at the *Guardian*," Widdowson included the observation that "the value of R+D spend is confirmed as **fully expensed**." Finally, while the heading of the questioned email **Strictly Private and Confidential** is very rare in company emails, it does occur in another email about this same *Guardian* investigation sent by Widdowson to Gavalda and then forwarded by Gavalda to the Executive Board on August 13, 2004:

MaxiSoft - THE HEALTH iNNOVATOR

Strictly private and confidential

In other words, all of the core vocabulary that is highlighted in the questioned email below is vocabulary that Widdowson also uses in other emails concerned with the problem of press coverage:

Strictly private and confidential

As we discussed on the telephone, it would appear that MaxiSoft is currenltly under attack from some quarter. There are various rhumours flying around that we anticipate will receive some press coverage over the comming days. We do not know the source of these rhumours, which may be from disgruntled (current/former) employees or unsuccessful competitors.

One of the **rhumours** being **peddled** is that becase of the delay in the finalisation of the HIS contract, we may have **recognised** some **revenue** assocaited with that work. However, I reassure you that such allegations are completely false and that we will refute and defend any such allegations. In addition, all the cost of supporting the HIS bid to date have been **fully expensed**. This

issue may not be raised in the press, but I thought I would let you know just in case.

To summarize: six central vocabulary choices made by the author of the questioned email occur in other emails on the same topic written by Shuy and three of them also occur in both the Goggin notes of the telephone conversation and in Caldas's notes. By contrast, there are no examples of Goggin making any of these vocabulary choices in his emails at this time.

Words and Phrases	**Goggin emails**	Goggin Notes	**Question-ed email**	**Widdowson *Guardian* emails**	**Caldas notes**
attack	NO	**YES**	**YES**	**YES**	**YES**
Recognise(d) + revenue	NO	**YES**	**YES**	**YES**	**YES**
Disgruntled + employee(s)	NO	**YES**	**YES**	**YES**	**YES**
Peddle + rumour(s)	NO	NO	**YES**	**YES**	NO
fully expensed	NO	NO	**YES**	**YES**	NO
Strictly private and confidential	NO	NO	**YES**	**YES**	NO

Table 5: Comparison of Occurrences of Six Crucial Linguistic Encodings

2. Finding Regarding Others as Candidate Authors

Significant lexical choices in the questioned email are **consistent** with choices Widdowson makes elsewhere, particularly in emails about the problem with the *Guardian* journalist. In addition, these coselections do not occur in emails sent by anyone else and so are **distinctive**.

3. Orthography

While the content and expression of the questioned email share important features with other texts authored by Widdowson, the frequency of mistakes is certainly atypical of his normal production, which displays only the occasional mistake like "furture" in the August 13th document. Thus, Widdowson is not an obvious candidate for typist of the email.

I was asked to consider the possibility that the questioned email had been dictated to Ms. Gavalda and, as noted above, I was provided with a set of her minutes. The task of typing a dictated email is in some ways very similar to taking minutes—in both cases, it is the conversion of the spoken content of others into typewritten form.

A comparison of the type and frequency of the mistakes in the questioned email with those in a randomly selected set of Ms. Gavalda's minutes produced in September 2003 identifies her as a candidate typist. Below are some mistakes and errors from these minutes. It will be seen that she makes mistakes in all of the five categories identified above:

1. metathesis of letters: p**al**ce; stren**tg**hs; addi**otn**al; terr**orit**ires; surp**ir**ing; abil**ti**y; ju**en**; fa**ri**;

2. omission of letters: announ**cm**ent; arrange**mn**t; laun**ci**ng; tak**ig**; d**ic**ussion; a**co**untable; termi**nt**e; **rs**ource po**st**ion; s**ta**tegy; surpi**ri**ng; expec**ti**ons; ri**ni**ng; contrac**tr**;

3. double keying: h**aa**ve; th**ee**m;

4. additional letters: dec**f**ision; et**e**c; meetin**i**g; dama**n**ges; incentivisi**e**d; analysts**t**; fina**j**ncial; happen**d**ing; rsour**e**ce; announcmen**e**t; re**n**venue; prod**oc**uct;

5. spelling error: hus**s**le, (hus**t**le); dis**b**ute, (dis**p**ute); **ph**ar**s**e, (**f**ar**c**e);

To convey an impression of the sheer frequency of Ms. Gavalda's mistakes, I have pasted below an extract from another set of her minutes dated April 7, 2004:

> PM – updated on the TAW note. Have asked for the **fucnational** heads to prepare a little script and have had two in, awaiting the rest.

Discssuion whether one or individual – one but will include individual ones as well.

Can get **stared** on the employee representatives, ought to get going – RR taking that forward around payrolls. Can be used for redundancy as well.

Making good progress with carrying on the **templte** meetings (**identifyinig** headcount reductions). Driven by accounts -0 drop date 26 April 2004. Needs to be done within the next week. Meetings agreed. Still waiting for date from RK. Can it be done virtually – Tuesday via telephone with. Sibsons are over in India – can do it over in Chennai.

TAW – make sure everyone is clear on the process. **Logalical** process of – database – mapped everyone to the new structure, all arrived on Monday. TAW, SPG and PM – biggest concern is in respect of NPfIT engagement and RK spoken to DR – **thinkging** moving forward – major conflicts and outstanding issues – who is involved where and what does this mean in respect of the mappings.

NP **struutre** needs to encompass the central solution team (software delivery team) that sits between the **rpodocut** business and NP team (**deploymnete** or **engagtement** team)

Confirm structures and names against the structures

TAW – np **strucurre** – most difficult area – where are we up to and when will it be finished. DR spoke to PM – RK, DR and RB – main area with regard to product delivery components. Central solution team is now effectively in 3 **component**, solution definition (identifying futures and obligations), manufacture and design and two **componesnt** solution delivery and support.

4. Finding Regarding Ms. Gavalda as a Candidate Typist

The range and nature of the mistakes in the questioned email are **compatible** with the mistakes that Ms. Gavalda makes in her contemporaneous minutes. In addition, the frequency is **distinctive**.

VI. OPINIONS

Opinion 1: The distinctive linguistic features of the questioned email are **not compatible** with Mr. Shuy's usage in other attested emails.

Opinion 2: The distinctive linguistic features of the questioned email are **not compatible** with Mr. Goggin's usage in other attested emails.

Opinion 3: The linguistic features of the questioned email are **compatible** with Mr. Widdowson's usage in other attested emails and with items in the notes made by recipients of two telephone conversations. These linguistic features are **distinctive.**

Opinion 4: The orthographic features of the questioned email are **compatible** with Ms. Gavalda's usage in contemporaneous minutes. These features are **distinctive.**

VII. CODA

Essentially, my expert report ended at this point, and the evidence I gave in court was based closely on it. However, I was unhappy that my evidence lacked any discussion of the frequency or rarity of the linguistic items I had claimed were crucial to the attribution of authorship. The analysis therefore was vulnerable to a cross-examiner suggesting that my analysis was not replicable and thus its credibility depended too much on my own credibility as an expert.

By a fortunate coincidence after I wrote the draft of my Workshop paper, I became aware of the work of doctoral student David Wright, who is using the Enron email database to develop computerized authorship attribution tools. Like me,

Wright is interested in the classificatory and attributory value of lexical as opposed to grammatical items. Thus his analyses, like mine, exclude function words such as articles, determiners, pronouns, and prepositions, which figure prominently in the analytic tools of many of the other authors in this volume.

Wright set out to investigate the degree of lexical similarity between different datasets and authors by examining the number of lexical types shared in the emails of selected Enron employees and then using the simple similarity metric Jaccard's coefficient[14] to evaluate the significance of his findings.

In an early exploratory study, he focused on the emails produced by a closed set of four Enron traders.[15] He found:

> [Even though] the writers were all men of working age, all shared occupational and institutional goals, were writing on largely the same topics and within the same register, when [their sets of emails] were compared with each other the Jaccard similarity scores were low. [This clearly indicated] that, despite being socially and professionally very similar, the four authors had their own distinctive and identifiable lexicons.[16]

Blind testing demonstrated that the four authors could indeed be distinguished from each other by means of their individual lexical choices. This clearly has important implications for forensic authorship identification and attribution. Wright tested his method by setting out to match sets of 100 emails to the original author and was able to do so with a very high success rate.[17] In my case, there were by this point only two potential authors, Widdowson and Goggin (Shuy having already been

---

[14] This method is discussed in some detail in Grant's paper. Tim Grant, *TXT 4N6: Method, Consistency, and Distinctiveness in the Analysis of SMS Text Messages*, 21 J.L. & POL'Y 467, 482 n.44 (2013).

[15] David Wright, *Existing and Innovative Techniques in Authorship Analysis: Evaluating and Experimenting with Computational Approaches to "Big Data" in the Enron Email Corpus*, 3D EUR. CONF. INT'L ASS'N FORENSIC LINGUISTS, Oct. 2012.

[16] David Wright, *Measuring Lexical Similarity for Authorship Identification: An Enron Email Case Study*, 28 LITERACY & LINGUISTIC COMPUTING (forthcoming 2013).

[17] *Id.*

discounted) and there is only one email, so the statistical route is not open to me. However, the question remains of whether the single email contains sufficient distinctive lexical information to make an attribution.

In undertaking this later analysis, I drew on a methodology proposed in Grant's article in this volume—a methodology which he developed for categorizing text messages.[18] Like me, Grant was working on a case with only two possible authors, but his data consisted of text messages.[19] Working from the known to the unknown, he took the two sets of known text messages and examined them in order to discover "whether there were features that discriminated consistently to some degree between the two writers in their known texts."[20] Grant only focused on features which were used predominantly by one author or the other and used "a rate of more than sixty-six percent of its total occurrence" as his criterion.[21]

Because in my case there was only one questioned email but vast numbers of comparison emails, I decided to restrict analysis to all and only the emails sent during a seven-month period, three months before and three months after the month in which the questioned email was sent. What I set out to do was, like Wright, to discover whether the lexical selections made by the author of the email were compatible with the usage of Goggin or of Widdowson. I decided to use Grant's criterion of majority usage to classify those items that occurred in both sets of emails as being characteristic of the usage of one of the authors, but I raised the required classificatory level of usage to a minimum of seventy-five percent.

My task was further complicated because while Grant had roughly equivalent sets of texts to compare, Goggin had produced over 2.5 times as many emails as Widdowson in the seven-month period—3,150 as compared with 1,234. For this reason, the raw scores for Goggin were reduced by sixty percent to normalize the frequencies before the comparison was made.

---

[18] Grant, *supra* note 14.

[19] *Id.*

[20] *Id.* at 480.

[21] *Id.*

Then the usage scores for all of the lexical items in the questioned email were compared. The scores for some items showed little difference in usage, but the relative frequencies of others were markedly different. Table 6 below shows first the items that were used only or more frequently by Widdowson (indicated in bold), then the Goggin items. It will be evident that there are many more distinctively Widdowson items in the list, and it becomes clear that the questioned email was composed using many more Widdowson than Goggin items.

Features (Normalized)	**Total in 1243 TW emails**	**40% of Total in SG emails**	**Total**	**Percent in TW emails**	**Percent in SG emails**
Recognise + revenue	**7**	**0**	**7**	**100**	**0**
Peddle	**2**	**0**	**2**	**100**	**0**
Attack	**1**	**0**	**1**	**100**	**0**
Coming Days	**1**	**0**	**1**	**100**	**0**
Competitor	**1**	**0**	**1**	**100**	**0**
Disgruntled	**1**	**0**	**1**	**100**	**0**
Former employee	**1**	**0**	**1**	**100**	**0**
Fully expensed	**1**	**0**	**1**	**100**	**0**
Rumour	**1**	**0**	**1**	**100**	**0**
Strictly Private and Confidential	**1**	**0**	**1**	**100**	**0**
It would appear	**7**	**0.4**	**7.4**	**95**	**5**
To date	**14**	**1.2**	**15.2**	**92**	**8**
Delay	**15**	**2**	**17**	**88**	**12**
Best Regards	**3**	**0.4**	**3.4**	**88**	**12**
Press coverage	**2**	**0 .4**	**2.4**	**88**	**12**
In addition	**10**	**2.4**	**12.4**	**80**	**20**
Currently	**16**	**4**	**20**	**80**	**20**
Employee	**3**	**0.8**	**3.8**	**79**	**21**
Issue + raise	**3**	**0.8**	**3.8**	**79**	**21**
Just in case	1	3.6	4.6	22	78
Reassure	0	0 .4	0.4	0	100
Completely	0	2	2	0	100

Table 6: Preferred Vocabulary Items for Widdowson and Goggin

What is evident in the highlighted version of the questioned email below is that a significant amount of the lexis is lexis that occurs predominantly in emails written by Widdowson (indicated in **bold)**, whereas only three items are typical Goggin items, (indicated in *italic*).

> Strictly private and confidential
>
> As we discussed on the telephone, it would appear that MaxiSoft is currenltly under attack from some quarter. There are various rhumours flying around that we anticipate will receive some press coverage over the comming days. We do not know the source of these rhumours, which may be from disgruntled (*current*/former) employees or unsuccessful competitors.
>
> One of the rhumours being peddled is that becase of the delay in the finalisation of the HIS contract, we may have recognised some revenue assocaited with that work. However, I *reassure* you that such allegations are *completely* false and that we will refute and defend any such allegations. In addition, all the cost of supporting the HIS bid to date have been fully expensed. This issue may not be raised in the press, but I thought I would let you know *just in case*.
>
> Best Regards

CONCLUSION

Unlike Forensic Phoneticians, forensic linguists are never going to have reliable population statistics to enable them to talk about "the frequency or rarity of particular linguistic features." I would argue, however, that the work of Wright and Grant opens a way to derive reliable and usable data about individual linguistic usage that can be applied in cases of authorship attribution. With tools like these, linguists can begin to make statements about frequency and likelihood of occurrence and, in cases where the data permits a Jaccard analysis, provide rigorous probability statistics.