Donlan, L. and Nini, A. (2022). Forensic authorship analysis of the Ayia Napa rape statements. In Picornell, I., Perkins, R., and Coulthard, M. (eds) *Methodologies and Challenges in Forensic Linguistics Casework*. Hoboken, NJ: Wiley-Blackwell.

Forensic authorship analysis of the Ayia Napa rape statements

Lisa Donlan & Andrea Nini

1. Introduction

On 17th July 2019, a nineteen-year-old British woman, who is the defendant in the legal case considered in this chapter, attended a clinic in Ayia Napa, a resort town in Cyprus, and reported that she had been gang-raped by a group of Israeli men. The defendant claimed that she had been engaging in consensual sexual intercourse with one of the men in his hotel room when his friends entered the room without her consent and raped her. After she managed to escape, the defendant's friends took her to the medical clinic attached to the hotel. The doctor at the clinic decided to call the Cyprian authorities, and the defendant subsequently gave a statement documenting her attack. The next day, Cyprus police arrested twelve men on suspicion of raping the defendant, all of whom denied that the sex acts that took place were non-consensual. Seven of the men were later released, but, ten days later, five remained in police custody.

However, on 27th July 2019, the defendant was asked by the Cyprus authorities to attend the police station, ostensibly to clarify inconsistencies in her statement. The police station visit ended with the defendant retracting the original statement attesting to her rape and allegedly writing a new handwritten statement in which she confessed that the rape accusation was a lie. As a consequence, her alleged attackers were released without charge and allowed to return to Israel, while the defendant was arrested and charged with "public mischief" for making a false rape statement.

At her subsequent trial, the prosecution claimed that the rape retraction statement which led to the charge of public mischief had been willingly written and signed by the defendant. However, the defendant testified that her original statement documenting her rape was a true account of her experience. She maintained that, on the day she signed the retraction statement, she was afraid for her life, after having been detained by the police for hours and denied a lawyer. Although the fact that the defendant wrote the statement in her own handwriting is undisputed, the defendant claimed that she had been forced to write the

statement by the Cyprus police and that an officer dictated what she should write. To support this claim, she told the courts:

> This [retraction statement] is not in proper English. This is in Greek English. [...] It doesn't make grammatical sense [...] All the way through there isn't one sentence an English person would write (quoted in BBC News 2019, p.n.p.).

In other words, then, a key part of the defence's case rested on the claim that the language used in the retraction statement was not consistent with language likely to be used by the defendant (a native speaker of British English). Because no video or audio recording of either police interviews or statements was available, the only evidence the court could rely on was linguistic. This is a situation that is virtually identical to the very early cases that led to the birth of forensic linguistics, such as the Timothy Evans case or the Derek Bentley case, in which the questioned documents were precisely disputed confessions by the defendants that the linguistic analyses revealed were coerced by the police. Both of these cases occurred in the 1950s when police interviews and statements in the UK were not video or audio recorded.

At this point in the proceedings, the second author of this chapter was hired by the defence as an expert witness and asked to analyse the authorship of the retraction statement. The retraction statement was three paragraphs long. The first and last paragraphs, both in the defendant's own handwriting, were dictated to her by a police officer and this fact was undisputed by both parties. The defence instructed that it was only the second paragraph that was contested, and this is reproduced below as accurately as possible from the handwritten retraction statement:

> The report I did on the 17th of July 2019 that I was raped at ayia napa was not the truth. The truth is that I wasnt raped and everything that happened in that appartment was with my consent. The reason I made the statement with the fake report is because I did not know they were recording & humiliating me that night I discovered them recording me doing sexual intercourse and I felt embarrassed so I want to appologise, say I made a mistake.

The defence team intended to determine whether it was possible to prove the defendant's claim that she did not write the disputed paragraph. The defence's question was reformulated

by the linguist to match the reasoning of forensic linguistics as well as its limitations. From a forensic linguistics perspective, the pivotal research question was:

> ***Did the defendant compose this paragraph in her own words or was it dictated to her by a local police officer?***

which could be rephrased as an expression of likelihood that the linguistic evidence lends support to the prosecution hypothesis:

H_0 *The disputed paragraph in its entirety was composed by the defendant in her own words*

or the defence hypothesis:

H_1 *All or some parts of the disputed paragraph were not composed by the defendant in her own words but were dictated to her by someone who spoke English as a second language*

The following sections detail the methodology and the results of the forensic linguistic authorship analysis that was carried out to uncover the linguistic evidence in support of one or the other hypothesis.

2. Methodology

Ostensibly, the authorship question for this case is one of **authorship verification** (Koppel & Schler 2004; Koppel et al. 2012), or the confirmation or exclusion that a single named suspect (in this case, the defendant) is likely to be the creator of one or more texts of uncertain authorship (the retraction statement).

However, there were several key issues which prevented authorship verification methodologies from being of use in this case. Authorship verification is typically conducted using computational techniques that rely on the frequency with which linguistic features occur. This means that a relatively large body of existing texts known to be written by the suspect author are needed so that these identifying frequency profiles can be determined. In addition, this body of texts must be roughly comparable in terms of **register** to the disputed text. This is because it is widely acknowledged that an individual's language use will greatly vary across different communicative situations, known as 'registers' (Biber 1995; Biber 2012). The language one uses in a text message to a friend, for instance, differs markedly

from the language one uses when composing a formal letter to a superior. Therefore, to avoid the confounding effects of register variation, the comparison texts should belong to the same or very similar registers as the disputed texts. In other words, to perform authorship verification for this case, it would have been necessary to obtain a large number of other texts known to be penned by the defendant that were roughly comparable contextually to a police statement. It was not possible to obtain such a body of texts and thus traditional authorship verification methodologies could not be pursued in this case.

Instead, this research question was approached as an **authorship profiling** task. In contrast to authorship verification, which compares the linguistic style of a disputed text to the linguistic style of the suspect author, authorship profiling involves determining the likely characteristics of the person who composed the disputed text from their use of linguistic constructions. Authorship profiling is a branch of authorship analysis which is heavily entwined with the study of linguistic variation in society, or **sociolinguistics**. Broadly, sociolinguistic research is concerned with the investigation of how an individual's language is influenced by social and cultural factors. Gender, age, ethnicity, socioeconomic class, sexuality, native language, and location of birth are all factors which can correlate with an individual's language use (Labov 2001; Tagliamonte 2012; Chambers and Schilling-Estes 2013). Authorship profiling draws on linguistic research about the relationship between language use and social and cultural factors to build a socio-demographic profile of the likely background of the author of the text. For instance, a forensic linguistic authorship profile of the Unabomber's known writings predicted the age, education level, geographic origin, and places of residence of Ted Kaczynski, the man who was eventually convicted of the crimes (Shuy 2005, pp.181–182). In this case, the authorship question centres on whether there is linguistic evidence to support the idea that the text was authored by a speaker matching the linguistic profile of the defendant, an educated 19-year-old woman who was a native speaker of British English, which is consistent with H_0, or alternatively with the profile of a speaker of English as a second language, which is more consistent with H_1.

2.2 Authorship profiling method

The first step of the analysis consisted of the manual analysis of the disputed text to identify linguistic **constructions**, from an individual word to entire sentences, that can exhibit variation. One of the advantages of authorship profiling is that texts need not be of a

considerable length to provide sufficient data for analysis. From the eighty-five word statement, we were able to identify five constructions, discussed in detail in section 3, which we believed would be useful for building a profile of the author of the disputed text:

1. [DO [REPORT]]
2. [BE *not the truth*]
3. [APARTMENT]
4. [DISCOVER [NP V-*ing*]]
5. [DO [*sexual intercourse*]]

The notations used in this chapter follow conventions from linguistics. A word in capital letters indicates a **lexeme** or **lemma**, thus including all of the forms of a particular word. For instance, when we write DO, we are referencing *do* but also the third person singular present form *does*, the past tense form *did*, the continuous form *doing*, and the perfect form *done*. Meanwhile, square brackets indicate a linguistic construction and its internal syntactic structure, italics is used for word forms (as opposed to lemma forms), and abbreviations of syntactic constituents are indicated with two capital letters (e.g. NP = Noun Phrase).

The second step of the analysis was to investigate the five constructions identified. Two primary methods were used in this step. Firstly, we drew on the findings of existing literature concerning the constructions of interest. Secondly, we used **corpus linguistic** methodologies to explore the constructions in greater detailer. Corpus linguistics focuses on the scientific study of language through the analysis of a large collection of naturally occurring texts (referred to as a *corpus* or, in the plural form, *corpora*), typically using computational methods. The use of corpora is particularly important in forensic contexts because it can lead to the estimation of typicality and probabilities. In other words, it is possible to determine how common a construction is, or how likely it is for someone of a particular socio-demographic profile to use a construction, by looking at the relative frequency at which it occurs in appropriate corpora.

For instance, in the Derek Bentley case, the posthumous pardon was based in part on corpus evidence (Coulthard 2013). Indeed, it was shown that the language used in his confession (such as the repeated used of *I then*) was rare in general English and in a corpus of witness statements taken from the general public but very common in a corpus of writings by policemen (Coulthard 2013). This provided evidence to support Derek Bentley's testimony

that the police had "helped" him with his statement at the time and that the statement was not a verbatim written account of Bentley's spoken account of events, as claimed by the police.

The choice of which corpus to use for comparison has to be justified by the specific needs of a case. After considering the distinctive properties of this case, namely the involvement of a speaker of British English and the fact that the disputed text is a police statement, the following set of corpora was selected as comparison data:

Corpus	**Acronym**	**Size and composition**[1]	**Platform used to access the corpus**	**Reason for selection**
Timestamped Jozef Stefan Institute (JSI) web corpus 2014-2019	JSI	47 billion words of news articles in English from the Internet	https://www.sketchengine.eu/	One of the largest corpora available.
English Web 2015 (enTenTen15)	Ten15	18 billion words of English texts collected from the Internet	https://www.sketchengine.eu/	One of the largest corpora available
British National Corpus	BNC	112 million words of written and spoken English	https://www.sketchengine.eu/	Corpus balanced for registers and with data only from native speakers of British English
Spoken British National Corpus 2014	SBNC14	11 million words of spoken English	https://cqpweb.lancs.ac.uk/	More recent version of the BNC with data only from native speakers of British English
British Law Report Corpus	BLaRC	10 million words of judicial decisions in English	https://www.sketchengine.eu/	Sample of legal English
Global Web-Based English corpus	GloWbE	1.9 billion words of English web data	https://www.english-corpora.org/	Large corpus of all major global varieties of English
News on the Web Corpus	NOW	9 billion words of web-based newspaper and magazine data	https://www.english-corpora.org/	One of the largest corpora available.
CORE corpus	CORE	50 million words of web data	https://www.english-corpora.org/	Corpus categorised by registers
Corpus of Contemporary American English	COCA	560 million words of written and spoken English	https://www.english-corpora.org/	Corpus with data produced by native speakers of American English and balanced for registers

[1] All descriptions reflect the nature of the corpora at the time of access in October 2019. Many of these corpora are updated on a rolling-basis and may now be significantly larger than indicated here.

3. Results

C1: [DO [REPORT]]

In corpus linguistics, a *collocation* is defined as the relation connecting two words that occur in proximity to each other far more frequently than expected by chance alone (Sinclair 1991). For instance, the word *fast* is a collocate of *food*, with the phrase *fast food* denoting an easily-prepared processed meal served in hospitality settings. In contrast, the word *rapid*, a near-synonym of *fast*, does not correlate with *food*. Although the phrase *rapid food* is grammatically correct, most native speakers of English would not recognise this is an acceptable construction.

It is fairly well established that non-native speakers of a language often have problems producing constructions that are consistent with the collocational behaviour of words in the target language (Wray 2000; Nesselhauf 2003; Futagi et al. 2008; Granger and Bestgen 2014). In other words, although non-native proficient speakers of a language can produce *grammatical* sentences, they might not be able to produce *acceptable* sentences that are idiomatic. In the disputed paragraph of the retraction statement, the following sentence occurs:

> The **report I did** on the 17th of July 2019 that I was raped at ayia napa was not the truth.

Here, the noun *report* is used as the object of *did*, the past tense form of the verb *do*. In order to explore the extent to which this construction is likely to be used by a native speaker of English, we first investigated how frequently the noun *report* collocates with forms of the verb *do* (as in the example taken from the statement above) and how frequently the noun *report* collocates with forms of the verb *make* (as in *the report I made*).

Because such an investigation requires the ability to undertake syntactic searches, only the corpora hosted on the Sketch Engine platform were used in this analysis since the Word Sketch function of this web application allows for syntactic analysis.

Variant	JSI	Ten15	BNC	BLaRC
[DO [REPORT]]	26% (24,302)	25% (9,462)	32% (77)	0% (0)
[MAKE [REPORT]]	74% (69,852)	75% (28,207)	68% (165)	100% (49)

Table 1 - Relative frequency and frequency (in brackets) of the two variants [DO [REPORT]] and [MAKE [REPORT]] in the corpora chosen for analysis. The instances were retrieved using the Word Sketch function of Sketch Engine.

The results clearly indicate that the unmarked (or standard) default preference in English is for *report* to be controlled syntactically by *make*. Across the four corpora explored, *report* is more likely to correlate with *make* than *do* in between 68% to 100% of instances. In other words, then, the construction used in the report, in which the noun *report* collocates with the verb *do* is unusual among native speakers of English.

Moreover, it is very important to note that the figures reported in Table 1 also include instances of *do* as an auxiliary verb which are irrelevant to this analysis. In the disputed statement, the past tense form of DO is used as a main verb (*I did the report*). However, while the results returned by the corpora searches include relevant instances of DO as a main verb, there are also a notable number of instances where DO is used as an auxiliary verb (e.g. "**Did** the mainstream media **report** on this policy move and the debate about it?" JSI). Constructions where DO functions as an auxiliary verb are very different and non-comparable to the type of construction used in the report. The corpora used, however, do not always recognise the difference between DO functioning as a main verb and DO functioning as an auxiliary, and thus auxiliary instances could not be automatically excluded from the search. Ideally, these irrelevant auxiliary constructions would have been filtered and removed from the results manually. However, this was not possible due to the high number of instances retrieved. What this means for the analysis is that the findings above overestimate the relative frequency of DO correlating with the noun REPORT, and the real probability of observing [DO [REPORT]] is therefore even lower than what is shown in Table 1.

In sum, the use of [DO [REPORT]] instead of [MAKE [REPORT]] is marked and atypical in British English and other varieties of English.

C2: [BE not the truth]

In addition to the unusual [DO [REPORT]] construction, there is a second noteworthy construction in the opening sentence of the disputed paragraph of the retraction statement:

> The report I did on the 17th of July 2019 that I was raped at ayia napa **was not the truth**.

To articulate that the original statement provided was false, the author has used the construction [BE *not the truth*], where *truth* is a noun (and this is, therefore, a nominal construction). However, an alternative way of articulating this sentiment would be to use *truth* in its adjectival form [BE *not true*]. Using the adjectival construction, the statement would have read as:

> The report I did on the 17th of July 2019 that I was raped at ayia napa **was not true.**

A corpus analysis was performed to determine which of these two constructions are more likely to be used by English speakers.

Variant	JSI	Ten15	BNC	SBNC14	BLaRC	GloWbE	NOW	CORE	COCA
[BE not the truth]	1.9% (2,880)	2.1% (854)	2.4% (13)	1.6% (2)	0% (0)	2.2% (300)	1.8% (677)	1.9% (8)	2.1% (104)
[BE not true]	98.1% (151,026)	97.9% (40,751)	97.6% (526)	98.4% (120)	100% (41)	97.8% (13,425)	98.2% (36,885)	98.1% (420)	97.9% (4,897)
Queries	[lemma="be"] [word="not"] [word="the"] [word="truth"] [lemma="be"] [word="not"] [word="true"]								

Table 2 - Relative frequency and frequency (in brackets) of the two variants [BE *not the truth*] and [BE *not true*] in the corpora chosen for analysis. The queries used are reported at the bottom in CQL.

The results in Table 2 show that the nominal construction, of the kind used in the retraction statement, is far less likely to be used in English than the alternative adjectival construction. In none of the corpora does the nominal construction account for more than 2.4% of instances returned. It is important to note that the British Law Report Corpus (BLaRC), which consists of over ten million words of judicial decisions, contains no instances of the nominal construction. This finding strongly suggests that the nominal construction is not simply a more formal variant of the adjectival construction.

In conclusion, although the construction [BE *not the truth*] is grammatical, it is far less likely to be used in British English and other varieties of English than its adjectival variant [BE *not true*].

C3: [APARTMENT]

In the second sentence of the retraction statement, the author refers to the location of the alleged assault:

> The truth is that I wasnt raped and everything that happened in that **appartment** was with my consent.

The noun APARTMENT is predominantly an American English form (OED Online 2020). In British English, the variant typically used is the noun FLAT. Therefore, we focused on determining the likelihood of the American English APARTMENT being used by a speaker of British English.

Specifically, we explored the relative frequency of APARTMENT and FLAT in corpora which were either searchable by geographical region or which were collated entirely from texts produced by speakers of British or American English.

Variant	Ten15 (UK)	BNC	SBNC14	GloWbE (UK)	NOW (UK)	Ten15 (.us)	COCA	GloWbE (US)	NOW (US)
[apartment]	41% (10,205)	41% (1,814)	20% (190)	43% (10,143)	40% (36,490)	92% (7,937)	80% (43,229)	76% (13,704)	88% (104,827)
[flat]	59% (14,512)	59% (2,618)	80% (783)	56% (13,185)	60% (54,358)	8% (734)	20% (10,756)	24% (4,409)	11% (14,204)
Queries	[lemma="apartment"&tag="N.*"] [lemma="flat"&tag="N.*"]								

Table 3- Relative frequency and frequency (in brackets) of the two variants [APARTMENT] and [FLAT] in the corpora chosen for analysis. The queries used are reported at the bottom in CQL. The shaded area of the table shows the US data while the non-shaded area shows the UK data.

The results in Table 3 suggest that, on average, it is much less likely for a native speaker of American English to use the British noun FLAT than it is for a British person to use the American noun APARTMENT. Nonetheless, the results clearly indicate a preference for native speakers of British English to use FLAT, especially as demonstrated by the spoken data in the SBNC14.

In conclusion, someone with the linguistic background of the defendant (that is, a native speaker of British English) is marginally less likely to use the noun APARTMENT, which was seen in the retraction statement, when compared to the noun FLAT, which denotes the same meaning.

C4: [DISCOVER [NP V-*ing*]]

In the third sentence of the disputed paragraph, the following phrase occurs:

> I **discovered them recording** me doing sexual intercourse

The second part of this sentence will be analysed below; however, here the focus is on the relationship between the verb *discovered* and the complement clause which serves as its object.

A complement clause is a type of dependent clause (so-called because they cannot stand alone) which completes the meaning of a verb, adjective, or noun. There are four types of complement clauses in English: *that*-clauses, *ing*-clauses, *wh*-clauses, and *to*-infinitives (Biber et al. 1999, p.658). In the retraction statement, the past tense form of the verb DISCOVER (*discovered*) is used with a complement clause in the form of an *ing*-clause, or a V-***ing*** **complement clause** (*recording*).

There are two ways in which V-*ing* complement clauses can be used in English (Biber et al. 1999, p.740). Firstly, the verb can be followed directly by the V-*ing* complement clause (as in "She remembered **stealing the clock**"). Alternatively, the verb can be followed by a noun phrase [NP] and then the V-*ing* complement clause (as in "she remembered him **stealing the clock**"). It is this second type of construction that is seen in the retraction statement. Specifically, the past tense form of the verb *discover* (*discovered*) is followed by a noun phrase constituted by just a pronoun (*them*) and then a V-*ing* complement clause (introduced by the verb *recording*).

However, the verb DISCOVER is attested by Biber et al. (1999, p.663) only with *that*-clauses (e.g. I **discovered that** they were recording me). In contrast, the type of structure seen in the retraction statement, in which DISCOVER is followed by a V-*ing* complement clause, is not attested. Importantly, the voice of a verb (that is, whether it is active, as in the retraction statement, or if it is passive) affects the type of complementation that it can take and therefore the discussion in this section only covers the active voice of DISCOVER.

To determine if the [DISCOVER [NP V-*ing*]] construction used in the statement is as unusual in English as the literature suggests, a corpus search was also conducted. Specifically, the frequency of the alternative [DISCOVER [*that*-clause]] was compared to the frequency of [DISCOVER [NP V-*ing*]]. As none of the corpora used in this investigation are searchable by syntactic restrictions, we only searched for instances of pronominal noun phrases, or noun phrases that are constituted by just a single pronoun.

Corpus analysis (Table 4) shows that DISCOVER is much more likely to appear in English with a *that*-clause than the [NP V-*ing*] clause attested in the retraction statement. Moreover, the search does not include cases of complementation with the omission of the

complementiser *that* (e.g. "I discovered they were recording me") meaning that the reported figures for the [DISCOVER [*that*-clause]] construction here are conservative estimates.

Variant	JSI	Ten15	BNC	SBNC14	BLaRC	GloWbE	NOW	CORE	COCA
[DISCOVER [NP V-*ing*]]	0.6% (3,573)	0.5% (1,237)	0.5% (11)	0% (0)	0% (0)	0.5% (166)	16.5% (535)	0.4% (5)	0.6% (68)
[DISCOVER [*that*-clause]]	99.4% (595,469)	99.5% (239,066)	99.5% (2,005)	100% (37)	100% (98)	99.5% (36,173)	83.5% (2,702)	99.6% (1,134)	99.4% (11,194)
Queries	[lemma="discover"][tag="PP.?"][word=".*ing"&tag="V.*"] [lemma="discover"][word="that"]								

Table 4 - Relative frequency and frequency (in brackets) of the two variants [DISCOVER [NP V-*ing*]] and [DISCOVER [*that*-clause]] in the corpora chosen for analysis. The queries used are reported at the bottom in CQL.

Although [DISCOVER [NP V-*ing*]] is considerably less frequent than the *that*-clause alternative, it is still attested in most of the corpora. However, a closer exploration of the results in corpora which are searchable by register (BNC, CORE, and COCA), revealed that when [DISCOVER [NP V-*ing*]] was attested, it was typically in very specific contexts and it was almost exclusively found in fictional texts.

Egan (2008, p.154) reports that the verb DISCOVER with a V-*ing* complement clause denotes "the perception by the matrix verb subject of something which has been hidden from him or her […] usually an ongoing process". This is confirmed by the manual analysis of the recovered instances of this construction from the corpora, which indicate that the V-*ing* complementation is used when (1) the event described by the complement clause is suggested to be happening contemporaneously to the event of the main clause, and (2) that at the time of the discovery, the discoverer was not aware of the event in the complement clause. In other words, the use of DISCOVER with a V-*ing* clause is reserved for those cases in which both the 'discovering' and the action discovered are happening at the same time. For example:

a) Aunt Ilsa was in the library; she had a heavy cold at the time and I am tempted to say we **discovered her poring** over a map, but the inelegant truth is that she was searching the shelves for a misplaced book when we entered. (BNC)
b) She remembered wandering around a boyfriend's house talking to herself, and then **discovering him sitting** in the lounge. (BNC)
c) I noticed the motion of her tail and turned to **discover her stalking** me, only two feet away. (COCA)
d) he was frightened to **discover himself grasping** his wrist and considering how simple it would be to open a vein (COCA)

However, the example in the disputed paragraph does not conform to this pattern. The paragraph unquestionably states that the writer did not know at the time of the events described in the complement clause that she was being recorded:

> The reason I made the statement with the fake report is because I did not know they were recording & humiliating me that night I discovered them recording me doing sexual intercourse

Therefore, the time of the clause headed by DISCOVER ("I discovered them recording me doing sexual intercourse") can only refer to a non-contemporaneous event. Both the literature on the topic and our corpus analysis strongly confirm that native speakers of British or American English exclusively use a *that* complement clause when describing non-contemporaneous events.

It is very important to stress at this point that native speakers of a language do not *learn* these rules explicitly and, indeed, if they are asked what rules they follow to choose complement clauses for a verb they are not able to spell them out. Instead, native speakers *acquire* these rules unconsciously by hearing and using language in real contexts. The use of this construction is thus not consistent with a person who is a native speaker of British English and it is more consistent with a person who speaks English as a second language, who is instead more likely to have acquired or even learned very specific grammatical rules of this kind.

C5: [DO [*sexual intercourse*]]

In the third sentence of the disputed paragraph, a construction occurs in which the present participle form of the verb DO (*doing*) takes the phrase *sexual intercourse* as its object:

> I discovered them recording me **doing sexual intercourse**

Here the verb DO is used to denote the speaker's participation in a past action (sexual intercourse). However, an alternative way of articulating this would be to use the present participle form of the verb HAVE (*having*):

I discovered them recording me **having sexual intercourse**

Therefore, a corpus search was performed to determine the relative frequencies with which DO and HAVE collocate with the phrase *sexual intercourse*. In the original results, there were some instances where DO functioned as an auxiliary verb (e.g. *In which state **does sexual intercourse** last the longest?* JSI) and thus represented a different construction to the one in the retraction statement. However, unlike with C.1, where it was not possible to filter out auxiliary instances because of the number of results returned, here, as a smaller number of results were returned by the corpus searches, it was possible to manually exclude auxiliary constructions. Thus, only relevant instances of the construction [DO [*sexual intercourse*]] are reported in Table 5 below:

Variant	JSI	Ten15	BNC	SBNC14	BLaRC	GloWbE	NOW	CORE	COCA
[DO [sexual intercourse]]	0.07% (9)	0.1% (3)	0% (0)	0% (0)	0% (0)	0.2% (2)	0.02% (1)	0% (0)	0% (0)
[HAVE [sexual intercourse]]	99.9% (13,460)	99.9% (2,860)	100% (108)	0% (0)	100% (51)	99.8% (1,240)	99.9% (3,633)	100% (50)	100% (241)
Queries	[lemma="do"][word="sexual"][word="intercourse"] [lemma="have"][word="sexual"][word="intercourse"]								

Table 5 - Relative frequency and frequency (in brackets) of the two variants [DO [*sexual intercourse*]] and [HAVE [*sexual intercourse*]] in the corpora chosen for analysis. The queries used is reported at the bottom in CQL.

The numbers in Table 5 indicate the verb DO is extremely unlikely to take the phrase *sexual intercourse* as its object in English. In none of the nine corpora explored, does the [DO [*sexual intercourse*]] construction account for any more than 0.2% of the results returned. Moreover, in the three corpora specifically centred on British English (BNC, SBNC14, BLARC), the variety of English spoken by the defendant, the [DO [*sexual intercourse*]] construction does not appear at all.

As corpora are composed of naturally occurring texts, there is often informative metadata available to help understand the context in which the language in question was produced. In this instance, exploring the metadata for the [DO [*sexual intercourse*]] results revealed that many of the few instances of this construction that were found were very likely authored by non-native speakers of British English. For instance, two of the nine results from the JSI corpus originated from *indiatimes.com* and another of the nine was from *thebalitimes.com.* Meanwhile, one of the two results from the GloWbe corpus originated from Pakistan.

Overall, then, the construction used in the retraction statement, in which DO takes the phrase *sexual intercourse* as its object, is extremely uncommon, and it is highly unlikely that a speaker of British English would author this construction. Instead, there is evidence that this construction is associated with authors who do not speak British English as their first language.

4. Summary

To summarise, our initial manual inspection of the disputed paragraph of the retraction statement revealed five constructions that exhibited variation and which we thus believed would help build a profile of the author. Drawing on corpus linguistics methodologies and previous linguistic research, we reached the following conclusions:

- the use of the construction [DO [REPORT]] instead of [MAKE [REPORT]] is marked and atypical in British English and other varieties of English and its presence is more consistent with a speaker of English as a second language;
- the use of the construction [BE *not the truth*] is far less likely to be used by a native speaker of British English and other varieties of English than its adjectival variant [BE *not true*] and therefore its presence in the disputed text is more likely to be explained by its author being a speaker of English as a second language;
- the noun [APARTMENT] is marginally less likely to be used by someone with the linguistic background of the writer and its presence in the disputed text is therefore marginally more likely to be explained by a speaker of a non-British variety of English or English as a second language being the author;
- the use of the construction [DISCOVER [NP V-*ing*]] in the context of the disputed paragraph is not consistent with a person who is a native speaker of British English and more consistent with a person who speaks English as a second language;
- the use of the construction [DO [*sexual intercourse*]] is not consistent with a person who is a native speaker of British English and it is instead more consistent with a person who speaks English as a second language.

Given these results, it was our opinion that the linguistic evidence strongly supports the defence hypothesis, H1: *the disputed paragraph is unlikely to have been composed by*

someone who matches the linguistic profile of the defendant and is more likely to have been composed by a person who spoke English as a second language.

The results discussed here were submitted in a report to the defendant's legal counsel on October 2019, and the second author of this chapter subsequently provided expert testimony at her trial in Cyprus the following month. Despite this forensic linguistic evidence, the Court ruled that the statement was admissible as a handwritten statement authored by the defendant without coercion. Subsequently, the defendant was then found guilty of public mischief in December 2019. At her sentence hearing in January 2020, she was given a suspended sentence of four months in prison and was ordered to pay £125 in legal fees. She was allowed to return home after the hearing, having by that point spent one month in a Cyprus prison and another six legally bound to remain in Cyprus while the trial unfolded. Her legal team have filed an appeal to the Supreme Court of Cyprus on the grounds that the defendant did not receive a fair trial and the trial breached both local and international laws (Justice Abroad 2019).

5. Conclusion

This chapter described the forensic linguistic evidence presented in court for this case of an alleged false allegation of rape. The chapter also demonstrates the way in which a forensic linguist approaches a case of this kind, both in helping the practitioner to frame the forensic question and in showing how a typical analysis of language for a forensic problem is carried out.

Although the results described in this chapter are exactly as presented in the report and in the testimony, the time passed between the analysis and the writing of this chapter and the space given by this publication format meant that these results could be written up more clearly and with more attention to detail for the benefit of an audience of non-linguists. The value added by this chapter, and by this volume as a whole, we believe, is therefore that the linguist can explain their mental and analytical process more clearly than what can be done in an expert witness report. By doing that, we hope that this chapter has contributed to shortening the knowledge gap between disciplines and fostering stronger collaborations.

Reference List

BBC News 2019. Ayia Napa: 'False rape claim' statement 'not proper English'. *BBC News* 16 October. Available at: https://www.bbc.co.uk/news/uk-50072902 [Accessed: 19 October 2020].

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

Biber, D. et al. 1999. *The Longman Grammar of Spoken and Written English*. Harlow: Longman.

Biber, D. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 8(1), pp. 9–37. doi: 10.1515/cllt-2012-0002.

Chambers, J.K. and Schilling-Estes, N. 2013. *The handbook of language variation and change*. Second edition. Hoboken, New Jersey: John Wiley & Sons.

Coulthard, M. 2004. Author Identification, Idiolect, and Linguistic Uniqueness. *Applied Linguistics* 25(4), pp. 431–447. doi: 10.1093/applin/25.4.431.

Coulthard, M. 2013. On the use of corpora in the analysis of forensic texts. *The international journal of speech, language and the law* 1(1), pp. 27–43. doi: 10.1558/ijsll.v1i1.27.

Coulthard, M. 2017. *An introduction to forensic linguistics*. Second edition. Abingdon, Oxon ; Routledge.

Egan, T. 2008. *Non-finite complementation: A usage-based study of infinitive and -ing clauses in English*. Amsterdam: Rodopi.

Futagi, Y. et al. 2008. A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning* 21(4), pp. 353–367. doi: 10.1080/09588220802343561.

George, A. 1990. Whose Language is it Anyway? Some Notes on Idiolects. *The Philosophical Quarterly (1950-)* 40(160), pp. 275–298. doi: 10.2307/2219723.

Granger, S. and Bestgen, Y. 2014. The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching* 52(3), pp. 229–252. doi: 10.1515/iral-2014-0011.

Grant, T. 2012. Txt 4N6: Method, Consistency, and Distinctiveness in the Analysis of SMS Text Messages Symposium. *Journal of Law and Policy* 21(2), pp. 467–494.

Grieve, J. et al. 2019. Attributing the Bixby Letter using n-gram tracing. *Digital Scholarship in the Humanities* 34(3), pp. 493–512. doi: 10.1093/llc/fqy042.

Juola, P. 2015. The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions. *Digital Scholarship in the Humanities* 30(suppl_1), pp. i100–i113. doi: 10.1093/llc/fqv040.

Justice Abroad 2019. Defence to Appeal after Teenager Found Guilty of Public Mischief After Reporting Group Rape in Cyprus. Available at: https://www.justiceabroad.co.uk/news/defence-to-appeal-after-teenager-found-guilty-of-public-mischief-after-reporting-group-rape-in-cyprus [Accessed: 19 October 2020].

Kniffka, H. 2007. Orthographic Data in Forensic Linguistic Authorship Analysis. In: Kniffka, H. ed. *Working in Language and Law: A German Perspective*. London: Palgrave Macmillan UK, pp. 192–234. Available at: https://doi.org/10.1057/9780230590045_9 [Accessed: 19 October 2020].

Koppel, M. and Schler, J. 2004. Authorship verification as a one-class classification problem. *Proceedings of the 21th International Conference on Machine Learning*, pp. 62–67. Banff, Alberta, Canada: ACM.

Koppel, M. et al. 2012. The "fundamental problem" of authorship attribution. *English Studies* 93(3), pp. 284–291. doi: 10.1080/0013838X.2012.668794

Labov, W. 2001. *Principles of linguistic change, volume 2: Social factors*. Oxford: Blackwell Publishers.

Nesselhauf, N. 2003. The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics* 24(2), pp. 223–242. doi: 10.1093/applin/24.2.223.

Nini, A. 2018. An authorship analysis of the Jack the Ripper letters. *Digital Scholarship in the Humanities* 33(3), pp. 621–636. doi: 10.1093/llc/fqx065.

OED Online 2020. apartment, n. *OED Online* . Available at: https://www.oed.com/view/Entry/9033 [Accessed: 19 October 2020].

Shuy, R.W. 2005. *Creating language crimes: how law enforcement uses (and misuses) language*. Oxford: Oxford University Press.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stamatatos, E. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60(3), pp. 538–556. doi: 10.1002/asi.21001.

Tagliamonte, S. 2012. *Variationist sociolinguistics: change, observation, interpretation*. Place of publication not identified: Wiley Blackwell.

Turell, M.T. and Gavalda, N. 2012. Towards an Index of Idiolectal Similitude (or Distance) in Forensic Authorship Analysis Symposium. *Journal of Law and Policy* 21(2), pp. 495–514.

Wray, A. 2000. Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics* 21(4), pp. 463–489. doi: 10.1093/applin/21.4.463.

Wright, D. 2013. Stylistic variation within genre conventions in the Enron email corpus: developing a textsensitive methodology for authorship research. *International Journal of Speech Language and the Law* 20(1), pp. 45–75. doi: 10.1558/ijsll.v20i1.45.

Donlan, L. and Nini, A. (2022). Forensic authorship analysis of the Ayia Napa rape statements. In Picornell, I., Perkins, R., and Coulthard, M. (eds) *Methodologies and Challenges in Forensic Linguistics Casework*. Hoboken, NJ: Wiley-Blackwell.

Wright, D. 2017. Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International Journal of Corpus Linguistics* 22(2), pp. 212–241. doi: 10.1075/ijcl.22.2.03wri.