Wright, D. (2021) Corpus approaches to forensic linguistics. In M. Coulthard, A. May and R. Sousa-Silva (eds.), The Routledge Handbook of Forensic Linguistics (2nd edn). London: Routledge, pp. 611-627.

# 37

# Corpus approaches to forensic linguistics

## Applying corpus data and techniques in forensic contexts

*David Wright*

### *Corpora in forensic linguistics*

Corpus linguistics is 'the study of language based on examples of real life language use' (McEnery and Wilson 1996: 1), with the examples collected, stored and analysed as a *corpus* (pl. *corpora*). Corpora can run into millions or even billions of words, and therefore require the use of specialised software to quantitatively and qualitatively analyse them. Corpus linguistics is a set of methods and procedures that can be applied in the analysis of a range of texts and contexts that forensic linguists may be interested in examining.

Since the advent of modern-day corpus linguistics, many fields have benefitted from its ability to identify patterns in text, add evidence to support qualitative analyses and explore large datasets in ways not previously possible. However, uptake in forensic linguistics has been relatively slow. This is likely due to a number of factors, not least the fact that the types of data that forensic linguists work with are often not in abundance. Whether it is courtroom or police interview transcripts, or evidential texts such as text messages, emails, letters or threats, the data (at least in most parts of the world) is scarce, and many researchers spend years to source and collect precious datasets, often after developing close working relationships with organisations or individuals who have access to data that are otherwise in short supply.

Some of the earliest and most seminal work in forensic linguistics is corpus-based. In the work which coined the term 'forensic linguistics', Svartvik (1968) used a corpus approach to analyse a set of disputed witness statements in a murder case. Similarly, in his analysis of the Derek Bentley statement, a watershed case for forensic linguistics, Coulthard (1994) used specialised corpora of ordinary witness statements and police statements, along with the much larger spoken element of the COBUILD corpus, to question the authorship of Bentley's disputed statement. With a few notable exceptions, including early adopters of corpus techniques such as Kredens (2002) in authorship analysis and Cotterill (2003) and Heffer (2005) in courtroom discourse analysis, there was relatively little corpus linguistic work in forensic linguistics in the twenty years since Coulthard (1994). The second decade of the twenty-first century, however, has seen a healthy increase in the amount of corpus-based

forensic linguistics work across a range of research areas. Such work has demonstrated three affordances offered to forensic linguistics by corpus approaches: (i) they allow for new perspectives on familiar genres, (ii) they offer possible solutions to methodological challenges, and (iii) they open up brand new avenues for research.

In terms of new perspectives, corpora and corpus methods have allowed for new approaches to familiar types of discourse in forensic linguistics. For instance, corpus linguistics has expanded the horizons of work in legal language, and the collection and analysis of large datasets of legal language have provided new insights to our understanding of legal texts. An example is Goźdź-Roszkowski and Pontrandolfo's (2017) collection of research investigating phraseology, lexical bundles and formulaic sequences in monolingual, multilingual and translated legislative and judicial discourses (see also Finegan and Lee; McAuliffe, this volume). Similarly, qualitative work in courtroom discourse has been augmented by the quantitative analysis made possible by corpus techniques, as in the examination of important high-frequency individual words and their usage across single trials or a number of trials (e.g. Tkačuková 2015; Szczyrbak 2016). In addition, corpora have been built and analysed which comprise some of the more underexplored aspects of trial discourse, such as closing arguments or summing up (e.g. Johnson 2014; Felton Rosulek 2015; Matoesian and Gilbert, this volume). Corpus linguistics can also provide new solutions to existing methodological challenges in forensic linguistics. One such area is statutory interpretation, where corpus techniques provide a transparent and rigorous alternative to the traditional reliance on judge's intuition, dictionaries and etymology (e.g. Lee and Mouritsen 2017; Solan and Gales 2018; Gries, this volume). Corpus linguistics has also been harnessed in authorship analysis research as a means by which to combine elements of existing stylistic and stylometric approaches, in terms of supporting or explaining statistical results of authorship attribution with a qualitative examination of author style (e.g. Wright 2017; Nini 2018). Meanwhile, Grant (2017) demonstrates how using a corpus method can be a valuable strategy in determining the meaning of slang words. The building and analysis of certain types of corpus have made available, for the first time, new datasets and investigations into areas of forensic linguistics not previously possible. One notable example is the digitization and online publication of *The Old Bailey Proceedings (1674–1913)* which has given rise to new insights into the historical courtroom (e.g. Archer 2014; Johnson 2018). Similarly, there now exist analyses of previously unexamined genres, such as malicious and threatening communications (e.g. Gales 2015), trolling (Hardaker, this volume) and online grooming (e.g. Chiang and Grant 2017).

Drawing on three case studies, this chapter aims to support the integration of corpora and corpus techniques into forensic linguistics by discussing some of the important considerations in building and designing forensic corpus data and demonstrating the ways in which established corpus linguistic techniques can be used in the analysis of such data.

### *Three case study corpora*

The three corpora drawn on in this chapter (Table 37.1) are 'specialised' corpora (Flowerdew 2004: 21) in that they all represent specific but different text-types, they each comprise data from a particular discourse community, and they were all collected to help answer precise research questions and aims. However, some readers may not consider these corpora to be 'forensic'. They do not consist of legal texts nor are they of texts or genres at any stage of the legal process, such as police interviews or courtroom talk, and they are not illegal or illicit texts and they do not constitute evidence in criminal or civil proceedings. They are, however, bound by their shared aim of improving 'the delivery of justice through language analysis' (MacLeod and Grant 2017: 173):

Table 37.1 Details of three case study corpora used in this analysis

| | **Seduction Forum Corpus (SFC)** | **Adolescent Harassment Reports (AHR)** | **Enron Email Corpus** |
|---|---|---|---|
| **Method of collection** | Data scraped from web using Python script | Dataset collected through a web-based app accessed by participants | Data released online, collected and pre-processed |
| **No. of texts** | 25,788 posts | 61 reports | 63,369 emails |
| **No. of tokens** | 26,527,412 | 1,512 | 2,462,151 |
| **Public domain** | Yes | No | Yes |

- The 'Seduction Forum Corpus' (SFC) was collected to examine the discourses of resistance and sexual consent in a specialised online community, and to determine whether language used within this community constitutes the incitement of violent offences against women and girls.
- The 'Adolescent Harassment Reports' (AHR) corpus was gathered to provide some initial insight into the street harassment experienced by young people aged 11 to 15 in England, in terms of what happened and how they responded.
- The 'Enron Email Corpus' is used to develop new methods of authorship analysis which attempt to combine quantitative and qualitative approaches to analysing authorial style.

Despite their shared aims, the three corpora diverge from one another in various ways. The differences in the means by which these data were collected, their size and composition and the necessary ethical considerations represent important points when sourcing and building forensic corpora.

## Corpus collection

The internet provides a rich set of possible data sets for forensic linguists. The Seduction Forum Corpus (SFC) and the Enron Email Corpus both already existed online, but were sourced in different ways. The SFC is made up of an entire discussion forum from the online 'Pick-Up Artist' community, a community committed to studying and applying the art of pickup and seduction. The forum is a popular one but has been given a pseudonym in this chapter to protect the identities of the forum members (discussed in more detail below). The data were collected from the forum using GNU *wget* web-scraping scripts, and saved as .json and .txt files. Meanwhile, the Enron Email Corpus was originally released into the public domain in 2003 by a federal judge as part of a database of 1.6 million documents following a criminal investigation surrounding the bankruptcy of the company. The corpus used here draws on that collected and prepared by Carnegie Mellon University (Cohen 2009) and contains emails sent by 176 Enron employees and is stored as individual emails and author sub-corpora in .txt format. Both SFC and the Enron corpus required some pre-processing and

'cleaning' before they were suitable for their respective (forensic) purposes. With both corpora, duplicate texts, web-associated HTML tags and unwanted metadata were all removed, leaving only the text and metadata that were relevant and useful for the analysis. In contrast, the texts in the Adolescent Harassment Reports (AHR) corpus were collected directly from adolescents who, as part of a project on street harassment (Betts et al. 2019), were invited to report any experiences of harassment they had in a given six to eight-week period by using a web-based app. The respondents were asked a series of multiple-choice questions about their experiences and were given the opportunity to describe the event(s) in their own words in a free-text comment box.

**Corpus size**

The corpora also vary in size. The SFC is the largest at over 26 million tokens, followed by the Enron Email Corpus at almost 2.5 million, with the AHR corpus at only 1,500. The latter may be seen to be pushing the boundaries of what can be considered a 'corpus' as it contains fewer words than are demanded by many undergraduate essays. But given that a corpus is a 'collection of texts' stored electronically (Baker 2006: 48), the collection of 61 reports satisfies this criterion. This does, however, raise an important question to consider when building a corpus for forensic purposes – how big should a corpus be or, perhaps, how small *can* it be? There is no straightforward answer to the optimum size of a corpus; rather, its size will be determined by a combination of the research questions being asked of it and restrictions of practicality. For example, the purpose of building the SFC was to observe discourses within the Pick-Up Artist community, and therefore any corpus used needed to represent at least one part of this community. At the same time, given its online nature, there were no restrictions on how much could be collected. In contrast, the size of the Enron corpus is limited to the number of emails in the original set released by the courts, and so the corpus was collected in its entirety to obtain any and all available data for each author. The AHR corpus was collected with the aim of learning as much as possible about children's experiences of harassment, and its collection was determined by practical restrictions such as access to participants, the willingness of participants to volunteer their experiences and the time-consuming nature of collecting such data.

**Ethical considerations**

Given the nature of the field, it is likely that forensic linguists will be attracted to data that is sensitive in some way, and for which there are likely to be important ethical and privacy considerations. This is true of all three case study corpora. First, and most straightforwardly, while the names and company email addresses of employees in the Enron corpus are visible, there is no other personally identifiable information included such as addresses or social security numbers, and Enron employees were able to request the removal of any emails from the dataset before it was released by FERC. At the time of collection, the forum from which the SFC is taken does not require registration to view posts; it is entirely accessible for the public to view online in the clear web (as opposed to the 'dark' web) and is indexed by major search engines. Therefore, although the members of the forum cannot give informed consent to the use of their posts, given that the forum is in the public domain it is not likely that they expect their posts to be hidden or private. Nevertheless, in the preparation of this corpus for analysis a series of additional steps have been taken to protect the identities of the forum users. The name 'Seduction Forum Corpus' omits the actual name of the forum, posts will be presented without usernames and any googleable verbatim quotes or extracts from the forum will be avoided. Lastly, when collecting the AHR corpus, as with any fieldwork

methodology, consent was obtained from participants. Because the data was being collected from people under the age of 18, consent was also obtained from the headteacher of the schools that were part of the project, and letters were sent to the children's parents to inform them of the study and to invite them to inform the schools if they did not want their child to participate in the research.

### *Applying corpus techniques*

The core tools, techniques and principles of corpus linguistics can be applied to each of the three case study corpora. The analyses here draw on five core principles of corpus linguistics: *keywords, collocates, concordances, word clusters* and *part-of-speech tagging*. Each of these techniques is applied to one of the three corpora to demonstrate how they reveal different things about the data and can help provide answers to forensically-relevant research aims. The corpus software used for this analysis is *Wordsmith Tools version 7* (Scott 2016).

### Keywords, collocates and concordances

Pick-Up Artists and their practices have been described as 'a movement that teaches men to assault and harass women' (Ratchford 2017), and the techniques used in achieving sexual 'success' with women have been determined as moving from 'a seduction script, focused on conversation and comfort, to a more aggressive and coercive approach reflecting characteristics of a rape script' (Denes 2011: 418). The analysis of this corpus seeks to identify whether any of the discourse present in SFC constitutes the incitement of sexual violence against women.

At over 26-million words, finding an appropriate place to start with an analysis of the SFC poses a formidable task. This is especially true for a corpus and a discourse community which the researcher is not familiar or acquainted with. Doing a direct search for specific words (e.g 'rape') risks missing more important and pervasive words and themes in the data and, in this case, overlooks the potentially indirect nature of inciteful language. A *keyword analysis,* however, can be a valuable first step in probing very large corpora, providing an overview of the lexical composition of the corpus and identifying words which are suggestive of potentially meaningful patterns and that offer routes for further exploration (Archer 2009: 2). Although a straightforward frequency list will invariably be dominated by grammatical words, and even the most common lexical words may not reveal anything meaningful about the discourse under examination (Baker 2006:123), a keyword analysis identifies words 'whose frequency is unusually high in comparison with some norm' (Scott 2016). These words often reflect the main concepts, topics or themes in a text or corpus. Table 37.2 shows the top 50 keywords emerging from the SFC when compared against the 450 million-token Corpus of Contemporary American English (COCA) (Davies 2012).

A number of semantic and grammatical categories emerge from the top 50 keywords into which almost all of the words can be grouped, and which help the analyst characterise the nature and content of the corpus. First, a wide variety of pronoun forms appear as keywords, including first person (*I'm, I've, me, I'd, I'll*), second person (*you're, you'll, you've*) and third person pronouns (*it's, she's, her, he's, they're*). The proliferation of personal terms is reflective of the involved and personal narrative registers represented in online texts of these kinds, particularly given that PUA forums are regularly used by members to share their tales of sexual encounters with women and to give advice, as in (1).

Table 37.2 Top 50 ranked keywords in the Seduction Forum Corpus (using log-likelihood)

| N | key word | Freq. | % | N | key word | Freq. | % |
|---|---|---|---|---|---|---|---|
| **1** | *i'm* | 81659 | 0.31 | **26** | *they're* | 11480 | 0.04 |
| **2** | *girls* | 107658 | 0.41 | **27** | *pussy* | 11723 | 0.04 |
| **3** | *it's* | 74304 | 0.28 | **28** | *sex* | 26460 | 0.10 |
| **4** | *girl* | 96673 | 0.36 | **29** | *fucking* | 14800 | 0.06 |
| **5** | *i've* | 43897 | 0.17 | **30** | *tinder* | 9960 | 0.04 |
| **6** | *you're* | 40562 | 0.15 | **31** | *he's* | 9157 | 0.03 |
| **7** | *game* | 78865 | 0.30 | **32** | *you'll* | 7881 | 0.03 |
| **8** | *her* | 267993 | 1.01 | **33** | *you've* | 7742 | 0.03 |
| **9** | *she's* | 31512 | 0.12 | **34** | *beta* | 8946 | 0.03 |
| **10** | *guys* | 53882 | 0.20 | **35** | *ass* | 11343 | 0.04 |
| **11** | *me* | 171663 | 0.65 | **36** | *dude* | 8941 | 0.03 |
| **12** | *my* | 185770 | 0.70 | **37** | *lol* | 7071 | 0.03 |
| **13** | *that's* | 26392 | 0.10 | **38** | *fucked* | 8179 | 0.03 |
| **14** | *fuck* | 28342 | 0.11 | **39** | *dating* | 11492 | 0.04 |
| **15** | *shit* | 29159 | 0.11 | **40** | *am* | 30738 | 0.12 |
| **16** | *women* | 75120 | 0.28 | **41** | *getting* | 29871 | 0.11 |
| **17** | *bang* | 21518 | 0.08 | **42** | *ltr* | 5815 | 0.02 |
| **18** | *i'd* | 17900 | 0.07 | **43** | *dudes* | 6370 | 0.02 |
| **19** | *chick* | 18914 | 0.07 | **44** | *banged* | 7001 | 0.03 |
| **20** | *get* | 108182 | 0.41 | **45** | *gonna* | 5726 | 0.02 |
| **21** | *guy* | 39365 | 0.15 | **46** | *what's* | 5780 | 0.02 |
| **22** | *chicks* | 16740 | 0.06 | **47** | *banging* | 6948 | 0.03 |
| **23** | *i'll* | 14742 | 0.06 | **48** | *pretty* | 20979 | 0.08 |
| **24** | *there's* | 12746 | 0.05 | **49** | *alpha* | 7575 | 0.03 |
| **25** | *date* | 24778 | 0.09 | **50** | *text* | 14369 | 0.05 |

(1) My first bit of advice is that *you* don't show your feelings for this girl. If *you* really want to do this river walk then fine, but *you* better get the kiss. *I* get the romantic vibe *you're* trying to create, but get *her* back to your place afterwards.

Second, besides the third person pronouns *she* and *her*, there are other keywords which refer to females (g*irl(s), women, chick(s)*) and other men (*dude(s), guy(s), alpha, beta*), which reflect the purpose of the SFC. Finally, a semantic category of keywords which emerges contains those words related to dating and sex, the main topic under discussion in the forum, and these reflect a preoccupation with sexual conquest (*game, fuck(ing), fucked, pussy, sex, ass, bang(ing/ed), date, dating, tinder, ltr, text*). Some of these terms are community-specific, such as the acronym *ltr* which stands for *long term relationship.*

Keywords are generally content words (Baker 2006: 127), and so it is notable that *her*, a function word, is the eighth keyword in SFC. Bearing in mind the aim of this analysis is to examine potentially inciteful discourses, the markedly high frequency of *her* warrants closer attention. A *collocation analysis* identifies which words commonly co-occur with a word under examination, and analysing these *collocates* can provide an insight into the contexts in which it is used and the discourses of which it is part. *Her* serves grammatically as both the object in a clause (e.g. *I saw <u>her</u>*) and as a determiner within a noun phrase (e.g. *I saw <u>her face</u>*). A frequency analysis of the collocates immediately before and after *her* reveals which verbs most frequently affect *her*, and which nouns are most commonly premodified by *her* (Table 37.3).

Table 37.3 Most frequent L1 VERB and R1 NOUN collocates of *her* in SFC

| L1 Collocates | | | R1 Collocates | | |
|---|---|---|---|---|---|
| **N** | **Word** | **Freq.** | **N** | **Word** | **Freq.** |
| **1** | *tell* | 5231 | **1** | *number* | 3583 |
| **2** | *get* | 4787 | **2** | *friends* | 3147 |
| **3** | *told* | 4607 | **3** | *friend* | 2363 |
| **4** | *let* | 2944 | **4** | *face* | 1934 |
| **5** | *give* | 2741 | **5** | *place* | 1796 |
| **6** | *make* | 2625 | **6** | *ass* | 1624 |
| **7** | *see* | 2564 | **7** | *phone* | 1553 |
| **8** | *fuck* | 2499 | **8** | *body* | 1288 |
| **9** | *take* | 2272 | **9** | *pussy* | 1272 |
| **10** | *ask* | 2229 | **10** | *mind* | 1202 |
| **11** | *bang* | 1684 | **11** | *head* | 1139 |
| **12** | *asked* | 1678 | **12** | *life* | 1093 |
| **13** | *call* | 1575 | **13** | *hand* | 1072 |
| **14** | *met* | 1564 | **14** | *boyfriend* | 1006 |
| **15** | *have* | 1452 | **15** | *eyes* | 937 |
| **16** | *text* | 1358 | **16** | *hair* | 906 |
| **17** | *like* | 1257 | **17** | *mouth* | 865 |
| **18** | *keep* | 1129 | **18** | *name* | 818 |
| **19** | *took* | 1124 | **19** | *parents* | 700 |
| **20** | *want* | 1054 | **20** | *family* | 677 |

The lexical verbs of which *her* is most commonly the object in SFC can be organised into material processes (*get, let, give, make, fuck, take, bang, met, keep, took*), verbal processes (*tell, told, ask(ed), call, text*) and mental processes (*see, like, want*). Some of these verbs are unsurprising given the nature of the forum, as they relate to communicating, meeting or sexual involvement with women and girls (e.g. *call, text, met, like, fuck, bang*). However, investigating some others more closely reveals some more malicious discourse. The most common material process *get*, for example, shows how the *get + her* collocation gives rise to some particularly predatory discourses:

1 If I **get her** alone with me, it's in the bag.
2 Always in busy places, the gym, **get her** drunk, get physical, **get her** home.
3 Even if she tells me she just wants to be friends, I try to **get her** drunk and isolated.
4 I was thinking go to the hotel, **get her** drunk and try and **get her** to an empty room.

These *concordance* lines show *get + her* in context and represent common patterns in the use of this collocation in SFC. Namely, there is evidence here of forum members advising each other to make women vulnerable before making sexual advances towards them by getting them *alone, drunk* or both. *Get her drunk* and *get her alone* occur 69 and 42 times respectively in SFC, suggesting that isolating and lowering the inhibitions of their female 'targets' is considered a helpful step in their pursuit of sexual gratification.

Similarly, *make* as the sixth most frequent L1 verb collocate is used within some patterns of unambiguous physical and sexual manipulation in which men advise each other to force women to behave in certain ways for their own sexual benefit:

1 I'm sure I can **make her** do whatever I want with time.
2 **Make her** do the sluttiest, most degrading shit you can imagine when you see her.
3 **Make her** do anal or get the fuck out.
4 I usually **make her** suck my dick while she's doing all that.
5 I get hard and **make her** suck my dick. Then I bang her as the television plays.
6 Then you can **make her** your sex slave, for you and only you.
7 But in the bed room you can **make her** your whore.

At the very least, this sample of concordance lines provides an insight into the male-dominated, patriarchal and misogynistic ideologies that underpin much of the discourse in the SFC. Forum members are encouraged to act as though women are their property, their 'sex slave' and 'whore'. Related to this are the examples from the corpus in which forum members appear to be encouraging other members to *make* women and girls perform certain sexual acts. Such acts are either packaged in vagueness, such as 'make her do whatever I want' or can be explicit and specific such as 'make her do anal' and 'make her suck my dick'. The specific choice of the pattern *make her* is notable insofar as it denotes force and exclusive male agency in causing something to happen, which in turn raises questions over consent in such circumstances.

Shifting attention to the R1 collocates of *her*, in which *her* is a determiner, the most dominant category of noun in SFC is words referring to various parts of the body: *face, ass, body, pussy, hand, eyes, hair* and *mouth*. While attention may be more immediately drawn to those sexualised body parts (i.e. *ass, body, pussy*), those which are less overtly related to sex are in fact used as part of highly sexualised discourses:

1 Do me a favour and spit in **her face** if you ever get a chance.
2 Then put your dick in **her face** and tell her what you want her to do.
3 I shoved **her face** in my crotch and she gave me a blow job.
4 Make out and then push **her head** toward your cock.
5 Just shove your dick in **her mouth** and tell her whatever you think she wants to hear.
6 How about grabbing her by her hair and shoving your cock in **her mouth**?
7 I choked her and pulled **her hair** hard enough so that it really hurt.
8 Don't forget to pull **her hair** and choke her.

*Face, head, mouth* and *hair* appear a combined total of 4,844 times as immediate R1 collocates of *her*. These words alone are relatively innocuous, but as even the very small sample of concordance lines here show, they are used within wider discursive patterns in SFC which show aggressive and abusive behaviours towards women and girls being recommended and encouraged. This includes the suggestion of forcing women to engage in sexual acts.

One final *her* body-part collocate worthy of note here is *hand*, which appears with a frequency of 1,072. A concordance analysis reveals a pattern of behaviour reported by forum

members in which they force a woman's hand to touch their penis, or wherein they advise such a course of action in the pursuit of sexual intercourse:

1 You have to take **her hand** and put it on your cock.
2 I take **her hand** and put it on my dick. I basically push it as far as she will go.
3 just before she resists just take **her hand** and put it on your dick.
4 Ten mins later take **her hand** and put it on your dick.

Forcing someone to engage in sexual activity in this way, without consent, amounts to sexual assault and therefore the encouragement to do just that amounts to incitement to commit a sexual offence.

The forensically-motivated aim of this analysis was to identify whether there are discourses present in SFC which represent incitement of violence against women and girls. Although it has only been possible here to scratch the surface of a 26-million word corpus, the analysis has demonstrated the ways in which keywords, collocations and concordances are useful tools for the analyst in probing and navigating a large and unfamiliar dataset. Importantly, *her* and subsequent collocates are unremarkable words at first glance, but a closer inspection of how these words and collocates are used within this community has revealed patterns of predatory and potentially abusive discourses emerging from the forum, glorifying sexual aggressiveness, and in which violent sexual offences against women and girls seem to be reported, encouraged and incited.

## Part-of-speech tagging

Often the types of corpora that forensic linguists are able to collect and analyse are much smaller than the 26-million words of the SFC. That is the case with the Adolescent Harassment Reports (AHR) corpus, which contains 61 reports of harassment experienced by young people aged between 11 and 15, and totals only 1,512 words. Even with such a small amount of data, there are techniques from corpus linguistics that can expedite useful analytical procedures that would be time-consuming to perform manually. Part-of-speech (POS) tagging is one such technique, in which automated software is used to 'tag' each word in a corpus for its part-of-speech (or 'word class'). This subsequently renders the corpus searchable by grammatical patterns as well as lexical ones, and because POS tags appear at much higher frequencies than individual lexical items, this can be particularly useful for smaller datasets with fewer overall tokens.

The AHR corpus was collected with the aim of understanding the types of street harassment young people experience and how they respond to it. Central to this analysis, therefore, are the verbs which are prevalent in the reports submitted. The AHR corpus was tagged using the free-to-use browser-based CLAWS tagger (Garside and Smith 1997). Despite having an accuracy rate of 96–97% (Garside and Smith 1997: 119), the tags assigned to each word in the AHR corpus were manually checked for accuracy, given that some of the reports included non-standard spellings. An example of a POS-tagged report is shown in (2) (PNP stands for personal pronoun, CJC for conjunction, AV0 for adverb and PRP for preposition; all the V codings are for verb forms).

(2) *They*_PNP *stopped*_VVD *and*_CJC *beeped*_VVD *and*_CJC *looked*_VVD *then*_AV0 *waved*_VVN *at*_PRP *me*_PNP

In the 60-tag set used in CLAWS, there are 25 different verb tags, all of which begin with

‘V’. Therefore, a wildcard query for ‘V*’ was run in *Wordsmith Tools* to identify all of the verbs present in the reports. While the verbs in the reports which have subjects other than the young people reflect the types of harassment they were victim to, the verbs for which they are the subjects indicate what the victim themselves did before, during and after the harassment incident. The corpus query found 137 different verbs in the corpus, totaling 341 tokens. These verbs were then divided into those for which the young people themselves *were* the subject and those for which they *were not*.

The verbs for which the young people themselves were not the subject were manually categorised according to the different types of harassment to which they related (Table 37.4). Note that the original spellings of words has been preserved in the table and in the reports presented here. The harassment type most commonly reported in the free-text comments are incidents involving people in cars beeping at the children, slowing down or stopping (Example 3). These are closely followed by incidents with some verbal interaction between the harasser and the child (4). This often involves the child being called names and men initiating a dialogue with young girls, complimenting them and inviting them into their cars. Next are types of harassment that do not involve any verbal interaction, in which children are watched, stared at, and pictured/videoed by someone on a mobile phone (5), or smiled, waved or pointed at (6). In a small number of cases, children are followed, chased and even cornered by their harassers (7), while acts of physical aggression are relatively rarely reported by children (8).

Table 37.4 Types of harassment expressed by verbs in the AHR corpus

| **Harassment type** | **Freq.** | **Verbs** |
|---|---|---|
| Involving vehicle | 29 | *beeped, stopped, slowed, turned (around), drove, speeding, horned, honked, driving, cycled, curbed.* |
| Verbal interaction | 25 | *said, called, shouted, laughing, asked, whispered, told, shouting, say, lafed, convincing, calling.* |
| Being watched | 15 | *looked, stared, watching, took, videoing, papped, staring.* |
| Non-verbal interaction | 12 | *smiled, waved, stuck, shrugged, showed, pulling (faces), pointing, pointed, whistled.* |
| Being followed | 9 | *followed, follow, following, cornering, coming, chased.* |
| Physicality | 5 | *grabbed, yanked, threw, hit.* |

(3) we crossed the road and the guy **beeped** his horn and stuck his middle finger up.

(4) I was walking to my friends house and I got horned at and I looked and two guys (men) **asked me** 'how old are you babe? It made me feel completely disgusted.

(5) I was with my sister and a friend, we were sitting at the back of the bus and **was stared at** a while after this fat dark skin male took out his camera and **started videoing us** we covered our faces and moved the [sic] reported his actions.

(6) A taxi driver was driving past then he slowed down and **did a creepy smile** at my face.

(7) I was in a shop in [the town] and this man started staring at me and my frinds [sic] and when we were about to leave he started to get his things and he started to **follow us** we ran away and we lost him.

(8) I was followed and once **touched** on my boobs and so people was laughing at me, I felt scared about it. And I was scared that they would take me.

Although there is a growing understanding of the types of stranger or street harassment that adult women are most often victims of (e.g. Kearl 2010), much less is known about the harassment experiences of adolescents. The results from this analysis therefore, albeit of a relatively small sample of reports, provide some important initial insights into the most common patterns of street harassment that young people are victim of. Indeed, it is perhaps notable that such clear patterns of harassment emerge *even though* the sample of reports is so small.

At the same time, although there has been a good deal of research attention on the interactional and behavioural techniques and strategies adult women use to cope with or respond to stranger harassment (e.g. Fairchild and Rudman 2008), there is less clarity on what resources young people rely on when faced with street harassment. An analysis of the verbs present in the AHR corpus for which the young people themselves are the subject provides some answers to this (Table 37.5).

Table 37.5 Verbs attributed to themselves in young people's reports of harassment in AHR corpus

| **Young person's actions** | **Freq.** | **Verbs** |
|---|---|---|
| Before event | 28 | *walking, going, sitting, waiting, riding, left, leave, crossed, talking.* |
| Response to event | 26 | *ran, walked, moved, went, run, move, lost, cycling, covered, said, shouted, call, spoke, reported, rang.* |

The young people reported themselves as subjects of verbs both before and after the harassment event. Those verbs reported *before* young people are harassed provide us with the immediate setting of the incident, and they almost all see the young person (and often their friends) *walking* or *going* to and from school, the shops, or each other's houses, as in (9).

(9) I was **walking** to school and out of my bag and some man grabbed it

By contrast, there is more variation in verbs used in the reports *after* the harassment and these reflect the young people's responses to what they had experienced. Most often, they took passive or evasive action, including running, moving or walking away (Example 10), which is in line with existing research on women's coping strategies for stranger harassment (Saunders et al., 2017). However, in addition to these passive strategies, children also report taking more active measures, including confronting their harassers (11), telling their parents (12) and, in a very small number of cases, informing the police (13). Previous research has found that these active strategies are less frequently employed by adults in response to harassment for fear of escalating violence or being disbelieved.

(10) I was in a shop in [the town] and this man started staring a me and my frinds [sic] and

when we were about to leave he started to get his things and he started to follow us we **ran away** and we lost him.

(11) ther was car on rode and he wasent moveing so I went behind it to corss and it nearly ran me over so I cross and **sed** what the hell man and he stered and lafed and drove away.

(12) When me and my best mate was walking some man in a car went by us, turned around and came back to follow us in his car. He said 'How old are you' my mate said 'why' and he said your beautiful so then I **rang my mum** and we tried running away.

(13) I was with my sister and a friend, we were sitting at the back of the bus and was stared a while after this fat dark skin male took out his camera and started videoing us we covered our faces and moved then **reported** his actions.

These findings provide new and valuable information on the types of street harassment young people experience in England and how they tend to linguistically and behaviourally react and respond. From a forensic linguistic point of view, this not only addresses an offence of (linguistic) harassment that has traditionally been overlooked by the field, but in practical terms it can inform education, policy and legislation regarding how best to protect vulnerable victims from harassment. This was a small sample, but the use of POS tagging revealed patterns in the discourse that were integral to providing initial answers to the research questions posed, that is the types of harassment experienced and the responses to it. POS tagging provides forensic linguists with an additional 'layer' of corpus-derived results that go beyond straightforward lexical patterns and can be applied to a dataset of any size.

**Collocation profiles and word clusters**

Quantitative approaches to authorship analysis which rely on feature sets at the lexical level distinguish authors and texts from one another by comparing the relative frequencies with which they use particular words or word combinations. In turn, the attribution of questioned texts to a given author is made on the basis of how closely the relative frequency of these features in the questioned texts matches that of a given known or training sample. Such an approach relies on a very narrow view of linguistic variation; the frequency with which an author uses a particular word (or set of words) does not account for differences in how they *use* those words or word combinations, which may be very distinctive. A corpus approach can bring to light stylistic differences between authors that may be overlooked by purely quantitative methods.

Content words have typically been avoided as style markers in quantitative, stylometric authorship analyses, as they are generally considered to be indicators of topic rather than authorial style. Despite the promising results of studies which have drawn on content words *as well as* function words (e.g. Jockers and Witten 2010), function words have largely been relied on for lexical-level authorship analysis. However, in a case of disputed email authorship within a company, Coulthard (2013) reports that distinctive use of content words, and in particular register-related content words, were useful in determining authorship. Drawing on the Enron Email Corpus, the analysis which follows shows that by using a corpus linguistic approach, content words can be central to the investigation of idiolect. The specific word which is the focus of this analysis is *deal*.

The word *deal* is embedded in the community register of Enron, as it relates to the core business of Enron as an energy trading company. *Deal* in this context refers to the verb sense:

'to carry on commercial transactions; to do business, trade, traffic', and the noun sense: 'an act of dealing or buying and selling; a business transaction' (*Oxford English Dictionary*). The importance of this word in Enron is attested by its frequency. *Deal* occurs 4,134 times, accounting for 0.17% of the total 2,462,151 tokens and is used by 125 of the 176 authors. Importantly for the purposes of this analysis, the proportion of 0.17% represents our 'base rate knowledge' (Turell and Gavaldà 2013) of how frequently *deal* is used in the Enron population under examination. For context, *deal* occurs 87,551 times in COCA, accounting for only 0.02% of the 464,020,256 tokens.

Authors in the Enron corpus can be distinguished from one another on the basis of how frequently they use *deal*. For instance, 42 of the 176 authors use *deal* with a higher relative frequency than that of the 0.17% base rate for the corpus. Amongst those 42 are two traders, Kate Symes and Daren Farmer. Farmer uses *deal* 286 times in an email set totalling 24,389 words (1.17%). Symes uses *deal* 682 times in a much larger total email set of 58,577 words (1.16%). Both of these relative frequencies are much higher than the base rate for the corpus, and are in fact the third and fourth highest relative frequencies in the corpus. If two authors who use *deal* fewer than ten times in much smaller email sets are excluded, Farmer and Symes become the two most prolific users of *deal*. In terms of frequency alone, there is not much to distinguish between Symes and Farmer. Any approach to analysing authorial style that relied solely on lexical frequency may not distinguish between these two authors, and therefore would not be able to correctly assign any questioned text to either. However, an analysis of how they *use* the word quickly reveals marked differences.

The 'collocational profile' (Sinclair 1996) of a word captures the words that are most frequently found in its immediate environment within a given span (in this case a ten-word span, five words to the left and five to the right). In order to observe differences in the ways in which Farmer and Symes used *deal*, their collocation profiles can be compared (Tables 37.6 and 37.7).

The words that are highlighted in bold are those which do not appear in the collocation profiles of the other author, and even the most cursory glance shows that the two authors tend to use *deal* in distinctive ways when compared with each other. There is not scope to discuss all the differences here, but a sample has been chosen for closer inspection.

One immediate and important difference between the two authors as revealed by their respective collocation profiles is that, whereas Farmer typically uses *deal* as an object, Symes' use of *deal* is normally as part of a longer noun phrase. For instance, one of Farmer's patterns which contains distinctive collocates in L1 and L2 position is his repeated use of the *I have* VERB + *deal*, which he uses 21 times in total, with *created* sitting in the verb slot most commonly (n=15), as in (14).

(14) Megan,
**I have created deal** # 1075281 to cover a sale from Cleburne to ENA. I also repathed Unify […]

The other six instances see *adjusted* (twice), *taken*, *extended*, *rolled* and *updated* filling the verb slot. As the collocation profile shows, Farmer prefers to use some of these other verbs, namely *extended* and *rolled* without the auxiliary, which are found five and six times respectively, as in (15).

(15) **I rolled deal** 150325 for the first 3 days of Jan. I expect this point to be zero for the rest of Jan.

Table 37.6 L5 to R5 collocational profile for *deal* in Farmer's data Enron Email Corpus (produced using the 'patterns' feature in *Wordsmith Tools* Concord tool)

| N | L5 | L4 | L3 | L2 | L1 | Centre | R1 | R2 | R3 | R4 | R5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | the | the | the | the | the | deal | **ticket** | to | the | the | the |
| 2 | is | to | to | **have** | this | | for | for | **cover** | for | of |
| 3 | deal | **we** | on | to | **created** | | to | is | this | this | deal |
| 4 | **we** | you | **can** | on | **new** | | with | the | is | to | in |
| 5 | to | this | of | for | to | | in | you | **should** | in | you |
| 6 | for | **should** | **have** | of | on | | has | **have** | for | is | to |
| 7 | **have** | changed | is | **pricing** | **spot** | | is | **should** | deal | be | on |
| 8 | you | **record** | deal | **change** | **rolled** | | and | this | **sitara** | **volume** | **meter** |
| 9 | **volume** | and | **price** | **corrected** | **extended** | | the | in | **zero** | **are** | **mtr** |
| 10 | **volumes** | with | and | **allocated** | **term** | | at | **under** | and | **gas** | **flow** |

Table 37.7 L5 to R5 collocational profile for *deal* in Symes' data Enron Email Corpus

| N | L5 | L4 | L3 | L2 | L1 | Centre | R1 | R2 | R3 | R4 | R5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | to | the | to | the | the | deal | is | **been** | changed | to | the |
| 2 | the | **kate** | the | on | this | | has | in | the | the | to |
| 3 | **thanks** | is | **just** | changed | in | | **numbers** | and | to | and | **it** |
| 4 | this | to | is | to | **peak** | | **entry** | is | in | **it** | is |
| 5 | **kate** | and | and | of | of | | and | to | **it** | is | and |
| 6 | and | this | **kate** | in | **last** | | **was** | be | on | **kate** | with |
| 7 | in | deal | **not** | **entered** | **forward** | | **number** | the | **mw** | **thanks** | you |
| 8 | **entered** | **thanks** | **i've** | is | and | | in | **kate** | is | **as** | **kate** |
| 9 | deal | **mike** | this | **that** | **that** | | **blotter** | **that** | and | at | be |
| 10 | is | **was** | deal | and | **correct** | | to | **was** | deal | deal | for |

These patterns are not found in Symes' *deal* collocation profile and so serve to distinguish Farmer from his colleague.

Meanwhile, although Farmer uses *deal* as the head of a noun phrase, he does so less frequently than Symes. Most typically, he uses *new deal* (n=13) and *spot deal* (n=8). However, the ADJECTIVE + *deal* pattern is far more characteristic of Symes' collocational profile. As shown in Table 37.7, *peak deal* (n=12), *last deal* (n=9), *forward deal* (n=9) and *correct deal* (n=5) constitute her most frequent uses of *deal* but are not found at all in Farmer's emails (e.g. 16).

(16) This was a futures contract done by Matt Motley, and I entered the **forward deal**. I wasn't aware that futures deals don't get confirmed. I only knew that when the contract expires, I need to enter a **forward deal** in EnPower to hold the trader's position […].

Difference in the collocational profiles of Farmer and Symes can also be observed in the most frequent collocates to the right of *deal* in the authors' data. There are major differences in the R1 position, for instance, wherein the only collocate distinctive of Farmer in this position is *ticket*, which he uses 14 times, as in (17).

(17) Nicole,
The correct price is 4.50. I changed the **deal ticket**.

This use of *deal* contrasts with Symes'; whereas Famer uses *deal* as a premodifier for *ticket*, Symes' most proliferate pattern is to use *deal* to premodify *number(s)*, which she uses 57 times, but this pattern is not present at all in Farmer's data (e.g. 18).

(18) Geir Solberg in Real Time is fixing this right now - I'll let you know the new **deal number** in just a second.
Thanks,
Kate

Word strings and word *n*-grams are now commonly part of the stylometrist's toolkit when conducting statistical authorship analysis, and such chunking of text in this way will capture an author's patterns of use of particular words. However, the pre-occupation of such approaches with *frequency* can overlook author-distinctive *uses* at the stylistic level. The corpus-based concept of collocational profiles offers an alternative perspective on authors' uses of a particular lexical feature; whereas the frequency with which a person uses a word may not be evidentially significant, the ways in which they use that word may be. As the analysis here has demonstrated, the collocational profiles of *deal*, an extremely common word embedded within the register of this particular discourse community, can vary greatly between authors (in this case, two people who have the same or a very similar job). Such an approach can unlock the potential for distinguishing between authors using content words, which, notwithstanding their inclusion in word strings, have been ignored by authorship analysts.

### *Conclusion*

The techniques demonstrated here are well-established in corpus linguistics and have been applied in other fields of linguistics for decades. Depending on the nature of the data and the aims of the research, they allow the forensic linguist to explore large, unfamiliar datasets, they offer techniques for enriching smaller datasets to reveal otherwise hidden patterns, and they provide a means by which quantitative and qualitative analyses can be combined.

However, the types of corpora forensic linguists are interested in are often hard to come by, especially as the field remains pre-occupied with the analysis of 'hidden' genres related to the legal process and criminal evidence. As the field expands, so too will the range of genres and text-types that fall within the remit of forensic linguists whose aim it is to improve the delivery of justice through language analysis. Indeed, the continued expansion of the field may depend, at least in part, on the broadening of the types of data that are subjected to forensic linguistic attention. None of the three corpora used in this chapter is 'forensic' in the traditional sense; yet they all hold potential for improving the delivery of justice, whether by gaining an understanding of the influence of online communication to incite violent offences against women, by protecting young people from street harassment, or developing new methods of authorship analysis to identify or eliminate suspect authors. The broadening of the scope of forensic linguistics opens up the possibility of collecting new and diverse corpora for forensic purposes and, as the recent surge in the use of corpus methods exemplifies, adopting such methods offers exciting and previously unexplored directions for forensic linguistics.

***Acknowledgements***

I would like to thank Ikechukwu Onyenwe for the technical support provided during the collection and preparation of the Seduction Forum Corpus.

***Further reading***

Baker, P. (2006) *Using Corpora in Discourse Analysis*. London: Continuum.

Kredens, K. and M. Coulthard. (2012) 'Corpus linguistics in authorship identification', in P. Tiersma and L. Solan (eds.) *The Oxford Handbook of Language and Law*, Oxford: Oxford University Press, 504–516.

Larner, S. (2015) 'From intellectual challenges to established corpus techniques: introduction to the special issue on forensic linguistics', *Corpora*, 10(2): 131–143.

Other chapters in this volume that draw on corpus methods and approaches are: Finegan, Gries, Hardaker, and McAuliffe.

***References***

Archer, D. (2009) 'Does frequency really matter?', in D. Archer (ed.) *What's in a Word-list?: Investigating Word Frequency and Keyword Extraction*, London: Routledge, 1–16.

——(2014) 'Historical pragmatics: evidence from The Old Bailey', *Transactions of the Philological Society*, 112(2): 259–277.

Baker, P. (2006) *Using Corpora in Discourse Analysis*. London: Continuum.

Betts, L. Harding, R. Peart, S., Sjolin Knight, C., Wright, D. and Newbold, K. (2019) 'Adolescents' experiences of street harassment: creating a typology and assessing the emotional impact', *Journal of Aggression, Conflict and Peace Research,* 11(1): 38–46.

Chiang, E. and Grant, T. (2017) 'Online grooming: moves and strategies', *Language and Law/Linguagem e Direito*, 4(1): 103–141.

Cohen, W. W. (2009). Enron Email Dataset [online]. Retrieved from http://www.cs.cmu.edu/~enron/.

Cotterill, J. (2003) *Language and Power in Court: A Linguistic Analysis of the OJ Simpson Trial*, Basingstoke/New York: Palgrave Macmillan.

Davies, M. (2012) The Corpus of Contemporary American English: 450 million words,1990-present [online] https://www.english-corpora.org/coca/

Heffer, C. (2005) *The Language of Jury Trial: A Corpus-Aided Analysis of Legal–Lay Discourse*, Basingstoke/New York: Palgrave Macmillan.

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

—— (2013) 'On admissible linguistic evidence', *Journal of Law and Policy,* 21(2): 441–466.

Denes, A. (2011) 'Biology as consent: problematizing the scientific approach to seducing women's bodies', *Women's Studies International Forum,* 34(5): 411–419.

Fairchild, K. and Rudman, L. A. (2008) 'Everyday stranger harassment and women's objectification', *Social Justice Research*, 21(3): 338–357.

Felton Rosulek, L. (2015) *Dueling Discourses: The Construction of Reality in Closing Arguments*, Oxford: Oxford University Press.

Garside, R. and Smith, N. (1997) 'A hybrid grammatical tagger: CLAWS4', in R. Garside, G. Leech and T. McEnery (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, London: Routledge, 102–121.

Goźdź-Roszkowski, S. and Pontrandolfo, G. (2017) *Phraseology in Legal and Institutional Settings: A Corpus-based Interdisciplinary Perspective*, London: Routledge.
Flowerdew, L. (2004) 'The argument for using English specialized corpora to understand academic and professional settings', in U. Connor and T. A. Upton (eds.) *Discourse in the Professions: Perspectives from Corpus Linguistics,* Amsterdam/Philadelphia: John Benjamins, 11–33.
Gales, T. (2015) 'The stance of stalking: a corpus-based analysis of grammatical markers of stance in threatening communications', *Corpora,* 10(2): 171–200.
Grant, T. (2017) 'Duppying yoots in a dog eat dog world, kmt: determining the senses of slang terms for the courts', *Semiotica,* 216: 479–495.
Jockers, M. L. and Witten, D. M. (2010) 'A comparative study of machine learning methods for authorship attribution', *Literary and Linguistic Computing,* 25(2): 215–223.
Johnson, A. (2014) '"Dr Shipman told you that ..." The organising and synthesising power of quotation in judicial summing-up', *Language and Communication,* 36: 53–67.
—— (2018) '"How came you not to cry out?" Pragmatic effects of negative questioning in child rape trials in the Old Bailey Proceedings 1730–1798', in D. Kurzon and B. Kryk-Kastovsky (eds.) *Legal Pragmatics,* Amsterdam/Philadelphia: John Benjamins, 41–64.
Kearl, H. (2010) *Stop Street Harassment: Making Public Places Safe and Welcoming for Women*. Santa Barbara, C.A.: Praeger.
Kredens, K. (2002) 'Towards a corpus-based methodology of forensic authorship attribution: a comparative study of two idiolects', in B. Lewandowska-Tomaszczyk (ed.) *PALC'01: Practical Applications in Language Corpora,* Peter Lang: Frankfurt am Mein, 405–437.
Lee, T. R. and Mouritsen, S. C. (2017) 'Judging ordinary meaning', *Yale Law Journal*, 127(4): 788–879.
MacLeod, N. and Grant, T. (2017) '"go on cam but dnt be dirty': linguistic levels of identity assumption in undercover online operations against child sex abusers', *Language and Law/Linguagem e Direito,* 4(2): 157–175.
McEnery, T. and Wilson, A. (2001) *Corpus Linguistics: An Introduction,* Edinburgh: Edinburgh University Press.
Nini, A. (2018) 'An authorship analysis of the Jack the Ripper letters', *Digital Scholarship in the Humanities*, 33(3): 621–636.
Ratchford, S. (2017) 'I tried to find out if Pick Up Artists are still influential in 2017', *Vice,* 25 August 2017. [online] available from: https://www.vice.com/en_au/article/j55bxd/i-tried-to-find-out-if-pick-up-artists-are-still-influential-in-2017
Saunders, B. A., Scaturro1, C., Guarino, G., and Kelly, E. (2017) 'Contending with catcalling: the role of system-justifying beliefs and ambivalent sexism in predicting women's coping experiences with (and men's attributions for) stranger harassment', *Current Psychology,* 36(2): 324–338.
Sinclair, J. M. 1996. 'The search for units of meaning', *Textus* 9(1): 71–106.
Scott, Mike. 2016. *Wordsmith tools version 7.* Stroud: Lexical Analysis Software.
Solan, L. M. and Gales, T. (2018) 'Corpus linguistics as a tool in legal interpretation', *Brigham Young University Law Review*, 2017(6), 1311–1358.
Svartvik, J. (1968). *The Evans Statements: A case for Forensic Linguistics*, Gotëborg: University of Gothenburg Press.
Szczyrbak, M. (2016) '*Say* and stancetaking in courtroom talk: a corpus-assisted study', *Corpora*, 11(2): 143–168.
Tkačuková, T. (2015) 'A corpus-assisted study of the discourse marker well as an indicator of

judges' institutional roles in court cases with litigants in person', *Corpora,* 10(2): 145–170.
Turell, M. T. and Gavaldà, N. (2013) 'Towards an indexof idiolectal similitude (or distance) in forensic authorship analysis', *Journal of Law and Policy*, 21(2): 495–514.
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): 212–241.