Nini, A. (2019). Corpus analysis in forensic linguistics. In Carol A. Chapelle (ed), The Concise Encyclopedia of Applied Linguistics, 313-320, Hoboken: Wiley-Blackwell

Corpus Analysis in Forensic Linguistics

Andrea Nini

Abstract

This entry is an overview of the applications of corpus linguistics to forensic linguistics, in particular to the analysis of language as evidence. Three main areas are described, following the influence that corpus linguistics has had on them in recent times: the analysis of texts of disputed authorship, the provision of evidence in cases of trademark disputes, and the analysis of disputed meanings in criminal or civil cases. In all of these areas considerable advances have been made that revolve around the use of corpus data, for example, to study forensically realistic corpora for authorship analysis, or to provide naturally occurring evidence in cases of trademark disputes or determination of meaning. Using examples from real-life cases, the entry explains how corpus analysis is therefore gradually establishing itself as the norm for solving certain forensic problems and how it is becoming the main methodological approach for forensic linguistics.

Keywords

Authorship Analysis; Idiolect; Language as Evidence; Ordinary Meaning; Trademark

This entry focuses on the use of corpus linguistics methods and techniques for the analysis of forensic data. “Forensic linguistics” is a very general term that broadly refers to two areas of investigation: the analysis of the language of the law and the analysis of language evidence in criminal or civil cases. The former, which often takes advantage of corpus methods, includes areas as wide as the study of the language of the judicial process and courtroom interaction (e.g., Tkačuková, 2015), the study of written legal documents (e.g., Finegan, 2010), and the investigation of interactions in police interviews (e.g., Carter, 2011). In contrast, the other component of forensic linguistics, sometimes called “investigative forensic linguistics,” covers areas in which linguistic theory and methods, as well as corpus analysis, are applied to solve forensic problems involving language data.

This entry will be concerned with the latter definition of forensic linguistics and with the advances made in this field on the use of corpus linguistics to solve real-world forensic problems. Investigative forensic linguistics is also the area that has most benefited from corpus linguistics since its early stages (e.g., Coulthard, 1994). In recent times, more forensic applications of corpus analysis have been tested and used in real-life cases and this entry contains an overview of these applications. So far, there are three main types of problems in investigative forensic linguistics that have benefited from corpus analysis: cases of texts of disputed authorship, trademark cases, and cases of disputed meanings. The remainder of this entry will outline the main advances in the use of corpus linguistics to solve forensic problems in these main areas.

Comparative Authorship Analysis

A common forensic question involving language evidence is the task of analyzing anonymous texts to gather evidence about their authorship. This task is often called *authorship analysis*. More specifically, when there already are one or more suspects who might have written the disputed text, the forensic linguist is called upon to perform a comparative analysis of the known and disputed texts to provide evidence for questions such as whether one among a pool of candidate suspect authors is the real author of the disputed text, or whether one suspect can be included or excluded as the possible author of the disputed text.

Indeed, questions of authorship are not strictly speaking only forensic, and solutions to authorship problems, for example in literary cases, have been sought since very ancient times. More recently, with the advent of large corpora and computational methods, a new field has emerged, *stylometry*, focused on the measurement of stylistic traits in order to attribute texts. Within this heavily corpus- and computational-based framework, many advances in the field of authorship analysis have been made by computer scientists who have used machine learning and similar computational techniques (Stamatatos, 2009; Juola, 2015). However, despite these advances, these new techniques are not always applicable in forensic scenarios because forensic cases often deal with data that are not perfectly controlled for size and genre, two basic prerequisites for stylometric techniques.

Recent research in the area of forensically realistic authorship scenarios, such as Johnson and Wright (2014) and Wright (2013, 2017), has so far demonstrated that the use of certain classic corpus analysis techniques can nonetheless be helpful for comparative authorship analysis. Johnson and Wright's data set consisted of a corpus of e-mails belonging to a company that went bankrupt following a scandal, which is a realistic scenario for testing the robustness of authorship techniques for a forensic problem. All of their studies were successful in showing that the use of *word n-grams,* short sequences of words collected using a moving window of *n* words, can be personal and idiosyncratic and can be relied on for attribution of short texts.

These results are particularly important since forensic work on short texts is becoming more common, given the widespread use of instant messaging and social media such as Facebook or Twitter. Some forensic linguistic research on these types of data has also been carried out using a corpus approach. For example, Grant (2013) proposed a method to analyze short texts based on mathematics borrowed from ecology and biology. He presented the results of his analysis of text messages involved in the case of Amanda Birks, who died in a house fire which forensic evidence suggested was not accidental. Other social media data have also been investigated: for example, Sousa Silva et al. (2011) examined microblogging messages from Twitter and discovered that despite their brevity, attribution is possible, especially when paralinguistic features such as emoticons are used. However, although corpus analysis in these cases proves useful, it is important to note that reliable results are only achievable provided that the corpora used reflect a representative background population, a task that at times might be very difficult (Ishihara, 2014).

The use of corpus linguistic methods is also likely to lead to concrete answers regarding the theoretical assumption of the existence of *idiolect*, each native speaker's distinct and individual version of a language (Coulthard, Johnson, & Wright, 2017). One of the cases that provided indirect evidence of the existence of idiolect was the Unabomber case, which involved a serial bomber, Ted Kaczynski, who sent explosives to universities and airlines between 1978 and 1995 in the United States. The case was solved in part through the use of language evidence, as Kaczynski was recognized by his brother through certain phrases and expressions that appeared in a manifesto published by the bomber. This language evidence was used to gain a search warrant which then led to Kaczynski's arrest. When this evidence was later challenged, it was demonstrated that a set of 12 words or phrases (e.g., *thereabouts, at any rate, in practice*) shared by the manifesto and Kaczynski's known writings were found only in other copies of the manifesto when searched on the Web. The identifying nature of word or grammatical sequences in combination has since then been confirmed by other corpus studies, such as Turell (2010) and Wright (2013, 2017). However, caution and appeal to further research has been called for by Larner's (2014, 2016) research on the use of formulaic sequences and on the use of the World Wide Web as reference corpus, as done in the Unabomber case. Despite these notes of caution, the value of corpus analysis in matters of comparative authorship analyses is evident and the results so far are promising.

Authorship Profiling

Another type of investigative forensic case that is often encountered by forensic linguists is the profiling of the anonymous author of a forensically relevant text, such as a malicious text. This type of case is still a kind of authorship analysis but one for which either no suspects or a large number of suspects are available, and therefore it is necessary to narrow down the investigation effort. Profiling is a task that requires a deep knowledge of linguistics and there is no established standard method by which to carry it out. Grant (2008) provides examples of these kinds of problems and explains that, so far, results have been achieved by using ad hoc searches of dialect databases or reference corpora, citing the example of the phrase *bad-minded*, which was found by him in a case to point to someone with a Jamaican background. Besides work in computational authorship profiling, which again mostly involves large data sets, a systematic work for forensic purposes has been Nini (2015), who tested an array of linguistic markers using experimental methods and corpus analyses. He found that age, gender, and especially social class can be estimated for an anonymous malicious text such as a threatening letter with degrees of accuracy ranging around 70%—provided, however, that the analysis takes into account the register of the text.

A common profiling question in the analysis of a disputed text written by non-native speakers of English is the determination of their first language. Similarly to general profiling of demographics, this task is also carried out using ad hoc methodologies, with recent advances only being done in computer science using large data sets. Despite its importance, this task, often called in the field of computer science "native language identification" (NLID), has again not been properly investigated within forensic linguistics. Recently, however, Perkins and Grant (2018) analyzed a corpus of blogs written by L1 (first language) Persian speakers as a reference data set to verify if a method based on interlanguage features can help in forensically similar short disputed texts. Their promising results point to the possibility of combining statistical with linguistics insights for better practice in these types of forensic cases.

Finally, an additional area of research is the understanding of the linguistic nature of those texts commonly involved in profiling questions, such threatening or abusive letters, ransom demands, stalking texts, and so forth. The study of these kinds of registers is an important point in the research agenda of investigative forensic linguistics, and it is significantly helped by the use of corpus methods and tools. The problem of dealing with these texts, however, is their rarity and sensitivity, which makes the collection of large and balanced corpora nearly impossible. Nonetheless, advances have been made for some of these forensic texts, such as for threatening texts. Gales (2011, 2015), for example, analyzed the texts contained in the Communicated Threat Assessment Reference Corpus (CTARC) and focused particularly on stance-taking linguistic devices, revealing that these are defining features of threatening texts. This finding was then independently confirmed by Nini (2017), using another corpus of malicious texts, the Malicious Forensic Texts corpus, in particular showing that threatening texts contain a more pervasive use of modality than nonthreatening texts.

These profiling problems can thus greatly benefit from corpus linguistics for several reasons, including the possibility of exploring data sets of naturally occurring data for reference, leading to better understanding of the language of malicious texts and to general methodologies that are more case-independent. However, the challenge for this area of investigative forensic linguistics is how to collect corpora that are large and stratified enough to be reliable, a challenge that can be tackled only with more collaboration between linguists and law enforcement authorities.

Trademark Disputes

Disputes over several issues involving trademarks have been another type of forensic linguistic problem in which corpus linguistics has been of great help. Good introductions to the use of linguistics in trademark disputes are Shuy (2002) and Butters (2008, 2010), who identifies three main issues for which knowledge of linguistics is useful: the judgment of the propriety of a trademark, the likelihood of confusion of a junior trademark with a senior trademark, and whether a trademark has been *genericized*, that is, it has become part of the language (e.g., *hoover*, *ping-pong*, *escalator*).

Although often the evidence in these cases revolves around the formal analysis of the sounds or grammatical forms of the trademarks, disputes of this kind can also be solved through the use of corpus analysis. The seminal work on the use of corpus linguistics for trademark disputes is Kilgariff's (2015) overview of his consultancy work, especially in cases of genericide. He explains that the question of genericide ultimately comes down to understanding whether, in the minds of the speech community, the trademarked word is perceived as a common noun or as a proper noun, with the former meaning that the trademark has become generic. Among his examples, Kilgariff (2015) describes his involvement in a case where the question was whether *botox* is a trademark or a generic noun used to indicate cosmetic surgery to remove wrinkles. To determine if a trademark is typically used as a proper or common noun, a corpus analysis such as the one proposed by Kilgariff (2015) involves both qualitative analysis of concordance lines and counting of hits, possibly employing automatic computational methods.

Although corpus analysis cannot possibly look into people's minds, through the use of real instances of language use the linguist can present evidence that indirectly points to the underlying cognitive reality of language users, a far more objective way of presenting evidence than the use of dictionaries or intuition.

Determination of Meaning

Finally, an emerging area for investigative forensic linguistics is the use of linguistic analysis to provide evidence regarding the meaning of a linguistic unit, such as a word, phrase, or sentence. In this area, the use of corpus analysis is key because it provides a window to actual instances of usage, as opposed to either native speaker intuition or dictionary evidence.

As an example of this type of work, Solan and Gales (2016) describe a corpus linguistic analysis in the case of *Shaw vs. United States*, where a man was charged with the crime of executing a scheme *to defraud a financial institution*. Shaw's scheme did not result in any loss for the bank, since the scheme only involved the bank for the transfer of money from the victim's account to his own. The question was therefore whether Shaw defrauded the bank or defrauded the victim, and, linguistically, whether the object of *defraud* is intended to be the entity suffering the loss. Solan and Gales (2016) explored the Corpus of Historical American English (COHA) (Davies, 2008) from 1870 to 2000 and gathered a random sample of concordances including the verb *defraud* for each decade, finding that in 98% of the examples the object of *defraud* is also the entity suffering the loss. Besides the advantage of being based on real examples of language use, the other key advantage of a corpus analysis—in contrast to intuition or dictionaries—is therefore that corpus analysis lends itself very well to quantification, as in this example.

Work on using linguistics and corpus linguistics to provide evidence on meaning is not only confined to ordinary meaning of standard English but recently has also been applied to decoding slang. Grant (2017) presents a case involving conspiracy to murder of a 15-year-old pregnant girl by her ex-boyfriend and an accomplice, two emerging grime music artists. The evidence in this case consisted of a chat log in which the two men talked about the murder using slang. Grant was called to testify on the meaning of a set of incriminating sentences contained in the chat log that contributed to the charges for conspiracy to murder. For example, one of the incriminating sentences was "I'll get the fiend to duppy her den," which contains elements of nonstandard varieties of English that need to be decoded for the jury. Grant considered several methods to perform this translation task, including reference material from the Web, slang dictionaries, existing academic literature, and ethnographic methods. However, as part of his work on the case, Grant also adopted a corpus methodology and collected a small but specialized corpus of 100,000 words from an online grime music forum. This data set was key in demonstrating the meaning of the word *duppy*, which, despite appearing only four times in Grant's corpus, was unquestionably found to mean *ghost*, from Jamaican English, when used as a noun and *to kill* when used as a verb.

In a similar fashion to trademark cases, therefore, the use of corpus analysis as a way to solve semantic ambiguity issues is one of the most reliable and objective tools in the forensic linguist's toolbox.

Conclusion

This brief overview of the use of corpus analysis within forensic linguistics and the analysis of linguistic evidence suggests that corpus linguistics not only is a promising method but has the potential to become the most important way to analyze linguistic evidence in the future.

It is important to note that the areas covered in this entry are not the only ones that benefit from corpus analysis. New areas of application of corpus linguistics to the analysis of linguistic evidence arise as the needs of legal professionals or investigators become clear and as technology and society change. For example, a new emerging area of forensic linguistics is the analysis of chat transcripts in cases of child abuse, either for investigative purposes to identify the perpetrator, or to provide undercover police officers with a way to disguise their writing style when impersonating a victim (MacLeod & Grant, 2017). Although the data in these very sensitive cases are often not available in large quantities, recent corpus studies show that linguistic analysis can be very useful. For example, Chiang and Grant (2017) analyzed a corpus of seven transcripts containing interactions between online sex offenders and users of the Perverted Justice Website, who pretend to be victims for child groomers in chat rooms to help law enforcement. Through the analysis of this data set, Chiang and Grant were able to propose a model of grooming interaction consisting of 14 rhetorical moves, a significant step in the advancement of understanding this type of interaction.

Generally speaking, therefore, corpus linguistics will be decisive in establishing forensic linguistics as a fully fledged forensic science, and it is possible to predict that the use of larger corpora, combined with more sophisticated computational methods to analyze them, will become the norm in the future of forensic linguistics.

References

Butters, R. R. (2008). Trademarks and other proprietary terms. In J. Gibbons & M. T. Turell (Eds.), *Dimensions of forensic linguistics* (pp. 231–47). Amsterdam, Netherlands: John Benjamins.

Butters, R. (2010). Trademarks: Language that one owns. In M. Coulthard & A. Johnson (Eds.), *The Routledge handbook of forensic linguistics* (pp. 351–64). Abingdon, England: Routledge.

Carter, E. (2011). *Analysing police interviews: Laughter, confessions and the tape*. London, England: Continuum.

Chiang, E., & Grant, T. (2017). Online grooming: Moves and strategies. *Language and Law/Linguagem e Direito*, *4*(1), 103–41.

Coulthard, M. (1994). On the use of corpora in the analysis of forensic texts. *International Journal of Speech Language and the Law*, *1*(1), 27–43. doi: 10.1558/ijsll.v1i1.27

Coulthard, M., Johnson, A., & Wright, D. (2017). *An introduction to forensic linguistics*. London, England: Routledge.

Davies, M. (2008). *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. Retrieved March 27, 2019 from https://corpus.byu.edu/coca/

Finegan, E. (2010). Corpus linguistic approaches to “legal language”: Adverbial expression of attitude and emphasis in Supreme Court opinions. In M. Coulthard & A. Johnson (Eds.), *The Routledge handbook of forensic linguistics* (pp. 65–77). London, England: Routledge.

Gales, T. (2011). Identifying interpersonal stance in threatening discourse: An appraisal analysis. *Discourse Studies*, *13*, 27–46. doi: 10.1177/1461445610387735

Gales, T. (2015). Threatening stances: A corpus analysis of realized vs non-realized threats. *Language and Law/Linguagem e Direito*, *2*(2), 1–25.

Grant, T. (2008). Approaching questions in forensic authorship analysis. In J. Gibbons & M. T. Turell (Eds.), *Dimensions of forensic linguistics* (pp. 215–31). Amsterdam, Netherlands: John Benjamins.

Grant, T. (2013). TXT 4N6: Method, consistency, and distinctiveness in the analysis of

SMS text messages. *Journal of Law and Policy*, *21*, 467–94.

Grant, T. (2017). Duppying yoots in a dog eat dog world, kmt: Determining the senses of slang terms for the Courts. *Semiotica*, *2017*(216), 87–106. doi: 10.1515/sem-2015-0082

Ishihara, S. (2014). A likelihood ratio-based evaluation of strength of authorship attribution evidence in SMS messages using N-grams. *International Journal of Speech, Language and the Law*, *21*(1), 23–49. doi: 10.1558/ijsll.v21i1.23

Johnson, A., & Wright, D. (2014). Identifying idiolect in forensic authorship attribution: An n-gram textbite approach. *Language and Law/Linguagem e Direito*, *1*(1), 37–69.

Juola, P. (2015). The Rowling case: A proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities*, *30*. doi: 10.1093/llc/fqv040

Kilgariff, A. (2015). Corpus linguistics in trademark cases. *Dictionaries: Journal of the Dictionary Society of North America*, *36*(1), 100–14.

Larner, S. (2014). *Forensic authorship analysis and the World Wide Web*. Basingstoke, England: Palgrave Macmillan.

Larner, S. (2016). Using a core word to identify different forms of semantically related formulaic sequences and their potential as a marker of authorship. *Corpora*, *11*(3), 343–69. doi: 10.3366/cor.2016.0099

MacLeod, N., & 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), 157–75.

Nini, A. (2015). *Authorship profiling in a forensic context* (Unpublished doctoral dissertation). Aston University, England.

Nini, A. (2017). Register variation in malicious forensic texts. *International Journal of Speech, Language and the Law*, *24*(1). doi: 10.1558/ijsll.30173

Perkins, R., & Grant, T. (2018). Native language influence detection for forensic authorship analysis: Identifying L1 Persian bloggers. *International Journal of Speech, Language and the Law*, *25*(1), 1–20. doi: 10.1558/ijsll.30844

Shuy, R. W. (2002). *Linguistic battles in trademark disputes*. Basingstoke, England: Palgrave Macmillan.

Solan, L. M., & Gales, T. (2016). Finding ordinary meaning in law: The judge, the dictionary or the corpus? *International Journal of Legal Discourse*, *1*(2), 253–76. doi: 10.1515/ijld-2016-0016

Sousa Silva, R., Laboreiro, G., Sarmento, L., Grant, T., Oliveira, E., & Maia, B. (2011). ’twazn me!!! ;(’ Automatic authorship analysis of micro-blogging messages. *Lecture Notes in Computer Science*, *6716*, 161–8.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538–56. doi: 10.1002/asi.21001

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), 145–70. doi: 10.3366/cor.2015.0072

Turell, M. T. (2010). The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *International Journal of Speech Language and the Law*, *17*(2), 211–50.

Wright, D. (2013). Stylistic variation within genre conventions in the Enron email corpus: Developing a text-sensitive methodology for authorship research. *International Journal of Speech Language and the Law*, *20*(1), 45–75.

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), 212–41. doi: 10.1075/ijcl.22.2.03wri

Suggested Readings

Coulthard, M., & Johnson, A. (Eds.). (2010). *The Routledge handbook of forensic linguistics*. London, England: Routledge.

Gibbons, J., & Turell, M. T. (Eds.). (2008). *Dimensions of forensic linguistics*. Amsterdam, Netherlands: John Benjamins.

Olsson, J. (2009). *Wordcrime: Solving crime through forensic linguistics*. London, England: Continuum.

Shuy, R. W. (2008). *Fighting over words: Language and civil law cases*. Oxford, England: Oxford University Press.

Solan, L. M., & Tiersma, P. M. (2005). *Speaking of crime: The language of criminal justice*. Chicago, IL: University of Chicago Press.