Forensic Linguistics:
The potential of language for law enforcement in the digital age

Rui Sousa-Silva
University of Porto – Faculty of Arts and Humanities[1]

Abstract
The recent technological developments have granted citizens worldwide access to the Internet, including in handheld devices, and offered them new communication possibilities. Nevertheless, they have also exposed them to more cybernetic attacks, as criminals gained new opportunities for cybercriminal practice. The (perceived) increase in the number of cyberattacks faces Law Enforcement with two major challenges: firstly, the higher the volume of cyberattacks, the harder it is to dedicate the necessary resources, including human, to fight them; secondly, the range of sophisticated stealth technologies used by cybercriminals to remain anonymous online hamper the work of the forces. This paper argues that, since (cyber)criminals use language to communicate, their anonymisation can be undermined by the language that they use because language use is idiosyncratic, so each speaker makes a particular use of their language (Coulthard, 2004). This is enabled by Forensic Linguistics, which can be broadly defined as the application of linguistic analyses in legal or Law Enforcement contexts. This article presents two illustrative cases of cybercrime to show the potential of the forensic linguistic analysis. The first is the case of an anonymous set of text messages spreading defamatory contents, whose linguistic analysis enabled the sociolinguistic profiling of the author, and hence narrow down the pool of suspects. The second presents a cross border cybercriminal practice: fraudulent and deceptive messages sent to citizens for purposes of extortion. The article concludes by discussing the potential of the linguistic analyses in the fight against (cyber)-crime, and making recommendations for Law Enforcement.

Keywords: cybercrime, threatening language, darkweb, anonymity, investigative linguistics

Introduction

Over the last decades, the world has witnessed unprecedented technological developments that have, among others, granted citizens worldwide immediate access to the Internet, including in handheld devices. As a result, new communication possibilities emerged, with obvious advantages for users, who have gained immediate access to information; additionally, anyone, virtually anywhere, has been granted the power to post, share or comment on anything at any time. As they grew more acquainted with technology, users have gained a more prominent participatory role in society. Nevertheless, the new possibilities offered by technology have also exposed citizens to more cybernetic attacks, i.e. cybercrime.

1 rssilva@letras.up.pt

Cybercrime is a borderless issue that can be classified in three broad definitions: (a) crimes specific to the internet, such as attacks against information systems or phishing (e.g. fake bank websites to solicit passwords enabling access to victims' bank accounts); (b) online fraud and forgery: large-scale fraud that can be committed online through instruments e.g. identity theft, phishing, spam and malicious code; and (c) illegal online content, including child sexual abuse material, incitement to racial hatred, incitement to terrorist acts and glorification of violence, terrorism, racism and xenophobia[2].

The overall preparedness of the users for the current technologically connected world was tested over the last two years, when, due to the COVID-19 pandemic, the world went massively on lockdown, and people everywhere had to quickly adapt to living a significant part of their lives online; office work was replaced to a large extent by telework, in-person education gave way to online learning and teaching, online meetings replaced face-to-face meetings, and shopping was superseded by online shopping. Leading online lives was a way of mitigating at least part of the negative impact of the pandemic. The sudden move from in-person to online daily activities came at a cost: the massive use of online platforms put a strain on technological systems and infrastructure, which were not ready for the boom of users; hardware often failed to meet the increasingly demanding needs of users; and software revealed vulnerabilities that were previously unimaginable. Simultaneously, social practices had to be adapted and adjusted to meet the requirements of the so-called 'new normal'. This was not problematic for digital natives and tech-savvy users, and digital immigrants, who were expected to struggle to adapt, appear to have coped surprisingly well with this technological leap. This readiness was only apparent, because, under the surface, they remained digital immigrants whose self-perceived competences left them vulnerable to criminals, who in turn found in this new scenario unprecedented opportunities for cybercriminal practice. Cybercrime thus became more evident, by attracting the public and the media attention.

As a result of the growing number of cyberattacks, and of their diverse nature, Law Enforcement is faced with two major challenges. Firstly, the higher the volume of cyberattacks, the harder it is for Law Enforcement to dedicate the necessary resources (including human) to fight them. Additionally, as the volume of cybercrimes increases, so does the diversity and variety of such crimes, which in turn demands a constant realignment of Law Enforcement. Secondly, the range of sophisticated stealth and obfuscation technologies used by cybercriminals to help them remain anonymous online have a serious negative impact on the work of Law Enforcement. In extreme cases, where highly sophisticated means are used to cover for any traces of their online crimes, the positive identification of those criminals may be very hard, even nearly impossible; in other cases, access to crucial data – including metadata – may be barred by data holders, such as big techs, or even by the Law, leaving the forces with very little tangible data to investigate (cyber)crimes. Hence, granting legal access to data is essential to investigate cases of cybercrime.

An often-underestimated type of such data is language. As previously argued (Sousa-Silva, 2017), despite their anonymisation efforts, in a significant proportion of crimes (cyber)criminals resort to communication, and consequently use language in their criminal practice to communicate with victims, fellow criminals, or others. By doing so, they ignore the potential of language data to identify them (just like, metaphorically speaking, a 'linguistic fingerprint'). Indeed, as has been theoretically and empirically demonstrated, use of language is idiosyncratic, so every speaker of a language has a particular way of speaking and writing that distinguishes them from other speakers of the same language(s) (Coulthard, 2004). This field, which is known as forensic authorship analysis, is one of the many different applications of forensic linguistics.

The potential of forensic linguistic analysis for law enforcement

Language, which can be briefly defined as the system that humans have developed and use to communicate, is at the basis of the field of scientific enquiry known as Linguistics: the science that studies language structures and its use (see e.g. Finegan, 2008).

Linguistics as a (forensic) science has been victim of two mistaken assumptions. The first is that, because language is a social science, linguistics is often accused of being 'subjective' and lacking the validity and reliability criteria required by science, which prevents con-

2 See for further reference the website of the EU Commission at https://ec.europa.eu/home-affairs/cybercrime_en.

clusions to be measured and quantified, and error rates to be known. The second is that, because language is wrongly seen to be subjective, speakers frequently take themselves to be able to analyse language scientifically simply because they are native speakers (this fallacy is sometimes phrased as 'I could call upon the expertise of a linguist, but why would I need one if I can also read and write'?). Both these assumptions are obviously wrong; linguistics is indeed a science, it is objective, and it is bound by principles of validity and reliability, in much the same way as any other science. Additionally, although it is still very hard – if possible – to establish a known error rate, linguistic patterns can be measured and quantified, if the volume of data so permits. Moreover, no matter how proficient a native speaker of a language may be, their competence cannot compare to that of linguists, who have a deep knowledge of language acquisition and language structures, as well as of how language is used, depending on purposes, participants, contexts, etc. It is this knowledge of linguistics that is at the basis of Forensic linguistics.

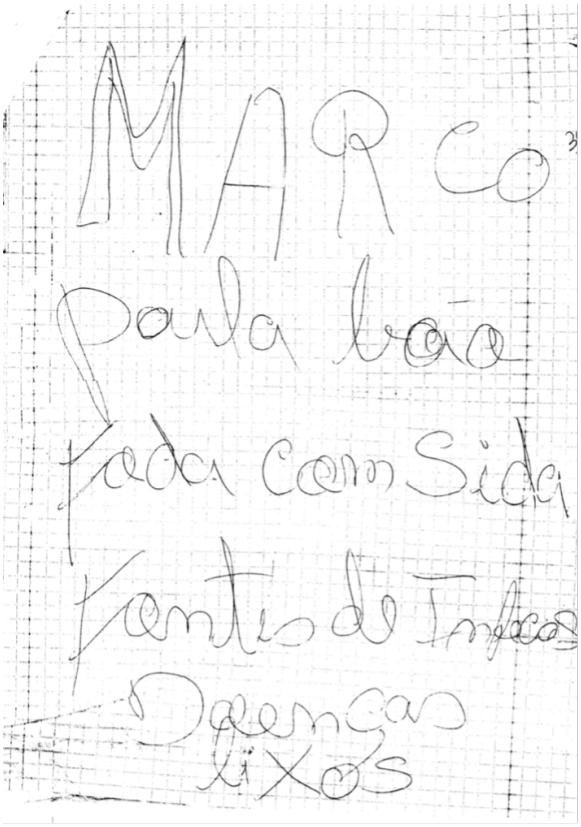
Linguistics can be approached from two main perspectives: a theoretical, which seeks to provide explanations for observable or possible linguistic phenomena; and an applied perspective, which focuses on language use in social interaction. Forensic linguistics is part of the latter, i.e., it is the branch of applied linguistics that consists of applying theories, knowledge and expertise of language sciences in forensic contexts (see e.g. Coulthard & Sousa-Silva, 2016) – whether to assist the courts of law, the investigative process or other issues that are of interest to the 'forum' in the traditional sense (i.e., the society in general (Turell, 2013)). Forensic Linguistic can thus be defined in a broad sense and in a narrow sense. In a broad sense, it subsumes three sub-areas: the study of the written language of the law, including interpretation of the language used in legal texts, such as laws and contracts, or the comprehensibility of legal language; the study of interaction in the legal process, which, in criminal cases, may include communications from phone calls to the police or to the emergency services, as well as police interrogation and interviewing, or interaction in a court of law; and the study of language as evidence (Coulthard, May, & Sousa-Silva, 2021), which includes, among others, authorship analysis (to establish which of a set of suspects is the most likely author of an incriminating anonymous text, or whether a suspect can be excluded as the author of that text), plagiarism detection and analysis (to help establish whether a text has been produced independently, and hence is original, or whether it was based on someone else's text), analysis of disputed meanings (in order to establish the more likely meaning of a disputed utterance), or sociolinguistic profiling (in order to establish the sociolinguistic features of the author of a suspect text). In a narrow sense, Forensic Linguistics is limited to the latter application, i.e., the study of language as evidence, including as an assistance to the investigation.

In the following two sections, two cases of cybercrime that illustrate the potential of the forensic linguistic analysis to uncover (cyber)criminal activities are presented and discussed, which are relevant for law enforcement. The first is the case of an anonymous set of text messages spreading defamatory contents, whose linguistic analysis enabled the investigation to establish the sociolinguistic profiling of the author, and consequently to narrow down the pool of suspects. The second case discusses a cross border cybercriminal practice: fraudulent and deceptive messages sent to citizens for purposes of extortion.

Establishing the sociolinguistic profiling of suspects

The first case in point is one of cyber-stalking. A set of anonymous text messages were sent from a pre-paid, unregistered mobile phone number spreading defamatory contents; they stated that a man (Marco) was HIV positive. An anonymous handwritten note (see Figure 1) was also circulated, reading that he and a woman he'd been seeing were 'spreading the disease.'

Figure 1: Handwritten anonymous message.



The complexity of the case was furthered by the fact that, if there were suspects, a set of texts known to have been written by each of them could be collected and compared to the suspect texts; safe assumptions could then be made as to whether the writing style of any of the suspects matched the writing style of the anonymous texts, as is common in authorship analysis tasks. However, since there were no suspects, no texts were available for comparison. Therefore, at best the investigation could try and establish the sociolinguistic profiling of the anonymous author(s) to narrow down the pool of suspects. Typically, sociolinguistic profiling (Coulthard & Sousa-Silva, 2016), which should not be confused with psychological profiling, is requested when the investigators do not have strong hypotheses about the identity of the author(s) of the suspect texts. The linguist is asked to find linguistic clues in the text that help establish social features of the author that reflect on the language used, such as age, gender, social and regional background, or level of education, among others. The purpose of sociolinguistic profiling is to identify the linguistic features of the group (in linguistics, the sociolect) to which the anonymous speaker may belong, rather than the linguistic features of the individual speaker (in linguistics, their idiolect). In other words, the aim of sociolinguistic profiling is not to find the exact (or even the most likely author) from all speakers of a language.

One of the challenges forensic linguists commonly face is related to the amount of data available for analysis. Ideally, linguists need considerable volumes of data to extract patterns from texts, and hence make safe assumptions about the writer. However, in forensic cases, the volume of text for analysis is usually small. In this particular case, the linguistic analysis focused on three sets of texts: the first set consists of text messages sent from an unregistered phone number (448 words in total); the second set consists of text messages sent from a second unregistered phone number (122 words in total); and the third set consists of the handwritten message shown in Figure 1 (16 words). The linguist was thus asked two questions by the investigation: (1) whether the three sets of texts were written by the same person; and (2) if they were written by the same person, whether there are some clues in the text that enable the identification of social characteristics (sociolinguistic profile) of the author.

The linguistic analysis of the three sets of texts revealed that they are highly likely to have been written by the same person because they share a large number of atypical linguistic patterns, including: use of slang and swear words, lack of prepositions, lack of punctuation (especially at the end of sentences), missing spaces between words, homophonic substitution (i.e., the correct spelling is replaced by how the words are pronounced), lack of accents in words, spelling errors[3] and lack of agreement in gender and number (in accordance with Portuguese grammar). Each of these patterns, individually, may not be relevant, since speakers from the same speech community, and who share identical social backgrounds, can share particular linguistic patterns, regardless of how idiosyncratic they may be. However, when used in combination with other idiosyncratic patterns, they can be highly identifying (or 'idiolectal', in linguistic terms), thus contributing to build the idiolectal style (Turell, 2010) of the writer. In this case, since the three sets of anonymous texts share identical linguistic features, it can be safely

3 The words 'error' and 'mistake' are used here with two distinct meanings: 'mistake' is used to refer to instances where an error is introduced by accident (as happens, for instance, with typos), regardless of the speaker's linguistic competence, whereas the word 'error' is used to refer to instances where those mistakes are made systematically, and hence do not result from accidental production.

assumed that they were written by the same person (even though the third set, the handwritten message, is very short, hence sharing fewer linguistic patterns). The following features, which are shared mostly by sets 1 and 2, are particularly idiolectal and hence relevant: (1) use of slang and swear words (e.g., 'merda'); (2) lack of agreement in number and gender (e.g. '*as merda* do' or '*dois bêbado*'); homophonic substitution (e.g., 'vo' for 'vou', 'inferniza' for 'infernizar', 'emcomoda' for 'incomodar'); and (3) misspelt words (e.g. 'emcomoda' for 'incomodar' or 'vo te' for 'vou-te').

In addition, these messages also include a unique phrase that is highly idiolectal: 'homem de sida' (literally translated into English as 'man of aids' to mean 'man with aids'). An example of a sentence where this phrase is used is 'Puta paga hotel para foder com homem de sida' (literally translated as 'Bitch pays hotel to fuck man of aids'). This phrase, which reads odd to any native speaker of Portuguese, is unique: when this analysis was first conducted, the exact phrase 'homem de sida' did not return any hits in Google[4]. What makes this phrase so unique is the use of the preposition, 'de' (English 'of'); although the words 'homem' (English 'man') and 'sida' (English 'aids') tend to keep company to each other very often (in linguistic terms, they are said to collocate very frequently), the grammatically correct preposition to be expected is 'com' (English 'with') and not 'de'. However, this phrase is used several times in different messages across the two sets, which demonstrates that its use is neither accidental, nor the result of an odd mistake; rather, its use is systematic, so the use of the correct alternative is not under the control of the writer.

Altogether, the analysis of these linguistic patterns provides us with several sociolinguistic clues to the origin and social characteristics of the writer, who is highly likely to be a woman in her mid-20s to mid-30s, with a low level of education, and from a low socioeconomic background. These patterns also indicate that the writer, most probably a black woman, originates from a Portuguese-speaking African country, highly likely, Angola. These patterns help narrow down the pool of suspects, by establishing that the writer probably belongs to a particular group of people, although they do not allow the analysis to precisely identify the individual writer of the questioned messages. This identification is only possible after the investigation has narrowed down the pool of suspects to just a few (typically two or three) writers, and a comparison is made between the questioned messages and sets of texts that are known to have been previously written by each of the suspects. When such an analysis is conducted, the unique phrase 'homem de sida' can potentially be highly idiolectal, and hence discriminatory, to identify the individual writer.

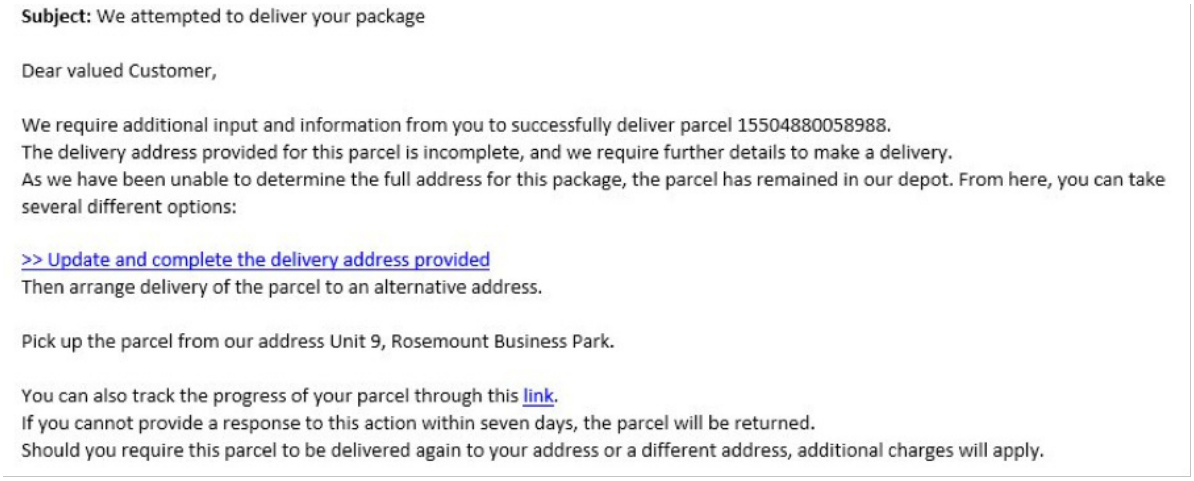
These findings, of course, need to be interpreted with caution because language is fluid, and although different social groups tend to share stable sociolinguistic patterns (see e.g. Labov, 1972), some features may span beyond those groups and be used by individual members of other groups. For this reason, sociolinguistic profiling is a very valuable tool for investigative purposes, but can hardly ever be used as evidence; for evidential purposes, forensic authorship analyses are more reliable.

Language use in cross-border cybercriminal practice

The previous section showed that sociolinguistic profiling consists of identifying a set of features that are typical of a certain sociolect, i.e., characteristics that are shared by a group of people from the same speech community. From a linguistics perspective, it is thus common for groups of criminals to share the same sociolect, that is, the same group of features. Therefore, an analysis identical to the one that is used for sociolinguistic profiling can also be relevant to identify cross border cybercriminal practices. Unlike traditional criminal practices, where criminals were, for the most part, geographically close, criminal groups are now expected to gather and operate cross-border. Therefore, it can be argued that technology has powered new, global forms of cybercriminal practices, which cross territories and jurisdictions. These practices may include, though not exclusively, threats, extortion, fraud, or cybercrimes such as cyber-trespass, cyber-fraud, cyber-piracy, cyber-porn and cyber-paedophilia, cyber-violence or cyber-stalking (see e.g. Wall, 2001), as well as scams, spoofing and phishing. Figures 2 and 3, written in English and in Portuguese, respectively, illustrate such criminal practices.

4 At the time of writing, Google only returns two results, both of which point to a book chapter where this case is mentioned to discuss linguistic identities.

Figure 2: Phishing email (in English).



Subject: We attempted to deliver your package

Dear valued Customer,

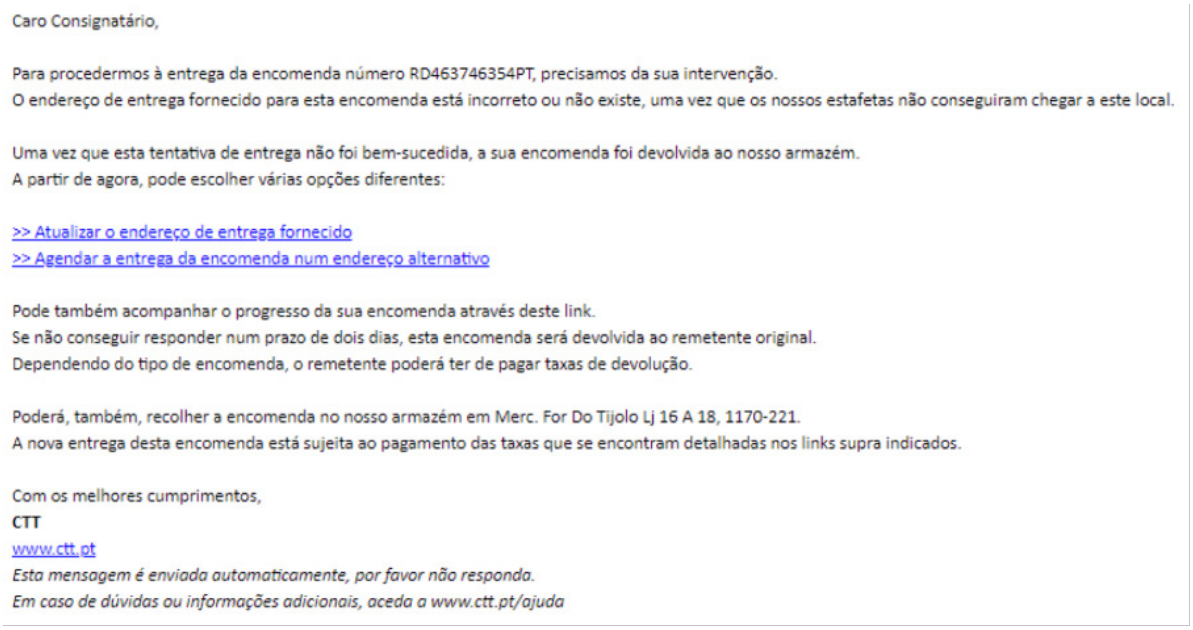
We require additional input and information from you to successfully deliver parcel 15504880058988.
The delivery address provided for this parcel is incomplete, and we require further details to make a delivery.
As we have been unable to determine the full address for this package, the parcel has remained in our depot. From here, you can take several different options:

>> Update and complete the delivery address provided
Then arrange delivery of the parcel to an alternative address.

Pick up the parcel from our address Unit 9, Rosemount Business Park.

You can also track the progress of your parcel through this link.
If you cannot provide a response to this action within seven days, the parcel will be returned.
Should you require this parcel to be delivered again to your address or a different address, additional charges will apply.

Figure 3: Phishing email (in Portuguese).



Caro Consignatário,

Para procedermos à entrega da encomenda número RD463746354PT, precisamos da sua intervenção.
O endereço de entrega fornecido para esta encomenda está incorreto ou não existe, uma vez que os nossos estafetas não conseguiram chegar a este local.

Uma vez que esta tentativa de entrega não foi bem-sucedida, a sua encomenda foi devolvida ao nosso armazém.
A partir de agora, pode escolher várias opções diferentes:

>> Atualizar o endereço de entrega fornecido
>> Agendar a entrega da encomenda num endereço alternativo

Pode também acompanhar o progresso da sua encomenda através deste link.
Se não conseguir responder num prazo de dois dias, esta encomenda será devolvida ao remetente original.
Dependendo do tipo de encomenda, o remetente poderá ter de pagar taxas de devolução.

Poderá, também, recolher a encomenda no nosso armazém em Merc. For Do Tijolo Lj 16 A 18, 1170-221.
A nova entrega desta encomenda está sujeita ao pagamento das taxas que se encontram detalhadas nos links supra indicados.

Com os melhores cumprimentos,
CTT
www.ctt.pt
Esta mensagem é enviada automaticamente, por favor não responda.
Em caso de dúvidas ou informações adicionais, aceda a www.ctt.pt/ajuda

The two emails, supposedly sent from legitimate post/parcel services, inform the recipient that a parcel could not be delivered to them because the address was incomplete (or incorrect, in the case of the email in Portuguese). The similarities between the two messages, despite their being written in two different languages, are striking, both in form and in contents. The Portuguese message even includes a reference to the official post website, which makes it more credible. However, both emails are phishing messages: "a fraudulent electronic communication that appears to be a genuine message from a legitimate entity or business for the purpose of inducing the recipient to disclose sensitive personal information" (Garner, 2009, p. 1263), such as login details, passwords or bank details. These deceitful communications usually attempt to route the user to false websites, where they are encouraged to provide confidential data.

Other deceitful communications include emails apparently sent from one's own email address stating that the sender is in full control of the computer, after malware has been installed upon visiting adult websites. Figures 4, in English, and 5, in Portuguese, illustrate these messages.

Figure 4: Extortion email (in English).

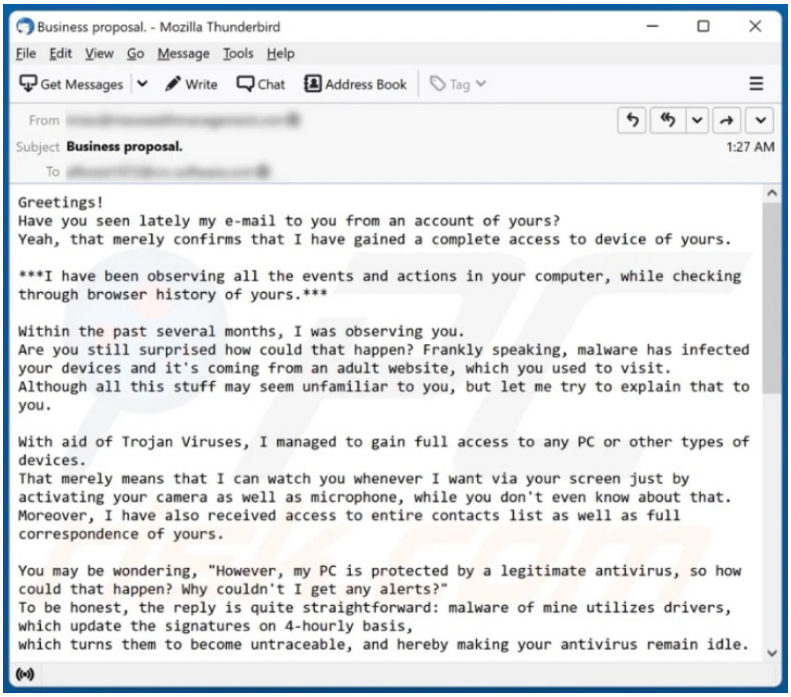
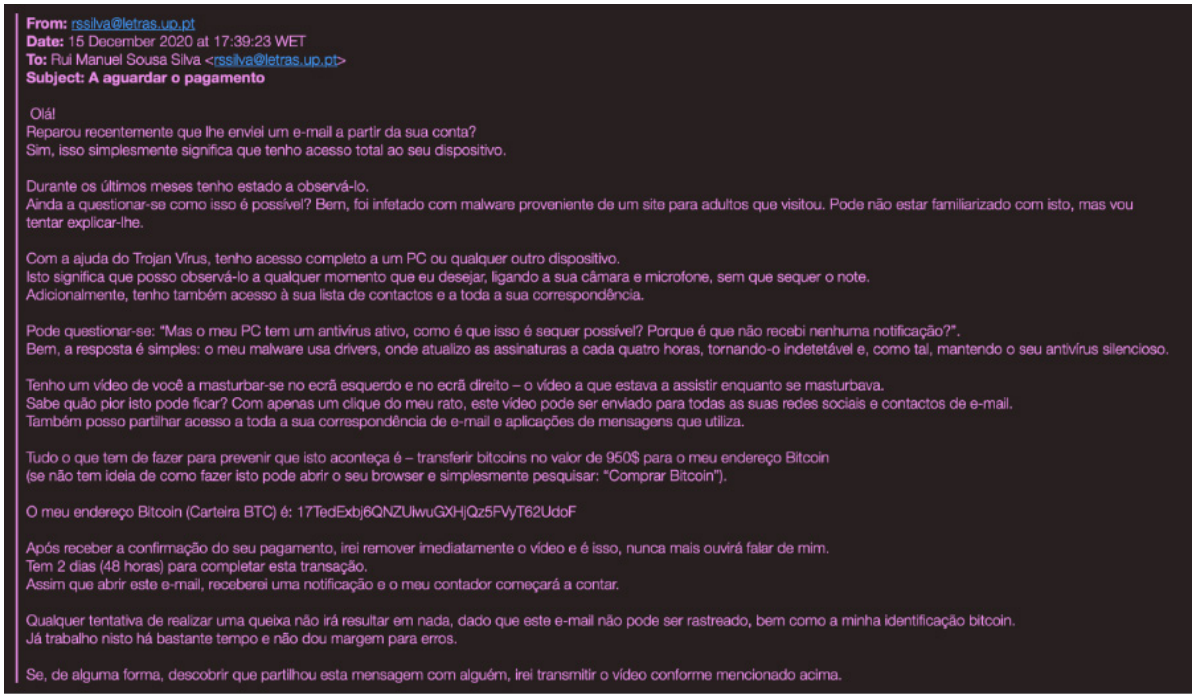


Figure 5: Extortion email (in Portuguese).



Communications of this type, which are attempts of extortion, are usually accompanied by a 'business proposal' or 'request for payment' and a ransom note stating that, if a sum is not paid (typically in crypto currency), then videos recorded by the sender showing immoral activities will be published or sent to the all the recipients in the victim's contact list. Although these messages are known to be fraudulent by a large part of the population, some users still worry that someone might have gained access to their computer,

so whether they have performed the action described or not is irrelevant to them; consequently, many victims still pay the sum demanded.

In both cases, the messages share some linguistic features that enable the identification of patterns of fraudulent and deceptive messages sent to recipients; in other words, a thorough forensic linguistic analysis allows the identification of features of language that enable the identification of the sociolect of the (cyber) criminals. Fraudulent and deceptive messages of this type traditionally contained a vast array of errors at all levels of language, including grammar, spelling, and punctuation. Over time, however, the quality of the deceitful text improved, and currently these communications very rarely include serious linguistic errors. Nevertheless, a careful reading and analysis of the texts reveals inconsistencies at the levels of cohesion (i.e., the relationship between items in a text) and coherence (i.e., the relationship between the items in the text and the extra-textual world), as well as minor grammatical mistakes. For example, the sentence 'I have gained a complete access to device of yours', although understandable to any speaker of English, is clearly not grammatically correct: 'a' in 'a complete access' is in excess, while 'to device of yours' is missing an article ('to a device of yours' would be more appropriate) or, even more appropriately, a possessive pronoun (e.g., 'to your device'), since the sender refers specifically to that same computer. Another grammatical mistake can be found in the sentence *'Although* all this stuff may seem unfamiliar to you, *but* let me try to explain that to you': in this sentence, the use of the two conjunctions ('although' and 'but', in italics) makes the sentence agrammatical. Examples like these abound in the texts.

It is also worth noting that the texts reveal peculiar patterns at the level of syntax (i.e., in sentence structure), which show that they were not originally written in that language. For instance, the structure of the sentences of the extortion text in Portuguese (Figure 5) is typical of English, so native speakers of the language (even non-linguists) will feel that the text is unidiomatic (or unnatural). Non-speakers of Portuguese can test this hypothesis by machine-translating the text into English: the more linguistically correct is the machine-translated text (called in translation studies the 'target text'), the closer the syntax of the source text (in this case, the Portuguese) is to English; conversely, the more the syntax of the target text differs from English syntax, the more likely it is that the source text has not been originally produced in English (see e.g. Sousa-Silva, 2013). The machine translated version, however, does not show major issues, which reveals that it is very close to English syntax.

These cases show that forensic authorship analyses, as well as an analysis identical to the one conducted in cases of sociolinguistic profiling, allows the identification of linguistic patterns that are typical of cross-border (cyber)criminal communications.

Linguistic analysis of disputed meanings

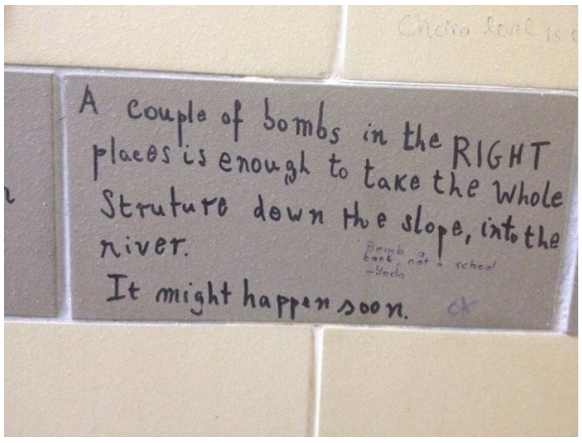
A relevant area of research in Forensic Linguistics is the analysis of disputed meanings, which consists of establishing the meaning of a textual element (such as a word, a phrase, or a sentence), confirming or rejecting the meaning associated with it, or analysing its linguistic uniqueness. Meanings are crucial because they underlie all instances of interaction among the speakers of a language and work to guarantee the communication among these speakers. In forensic contexts, the analysis of disputed meanings includes the study of suspect or illegal communications, cybercriminal messages, defamatory contents, text that infringes the 'property' of certain words (as in cases of plagiarism, copyright infringement or trademark disputes), as well as detection of hate speech and threatening messages (or, conversely, false threats).

Analysing disputed meanings can be problematic. When speakers of a language want to learn the meaning of a word, they usually refer to dictionaries, as these are supposed to compile the meanings of all the words in a language. Nevertheless, dictionaries do not suffice: firstly, new meanings emerge every day, either because new words are created, or because old words are re-signified (i.e., existing words can be given new meanings); secondly, dictionaries include the standard meaning(s) for words, but the precise meaning of an utterance can only be established in context. For instance, the sentence 'The bus was late.' can be used simply to inform the reader that the bus did not arrive on schedule, or – if the speaker is late – operate as a justification for their being late. Therefore, lexicographic definitions – the ones provided in dictionaries – can be useful to give speakers a general idea of the meaning of a word, but the precise meaning of an utterance always depends on its context, including setting, participants, purpose, etc.

Figure 6: Disputed meaning of a threatening utterance.

Figure 6 shows an illustrative example of disputed meanings. The message was written on the bathroom wall of FLUP ('Faculdade de Letras da Universidade do Porto', the Faculty of Arts and Humanities of the University of Porto). The utterance starts with the word 'Beware', thus cautioning the reader about something. The remaining of the message, however, is of a more informative nature, so most readers, when asked whether this utterance is a threat, will likely say it isn't. This message, however, is accompanied by another one, shown in Figure 7:

Figure 7: Disputed meaning of a threatening utterance.



The message illustrated in Figure 6 will gain a new meaning after reading the message in Figure 7: that of a threat. In combination, the two messages state the intention to conduct a certain (violent) act, convey the belief that this act will have negative consequences on the recipient, and have the intention to intimidate (Fraser, 1998). This threat is strengthened by the choice of words (e.g., 'bombs'), by the final sentence ('It might happen soon.'), and by contextual information: the Faculty is geographically located in the valley of the river Douro, hence the reference to 'down the slope'. This example shows that meanings are largely context-dependent, so an appropriate analysis of disputed meanings is essential, especially in investigative contexts.

Final remarks and recommendations for Law Enforcement

Language underlies all acts of human communication, including in criminal contexts, where it is crucial to interact with both victims and other criminals. Linguistic analysis is therefore a powerful tool in investigative contexts because criminals ignore that they can be identified by the language that they use – and even if they become aware of this fact, disguising one's language is usually not within the control of the speaker. Nonetheless, the power of linguistic analysis in forensic contexts has been underestimated, in no small part due to the mistaken assumption that, if we all learn the same language from the same books, then we all speak the language exactly the same; as has been empirically demonstrated, each speaker/writer of a language makes an idiosyncratic use of their language – their own 'idiolect,' in linguistic terms (Coulthard, 2004) – and that particular use is identifying. Therefore, this article strongly argues that linguistic analysis is crucial when investigating criminal activities, in general, and cybercriminal practices in particular, including: acts of cyber-violence; defamation; cyber-threats; dissemination of dangerous material; online harassment, cyber-bullying, cyber-stalking, or sexting; cyber-terrorism; hate speech; copyright infringement and piracy; and child pornography.

Despite its relevance for investigative purposes, research in forensic linguistics is frequently limited by access to data, and consequently insufficiently studied; in academic contexts, researchers can investigate and explore hypotheses using ordinary, naturally occurring data, so that the methods and techniques developed can later be used in forensic cases, if necessary. Ideally, however, such methods will be more reliable if developed and tested on real forensic data.

This article therefore concludes by making some recommendations for law enforcement: the first is that qualified forensic linguistics scholars are usually open to research collaboration with the forces, so police investigation, too, can take advantage of such research. This collaboration can start, for instance, by sharing forensic data for analysis. Notwithstanding the fact that

there are often access restrictions, including legal, to real forensic data, gains for the forces are potentially significant if authorisation is cleared.

My second recommendation is related to training: law enforcement officers do not usually have in-depth training in forensic linguistic analysis, and neither are all of them expected to further their knowledge in the short run; instead, cooperation with forensic linguistics scholars to provide expertise in the field to assist with real cases can be coupled with the offer of training activities for the forces, so as to allow officers to gain at least some basic knowledge of the relevance of linguistic analysis in forensic contexts.

In the future, technology will be increasingly integrated with human communication, which means that the boundaries between crime and cybercrime will tend to fade; therefore, language (and its analysis) will play a core role in the fight against crime. This is the future of digital age, so may law enforcement be ready for it.

Acknowledgements

This work was partially supported by grant SFRH/BD/47890/2008 and post-doctoral research grant SFRH/BPD/100425/2014, FCT-Fundação para a Ciência e Tecnologia, Portugal, co-financed by POPH/FSE, and by national funds by FCT – Fundação para a Ciência e a Tecnologia, I.P., project UID/00022/2020. The present research was conducted in cooperation with the Cybercrime Office of the Prosecutor General's Office.

References

- Coulthard, M. (2004) Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, *24*(4), 431–447.
- Coulthard, Malcolm, May, A., & Sousa-Silva, R. (Eds.) (2021). *The Routledge Handbook of Forensic Linguistics* (2.ª ed.). London and New York: Routledge.
- Coulthard, Malcolm, & Sousa-Silva, R. (2016) Forensic Linguistics. Em R. J. Dinis-Oliveira & T. Magalhães (Eds.), *What are Forensic Sciences? – Concepts, Scope and Future Perspectives*. Lisbon: Pactor.
- Finegan, E. (2008) *Language: Its Structure and Use* (6th, Inter ed.). Australia; United Kingdom: Wadsworth.
- Fraser, B. (1998) Threatening revisited. *Forensic Linguistics*, *5*(2), 159--173.
- Garner, B. A. (2009) *Black's Law Dictionary* (9th ed.). St. Paul, MN: West.
- Labov, W. (1972) *Sociolinguistic patterns*. Oxford: Basil Blackwell.
- Sousa-Silva, R. (2013) *Detecting plagiarism in the forensic linguistics turn* (PhD Thesis). Aston University.
- Sousa-Silva, R. (2017) *CybercrimeLab: A (computational) forensic linguistics approach against cybercrime*. Conference presentation at CEPOL 2017 Research and Science Conference INNOVATIONS IN LAW ENFORCEMENT Implications for practice, education and civil society, Budapest, Hungary.
- Turell, M. T. (2010) The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *The International Journal of Speech, Language and the Law*, *17*(2), 211–250.
- Turell, M. T. (2013) Presidential Address. In *Proceedings of the 3rd European Conference of The International Association of Forensic Linguists on the theme of «Bidging the Gaps between Language and the Law»*. Porto: Universidade do Porto - Faculdade de Letras.
- Wall, D. S. (2001) Cybercrimes and the Internet. In *Crime and the Internet* (pp. 1–17). London and New York: Routledge.