Using digital humanities and linguistics to help with terrorism investigations

Julien Longhi

Title : ***Using digital humanities and linguistics to help with terrorism investigations***

Author: Julien Longhi, CY Cergy Paris université

Julien.Longhi@cyu.fr
Institute of digital humanities
33 Bd du Port
F-95000 Cergy-Pontoise

Corresponding author: Julien Longhi

Abstract

This article seeks to offer a response to the digital transformation of forensic science by employing a tool-based linguistic analysis, integrated into the paradigm of digital humanities. It is a way to scientifically model the analysis of digital texts using digital methods. Computer science comes in support of linguistic skills in order to deal with investigative situations and help analyze criminal acts. It presents a case report thanks to the analysis of a corpus made up of 23 texts relating to criminal acts related to suspected terrorist groups with links to the far left. The goal is to help investigators by providing results which can help find stylistic similarities or exclusions between texts and thus potentially between the authors of those texts, offering authors profiling hypothsesis that may be included in the investigation process. While linguistics alone cannot solve such cases, a better understanding of language data, including topics, style and grammar, bring additional clues that can be very useful information in the investigation of crimes (linguists can "translate" information to investigators, so that il can be integrated to the investigation). Digital tools provide a form of objectification since they are based on statistical calculations which reveal regularities that are otherwise invisible to the naked eye. These tools, when used properly in investigations, can prove invaluable in extracting "clues" from the linguistic "traces" that make up texts.

Keywords

Linguistics, corpus, forensic linguistics, textometry, stylometry

Introduction

Humanities and social sciences and especially the emerging field of forensic linguistics are today becoming extremely important in the mechanisms for understanding threats. If, as Ribaux, Walsh, and Margot [1] explain, "it is recognized that forensic case data is still poorly integrated into the investigation and the crime analysis process, despite evidence of its great potential in various situations and studies", this is probably even more true for linguistic data. Following these authors who wish a change of attitude "in order to accept an extended role for forensic science that goes beyond the production of evidence for the court", this paper propose methods that are concretely applicable in risk contexts, in order to detect, characterize, and prevent threatening phenomena. It is possible to consider linguistics as a part of forensic science because it can follow the same process of collecting, analyzing, comparing, interpreting, and reporting linguistic clues in a forensic context. This approach is particularly similar to the typical process for source attribution, ACE-V (used in the field of documents in the 1970s and now systematically used in the field of digital traces):

A = analysis (extraction of characteristics);

C = comparison of characteristics (e.g. between two texts), identification of concordances and

discrepancies;

E = assessment of the value of matches (e.g. frequency, rarity, identifying power) and discrepancies (can they be explained or not, if not then exclusion from a common source).

Results allow to answer typical reconstruction questions in investigations:

- Who: can we extract or deduce information about the author(s) of illegal activities (a specific author, a group; or extract information about the author(s), for example age, gender, level of education, etc.)?;
- What: what are the corpora about? Can we use semantic features to understand the specificities of a document?
- When: linguistic features can give information on the event/facts (temporality, anteriority/posteriority, etc.).
- Where: sociolinguistic and geographic features can help to understand the origin of the author(s) or the place where they were when they produced the texts;
- How: stylistic features can help to understand how a text was produced (for example with a genetic approach);
- Why: performativity, intentionality are ways to examine what the goal of a text is.

The proposed paper aims to answer these questions through the prism of linguistics, by analyzing corpora. It will seek to answer, more or less precisely for the moment, these questions: can we characterize one or more authors, or groups of authors (who?); can we reveal thematic regularities in the corpora (what?); do we have geographic or temporal information in the texts studied (where / when?); can we make inferences about the author's intentions and about the means used (why / how)? In addition, the originality of this work is that it applies to texts in French, whereas most studies in the field focus on analyzing English. Its impacts allow bridging the gap, with extensions in future works, between the qualitative treatment of the threat and its automated processing through large corpora. More concretely, I shall describe a method for analyzing linguistic meaning based on identifying expression and content regularities from a statistical point of view. This method is part of a forensic science perspective, seeking to lay the foundations of a forensic linguistics which combines quantitative identification with qualitative analysis. The analysis corpus is made up of 23 texts relating to criminal acts which concern especially terrorist groups with links to the far left. The goal is to be able to help investigators with these cases, providing results which can help find stylistic similarities or exclusions between texts and therefore potentially between the authors of those texts.

To achieve the set goals, we need new improved tools that ally qualitative and quantitative analyses. This paper is a case example that can be generalized. To start with the way we can use the tools, Chaski [2] presents different approaches in forensic linguistics:

- Qualitative vs. quantitative;
- Prescriptive vs. descriptive;

- Analyst-based vs. Machine-based;
- Theory-based vs. practice-based.

The adoption of a quanti-qualitative approach made it possible to go beyond the limits of the two separate approaches. This global approach to analysis has been part of the lineage of my work for more than ten years ([3] [4] [5]), with various fields of application. The research initiated in 2015 around security issues testifies more than ever to the relevance of a longitudinal linguistic study: all levels of linguistic analysis can be mobilized to detect, identify, and characterize a threatening phenomenon. I will present different aspects in the "method" section.

For Chaski [6], dealing with a problematic close to this paper ("Determining who was at the keyboard"), authorship attribution can be approached through several avenues:

- biometric analysis of the computer user;
- qualitative analysis of "idiosyncrasies" in the language in question and known documents;
- quantitative, computational stylometric analysis of the language in question and known documents.

The second approach, "known as forensic stylistics, could be quantified through databasing", because "without the databases to ground the significance of stylistic features, the examiner's intuition about the significance of a stylistic feature can lead to methodological subjectivity and bias". The third approach, stylometry, "is quantitative and computational, focusing on readily computable and countable language features, e.g. word length, phrase length, sentence length, vocabulary frequency, distribution of words of different lengths". In these two cases, I follow Champod [7], about "Meaning of Identification/Individualization": "the problem of inferring identity of source is more complex than a simple dichotomy between class and individual characteristics", because, from a logical point of view, "the strength to be attached to forensic findings is essentially relative to the case and its value is best expressed using a likelihood ratio. The question of the size of the relevant population – which impact on prior probabilities – and decision thresholds are finally outside the province of forensic scientists but rightly belong to the fact finder".

With the preliminary results presented in this paper, the proposed approach fits the third approach described by Chatsky: it focuses on computable and countable language features (lexical and grammatical; I don't have a database as in the 2nd approach, but the first corpus of 23 texts can be augmented in future work, to become a reference database). From a general point of view, I follow Coulthard [8] who explains that "there are no specific forensic linguistic tools and […] the best training for a forensic linguist is a course in descriptive and applied linguistics": "each case will normally require a different selection from the basic descriptive linguist's toolbox". Chapter 6 of Coulthard and Johnson [9] has examples of forensic analyses focusing on morphological meaning, syntactic complexity, lexico-grammatical ambiguity, lexical meaning, pragmatic meaning, speech-to-writing transformation, narrative analysis and features of non-native language usage" (see

also [10] for more examples). With these variables, we can try to evaluate their intra and inter-variability: by comparing the intra-variability of the characteristic observed within the texts of the same author with the inter-variability of the characteristic within the text of different authors, we can highlight the useful characteristics (those which are less intra-variable than inter-variable), with the aim of reducing as much as possible the population of possible authors for a text.

Material

The material was given by the French Gendarmerie in order to help investigators with cases involving criminal acts which concerned especially terrorist groups with links to the far left. There are websites where these acts are claimed in writing, giving accounts of the facts (sometimes accompanied by photos), interpreting those facts, justifying them, etc. These texts are of course anonymous, and at this stage, it is not possible to know if the author of the (online) post is the author of the text, whether the author of the text is (part of) the terrorist (group), etc. Nevertheless, being able to characterize the style of these texts, establish similarities, show differences, or linguistically characterize the web pages could help investigators by formulating hypotheses on the possible number of authors or the probability that, for example, texts x, y, z have one, two, or three authors. This addition of linguistic characteristics should help investigators, if they can integrate them into analyzes which are at the same time based on other characteristics (for example physical, in connection with the examination of the crime scene).

Formally, the pages I analyzed came in this form:

Figure 1: Example of a claim site[1]

1 This page is accessible at: https://sansattendre.noblogs.org/post/2018/04/29/couflens-france-contre-le-projet-de-mine-de-tungstene-26-avril. The website is called "Sans Attendre Demain" ("Not waiting until tomorrow") and the title of the article reads: "Couflens, France : Contre la réouverture d'une mine de tungstène – 26 avril" ("Couflens, France: Against the reopening of a

In this page we can identify a lot of information that is potentially useful to investigators (information which, in the corpus, will become metadata): the place (Couflens), the date of the event (26 April), the date of the post (29 April 2018), the subject (opposition to (*contre* or "against") the reopening of a tungsten mine). With this example, we see the link with the other information (which makes it possible to answer the questions "who?", "When?", "Where?") not from digital elements (which can be false if the authors take precautions, such as using the darknet) but linguistic information.

Reading the article's introductory paragraph provides us with further details:

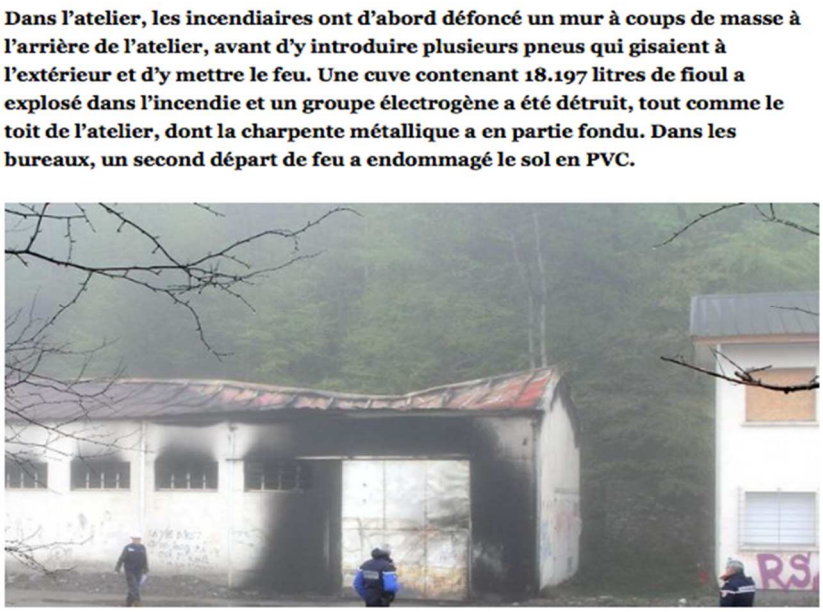

Dans l'atelier, les incendiaires ont d'abord défoncé un mur à coups de masse à l'arrière de l'atelier, avant d'y introduire plusieurs pneus qui gisaient à l'extérieur et d'y mettre le feu. Une cuve contenant 18.197 litres de fioul a explosé dans l'incendie et un groupe électrogène a été détruit, tout comme le toit de l'atelier, dont la charpente métallique a en partie fondu. Dans les bureaux, un second départ de feu a endommagé le sol en PVC.

Figure 2: Beginning of the article illustrated with a photo

Thus, the word *incendiaires* meaning "arsonists" tells us that this is about an arson incident, that a wall has been *défoncé* or "smashed down", for example.

Methods

Textometry offers an instrumented approach for dealing with corpora, combining quantitative syntheses and analyses which include text [11]. From a functional standpoint, textometry implements differential principles. This approach highlights the similarities and differences observed in a corpus according to the representation dimensions considered (lexical, grammatical, phonetic, prosodic, etc.). In addition to providing sorting procedures and statistical calculations for the study of digital corpora of texts, textometry establishes contextual and contrastive modeling. Thus, the text is characterized by its words in relation to their use in the corpus; the word is characterized by its

tungsten mine – 26 April"). The website pages were provided by the Gendarmerie in a spreadsheet together with several other pieces of information. The paper has been authorized to use the online pages and the data in them, but the additional information from the Gendarmerie cannot be published.

co-occurrences, etc. [12]. Textometry is particularly relevant to corpus exploitation in humanities and social sciences. It simultaneously enables a detailed and global observation of different texts while remaining close to them, and highlights the fact that language is an important observation field for these disciplines. The 23 texts contain a total of 12109 occurrences, 2534 forms with an average of 526 occurrences per text. This corpus size, although modest, is well suited to textometric methods, which makes it possible to compare corpora without having to resort to large training corpora. I consider it is more appropriate to use a word-frequency approach rather than bag-of-words or topic modeling for two reasons:

1) The size of the corpus and the nature of the data (specific texts without a big database to train a model) can be analyzed within the prism of digital humanities using corpus exploration rather than more computational approaches;
2) The analysis of textual data, according to Mayaffre and Vanni [13], "is essentially paradigmatic (tokenization, then selection of tokens according to the dictionary of frequencies, according to the morpho-syntactic label, according to the initial, etc.)" and artificial intelligence "is syntagmatic in essence (contextualization, that is to say the combination or sequentialization or convolution of the token caught in the linearity of the text)", which allows us to take advantage of both the linguistic and grammatical specificities of the corpus, and also the link with the metadata.

Ongoing work aims to combine the two methods and in particular to use convolutional neural networks.

In order to analyze the corpus in a way that would make it possible to analyze both the text produced and the contextual elements and investigative data, I tagged the corpus by putting "variables" (preceded by *) at the top of each document:

```
**** *ville_LIMOGES *faits_Incendie_VL_GIE *site_NANTES_INDYMEDIA
[Limoges] Attaque incendiaire d'une caserne de gendarmerie
Limoges. Dans la nuit du 25 au 26 octobre 2016 nous avons incendié la voiture d'un
chef de la gendarmerie des Tuilières garée dans l'enceinte de la caserne.

Limoges. Dans la nuit du 25 au 26 octobre 2016 nous avons incendié la voiture d'un
chef de la gendarmerie des Tuilières garée dans l'enceinte de la caserne.

Nous avons commis cet acte de sabotage en solidarité avec les migrant-e-s de la lande
de calais.

Cette semaine l'état français a déplacé de force et détruit les lieux de vies de

**** *ville_CREST *faits_Incendie_locaux_Enedis *site_NANTES_INDYMEDIA


A propos de dialogue, de solidarité et d'attaque
Attaque incendiaire pour un mois de juin dangereux

 En tant qu'individualistes il est difficile de parler de solidarité parce que nous
ne voulons pas l'exprimer à un groupe mais à des individualités desquelles nous nous
sentirions suffisement proches pour vouloir établir un dialogue.

Nous nous sentons solidaires des personnes qui de part leurs actes et leurs discours
```

Figure 3: Corpus and tagging

For this preliminary work I chose to enrich the data with three types of metadata: the city where the event took place (e.g. *ville_LIMOGES), the type of incident (e.g. *faits_incendie_VL_GIE) as characterized by the investigators, and the website where the text was published (e.g. *site_NANTES_INDYMEDIA).

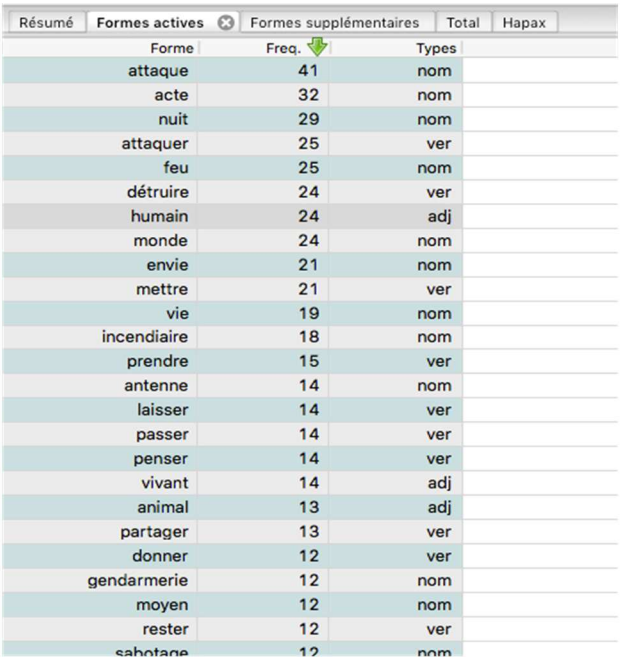
These metadata correspond to information given by the Gendarmerie about the cases in an xls file; for example, for incidents (*faits_XXX), they provided facts like:

- Fire on Enedis premises
- Damage to forestry machinery
- Barracks fire
- Damage to ecodistrict
- Destruction of hunting stand
- Wind-turbine fire
- Etc.

The addition of * before tags was due to my using the Iramuteq software which provides a set of analysis procedures for the description of a textual corpus. IRAMUTEQ is a GNU GPL (v2) licensed software that provides users with statistical analysis on text corpora and tables composed by individuals/words. It is based on R software and on python language. It is an R INTERFACE for the multidimensional analysis of texts and questionnaires.

One of its principal methods is Alceste. This allows a user to segment a corpus into “context units”, to make comparisons and groupings of the segmented corpus according to the lexemes contained within it, and then to seek “stable distributions” [14]. In addition to the Alceste method, Iramuteq provides other analysis tools including prototypical analysis, similarity analysis, and word-cloud analysis. All of these methods allow the users of this tool to map out the dynamics of the discourses of the different subjects engaged in interaction [15].

The corpus relating to real cases which the Gendarmerie provided contained 23 texts. These 23 texts were formatted as shown in Figure 3 and then subjected to a statistical analysis. A first basic analysis can be conducted on the frequency of forms (lemmas, see Table 1):



Résumé | Formes actives | Formes supplémentaires | Total | Hapax

Forme	Freq.	Types
attaque	41	nom
acte	32	nom
nuit	29	nom
attaquer	25	ver
feu	25	nom
détruire	24	ver
humain	24	adj
monde	24	nom
envie	21	nom
mettre	21	ver
vie	19	nom
incendiaire	18	nom
prendre	15	ver
antenne	14	nom
laisser	14	ver
passer	14	ver
penser	14	ver
vivant	14	adj
animal	13	adj
partager	13	ver
donner	12	ver
gendarmerie	12	nom
moyen	12	nom
rester	12	ver
sabotage	12	nom

Table 1: Frequency of lemmas[2] in the corpus

Methodologically, and from the perspective of a criminal investigation, this software is interesting because it can provide tools for exploring the corpus and not just quantified results. Thus, a user can go back to the corpus in order to see the various forms taken by a lemma or get a concordance, that is, see the forms as they appear in context (examples in Figure 4):

2 A lemma makes it possible to group together different occurrences of a word under the same unit (for example, in Figure 5, the lemma *attaque* ("attack") groups together the occurrences *attaque*, *attaques*; *attaquer* ("to attack") groups together the various conjugations of the verb; and so on).

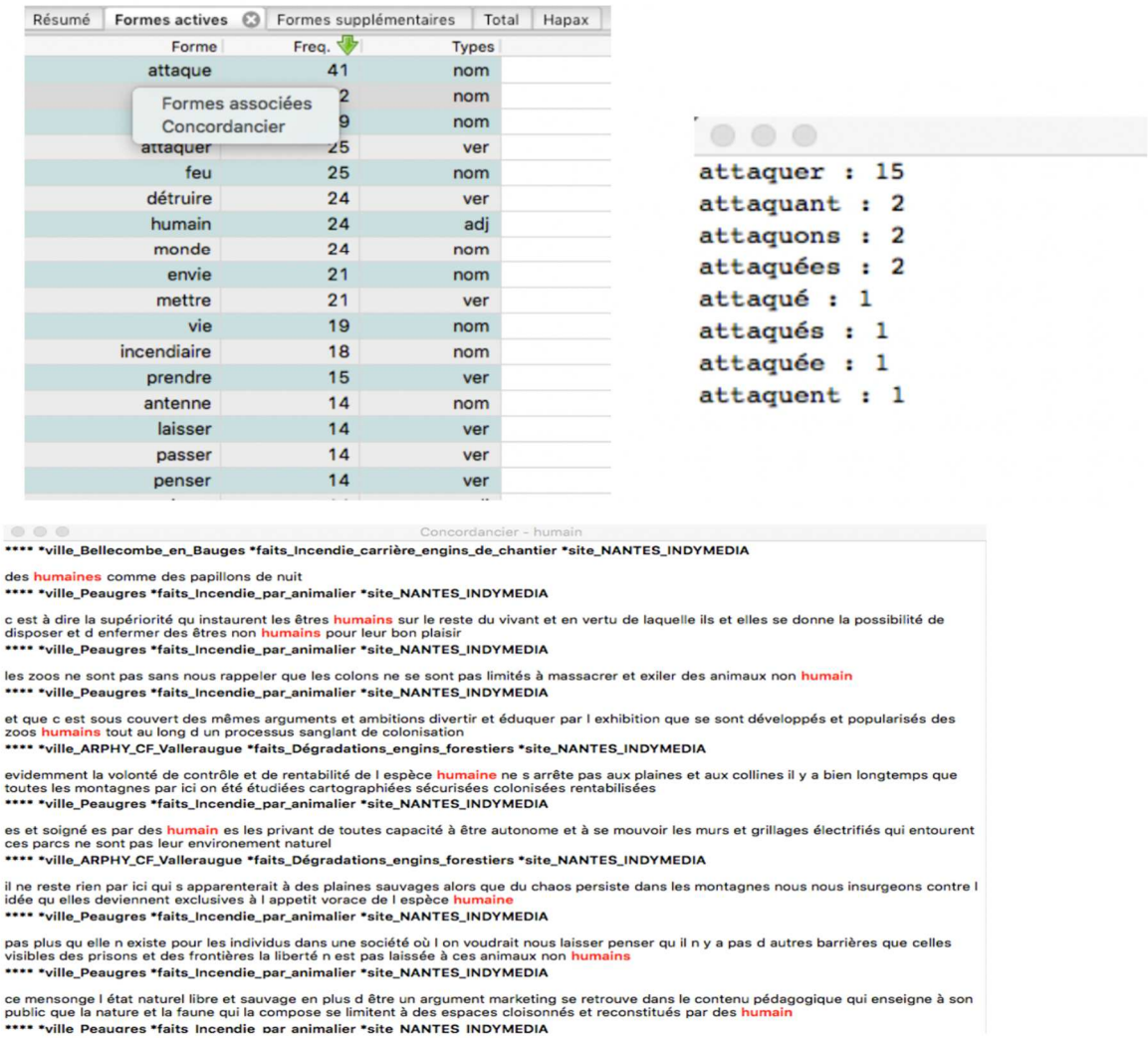


Figure 4: Corpus exploration (software functions), associated forms different forms of the same lemma), and concordance

This is very much an approach to collecting, analyzing, comparing and interpreting data – in this case textual data – which belongs to the forensic science approach, seeking to establish a method that combines identifying phenomena, returning to corpora, conducting linguistic analyses, and exploring phenomena. This first type of analysis has the advantage of providing the users of this software, and therefore potential investigators, with tools for exploring corpora, in particular procedures for counting, reorganizing data, and observing certain quantified phenomena. With this software "*Lexical Features*" ("single-content words as document features" [16]) but also grammatical features [17] were used.

With lexical features, we can use a Reinert-type classification proposed by the Iramuteq software: "this classification, implemented for the first time in the Alceste® software [14], makes it possible to highlight lexical worlds. These discourse structures assume that a statement is a stance that is dependent on the subject but also on its activity and context" [18]. At a methodological level, the vocabulary of the corpus is "used to build a double-entry table listing the presence/absence of the

full forms selected in/from the segments; a series of bi-partitions are [then] performed on this table based on a factorial analysis of correspondences." These classifications are very useful for understanding the topics of a corpus through the lexical worlds that compose them. They can be represented with descending hierarchical classifications (DHC).

In terms of the visualization and representation of the results, one can use the factorial analysis of correspondences (FAC): this is a statistical method "that can be applied to contingency tables such as tables resulting from counting different types of vocabulary (table rows) in different parts (table columns) of a corpus of texts" [19]. We start by "calculating a distance (known as the χ^2 distance) between each pair of texts making up the corpus. These distances are then broken down into a hierarchical succession of factorial axes. […] This method helps obtain synthetic representations of both the distances calculated between texts and those that can be calculated between the textual units that make them up." It is nevertheless important to note that while "the main advantage of FAC lies in its ability to extract from vast data tables that are difficult to grasp simple structures that can approximately reflect large underlying oppositions within a corpus of texts", this is only an "approximation" and the results of previous functions (calculations, tables of figures) must be precisely considered. These visualizations are an issue in the context of research in numerical humanities, which aim, in particular, to make complex results comprehensible through visualizations which are based on metrics and rigorous calculations.

Results and discussion

Two types of results, based on the methods described, can be provided: a lexical classification and a calculation of the proximity/distance between texts.

Lexical classification

With the lexical classification, it is possible to identify major topics (or lexical worlds, [20]) which could help group some of the texts together or objectively take into account the subject matter of the texts analyzed. Representing the data in the form of descending hierarchical classifications (DHC) in Figure 5 or factorial correspondence analyses (FCA) will then help visualize the results of the algorithm (see Figure 6).

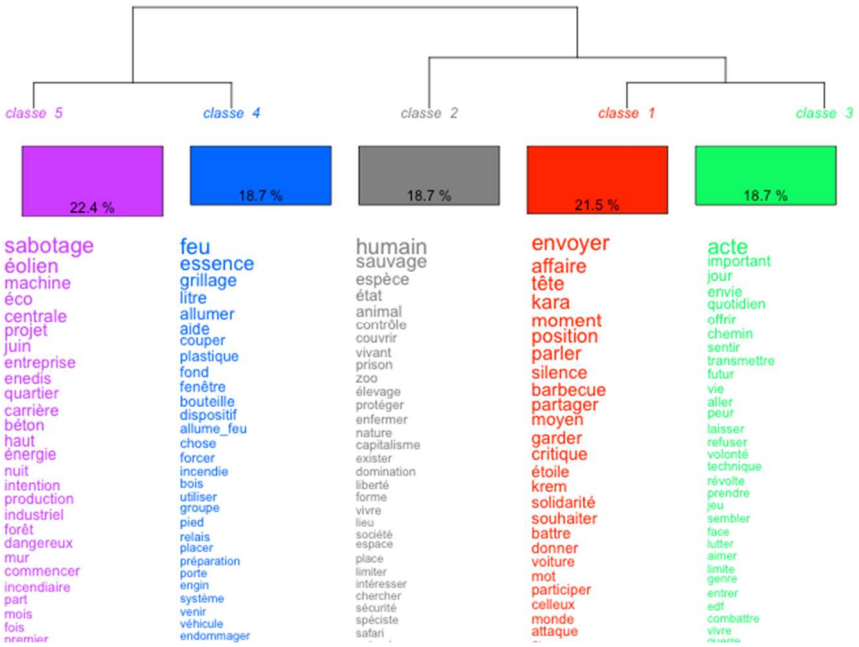


Figure 5: Corpus DHC (classification for the whole corpus)

We can thus see that 5 main classes could be identified: these classes, which were based on "text segments", showed that, for example, 18.7% of the text segments within the corpus fell under class 4 related to arson incidents (what the user could interpret through the words *feu* or "fire", *essence* or "petrol", *allumer* or "to light", etc.). 18.7% of the segments also referred to acts connected to animal species in class 2 (*humain* or "human", *sauvage* or "wild", *animal*, etc.).[3] This dendrogram can also be projected onto axes to help identify proximities which may exist between different parts of the corpus (in this case, each of the claim texts) by virtue of their shared connection to lexical worlds (see Figure 6).

3 Other topics are related to energy issues (class 5), the political dimension of actions (class 3) and the issue of discrimination (class 1). The class numbers have no meaning, but their proximity in the dendogram is based on a statistical proximity criterion

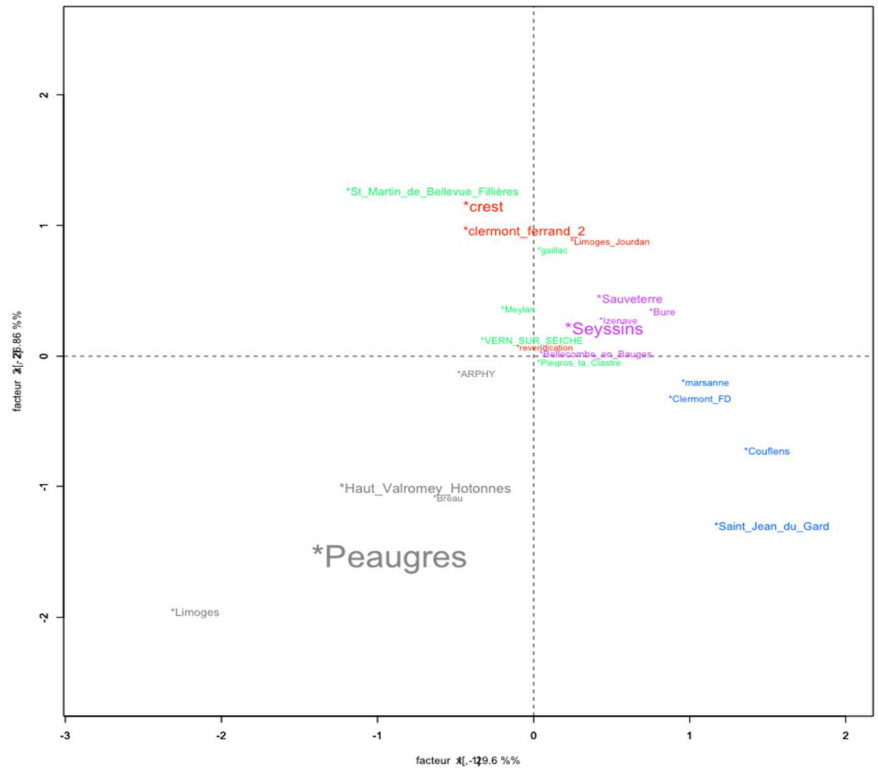


Figure 6: Corpus FCA (representing the *towns according to the topics of the CHD)

This way of representing the data does not really allow us to make hypotheses on the stylistic similarity between the texts but rather to see which texts could fall under the same theme (e.g. the four in blue would be related to arson incidents).

While the regularities identified here are mostly thematic and could be the subject of a manual classification, using forensic linguistics in order to see the doxas specific to different texts (i.e. the regularities underlying these texts, whether in relation to what is said or how it is said) helps us to go further in our investigations. Indeed, the links between topics and style is a complex subject. Mikros [16] distinguishes three types of thematic analyses in a quantitative perspective: Intra-Topic (IT), Single Cross-Topic (SCT) and Multiple Cross-Topics (MCT). In our case, the advantage of textometry is that the analysis is conducted on the existing corpus, which allows us to do calculations and make connections on the basis of the texts' vocabulary. Thus, the principle of differentiality helps us to measure distances and proximities rather than make classifications, which is relevant to the task at hand. The FCA (Figure 6) shows that classes 2 and 4 seem well separated from the others, while overlaps exist between some others.

Proximity/distance between texts

For this analysis we could use the software to calculate various specificities of each one of the texts so we may measure proximities or differences between them. These calculations could be based on different contexts which were coded using the following metadata:

- Towns as units
- Events as units
- Websites as units.

This linguistic approach brings more information than basic statistical approach (already well known in these contexts) because it uses specific and adapted statistics for textual analysis,[4] and brings different features (linguistic forms, grammatical categories, etc.). Here, the interpretation of the quantitative results by a linguist makes it possible to assess their significance (knowing that they are not random). This feature produces a factorial correspondence analysis in a contingency table which cross-tabulates active forms (common names, proper names, adjectives etc.) and function words (prepositions, articles) with variables. In this way, each form was assigned a (positive or negative) value in each text, which indicated whether the form was over- or underused in that particular text (Table 2):

Formes	Formes banales	Types	Fréquences des formes	Fréquences des types	Fréquences relatives des formes	Fréquences relatives des types	AFC
formes	*ville_ARPHY_CF_Valleraugue		*ville_Bellecombe_en_Bauges		*ville_Bréau	*ville_Bure	
que	4.3508		0.507		-0.5396	-0.8608	
montagne	3.5347		-0.1438		-0.0379	-0.107	
nous	2.8267		-0.244		0.6472	0.2788	
les	2.6641		0.4038		-0.4307	-0.3407	
parce_que	2.1662		-0.4977		-0.1312	0.2414	
si	2.1193		-0.2747		-0.0724	-0.2043	
parler	1.8249		1.3945		-0.0379	-0.107	
vouloir	1.7874		-0.7474		-0.1971	-0.1943	
rester	1.6462		-0.1569		-0.0414	-0.1167	
leurs	1.364		-0.3009		0.7773	-0.2238	
c	1.3434		-0.4977		-0.1312	-0.3702	
y	1.3434		-0.1619		-0.1312	0.688	
monde	1.2713		-0.314		-0.0828	-0.2336	
elles	1.1855		-0.3271		-0.0863	0.3676	
sauvage	1.177		-0.1438		-0.0379	-0.107	
nos	1.0446		-0.5291		-0.284	-0.8011	
tout	0.9473		-0.3009		0.7773	-0.2238	
il	0.9151		-0.8923		0.3785	0.3504	
vie	0.8592		0.3608		-0.0655	-0.1848	
et	0.8013		-0.3828		0.3445	0.4093	
ne	0.7752		1.1004		0.4738	-0.2443	
élevage	0.7579		-0.1307		-0.0345	-0.0972	
savoir	0.7579		1.4733		-0.0345	-0.0972	
toujours	0.6923		-0.2092		-0.0552	0.5212	
haut	0.6673		-0.1438		-0.0379	-0.107	
même	0.6622		-0.2878		-0.0759	-0.2141	

Table 2 Specificity table

4 The reference website on textometry defines the principle of calculating specificity as follows: it consists of "normalizing by dividing by the size of the part [which] makes us consider implicitly (or not) that the relative frequencies are representative of the original frequencies (before dividing by size). To this end, by committing as few errors as possible outside any complementary information, relative frequency can be taken to be the likeliest number of occurrences in a part of any size according to a law of normal occurrence. In a way, relative frequency is considered as a mode of normal probability distribution (the middle of the Gaussian bell curve where it is at its highest and therefore its likeliest), i.e. the average (cf. properties of the normal law: average, standard deviation, etc.). However, it so happens that the occurrence probability of a graphic form – or, more generally, a CQL expression – in a part has no reason to behave according to a normal law. That is to say, its distribution does not have to resemble a perfect Gaussian bell curve with an average, standard deviation, etc." Source: http://textometrie.ens-lyon.fr/html/doc/manual/0.7.9/fr/manual43.xhtml

Table 2 is of course hard to read and understand since it is made up of 23 columns (one for each text) and very many rows (corresponding to the number of forms). Using the method suggested by Ratinaud [22], it "shows the result of a correspondence analysis carried out on the full lexical table which partitions this corpus" according to the towns where the events occurred. This table is a simple contingency table with facts in columns and words in rows. The cells in the table contain the frequency of each word in each row.

The analysis "allows us to project on a 2-dimensional plane the relations" between the towns on the basis of their lexical co-occurrences, i.e. their tendency to have (or not to have) the same words in their texts. With Iramuteq, we could then calculate the distance between the texts according to these variables (textual statistics based on Labbé's intertextual distances). Labbé & Labbé [21][5] suggest to consider the "frequency of each type, that is to say, the entire texts (we use the adjective "textual" in order to show that the calculation is on N and not only on V or on a part of V)". Their metric "measures whether two or several texts are relatively far from one another", and their paper presents "a good approximation of the distances between several texts", with mathematical justifications.

This graphic visualization of the results makes it easier to interpret the proximities or distances between texts Figure 8).

5 Their work "consider the difference between "token" and "type". The token is the smallest measurable element in a text, and the «type» forms the vocabulary's basic element. For instance, the longest novel in French, Les misérables is made up of half a million tokens : its length or extent (noted N), while its vocabulary (noted V) is made up of less than 10 000 normalized and tagged types".

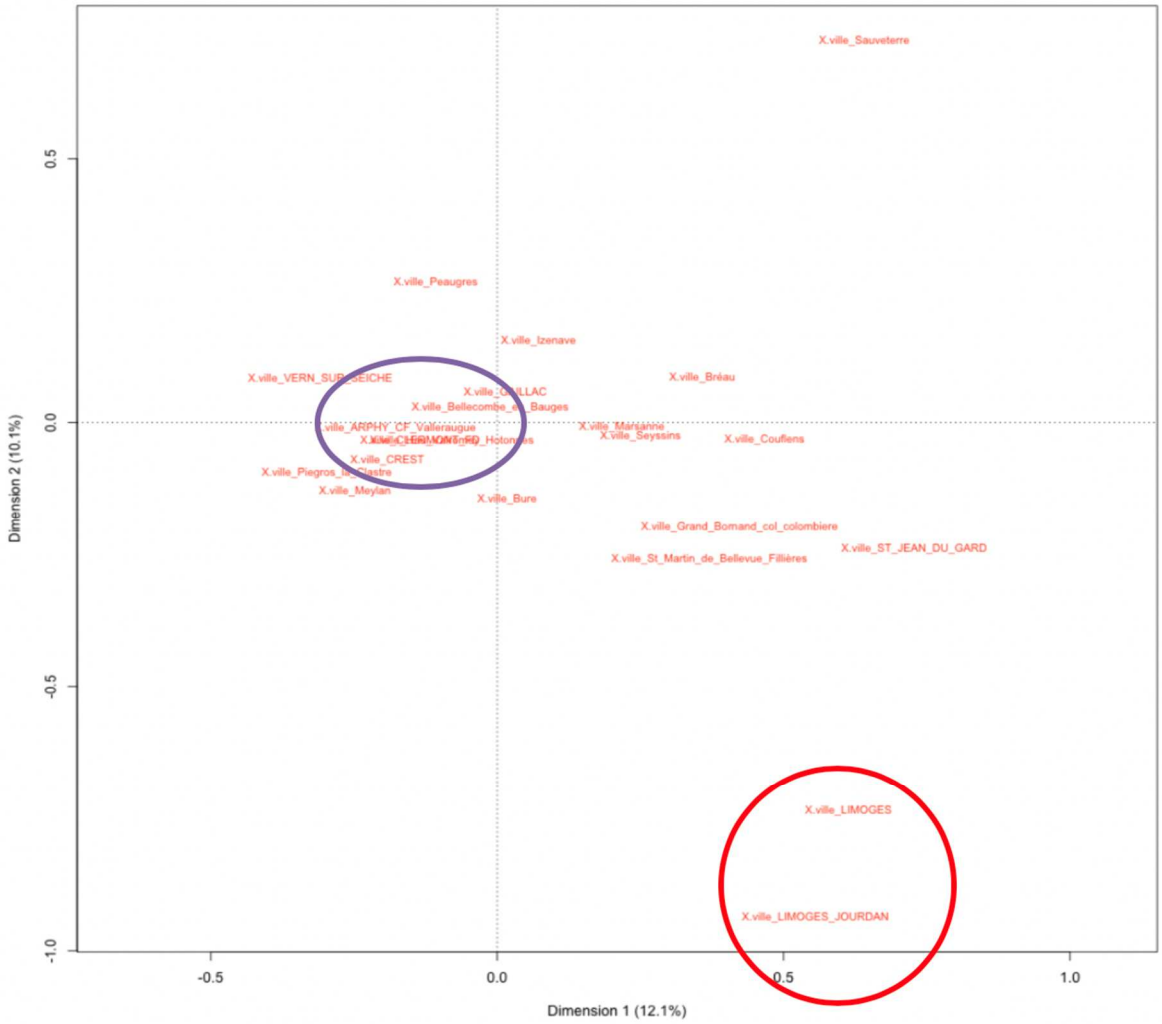


Figure 8: FCA on the texts' lexical specificities

A graph of this kind needs to be interpreted very cautiously (because of the instability of linguistic features, that it is necessary to verify by a return to the corpus, in particular because of the phenomena of polysemy), however, it does help us to consider grouping together texts on the basis of their lexical proximity, or excluding them on the basis of their distance. At this stage it is possible to come up with groups of variables (an example of which is shown in Figure 10 – all that is required is to note the proximity or distance between sets). We have seen above that the texts could also be grouped together based on their topics (towns and cities could be grouped together because the lexicon relating to the topics "saturated" the texts with numerous specific terms – for example feu ("fire"), essence ("petrol"), allumer "to light" for arson incidents). In order to examine the style independently of the topics, we could conduct the same type of analysis by taking into account grammatical of the words, rather than lexical aspects (i.e. the proportion of common names, definite articles, indefinite articles, adjectives, etc.).

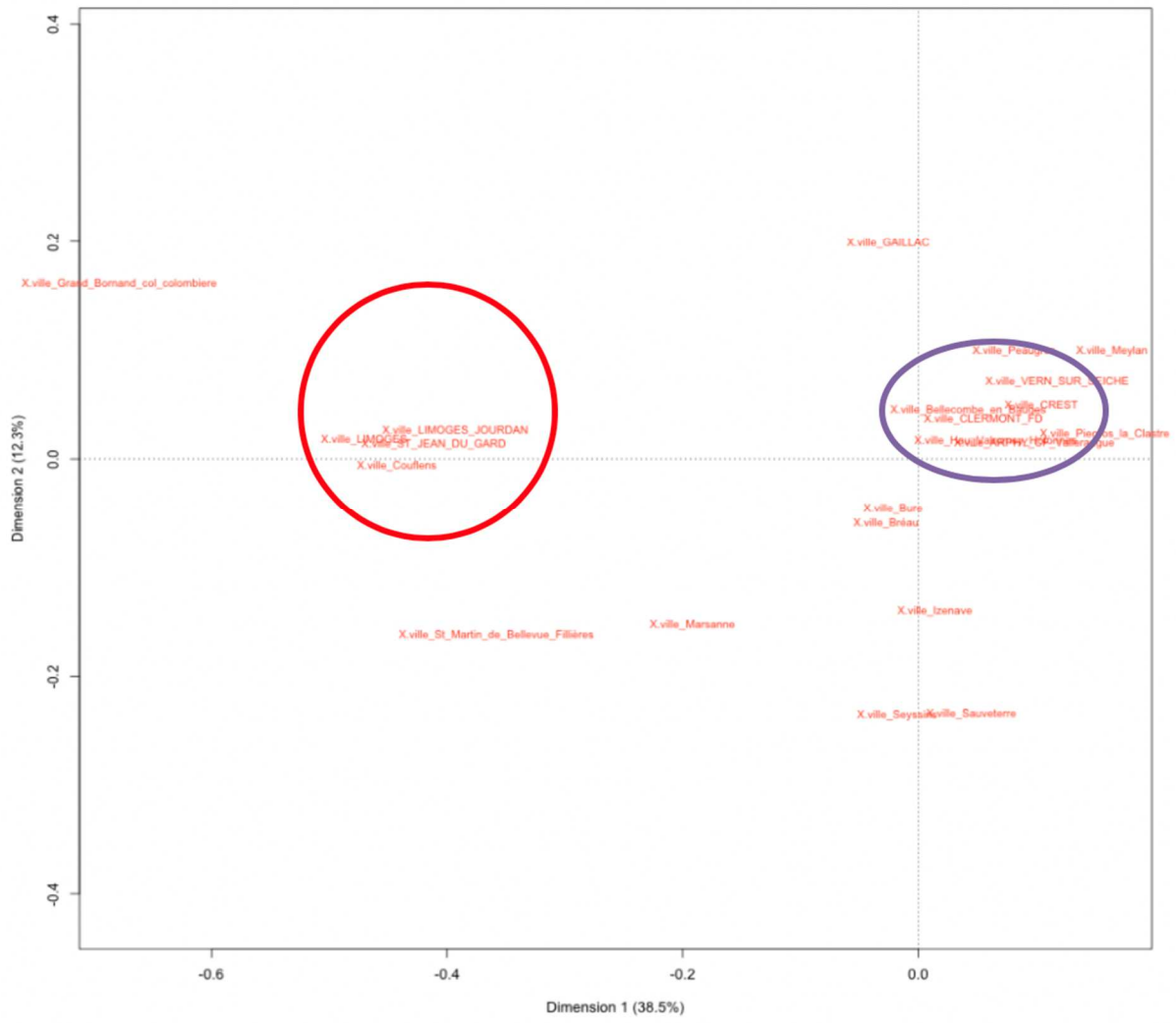


Figure 9: FCA on the texts' grammatical specificities

In Figure 9, we can observe other proximities, based on grammatical features. If the user wants more details about the proximities, he can have a look at table 2 and compare specificities of the text (grammatical specificities showed in figure 9, lexical specificities in Figure 8).

We can see the general grammatical specificities (categories) on a graphic (Figure 10):

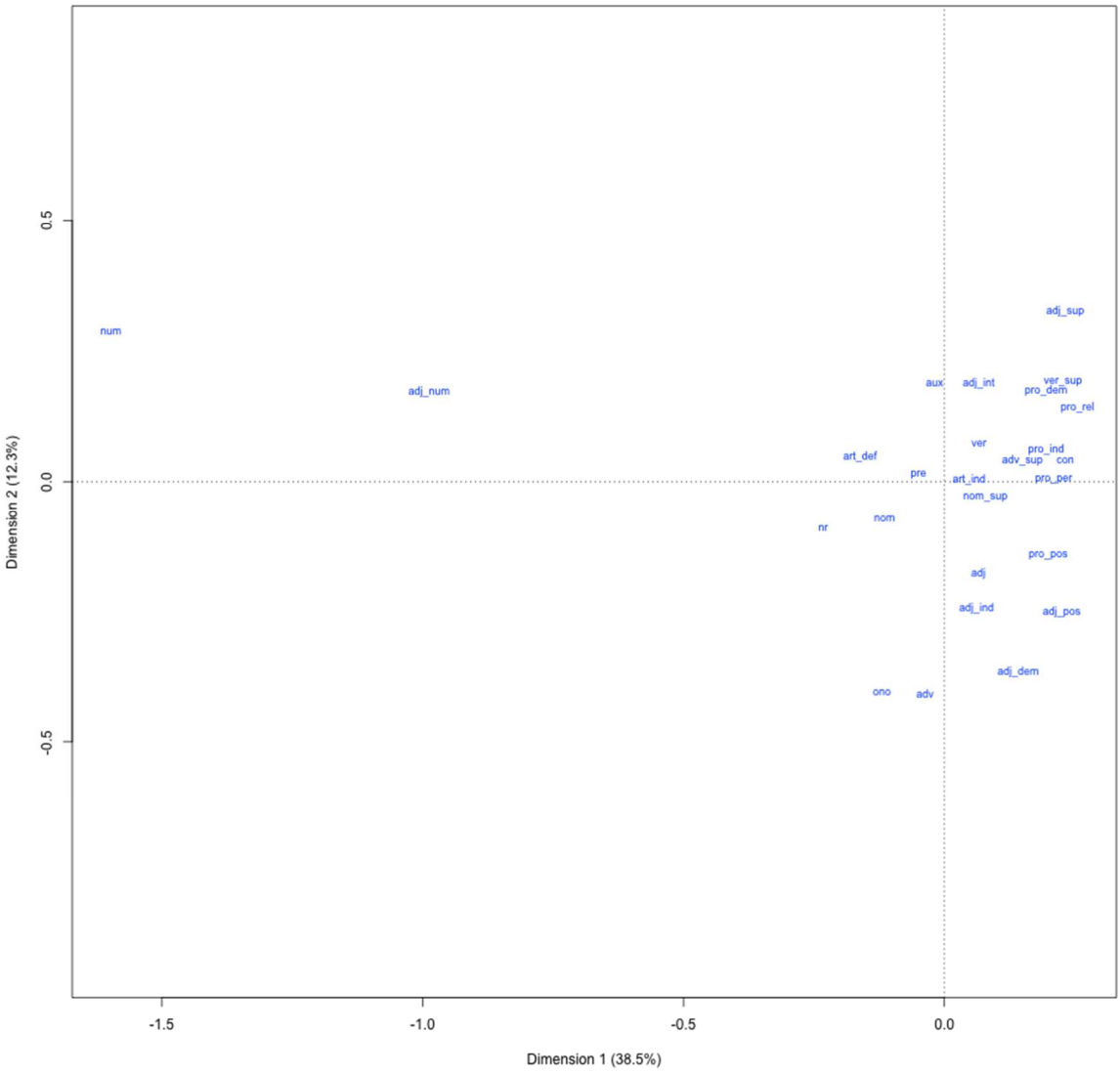


Figure 10: FCA with the grammatical features

We can thus see the linguistic categories characteristic of the different planes of the graphs, in order to be able to know what characterizes a given relation.

It is my hypothesis that cross-referencing Figures 10 and 11 would allow a user/investigator to identify similarities and exclusions by comparing the texts both lexically and grammatically. Thus, in the two graphs above, the texts relating to the events in Limoges and Limoges_Jourdan appeared close to each other (maybe because they use "adj_num" specifically), which could help investigators test the hypothesis of a same author/group of authors for these two incidents. Similarly, in the same two graphs there were proximities between the cities of Clermont_FD, Crest, and Bellecombe_en_Bauges (with other grammatical or lexical particularities: "pro_ind", pro_rel", "pro_dem", which shows the importance of pronouns). If we can clearly see the potential of such an approach, further step can be taken by providing a tool adapted to the needs of the Gendarmerie, with a specific interface and functionalities for identifying authors and comparing documents. This would

include superimposing specific data and metadata on the same graph, in order to more easily observe the connections, and the possible reasons for these connections. This is precisely what was achieved during a research project funded by CHEMI (Centre des Hautes Études du Ministère de l'Intérieur) as part of the 2020 call for proposals in strategic and prospective studies and innovations. The development of a prototype called "Text Print"[6] made it possible to lay the foundations of a method and tool adapted to this topic. This tool is still under development but what can be highlighted, apart from documentary and archival benefits, is the implementation of the functionalities mentioned in this article:

- N-grams of terms, grammar, and characters;
- Implementing an algorithm for calculating specificities;
- Implementing a classification method based on a Convolutional Neural Network trained on specific data (for the moment short online messages).

For example, with the same data used in this paper, the factorial correspondence analyses of the specificities can directly show on the graphic representation (Figure 11) the characteristic words of the parts (in red), and the variables (in blue), and here I choose to combine the lexical words and the grammatical words:

6 This project was carried out together with Alexandra Freeman (doctoral student at CY on copyright attribution issues), Jérémy Demange, prototype developer, and Trang Lam, Master's intern.

specificites > data-villes.txt

	Mots	ville_ARPHY_CF_Vallerangue.txt	ville_Bellecombe_en_Bauges.txt	ville_Bréau.txt	ville_Bure.txt	ville_CLERMONT_FD.t
2	controle	3.792301230956709	0.001	0.001	0.001	1.006990447665714
3	montagne	3.659828948273045	0.001	0.001	0.001	0.001
4	si	3.342262492187802	0.001	0.001	0.001	1.8490874131644521
5	que	3.261840258467336	1.3983657844629935	0.001	0.001	0.001
6	evidemment	2.6941607407313124	0.001	0.001	0.001	0.001
7	mot	2.419324640654438	3.270363854371355	0.001	0.001	0.001
8	zone	2.3234079719617484	0.001	0.001	0.001	0.001
9	parce	2.301403655993478	0.001	0.001	1.1167063785608	0.939851370482805
10	saluer	2.0494427333862726	0.001	0.001	0.001	0.001
11	sauvage	2.018433652942492	0.001	0.001	0.001	0.001
12	crever	1.7261059760907853	0.001	0.001	0.001	0.001
13	marteau	1.7261059760907853	0.001	0.001	0.001	0.001
14	emotion	1.7261059760907853	0.001	0.001	0.001	0.001
15	avancer	1.7261059760907853	0.001	0.001	0.001	0.001
16	excitation	1.7261059760907853	0.001	0.001	0.001	0.001
17	petit	1.506636050014152	0.001	0.001	0.001	0.9172224960183842
18	architecte	1.449100598137726	0.001	0.001	0.001	1.597236977541631
19	chaos	1.2510972748129245	0.001	0.001	0.001	0.001
20	temps	1.2510972748129245	0.001	0.001	0.001	2.548648350910009
21	finir	1.1677727834200837	0.001	0.001	0.001	0.001
22	monde	1.1587648747016484	0.001	0.001	0.001	0.7988774458423346
23	esperer	1.0986704479068339	0.001	0.001	0.001	1.2389259434087565
24	vivre	1.079990116310072	0.001	0.001	0.001	0.001
25	alors	1.065349179311053	0.001	0.001	0.001	0.001
26	haut	1.065349179311053	0.001	0.001	0.001	0.001

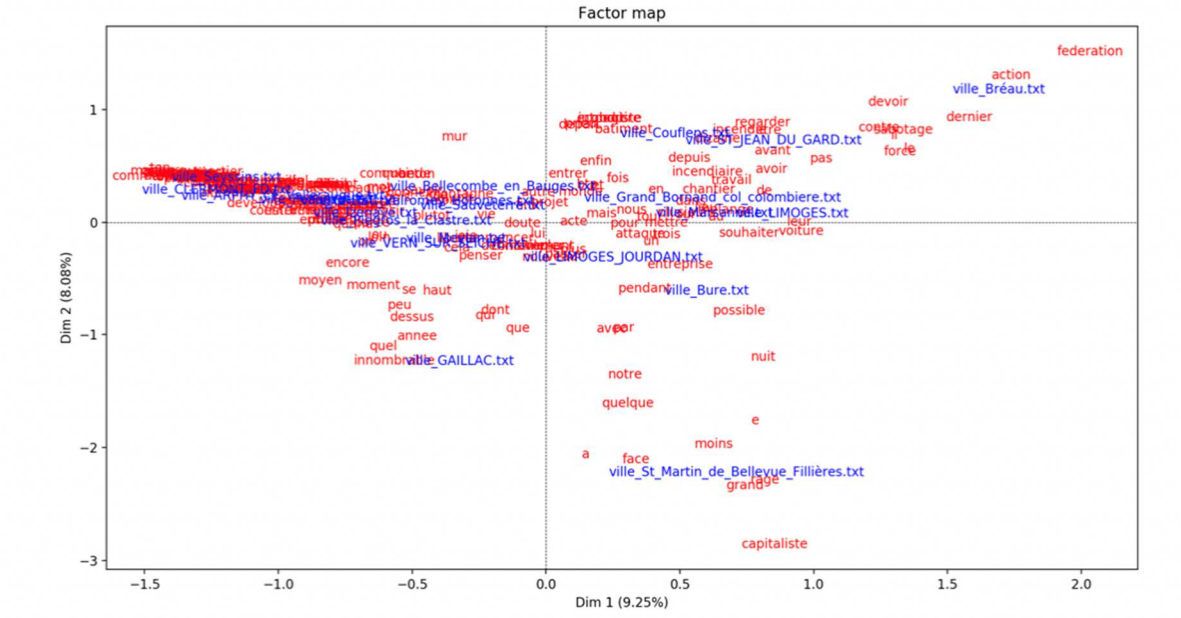


Figure 11: calculation and FCA with the prototype “Text Print”

Here, we distinguish texts that seem distant from others because of their vocabulary (Saint Martin de Bellevue, Bréau, Gaillac), and we can also make connections, on the basis of lexical words or tool words. The development of such a tool goes hand in hand with a process of mutual acculturation between academic linguists and Gendarmerie experts, aimed at generating new skills in forensic linguistics – a field which is currently absent from investigation teams. Work therefore needs to be done in the long run so that expertise and tools can be provided which can tackle scientific issues pertaining to the community of linguists and meet the expectations and needs of investigators. This means in particular that linguistic results are taken into account in an investigation, and investigators can provide linguists with clear feedback on the usefulness of the analyses, measures, and support they gave during the investigation. Today, I particularly work together with the technicians of the

Documents Department of the Criminal Research Institute of the National Gendarmerie (IRCGN)[7] who conduct physico-chemical examinations to determine the composition of the paper supports and inks and, if necessary, carry out comparative reviews. They analyze documents from the perspectives of support, inks and printing techniques. For example, they look for latent traces of fulling, determine printing techniques and authenticate documents. Handwriting and signing fall into the category of printing techniques: these are therefore personal and detailed human products. In this case, as in the forensic linguistics proposed in this article at the level of style, each writer personalizes the writing model that has been taught and develops a personal writing. The careful examination of a sufficient and comparable sample can allow the technician to determine if a contentious writing was produced by a given person (the writer). In the same way, the examination of a textual corpus can allow the linguist to determine if a contentious writing was produced by such a person.

Of course, other types of measurement are also possible. For example, in order to grasp certain specific turns of phrase or ways of saying something, we could use the "repeated segments" feature which helps us see specific uses of strings of several terms (Table 3):

formes		
l affaire de la voiture de	3	6
affaire de la voiture de	3	5
l affaire de la voiture	3	5
affaire de la voiture	3	4
de la voiture de	3	4
l affaire de la	3	4
affaire de la	3	3
d antennes relais	3	3
de la voiture	3	3
l affaire de	3	3
l attaque nous	3	3
la voiture de	4	3

Table 3: Repeated segments (class 1)

This level of analysis is also studied by Wright and Johnson [23]: these collocations and sequential strings of lexical words "can be called 'n-gram textbites', small portions of text that characterise the writing of a particular author". For the authors, these repeated segments "reinforce the benefits of a triangulation of approaches: corpus stylistic, statistical, computational, and case study". The corpus brings attention to textual data; stylistics allows attention of textual diversity; IT offers computation

7 In my collaboration with Cecile Jacob from the IRCGN (https://theconversation.com/la-linguistique-appliquee-aux-enquetes-criminelles-comment-ca-marche-143127#comment_2294798), I drew parallels between the methods of writing comparison and linguistics in order to validate my approach from a forensic point of view.

of data applied to the case study. In this case, the 3 occurrences of the segment "l affaire de la voiture de" are found in three different texts, which can then lead us to investigate more precisely the similarities between these three texts, and therefore between the three separate cases:

1) **** *ville_Meylan *faits_Incendie_caserne *site_SANS_ATTENDRE_DEMAIN
2) **** *ville_LIMOGES_JOURDAN *faits_Incendie_5_VL_GIE *site_NANTES_INDYMEDIA
3) **** *ville_ARPHY_CF_Valleraugue *faits_Dégradations_engins_forestiers *site_NANTES_INDYMEDIA

Here, textometry provides different ways of conducting syntactic and semantic analyses, which can provide information for observing and interpreting corpora. Using a rigorous method based on building and exploring corpora, linguistics can be a branch of forensic science and help investigators when textual data are exhibits in a case. With these results, investigators can use new hypotheses and, sometimes, find new connections that they would not have spontaneously thought of (for example Limoges and Limoges_Jourdan, or Meylan, Limoges_Jourdan and Arphy_CF_Vallerauge).

Discussion

The discussion concerns the specificity of the work presented here in light of ongoing research conducted in forensic linguistics. According to Coulthard [8], “there is comparatively little work [done] in France and on the French language”. This particularity is probably linked to the status of experts in French courts of law, but it is also, I believe, related to the French scientific context, in particular the fields of linguistics and discourse analysis.

Indeed, in the French intellectual and scientific context, general linguistics was a pioneering and leading science in the birth and development of structuralism (starting with Ferdinand de Saussure’s *Course in General Linguistics* in 1916). The view according to which socially stable representations attached to a topic impact discourse productions has played an important role in the French research community. As a linguist, Saussure is interested in languages as symbolic formations allowing to signify the specificities of cultures and societies [24]. Foucault [25] states that discourse is a social practice underpinned by areas of conventional knowledge called discursive matrices (*formations discursives*). A discursive matrix of social stereotypes related to work would, in France, comprise predicates concerning fatigue, money, freedom, constraint, etc. The existence of such matrices would explain why these predicates seem to come up with notable frequency in discourse productions. The study also addresses sociological issues linked to social structure and citizens’ perceptions. In *Distinction*, Bourdieu [26] aimed to describe the process of entering the social space, and the question of the passage from abstract situations to objective structures, which are nevertheless provisionally

reified. If we transpose that analysis of social structure and the analysis of discursive structure, this raises the problem of the reification of forms of discourse since the analyst can lose sight of discursive dynamics (the process of making discourse which is not only a result). Given that meaning is built by different dynamic processes, and according to different constraints, it is crucial to analyze it in context, in relation with the communities that are behind it or that interact, especially online. Discourse is therefore a field that semiotizes one's understanding of the social world and talking about discourse will include knowledge that speakers have about the social world and the semiotic contribution they bring to their discourse.

The concept of discourse as a semiotization of representations of reality is relevant here since it allows us to conceive of both lexical and grammatical linguistic data as traces left by subjects in their texts. These regularities, similarities, and differences are all elements that can turn traces into clues. Indeed, an expert linguist can make the transition from (linguistic) "trace" to "clue" from a strictly semiotic point of view rather than a legal one [27]. It seems necessary to develop real skills around the notions of textual semiosis, discursive spaces, or genericity, in order to understand how the linguistic "traces" left in these documents can become linguistic "clues" (in a semiotic way), which can then acquire an interpretable status within the framework of a procedure or an investigation. In some cases a "trace" is mediated, indirect, and transcribed (hearings, reported comments), while in others it is direct but strongly constrained by generic structures and a sociolect (autopsy reports, scene description). However, in all these cases the attention paid to these phenomena could help make the link to forensic science concepts of intra-variability (variability in the language of one people) vs. inter-variability (variability in the language of different peoples). In our study case, features for which inter-variability is larger than intra-variability can be used for identification/association purposes. For that, a clear distinction needs to be made between clue and trace: for Margot [28] "the definition of a clue corresponds to a probabilistic view of things":

> The material and physical nature of a trace makes it experimentally possible to measure the probability of finding a type of trace in a given situation. This probability varies greatly with the versions of the facts, being dependent on circumstances. It is only on the basis of the knowledge of the possible versions of the facts, or propositions, that the value as a clue or the quality of the information provided by the trace can be measured or assessed (p.79).

We can therefore conclude that there is strong agreement between this explanation of the relationship between trace and clue and the way in which the textometric analysis presented in this paper has worked: indeed, calculating specificity allows us to calculate the probability that an event will occur as many times as we effectively observe it in the part or even more frequently, up to the size of the part. Thus, a linguistic trace meets Margot's definition: "Trace is only an object with no meaning of its own. Its link to a case, and to reasonable hypotheses explaining its presence, in a way gives it its fundamental "*raison d'être*". It is the observed result that makes the reasoning possible, an inference

about a past fact. Thus, a trace becomes a sign when it is used for investigative purposes, or a clue when it is involved in a reconstruction or demonstration" (p.86). The term "sign" is particularly interesting since it is the founding term of semiology, the general science mentioned by Saussure, to which linguistics belongs. One of the additional interests of this research is therefore to offer an application to French corpora, which could be of interest to investigators in the very large French-speaking context (300 million speakers of which 235 million make daily use)

Conclusions

This article has presented the benefits of using linguistics, equipped with computing resources, in the analysis of online criminal content. Of course, linguistics alone cannot solve such cases. Nevertheless, it is undeniable that the corpus of texts analyzed above offers certain prospects for identifying similarities and differences, proximities and distances between texts. These results can then be submitted to the investigators who can work with linguists to look at different hypotheses in terms of topics, style, and grammar and to better understand language data. Digital tools provide a form of objectification since they are based on statistical calculations which reveal regularities that are otherwise invisible to the naked eye. They also make it possible to benefit from tools developed by linguists (grammatical analyzers, concordancers, dictionaries) and more broadly by the digital humanities (data visualization, corpus exploration). These tools, when used properly in investigations, can be invaluable in extracting "clues" from the linguistic "traces" that make up texts. Some methodological limitations can be easily removed since the proposed approach makes it possible to work both on languages other than French (the software presented here has versions in other languages too, working in the same way since the statistical approach does not require training a model on a particular language) and on larger corpora. If the results seem very case specific, and preliminary at this stage, future research will consolidate the methodology and test it on other cases.

Bibliographical References:

[1] Ribaux O., Walsh Simon J., Margot P. (2006), "The contribution of forensic science to crime analysis and investigation: Forensic intelligence", *Forensic Science International* 156 (2006) p.171–181.

[2] Chaski, C. E. (2008), "The Computational-Linguistic Approach to Forensic Authorship Attribution", In Frances Olsen, Alexander Lorz, and Dieter Stein, (Eds.), *Law and Language: Theory and Practice.* Düsseldorf University Press, https://www.semanticscholar.org/paper/The-Computational-Linguistic-Approach-to-Forensic-Chaski/b8bb5ee24b7c5084f61a745153ab1ae2300c1792

[3] Longhi J (2012), “Types de discours, formes textuelles et normes sémantiques: expression et doxa dans un corpus de données hétérogènes”, *Langages*, n°187, p.41-58.

[4] Longhi J. (éd., 2017), Digital Humanities, Corpora and Meaning, *Questions de communication*, n°31.

[5] Longhi J. (2020), “Proposals for a Discourse Analysis Practice Integrated into Digital Humanities: Theoretical Issues, Practical Applications, and Methodological Consequences”, Languages, 5(1), https://doi.org/10.3390/languages5010005

[6] Chaski, C. E. (2005), “Who's at the keyboard? Authorship attribution in digital evidence investigations”, *International journal of digital evidence*, 4(1), p. 1-13.

[7] Champod, C. (2013), “Overview and Meaning of Identification/Individualization”, in *Encyclopedia of Forensic Sciences*, p.303-309.

[8] Coulthard M. (2010), “Forensic Linguistics: the application of language description in legal contexts”, *Langage et société*, 2010/2 (n° 132), p. 15-33.

[9] Coulthard M. & Johnson A. (2007), *An Introduction to Forensic Linguistics. Language in Evidence*, London: Routledge

[10] Coulthard, M. (2013), “On Admissible Linguistic Evidence”, 21, *J. L. & Pol'y*, https://brooklynworks.brooklaw.edu/jlp/vol21/iss2/8

[11] Lebart, L. & Salem, A. (1994), *Statistique textuelle*, Paris, Dunod.

[12] Pincemin B. (2012), “Sémantique interprétative et textométrie– Version abrégée”, *Corpus* [En ligne], 10 | 2011. Mis en ligne le 15 juin 2012, consulté le 13 mai 2016. URL: http://corpus.revues.org/2121.

[13] Mayaffre, D. & Vanni, L. (2020), “Objectiver l'intertexte ? Emmanuel Macron, deep learning et statistique textuelle”, *JADT 2020*, Jun 2020, Toulouse, France.

[14] Reinert, M. (1983), “Une méthode de classification descendante hiérarchique: Application à l’analyse lexicale par contexte”, *Les Cahiers de L’analyse des Données VIII*, p. 187–98.

[15] Reinert, M. (1999), “Quelques interrogations à propos de l’objet d’une analyse de discours de type statistique et de la réponse ‘Alceste’ ”, *Langage et société* 90 (1), p. 57-70.

[16]_Mikros, G. (2013), “Authorship Attribution and Gender Identification in Greek Blogs”, https://www.researchgate.net/publication/236583622_Authorship_Attribution_and_Gender_Identification_in_Greek_Blogs

[17] Charaudeau, P.(1992), *Grammaire du sens et de l’expression*, Paris: Hachette Éducation

[18] Loubère, L. (2016), “L’analyse de similitude pour modéliser les CHD”, *JADT 2016 Conference*, Nice, France, June 7–10; Available online: http://lexicometrica.univ-paris3.fr/jadt/jadt2016/01-ACTES/83440/83440.pdf (accessed on 15 November 2019).

[19] Salem, A.. n.d. b. Tutoriels pour L’analyse Textométrique. Available online:

http://lexicometrica.univ-paris3.fr/numspeciaux/special8/tutoriel1.pdf (accessed on 15 November 2019).

[20] Ratinaud P. & Marchand P. (2015), "Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014)", *Mots. Les langages du politique* [En ligne], 108 | 2015. URL : http://journals.openedition.org/mots/22006

[21] Labbé C. & Labbé D. (2001) "Inter-Textual Distance and Authorship Attribution. Corneille and Molière", *Journal of Quantitative Linguistics*, Taylor & Francis (Routledge), 2001, 8 (3), p.213-231.

[22] Ratinaud P. (2016) "The Brilliant Friend(s) of Elena Ferrante: A Lexicometrical Comparison between Elena Ferrante's Books and 39 Contemporary Italian Writers". In Tuzzo A. & Cortelazzo M. (dir.), *Drawing Elena Ferrante's Profile*, Padova, Italy, p. 97-110. URL: http://www.padovauniversitypress.it/publications/9788869381300

[23] Wright, D. & Johnson, A. (2014), "Identifying idiolect in forensic authorship attribution: an n-gram textbite approach", *Language and Law (Linguagem e Direito)*,1, p. 37-69.

[24] Kharbouch A. (2014), "La séméiotique de Peirce et la sémiologie de Saussure", *Actes Sémiotiques* [online], 117, 2014, accessed 09/07/2020, URL : https://www.unilim.fr/actes-semiotiques/5218

[25] Foucault M. (1969), *L'archéologie du savoir*, Paris, Gallimard.

[26] Bourdieu P. (1979), *La distinction. Critique sociale du jugement*, Paris, éd. de Minuit, coll. "Le sens commun".

[27] Renaut L., Ascone L., & Longhi J. (2017), "De la trace langagière à l'indice liguistique : enjeux et précautions d'une linguistique forensique", *Études de linguistique appliquée: revue de didactologie des langues-cultures*, Klincksieck, 2017.

[28] Margot, P. (2014), "Traçologie: la trace, vecteur fondamental de la police scientifique", Revue internationale de criminologie et de police technique et scientifique, p. 72-97.

Website:
http://textometrie.ens-lyon.fr/html/doc/manual/0.7.9/fr/manual43.xhtml

Acknowledgements: I wish to thank the French Gendarmerie for our discussions and the links provided to the topic of this article. The discussions we had were particularly stimulating. I would also like to thank CY Cergy Paris université, which funded research in connection with the Gendarmerie within the framework of INEX, as well as the Ministry of the Interior (CHEMI IRITA project) and the University Institute of France (IUF)