



**HAL**  
open science

# Aspects textuels de la procédure judiciaire exploitée en analyse criminelle et perspectives pour son traitement automatique

Lucie Gianola

► **To cite this version:**

Lucie Gianola. Aspects textuels de la procédure judiciaire exploitée en analyse criminelle et perspectives pour son traitement automatique. Linguistique. Université de Cergy-Pontoise, 2020. Français. NNT: . tel-02522680

**HAL Id: tel-02522680**

**<https://theses.hal.science/tel-02522680>**

Submitted on 27 Mar 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



UNIVERSITÉ DE CERGY-PONTOISE

THÈSE DE DOCTORAT

*en*

Sciences du langage

École doctorale 284 : Droit et Sciences Humaines (DSH)

Centre de recherches AGORA (EA 7392)

---

**Aspects textuels de la procédure judiciaire  
exploitée en analyse criminelle et  
perspectives pour son traitement  
automatique**

---

*présentée par :*

Lucie GIANOLA

28 février 2020

*Jury :*

M. Laurent CHARTIER

Colonel de Gendarmerie, Examineur

M. Julien LONGHI

Professeur, Université de Cergy-Pontoise, Directeur

Mme Sylvie MONJEAN-DECAUDIN

Professeur, Université Paris IV, Rapporteuse & Présidente

M. Patrick PAROUBEK

Ingénieur de recherche HDR, LIMSI (UPR 3251), Rapporteur

Mme Bénédicte PINCEMIN

Chargée de recherche, IHRIM (UMR 5317), Examinatrice

M. Olivier RIBAUX

Professeur, Université de Lausanne, Examineur



*Entre la soutenance et le dépôt définitif de ce manuscrit, mon  
père Marc Gianola a été emporté par le cancer.  
À lui, qui a été et sera toujours notre roc.*



## Résumé

L'analyse criminelle est une discipline d'appui aux enquêtes pratiquée au sein de la Gendarmerie Nationale. Elle repose sur l'exploitation des documents compilés dans le dossier de procédure judiciaire (auditions, perquisitions, rapports d'expertise, données téléphoniques et bancaires, etc.) afin de synthétiser les informations collectées et de proposer un regard neuf sur les faits examinés. Si l'analyse criminelle a recours à des logiciels de visualisation de données (i. e. Analyst's Notebook d'IBM) pour la mise en forme des hypothèses formulées, la gestion informatique et textuelle des documents de la procédure est entièrement manuelle. Or, l'analyse criminelle s'appuie entre autres sur le concept d'entités pour formaliser son travail.

La présentation du contexte de recherche détaille la pratique de l'analyse criminelle ainsi que la constitution du dossier de procédure judiciaire en tant que corpus textuel.

Nous proposons ensuite des perspectives pour l'adaptation des méthodes de traitement automatique de la langue (TAL) et d'extraction d'information au cas d'étude, notamment la mise en parallèle des concepts d'entité en analyse criminelle et d'entité nommée en TAL. Cette comparaison est réalisée sur les plans conceptuels et linguistiques. Une première approche de détection des entités dans les auditions de témoins est présentée.

Enfin, le genre textuel étant un paramètre à prendre en compte lors de l'application de traitements automatiques à du texte, nous construisons une structuration du genre textuel « légal » en discours, genres et sous-genres par le biais d'une étude textométrique visant à caractériser différents types de textes (dont les auditions de témoins) produits par le domaine de la justice.



## Abstract

Criminal analysis is a discipline that supports investigations practiced within the National Gendarmerie. It is based on the use of the documents compiled in the judicial procedure file (witness interviews, search warrants, expert reports, phone and bank data, etc.) to synthesize the information collected and to propose a new understanding of the facts examined. While criminal analysis uses data visualization software (i. e. IBM Analyst's Notebook) to display the hypotheses formulated, the digital and textual management of the file documents is entirely manual. However, criminal analysis relies on entities to formalize its practice.

The presentation of the research context details the practice of criminal analysis as well as the constitution of judicial procedure files as textual corpora.

We then propose perspectives for the adaptation of natural language processing (NLP) and information extraction methods to the case study, including a comparison of the concepts of entity in criminal analysis and named entity in NLP. This comparison is done on the conceptual and linguistic plans. A first approach to the detection of entities in witness interviews is presented.

Finally, since textual genre is a parameter to be taken into account when applying automatic processing to text, we develop a structure of the « legal » textual genre into discourse, genres, and sub-genres through a textometric study aimed at characterizing different types of texts (including witness interviews) produced by the field of justice.



## Remerciements

On évoque souvent la thèse comme une course, un marathon qui s'achève au terme de trois, quatre, cinq années. Dans mon cas, j'ai plutôt le sentiment que ces trois ans sont l'aboutissement (mais pas la fin) d'un parcours modelé par de nombreuses personnes rencontrées depuis bien plus longtemps qui ont affûté mon goût d'apprendre.

J'ai eu la chance à plusieurs reprises de trouver dans mon parcours scolaire et universitaire des professeurs extraordinaires qui m'ont appris à lire, à écrire, à penser. Même si elles n'ont pas contribué directement à ce travail de recherche, je les compte parmi celles et ceux qui m'auront donné les moyens d'arriver au bout, ou plutôt à *un* bout de ma curiosité. Merci en particulier à Anne-Pascale BOUBE qui a été bien plus qu'une prof de philo.

Les travaux de Mmes Sylvie MONJEAN-DECAUDIN et Bénédicte PINCEMIN m'ont inspirée et guidée dans mes recherches. Je suis heureuse de pouvoir les compter dans mon jury et les en remercie.

Je remercie M. Olivier RIBAUX de m'avoir très gentiment reçue à Lausanne et de m'avoir orientée dans une vision scientifique de l'enquête et de l'analyse criminelle.

Merci au Colonel Laurent CHARTIER d'avoir immédiatement accepté de prendre part à ce jury.

M. Patrick PAROUBEK, en plus d'être rapporteur de ce manuscrit, a guidé il y a longtemps les premiers pas que je faisais par hasard dans la recherche. Il ne fait aucun doute que sans cette rencontre (et celle de tous les stagiaires, doctorants et chercheurs que j'ai côtoyés au LIMSI), j'aurais pris un autre chemin. Je le remercie de sa confiance envers moi plusieurs fois renouvelée.

Je remercie mon directeur de thèse M. Julien LONGHI de m'avoir fait confiance et de m'avoir recrutée pour ce projet.

Merci aux relecteurs anonymes du volume 60, numéro 3 de la revue TAL consacré aux Humanités Numériques pour leurs retours riches et constructifs.

Le major Christophe KRUCKER a initié ce projet et a partagé avec moi son expérience d'analyste criminel, de la procédure et des enquêtes. Je le remercie chaleureusement et lui souhaite de belles réussites apicoles.

Je remercie les analystes criminels de l'équipe DSAC du SCRC pour leur accueil et leur disponibilité. J'ai été honorée de mettre mes recherches au service de leur travail méconnu et pourtant potentiellement crucial dans des cas souvent douloureux. Merci à la commandante Léa JANDOT et à la lieutenant Peggy BONNET pour leur intérêt pour mon travail, leur soutien et leur optimisme.

L'adversité fédère : je remercie mes camarades doctorantes et docteur de Cergy Laurène, Manon et Laura, toujours prêtes pour une réunion de crise autour d'une pizza. La recherche est un milieu particulier mais j'y ai aussi trouvé beaucoup de solidarité et d'entraide.

En me tolérant dans ses bureaux, l'ERTIM de l'Inalco m'a offert un précieux lieu de socialisation pendant plus d'un an et demi. Pour m'avoir accueillie presque comme une membre à part entière, je remercie les adultes Jean-Michel, François, Damien, et Mathieu qui a prêté l'oreille à mes plaintes, qui a nourri mes questionnements, et qui m'a éclairée sur le sens de la thèse (mais pas sur le sens de la vie). Les enfants (et anciennes camarades de master...!) Amélie, Jennifer, Qin Ran, Li Yun, pour leur solidarité, les discussions existentielles sur la thèse et sur le reste, les coups de pouces de toutes sortes. Bénédicte pour avoir partagé avec moi sa connaissance de la procédure et des tribunaux, son *Vocabulaire Juridique* et une traversée du lac d'Annecy à la nage.

En effet, ce que j'ai fait en dehors de la thèse a autant compté pour mener le projet à bien que la recherche en elle-même.

Alors, je remercie les membres de la team Bribri pour leur bonne humeur qui me porte depuis *l'annus horribilis* 2015. Merci à Alice, qui a tenu à me conduire à mon premier jour de thèse et pour qui j'aurai l'honneur de signer un papier administratif dans quelques mois. Merci à sa mère la formidable Brigitte DESSUTTER.

Ayant vécu la thèse comme une galère, j'ai décidé d'apprendre à ramer : merci à mes camarades d'aviron de la Société Nautique d'Enghien notamment son président Vincent et son entraîneur Frédéric qui m'ont appris les secrets d'un « bon coup de pelle », et à mon équipage habituel du 408 pour les sorties du samedi sous le soleil, la pluie, le vent.

De manière générale, je remercie tous ceux et celles qui m'ont apporté de l'air dans ce long tunnel, tous ceux qui m'enrichissent d'une façon ou d'une autre, qui

me transmettent des choses, tous ceux qui alimentent mon éclectisme et mes lubies. Parmi eux, je remercie ma psychothérapeute de m'aider à structurer tout ça. Merci aussi à Philippe JAENADA, anacrim qui s'ignore, de m'avoir expliqué comment il avait écrit *La Serpe*.

J'ai une pensée pour ceux qui ont contribué à faire de moi la personne que je suis mais qui ne sont plus là pour le voir : ma merveilleuse et regrettée grand-mère Solange GIANOLA, qui m'a tellement encouragée dans ma curiosité, et mon grand-père le commissaire divisionnaire Roger GIANOLA, qui naturellement n'aimait pas beaucoup les gendarmes, jusqu'au début de ma thèse. Une pensée aussi pour mes grands-parents Solange et Georges CHARRON, pour qui ce que je fais dans la vie est un peu obscur mais sans lesquels je n'aurais été qu'une petite fille de la ville comme les autres. Avec eux, je remercie mes parents Catherine et Marc GIANOLA qui m'ont fait suffisamment confiance pour toujours me laisser libre de choisir ma voie, qui m'ont soutenue en particulier quand l'horizon de mes études et de l'avenir n'était pas très clair, et qui ne sourcillent pas lorsque je leur annonce mes projets. Merci à mon frère Guillaume et à ma sœur Emma d'avoir fait bloc dans l'imprévu des derniers mois.

Le genre « Remerciements » s'achève souvent sur la mention de la patience et de l'écoute prodiguées par l'être aimé. Mais pour toi, Swen RIBEIRO, mon collègue, ami et compagnon, il serait insuffisant et ingrat de m'en tenir à de telles platitudes lorsque ton appui s'est étendu aussi à un inestimable soutien technique, scientifique et logistique. Merci de me connaître si bien et de me suivre dans mes lubies, qu'il s'agisse de quitter un CDI pour commencer une thèse, d'apprendre l'estonien ou de repasser le bac à 30 ans. Sans toi, je n'aurais pas pu, et bien entendu, les mots sont faibles.

*Sic parvis magna*



« Surtout, il lit le dossier, tout le dossier. [...] Il sait presque tout de son intimité, rien n'est plus impudique, obscène même, qu'un dossier d'instruction. Il y a les témoignages de ses nombreux amis, ceux de sa famille, toute l'enquête qui a radiographié sa vie, mais il y a aussi ses mots à elle. Ceux qu'elle couchait frénétiquement sur les pages de son journal intime, de son écriture ronde, presque enfantine, l'écriture d'une jeune femme amoureuse et triste, qui abdiquait sa fierté, sa raison, son indépendance devant lui, l'amant, et qui se reprochait à la fois d'être trop soumise et de ne pas l'être assez. »

Pascale ROBERT-DIARD, *La Déposition*



# Avertissement

Ce manuscrit présente des extraits authentiques de documents issus de dossiers de procédures judiciaires criminelles. Nous citons ces extraits soit par capture d'écran pour illustrer la topologie du document, soit par copier/coller.

Afin de respecter la vie privée des personnes impliquées et le secret de l'instruction, deux stratégies d'anonymisation ont été adoptées. Lorsque des captures d'écran sont insérées comme images, les noms, lieux, dates, éléments portant atteinte à la confidentialité des faits sont camouflés ou floutés. Dans le cas où le texte est intégré par copier/coller, les éléments portant atteinte à la confidentialité sont remplacés par d'autres éléments similaires : un prénom par un autre prénom, un nom par un autre nom, une ville par une autre ville, etc.

Ce travail d'anonymisation a été effectué manuellement uniquement pour les besoins de ce manuscrit.



# Table des matières

<b>Résumé</b>	<b>i</b>
<b>Abstract</b>	<b>iii</b>
<b>Remerciements</b>	<b>v</b>
<b>Avertissement</b>	<b>xi</b>
<b>Table des matières</b>	<b>xiii</b>
<b>Introduction</b>	<b>xvii</b>
<b>I Enquête et analyse criminelle</b>	<b>1</b>
1 Police judiciaire et enquête . . . . .	1
2 Principe de l'analyse criminelle . . . . .	5
3 Les entités en analyse criminelle . . . . .	7
3.1 Le cadre des entités . . . . .	8
3.2 Types d'entités . . . . .	8
3.3 Relations entre les entités . . . . .	9
4 Les schémas relationnels et événementiels . . . . .	10
5 L'analyse criminelle à la Gendarmerie Nationale . . . . .	14
5.1 Les types d'enquête . . . . .	15
5.2 Cadre juridique . . . . .	16
5.3 Formation des analystes criminels . . . . .	17
5.4 Pratique de l'analyse criminelle à la Gendarmerie Nationale . .	18
6 Enjeux, contraintes et difficultés . . . . .	19
6.1 Sur la méthode . . . . .	19
6.2 Sur la matière . . . . .	20

6.3	Sur l'organisation documentaire . . . . .	21
6.4	Sur le texte et le contenu . . . . .	22
7	Conclusion : que proposer à l'analyse criminelle? . . . . .	23
<b>II De la procédure judiciaire criminelle au corpus textuel</b>		<b>25</b>
1	Qu'est-ce qu'une procédure judiciaire? . . . . .	26
2	Les pièces de procédure . . . . .	27
2.1	Documents réglementaires et de procédure . . . . .	28
2.2	Documents d'information . . . . .	29
2.3	Synthèse . . . . .	44
3	Le texte de la procédure judiciaire . . . . .	45
3.1	Le concept de corpus selon la linguistique de corpus . . . . .	45
3.2	Corpus réflexifs, corpus hétérogènes . . . . .	47
3.3	Et la procédure? . . . . .	47
3.4	Constitution du corpus de recherche . . . . .	49
4	Conclusion . . . . .	51
<b>III État de l'art</b>		<b>53</b>
1	Le texte, l'ordinateur, l'humain . . . . .	54
1.1	Humanités numériques . . . . .	55
1.2	Linguistique de corpus & textométrie . . . . .	57
1.3	Quel paradigme textuel pour l'analyse criminelle? . . . . .	62
2	« Entités », « entités nommées » et « descriptions définies » . . . . .	63
2.1	En analyse criminelle . . . . .	63
2.2	En traitement automatique des langues . . . . .	64
2.3	Détection d'entités dans des textes liés au domaine criminel . . . . .	71
2.4	Lier entités criminelles et entités nommées . . . . .	72
3	Synthèse . . . . .	74
4	Conclusion . . . . .	78
<b>IV Les entités en analyse criminelle</b>		<b>79</b>
1	Catégories et réalisations linguistiques correspondantes . . . . .	80
1.1	Description des entités . . . . .	81

1.2	Synthèse : formes des entités dans les auditions et facteurs d'influence . . . . .	101
2	Détection d'entités nommées criminelles . . . . .	103
2.1	Conception . . . . .	104
2.2	Mise en œuvre . . . . .	111
2.3	Résultats et discussion . . . . .	117
3	Synthèse . . . . .	120
4	Conclusion . . . . .	123
<b>V</b>	<b>Le texte de la procédure judiciaire : l'exemple des auditions</b>	<b>125</b>
1	L'audition d'enquête : pratique et productions . . . . .	126
1.1	Entendre, interroger, rapporter . . . . .	126
1.2	Le discours de l'audition et le discours du procès-verbal . . . . .	128
1.3	Le procès-verbal d'audition : régularité et narrativité . . . . .	131
1.4	Conclusion . . . . .	140
2	Étude textométrique pour une catégorisation en genres . . . . .	140
2.1	<i>LegalNLP</i> , traitement automatique des langues et genre textuel	140
2.2	Corpus . . . . .	145
2.3	L'analyse factorielle des correspondances . . . . .	146
2.4	Analyse . . . . .	148
2.5	Critique de l'étiquetage en parties du discours . . . . .	153
2.6	Conclusion . . . . .	154
3	Conclusion . . . . .	156
	<b>Conclusion générale</b>	<b>159</b>
	<b>A Guide pour l'annotation manuelle des entités</b>	<b>165</b>
	<b>B Graphes Unitex</b>	<b>171</b>
	<b>C Tableau de contingence des parties du discours</b>	<b>179</b>
	<b>Glossaire</b>	<b>181</b>
	<b>Table des figures</b>	<b>189</b>

<b>Liste des tableaux</b>	<b>193</b>
<b>Bibliographie</b>	<b>195</b>

# Introduction

La fiction policière est un objet culturel si populaire que tout un chacun se fait une idée de ce en quoi consiste une enquête policière. Autrefois concentrée sur les personnages de grands enquêteurs et détectives tels que Sherlock Holmes, le commissaire Maigret ou Hercule Poirot, la production culturelle sur le sujet comprend aussi aujourd'hui de nouveaux formats consacrés aux récits d'affaires réelles, comme les séries documentaires et les podcasts de *true crime*. L'omniprésence dans le paysage culturel de narrations policières suscite chez le public des mythes comme celui que la résolution d'une affaire criminelle se fait nécessairement à l'aide de méthodes et de techniques de laboratoire analysant des traces laissées sur la scène de crime, approches garantissant de confondre le coupable presque à tous les coups et sans aucun doute sur son implication. Dans la réalité, une enquête criminelle engage de nombreux acteurs professionnels sur un laps de temps qui s'étale de plusieurs mois à plusieurs années. Du point de vue policier, sa résolution ne se trouve pas dans la désignation d'un suspect mais dans la présentation au procès d'indices forts et concordants permettant d'expliquer les faits considérés. Au-delà de l'analyse des éléments matériels, la collecte et la gestion de l'information au cours de l'enquête constituent un enjeu fondamental qui ne se réduit pas à discuter avec une poignée de témoins ou à un interrogatoire plus ou moins musclé.

Le principe d'une enquête revient à accumuler le plus d'information possible en lien ou possiblement en lien avec les faits depuis différents types de sources : d'une part des éléments matériels, mais aussi des témoins, des télécommunications, des constatations des enquêteurs. On en tirera des éléments pour construire des hypothèses, en réfuter d'autres, mettre une personne en ou hors de cause. Gérer l'information de l'enquête est un enjeu stratégique tout aussi important que les enjeux liés à l'exploitation des traces matérielles.

En France, la procédure judiciaire et l'enquête criminelle passent par l'écrit : les

différentes étapes des investigations et du processus judiciaire donnent lieu à la production de pièces qui documentent les actes réalisés, et qui une fois compilés constituent le dossier de procédure. Il est délicat de donner un ordre de grandeur, mais on peut affirmer que chaque procédure contient au moins plusieurs centaines de documents. C'est face à la profusion d'éléments consignés sur papier que la gestion de l'information de l'enquête devient une problématique fondamentale en lien étroit avec la gestion de documents textuels.

Lorsque l'information devient difficile à assimiler pour les enquêteurs, l'intervention d'analystes criminels peut être sollicitée afin de reprendre le dossier de procédure et de l'examiner à la recherche de nouvelles pistes ou d'éléments permettant de consolider une hypothèse. On pourra par exemple chercher à reconstituer l'emploi du temps d'une personne en particulier à l'aide de ses propres déclarations mais également en confrontation avec les témoignages d'autres personnes, ou encore reconstruire la hiérarchie d'un réseau. Telle qu'elle est pratiquée à la Gendarmerie Nationale, cette approche consiste à procéder à un examen approfondi des éléments collectés pour mieux orienter les investigations, et ne peut s'appuyer aujourd'hui sur des outils informatiques dédiés. La recherche de l'information dans le dossier de procédure occupe une partie importante du temps de travail des gendarmes analystes criminels. Parmi les contraintes pesant sur ce travail, on peut citer la numérisation des dossiers, leur organisation interne, la forme de l'information dans les documents (tableaux de données, images, texte libre).

Or ces aspects ne sont pas inconnus des chercheurs en sciences humaines et en informatique depuis longtemps confrontés à des problématiques similaires de gestion et d'exploitation d'information. Il semble donc judicieux d'ouvrir l'analyse criminelle aux disciplines étudiant les textes et leur compréhension assistée par ordinateur afin de repenser ses outils et ses méthodes. La linguistique de corpus et l'extraction automatique d'information entre autres ont développé des approches qu'il faut chercher à réinvestir dans le traitement des procédures judiciaires. Les bénéfices potentiels consistent en une charge de travail mieux répartie pour l'analyste entre repérage de l'information et synthèse, moins de temps et plus de facilité à rechercher l'information.

Dans une telle perspective, la problématique devient celle d'adapter des approches et solutions existantes aux spécificités d'un nouveau domaine. Pour le cas de l'analyse criminelle, nous devons prendre en compte le contexte de sa pratique, les caractéristiques des données traitées, les méthodes et concepts déjà en place. À travers une étude approfondie de ces aspects et en nous appuyant à la fois sur l'expertise des analystes criminels et des exemples de procédures, nous espérons tracer les possibilités et les modalités tendant vers une meilleure gestion des données textuelles de la procédure judiciaire exploitée en analyse criminelle.

Ce travail de thèse réalisé en partenariat avec une équipe d'analyse criminelle de la Gendarmerie Nationale tente donc de concilier les apports potentiels de méthodes de gestion automatique des textes aux objectifs et contraintes de l'analyse criminelle. Trois disciplines se rencontrent dans cette recherche : linguistique, informatique et analyse criminelle. La linguistique et l'informatique ayant déjà l'habitude de dialoguer dans le domaine du traitement automatique des langues, nous soulignerons en particulier les aspects linguistiques et textuels propres à la procédure judiciaire d'une part, et la confrontation des approches de traitement automatique des langues à de nouvelles formes de concepts déjà explorés.

Le manuscrit est organisé en cinq parties. Dans un premier temps, nous présentons le contexte de recherche en décrivant l'insertion de l'analyse criminelle dans l'enquête (en particulier dans le cas de la Gendarmerie Nationale), ses outils, pratiques, méthodes et concepts, notamment celui d'entité. Nous procédons ensuite à la description du dossier de procédure judiciaire, des documents qui le composent et de leurs aspects textuels dans l'objectif de déterminer dans quelle mesure et selon quelles modalités est-il possible de l'envisager comme corpus textuel à la fois afin de l'employer pour développer des approches automatiques que pour l'étude de la langue de la procédure. Suite à la délimitation du champ de recherche et de sa matière, nous établissons ensuite un état de l'art relatif aux disciplines impliquées dans la problématique : nous y abordons les pratiques en humanités numériques, linguistique de corpus, en recherche et extraction d'information et présentons quelques exemples d'application au domaine criminel. L'avant-dernier chapitre est consacré à la mise en relation des concepts de traitement automatique des langues utiles à l'analyse criminelle, à savoir les concept d'entités nommées et d'entités criminelles, et au

test d'une première approche d'extraction automatique. Enfin, le dernier chapitre propose une étude du genre textuel de documents produits au sein de la justice au sens large. Cette étude, replacée dans le sillage d'autres travaux de caractérisation du genre textuel judiciaire, permet la structuration du genre en genres et sous-genres, afin de proposer un cadre aux futurs développements de technologies du langage appliquées au domaine du droit et de la justice.

## Chapitre I

# Enquête et analyse criminelle

### *Le cas de la Gendarmerie Nationale*

Pour situer la recherche menée dans cette thèse, nous commencerons par présenter le cadre dans lequel celle-ci s'inscrit, celui de la pratique de l'[analyse criminelle](#) au sein de la [Gendarmerie nationale](#). Pour cela, nous aborderons les missions des forces policières en nous concentrant sur le cas de la [police judiciaire](#), en particulier l'[enquête](#), ce qui nous permettra d'expliquer l'insertion et les objectifs de l'[analyse criminelle](#). Nous détaillerons ensuite les pratiques de l'[analyse criminelle](#) à la [Gendarmerie nationale](#) et au [PJGN](#) puisqu'il s'agit de notre cas d'étude. Les observations formulées seront basées essentiellement sur les échanges ayant eu lieu avec les personnels [analystes criminels](#) côtoyés au cours de nos travaux. Enfin, nous discuterons des difficultés rencontrées par cette pratique d'[enquête](#).

## 1 Police judiciaire et enquête

En France, les forces de police sont composées de deux organisations : la Police Nationale, dont les membres sont civils, et la [Gendarmerie nationale](#), dont les membres sont militaires. Au-delà de cette différence de statut (et donc d'organisation hiérarchique), leurs missions sont globalement les mêmes, les différences résidant principalement dans la répartition des zones d'action : urbaines pour la police, rurales pour les gendarmes, et dans les habitudes de travail.

Les missions des forces de police comprennent d'une part les actions de sécurité qui veillent au maintien de l'ordre public et relevant de la police administrative, et

la **police judiciaire**, dont l'objectif est défini par les articles 12 et suivants du **Code de procédure pénale**<sup>1</sup>, et plus précisément par l'article 14 :

Elle est chargée, suivant les distinctions établies au présent titre, de constater les infractions à la loi pénale<sup>2</sup>, d'en rassembler les preuves et d'en rechercher les auteurs [...].

LIZUREY (2006, p. 168) estime que la mission de **police judiciaire** représente 60%<sup>3</sup> de l'activité opérationnelle de la Gendarmerie nationale, et que celle-ci constate près de 30% de la délinquance totale.

Cette mission de **police judiciaire** est liée intrinsèquement à la pratique de l'**enquête**. Par **enquête**, nous entendons ici le principe général de recherche de preuves dans l'objectif d'apporter la lumière sur des faits répréhensibles, car nous verrons dans la suite de ce chapitre que formellement, plusieurs types d'**enquête** sont prévus par le droit français.

Si le principe de l'**enquête** criminelle est généralement connu étant donné sa popularité dans les œuvres de fiction, il mérite toutefois d'être défini. Nous reprenons pour cela la proposition de ROSSY et al. (2019, p. 429) :

L'enquête judiciaire consiste d'abord à recueillir et à traiter des informations afin de détecter les délits, d'établir les faits, de qualifier les infractions, de trouver les auteurs (les identifier et les localiser) et de les présenter à l'intention d'un tribunal qui décidera au moyen de ces preuves.

Les auteurs structurent le déroulement de l'**enquête** en trois temps :

- Détection : phase d'identification et de localisation qui mène à l'arrestation,
- Structuration : phase qui établit les faits et les qualifie, menant à la mise en accusation,
- Évaluation : phase d'évaluation et de jugement.

Les phases de détection et de structuration sont relatives à l'investigation policière tandis que celle de jugement dépend du tribunal.

---

1. Légifrance : <https://www.legifrance.gouv.fr/affichCode.do?idArticle=LEGIARTI000006574849&idSectionTA=LEGISCTA000006167411&cidTexte=LEGITEXT000006071154&dateTexte=20191212> [consulté le 12 décembre 2019].

2. Le droit pénal régit les rapports entre les individus et la société, et le droit civil régit les rapports entre individus.

3. Un chiffre toutefois réduit à 40% sur le site de la **Gendarmerie nationale** : <https://www.gendarmerie.interieur.gouv.fr/Notre-institution/Generalites/Nos-missions/Police-judiciaire/Police-Judiciaire> [consulté le 13 décembre 2019].

Pour collecter les éléments nécessaires à l'enquête, les enquêteurs disposent d'un éventail de moyens d'action encadrés par la loi qui exploitent deux types de sources : les sources matérielles et les sources humaines. Dans les sources matérielles figurent les éléments physiques, qu'on appelle des traces : empreintes digitales, traces biologiques comme l'ADN, objets, vêtements, armes, matériel informatique, etc. MARGOT (2014) définit la trace ainsi :

Marque, signal ou objet, la trace est un signe apparent (pas toujours visible à l'œil nu). Elle est le vestige d'une présence et/ou d'une action à l'endroit de cette dernière. [...] Plusieurs éléments essentiels ressortent de cette définition : 1) elle est matérielle, elle existe indépendamment de toute signification; 2) elle nous vient du passé, un passé que l'on ne saurait faire revivre; 3) elle est incomplète, imparfaite (vestige); 4) elle n'appartient pas à l'environnement habituel de l'endroit où elle se trouve (elle est l'effet d'une activité en un endroit, à un moment); 5) elle contient une information (signe) sur sa source et finalement 6) sur l'action qui l'a produite.

Ces éléments matériels, qui sont le résultat de l'activité délictuelle ou criminelle, sont collectés et placés sous scellé afin de garantir leur préservation en l'état. Leur exploitation peut demander l'intervention de spécialistes de divers domaines qui sont capables d'en tirer de l'information. On recourra par exemple aux soins d'un informaticien pour extraire le contenu d'un téléphone ou d'un ordinateur. De plus, l'expansion des technologies numériques dans la société confronte les services de police à une nouvelle forme de trace, les traces numériques, qui représentent un défi pour leur exploitation comme les télécommunications, les réseaux sociaux, les localisations GPS, etc.

Les sources humaines sont les témoins entendus dans le cadre de l'enquête. Leur contribution à la progression des investigations consiste à confier les informations qu'ils détiennent et qui pourraient être utiles. Les informations sont fixées dans des procès-verbaux d'audition, sur lesquels nous reviendrons en détails.

D'après les traces et les informations collectées, les enquêteurs orientent la suite

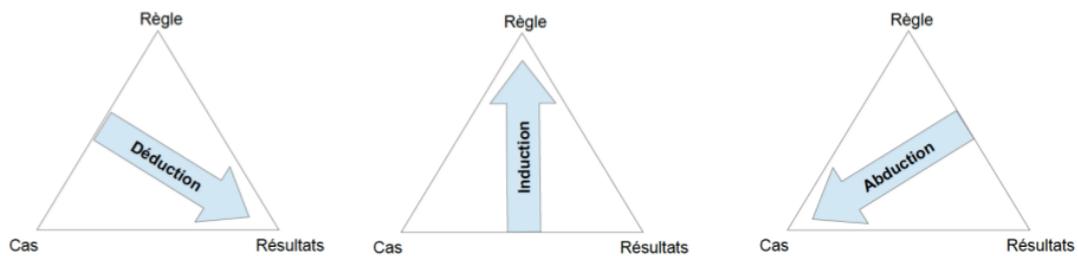


FIGURE 1 – Modélisation des formes de raisonnement élémentaires  
issue de BARLATIER (2017, p.15)

des investigations et construisent des hypothèses dans une logique hypothético-déductive (ROSSY et al., 2019, p. 432) basée sur trois formes élémentaires de raisonnement : l'abduction, la déduction et l'induction (RIBAUX (2014, p. 167) inspiré des travaux du sémiologue Charles Sanders Peirce), qui agencent triangulairement les connaissances (« Règle » sur la figure 1, c'est-à-dire les renseignements, la connaissance des situation criminelles), l'activité (les faits considérés), et les **traces**.

Dans l'abduction, on part des connaissances et des **traces** pour développer une explication de l'activité.

Dans la déduction, on part d'une activité pour déduire les effets qui en découlent dans une démarche expérimentale.

Dans l'induction, on part d'une activité et de ses **traces** pour dégager des connaissances.

Ces trois modes d'inférence sont employés alternativement dans l'**enquête**, afin de déterminer, concernant les faits, qui en est la cause, quand, comment et pourquoi ont-ils eu lieu.

Qu'elle qu'en soit la source, matérielle ou humaine, la collecte et la gestion de l'information est l'étape préalable aux trois modes de raisonnement. Or, l'information impliquée dans une affaire criminelle peut rapidement dépasser les capacités cognitives, par son volume, sa variété et la diversité de ses formes. Il est alors nécessaire de l'organiser, pour une meilleure exploitation. Pour répondre à cet objectif, on peut employer les méthodes de l'**analyse criminelle**, que nous présentons dans la section suivante.

## 2 Principe de l'analyse criminelle

Le principe général de l'analyse criminelle est de structurer l'information récoltée à propos de faits criminels et délictueux pour en faciliter la compréhension. On peut l'appliquer dans un objectif de police judiciaire et d'enquête, on parle alors d'analyse criminelle opérationnelle, ou bien dans un objectif d'action de sécurité en cherchant les répétitions de faits (les *patterns* : par exemple une série de cambriolages) afin d'anticiper leur développement, il s'agit alors d'analyse criminelle stratégique (CUSSON et CORDEAU, 1994 ; RIBAU, 2014, p. xxvi). Nos propos ne portant que sur l'analyse criminelle opérationnelle, nous y ferons désormais référence comme l'analyse criminelle.

ROSSY (2011, p. 7) en fait la description suivante :

L'analyse criminelle opérationnelle décrit un processus général qui, par une exploitation structurée, aide à gérer les information accessibles afin d'en maîtriser les flux et de garder la vue d'ensemble, cherche à structurer les raisonnements et en particulier à en minimiser les biais et, finalement, met en perspective les informations pour les présenter aux décideurs. Le caractère systématique de l'analyse criminelle opérationnelle réside dans sa démarche, fondée sur un ensemble organisé de méthodes et de principes qui relèvent de la gestion de l'information.

Dans le cas de l'application de l'analyse criminelle à une enquête, elle consiste à identifier l'information au sein de la procédure, à l'organiser, et à la représenter visuellement pour en faciliter la compréhension en fournissant une vue globale du contenu du dossier de procédure ou des faits examinés. L'objectif, par l'extraction et la mise en valeur de l'information essentielle et pertinente de la masse du dossier, est de suggérer des investigations et des pistes potentielles aux enquêteurs. L'analyse criminelle ne cherche pas nécessairement à résoudre les faits, elle creuse les éléments et apporte un regard neuf sur le dossier à un moment donné de l'enquête.

L'extraction de l'information modélise celle-ci sous le concept d'entité, une entité pouvant être à peu près n'importe quel objet ou paramètre du monde réel évoqué dans les faits. La visualisation de l'information prend la forme de schémas de synthèse qui agencent les entités entre elles selon plusieurs paradigmes en fonction des

objectifs visés en particulier.

L'enquête et l'analyse criminelle étant des processus de renseignement, il paraît utile de cerner la production et la circulation de l'information dans ce cadre. Pour cela nous reprenons le modèle « DIKW » (*data*, les données, *information*, l'information, *knowledge*, la connaissance, et *wisdom*, la sagesse) synthétisé par ROWLEY (2007). Dans ce modèle, les données sont des matériaux bruts, non organisés et non traités (*unprocessed*) (ERMINE et al., 2012). L'information est constituée de données organisées, traitées dans un but particulier. Le savoir enfin émanerait de la synthèse d'informations issue de diverses sources, de l'étude et de l'expérience. D'après ces éléments, nous proposons le flux suivant :

- Les témoins et experts produisent des données sous forme de témoignages et de rapports,
- Ces données collectées et ordonnées par les enquêteurs deviennent des informations,
- Les informations synthétisées et représentées par les analystes criminels deviennent du savoir,
- Le savoir est transmis au cours du procès.

Cette configuration correspond à la description établie par RIBAUX (2014, p. 369) :

Dans la conception de l'analyste criminel, les données recueillies constituent le matériel brut. Il s'agit de les façonner progressivement pour produire de l'information utile à la prise de décisions dans une enquête par une succession d'étapes.

En ce qui concerne la sagesse, au sujet de laquelle ROWLEY, 2007 souligne le manque de consensus, on pourrait la recouper avec celle de l'expérience professionnelle : chaque cas traité par les enquêteurs et les analystes nourrit leur expérience, qu'ils pourront mettre à profit dans le traitement d'affaires futures. Cette conception rejoint celle de JESSUP et VALACICH (2002) cités par J. ROWLEY :

JESSUP and VALACICH see wisdom as accumulated knowledge, which allows you to understand how to apply concepts from one domain to

new situations or problems.<sup>4</sup>

Néanmoins, la frontière entre *data* et *knowledge* est sujette à controverse. J. ROWLEY rapporte plusieurs travaux selon lesquels c'est le récepteur humain qui détermine s'il s'agit de *data* ou de *knowledge*. Dans ce cas, on pourrait proposer, en plus du flux ci-dessus centré sur le point de vue des enquêteurs, un flux complémentaire centré sur le point de vue de l'**analyste criminel** :

- Les enquêteurs produisent les données sous la forme des actes de **procédure**,
- Les **analystes criminels** extraient l'information de la masse des données,
- Les **analystes criminels** synthétisent l'information qui devient du savoir,
- Le savoir sur les faits est transmis au procès.

Ces préoccupations relevant plus de la problématique de la gestion des connaissances au cours de l'**enquête** que de notre étude proprement dit, nous ne pousserons pas la réflexion au-delà dans ce manuscrit. Il pourra nous arriver par la suite d'employer indifféremment *données* et *information(s)* pour évoquer la matière traitée en **analyse criminelle**.

### 3 Les entités en analyse criminelle

L'**analyse criminelle** se concentre sur le traitement des informations rapportées par les acteurs de l'**enquête** et consignées dans le dossier de **procédure** plutôt que sur la découverte d'indices matériels. Il est impossible de prévoir dans une **enquête** quel critère, quel objet sera celui portant une information cruciale. Néanmoins, les grandes lignes des faits examinés peuvent être dégagés grâce aux entités, qui servent de support à l'analyse. Par « entité », l'**analyse criminelle** entend toute référence à un objet matériel ou conceptuel de la vie réelle dont il est fait mention dans la **procédure**. Le chapitre **IV** est consacré à la description conceptuelle et linguistique détaillée de ces entités et à leur comparaison avec les « **entités nommées** » que l'on traite en **TAL**.

En pratique, n'importe quel objet ou sujet peut constituer une entité telle que l'entend l'**analyse criminelle** : personnes, entreprises, dates, bâtiments, lieux, groupes ou associations, adresses, moyens de locomotion... Certaines se retrouvent

---

4. « JESSUP et VALACICH voient la sagesse comme un savoir accumulé, qui permet de comprendre comment appliquer des concepts d'un domaine à de nouvelles situations ou problèmes. », traduction par nos soins assistée du logiciel de traduction automatique en ligne DeepL : <https://www.deepl.com/translator> [consulté le 13 décembre 2019].

dans toutes les affaires traitées, on les considère donc comme fondamentales. Elles constituent le socle commun de l'analyse criminelle.

### 3.1 Le cadre des entités

L'analyse criminelle, et l'enquête en général, repose sur un cadrage des investigations dont le premier niveau est la scène de crime. Cette scène, délimitée par les enquêteurs lors de leur arrivée sur place<sup>5</sup>, alimente l'enquête de ses premiers éléments, c'est-à-dire de ses premières entités : la date, le lieu, la victime, les objets. Une fois les entités du cadre exploité, celui-ci est agrandi à la recherche de plus d'information, incluant de plus en plus d'entités.

Cette notion de cadre de l'enquête est présentée par RIBAUX (2014, p. 373) reprenant la notion de « cercle des entités » (*frame*) issue de KIND (1987), qui permet de circonscrire l'enquête en trois dimensions : les lieux, le temps et les entités. Pour donner une illustration concrète, O. RIBAUX utilise p. 199 et p. 390 le cas du viol et du meurtre de Caroline Dickinson dans une auberge de jeunesse en 1996<sup>6</sup>. Dans cette affaire, les investigations ont été concentrées localement au premier abord, avant d'être élargies au territoire national à la recherche d'autres faits similaires dans des auberges de jeunesse. Il s'agit de l'une des premières affaires françaises où les analyses ADN ont joué un rôle déterminant : d'abord réalisés sur les hommes séjournant dans l'auberge, les prélèvements ont ensuite été effectués sur les hommes de la commune, puis sur ceux de la région, élargissant ainsi le cadre des entités.

### 3.2 Types d'entités

**Personnes** Qu'il s'agisse de la victime, ses proches, la ou les personnes suspectées des faits, ou encore les témoins, de nombreuses personnes concernées de près ou de loin par les faits sont évoquées dans les dossiers de procédure. Il faut retrouver leur position spatio-temporelle au moment des faits, mais aussi leurs liens avec la victime ou les autres personnes mentionnées afin de comprendre leur implication.

---

5. Dans le cas des enquêtes menées par la gendarmerie, cette délimitation est rapportée dans la procédure par le procès-verbal de transport constatations mesures prises que nous présentons à la section 2.2.1 du chapitre II.

6. Pour un résumé de l'affaire, consulter Caroline Dickinson : la preuve par l'ADN, *Le Monde*, 17 août 2006 et l'émission radiophonique Caroline Dickinson, une enquête historique, *Affaires sensibles*, France Inter, 17 août 2016.

Les témoignages constituent la source principale des informations relatives aux faits et gestes des personnes.

**Lieux & Dates** Dans cette perspective de localisation spatio-temporelle des acteurs, les lieux et les dates sont bien évidemment d'une importance majeure. On peut les extraire de différents documents, comme les bornages téléphoniques par exemple. Lorsqu'on les tire des **auditions** de **témoins**, leur mention peut poser des problèmes d'ambiguïté, comme nous le verrons en détail dans le chapitre IV.

**Numéros de téléphone** Les numéros de téléphone sont aussi une source d'information, notamment dans les affaires impliquant des réseaux comme le trafic de stupéfiants. L'analyse des communications téléphoniques peut permettre dans ce type de cas d'en reconstituer la structure hiérarchique. À cette fin, les factures détaillées de téléphonie obtenues auprès des opérateurs constituent la source essentielle de l'information, mais il est également utile de repérer les numéros de téléphone dans les **auditions** de **témoin** car ils sont souvent associés au nom du titulaire de la ligne, ce qui n'est pas le cas dans les factures.

**Moyens de locomotion** Le dernier type d'entité fondamentale recouvre les moyens de locomotion, et plus particulièrement les véhicules. Les véhicules, étant associés à une autre entité comme une personne ou une entreprise, peuvent permettre de remonter vers un **témoin** ou une personne suspectée. La particularité des véhicules motorisés par rapport aux autres entités est d'être référencés par un numéro d'immatriculation qui, s'il est connu, permet son identification formelle.

### 3.3 Relations entre les entités

L'intérêt du repérage des entités n'est pas simplement de les extraire et de les lister : il faut mettre en lumière la nature des rapports qu'elles entretiennent entre elles. Les relations entre entités sont conceptualisées en deux types : propres à un état, ou propres à un acte. Les relations d'état sont par exemple les liens familiaux (X est la mère de Y, Y est le mari de Z...) ou sociaux (propriétaire, gérant, employé,

titulaire...). Les relations d'acte représentent des actions ou événements : X vend Y, Z tue W, etc.

Une fois le repérage et la compréhension des liens entre entités réalisés, l'[analyste criminel](#) acquiert une connaissance globale du dossier. Les entités et leurs liens sont insérés dans une base de données. Après le temps de lecture, le temps de saisie manuelle des informations est l'autre étape chronophage du processus. La base de données est intégrée à un logiciel de représentation graphique et l'étape suivante de l'[analyse criminelle](#) commence : la représentation schématique des informations.

## 4 Les schémas relationnels et événementiels

ROSSY (2011, p.42) distingue six types de schémas exploités en [analyse criminelle](#) : schémas relationnels, schémas de flux, schémas d'activité, schémas d'événements, schémas chronologiques, et schémas de flux temporels. La nature des dossiers habituellement confiés au [analystes criminels](#) de la [Gendarmerie nationale](#) fait qu'ils ont essentiellement recours à deux types de schémas : les schémas événementiels et les schémas relationnels.

Les schémas événementiels (figure 2) représentent des séquences d'événements positionnés dans le temps. Ils permettent par exemple de comparer l'emploi du temps de la victime avec celui d'une personne suspectée pour voir si leurs chemins se sont croisés, et sont également utiles pour relever les incohérences entre différents emplois du temps ou témoignages.

Les schémas relationnels (figure 3) représentent des réseaux de relations. Ils sont particulièrement utiles dans les affaires de bandes organisées avec des structures hiérarchiques, comme le trafic de stupéfiant ou pour l'analyse de contacts téléphoniques.

Selon l'orientation de l'analyse définie, l'[analyste criminel](#) réalise un ou plusieurs schémas. Cependant, ces schémas ne sont pas la finalité ultime de l'[analyse criminelle](#), mais un support graphique de l'analyse complété par un [procès-verbal](#) de synthèse, dans lequel l'[analyste criminel](#) livre le raisonnement qu'il a construit pour expliciter les faits et propose plusieurs hypothèses de travail : de nouvelles pistes pour la poursuite des investigations.

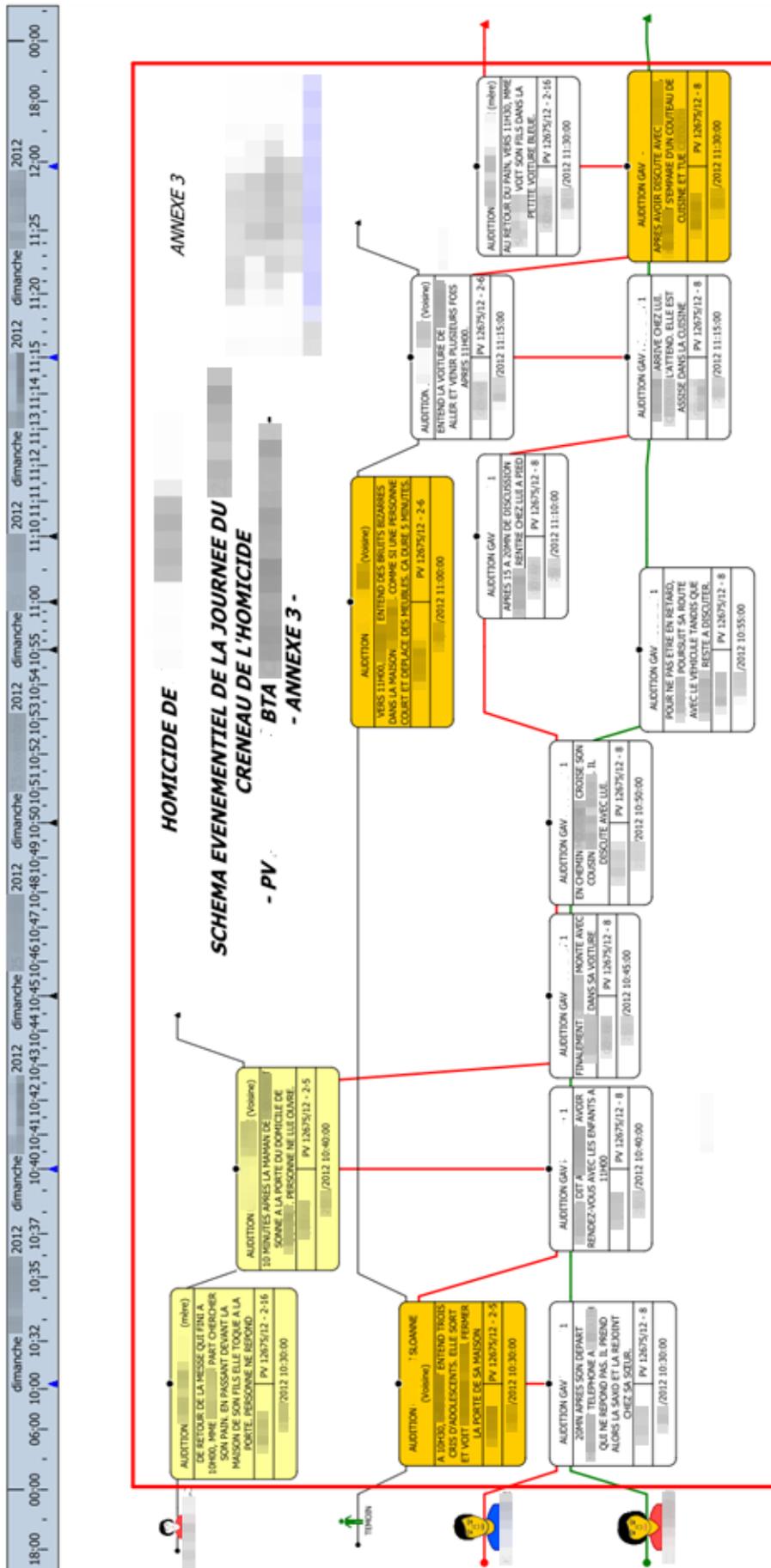


FIGURE 2 – Exemple de schéma événementiel confrontant plusieurs témoignages et les faits.

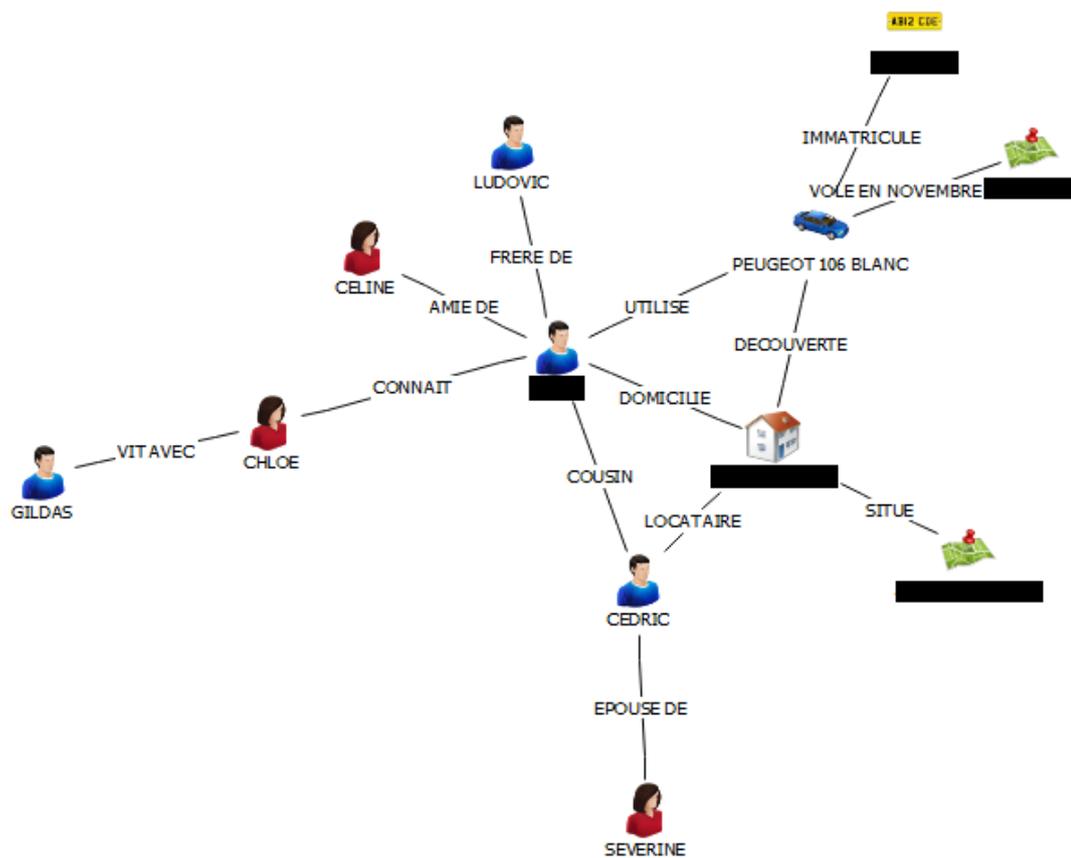


FIGURE 3 – Exemple de schéma relationnel simple

Pour la réalisation des schémas, l'outil principal est un logiciel de représentation graphique et d'analyse de réseaux développé par la société I2 puis racheté par IBM intitulé *Analyst's Notebook*<sup>7</sup>. Selon le site commercial de son constructeur, ce logiciel « fournit une analyse visuelle multidimensionnelle qui permet aux analystes de détecter rapidement les connexions et tendances cachées dans les données ». En pratique, les *analystes criminels* construisent une base de données avec l'information collectée dans la *procédure* qui est ensuite injectée dans *Analyst's Notebook*, et peuvent ensuite élaborer des schémas utilisant ces informations. L'interface propose un panel d'icônes et de structures permettant de représenter des nœuds, arêtes et cadres : les nœuds correspondent aux entités, les arêtes aux liens, et les cadres mettent en valeur des ensembles d'entités. L'intérêt de la conception graphique proposée par *Analyst's Notebook*, en plus de donner une visualisation différente des faits qui permet de les « saisir en un coup d'œil », est de révéler des proximités entre entités qui peuvent échapper à la lecture du texte : par exemple, repérer si deux entités sont connectées par l'intermédiaire d'une autre.

---

7. Lors de l'interpellation de suspects dans le cadre de l'affaire Grégory Villemin en juin 2017, l'usage d'un logiciel appelé « Anacrim » a été mis en avant et largement repris par les médias. Nous tenons à rectifier cette inexactitude : il n'existe pas de logiciel « Anacrim ». Le terme « anacrim » est un diminutif utilisé pour désigner la méthode de travail de l'*analyse criminelle*. Quant au logiciel, il s'agit d'*Analyst's Notebook*. Pour plus de détails, consulter : « L'intelligence artificielle s'insinue dans l'analyse criminelle », *Le Monde*, 17 juillet 2017, [https://www.lemonde.fr/sciences/article/2017/07/17/1-intelligence-artificielle-s-insinue-dans-l-analyse-criminelle\\_5161548\\_1650684.html](https://www.lemonde.fr/sciences/article/2017/07/17/1-intelligence-artificielle-s-insinue-dans-l-analyse-criminelle_5161548_1650684.html) [consulté le 17 avril 2019], et « Expliquez-nous... Les logiciels d'analyses criminelles Anacrim et Salvac », *Franceinfo*, 30 mars 2018, [https://www.francetvinfo.fr/replay-radio/expliquez-nous/expliquez-nous-les-logiciels-d-analyses-criminelles-anacrim-et-salvac\\_2658774.html](https://www.francetvinfo.fr/replay-radio/expliquez-nous/expliquez-nous-les-logiciels-d-analyses-criminelles-anacrim-et-salvac_2658774.html) [consulté le 16 mai 2019].

## 5 L'analyse criminelle à la Gendarmerie Nationale

Selon les chiffres du ministère de l'Intérieur rapportés par MUCCHIELLI (2006), en 2000, 79% des homicides enregistrés, 77% des tentatives d'homicides et 82% des coups mortels sont élucidées par les services de [police judiciaire](#) de la police et de la gendarmerie. Dans les cas qui ne sont pas résolus aussi aisément, la [Gendarmerie nationale](#) peut faire appel à ses [analystes criminels](#).

Fondée dans les années 60 aux États-Unis, puis développée dans les années 80 au Canada, au Royaume-Uni, aux Pays-Bas et en Belgique, l'[analyse criminelle](#) à été adoptée par la [Gendarmerie nationale](#) au milieu des années 90 (MARIN et al., 2003). Actuellement on compte en France environ trois cents [analystes criminels](#) affectés sur l'ensemble du territoire en unités opérationnelles, essentiellement dans les unités de recherches locales (Sections de Recherches, Sections d'Appui Judiciaire, Brigades Départementales de Renseignement et d'Information Judiciaire<sup>8</sup>...), mais aussi au sein des services centraux (Bureau de la lutte anti-terroriste, gendarmeries spécialisées).

Nos recherches ont été accueillies au sein du Département Sciences de l'Analyse Criminelle ([DSAC](#)), du Service Central de Renseignement Criminel ([SCRC](#)) intégré au Pôle Judiciaire de la Gendarmerie Nationale ([PJGN](#)). Le [PJGN](#) est un centre de [police judiciaire](#) relevant de la [Gendarmerie nationale](#), composé principalement de l'Institut de Recherche Criminelle de la Gendarmerie Nationale ([IRCGN](#)) dont l'expertise concerne les aspects police scientifique, et du Service central de renseignement criminel ([SCRC](#)), qui gère les aspects renseignements.

Le [DSAC](#) est composé d'une dizaine d'[analystes criminels](#) et dispose d'une compétence nationale qui lui permet de traiter des dossiers issus de l'ensemble du territoire. Le département collabore également à l'international notamment avec l'agence [Europol](#), les services de la police belge, et l'École des Sciences Criminelles de l'Université de Lausanne.

Les dossiers pris en charge sont principalement des dossiers criminels d'atteinte aux personnes (homicides, disparitions) ou des dossiers de criminalité organisée. Dans le cas du trafic de stupéfiant par exemple, l'[analyse criminelle](#) est employée

---

8. Pour une présentation détaillée de la structure de la [Gendarmerie nationale](#), nous renvoyons à LIZUREY (2006)

pour comprendre la structure du réseau pour établir les responsabilités ou pour explorer les réseaux financiers en cas de blanchiment d'argent. Enfin, les dossiers peuvent être aussi bien actuels que plus anciens : ce qu'on appelle (abusivement) des *cold cases*.

Telle que la pratique a lieu à la [Gendarmerie nationale](#), la mission de la personne en charge de l'[analyse criminelle](#), appelé [analyste criminel](#), est de procéder à une lecture très minutieuse de l'ensemble de la [procédure](#) afin de reconstituer le déroulement des infractions, le positionnement spatio-temporel des protagonistes, l'emploi du temps d'un individu, ou encore de mettre en lumière les contradictions entre différents témoignages ou la réalité des faits.

### 5.1 Les types d'enquête

En France, trois cadres d'[enquête](#) sont prévus par la loi : l'[enquête préliminaire](#), l'[enquête de flagrance](#), et l'[information judiciaire](#). L'[analyse criminelle](#) peut se situer dans les trois cadres.

L'[enquête préliminaire](#)<sup>9</sup> se fait soit à l'initiative des services de [police judiciaire](#), soit à la demande du [procureur de la République](#). Comme son nom l'indique, elle précède l'ouverture éventuelle d'une [information judiciaire](#).

L'[enquête de flagrance](#)<sup>10</sup> concerne des faits pris en [flagrant-délit](#)<sup>11</sup> punis d'une peine d'emprisonnement. Cette phase d'une durée de huit jours (prolongeable dans certains cas) cherche à profiter du caractère récent des faits dans l'objectif d'une résolution rapide : les [témoins](#) se trouvent à proximité, leurs souvenirs sont frais, il est encore possible de collecter des preuves, etc. À ce stade, les enquêteurs possèdent tous les droits de coercition : ils réalisent les [actes d'enquête](#) puis en rendent compte au [procureur de la République](#), ce qui leur confère une plus grande liberté d'action (SÉNAT, 2009).

---

9. Définie par les articles 75 et suivants du [Code de procédure pénale](#) : <https://www.legifrance.gouv.fr/affichCode.do?idSectionTA=LEGISCTA000006151877&cidTexte=LEGITEXT000006071154> [consulté le 12 décembre 2019].

10. Définie par les articles 53 et suivants du [Code de procédure pénale](#) [https://www.legifrance.gouv.fr/affichCode.do;jsessionid=6EF043DDDC7C076952B65C61468D03C.tplgfr42s\\_2?idSectionTA=LEGISCTA000006151876&cidTexte=LEGITEXT000006071154&dateTexte=20190410](https://www.legifrance.gouv.fr/affichCode.do;jsessionid=6EF043DDDC7C076952B65C61468D03C.tplgfr42s_2?idSectionTA=LEGISCTA000006151876&cidTexte=LEGITEXT000006071154&dateTexte=20190410) [consulté le 11 décembre 2019]

11. Ou crime flagrant.

Si à l'issue de l'**enquête de flagrance**, les faits ne sont pas élucidés et nécessitent la poursuite des investigations, l'**enquête** est placée sous la direction d'un **juge d'instruction** qui ouvre une **information judiciaire**. Le **juge d'instruction** peut procéder lui-même à tous les **actes d'enquête**, mais il peut aussi déléguer ses pouvoirs aux services de **police judiciaire** par le biais d'une **commission rogatoire**<sup>12</sup>. Les **actes d'enquête** sont alors ordonnés par le **juge d'instruction**. La dynamique de prise d'initiative est inversée par rapport à l'**enquête de flagrance**.

Dans le cas de l'**enquête de flagrance**, l'**analyste criminel** peut être présent sur les lieux et son travail se fait alors au fil de la réalisation des **actes d'enquête**. Dans le cas de la **commission rogatoire**, le travail d'analyse est plutôt basé sur les documents déjà produits par les investigations réalisées mais de nouveaux éléments fournis par les enquêteurs peuvent également venir les compléter et les **analystes criminels** ont la possibilité de contribuer eux-mêmes aux investigations aux côtés des enquêteurs et de réaliser des **actes d'enquête** (par exemple : mener une **audition**, ou parcourir le chemin emprunté par une victime avant sa disparition pour évaluer le temps nécessaire au trajet). L'enjeu de nos travaux de recherche se situe plutôt au stade de la saisie des **analystes criminels** sur **commission rogatoire** que dans le temps de la flagrance, lorsque l'**analyse criminelle** repose sur les documents produits par les investigations et consignés dans le dossier de **procédure**.

## 5.2 Cadre juridique

Plusieurs textes précisent le cadre légal d'exercice des analystes criminels :

- Les articles 230-13 à 27<sup>13</sup> et R40-39 à 41<sup>14</sup> du **Code de procédure pénale** (loi du 14 mars 2011 d'orientation et de programmation pour la performance de la sécurité intérieure);

---

12. Définie par les articles 151 et suivants du **Code de procédure pénale** [https://www.legifrance.gouv.fr/affichCode.do;jsessionid=EC897721353DA45774D2F79638ABD888.tplgfr42s\\_2?idSectionTA=LEGISCTA000006167428&cidTexte=LEGITEXT000006071154&dateTexte=20190410](https://www.legifrance.gouv.fr/affichCode.do;jsessionid=EC897721353DA45774D2F79638ABD888.tplgfr42s_2?idSectionTA=LEGISCTA000006167428&cidTexte=LEGITEXT000006071154&dateTexte=20190410) [consulté le 11 décembre 2019]

13. Légifrance : Articles 230-13 à 27 du **Code de procédure pénale** <https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006071154&idArticle=LEGIARTI000023709625&dateTexte=&categorieLien=cid> [consulté le 20 mars 2019]

14. Légifrance : Articles R40-39 à 41 du **Code de procédure pénale** [https://www.legifrance.gouv.fr/affichCode.do;jsessionid=30B70AE4E1CCA85899C8D19EB80B6598.tplgfr42s\\_2?idSectionTA=LEGISCTA000025833613&cidTexte=LEGITEXT000006071154&dateTexte=20190320](https://www.legifrance.gouv.fr/affichCode.do;jsessionid=30B70AE4E1CCA85899C8D19EB80B6598.tplgfr42s_2?idSectionTA=LEGISCTA000025833613&cidTexte=LEGITEXT000006071154&dateTexte=20190320) [consulté le 20 mars 2019]

- Le décret n°2012-687 du 7 mai 2012 relatif à la mise en œuvre de logiciels de rapprochement judiciaire à des fins d'analyse criminelle<sup>15</sup>;
- La décision n°2011-625 DC du Conseil Constitutionnel en date du 10 mars 2011<sup>16</sup>;
- La délibération n°2011-420 de la CNIL en date du 15 décembre 2011<sup>17</sup>

L'intervention de l'analyse criminelle est limitée à une enquête dont le périmètre est déterminé par le magistrat : l'analyse ne peut pas porter sur des éléments non mentionnés en procédure. Les analystes criminels ne peuvent pas rapprocher des dossiers partageant des similitudes sans accord des magistrats (une base de données de travail correspond à une affaire traitée). Toute information mentionnée dans les résultats de l'analyse (détaillés à la section 5.4.2) doit figurer dans les documents de la procédure, et l'on doit être capable de tracer son origine dans la procédure. Cet aspect est important à tout moment du travail, mais il revêt une importance capitale lors du procès, où les raisonnements du travail d'analyse criminelle peuvent être mis à mal par les parties engagées.

### 5.3 Formation des analystes criminels

Les analystes criminels sont des gendarmes ayant une expérience du terrain judiciaire, disposant de la qualification d'officier de police judiciaire (OPJ).

La formation « analyste criminel opérationnel » est dispensée au Centre National de Formation de la Police Judiciaire (CNFPJ) de Rosny-Sous-Bois, en partenariat avec l'Université de Troyes. La formation, d'une durée de six semaines, est entrecoupée par six mois d'application en unité. Le premier bloc de formation en quatre semaines aborde les connaissances pratiques et techniques et la maîtrise des outils logiciels. Lors du deuxième bloc de formation, en deux semaines, les techniques d'analyse criminelle sont consolidées et le stagiaire est évalué. Le passage en unité doit permettre le traitement d'un dossier.

---

15. Légifrance : Décret n°2012-687 <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000025879033&dateTexte=20190310> [consulté le 20 mars 2019]

16. Site du Conseil Constitutionnel : Décision n°2011-625 DC <https://www.conseil-constitutionnel.fr/decision/2011/2011625DC.htm> [consulté le 20 mars 2019]

17. Légifrance : Délibération n°2011-420 [consulté le 20 mars 2019]

## 5.4 Pratique de l'analyse criminelle à la Gendarmerie Nationale

L'analyse criminelle telle que nous l'avons observée au sein de la Gendarmerie nationale et plus spécialement au DSAC du SCRC est structurée en plusieurs étapes et s'appuie sur quelques concepts essentiels. La configuration de travail la plus fréquemment rencontrée y est l'étude de cas, c'est-à-dire l'étude d'une affaire en particulier.

### 5.4.1 Étude de faisabilité

Avant toute prise en charge, le dossier est examiné pour évaluer l'apport potentiel de l'analyse criminelle. Un analyste criminel en fait une lecture superficielle évaluant la quantité, la qualité de l'information contenue dans la procédure, les investigations déjà menées et celles qui n'ont pas été effectuées, le temps de travail nécessaire, etc. Un rapport de quelques pages est réalisé afin de rendre la décision.

Si à l'issue de cette évaluation, les analystes estiment que l'analyse criminelle peut apporter quelque chose au dossier et que sa prise en charge est acceptée, les pièces à disposition sont listées et les objectifs de l'analyse sont définis avec le magistrat ou le directeur d'enquête. Par exemple, on peut décider de focaliser l'attention sur les déplacements géographiques, sur une fenêtre temporelle en particulier, sur les contacts téléphoniques ou sur l'environnement relationnel de la victime ou d'un suspect. Le travail d'analyse criminelle proprement dit peut commencer ensuite : les analystes, par équipe de deux, s'attellent à la lecture et à l'extraction de l'information des documents. Nous présentons dans les parties suivantes le concept d'entité et la construction de schémas, sur lesquels le processus s'appuie.

### 5.4.2 Hypothèses de travail et restitution de l'analyse

Au-delà de la réalisation des schémas, le véritable objectif de l'analyse criminelle, par le biais d'un regard neuf et d'une synthétisation de la procédure, est de proposer des pistes d'investigation nouvelles. Ce rendu prend la forme d'un procès-verbal de synthèse présentant plusieurs hypothèses appuyées par les schémas. Les synthèses restent ouvertes en présentant plusieurs scénarios, il ne s'agit pas d'établir

une conclusion ferme. Prenons le cas d'une disparition : selon les éléments du dossier, il sera peut-être possible d'envisager un suicide, un enlèvement par un proche, ou une mauvaise rencontre. Les scénarios sont étayés par des éléments concrets référencés en [procédure](#) et l'[analyste criminel](#) les classe selon leur vraisemblance (toujours en s'appuyant sur les éléments de la [procédure](#)).

Pour la présentation des résultats, les schémas sont imprimés et peuvent mesurer plusieurs mètres de long. Les [analystes criminels](#) présentent et discutent les résultats de l'analyse avec le magistrat et les enquêteurs (avec lesquels par ailleurs ils sont restés en contact tout au long du processus).

Les [analystes criminels](#) sont aussi parfois convoqués pour témoigner lors du procès. Ils doivent être capables de justifier les orientations de leur travail et faire preuve d'une grande maîtrise des points pouvant faire l'objet d'attaques des différentes parties du procès.

## 6 Enjeux, contraintes et difficultés

La présentation du contexte de pratique et de la méthodologie appliquée en [analyse criminelle](#) a permis de cerner globalement son rôle au cœur du processus d'[enquête](#) et son fonctionnement. Nous allons maintenant présenter les obstacles aussi bien pratiques que théoriques auxquels elle se heurte, et nous poserons les bases d'une réflexion sur sa proximité avec des champs de recherche liés au texte et à la langue afin d'en dresser les perspectives de développement.

### 6.1 Sur la méthode

Nous souhaitons souligner le caractère fortement humain du travail produit en [analyse criminelle](#). L'objectivité de l'[enquête](#) est un but à atteindre mais quels qu'en soient les acteurs, enquêteurs, magistrats ou analystes, tous agissent avec leur propre contexte. Le résultat de l'[analyse criminelle](#) est le fruit de réflexions et d'interprétations humaines, avec ce que cela comporte de subjectivité et de mise à profit de l'expérience individuelle et en aucun cas le produit d'un logiciel.

D'un point de vue pratique, la méthode de travail requiert concentration, minutie et attention aux détails. Dans des conditions idéales, une longueur de traitement de

l'ordre de un à deux ans voire plus est à prévoir, en raison du temps de lecture et du temps de saisie des informations.

Théoriquement, la méthode préconise une première lecture intégrale du dossier, puis l'intégration de l'ensemble des informations de la [procédure](#). En réalité, les analystes adaptent la méthode à chaque situation et selon leur expérience : ils peuvent par exemple intégrer les informations au fil de la lecture et sélectionner les informations utiles, surtout si le temps qui leur est alloué pour réaliser l'analyse est court. L'[analyse criminelle](#) est soumise à des contraintes d'efficacité en termes de temps de traitement, que l'on se situe dans le cas de faits récents ou anciens. Nous retenons deux aspects en ce qui concerne la méthode : l'implication et l'importance du raisonnement humain dans un processus contraint dans le temps (entre autres limitations).

## 6.2 Sur la matière

Concernant la matière traitée, le problème principal vient de la numérisation des documents. Bien que les [procédures](#) soient communiquées aux [analystes criminels](#) le plus souvent au format électronique (gravées sur CD), il s'agit de documents papiers scannés et non pas de fichiers informatiques originaux, traitements de texte ou [PDF](#). En effet, le droit français exige la signature manuscrite des dépositions, et d'autre part, il n'est pas rare que les documents présentent des mentions manuscrites (voire, des paragraphes ou documents entiers rédigés à la main). Ces documents sont passés par un logiciel de reconnaissance optique de caractères ([OCR](#)), dont la qualité des résultats est variable, et ce, même sur des affaires relativement récentes.

La question de la numérisation des services judiciaires semble être prioritaire pour le ministère de la Justice à l'heure actuelle<sup>18</sup>, mais en ce qui concerne l'[analyse criminelle](#) les réformes futures ne pourront de toutes façons rien au cas des [cold cases](#), les affaires anciennes, auxquelles le [DSAC](#) est souvent confronté. La réactivation d'une affaire ancienne ou la prise en charge en [analyse criminelle](#) d'une affaire de longue haleine demande en premier lieu un travail de numérisation des documents de la [procédure](#). La qualité des documents papier influe sur la qualité de

---

18. Voir par exemple : La justice multiplie les chantiers numériques, *Le Monde*, 16 octobre 2018, [https://www.lemonde.fr/police-justice/article/2018/10/16/la-justice-multiplie-les-chantiers-numeriques\\_5370085\\_1653578.html](https://www.lemonde.fr/police-justice/article/2018/10/16/la-justice-multiplie-les-chantiers-numeriques_5370085_1653578.html) [consulté le 9 avril 2019]

l'OCR, un procédé initialement imparfait même lorsque les documents papier sont de bonne qualité. Dans le cas des *cold cases* justement, la *procédure* peut avoir circulé entre de nombreux services, être passée des services de la police à ceux de la gendarmerie (ou inversement), les documents peuvent avoir été photocopiés, puis scannés, réimprimés, photocopiés de nouveau. Autant d'opérations qui dégradent la qualité des documents lorsqu'ils arrivent aux mains des *analystes criminels*. La circulation de la *procédure* entre acteurs de l'enquête représente donc potentiellement un obstacle à son traitement : si la reconnaissance des caractères n'est pas la hauteur, l'efficacité de simples recherches dans le texte est impactée, ainsi que tout traitement informatique du texte.

### 6.3 Sur l'organisation documentaire

Il n'y a pas de norme d'organisation des documents de *procédure*. Lorsque le dossier parvient à l'*analyste criminel*, les fichiers peuvent être classés par ordre chronologique de production, ou selon des thématiques propres à l'enquêteur, ou encore ne présenter aucune logique particulière. Le découpage des dossiers en fichiers informatiques n'est pas non plus normé. Un fichier peut contenir une seule pièce de *procédure* (par exemple, une *audition*) ou mélanger plusieurs pièces qui ne sont pas forcément toutes du même type<sup>19</sup>. Les pièces du dossier doivent être cotées, mais les pratiques de cotation ne sont pas uniformisées : les enquêteurs réalisent une première cotation, puis transmettent le dossier au magistrat dont les services peuvent réaliser une nouvelle cotation. Cela donne lieu à des référencements de pièces qui peuvent diverger ou se superposer, dans le cas par exemple d'affaires archivées puis réactivées, or comme nous l'avons indiqué, les informations retenues doivent pouvoir être reliées aux pièces de *procédure* dont elles sont issues. Ce manque de coordination et de normes dans l'organisation des documents de la *procédure* complexifie la prise en main du dossier par l'*analyste criminel* qui ne peut pas toujours compter sur un classement clair pour se retrouver dans les fichiers .

---

19. Le DSAC déjà été confronté à un dossier de *procédure* contenu entièrement dans un seul fichier PDF de plus de sept mille pages.

#### 6.4 Sur le texte et le contenu

Le dossier de **procédure**, dont nous détaillerons la composition au chapitre II, rassemble une variété de documents dont une bonne part est textuelle. La pratique de l'**analyse criminelle** entretient donc un lien étroit avec une matière textuelle au volume conséquent et riche en information : c'est en même temps un obstacle et un atout. Le volume textuel, si l'on ne dispose pas d'outils adéquats pour le manipuler, peut sérieusement compliquer la tâche : repérer un élément, suivre son évolution dans le texte, le retrouver plus tard... sont autant d'opérations simples qui peuvent devenir extrêmement pénibles si l'on n'a pas d'autre possibilité que celle de parcourir et lire les documents un à un.

La masse d'information quant à elle place en quelque sorte l'**analyste criminel** dans la peau d'un chercheur d'or : après avoir amassé le plus d'éléments possible, il faut trier le bon grain de l'ivraie, sélectionner un détail et en écarter un autre. Un tri qui peut se révéler très difficile à opérer tant il est impossible d'anticiper l'importance de tel ou tel élément dans une affaire.

La culture populaire, par le biais des romans policiers, thrillers, polars, nous a habitués à envisager l'**enquête** policière comme le suivi d'une piste que le fin limier parcourt au gré des témoignages et des indices, avec éventuellement une ou deux réorientations pour cause de « fausse piste ». Dans la réalité, nous avons constaté que le travail d'**enquête** fonctionne par l'élimination d'innombrables possibilités, élimination qui peut avoir à se baser sur plusieurs documents ou témoignages.

Prenons l'exemple d'une personne retrouvée morte sur la voie publique sans **témoin** oculaire des faits. L'**autopsie** révèle que la victime s'est défendue. Une **réquisition** est alors adressée au service des urgences orthopédiques de l'hôpital local afin d'obtenir la liste des personnes ayant présenté des blessures aux mains à compter de la date déduite de la mort. La liste est confrontée au fichier des antécédents judiciaires. Si un profil est repéré, la personne est convoquée pour justifier son emploi du temps dans une déposition. Éventuellement, d'autres personnes pouvant certifier ces dires sont convoquées aussi. Tous ces témoignages sont consignés dans le dossier de **procédure**. Même si en soi, ils n'apportent pas d'éléments nouveaux, ils attestent que cette piste a été explorée puis écartée sur la base d'éléments solides.

Chacune de ces possibilités explorée génère un volume documentaire et un bruit parasitant des informations plus essentielles. Néanmoins il n'est pas question de se passer de ces recherches, comme nous l'avons dit. La question est plutôt : comment optimiser le repérage des informations ? Comment construire un balisage du contenu qui permette à l'[analyste criminel](#) de saisir la teneur des documents au fil de son travail ? Pour filer la métaphore du chercheur d'or : quel est le meilleur tamis et comment le construire ? La théorie de l'[analyse criminelle](#) fournit elle-même une réponse partielle à cette question car elle exploite des concepts qui sont des clés d'accès au texte : les entités, dont certains types fondamentaux se retrouvent dans pratiquement toutes les affaires, doivent permettre un premier élagage de la masse documentaire.

## 7 Conclusion : que proposer à l'analyse criminelle ?

L'[analyse criminelle](#), en tant que pratique plaçant un être humain face à un ensemble de documents duquel il doit extirper de l'information, partage des points communs avec la linguistique de corpus et la fouille de textes. Nous avons identifié les difficultés qui lui sont propres : le rôle fondamental du raisonnement humain, des contraintes temporelles de traitement, une matière dense et difficile d'accès.

Ce contexte est propice à de nombreux axes de développement allant vers l'automatisation de la gestion de l'information et sa représentation graphique. À ce sujet, on peut constater que les travaux de TAPASWI et al. (2014) offrent un parallèle intéressant avec le cas qui nous intéresse en cherchant à construire automatiquement les lignes de vie de personnages de séries télévisées d'après les scripts des épisodes. Cependant, la nature de la donnée traitée est moins complexe que dans notre cas, et le domaine d'application, les séries télévisées, n'impliquent pas de conséquences légales.

Nous avons aussi l'exemple de pratiques professionnelles assistées par ordinateur comme la traduction, dans lesquelles le temps de travail consacré à la post-édition peut s'avérer plus long et plus pénible que lorsque la tâche est réalisée entièrement par l'être humain. Dans de tels cas, la question du bénéfice réel de la technologie se pose.

Enfin, l'engouement pour les technologies d'apprentissage automatique s'accompagne parfois d'un manque de discernement médiatique sur la prétendue impartialité des algorithmes. Il est pourtant fortement soupçonné que les algorithmes appliqués dans des contextes judiciaires reproduisent des discriminations basés sur des critères sociaux-économiques, de genre ou de race<sup>20</sup>.

Par conséquent, nous considérons que le rôle central de l'humain au coeur de l'analyse criminelle doit être conservé. POIBEAU (2014b) souligne le rôle de l'analyste humain au sein des systèmes d'extraction d'information en rappelant que les outils automatiques ne se substituent pas à l'expertise humaine, et que l'analyse et l'évaluation d'informations extraites automatiquement reste l'apanage de l'expert. Dans notre cas, les analystes criminels sont des professionnels dotés d'une expérience du terrain judiciaire et d'une formation complémentaire leur conférant une expertise dont il serait dommage et risqué de se passer. Il est illusoire de penser qu'il soit souhaitable et possible d'automatiser en profondeur le travail d'analyse criminelle. Nous orientons donc nos recherches vers l'adaptation des pratiques d'exploration textuelle et de fouilles de texte pour concevoir de nouveaux outils de recherche d'information tout en concevant une nouvelle méthodologie d'approche des documents de la procédure, en particulier des documents textuels. Dans cette optique, le raisonnement reste à la charge de l'humain, mais la collecte des éléments doit être facilitée.

Nous nous attacherons dans les chapitres suivants à détailler autant que possible la nature des données traitées en analyse criminelle, puis, armée de ces observations, nous réaliserons un état de l'art des techniques et approches pertinentes pour notre problématique, avant de décrire les concepts propres à l'analyse criminelle et à les relier à des concepts comparables rencontrés dans les disciplines de gestion automatique du texte. Nous présenterons ensuite une approche préliminaire d'extraction automatique d'information adaptée au cas de l'analyse criminelle, et enfin nous tenterons de définir la nature de la donnée textuelle de la procédure judiciaire en confrontation avec d'autres textes du domaine de la justice.

---

20. Voir : Just how transparent can a criminal justice algorithm be? <https://slate.com/technology/2018/07/pennsylvania-commission-on-sentencing-is-trying-to-make-its-algorithm-transparent.html?via=gdpr-consent>, *Slate*, 3 juillet 2018 [consulté le 02/05/2019].

## Chapitre II

# De la procédure judiciaire criminelle au corpus textuel

Après avoir présenté le contexte de recherche, il est nécessaire de procéder à la description des données exploitées en [analyse criminelle](#) et qui serviront de base à notre étude. Pour cela, nous avons étudié plusieurs dossiers de [procédures](#) d'atteintes aux personnes (homicides, disparitions) comme ceux traités par le [DSAC](#).

Avant d'entamer leur description, précisons que nous avons pu les consulter au sein des bâtiments du [PJGN](#), sous le contrôle des officiers et sous-officiers responsables du Département, et dans le cadre d'une convention de partenariat entre le [PJGN](#) et l'université de Cergy-Pontoise comportant une clause de confidentialité. La présentation des données sera donc d'ordre général, sans description précise des faits investigués, sans mention d'information portant atteinte à la vie privée des personnes impliquées ou d'éléments permettant d'identifier l'affaire. Les extraits cités pour des besoins d'illustration le seront après masquage ou pseudonymisation <sup>1</sup>.

L'objectif de ce chapitre est de décrire le contenu d'une [procédure](#) judiciaire criminelle telle qu'elles sont confiées aux [analystes criminels](#), puis d'en dégager les aspects textuels afin de déterminer d'une part leur potentiel en tant que ressource d'étude et de développement, et d'autre part les modalités d'automatisation de leur traitement.

---

1. La pseudonymisation consiste à remplacer une information par une autre équivalente afin de préserver l'anonymat (une date par une autre date, un prénom par un autre prénom, etc.). La pseudonymisation n'est réalisée que pour les besoins de ce manuscrit, les travaux ont été réalisés sur le texte original.

## 1 Qu'est-ce qu'une procédure judiciaire ?

Les chroniqueurs judiciaires font parfois mention lorsqu'ils relatent un procès de la masse des dossiers d'archives qui compile le travail d'investigation ayant été mené jusque-là. Si l'on peut supposer que la numérisation des services de la justice permettra à l'avenir de se passer du papier, les **procédures** sont encore parfois confiées aux analystes criminels sous la forme de plusieurs boîtes d'archives, comme à la figure 4. Néanmoins, ce cas de figure se raréfie, et même les **cold cases** circulent aujourd'hui sous forme électronique<sup>2</sup>. Au-delà de sa forme matérielle, commençons par définir la **procédure** judiciaire.



FIGURE 4 – Un dossier de **procédure** relatif à une disparition au format papier

La « **procédure** judiciaire » est l'ensemble des démarches effectuées pour mener une action en justice. Son déroulement, selon les configurations, est encadré par les articles 53 et suivants du **Code de procédure pénale** pour l'**enquête de flagrance**<sup>3</sup>, 75 et suivants pour l'**enquête préliminaire**<sup>4</sup>, 74 et suivants pour les découvertes de cadavres et les disparitions inquiétantes<sup>5</sup> et 151 pour la commission rogatoire<sup>6</sup>.

Le dossier de **procédure** judiciaire rassemble les documents compilés par les enquêteurs dans l'objectif de la manifestation de la vérité.

2. Pour les **cold cases**, la numérisation des documents devient d'ailleurs une étape préalable à la reprise des investigations.

3. Légifrance : [https://www.legifrance.gouv.fr/affichCode.do;jsessionid=0FB4EA802845B5D4FF4651709A000D1C.tplgfr42s\\_2?idSectionTA=LEGISCTA000006151876&cidTexte=LEGITEXT000006071154&dateTexte=20040310#LEGIARTI000006575016](https://www.legifrance.gouv.fr/affichCode.do;jsessionid=0FB4EA802845B5D4FF4651709A000D1C.tplgfr42s_2?idSectionTA=LEGISCTA000006151876&cidTexte=LEGITEXT000006071154&dateTexte=20040310#LEGIARTI000006575016) [consulté le 11 décembre 2019]

4. Légifrance : [https://www.legifrance.gouv.fr/affichCode.do;jsessionid=0FB4EA802845B5D4FF4651709A000D1C.tplgfr42s\\_2?idSectionTA=LEGISCTA000006151877&cidTexte=LEGITEXT000006071154&dateTexte=20040310](https://www.legifrance.gouv.fr/affichCode.do;jsessionid=0FB4EA802845B5D4FF4651709A000D1C.tplgfr42s_2?idSectionTA=LEGISCTA000006151877&cidTexte=LEGITEXT000006071154&dateTexte=20040310) [consulté le 11 décembre 2019]

5. Légifrance : <https://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000020632053&cidTexte=LEGITEXT000006071154&dateTexte=20090514> [consulté le 11 décembre 2019]

6. Légifrance : [https://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=62F4026FB69ECD7B0C8350B6FB04D828.tplgfr42s\\_2?cidTexte=LEGITEXT000006071154&idArticle=LEGIARTI000006575360&dateTexte=20191211&categorieLien=cid#LEGIARTI000006575360](https://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=62F4026FB69ECD7B0C8350B6FB04D828.tplgfr42s_2?cidTexte=LEGITEXT000006071154&idArticle=LEGIARTI000006575360&dateTexte=20191211&categorieLien=cid#LEGIARTI000006575360) [consulté le 11 décembre 2019]

Dans le cas de la [commission rogatoire](#), le cadre juridique de production de ce dossier est précisé par l'article 81 du [Code de procédure pénale](#)<sup>7</sup> (extrait) :

Le [juge d'instruction](#) procède, conformément à la loi, à tous les **actes d'information** qu'il juge utiles à la manifestation de la vérité. Il instruit à charge et à décharge.

Il est établi une **copie de ces actes** ainsi que de toutes les **pièces de la procédure** ; chaque copie est certifiée conforme par le greffier ou l'officier de police judiciaire commis [...]. Toutes les pièces du dossier sont **cotées** par le greffier au fur et à mesure de leur rédaction ou de leur réception par le juge d'instruction.

Toutefois, si les copies peuvent être établies à l'aide de procédés photographiques ou similaires, elles sont exécutées à l'occasion de la **transmission du dossier**. Il en est alors établi autant d'exemplaires qu'il est nécessaire à l'administration de la justice. Le greffier certifie la conformité du dossier reproduit avec le dossier original [...].

Dans cet article sont introduites les notions d'acte d'information, de pièce de [procédure](#), et il est précisé quelques modalités de leur circulation. On comprend que les actes d'information, qui sont réalisés au cours de l'[enquête](#), débouchent sur la production de documents appelés pièces de [procédure](#) qui relatent ces actes, et que l'ensemble des pièces de [procédure](#) constitue le dossier. Dans la suite de ce chapitre, nous nous efforçons de donner un aperçu de la nature de ces documents, qui constituent donc la base du travail des [analystes criminels](#).

## 2 Les pièces de procédure

Nous avons répertorié une quinzaine de types de documents constituant des pièces de [procédure](#), que nous avons rassemblés en deux catégories (tableau 1). D'un côté, la catégorie des documents réglementaires et de [procédure](#) constituée par les pièces liées au déroulement légal et opérationnel de l'[enquête](#), et de l'autre les documents d'information, qui convoient des éléments d'information relatifs au faits.

7. Légifrance : [Article 81 du Code de procédure pénale](#) [consulté le 09 mai 2019].

Documents réglementaires et de procédure	Documents d'information
<ul style="list-style-type: none"> <li>- Réquisitions,</li> <li>- Bordereau d'envoi,</li> <li>- Inventaire des pièces à conviction,</li> <li>- PV de notification de garde à vue,</li> <li>- PV de saisie,</li> <li>- Certificats médicaux préalables à la garde à vue</li> </ul>	<ul style="list-style-type: none"> <li>- PV transport constatations mesures prises,</li> <li>- PV d'audition de témoin et de garde à vue,</li> <li>- PV d'actes d'enquête divers (investigations, renseignements, perquisition...),</li> <li>- PV de synthèse,</li> <li>- Rapports d'expertises,</li> <li>- Retours des réquisitions : téléphoniques, bancaires, péages, compagnies d'assurance, etc.,</li> <li>- Documents graphiques : photos et vidéos,</li> <li>- Autres documents.</li> </ul>

TABLE 1 – Récapitulatif des documents rencontrés en **procédure** traitées en **analyse criminelle**.

Les sous-parties suivantes sont consacrées à la description des pièces, en détaillant notamment leur format, leur rôle pour l'**enquête**, et leur émetteur.

L'inventaire de la matière documentaire que nous allons réaliser dans cette sous-partie n'est pas exhaustif et ne constitue en aucun cas un référentiel rigide de la constitution de la **procédure**. Il s'agit d'un aperçu livré sur la base de nos observations qui de plus ne concerne que les pratiques de la **Gendarmerie nationale**. Dans la Police les pratiques et les formats de documents diffèrent, ce qui d'ailleurs dans notre objectif est à prendre en compte lorsqu'une affaire « circule » entre Police et Gendarmerie, puisque le dossier mélange alors des pièces issues des deux institutions.

## 2.1 Documents réglementaires et de procédure

Les documents réglementaires et de **procédure** sont des documents ne convoyant pas en tant que tel des informations à propos aux faits. Leur fonction est de garantir la légalité et la conformité des opérations ainsi que le respect des droits des personnes impliquées.

Nous classons dans cette catégorie :

- Les bordereaux d'envoi : il s'agit de documents faisant office de sommaire et de liste des pièces lors de la transmission du dossier,
- Les inventaires de pièces à conviction, faisant la liste des **scellés**,

- Les **réquisitions** de tiers et d'experts (figure 5) afin d'obtenir des informations, une assistance ou des moyens matériels, elles sont adressées aux opérateurs téléphoniques, établissements bancaires, médecins légistes, maires, pompes funèbres, laboratoires photographiques, interprètes, etc.
- Les certificats médicaux préalables à la **garde à vue**, qui attestent que l'état de santé de la personne est conforme avec la mesure de **garde à vue**,
- Les **procès-verbaux** de notification de **garde à vue**, qui en consignent les modalités : heure de début et de fin, notification des droits de la personne gardée à vue, consultation d'un médecin, assistance d'un avocat, temps de repos, etc.,
- Les **procès-verbaux** de saisie, qui documentent la collecte d'un élément matériel placé sous **scellé**.

Au-delà de leur fonction au sein de la **procédure**, nous avons regroupé ces documents car ils sont semi-structurés : soit comme des formulaires, soit sous la forme d'une langue très normée comme la figure 5 l'illustre, soit parce qu'ils constituent des inventaires d'éléments de la **procédure**. Leur rôle au sein de la **procédure** ne concerne pas directement les faits, mais ils fournissent des informations sur le déroulé des investigations. C'est pourquoi nous avons décidé de les considérer à part du reste de la **procédure**.

## 2.2 Documents d'information

Nous regroupons comme documents d'information des documents convoyant des informations directement liées aux faits ou des comptes-rendus d'opérations menées par les enquêteurs. Ces documents consignent une information dense et non-structurée qui est le matériau informationnel au cœur de l'**enquête**. Bon nombre d'entre eux sont des **procès-verbaux**.

Le dictionnaire Larousse en ligne propose la définition suivante pour le terme **procès-verbal** :

Acte écrit rédigé par un magistrat, un officier ou agent de police judiciaire, un officier public, qui rend compte de ce qu'il a fait, entendu ou

République Française – Ministère de la Justice  
**TRIBUNAL DE GRANDE INSTANCE DE [REDACTED]**  
*Parquet du Procureur de la République*

D 606  
C

### REQUISITIONS

Le Procureur de la République près le Tribunal de Grande instance de [REDACTED]

Vu la découverte le [REDACTED] à [REDACTED]

du cadavre de : [REDACTED]

Vu l'enquête suivie :

<input type="checkbox"/> Vu l'article 74 du Code de Procédure Pénale Attendu que les causes de la mort apparaissent inconnues ou suspectes	<input checked="" type="checkbox"/> Vu l'article 60 - 77-1 du Code de Procédure Pénale Attendu qu'au vu des premiers renseignements apportés par l'enquête il y a lieu de faire procéder à des constatations, examens techniques ou scientifiques
---	--

### REQUIERT

■ Monsieur le Professeur [REDACTED] Expert inscrit sur la liste de la Cour d'Appel de [REDACTED]

#### AUX FINS DE

- ✕ • Procéder à l'analyse des prélèvements qui font l'objet des scellés n° 1A et 16A, 17 et 20 A.
- ✕ • Procéder sur ces prélèvements à la recherche de sperme - de sang d'origine humaine ;
- ✕ • Procéder sur ces prélèvements à toutes analyses et recherches d'empreintes génétiques ;
- Procéder à tous prélèvements de comparaison utiles sur la personne de [REDACTED] et sur ces derniers prélèvements à toutes analyses et recherches d'empreintes génétiques aux fins de les comparer aux résultats obtenus par les analyses exécutées sur les scellés ;
- ✕ • dresser un rapport d'ensemble détaillé des opérations et des conclusions.

Les ouvertures nécessaires de scellés seront faites conformément aux dispositions de l'article 60 alinéa 3 du Code de Procédure Pénale.

Fait au Parquet, le [REDACTED]  
 Le Procureur de la République,

[REDACTED]

[REDACTED] - [REDACTED] Cedex  
 [REDACTED] - Fax : [REDACTED]

FIGURE 5 – Exemple de réquisitions d'expert

constaté dans l'exercice de ses fonctions. (Abréviation familière : P.-V.)<sup>8</sup>

Cette définition est en accord avec l'article 429 du [Code de procédure pénale](#)<sup>9</sup> :

Tout procès-verbal ou rapport n'a de valeur probante que s'il est régulier en la forme, si son auteur a agi dans l'exercice de ses fonctions et a rapporté sur une matière de sa compétence ce qu'il a vu, entendu ou constaté personnellement. [...]

Le [procès-verbal](#) est donc un document officiel rédigé par une personne qualifiée et rapportant des faits qu'elle a perçus personnellement.

En [Gendarmerie nationale](#), les [procès-verbaux](#) s'organisent selon une structure verticale<sup>10</sup> (figures 6, 7) : un en-tête mentionne les références de l'unité et la nature du document, suivie des date, heure, lieu de la rédaction, suivies des nom, grade et affectation du ou des auteurs du document, et enfin du cadre légal des opérations rapportées (cadre général et renseignements sur la commission rogatoire). Ensuite vient le corps du texte à proprement parler, et enfin le document s'achève par la ou les signatures et visas nécessaires.

Néanmoins, cette structure n'est pas rigide et s'adapte selon les besoins et la finalité du [procès-verbal](#). Par exemple, les [auditions de témoins](#) comportent une partie de renseignements d'état-civil à propos de la personne auditionnée.

### 2.2.1 Procès-verbaux de transport constatations et mesures prises

Il s'agit d'un document, qui, comme son nom l'indique, relate le déplacement et les premières mesures adoptées par un ou des gendarmes suite au signalement ou à la découverte de faits criminels. Lors de la découverte d'une scène de crime par exemple, les techniciens en identification criminelle ([TIC](#)) se rendent sur les lieux et procèdent aux actes de police technique et scientifique comme le prélèvement et la protection des éléments matériels ou la prise de mesures et de photographies de la scène de crime. Ce document peut se structurer en plusieurs rubriques : la saisine, la

---

8. [https://www.larousse.fr/dictionnaires/francais/procès-verbal\\_procès-verbaux/64067](https://www.larousse.fr/dictionnaires/francais/procès-verbal_procès-verbaux/64067) [consulté le 11 juin 2019]

9. Légifrance : article 429 du [Code de procédure pénale](#) <https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006071154&idArticle=LEGIARTI000006576551> [consulté le 11 juin 2019]

10. En comparaison, les [procès-verbaux](#) de Police comportent une marge à gauche où figurent les références et l'objet des investigations.

description de la situation à l'arrivée des enquêteurs, les mesures prises, la liste des personnes présentes, la situation et l'environnement des lieux (topographie régionale, d'ensemble et rapprochée), les conditions climatiques, le corps du délit (dans le cas d'un homicide, position du cadavre, habillement, éventuellement éléments matériels le camouflant, blessures visibles), les prélèvements effectués, les constatations en présence du [médecin légiste](#), les saisies d'éléments matériels et enfin les conclusions de cette première étape de l'[enquête](#). Deux paragraphes d'exemple ci-dessous sont tirés d'un [procès-verbal](#) de transport, constatations et mesures prises suite à la découverte d'un cadavre.

### 1- SAISINE

Le 03 juillet 2012 à 15 heures 15, nous recevons un appel téléphonique de la Brigade de Recherches d'ETAMPES nous informant de la découverte d'un corps humain enterré dans la forêt au Nord de l'agglomération de DOURDAN (91). Cette unité demande notre concours aux fins de procéder aux constatations et opérations de police technique.

Nous nous transportons immédiatement sur les lieux en compagnie du Major GAGNAC commandant notre unité.

[...]

4) - position générale du corps :

Le cadavre est en position décubitus dorsal légèrement désaxé côté droit. L'ensemble tête-tronc, membre inférieur gauche se trouvent sensiblement dans le prolongement. La tête est légèrement inclinée sur la gauche. Le corps présente une légère cambrure au niveau des reins ainsi qu'une faible torsion au niveau du thorax. Le membre inférieur droit est relevé, et fait un angle droit à l'articulation du genou. Le talon droit se trouve au niveau de la partie médiane de la face interne du mollet gauche.

Les [procès-verbaux](#) de transport, constatation et mesures prises sont donc des documents rapportant un certains nombres d'opérations en s'appuyant sur une pratique formalisée : l'enquêteur est saisi, se déplace, décrit la situation et prend des mesures. Le texte du document est à la fois narratif et descriptif (situation et opérations), organisé en paragraphes, exprimé à l'aide de phrases complètes rédigées et

comporte plusieurs pages.

### 2.2.2 Procès-verbaux d'auditions et d'interrogatoires

Au cours d'une **enquête**, deux sortes d'**auditions** peuvent être menées : les **auditions** de **témoins**, où la personne entendue, libre, se présente sur convocation ou de sa propre initiative, et les **auditions** de personnes mises en cause, où la personne entendue est gardée à vue ou **mise en examen**.

Les conditions d'**audition** des premières sont précisées par les articles 101 à 113<sup>11</sup> et 113-1 à 113-8<sup>12</sup> du **Code de procédure pénale**. Quant aux personnes mises en cause, il s'agit des articles 114 à 121 du **Code de procédure pénale**<sup>13</sup> dans le cadre de la **commission rogatoire**, 77 et suivants dans le cadre de l'**enquête préliminaire**<sup>14</sup>, et 63 et suivants dans le cadre de l'**enquête de flagrance**<sup>15</sup>.

En pratique, les **auditions** sont réalisées par deux ou trois personnels ayant la qualification d'**officiers de police judiciaire**, ou par le **juge d'instruction** pour les personnes **mises en examen**. L'**audition** est une conversation entre les enquêteurs et le **témoin**. La personne entendue s'exprime oralement, les enquêteurs posent des questions, et les propos sont pris en note par l'un des enquêteurs au fil de la discussion. Selon les rapports de la personne entendue avec les faits investigués, l'**audition** peut concerner toutes sortes de sujets. On interrogera les proches sur les éléments psychologiques, la description du caractère, des habitudes, des connaissances, du train de vie de la victime, pour les **témoins** oculaires, on s'intéressera aux personnes, aux lieux, aux horaires et à tout élément ayant pu attirer leur attention.

---

11. Légifrance : Articles 101 à 113 du **Code de procédure pénale** [https://www.legifrance.gouv.fr/affichCode.do;jsessionid=24358CA45B3E6900D6BE9EC1D3A3D9D9.tplgfr42s\\_2?idSectionTA=LEGISCTA000006182923&cidTexte=LEGITEXT000006071154&dateTexte=20190510](https://www.legifrance.gouv.fr/affichCode.do;jsessionid=24358CA45B3E6900D6BE9EC1D3A3D9D9.tplgfr42s_2?idSectionTA=LEGISCTA000006182923&cidTexte=LEGITEXT000006071154&dateTexte=20190510) [consulté le 10 mai 2019].

12. Légifrance : Articles 113-1 à 113-8 du **Code de procédure pénale** [https://www.legifrance.gouv.fr/affichCode.do;jsessionid=24358CA45B3E6900D6BE9EC1D3A3D9D9.tplgfr42s\\_2?idSectionTA=LEGISCTA000006182888&cidTexte=LEGITEXT000006071154&dateTexte=20190510](https://www.legifrance.gouv.fr/affichCode.do;jsessionid=24358CA45B3E6900D6BE9EC1D3A3D9D9.tplgfr42s_2?idSectionTA=LEGISCTA000006182888&cidTexte=LEGITEXT000006071154&dateTexte=20190510) [consulté le 10 mai 2019].

13. Légifrance : Articles 114 à 121 du **Code de procédure pénale** [https://www.legifrance.gouv.fr/affichCode.do;jsessionid=24358CA45B3E6900D6BE9EC1D3A3D9D9.tplgfr42s\\_2?idSectionTA=LEGISCTA000006167425&cidTexte=LEGITEXT000006071154&dateTexte=20190510](https://www.legifrance.gouv.fr/affichCode.do;jsessionid=24358CA45B3E6900D6BE9EC1D3A3D9D9.tplgfr42s_2?idSectionTA=LEGISCTA000006167425&cidTexte=LEGITEXT000006071154&dateTexte=20190510) [consulté le 10 mai 2019].

14. Légifrance : <https://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006575132&cidTexte=LEGITEXT000006071154&dateTexte=19930301> [consulté le 11 décembre 2019]

15. Légifrance : [https://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=62F4026FB69ECD7B0C8350B6FB04D828.tplgfr42s\\_2?idArticle=LEGIARTI000006575065&cidTexte=LEGITEXT000006071154&dateTexte=20001231](https://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=62F4026FB69ECD7B0C8350B6FB04D828.tplgfr42s_2?idArticle=LEGIARTI000006575065&cidTexte=LEGITEXT000006071154&dateTexte=20001231) [consulté le 11 décembre 2019]

Des dispositions spécifiques sont prévues pour certains types de publics : les **auditions** de mineurs doivent être filmées, tout comme les **auditions** de personnes mises en cause (voir section 2.2.9). Ces enregistrements sont versés au dossier de **procédure**. Si la personne n'est pas capable de s'exprimer en français, il peut être fait appel à un interprète en langue des signes ou en langue étrangère ou régionale. Enfin, la personne entendue peut être assistée par un avocat.

La figure 6 présente la première page d'une **audition** de **témoin**. Les propos tenus lors de l'**audition** sont consignés après les différents éléments de contexte que nous avons décrits ci-dessus. Au bas de chaque page, la personne auditionnée et les **officiers de police judiciaire** apposent leurs signatures. La longueur des **procès-verbaux** d'**audition** varie de une à plusieurs dizaines de pages (notamment pour les **auditions** de **garde à vue**).

Dans l'exemple 6, les propos sont consignés sous la forme d'un dialogue entre les gendarmes et le **témoin** entendu. Bien que se calquant sur l'échange oral de l'**audition**, cette forme dialogique du texte du **procès-verbal** n'est pas systématique : nous avons rencontré des **auditions** consignées entièrement sous la forme de récit à la première personne, comme cela est préconisé notamment par la méthode d'interrogatoire ProGREAI (BLANCHET et al., 2013).

Contrairement aux autres documents d'information de l'**enquête**, la pratique de l'**audition** a déjà fait l'objet de recherches en sciences humaines. Plusieurs domaines l'ont étudiée : le droit en étudie la constitutionnalité et les rapports entre **audition** libre et **audition** sous contrainte (LAMY, 2012), la psychologie a produit des recherches sur les entretiens cognitifs (BRUNEL et al., 2013) ou les **auditions** d'enfants (VERKAMPT et al., 2010; VILAMOT et al., 2010; VERKAMPT, 2013), et dans une certaine mesure, la linguistique s'est également penchée sur le sujet par le biais de l'interprétation (DRIESEN, 2016). Hormis ces travaux, nous n'avons pas rencontré de recherches qui se seraient intéressées à la matière textuelle produite par les **auditions** d'**enquête**. Or, l'**audition** d'**enquête**, par la quantité d'information qu'elle introduit dans le dossier, les conditions linguistiques de sa réalisation, et la singularité des pièces qu'elle produit, mérite une attention toute particulière. Nous avons tâché de l'étudier et d'en proposer une définition en tant que genre textuel au sein du chapitre V afin de cerner les enjeux à la fois pratiques et théoriques de cette technique

d'enquête.

### 2.2.3 Procès-verbaux des actes d'enquête

Sous la catégorie des **procès-verbaux** d'actes d'enquête, nous rassemblons différents types de documents : les comptes-rendus de **perquisition**, de porte-à-porte, de ratissage (fouille d'une zone à la recherche d'éléments matériels), de **tapissage**, les filatures, les écoutes téléphoniques, les contrôles d'identité, les relevés de plaques de véhicules... Dans ce type de documents, les enquêteurs indiquent toute opération menée dans l'objectif de la recherche d'éléments ou de témoignages, ou consignent des faits qu'ils ont constatés pouvant intéresser les investigations. Ces **procès-verbaux** peuvent aller de une à plusieurs pages selon les opérations rapportées.

L'exemple de la figure 7 présente les éléments typiques du **procès-verbal** avec les en-têtes suivies du récit des opérations menées par l'auteur, qui est seul à signer le document. Les propos éventuellement rapportés par des **témoins** le sont au discours indirect, contrairement aux **auditions**. Le contenu est textuel et varié, allant de phrases complètes rédigées à des listes d'éléments d'intérêt (identités, immatriculations de véhicules, etc.).

### 2.2.4 Procès-verbaux de synthèse

Nous avons déjà évoqué les **procès-verbaux** de synthèse dans le chapitre I section 5.4.2. Leur rédaction n'est pas réservée au stade de l'**analyse criminelle**, mais peut intervenir à d'autres moments de l'**enquête**, par exemple à la clôture de l'**enquête de flagrance**. Le **procès-verbal** de synthèse fait le bilan des investigations réalisées et des moyens mis en œuvre au moment de la rédaction et produit des hypothèses de travail appuyées sur les éléments référencés en **procédure**, comme l'illustre l'extrait page 38.

<b>GENDARMERIE NATIONALE</b>		<b>PROCEDURE SUR COMMISSION ROGATOIRE PROCES - VERBAL D'AUDITION</b>	N° Pièce <i>05/AL</i>	N° Feuillet 1
Compagnie ou escadron SECTION RECHERCHE				
Unité				
C.U.	Procès Verbal			
	<b>907/07 BT</b>			
(ANALYSE ET REFERENCES)				
<b>Audition</b> _____ <b>– Père de</b> _____ <b>et</b> _____ <b>dont</b> _____ <b>avait la charge</b>				
Le _____ à dix heures et trente minutes,				
Nous soussigné Mdl-Chef _____ de la SR _____ et Gendarme _____, de la BR _____, Officiers de police judiciaire,				
Vu les articles 16 à 19 et 151 à 155 du Code de Procédure Pénale.				
Nous trouvant à _____, 10 rue _____, au _____, rapportons les opérations suivantes :				
<b>RENSEIGNEMENTS SUR LA DELEGATION</b>				
Date et numéro - NOM et fonction du magistrat				
N° _____, en date du _____ délivrée par Madame _____, Vice-présidente chargée de l'Instruction au TGI de _____				
Information ouverte contre :				
X				
Pour (qualification des infractions) :				
Homicide volontaire				
Transmission :				
Mission :				
<b>ENQUETE</b>				
Nous faisons comparaître devant nous, le témoin ci-après nommé et lui donnons connaissance des faits pour lesquels sa déposition est requise. ----				
Après avoir prêté serment de dire toute la vérité, rien que la vérité, le témoin, entendu séparément dépose ainsi qu'il suit :				
Je me nomme: _____				
née le _____ à _____ (FRANCE). Nationalité française. Célibataire				
Filiation: née de _____ et _____				
Je demeure: 10 rue _____ à _____ (FRANCE) et 18 rue de _____ à _____, chez sa soeur _____, tél: 06. _____				
Profession : Peintre en bâtiment				
---Je suis la père de _____ née le _____, d' _____ née le _____ et de _____ né le _____				
<b>Concernant ses filles _____ et _____ :</b>				
<b>Question:</b> Pouvez-vous nous donner les raisons du placement de vos deux filles, _____ et _____ ?				
<b>Réponse:</b> Comme ma femme a eu un accident de scooter le _____ et qu'elle est hospitalisée et que moi je n'ai pas encore reconnu les enfants, j'ai téléphoné à une assistante sociale de _____ pour trouver une solution. D'autant plus qu'avec mon travail je ne peux pas m'en occuper. Je suis passé avec les filles au centre médico-social de _____ le _____, il me semble. J'ai vu M. _____ et une assistante sociale. Ils m'ont proposé le placement dans une famille d'accueil. J'ai dit que ma femme en aurait pour 1 an de convalescence, et on m'a dit que les filles seraient placés pour une durée minimum de 1 an. ----				
Pour moi c'était une bonne solution, au moins les enfants étaient placés. Nous avons un fils, _____, qui est également placés. C'est _____ qui avait demandé que _____ soit placé, car elle avait des problèmes pour s'en occuper, car il était très actif et ne faisait que des conneries. ----				
J'ai assisté à la décision de placement du Juge au Tribunal à _____ . Il me semble que c'était salle 122 ou 126. M. _____ était également présent ainsi que l'assistante sociale de _____, Mme _____, ----				
La personne entendue			Les OPJ	
_____			_____	

FIGURE 6 – Exemple d'audition de témoin

<b>GENDARMERIE NATIONALE</b> <small>GROUPEMENT</small> [REDACTED] <small>UNITE</small> [REDACTED] <small>BT</small> [REDACTED] <small>C.U.</small> : [REDACTED] <small>PV N°</small> 905/2000	<b>PROCEDURE SUR COMMISSION ROGATOIRE</b>  <b>PROCES-VERBAL DE RENSEIGNEMENT</b>	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 60%;"><small>Pièce</small> 21/1</td> <td style="width: 40%;"><small>Feuille</small></td> </tr> </table>	<small>Pièce</small> 21/1	<small>Feuille</small>
<small>Pièce</small> 21/1	<small>Feuille</small>			

Nous soussignés, [REDACTED], Gendarme, Officier de police judiciaire,  
 en résidence à la brigade de [REDACTED]

Vu les articles 16, 17 à 19, 151 à 155 du Code de Procédure Pénale,  
 Rapportons les opérations suivantes que nous avons effectuées en exécution de la commission rogatoire jointe  
 délivrée en application de l'article 18 alinéa 4 du Code de procédure pénale.-----  
 Nous trouvant à [REDACTED]. Après en avoir informé le magistrat mandant à [REDACTED] ainsi que l'officier de police  
 judiciaire de territorialement compétent :-----

● **RENSEIGNEMENTS SUR LA DELEGATION**      Date [REDACTED]      N° 200/0068

Nom et fonction du magistrat :  
 Mme [REDACTED], Vice-président chargée de l'instruction au tribunal de  
 grande instance de [REDACTED]

---

Vu l'information ouverte pour :  
 Homicide volontaire

---

A l'encontre de :  
 .....X.....

---

● **TRANSMISSIONS :**      Date      N°

Grade, Nom, Prénom :  
 Fonction :

---

● **MISSION : VOIR COMMISSION ROGATOIRE JOINTE**

Le mercredi [REDACTED]

---Dans le cadre de la commission rogatoire de référence, nous nous portons à [REDACTED] à la  
 brigade de Gendarmerie afin de contrôler un individu ayant une ressemblance avec le portrait robot  
 n° 5.-----

---L'intéressé M. [REDACTED], commerçant à [REDACTED] se présente à notre unité à  
 [REDACTED]. Nous procéderons à son audition ce même jour. Il sera de même effectué avec son accord  
 à un tapissage puis sur prescription du juge d'instruction à un prélèvement sanguin.  
 (CF pièce 21/1, 21/2, 21/4).-----

---Le témoin, [REDACTED], ne reconnaît pas [REDACTED] qui par ailleurs elle  
 connaît en temps que personne demeurant sur la commune de [REDACTED].-----  
 (CF pièce 21/3)

---Le cousin de [REDACTED], [REDACTED] sera également entendu dans  
 le cadre de la présente vérification.  
 (CF pièce 21/6).

---Le [REDACTED] nous sommes destinataires des résultats d'analyse de l'ADN de  
 [REDACTED]. La comparaison est négative.

---Fait et clos à [REDACTED] le [REDACTED]

*L'Officier de police judiciaire*

FIGURE 7 – Procès-verbal de renseignement

**- Les hypothèses de travail :**

Le suicide ayant été écarté d'emblée, les premières hypothèses de travail ont pu être échaffaudées à l'aide des témoignages puis sur les axes habituels d'investigations d'enquêteurs.

[...]

**HYPOTHESE 2 :**

Ces derniers ont quitté les lieux à 13H00 en emportant des sacs et en utilisant leur voiture. Ce départ pouvait paraître suspect d'autant qu'ils savaient que la jeune fille était recherchée. Ils sont en outre des témoins importants quant à l'emploi du temps de la victime. - Cf pièces n°15, 16, 26, 47, 48, 49, 50 -

**2.2.5 Rapports d'expertise**

Les rapports d'expertise sont obtenus sur **réquisition**. Différents types de rapports d'expertises peuvent figurer dans la **procédure** avec pour objectif de fournir des éléments de preuve scientifique : dans les cas d'homicides, les examens médico-légaux comme l'**autopsie** permettent de dater le décès, d'en identifier la cause, ainsi que de récupérer des **traces** biologiques du ou des auteurs. Les expertises psychiatriques dressent le profil psychologique d'un individu, permettant notamment de décrire sa personnalité ou de déterminer sa responsabilité pénale.

Les expertises complémentaires portent sur les prélèvements biologiques lors de l'autopsie, comme le prélèvement des organes pour procéder à des analyses biologiques (**traces** de sperme, de sang, d'**ADN**), toxicologiques (alcool, produits stupéfiants). Les **scellés** collectés sont aussi analysés à la recherche de ce type de **traces**. Les analyses sont documentées par des rapports versés au dossier, éventuellement agrémentés de planches photographiques (notamment pour les **autopsies**), des documents qui se distinguent de ceux que nous avons décrit jusqu'à présent. Comme le montrent les extraits aux figures 8 et 9, ils relèvent d'une pratique professionnelle spécifique et convoient une information technique dont la compréhension nécessite des compétences médicales ou des connaissances en chimie et biologie. Ces rapports présentent souvent des résultats sous forme de tableaux et de listes de données chiffrées.

### I - TECHNIQUES

- **La recherche de l'alcool éthylique** a été réalisée sur l'échantillon de liquide séro-hématique par chromatographie en phase gazeuse avec un système d'injection par espace de tête (Headspace) de l'alcool volatilisé dans un flacon scellé, séparation sur colonne capillaire HP Wax et détection par ionisation de flamme en utilisant le propanol 1 comme standard interne, selon la méthode officielle (arrêté du 6 mars 1986). La limite de détection est de 0,01 g/l. Cette technique permet l'identification et la quantification de l'éthanol, du méthanol et de l'isopropanol.

FIGURE 8 – Extrait de rapport d'analyse présentant une technique

### POUMONS

- Les différents fragments pulmonaires ont été débités en coupes semi-sériées
- Les poumons ont perdu leur caractère aérien.
- Les alvéoles et cloisons interalvéolaires ont à peu près complètement disparu car elles sont noyées dans une sérosité hémorragique.
- Quand les septa interalvéolaires sont visibles, ils sont souvent rompus.
- Les bronches et les bronchioles ont conservé leur revêtement de type cylindrique respiratoire et elles sont libres de tout obstacle.
- Les vaisseaux sont tous dilatés et gorgés de sang.

FIGURE 9 – Extrait de rapport d'analyse d'organe

**Autopsies.** La Société Française de Médecine Légale a publié en 2009 des lignes directrices pour la rédaction de rapports d'autopsie (SOCIÉTÉ FRANÇAISE DE MÉDECINE LÉGALE, 2009), dans lesquelles est soulignée la difficulté d'harmonisation face à la diversité des formations et des pratiques. On y préconise une structure en plusieurs rubriques rappelant la situation de découverte, décrivant l'examen externe (figure 10) puis l'examen interne (l'autopsie médico-légale), les prélèvements effectués et enfin la conclusion. Le rapport d'autopsie est rédigé dans une langue faisant appel au vocabulaire médical, tout en apportant en conclusion des interprétations accessibles à un lecteur non-expert (figure 11 et extraits suivants).

- les multiples lésions de la nuque indiquent que la tête de la victime devait alors être fléchie en avant, c'est-à-dire que la victime était à ce moment-là soit décédée, soit sans connaissance ; là encore, cette topographie lésionnelle indique une victime non pas debout, mais assise.

Suite à la levée de corps et aux opérations thanatologiques les Experts

## A/ EXAMEN EXTERNE :

SEXE : Féminin  
 HABITUS : corpulence mince.  
 Corps plein de sable.

### PHÉNOMÈNES CADAVÉRIQUES :

- RIGIDITE : marquée et généralisée.
- LIVIDITES : majoritairement de siège postérieur et très violacées, avec respect des zones d'appui et minoritairement sur la face antérieure du tronc.
- TACHE VERTE ABDOMINALE : non développée.

### TÊTE :

- Sable dans la cavité buccale. Présence d'un cerclage des arcades dentaires indiquant un suivi orthodontique.
- Léger écoulement rosé par les narines, lors de la mobilisation du corps.
- Etat congestif des vaisseaux sous-conjonctivaux avec pétéchies à la partie supérieure des globes oculaires
- Une excoriation superficielle, punctiforme, entre les sourcils..
- Intégrité apparente des os propres du nez.
- Une excoriation punctiforme de 0,3 cm sur la partie gauche du milieu de l'arrête nasale.

FIGURE 10 – Extrait de rapport d'autopsie

évoquent une mort par asphyxie mécanique, engendré par le double mécanisme d'une strangulation et d'une suffocation. Actuellement il leur est impossible de déterminer s'il y a eu agression sexuelle.

### 2.2.6 Factures téléphoniques

Les factures téléphoniques détaillées (couramment appelées « fadet » par les enquêteurs) listent les communications d'un client et d'autres informations intéressantes : pour un client donné, on y voit les numéros utilisés, les numéros contactés avec horodatage, le numéro IMEI<sup>16</sup> du boîtier téléphonique depuis lequel la communication a été passée ou reçue (ce qui permet d'identifier l'utilisation d'un même téléphone par plusieurs numéros), le relais téléphonique par lequel la communication est passée, le sens de la communication et la nature de l'échange. Ces documents étaient encore récemment adressés sur [réquisitions](#) par les opérateurs téléphoniques. En France, depuis avril 2017 une agence du ministère de la justice spécialisée dans

16. Le numéro IMEI est un numéro d'identification des appareils de téléphonie mobile.

## CONCLUSION

1°/ Il s'agit du corps de ██████████, née le ██████████, découverte sans vie, le ██████████, sur ██████████ 2 kilomètres au nord de ██████████ et porté disparue depuis la fin d'après-midi ██████████.

2°/ D'après les Enquêteurs, le corps été recouvert de sable.

3°/ La levée de corps, effectuée le jour même vers 19 heures, par le docteur P. ██████████ permettait de noter : un corps en décubitus dorsal, froid, rigide et portant des lividités très cyanotiques à la face postérieure du corps et paradoxalement sur la face antérieure du tronc, laissant évoquer une mobilisation secondaire du corps. La mort pouvait être estimée comme remontant 10 à 24 heures avant le ██████████ 19 heures.

FIGURE 11 – Extrait de conclusion de rapport d'autopsie

IMSI	Date appel	Heure début	Durée	Numéro appelant	Numéro appelé	N° de série du téléphone	Numéro Cellule	Sens d'appel	Nature d'appel
208 ██████████	██████████	11:27:26	00:00:19	3369 ██████████	3365 ██████████	3543 ██████████	208 ██████████	Sortant	Voix
208 ██████████	██████████	13:27:12	00:00:01	3369 ██████████	3361 ██████████	3543 ██████████	208 ██████████	Sortant	Voix
208 ██████████	██████████	13:27:35	00:00:31	3361 ██████████	3369 ██████████	3543 ██████████	208 ██████████	Entrant	Voix
208 ██████████	██████████	14:11:31	00:00:00	3369 ██████████	3362 ██████████	3543 ██████████	208 ██████████	Sortant	SMS
208 ██████████	██████████	14:11:32	00:00:00	3362 ██████████	3369 ██████████	3543 ██████████	208 ██████████	Accusé de réception	SMS
208 ██████████	██████████	14:12:55	00:00:00	3362 ██████████	3369 ██████████	3543 ██████████	208 ██████████	Entrant	SMS

FIGURE 12 – Extrait de facture téléphonique détaillée

l'interception des communications numériques a été mise en place, l'Antenj. Cette plateforme pilote la plate-forme nationale des interceptions judiciaires (PNIJ), un portail qui permet aux enquêteurs d'obtenir des informations liées aux télécommunications via une demande en ligne, dont entre autres les factures détaillées. En plus de faciliter l'accès aux données de télécommunication, la mise en place de cette plateforme a permis l'uniformisation des formats de documents communiqués par les opérateurs, et ainsi le développement d'utilitaires de traitement des données issues des factures téléphoniques, ce qui n'était pas le cas auparavant.

La figure 12 est un extrait de facture téléphonique, où l'on peut voir que les factures sont structurées sous la forme de tableaux de données où chaque colonne correspond à un type d'information et chaque ligne à une communication.

### 2.2.7 Relevés bancaires

Les relevés bancaires sont fournis sur [réquisition](#) par les établissements bancaires afin de constater les mouvements de fonds : virements, salaires, prestations sociales,

20.01	Carte	Air France Roissy Cdg Cede	16/01	-245,52	<input type="checkbox"/>
20.01	Carte	Amazon Premium PARIS	18/01	-49,00	<input type="checkbox"/>
20.01	Carte	SNCF Paris	19/01	-3,40	<input type="checkbox"/>
22.01	22.01	Avoir	Carte Amazon Premium	21/01	49,00 <input type="checkbox"/>
22.01	Carte	Match Dac	21/01	-55,00	<input type="checkbox"/>
23.01	23.01	Prlv	Free Mobile	-19,99	<input type="checkbox"/>
27.01	27.01	Carte	Apple Itunes Store-eu	24/01	-17,99 <input type="checkbox"/>
27.01			25/01 11H42	-20,00	<input type="checkbox"/>

FIGURE 13 – Extrait de relevé bancaire. La mention « RET DAB » indiquant un retrait au distributeur automatique a été surlignée par l’enquêteur (dernière ligne), suivie de la localité du retrait (camouflée).

retraits d’espèces, émissions de chèques, pensions, etc. Les retraits en espèces permettent par exemple de vérifier la présence géographique d’une personne à un moment donné, et les transferts d’argent attestent de contacts entre personnes. On peut également identifier des dépenses suspectes, comme des achats d’objets potentiellement impliqués dans les faits (figure 13).

Les relevés bancaires présentent sur une ligne la date, la nature de l’opération, son objet et le montant entrant ou sortant. Tout comme les données téléphoniques, ce sont des documents structurés sous forme de tableaux, mais à l’inverse, les formats ne sont pas uniformisés entre établissements bancaires.

### 2.2.8 Autres documents versé au dossier de procédure

Toutes sortes d’autres documents peuvent être versés au dossier de [procédure](#) par les enquêteurs. Nous avons relevé une très grande variété de documents, souvent ajoutés en annexe de [procès-verbaux](#) :

- Avis de recherche,
- Lettres anonymes, de revendication,
- Documents administratifs : actes d’état-civil, permis d’inhumér, copies de livrets de famille, dossiers pénitentiaires,
- Coupures de presse,
- Correspondance avec les représentants des parties : notamment les demandes d’actes,
- Documents fournis par les services de police municipale ou des sociétés de sécurité : compte-rendus de ronde par exemple,
- Documents commerciaux : fichiers clients,

- Documentation diverse, catégorie au sein de laquelle nous avons rencontré de la documentation sur des sectes et des éléments bibliographiques sur des techniques d'enquête.

Cette liste, bien entendu, est loin d'être exhaustive. N'importe quel document peut être ajouté au dossier pourvu qu'il puisse intéresser l'enquête.

### 2.2.9 Documents photographiques, vidéos ou graphiques divers

De nombreuses étapes des investigations sont documentées par des photographies et des vidéos : prises de vue de la scène de crime et de ses environs (y compris des vues aériennes), photos du cadavre, photos de l'autopsie, photos de véhicules, de scellés, d'éléments matériels, portraits de personnes suspectes, photographies de tapissage, photographies prises lors de reconstitutions, etc., portant parfois des annotations ou des mentions. En plus de ces photos visant à documenter l'enquête, des photos issues d'appareils appartenant à des personnes impliquées peuvent être collectées comme pièces à conviction. Elles sont soit intégrées à des documents sous forme de planches photographiques, soit rassemblées dans des répertoires informatiques.

Les portraits-robots sont des représentations graphiques du visage de personnes recherchées, établies d'après une description fournie par un témoin. De nos jours, ils sont réalisés à l'aide de logiciels spécialisés qui permettent d'assembler des traits physiques.

Pour les vidéos, on trouve les enregistrements des gardes à vue<sup>17</sup> et de certaines auditions, ou encore des extraits de caméras de vidéo-surveillance.

Enfin, des cartes type IGN ou état-major peuvent être ajoutées au dossier, tout comme des plans manuscrits ou des schémas produits par des témoins.

Les documents graphiques, de natures variées, apportent des informations complémentaires aux autres documents de la procédure. Leur fonction peut être illustrative (plans, cartes, photos des lieux et des faits), légale (captation des auditions),

---

17. Les conditions d'enregistrement sont précisées aux articles 64-1 <https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006071154&idArticle=LEGIARTI000006575093&dateTexte=&categorieLien=cid> et D15-6 <https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006071154&idArticle=LEGIARTI000006514842&dateTexte=&categorieLien=cid> du Code de procédure pénale. [consultés le 25 juin 2019]

d'investigation ([portraits-robots](#)), ou être des éléments de preuve en tant que tels (vidéo-surveillance).

### 2.3 Synthèse

Le volume de chaque type de document au sein de la [procédure](#) est dépendant de l'affaire traitée. Les pratiques d'[enquête](#) ont recours à l'analyse des télécommunications et des mouvements de fonds en ce qui concerne le trafic de stupéfiants. Dans ces cas, les données téléphoniques et bancaires représenteront donc un volume de documents important au sein du dossier de [procédure](#), alors que dans d'autres types d'affaires, les [auditions](#) auront la place centrale. Les autres types de documents, notamment les rapports d'expertise, sont moins nombreux et ne sont pas la source principale de l'information exploitée par l'[analyse criminelle](#).

Dans cette section, nous avons présenté un panorama non-exhaustif des documents qui constituent le dossier de [procédure](#) judiciaire traité en [analyse criminelle](#). Nous avons pu constater une grande variété dans la nature des documents, aussi bien en ce qui concerne leurs conditions de production, leurs fonctions, leurs formes et leurs émetteurs. À propos des émetteurs, il faut noter pour la poursuite de notre réflexion la multiplicité des acteurs impliqués dans la production des documents, membres de l'institution judiciaire ou civils réquisitionnés : enquêteurs, techniciens, magistrats, médecins, experts... Cet aspect fait de l'élaboration du dossier de [procédure](#) un processus collaboratif orchestré par la direction de l'[enquête](#).

Malgré ces aspects, tous les documents de la [procédure](#) visent à clarifier les faits investigués. Ils partagent donc au moins, dans les grandes lignes, une finalité commune. Dans la suite de ce chapitre, nous essaierons de déterminer les aspects qui rapprochent le dossier de [procédure](#) d'un [corpus](#) textuel, ce qui permettra par la suite d'envisager l'adaptation des méthodes de la linguistique de [corpus](#) à la pratique de l'[analyse criminelle](#).

### 3 Le texte de la procédure judiciaire

La description de la matière rassemblée dans le dossier de **procédure** judiciaire, remise dans la perspective des besoins de l'**analyse criminelle** présentée au chapitre I, conduit à envisager les directions suivantes :

- Le développement d'approches automatiques d'**extraction d'information**,
- À terme, la conception d'une méthodologie d'accès au texte inspirée de la linguistique de **corpus** et de l'analyse des données textuelles (**ADT**).

Nous concevons ces deux directions comme complémentaires : il s'agirait d'élaborer un outil d'exploration textuelle *ad hoc* pour les **analystes criminels**, impliquant, entre autres possibilités d'application de technologies de traitement automatique des langues (**TAL**), un repérage automatique de certaines informations et permettant une navigation facilitée au sein des textes.

Afin de situer nos données de travail au regard de ces deux dernières disciplines, il convient de s'attarder sur un concept qui leur est fondamental : celui du **corpus**. Toute archive documentaire ne constituant pas nécessairement un **corpus**, nous souhaitons déterminer si et selon quelles modalités le dossier de **procédure** judiciaire se prête à l'application des méthodologies en question. Pour cela, nous considérerons la (les) définition(s) du **corpus** selon la linguistique de **corpus** ainsi que deux types de **corpus**.

#### 3.1 Le concept de corpus selon la linguistique de corpus

La linguistique de **corpus** est une branche de la linguistique qui prend comme objet d'étude des **corpus**, c'est-à-dire des ensembles de textes authentiques, au contraire d'autres branches de la linguistique qui travaillent à des niveaux différents, par exemple celui de la phrase (l'énoncé), et sur des exemples construits pour la démonstration (artificiels).

Avec le développement de méthodes informatiques appliquées au texte, les pratiques de la linguistique de **corpus** comme le relevé d'occurrences, les concordances, ou l'application de méthodes statistiques comme l'analyse factorielle des correspondances (**AFC**) (BENZÉCRI, 1973; SALEM, 1986; LEBART et SALEM, 1994; HABERT et al., 1997) ont été décuplées. Aujourd'hui, de nombreux logiciels et outils s'adressent

aux linguistes plus ou moins familiers de l'informatique pour leur offrir des voies d'entrées dans les **corpus**. Le développement technique s'est accompagné d'une réflexion théorique sur la constitution des **corpus**, les modes d'approche (*corpus-based* ou *corpus-driven*), la validité des pré-traitements (annotations, lemmatisation), la représentativité des résultats, etc. Les pratiques de la linguistique de **corpus** et leur potentiel dans notre cas d'étude seront plus amplement présentés dans le chapitre III.

L'un des grands questionnements de la linguistique de **corpus** et qui nous intéresse tout particulièrement réside dans la définition même du **corpus** : SINCLAIR (1991) le définit comme

A collection of naturally occurring language text, chosen to characterize a state or variety of a language,<sup>18</sup>

MAYAFFRE (2002) renvoie à la littérature spécialisée et à une définition encyclopédique selon laquelle un **corpus** est

un rassemblement de textes ou une collection de textes regroupés sur la base d'hypothèses de travail en vue de les interroger

À propos de la structure du **corpus**, RASTIER (2004) oppose conception documentaire à conception philologique-herméneutique, et propose la définition suivante :

Un **corpus** est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications.

GARRIC et LONGHI (2012) indiquent qu'il n'existe pas de consensus sur la notion de **corpus** tout en proposant la définition suivante :

Un **corpus** est défini comme un ensemble raisonné de textes, structuré par une cohérence interne.

De ces définitions, on retient l'idée d'un **rassemblement** de **textes** qui n'est pas le fruit du hasard puisque le **corpus sert une intention** et répond à une **structure interne**.

---

18. « Une collection de textes linguistiques naturels, choisis pour caractériser un état ou une variété d'une langue » (traduction libre)

Ces définitions presque dictionnairiques sont complétées dans la littérature par des critères de bonne formation parmi lesquels on rencontre homogénéité, contrastivité, diachronicité selon MAYAFFRE (2002), signifiante, acceptabilité, exploitabilité selon PINCEMIN (2012a). L'objectif de ces critères est de légitimer les observations du chercheur en garantissant la qualité de la production du *corpus* exploité : le critère d'acceptabilité par exemple évalue entre autres la représentativité des données considérées afin d'objectiver au maximum les observations.

### 3.2 Corpus réflexifs, corpus hétérogènes

Parmi les différents *corpus* impliqués en linguistique de *corpus*, deux types en particulier attirent notre attention : les *corpus* réflexifs et les *corpus* hétérogènes.

Le concept de *corpus* réflexif a été établi par D. MAYAFFRE dans son article précédemment cité. La réflexivité du *corpus* tiendrait dans « le fait que ses constituants [...] renvoient les uns aux autres pour former un *réseau sémantique* performant dans un tout (le *corpus*) cohérent et auto-suffisant ». Connectée à l'hypertextualité et à l'intertexte, cette notion mêle la question de la clôture du *corpus* et celle des ressources nécessaires à son interprétation. L'exemple employé par l'auteur est celui du champ des discours politiques, dans lequel les productions verbales s'inscrivent dans un réseau de discours déjà énoncés.

En ce qui concerne l'homogénéité, PINCEMIN (2012a) déclare qu'elle ne doit pas être une condition de prise en charge d'un *corpus*, pourvu que l'on dispose d'une bonne connaissance de ses défauts et que l'on en fasse un usage éclairé, ce que l'auteur appelle le critère d'*interprétabilité*. L'intérêt de la proposition est de dépasser les critères de bonne formation d'un *corpus* en posant un cadre pour l'étude de *corpus* imparfaits. Deux formes d'hétérogénéité sont rapportées : des aspects techniques tels que l'encodage des fichiers ou les questions d'échantillonnage, et des aspects de variation linguistique liés, notamment, au genre textuel.

### 3.3 Et la procédure ?

L'hétérogénéité des composants du dossier de *procédure* est flagrante à la lecture de la première section de ce chapitre : les contenus, genres, fonctions, volumes

et émetteurs variés des documents génèrent une hétérogénéité sur les plans aussi bien techniques que linguistiques. Que ce soit à propos de la méthode de l'[analyse criminelle](#) ou de notre question de recherche (pour rappel, l'adaptation d'approches automatiques à la pratique de l'[analyse criminelle](#)), cette forte hétérogénéité des documents et du texte est un aspect crucial à prendre en compte dans notre perspective de recherche, car c'est aux outils et aux méthodes de s'adapter aux données et pas l'inverse (PINCEMIN, 2012a).

L'aspect réflexif est moins évident mais néanmoins intéressant à souligner car il se retrouve sur deux plans.

Tout d'abord, comme nous l'avons décrit dans le cas des [procès-verbaux](#) de synthèse par exemple (en 2.2.4), les pièces de [procédure](#) peuvent référer explicitement les unes aux autres en mentionnant leurs cotes. On retrouve ainsi le critère de ressource d'interprétation interne : une information est avancée sur la base d'un autre passage du [corpus](#), les informations peuvent être interprétées en regard les unes des autres. L'aspect d'auto-suffisance est même prévu par la loi puisque l'[analyse criminelle](#) et plus largement l'action judiciaire reposent sur l'exploitation des éléments actés dans le dossier de [procédure](#) exclusivement.

D'autre part, les [procès-verbaux](#), dans leurs en-têtes légaux, font mention des textes de loi cadrant la production du document (exemple à la figure 7 : « Vu les articles 16, 17 à 19, 151 à 155 du [Code de procédure pénale](#) »). Le texte de loi est la ressource interprétative des aspects légaux du document, que chaque partie impliquée dans sa constitution est, théoriquement, censée connaître puisque nul n'est censé ignorer la loi. Les textes de loi ne sont pas inclus à proprement parler dans le dossier de [procédure](#). Mais est-il nécessaire et pertinent d'intégrer la loi, texte public par excellence, référencé et structuré, que nul n'est censé ignorer, et surtout pas les professionnels de la justice, au dossier de [procédure](#) ?

Réflexif et hétérogène, il reste néanmoins à caractériser le caractère de [corpus](#) du dossier de [procédure](#). Si l'on reprend les éléments définitoires énoncés à la sous-partie 3.1 :

- Rassemblement : comme archive ou comme [corpus](#), le dossier de [procédure](#) compile des documents,

- Textes : tous les documents ne sont pas des textes, mais une partie non-négligeable en est,
- Finalité : tous concourent à un même objectif, celui de faire la lumière sur des faits criminels,
- Structure : si elle manque de normalisation, la structure du dossier de **procédure** n'est pas inexistante pour autant. Au minimum, elle correspond au découpage en pièces de **procédure**.

Globalement, le dossier de **procédure** remplit les critères définitoires du **corpus** précédemment dégagés. La finalité est un aspect qui diverge des définitions du **corpus** produites par la linguistique textuelle, puisqu'elle n'est pas d'étudier des phénomènes linguistiques. On peut la réajuster dans la perspective de notre recherche, dans ce cas, la finalité du dossier de **procédure** devient d'être un support de développement d'approches de **TAL** et de description des phénomènes linguistiques de l'**enquête**. Ceci posé, nous pouvons nous intéresser plus précisément aux manipulations et pré-traitements nécessaires avant son utilisation pour notre recherche.

#### 3.4 Constitution du corpus de recherche

Est dit *textuel* ce qui procède d'une langue, se transcrit, s'articule en lettres et en mots dans une écriture, par opposition à ce qui relève d'autres médias : les images, les sons. (PINCEMIN, 1999, p. 135)

Dans le jargon des bases de données, les champs textuels s'opposent aux champs factuels et numériques. Ce qui est factuel, c'est ce qui prend sa valeur parmi un ensemble donné d'alternatives (vrai / faux, codes départementaux, date, répertoire de noms d'auteurs, etc.). Pour ce qui est textuel, il n'y a pas de liste de possibilités prévues, la seule contrainte est en général une longueur maximale. (*ibid.* p. 136)

Resituons notre recherche : les **procédures** que nous avons à notre disposition sont donc, de notre point de vue, à envisager comme un **corpus** que nous allons employer comme « bac à sable » de test et d'adaptation de méthodes de **TAL**, en

orientant nos travaux par la réalisation d'une étude des textes de la [procédure](#) judiciaire, notamment des réalisations linguistiques des entités qui intéressent l'[analyse criminelle](#).

Nous sélectionnerons une partie de la [procédure](#) pour remplir ce rôle, afin de retenir les documents qui s'y prêtent le mieux, aussi bien en termes d'exploitabilité immédiate que de l'intérêt des informations qu'ils comportent.

Il ressort de la présentation des pièces de [procédure](#) réalisée lors de la première partie de ce chapitre que :

- Tous les documents ne sont pas des textes (documents graphiques),
- Certains documents sont des extractions de bases de données (documents téléphoniques et bancaires),
- Certains documents sont des textes ou comportent du texte, mais leur statut de documents réglementaires réduit leur apport en information et leur confère une forme rigide et procédurale (bordereaux d'envoi, inventaires des pièces à conviction, [procès-verbaux](#) de notification de [garde à vue](#), [réquisitions](#), certificats médicaux de [garde à vue](#), [procès-verbaux](#) de saisie),
- Certains consignent des pans de texte libre ([procès-verbaux](#) transport constatations mesures prises, synthèse, [actes d'enquête](#), [auditions](#), rapports d'expertise).

La dernière catégorie, celle des documents comportant des pans de texte libre, est donc celle qui paraît le mieux convenir à notre objectif, néanmoins, le volume de chaque type de document en son sein est variable. Pour une affaire donnée, on trouvera un [procès-verbal](#) transport constatations mesures prises, une ou deux synthèses selon le stade de l'[enquête](#), un nombre indéterminé d'[actes d'enquête](#) et d'[auditions](#). En ce qui concerne les expertises, leur nature et leur nombre sont imprévisibles, et bien qu'elles comportent des pans de texte, une partie de leur contenu peut aussi prendre la forme de tableaux de données et d'images, ce qui ne correspond pas à nos contraintes. Le document d'expertise le plus exploitable serait le rapport d'autopsie, mais quel que soit le nombre de victimes de l'affaire étudiée, il sera toujours vraisemblablement inférieur au nombre de [témoins](#) auditionnés. Dans les [procédures](#) dont nous disposons, chaque affaire comporte une autopsie, ce qui ne représente pas un volume suffisant pour être traité selon les méthodes que nous envisageons.

L'étude des rapports d'expertise criminalistique, qui n'est pas dénuée d'intérêt, demanderait la constitution d'un **corpus** spécifique de documents collectés à travers différentes affaires.

Dernier critère décisif : les **analystes criminels** rapportent se baser en grande partie sur le contenu des **auditions** pour produire leurs analyses. Ce type de document représente en effet un volume conséquent de la **procédure** (dans nos exemples, de plusieurs centaines à plusieurs milliers de documents), convoie une information dense, non-structurée, et porte des enjeux linguistiques très spécifiques qui méritent une étude approfondie.

En conséquence, nous retenons la catégorie des **auditions** pour la poursuite des travaux. Cette catégorie catalyse nos différents intérêts de recherche. D'une part, les « aspects **TAL** » sont satisfaits en constituant une quantité de données pertinente et en rapportant les informations qui intéressent l'**analyse criminelle**, et d'autre part, les « aspects linguistiques » sont remplis car les **auditions** constituent un type de texte singulier dont la description et la compréhension pourront alimenter le développement des méthodes informatiques.

Parmi les **procédures** disponibles pour nos recherches, une seule présente une qualité de reconnaissance de caractères exploitable, c'est-à-dire ne nécessitant pas de correction et permettant son traitement informatique. Cette **procédure** contient 370 **auditions**, qui feront désormais office de **corpus** de travail.

## 4 Conclusion

Dans ce chapitre, nous avons détaillé la composition du dossier de **procédure** judiciaire exploité par l'**analyse criminelle** d'après des exemples d'affaires réelles.

L'enjeu était de structurer l'organisation de la **procédure** en types de documents, puis d'en cerner les aspects textuels afin de délimiter à la fois le futur champ d'application des outils et les ressources nécessaires à leur développement. Cette phase de description a démontré la singularité du dossier de **procédure** judiciaire, rassemblant des documents de formes et de fonctions très variées, émanant d'un grand nombre d'acteurs différents, mais qui sont tout de même tous produits et consignés dans un seul et même but. La variété des documents, les volumes et natures de texte

qu'ils contiennent imposent un tri à la fois pour notre étude et pour l'intégration des méthodes automatiques de traitement, pour proposer dans un premier temps des observations et des pistes de développement. Guidée par l'ajustement entre l'intérêt, la disponibilité et l'exploitabilité des données, nous faisons le choix d'utiliser un type particulier de documents, les [auditions](#) de [témoins](#).

Maintenant que contexte et matière ont été présentés en détails, nous consacrons le chapitre suivant à constituer un état de l'art adapté à nos préoccupations, c'est-à-dire aux questions de linguistique outillée et d'extraction automatique d'information que nous considérerons dans la perspective de l'[analyse criminelle](#).

## Chapitre III

# État de l'art

La présentation du contexte de recherche et des données de travail nous a permis de constater que notre problématique se situe à la frontière entre plusieurs domaines : la linguistique et l'informatique, rassemblés dans le domaine établi du traitement automatique des langues (TAL) d'une part, et l'analyse criminelle d'une autre part, en tant que domaine d'application des deux disciplines précédentes<sup>1</sup>.

À travers la présentation de la matière exploitée, l'application de la problématique s'est resserrée sur un type de texte en particulier, les **auditions**, qui présentent un intérêt particulier pour les **analystes criminels** et serviront donc de **corpus** de d'étude et de développement.

Comme on le verra dans ce chapitre, l'état de l'art du traitement des données textuelles dans le domaine criminel est peu fourni, nous inscrivons donc notre travail comme une première étape d'un processus plus large qui vise à concevoir et développer des outils informatiques et linguistiques pour l'**analyse criminelle**. Nous concentrerons notre attention sur un angle en particulier : la reconnaissance d'**entités nommées**. En effet, cet angle nous permet de coupler un concept clé de l'**analyse criminelle** et un sous-domaine du TAL faisant l'objet de recherches depuis de nombreuses années. Le concept d'**entités nommées** du TAL et celui d'**entités** de l'**analyse criminelle** seront comparés afin de vérifier leur recoupement et les implications techniques et linguistiques à prendre en compte pour de futurs développements, tout en

---

1. Cette manière d'envisager l'analyse criminelle, et plus largement les sciences criminelles comme domaine d'application peut faire l'objet d'un débat qui se connecte aussi à la constitution des sciences criminelles comme domaine de recherche scientifique propre. Étant donné notre arrière-plan scientifique, nous estimons que notre situation est celle de la mise en application de connaissances du domaine de la linguistique informatique au cas de l'analyse criminelle, d'où une telle formulation.

interrogeant la pertinence conceptuelle et l'apport méthodologique de la détection automatique d'entités nommées en analyse criminelle.

## 1 Le texte, l'ordinateur, l'humain

Ce travail de recherche trouve sa source dans l'intérêt de l'analyse criminelle pour l'apport des méthodes informatiques afin d'améliorer ses conditions et moyens de travail, une préoccupation qui rappelle celle des sciences humaines et sociales ayant mené au développement du champ des humanités numériques.

Notre cas d'étude illustre en effet les liens entre le texte, l'ordinateur et l'humain. C'est en cela que nous rapprochons notre sujet de recherche de celui des humanités numériques : le besoin commun de conception de nouveaux outils et de nouvelles méthodes pour l'exploitation de documents textuels ouvre la réflexion sur le rapport triangulaire entre le texte, objet, l'humain, sujet, et l'ordinateur, outil. À ce sujet PINCEMIN (1999, p. 129), dans le cadre d'une thèse visant à concevoir un outil de diffusion de textes dans une entreprise, affirmait déjà :

Quelle place donner à la machine dans l'analyse des textes, quelle aide peut-elle apporter? L'idée d'une *compréhension* automatique est ici rejetée. L'apport de l'ordinateur tient à ses capacités en termes de mémoire, de manipulation systématique et de vitesse de calcul.

L'ordinateur est placé dans son rôle d'instrument, dont les capacités sont exploitées par un sujet afin de manipuler plus efficacement un objet. Toute idée d'interprétation ou de prise de décision « automatique » est exclue de cette définition.

C'est dans une optique similaire que nous souhaitons placer nos recherches pour l'analyse criminelle : actuellement, les analystes criminels manipulent le texte « manuellement », c'est-à-dire via l'explorateur de fichiers en ouvrant et parcourant les fichiers un à un, ou avec des outils de recherche limités de type moteur de recherche de bureau ou le raccourci de recherche CTRL + F. La phase d'extraction et d'interprétation des informations est effectuée entièrement par les analystes criminels. L'ordinateur doit être intégré comme outil de manipulation, alors que le travail d'analyse doit être laissé aux analystes criminels.

Ainsi, l'objectif de notre recherche étant d'approfondir l'utilisation de l'ordinateur comme outil de gestion du texte dans le rapport entre l'analyste et les documents de la *procédure*, nous présentons dans les sous-parties suivantes deux champs de recherche en sciences humaines et sociales confrontés à une ambition similaire, les humanités numériques d'une part, et la linguistique de *corpus* et la *textométrie* d'autre part, ainsi que les stratégies et méthodologies qu'ils ont adoptées pour y parvenir.

### 1.1 Humanités numériques

On fait habituellement remonter l'émergence des humanités numériques (*digital humanities* en anglais) au projet d'*Index* des œuvres de Saint Thomas d'Aquin par R. BUSA (BERRA, 2015 ; CITTON, 2015). Cherchant à exploiter les apports des technologies numériques aux sciences humaines et sociales, leur définition et leurs frontières sont sujettes à discussion, néanmoins leur préoccupation principale peut être dégagée comme étant celle de l'étude, de la gestion et de la transmission des savoirs et connaissances dans une ère de l'information globalisée et électronique, ou pour le dire très simplement, comment combiner sciences humaines et sociales et technologies informatiques et de l'information. Citons comme définition celle de DACOS (2010) :

Les *digital humanities* désignent une transdiscipline, porteuse des méthodes, des dispositifs et des perspectives heuristiques liés au numérique dans le domaine des Sciences humaines et sociales.

En France, ce *Manifeste des Digital Humanities* fait office de référence dans la définition des humanités numériques et de leurs orientations, en faisant un état des lieux de la situation et en posant les orientations souhaitées pour leur avenir, notamment leur aspect ouvert et collaboratif. Depuis lors, les humanités numériques se sont développées en France et en Europe sous la forme de structures et de plateformes (BERRA et al., 2017), de formations universitaires, de groupes et centres de recherches, de revues... dont DACOS et MOUNIER (2015) proposent un panorama.

De la numérisation de parchemins médiévaux à la mise à disposition des savoirs et connaissances pour le grand public, en passant par les pratiques de recherche numériques comme la tenue de carnets de recherches, les multiples facettes des projets en humanités numériques rendent leur illustration concise difficile. Leur développement a amené à nombreux débats sur les rapports entre sciences dites « dures », représentées essentiellement par l'informatique, et les sciences humaines et sociales, avec à la clé la question de la structuration de la collaboration de ces horizons que l'on peut au premier abord croire comme diamétralement opposés. L'enjeu de ces débats est de dépasser une conception des humanités numériques où les ingénieurs et personnels techniques sont envisagés par les porteurs de projet en sciences humaines et sociales comme de simples pourvoyeurs de services informatiques.

C'est en cela que les humanités numériques intéressent notre projet de recherche : leur développement ne se réduit pas au développement de solutions techniques mais s'accompagne de la construction d'une théorie des savoirs et des pratiques numériques (GANASCIA, 2015), ce que MOUNIER (2017) souligne comme étant la particularité du monde de la recherche en comparaison des autres métiers ayant également intégré les évolutions technologiques.

L'analyse criminelle doit se placer de manière semblable : consciente des potentialités des technologies numériques dans la gestion des informations, elle souhaite les exploiter afin d'améliorer ses performances. Mais cette évolution ne peut être que purement technique, elle doit s'appuyer sur une réflexion théorique sur les attentes et les modalités de mise en place de ces solutions. Si l'on reprend GANASCIA (2015) :

The Digital Humanities [...] propose systematic and technologically equipped methodologies in activities where, for centuries, intuition and intelligent handling had played a predominant role.<sup>2</sup>

On constate que l'auteur évoque des *méthodologies systématiques et équipées*, venant en appui d'activités intellectuelles, une configuration dans laquelle se trouve déjà l'analyse criminelle dans une certaine mesure. Pour la progression de cet équipement, il s'agit d'éviter l'introduction d'« usines à gaz » coûteuses en développement,

---

2. « Les Humanités Numériques (...) proposent des méthodologies systématiques et technologiquement équipées dans des activités où, depuis des siècles, l'intuition et la manipulation intelligente ont joué un rôle prédominant. », traduction libre et assistée par le logiciel de traduction automatique en ligne DeepL : <https://www.deepl.com/translator> [consulté le 16 décembre 2019].

difficiles à manipuler et à la pertinence relative, dans une perspective d'expérimentation.

En conséquence, plusieurs parties de ce manuscrit sont consacrées à une étude approfondie des données et des pratiques de l'[analyse criminelle](#), afin de nourrir la réflexion préalable au développement d'outils techniques.

## 1.2 Linguistique de corpus & textométrie

La linguistique de [corpus](#) est une branche de la linguistique étudiant une langue « réelle », par contraste avec une linguistique pratiquée sur des exemples « artificiels », c'est-à-dire construits expressément à titre d'illustration. Pour étudier cette langue réelle, la linguistique de [corpus](#) a recours à des [corpus](#) de données orales ou écrites authentiques (ESHKOL-TARAVELLA et LEFEUVRE-HALFTERMEYER, 2017). Si elle existe depuis le milieu du XX<sup>ème</sup> siècle, elle a connu un regain d'intérêt depuis l'émergence de technologies informatiques qui ont permis non seulement la collecte et la mise en forme de grands volumes de données linguistiques, mais aussi leur traitement automatique et systématique (LÉON, 2005; POIBEAU, 2014a). De plus, l'explosion des moyens de communication numériques a engendré un volume de productions linguistiques (essentiellement textuelles) que la linguistique de [corpus](#) est à même de traiter. En cela, la linguistique de [corpus](#) s'est reliée au mouvement des humanités numériques.

La linguistique de [corpus](#) s'oppose également à une pratique de la linguistique limitée au seuil de la phrase, car elle étudie les textes dans leur ensemble et en prenant en compte leurs structures internes et externes (RASTIER, 2004). Enfin, son ambition, entre autres choses, est d'assurer la validité des observations réalisées par la représentativité statistique des résultats obtenus.

La disponibilité de moyens informatiques a permis à la linguistique de [corpus](#) de se développer et de renouveler ses problématiques : les principales sont liées à la conception et l'adaptation de méthodes statistiques pour le traitement du texte ou encore aux modalités de constitution de [corpus](#) d'étude (questions de collecte des données, de nettoyage, de formatage, de compilation, de mise à disposition de la communauté des chercheurs...).

La linguistique de **corpus** entretient des liens étroits avec la **textométrie**, une approche spécifiquement centrée sur l'analyse de données textuelles (**ADT**), alors que l'objectif de la linguistique de **corpus** est de décrire et modéliser tout type de données langagières. La **textométrie** repose sur un principe différentiel mettant en valeur les points de convergence et de divergence dans les **corpus** (SÖZE-DUVAL, 2008; PINCEMIN, 2012a; PATIN et al., 2016) en se basant sur un découpage préalable en unités (qu'il s'agisse des mots, des **parties du discours** ou encore des **lemmes**) et cherche à développer « de nouveaux modèles statistiques pour rendre compte de caractéristiques significatives des données textuelles » (PINCEMIN et HEIDEN, 2008). La **textométrie**, inscrite dans le champ de l'**ADT**, trouve toutes sortes d'objectifs et d'applications en sciences humaines et sociales<sup>3</sup> : analyses de questionnaires, du discours politique ou médiatique, de **corpus** historiques, d'entretiens psychologiques... PINCEMIN (2012b) synthétise la distinction entre linguistique de **corpus** et **textométrie** ainsi :

Une étude qui utilise une approche et des outils textométriques, et qui vise à observer et décrire des phénomènes linguistiques en **corpus**, peut donc relever à la fois de la linguistique de **corpus** et de la **textométrie**, sans pour autant qu'aucun de ces deux courants ne subsume l'autre.

À titre d'exemple, citons le travail de HO ĐINH (2017), consacré à l'étude de forums de discussion en ligne au sujet du VIH en français et en vietnamien. L'étude est basée sur un **corpus** d'environ 10 millions de mots structurés en quatre sous-parties. Les discours institutionnels (gouvernements, instances de santé) ont été comparés avec les discours informels (forums). L'auteure a également proposé une « cartographie » des forums en rubriques, interventions et catégories de conversations. Cette contribution a permis de caractériser les échanges dans les espaces du web social, tout en proposant une analyse comparative des discours sur le VIH selon plusieurs angles (France/Vietnam, institutionnels/informels).

3. Voir par exemple les actes des Journées d'Analyses des Données Textuelles (JADT) :

— 2014 <http://lexicometrica.univ-paris3.fr/jadt/jadt2014/>,

— 2016 <http://lexicometrica.univ-paris3.fr/jadt/jadt2016/>,

— 2018 <http://lexicometrica.univ-paris3.fr/jadt/JADT2018/actes-jadt18.pdf>

[consultés le 12 décembre 2019].

Quelques exemples de techniques fréquemment utilisées en :

- Les concordanciers permettent l’affichage d’un mot recherché dans son contexte d’apparition (figure 14),
- Les cooccurrences : pour un mot ou une forme donnés, le calcul des cooccurrences fournit les mots ou formes statistiquement sur-représentés dans le même contexte (figure 15),
- Les [segments répétés](#) identifient des suites de formes graphiques apparaissant plus d’une fois (figure 16),
- Les spécificités calculent, pour un [corpus](#) divisé en plusieurs sous-parties, quelles unités sont les plus caractéristiques, statistiquement les plus associées à chaque sous-partie. À la figure 17, les spécificités sont calculées entre quatre mois du quotidien L’Est Républicain (ATILF, 2018) de l’année 1999, classées par mots les plus spécifiques du mois d’août : période à laquelle eut lieu l’éclipse totale de soleil du 11 août 1999 ainsi que la sortie dans les cinémas des films *Wild Wild West* et *Coup de Foudre à Notting Hill*. En comparaison, les spécificités du mois de juillet, à la figure 18, sont marquées par la présence du roi Hassan II en raison d’une visite officielle en France à l’occasion du 14 juillet, ainsi que de son décès neuf jours plus tard le 23 juillet 1999. Quant aux sigles en trois lettres et aux mentions abrégées de pays, il s’agit d’équipes du Tour de France et de leur nationalités.

Requête : année

Réf	Contexte gauche	Pivot	Contexte droit	
1	0001	Je suis rempli de l'espoir que cette	année	nous sera propice, parce que nous avons fait beaucoup au cours
2	0001	De ce que nous avons accompli pendant l'	année	qui se termine, les résultats sont apparents. Pour l'année
3	0001	, les résultats sont apparents. Pour l'	année	qui commence, je souhaite à la France, à l'Algérie
4	0002	en notre nom à tous, une bonne	année	à la France. Je le fais en toute confiance. Non
5	0002	. Non point que 1961 doive être une	année	sans épreuves. Au contraire, rien n'annonce qu'elle se

FIGURE 14 – Exemple de concordance issue du [corpus](#) des vœux présidentiels sur le [portail TXM](#)

Requête: année

Propriétés des cooccurrences : word  Seuils : Fmin ≥ 2 Cmin ≥ 2

Contexte :  forme  structure lb  Contexte gauche actif  Cont

de -10 à 0 et de 0 à 10

Cooccurent	Fréquence	Cofréquence	+ Indice	Distance moyenne
bonne	56	41	32	1,7
Vive	74	45	30	5,4
heureuse	25	25	26	1,1
Bonne	25	22	21	,8
!	148	52	20	6,1
nouvelle	50	25	14	1,0
année	258	61	14	5,9
cette	140	42	13	1,8
une	446	84	13	2,4
achève	15	13	12	3,2
vous	387	73	11	5,3
République	82	27	10	5,7
Mes	76	25	9	7,3
souhaite	70	23	8	4,0

FIGURE 15 – Exemple de cooccurrence pour le mot « année » dans le corpus des vœux présidentiels (TXM)

	A	B	C
1	Longueur	Segment	Fréquence
2		4 a minute or two	11
3		4 she said to herself	16
4		4 said the Mock Turtle	19
5		3 the March Hare	28
6		3 the Mock Turtle	49
7		3 the White Rabbit	20
8		3 she said to	17
9		3 she went on	16
10		3 said the Cat	14
11		3 said the Caterpillar	18

FIGURE 16 – Segments répétés tirés d'*Alice au pays des Merveilles* obtenus à l'aide de Lexico5

r<sup>1</sup> Mai-Août: word

Unités	Fréquence T 20404420	Mai t=4188855	score	Juin t=8081318	score	Juillet t=2306144	score	Août t=5828103	score
août	17150	587	000,0	2110	000,0	3546	272,3	10907	1 000,0
h	252917	48560	-62,8	91619	-270,5	30272	25,7	82466	1 000,0
éclipse	2027	58	-123,3	140	-249,8	319	8,9	1510	1 000,0
septembre	10774	722	000,0	2096	000,0	964	-15,1	6992	1 000,0
Wild	910	3	-84,5	4	-189,6	4	-40,5	899	1 000,0
West	553	13	-37,0	17	-92,2	16	-12,5	507	215,0
Notting	382	0	-38,1	0	-83,7	0	-19,9	382	207,9
lunettes	988	31	-58,0	87	-105,7	165	6,6	705	169,5
Hill	435	5	-35,2	18	-67,0	15	-8,6	397	166,8

FIGURE 17 – Spécificités calculées sur quatre mois du quotidien L'Est Républicain de l'année 1999, classées par mots les plus spécifiques au mois d'août.

r<sup>1</sup> Mai-Août: word

Unités	Fréquence T 20404420	Mai t=4188855	score	Juin t=8081318	score	Juillet t=2306144	score	Août t=5828103	score
juillet	11938	1082	-253,0	6358	198,7	3082	1 000,0	1416	-1 000,0
août	17150	587	000,0	2110	000,0	3546	272,3	10907	1 000,0
Esp	495	63	-5,4	59	-41,8	287	137,7	86	-8,3
Fra	942	158	-2,7	137	-63,1	370	107,5	277	0,5
Hassan	210	6	-13,4	11	-30,1	161	106,6	32	-5,4
Ita	702	193	5,2	109	-43,3	293	93,0	107	-16,1
TEL	101	0	-10,1	0	-22,1	100	92,7	1	-13,1
LOT	98	1	-8,4	1	-19,6	96	87,3	0	-14,3
BAN	86	0	-8,6	0	-18,8	86	81,4	0	-12,6
BIG	86	0	-8,6	0	-18,8	86	81,4	0	-12,6

FIGURE 18 – Spécificités calculées sur quatre mois du quotidien L'Est Républicain de l'année 1999, classées par mots les plus spécifiques au mois de juillet.

De nombreux logiciels se trouvent sur le « marché » de la linguistique de **corpus** aujourd'hui. Pour n'en citer que quelques uns parmi les plus populaires : TXM<sup>4</sup> (HEIDEN et al., 2010), Lexico<sup>5</sup>, Iramuteq<sup>6</sup>, AntConc<sup>7</sup>, Hyperbase<sup>8</sup>, Voyant Tools<sup>9 10</sup>... Chacun de ces outils présente ses propres spécificités et se situe dans l'une des sub-divisiones de l'analyse des données textuelles (ADT) (**textométrie**, **lexicométrie**, **logométrie**). Il est recommandé, avant d'avoir recours à l'un ou l'autre, de connaître un minimum leurs fonctionnements afin de choisir celui qui se prêtera le mieux aux objectifs de l'étude. De manière générale, l'utilisation d'un outil informatique n'est pas une fin en soi et ne doit pas dispenser d'une analyse des résultats obtenus : des sorties logicielles ne constituent pas des résultats si elles ne sont pas interprétées et analysées par le chercheur. Ces logiciels permettent de plus de combiner les approches quantitatives et qualitatives : les résultats sous forme de calculs statistiques guident le retour au texte pour forger des hypothèses (PINCEMIN, 2012a).

En somme, la linguistique de **corpus** et la **textométrie** sont dotées d'un équipement théorique et pratique pour l'analyse de grands volumes de données textuelles, dans des objectifs d'études linguistiques ou non. Les deux approches ont pour cela recours à des logiciels spécifiques qui permettent l'exploration des textes via une interface graphique de navigation et de recherche. Il nous semble que la mise en place de tels outils, même en l'état, et assortie d'une formation adéquate, représenterait déjà pour les **analystes criminels** un progrès sensible, en permettant par exemple la recherche de noms apparaissant dans des contextes similaires, des recherches simples ou complexes grâce aux cooccurrences et aux concordances. Le principe différentiel de la **textométrie**, de son côté, pourrait permettre d'obtenir un aperçu des thématiques des documents.

---

4. <http://textometrie.ens-lyon.fr/?lang=fr> [consulté le 11 décembre]

5. <http://www.lexi-co.com/L5Presentation.html> [consulté le 11 décembre]

6. <http://www.iramuteq.org/> [consulté le 11 décembre]

7. <https://www.laurenceanthony.net/software.html> [consulté le 11 décembre]

8. <http://ancilla.unice.fr/> [consulté le 11 décembre]

9. <https://voyant-tools.org/> [consulté le 11 décembre]

10. Pour une étude comparative de plusieurs logiciels de **textométrie**, voir (PINCEMIN, 2018)

### 1.3 Quel paradigme textuel pour l'analyse criminelle ?

Les humanités numériques, la linguistique de [corpus](#) et la [textométrie](#) nous apportent des éléments de réflexion pour préparer l'introduction d'une nouvelle manière d'appréhender les documents textuels en [analyse criminelle](#). Pour reprendre chaque angle de notre triangle outil-sujet-objet :

Premier point, l'outil : les moyens informatiques de type bureautique, combinés à l'utilisation d'un logiciel d'analyse de données, apportent une assistance à la pratique de l'[analyse criminelle](#). Néanmoins l'apport de ces moyens n'est pas suffisant pour couvrir l'ensemble des phases de travail. Les outils de linguistique de [corpus](#) et de [textométrie](#) permettent de faciliter le traitement des documents textuels, d'en obtenir une vue d'ensemble, et de mettre en place des approches qui pourraient s'avérer utiles à la pratique de l'[analyse criminelle](#). Le versant théorique de ces deux disciplines, combiné à celles actuellement développées en humanités numériques apporterait un cadrage épistémologique de ces nouvelles pratiques.

Deuxième point, le sujet : l'analyste est l'acteur de la recherche d'information et de la construction du raisonnement basé sur ces informations. Ce principe d'[analyse criminelle](#) correspond à la pratique de la linguistique de [corpus](#) selon laquelle le chercheur interprète les résultats produits par les logiciels : les logiciels de linguistique outillée ne fournissent pas une *réflexion*, et dans la même perspective, le raisonnement, la construction des hypothèses doivent rester le fruit du travail des [analystes criminels](#). Pour outiller l'[analyse criminelle](#), il faudra adapter des techniques et théories à ses objectifs et à la nature du texte traité, ce qui demandera probablement également de former les [analystes criminels](#), tout en gardant en tête que le cœur de métier de l'[analyste criminel](#) n'est pas la manipulation de textes : le choix opéré doit rester maniable pour un utilisateur formé mais non-expert.

Troisième point, l'objet : l'étude de la nature des documents traités en [analyse criminelle](#) nous a conduite à concentrer nos efforts sur les [auditions de témoins](#), qui représentent un sous-ensemble riche en information non-structurée. Nous avons également dégagé l'utilisation des entités comme clés d'entrée dans le texte, un concept que nous décrirons dans la suite de ce chapitre. D'autre part, au-delà des applications en [analyse criminelle](#) directement visées par cette recherche, les [auditions de](#)

**témoins** constituent un type de texte singulier qui n'a jusqu'à présent fait l'objet que de peu de recherches en français. Le chapitre V est donc consacré à une étude détaillée du texte des **auditions** de **témoins**, car nous espérons que leur caractérisation amènera non seulement des éléments pour leur traitement automatique, mais également des pistes de réflexion sur la production de ces textes : la prise en note, le dispositif de l'**audition**, la forme du texte consigné seront évoqués.

## 2 « Entités », « entités nommées » et « descriptions définies »

Comme nous l'avons expliqué dans le chapitre I, l'**analyse criminelle** fonctionne entre autres à l'aide du concept d'entités. Or, on rencontre en TAL (plus précisément en extraction d'information) le concept d'**entité nommée**. Au-delà de ses aspects techniques, la conceptualisation des **entités nommées** a amené les informaticiens à dialoguer avec les linguistes afin de définir les **entités nommées**, leur fonctionnement, et les formes qu'elles prennent. Dans cette partie, nous présentons puis comparons ces deux concepts, ainsi que le concept annexe de **description définie**.

### 2.1 En analyse criminelle

Dans le domaine criminel, le concept d'entité sert à désigner tout élément impliqué dans l'affaire étudiée. ROSSY (2011, p. 37) évoque des « entités d'intérêts », définies comme des « chose[s] possédant [des] existence[s] distincte[s] et identifiable[s] », en reprenant CHEN (1976), qui définit l'entité comme une chose pouvant être distinctement identifiée. Il propose une revue de ce que recouvre le terme « entité » dans le domaine de la recherche criminelle. Il en ressort que le concept s'applique à une grande variété d'objets : personnes, véhicules, adresses, organisations, événements, « choses », lieux, numéros de téléphone, types d'infraction, armes, drogues, documents, comptes bancaires, etc.

RIBAUX (2014, p. 373) quant à lui, définit les entités comme suit :

L'entité peut être une **trace** (p. ex. une transaction comptable) ou représenter une **trace** (p. ex. numérisation d'une **trace** digitale), un objet

concret (p. ex. un véhicule) ou abstrait (p. ex. une entreprise, une institution, une profession), une entité virtuelle (pseudo sur internet, un programme informatique), un type d'objet (p. ex. une marque et un modèle d'imprimante), ou une personne. [...] Il s'agit de délimiter les choses dont on veut parler<sup>11</sup> pour résoudre le problème.

On comprend donc que l'utilisation des entités permet de formaliser le processus de la recherche criminelle : face à des faits criminels, l'enquêteur recherche des entités, les situe dans le cadre présenté en section 3.1 du chapitre I, et produit ensuite des liens entre elles. Les entités de l'enquête ne sont pas nécessairement expressément identifiées ou nommées, elles sont conceptualisées.

Pour illustrer cela, RIBAUX (2014) détaille tout au long de son ouvrage deux exemples d'affaires criminelles dans laquelle des faits ont été imputé à un seul et même auteur au cours des investigations, l'affaire des animaux mutilés, et celle du fantôme de Heilbronn. Dans les deux cas, des faits criminels en grand nombre sont imputés au fil de la progression de l'enquête et de la commission des faits à un même auteur non-formellement identifié, mais *conceptualisé* par les enquêteurs, sur la base de différents indices et traces (similarité des faits pour l'un et traces ADN pour l'autre). Or il s'est avéré dans les deux cas qu'il n'y avait pas d'auteur unique, et que la recherche d'un tel auteur unique a fait obstacle à la manifestation de la vérité.

Si le concept d'entité en analyse criminelle n'est donc pas intrinsèquement lié au texte produit par les investigations, le travail d'analyse criminelle au PJGN est réalisé comme nous l'avons expliqué par l'exploitation de documents dont beaucoup sont textuels. Pour les analystes criminels du PJGN, la conceptualisation des entités se fait par l'exploration du texte de la procédure.

## 2.2 En traitement automatique des langues

Cette section consacrée à la description et à l'historique du concept des entités nommées en TAL sera essentiellement basée sur l'ouvrage de NOUVEL et al. (2015) qui présente le sujet et ses enjeux avec exhaustivité.

---

11. Ici, l'auteur reprend KIND (1987) sur la notion de cadrage des entités ou *frame*.

**Historique** Le TAL s'intéresse depuis la fin des années 80 à l'extraction automatique d'entités nommées (*named-entities recognition* en anglais, ou NER) du texte en langage naturel, au point que cette thématique se soit structurée en sous-domaine, celui de l'extraction d'information. Le principe de base est simple : il s'agit de repérer automatiquement des éléments d'intérêt préalablement définis dans du texte en langage naturel. Historiquement, on fait remonter cette tâche aux campagnes de la *Message Understanding Conference* (MUC), dont la première eu lieu en 1987 (GRISHMAN, 1997; NOUVEL et al., 2015, p. 12). L'objectif initial de ces campagnes était de comprendre les textes automatiquement, notamment afin de constituer des bases d'information, une tâche d'une grande complexité qui fut donc découpée en plusieurs sous-tâches, dont la détection d'entités nommées. Depuis, de nombreuses campagnes et programmes internationaux (MUC (GRISHMAN et SUNDHEIM, 1996), ACE (DODDINGTON et al., 2004)) et nationaux (ESTER 1 & 2, QUAERO, ETAPE (LE PEVEDIC et MAUREL, 2016; GALLIANO et al., 2006)) se sont consacrés à la détection d'entités nommées.

**Types d'entités** L'essentiel des publications de recherche actuelles au sujet des entités nommées sont dédiées à la présentation d'approches de détection des entités nommées, à leur normalisation ou à l'établissement de relations entre entités, en somme à proposer des contributions techniques. Dans bon nombre de ces publications, les entités nommées sont réduites aux noms propres et à certaines expressions de temps et de quantités, rejoignant là le modèle de la campagne MUC initial, qui cherchait à détecter noms propres (de lieux, de personnes et d'organisations dans une catégorie intitulée ENAMEX), expressions temporelles (dates et horaires dans une catégorie TIMEX) et expressions numériques (pourcentages, valeurs monétaires dans une catégorie NUMEX) (CHINCHOR et ROBINSON, 1998). Selon le domaine considéré, ces catégories générales sont ajustées. Par exemple, dans le cas du traitement automatique de la langue médicale, on recherchera des noms de gènes, de traitements, ou des posologies de médicaments.

**Définitions** La définition du concept d'entité nommée n'est pas aisée. NOUVEL et al. (2015, p. 22) soulignent la « quasi-absence de définition de ce concept » ainsi

qu'un « flottement historique quant à l'appellation « *entité nommée* » et à la portée de cette dénomination ». Les auteurs rapportent la multiplicité des « formules définitives » proposées par les campagnes d'annotation, qui peuvent néanmoins être regroupées en deux tendances : sémasiologique ou onomasiologique. L'approche sémasiologique part des expressions linguistiques pour déterminer le concept (du signe vers le concept), dans une perspective sémantique, quand l'approche onomasiologique part du concept pour aller vers les expressions linguistiques qui l'expriment, dans une perspective syntaxique (EHRMANN, 2008, p. 83-84). Dans les deux cas, on cherche à définir quelles formes annoter pour quels concepts.

EHRMANN (2008, p. 168) propose la définition suivante :

Étant donné un modèle applicatif et un *corpus*, on appelle *entité nommée* toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le *corpus*.

On retrouve ici la définition préalable des entités recherchées dans un modèle, ainsi que la notion cruciale de *référence*.

**Référence & descriptions définies** NOUVEL et al. (2015, p. 31) définissent la référence comme « le lien qui existe entre une expression linguistique et un élément du monde ». Si l'on considère les exemples suivants :

*Albert Grimaldi*

*Albert II*

*le prince de Monaco*

Bien que les deuxième et le troisième exemples ne puissent être qualifiés de noms propres, un humain est tout à fait capable de distinguer ces segments comme une entité de type personne, et de comprendre qu'ils réfèrent tous à la même entité du monde réel. Ce sont des expressions linguistiques autonomes, évoquant un référent par leurs seules ressources (NOUVEL et al. (2015, p. 50) citant CHAROLLES (2002)). Or les noms propres ne sont pas les seules structures fonctionnant ainsi, le TAL a également relevé le cas des *descriptions définies*, catégorie à laquelle on peut rattacher le troisième exemple ci-dessus.

Les **descriptions définies** correspondent à des expressions linguistiques de type *le* + nom. Par exemple : « le collègue de Pierre » est une **description définie**. Des éléments de définition ont été proposés à leur sujet notamment par KLEIBER (1981) qui énonce deux restrictions à leur égard : le nom dans la **description définie** doit permettre d'identifier un élément en particulier (« individuant ou *globalisant réifiés en discontinu* c'est-à-dire réduisant à l'état d'objet une notion abstraite », NOUVEL (2012, p. 44)), et comporter des indices ou points référentiels déterminant la signification. NOUVEL et al. (2015, p. 44) le résumant ainsi :

Si le substantif suppose une classe d'individus (restriction 1), l'article défini pointe lui sur un particulier précis et unique de cette classe, cette unicité devant être vérifiée empiriquement.

Ainsi l'exemple « le collègue de Pierre » est une **description définie** car elle pointe dans la classe générique « collègues » un particulier qui porte la caractéristique d'être « de Pierre ».

Le **TAL**, dans le cadre de la tâche de reconnaissance des **entités nommées**, s'est donc tout particulièrement intéressé au cas de différents types d'expressions linguistiques identifiant des individus ou particuliers<sup>12</sup>.

**Tâches et approches** Cette phase de conceptualisation des entités, en lien avec l'objectif final du projet, est nécessaire à la définition des objets pris en compte dans le développement des approches d'extraction. Comme NOUVEL et al., 2015, p. 22 l'affirment, deux questions essentielles doivent se poser avant de réaliser, évaluer, et utiliser un système de reconnaissance d'**entités nommées** : quelles entités prendre en compte (détermination des catégories), et comment les annoter (à quelles réalisations linguistiques elles correspondent).

Les approches mises en œuvre pour la détection d'**entités nommées** se regroupent en trois types de stratégies : les approches statistiques, les approches symboliques, et les approches hybrides. Elles sont détaillées avec une grande clarté dans

---

12. Étant donné l'objectif de notre recherche, nous avons limité l'état de l'art au **TAL**, toutefois l'expression de la référence est un sujet à part entière sur lequel on pourra consulter les travaux de sémantique de G. KLEIBER que nous avons déjà cités, ou encore TAMBA (1983), CADIOT (1991), DUBOIS (1997).

le chapitre 2 de la thèse de GROUIN (2013), qui les applique dans un objectif d'anonymisation de documents cliniques, une tâche proche de la reconnaissance d'entités nommées.

Les approches statistiques reposent sur l'usage de grandes quantités de données utilisées comme exemples afin d'alimenter un système d'apprentissage automatique. (WISNIEWSKI, 2007) cité par GROUIN (2013, p. 75) définit l'apprentissage artificiel comme suit :

L'apprentissage artificiel a pour objectif la mise au point de programmes capables d'apprendre à partir de leur expérience, c'est-à-dire de changer leur structure interne ou la valeur de leurs paramètres en fonction de leur expérience de manière à améliorer leurs performances futures.

On fournit à un algorithme un corpus de données faisant office d'exemples, à partir duquel on dit que l'algorithme *apprend* ou *est entraîné*, puis l'algorithme est déployé sur un corpus de données qui lui sont inconnues et non-annotées, sur lequel il doit alors opérer des annotations ou prendre des décisions selon les caractéristiques qu'il a observées dans le corpus d'exemples.

Ce type d'approches suppose l'existence ou la création de corpus d'exemples, c'est-à-dire de données évaluées et annotées par des humains selon des règles précises par un guide d'annotation. Le guide d'annotation compile les modalités d'annotation, c'est-à-dire les types de réalisations linguistiques correspondant aux entités que l'on souhaite détecter. La constitution de ces données que l'on appelle données d'apprentissage ou données d'entraînement est un travail coûteux en temps et en ressources humaines qui fait l'objet de recherches concernant par exemple la rédaction du guide d'annotation, les questions éthiques des conditions de travail des annotateurs<sup>13</sup>, l'évaluation de la conformité des annotations relativement aux lignes directrices et de l'accord entre les différents annotateurs, etc.

Les approches symboliques, dites également à base de règles, reposent sur la définition de règles décrivant les entités à extraire et/ou sur l'utilisation de ressources externes : outils qui apportent une information supplémentaire tels que lemmatiseur, étiqueteur morpho-syntaxique, ou ressources linguistiques tels que lexiques,

---

13. Consulter à ce sujet les travaux de K. Fort sur le crowd-sourcing, la myriadisisation et Amazon Mechanical Turk : FORT et al., 2011

dictionnaires, listes de déclencheurs. Ces ressources et outils peuvent être projetés directement sur le **corpus** ou combinés entre eux. Contrairement aux approches statistiques, les approches symboliques ne nécessitent pas de **corpus** annoté mais mobilisent par contre des connaissances expertes pour la constitution des ressources adaptées au projet.

Les approches mixtes ou hybrides combinent les approches statistiques et symboliques, par exemple en utilisant un **corpus** annoté par des méthodes symboliques dont les résultats ont été contrôlés et validés pour servir de données d'apprentissage à un système statistique.

Chacune de ces approches présente ses propres intérêts et s'applique plus ou moins bien à chaque situation de recherche. Les approches symboliques sont privilégiées si l'on peut développer ou adapter des ressources linguistiques et dans les cas où l'on ne dispose pas de suffisamment de données d'entraînement pour alimenter un modèle statistique, car elles présentent l'avantage de pouvoir être déployées plus rapidement. Néanmoins, elles demandent plus de maintenance sur le long terme que les approches statistiques qui sont plus robustes et s'adaptent mieux aux cas nouveaux (MAUREL et al., 2011).

**Évaluation** L'évaluation des résultats de la reconnaissance automatique des entités se fait en général en comparant une portion du **corpus** annotée par un ou des humains avec les résultats produits. Les métriques les plus utilisées sont le **rappel**, la **précision** et la **F-mesure**.

La **précision** évalue la pertinence des entités reconnues :

$$P = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux positifs}}$$

Autrement dit : sur l'ensemble des entités reconnues, combien sont correctes.

Le **rappel** évalue le nombre d'entités pertinentes reconnues par rapport au nombre d'entités pertinentes totales :

$$R = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux négatifs}}$$

Précision	Rappel	F-mesure
0.9	0.6	0.72

TABLE 2 – Exemples de précision, rappel, F-mesure

Autrement dit : sur l'ensemble des entités du texte, combien sont correctement reconnues.

La **F-mesure** combine **rappel** et **précision** :

$$F = 2 \cdot \frac{\text{précision} \cdot \text{rappel}}{\text{précision} + \text{rappel}}$$

La **précision**, le **rappel** et la **F-mesure** prennent la forme d'un indice inférieur ou égal à 1, 1 équivalant à un score parfait. Dans l'exemple du tableau 2, on constate un décalage entre la **précision** et le **rappel**. On interprète ces résultats comme ceux d'un système générant peu de bruit, mais qui manque beaucoup d'éléments pertinents.

C. GROUIN, à la page 99 de son manuscrit indique que la **F-mesure** peut être pondérée selon l'application qui en est faite afin de privilégier le **rappel** ou la **précision**<sup>14</sup>. Par exemple, dans son cas d'étude qui est celui de l'anonymisation de documents cliniques, on préférera privilégier le **rappel**, c'est-à-dire que le système reconnaisse plus de choses qu'il ne devrait. En revanche, privilégier la **précision** revient à donner plus d'importance à l'exactitude des résultats au détriment de leur quantité. Dans notre cas et au stade actuel de nos recherches, il n'est pas possible de déterminer si la pondération de l'une ou de l'autre des valeurs est préférable. En effet, privilégier le **rappel** produit du bruit, on risquerait alors de gêner le travail de l'**analyste criminel** et l'utilité du système de reconnaissance des entités serait diminuée, mais privilégier la **précision** c'est prendre le risque que des entités pertinentes ne soient pas présentées à l'**analyste criminel**, ce qui de nouveau remettrait en cause l'intérêt du système. Par conséquent, lorsque nous ferons usage de la **F-mesure**, celle-ci ne sera pas pondérée, c'est-à-dire que **rappel** et **précision** y auront le même poids.

14. Dans ce cas, la formule devient

$$F = \frac{(1 + \beta^2) \cdot \text{précision} \cdot \text{rappel}}{\beta^2 \cdot \text{précision} + \text{rappel}}$$

où si  $\beta > 1$ , on privilégie le **rappel**, et si  $\beta < 1$ , on privilégie la **précision**.

### 2.3 Détection d'entités dans des textes liés au domaine criminel

Nous avons trouvé dans la littérature plusieurs contributions étudiant la détection d'entités dans du texte en langage naturel relevant du domaine criminel.

Deux concernent du texte qui n'est pas produit directement par les services de police. ARULANANDAM et al. (2014) ont utilisé des champs aléatoires conditionnels (CRF) pour détecter automatiquement les lieux de crimes dans des articles de presse en anglais en se concentrant sur les cas de vol. SCHRAAGEN et al. (2017) détectent six types d'entités (événements, localisations, divers, organisations, personnes, produits) à l'aide d'un outil de reconnaissance d'entités nommées généraliste dans des pré-plaintes en ligne en néerlandais.

CARNAZ et al. (2019) ont utilisé la suite Apache OpenNLP<sup>15</sup> pour la détection de quatre types d'entités (personnes, organisations, lieux, dates) dans des rapports de police en portugais, sans plus d'information sur la nature des données textuelles traitées.

CHAU et al. (2002) ont extrait à l'aide d'un réseau de neurones quatre types d'entités (personnes, adresses, drogues, biens personnels) dans 36 rapports en anglais du service de police de Phoenix (Arizona) liés aux stupéfiants.

KU et al. (2008) détectent quinze catégories, elles-mêmes divisées en 126 sous-catégories : acte/événement, lieu, personne, biens personnels, véhicules, armes, partie du corps, heure, drogue, chaussures, électronique, caractéristique physique, condition physique, chevelure et habillement, dans des documents de type narratif émanant de témoins et de la police. Leur approche repose sur l'utilisation d'un grand « lexique du crime »<sup>16</sup> combiné à des modules GATE<sup>17</sup>. La construction du lexique s'est faite sur des ressources linguistiques (FrameNet) et encyclopédiques (Wikipedia, Encarta). Le développement et le test de l'approche ont eu lieu sur des documents issus de blogs et de sites internet de *true crime* ou d'entraide dans le domaine de la justice<sup>18</sup>. Cette revue sommaire illustre la diversité des entités qui

15. Apache OpenNLP : <https://opennlp.apache.org/> [consulté le 11 décembre 2019].

16. « crime-specific lexicon »

17. GATE <https://gate.ac.uk/sale/tao/split.html> [consulté le 11 décembre 2019] est une plateforme de gestion textuelle permettant de réaliser de l'annotation et divers traitements linguistiques.

18. Notamment :

— alt.true-crime : <https://groups.google.com/forum/#!forum/alt.true-crime> [consulté le 10 décembre 2019],

— Crime-stoppers.org : <https://crime-stoppers.org/> [consulté le 10 décembre 2019],

intéressent l'**extraction d'information** à des fins de recherche criminelle et les écarts de couverture du sujet (de une à quinze catégories détectées). Nous relevons cependant qu'aucune de ces contributions ne décrit les entités recherchées, et les formes qu'elles prennent dans le texte contrairement aux recommandations que nous avons citées précédemment.

D'autre part, deux de ces références (ARULANANDAM et al. (2014) et SCHRAAGEN et al. (2017)) concernent des textes qui ne sont pas produits par des services de police ou dans le cadre d'une **enquête** : du texte journalistique et du contenu généré par l'utilisateur dans un système de dialogue. Or il a été démontré notamment par JACQUES et AUSSENAC-GILLES (2006) que le genre textuel est un paramètre à prendre en compte lors de l'application d'approches de **TAL**. NADEAU et SEKINE (2007) soulignent également le manque d'attention accordée au genre textuel et au domaine dans les approches de reconnaissance d'**entités nommées**, alors qu'il s'agit d'une cause de baisse de performance des systèmes. Sur les cinq références présentées ci-dessus, seules deux semblent concerner réellement du texte rédigé par des services de police (pour KU et al. (2008) la nature du texte traité n'est pas très claire et un certain nombre de sites dont le **corpus** est issu ne sont plus en ligne).

Cet aspect nous conforte dans la décision présentée au chapitre II de ne traiter qu'un seul type de documents figurant dans les **procédures** judiciaires. En effet, un domaine d'application donné, dans notre cas le domaine de l'**analyste criminel**, englobe différents genres textuels, ce qui influe sur la stratégie à adopter et sur les performances. En conséquence, une étude du genre textuel des **auditions** et de textes issus plus largement du domaine de la justice sera présentée au chapitre V.

## 2.4 Lier entités criminelles et entités nommées

Nous avons présenté dans cette section le concept d'entité en **analyse criminelle** et celui d'**entité nommée** en **TAL**. Dans cette section, nous chercherons à souligner les points communs entre les deux concepts, afin d'amorcer la conceptualisation des objets à rechercher dans notre étude.

---

— TheLaw.com : <http://www.thelaw.com/forums/> [consulté le 10 décembre 2019].  
Les autres sites mentionnés n'étaient plus en ligne.

Dans un souci de clarté, nous désignerons dans cette section les entités traitées en [analyse criminelle](#) par « entités criminelles ».

L'[enquête](#) partage avec la fouille de textes et l'[extraction d'information](#) l'objectif de répondre à des questions basiques permettant de comprendre les circonstances d'occurrence du propos, des questions que l'on rassemble parfois sous l'appellation de questionnement de Quintilien : qui, quoi, où, comment, combien, pourquoi ? (GROUIN (2013, p. 73) : « En matière de fouille de textes, le repérage des [entités nommées](#) sert traditionnellement [...] à répondre à des questions basiques (Qui? Quoi? Où? Quand? Comment?) » et RIBAU (2014, p. 371) : « Les investigations veulent répondre aux questions attribuées à Quintilien : qui, avec qui, avec quoi, où, quand, quoi, comment, pourquoi? »). Les entités criminelles et nommées, dans les deux processus et une fois organisées entre elles, sont supposées apporter les réponses à ces questions et donc la compréhension globale des faits présentés ou investigués. On peut déjà dégager une similarité de finalité des entités en [analyse criminelle](#) et en [TAL](#).

Un autre point commun peut être pointé dans la phase de conceptualisation. O. RIBAU dans un extrait repris page 63 mentionne la nécessité de « délimiter les choses dont on veut parler pour résoudre le problème ». Cette démarche de délimitation des entités intéressantes dans une affaire ou une [procédure](#) correspond à la phase de délimitation des [entités nommées](#) préalable à leur détection évoquée par NOUVEL et al. (2015).

Les entités criminelles et les [entités nommées](#) partagent une fonction référentielle inscrite dans un cadre conceptuel, celui de la [procédure](#) pour les entités criminelles, et celui du texte pour les [entités nommées](#). Toutes deux font un lien entre ce cadre et un objet du monde réel qu'il convient de représenter. Elles partagent également une variété de types difficile à circonscrire, tout en incluant des types d'intérêt supérieur comme les personnes, les dates et les lieux, car ces types d'entités constituent des clés fondamentales d'accès à l'information.

Concernant le « contenu » de l'entité, examinons ce passage de RIBAU (2014, p. 373-374) :

S'il s'agit bien d'un homicide, il y a au moins un auteur forcément. Cette

personne existe donc en tant qu'entité, mais sa description n'est pas immédiatement disponible. L'investigation cherchera d'abord à préciser à quoi il ressemble (profil), à trouver son identité et à le localiser.

Le processus d'enquête est présenté dans ce passage comme la collecte de traits caractéristiques de l'entité *auteur*, collecte menant à la découverte de l'identité de l'entité. En somme, partant d'une entité criminelle conceptualisée dans le cadre de l'enquête, l'accumulation et le croisement d'informations à son sujet doit mener à la découverte de l'entité *nommée* qui correspond à l'entité criminelle.

Par conséquent, les entités criminelles figurant en procédure peuvent avoir une existence sans être rattachées à une entité nommée, si l'on admet l'entité nommée comme une entité qui identifie. De plus, les entités criminelles à la différence des entités nommées ne sont pas intrinsèquement textuelles, elles peuvent être orales, visuelles ou schématiques selon le support sur lequel elles figurent (vidéo, image, schéma d'analyse criminelle) au sein de la procédure.

Dans le cadre de notre étude, nous sommes donc face au cas d'une *gestion textuelle des entités criminelles*. Notre tâche consiste donc à explorer et adapter les méthodes de reconnaissance des entités nommées et de leurs concepts reliés et de les expérimenter sur les entités criminelles. Dans cette perspective, nous devons passer par la phase de détermination des catégories de ce que l'on cherche, puis par celle de détermination de l'annotation, c'est-à-dire les formes que les entités recherchées prennent dans le texte. Ayant circonscrit notre recherche à un type particulier de document textuel de la procédure, les auditions de témoin, nous serons donc amenée à observer et décrire l'influence et les contraintes que ce type de document produit sur les réalisations linguistiques correspondant à des entités d'intérêt pour l'analyse criminelle.

### 3 Synthèse

Ce chapitre d'état de l'art a permis de passer en revue les disciplines touchant à notre sujet de recherche, à savoir la conception d'outils et de méthodes de traitement textuel pour l'analyse criminelle. Nous avons présenté les méthodes appliquées en humanités numériques et linguistique de corpus, les concepts d'entités en analyse

criminelle et d'entités nommées en TAL ainsi que les approches permettant leur détection automatique. Cet état de l'art a permis de rapprocher les concepts d'entité criminelle et d'entité nommée afin d'ouvrir la voie aux recherches pour la détection des entités criminelles dans le texte des procédures judiciaires. La littérature que nous avons rencontrée sur le sujet (présentée à la section 2.3) est peu fournie, et les cas présentés ne s'adaptent pas bien à notre situation en raison de la nature des textes traités, et parce qu'aucune de ces contributions n'est appliquée sur du français.

Deux directions de recherche peuvent être dégagées pour la recherche en extraction d'information appliquée à l'analyse criminelle : il s'agit d'une part d'adapter les méthodes de repérage et de collecte de l'information dans la procédure en accord avec les pratiques et les intérêts de l'analyse criminelle, et d'autre part, il est nécessaire de concevoir les modalités de la mise à disposition de ces informations aux analystes criminels. Si ces deux préoccupations peuvent sembler éloignées, elles n'en sont pas moins connectées du point de vue épistémologique et l'on peut les replacer encore une fois dans le dialogue produit dans le cadre des humanités numériques entre les sciences humaines et les sciences informatiques. En effet, les méthodes de TAL devront s'adapter aux contraintes d'un domaine où l'humain, selon nous, doit continuer à jouer un rôle d'interprétation puisque c'est à l'aide de son expérience du terrain judiciaire que les informations extraites pourront être valorisées dans l'analyse.

Ce parti-pris se situe dans la perspective évoquée par POIBEAU (2014b). Les travaux présents et futurs de conception de systèmes d'extraction d'information de la procédure judiciaire à des fins d'analyse criminelle doivent et devront ajuster performances informatiques et rendu à destination de l'humain. Dans le cas du TAL pour les sciences sociales, T. POIBEAU souligne le fait que la compréhension profonde du texte nécessite la production d'inférences hors de portée des systèmes actuels, qui restent à la surface des choses. Il n'est donc pas d'actualité de se passer de l'analyste humain et l'auteur s'en félicite. Nous ajoutons à cette limitation technique une considération judiciaire : il est admis que l'humain est faillible, l'erreur est donc possible au cours d'une enquête<sup>19</sup>, et les cas d'erreurs humaines sont fréquents dans les

---

19. Par « possible » nous n'entendons pas « tolérable », mais « du domaine des possibilités ».

**enquêtes** à travers le monde. Justifier une erreur ou un oubli causé par un système informatique n'en est pas (encore) à ce stade, et semble d'autant moins acceptable que l'on aime prêter aux méthodes informatiques un prétendu caractère infallible, ou tout du moins, une plus grande fiabilité qu'aux êtres humains. Toute personne s'étant frottée un tant soit peu à la programmation ou au développement informatique sait que la réalité est tout autre, il n'en reste pas moins que dans le cas d'une **enquête** criminelle, il semble extrêmement problématique d'invoquer comme prétexte à une erreur que « le logiciel s'est trompé ». C'est pourquoi nous insistons sur le rôle fondamental de contrôle, de validation et d'interprétation que l'**analyste criminel** doit conserver dans le traitement de la **procédure**. L'objectif des systèmes à développer doit être de reconnaître, d'extraire et de mettre à disposition les informations présumées d'intérêt que l'**analyste criminel** passe aujourd'hui beaucoup de temps à rechercher au sein des documents. Si le développement de ce type de système aboutit, on peut imaginer une pratique de l'**analyse criminelle** où le travail de l'analyste serait uniquement d'interpréter et de construire des hypothèses et non plus de parcourir les documents à la recherche des informations.

Il s'agit donc de concevoir une nouvelle méthode de pratique de l'**analyse criminelle**. Si l'on reprend cet extrait de POIBEAU (2014b) appliqué au cas d'un analyste linguiste employant des outils de TAL :

Le plus important consiste sûrement à déterminer la façon d'intégrer l'analyste dans la boucle d'analyse [...]. Les outils automatiques sont efficaces pour identifier des thèmes, donner une idée du contenu d'un **corpus** documentaire trop grand pour être étudié manuellement. Les outils peuvent aussi permettre de « sonder » le contenu, c'est-à-dire estimer le contenu probable en lançant des requêtes a priori pertinentes. C'est ensuite à l'analyste de définir l'information à extraire et, surtout, de décrire comment celle-ci peut être extraite. [...] C'est dans la façon de définir les problèmes, de choisir les bonnes techniques et surtout de faire collaborer le tout en un ensemble harmonieux que se situent les enjeux pour les sciences sociales.

Ici, T. POIBEAU considère le cas où l'analyste est à la fois le développeur des

solutions employées et la source des analyses et observations. En réalité, dans notre cas, deux rôles d'analystes distincts peuvent être dégagés : d'une part le notre, qui recouvre le développement du système et d'une méthodologie, et d'autre part, celui de l'[analyste criminel](#), qui concerne l'utilisation du système et la méthode d'accès à l'information de la [procédure](#). Dans les deux cas, le rôle de l'analyste n'est pas celui d'un pousse-bouton mais celui d'un véritable acteur de la recherche, qui comprend ses outils, qui sait les employer à bon escient selon les cas et les objectifs, et qui est capable d'avoir un regard critique et du recul à la fois sur les outils et sur les résultats.

HO ĐINH (2017, p. 57), reprenant EENSOO et VALETTE (2015) affirme que les méthodes et logiciels d'[ADT](#) s'opposent au [TAL](#) et à sa volonté d'automatisation, une idée reprise et approfondie par VALETTE (2016), qui examine et compare les pratiques et objectifs des deux champs. M. VALETTE souligne que l'un des objectifs du [TAL](#) est de réduire la part de l'intervention humaine dans le processus de traitement du texte, mais propose une opposition novatrice entre le travail humain et le travail de l'ordinateur :

Les tâches effectuées par les humains sont par tradition qualifiées [par le [TAL](#)] de *manuelles*. L'antonyme en est certes *automatique*, mais peut s'entendre également comme *intellectuel* [...].

Notre recherche, et le champ qui en naît, doivent chercher à conjuguer cette volonté d'automatisation et de reproductibilité du [TAL](#) pour les aspects [extraction d'information](#) avec la méthodologie d'interprétation de l'[ADT](#) et de la linguistique de [corpus](#). En somme : à l'ordinateur le travail de force d'extraction, à l'humain le travail intellectuel de réflexion.

Par ailleurs, cette articulation entre [TAL](#) et exploration des données textuelles sera également le paradigme conducteur de ce travail de thèse. En effet, comme nous l'avons déjà évoqué plus haut notre situation correspond à la description de T. POIBEAU où l'analyste (c'est-à-dire nous-même) s'appuie sur un sondage des données étudiées pour développer son approche (définir, décrire l'information à extraire et comment l'extraire).

## 4 Conclusion

L'*analyse criminelle* présente de nouveaux défis pour l'*extraction d'information*, des défis conceptuels liés à l'application des approches à un domaine de collecte d'information où la part d'incertitude est grande, et des défis relatifs à l'importance de la part de l'humain dans le travail d'*analyse criminelle*. Par ces aspects, la problématique d'outillage de l'*analyse criminelle* s'inscrit dans les mêmes préoccupations que celles des sciences humaines et sociales vis-à-vis du numérique et des nouvelles technologies que l'on rencontre dans la communauté des humanités numériques.

Étant donné l'état de l'art peu fourni sur le sujet, le travail nécessaire pour le développement d'un système ou d'une plate-forme complète s'étale de bout en bout de la chaîne de traitement du texte, c'est-à-dire du pré-traitement des données jusqu'à la restitution des résultats. Ce travail de thèse étant contraint par le temps, les moyens humains, techniques et financiers, nous avons dû sélectionner les aspects à traiter prioritairement. Les résultats de notre travail représenteront donc un premier défrichage de la détection des *entités nommées* pour l'*analyse criminelle* (chapitre IV), et une étude des documents pris en compte dans notre recherche (chapitre V).

## Chapitre IV

# Les entités en analyse criminelle

En suivant les pratiques de développement de systèmes de détection d'[entités nommées](#) que nous avons présentées au chapitre [III](#), ce chapitre sera constitué de deux parties : la première concerne la conceptualisation des entités criminelles dans le texte, et la deuxième présentera une approche exploratoire de leur détection.

Dans la première partie du chapitre, nous tâcherons de déterminer les types d'entités utiles pour l'[analyse criminelle](#) et nous illustrerons ces catégories à l'aide d'exemples anonymisés tirés des dossiers de [procédure](#) que nous avons pu étudier. Nous chercherons à apporter le meilleur aperçu possible de l'étendue des réalisations linguistiques auxquelles peuvent correspondre les entités criminelles. La conceptualisation sera l'occasion de comparer les caractéristiques des différentes catégories d'entités utiles à l'[analyse criminelle](#) ainsi que de comprendre les facteurs influençant leurs formes linguistiques.

Dans la deuxième partie nous développerons et testerons une approche minimale basée sur les données dont nous disposons et sur des ressources en libre accès. Le but est de démontrer que même une approche qui n'a pas les faveurs de la recherche en [TAL](#) actuellement représente déjà un progrès sensible à l'échelle des [analystes criminels](#).

En conclusion, nous analyserons le rapport entre la conceptualisation des entités et la pertinence de l'approche que nous avons construite. Nous évoquerons les possibilités futures pour la poursuite de la recherche en [extraction d'information](#) appliquée aux [procédures](#) judiciaires exploitées en [analyse criminelle](#).

## 1 Catégories et réalisations linguistiques correspondantes

Déterminer les catégories d'entités intéressantes pour l'analyse criminelle s'est fait en concertation avec les analystes du SCRC. Parmi les entités qu'ils manipulent, certaines le sont dans toutes les affaires, et ce sont ces entités que nous avons considérées comme prioritaires : il s'agit des numéros de téléphone, des dates, des personnes, des véhicules et des lieux.

Il n'est pas forcément souhaitable qu'un système de détection d'entités nommées pour l'analyse criminelle cherche à couvrir toutes les entités d'une procédure judiciaire, en premier lieu parce qu'une telle volonté de catalogage est peine perdue, et que tout détecter revient à ne plus rien trier et donc à remettre l'analyste sous le même volume de données que l'on cherchait initialement à organiser. De plus, si certaines entités se rencontrent dans toutes les affaires (les personnes, les dates et les lieux au minimum, si l'on admet une affaire criminelle comme le résultat d'une activité humaine inscrite dans le temps et dans l'espace), ce n'est pas le cas de toutes les entités pour autant, et l'importance de certaines varie selon le type d'affaire considéré. Par exemple, dans les affaires de criminalité en réseau du type trafic de stupéfiants, l'analyse des communications téléphoniques sert pour reconstruire la structure hiérarchique du réseau, ce qui ne sera pas nécessairement aussi utile dans des affaires de type homicide.

Parmi les exemples tirés de la littérature concernant la détection d'entités criminelles présentées à la section 2.3 du chapitre III, on remarque que plusieurs références comptaient des catégories « divers », « choses » ou « objets ». Il semble particulièrement difficile de délimiter conceptuellement ce type de catégorie, qui peut recouvrir d'autres types d'entités : un objet du quotidien peut être employé comme une arme, par exemple, ce qui est d'autant plus parlant quand on considère le concept légal d'arme par destination, précisé à l'article 132-75 du Code pénal<sup>1</sup> :

Est une arme tout objet conçu pour tuer ou blesser.

Tout autre objet susceptible de présenter un danger pour les personnes

---

1. Légifrance : Article 132-75 du Code Pénal <https://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006417499&cidTexte=LEGITEXT000006070719&dateTexte=20040310> [consulté le 21 octobre 2019]

est assimilé à une arme dès lors qu'il est utilisé pour tuer, blesser ou menacer ou qu'il est destiné, par celui qui en est porteur, à tuer, blesser ou menacer.

Une voiture, une batte de base-ball ou encore un marteau sont des objets qui peuvent faire office d'arme par destination bien qu'il ne s'agisse pas de leur usage initial. Réciproquement, la catégorie « arme » est tout aussi difficile à définir.

Nous concentrerons donc nos recherches sur les cinq types d'entités que nous avons énoncés plus haut. À cette fin, nous avons lu le [corpus](#) intégralement, puis la collecte des exemples s'est appuyée sur le logiciel de textométrie TXM (HEIDEN et al., 2010), qui permet de réaliser des recherches plein texte et des concordances textuelles. Pour rappel, à l'issue du chapitre II, nous avons resserré notre étude sur les [auditions](#) de [témoins](#). Nous disposons donc d'un [corpus](#) composé de 370 [auditions](#) de [témoins](#), totalisant environ 600 000 mots, sur lequel nous baserons nos expérimentations.

## 1.1 Description des entités

Dans le chapitre précédent, nous avons établi que notre recherche concernait la gestion textuelle des entités criminelles, c'est-à-dire les formes textuelles des entités criminelles. On évoquera donc à partir de maintenant des [entités criminelles textuelles](#). Dans les sous-parties suivantes, nous examinons chaque type d'entité retenu en les illustrant à l'aide d'exemples tirés du [corpus](#). L'objectif est de recueillir les réalisations linguistiques qui peuvent tomber sous les titres des catégories d'entités, et ainsi de comprendre les spécificités des [entités criminelles textuelles](#) dans les [auditions](#) de [témoins](#).

### 1.1.1 Numéros de téléphone

Les numéros de téléphones dans les [auditions](#) de [témoins](#) figurent dans la partie de renseignements d'état-civil, et parfois dans le corps même de l'[audition](#), par exemple si l'on demande à un [témoin](#) de fournir des numéros de téléphone de son entourage.

Concrètement, un numéro de téléphone correspond à une suite arbitraire de chiffres. En France métropolitaine, il s'agit d'une suite de cinq paires de chiffres, la première paire marquant l'origine géographique ou le type de ligne attribuée (01, 02, 03, 04, 05 pour l'origine géographique, 06, 07, 08, 09 pour les types de lignes). À ceci s'ajoute les indicatifs spéciaux pour les DOM-TOM (0590, 0596, 0594, 0262, 687)<sup>2</sup>.

Des numéros de téléphone étrangers peuvent également figurer dans la **procédure** (sur les factures détaillées) et dans les **auditions** (affaires frontalières, **témoins** résidant habituellement à l'étranger). Dans ce cas, les numéros de téléphone ne correspondent pas au modèle du plan de numérotation français, et s'ils figurent en **procédure**, sont agrémentés de leur indicatif international.

Ils sont pris en note sous la forme d'une suite de chiffres d'un seul tenant ou intercalés d'espaces, de points ou de tirets, parfois de barres obliques.

J'ai son numéro de téléphone portable à savoir le **06 00 00 00 00** ainsi que le numéro de téléphone de son domicile **03 00 00 00 00**.

Il m'a téléphoné avec le téléphone n°**06.00.00.00.00**.

J'exerce la profession d'ingénieur dans une société allemande. Mon numéro de téléphone est le **00 49 000 000 00 00**.

Mon numéro de téléphone est le **0600000000 / 0400000000**.

Je vous fournis son numéro de téléphone portable **06-00-00-00-00**.

Mon numéro de téléphone est le **05/00/00/00/00** et **06/00/00/00/00**.

Un développement intéressant serait de recouper les numéros de téléphone repérés dans les **auditions** avec les numéros figurant sur les factures détaillées. Néanmoins, le numéro de téléphone, en tant qu'entité formalisée répondant à un standard ne nécessite pas d'effort particulier de conceptualisation ni de délimitation textuelle. Par conséquent nous ne les détaillerons pas outre mesure.

2. Pour une explication détaillée du plan de numérotation téléphonique en France, consulter l'article Wikipédia à ce sujet [https://fr.wikipedia.org/wiki/Plan\\_de\\_num%C3%A9rotation\\_en\\_France](https://fr.wikipedia.org/wiki/Plan_de_num%C3%A9rotation_en_France) [consulté le 11 décembre 2019]

### 1.1.2 Dates & éléments temporels

Les dates sont très fréquentes dans les **auditions**, qu'il s'agisse des éléments d'état-civil relatifs au **témoign** entendu, du récit d'éléments biographiques ou de celui des jours concernant les faits examinés. Elles permettent de situer des événements dans le temps.

Nous dégageons deux paramètres modulant les types d'informations temporelles (appellation que nous préférons au terme de « dates » strictement) : la précision et l'échelle.

La précision concerne l'exactitude avec laquelle l'expression désigne une information temporelle. Par exemple, « en fin de matinée » et « à 11h35 » désignent une même période temporelle, mais la seconde expression est plus précise que la première.

L'échelle concerne la focalisation des faits rapportés : s'il s'agit d'éléments de vue d'ensemble ou d'éléments reliés à des faits précis. Typiquement, la notion d'échelle distingue les informations rapportées au niveau biographique ou au niveau de l'emploi du temps ou du déroulé d'une journée.

Précision et échelle ne sont pas en concurrence. On peut rapporter des éléments biographiques avec précision (par exemple, citer la date d'un mariage) et on peut citer des éléments d'emploi du temps avec imprécision (*cf.* l'exemple ci-dessus comparant « en fin de matinée » et « à 11h30 »).

L'obtention et l'usage d'informations temporelles dans les **auditions** peuvent s'inscrire dans différents objectifs et sont soumis à différentes conditions.

Par exemple, pour retracer le parcours d'une personne soupçonnée de faits anciens ou sériels, on examinera son *curriculum vitae*, ses emplois successifs, déménagements, mutations, errances, qui seront croisés avec les emplacements géographiques des faits que l'on cherche à lui imputer. La précision des informations peut être amoindrie par la distance temporelle entre l'**enquête** et les faits.

En revanche si l'on interroge une personne au sujet de faits récents, on peut obtenir de sa part le récit de son emploi du temps en termes de jours au moins. Pour ce qui concerne les horaires, en dehors des horaires de travail, de cours ou encore si l'on a un train à prendre, il n'est pas courant d'y prêter une attention soutenue.

Fréquemment les **témoins** ne sont pas capables de citer une heure précise et utiliseront à la place une période de la journée (« dans la matinée », « à midi », « en soirée », etc.).

Pour illustrer ceci, considérons les exemples suivants.

1. Le **jeudi 30 août 2004 à 17 heures 00 minute** Nous soussigné Gendarme, BOULANGER Patrick, Officier de Police Judiciaire en résidence à Section de Recherches de CAEN

2. **L’an deux mille sept, le vingt neuf septembre** Nous soussigné (s) LUBSCK Thierry, Adjudant et GRENIER Stéphanie, Gendarme, en résidence à la Section de Recherches de LILLE, Officiers de Police judiciaire [...]

3. Savez-vous quel a été l’emploi du temps précis de Ghislaine pour la journée du **dix neuf septembre 2009** ?

4. Philippe m’a parlé **jeudi 27/08/2017** un peu à près 17h alors qu’il s’apprêtait à aller chercher Léo à la crèche.

5. Je l’ai eu au téléphone le **10** ou le **11 septembre 2009** avant le week-end.

Les exemples 1, 2 et 3 et 4 comportent des dates précises et complètes, c’est-à-dire composées d’un jour, un mois, et une année, ce qui les rend non-ambiguës. L’exemple 5 mentionne également une date complète et explicite, mais le **témoin** a un doute sur celle-ci, dans cet exemple, la précision de l’information est affectée.

6. **Vendredi 12** au soir vers **21 h00**, j’ai eu un appel de la gendarmerie avec le portable d’Eric me disant qu’il ne pouvait pas venir travailler.

7. QUESTION : Est ce que vous travailliez **le jour de la mort de Dominique** ?

L’exemple 6 comporte une date incomplète, composée seulement d’un jour de la semaine et d’un chiffre. Ce genre d’exemple se rencontre dans des récits d’emploi du temps, dans lesquels le mois est sous-entendu car il a été mentionné plus haut. La date s’inscrit dans une progression logique liée au déroulé de l’emploi du temps. L’information reste précise. L’exemple 7 est ce qu’on pourrait appeler une date en

**description définie** : la date n'est pas explicite dans le texte, mais elle n'est pas ambiguë pour autant, l'expression correspond ici à la date des faits sur lesquels porte l'enquête.

8. Cette conversation a eu lieu en **début d'après-midi**. Je dirais vers **14H30-15H00**.

9. **Vers 12h45** je suis reparti au boulot à MOULINS. Je me rappelle qu'il y avait encore "attention à la marche" à la télé.

10. Après, je suis allé au magasin Super U de BRUGUIERES **avant la fermeture en fin d'après-midi**. Je suis ensuite rentré chez moi et j'ai du préparer à manger pour le retour de mon épouse.

Les exemples 8, 9 et 10 illustrent des cas d'horaires et de périodes de la journée. La mention de l'heure est teintée par l'incertitude du **témoin**, qui étaye l'heure mentionné à l'aide d'éléments extérieurs, comme à l'exemple 9 où le **témoin** cite une émission de télévision, ou à l'exemple 8 où il mentionne un intervalle horaire. En revanche, l'exemple 10 évoque une période de la journée sans horaire (bien que l'heure approximatif puisse être déduit en consultant les horaires du magasin visité). Dans tous ces exemples, l'information est moins précise que dans le cas des dates. Les informations fondées sur les horaires peuvent être plus facilement replacées sur une ligne chronologique, tandis que le sens d'expressions telles que « en début d'après-midi », « le soir », « en milieu de matinée » peut varier sensiblement d'une personne à une autre.

11. Avec quel véhicule votre père se rendait chez l'entreprise SIGMA à l'**époque de MOULINS**?

12. Puis il a rencontré Christine du **temps où il habitait à côté de STRASBOURG**.

13. Il y a **quinze à vingt ans** que Jacqueline travaille pour la DDASS et qu'elle héberge des enfants placés.

14. Elle s'est mariée à **21 ans** avec Didier GARCIA.

15. Question : Comment avez vous fait la connaissance de Séverine BERIN?

Réponse : C'est pendant les vendanges à Bergerac en **août 1993**.

Les exemples 11 et 12 concernent des périodes temporelles qui a priori peuvent aller de plusieurs mois à plusieurs années : ce sont des informations imprécises d'échelle large (d'autant plus imprécises que ces informations temporelles sont basées sur des éléments biographiques qui dépassent le cadre de la phrase dans laquelle elles sont mentionnées).

De même, l'exemple 13 concerne une période temporelle biographique définie relativement au présent. L'information est imprécise puisque le **témoin** emploie une fourchette d'années. L'exemple 14 apporte aussi une information biographique temporelle, mais cette fois on utilise l'âge de la personne pour situer un événement dans sa chronologie de vie. On peut en déduire une année, l'information est assez précise pour l'échelle biographique. Dans les deux cas, on en tire un ordre d'idée concernant les événements biographiques.

L'exemple 15 situe un événement avec la mention d'un mois et d'une année. Il s'agit de nouveau d'une information biographique, mais elle est plus précise que dans les exemples précédents.

Tous ces exemples ne sont pas exhaustifs de la variété rencontrée dans l'expression des éléments temporels au sein d'une **procédure** judiciaire. Ils permettent néanmoins de donner un aperçu des informations temporelles et des formes linguistiques prises par leurs expressions. Parmi les formes restantes, on trouve aussi les intervalles temporels, c'est-à-dire la combinaison de deux informations temporelles définissant un laps de temps. À l'issue de cet aperçu, on peut ajouter que les formes linguistiques des éléments temporels peuvent être divisés en trois catégories :

- Des éléments absolus, c'est-à-dire qui sont précis et que l'on peut replacer sur une chronologie indépendamment d'autres informations (c'est le cas des exemples 1, 2, 3 et 15),
- des éléments relatifs, c'est-à-dire dépendant d'une autre information pour être déduits (c'est le cas des exemples 6, 11, et 14),
- et des éléments temporels imprécis, c'est-à-dire désignant une période temporelle ou horaire qui n'est pas explicitement bornée (cas des exemples 8, 10, 11 et 12) ou qui présente un doute (cas des exemples 4 et 8).

Nous avons pu constater au travers des exemples que la catégorie des informations temporelles englobe plus que les dates, et que leurs réalisations linguistiques varient d'autant, notamment sous l'influence de la précision des informations et de l'échelle thématique en combinaison l'une avec l'autre, comme nous l'avons exposé en début de section.

### 1.1.3 Personnes

Le réseau de personnes entendu dans le cadre d'une affaire peut être très étendu et regroupe différents types de personnes : amis, famille, collègues pour les proches de la victime, membres du voisinage, **témoins** oculaires ou indirects, personnes venant témoigner spontanément à propos de faits indirectement connectés, etc. De nombreuses informations sont collectées sur les personnes, et de nombreuses personnes permettent de collecter l'information de la **procédure**. Nous allons donc nous intéresser aux formes prises par les mentions de personnes dans les **auditions**.

La mention d'une personne dans la **procédure** judiciaire sert en premier lieu à l'identifier, c'est-à-dire à déterminer son identité. L'identité, selon le dictionnaire Larousse, est un « ensemble des données de fait et de droit qui permettent d'individualiser quelqu'un (date et lieu de naissance, nom, prénom, filiation, etc.) »<sup>3</sup>, et ensuite à rassembler des informations à son sujet utiles aux investigations.

Les informations collectées par la **procédure** au sujet des personnes sont les éléments d'état-civil (âge, situation familiale et maritale, filiation), la situation sociale de la personne (situation professionnelle, situation financière, niveau d'études, engagements, loisirs, etc.), la description physique (dont l'état de santé), les éléments d'ordre psychologique (caractère, habitudes, relations avec l'entourage). Comme nous l'avons expliqué au chapitre III, l'objectif de la **procédure** concernant les personnes peut s'ordonner ainsi :

- Soit l'on dispose de l'identité de la personne, ce qui permet de l'identifier, dans ce cas on cherche à collecter des informations à son sujet pour comprendre son rôle dans les faits (de l'**entité nommée** vers l'entité criminelle),

---

3. <https://www.larousse.fr/dictionnaires/francais/identit%c3%a9/41420?q=identit%c3%a9#41315> [consulté le 24 octobre 2019]

- Soit l'on ne dispose pas de l'identité de la personne, dans ce cas, la collecte de l'information doit viser à la déterminer, en plus de la compréhension de son rôle dans les faits (de l'entité criminelle vers l'entité nommée).

En conséquence, les personnes peuvent apparaître sous différentes formes linguistiques dans le texte des auditions. Si l'on considère le nom et le prénom comme éléments basiques de l'identité, considérons les exemples suivants qui utilisent un nom de famille :

1. QUESTION : **Mme WOERTH** vous a t elle parlé de problèmes quelle aurait pu avoir ?
2. Question : Est ce que **M. PIÈGE** connaissait **Ghislaine WOERTH** ?
3. Lorsque le **couple KNOPE** habitait à Ivry-sur-Seine, nous allions souvent chez eux avec mes parents.
4. Je me nomme **ROUSSEL Chantal épouse GATTIAT**. Je suis née le 05 mars 1960 à VESOUL.
5. Nous avons rencontrés **M. et Mme FERRAND** et nous avons emménagé dans la maison environ un mois après en novembre 2001.
6. Je suis le frère de **Martine BASTY divorcée GALLARD**.
7. Je sais par ma soeur Anne que la **famille LEVASSEUR** qui habite rue de la poste à DIJON, près de la forêt, connaît la victime.

On constate que le nom est complété par le prénom, et/ou par des éléments complémentaires qui renseignent sur le titre de civilité, le statut marital. Contrairement à ce que l'on pourrait attendre au premier abord, un nom de famille n'est pas toujours immuable, notamment pour les femmes qui peuvent figurer sous leur nom de jeune fille, d'épouse ou de divorcée (exemples 3 et 4). De même, le nom de famille peut servir à désigner plusieurs personnes à la fois, comme à l'exemple 7 qui mentionne une famille complète, ou bien un couple comme aux exemples 3 et 5.

Beaucoup de personnes sont mentionnées uniquement par leurs prénoms. C'est le cas notamment des enfants, ou bien lorsque l'identité de la personne a déjà été énoncée auparavant dans le texte.

8. Que savez vous de la relation de **Camille** avec son ex-mari **Luc** ?

9. Par la suite sa copine, **Sarah**, l'emmenait et le recherchait lorsque son scooter était en panne.

10. Quelles étaient les habitudes de **Laurence** à la fin d'une séance de sport ?

11. Question : Concernant **Karine** et **Floriane** : leur avait-il de recevoir du monde à la maison ?

Réponse : **Karine** reçoit souvent son petit ami à la maison. **Floriane** ramenait ces copines d'école aussi de temps en temps. **Karine** avait un ancien copain, **Thomas** ; il y aussi **Hervé** et **Ludovic**, qui sont boulangers, qui sont déjà venus à la maison. **Karine** est assez réservée, elle n'a pas beaucoup d'amis.

Autre cas de figure avec l'exemple suivant : le **témoin** mentionne une personne qu'il connaît sous une appellation de type surnom. Le **témoin** est incertain de la nature du nom qu'il utilise.

12. Jeudi nous étions accompagné d'un jeune dont le père s'appelle ou est surnommé " **NEIL** " et qui habite à OBERNAI. Mon père ramène du bois à " **NEIL** " et son fils **Kylian** était venu avec nous.

Dans tous les cas que nous venons de citer, le **témoin** est capable de nommer la personne qu'il mentionne, que ce soit à l'aide d'un segment non-ambigu comme un nom de famille et un prénom, un titre de civilité et un nom, ou bien un prénom seul faisant abstraction du nom de famille associé, ou bien à l'aide d'un segment ambigu comme un prénom sans précision préalable du nom de famille associé, ou avec un surnom. Dans tous les cas, l'identification de la personne est soit immédiate, soit possible avec les autres éléments apportés par le **témoin**.

Dans le chapitre **III**, nous avons repris la définition de la **description définie** de NOUVEL et al. (2015). On rencontre des mentions de personne fonctionnant ainsi, comme l'illustrent les exemples suivants :

13. Stéphanie est retournée avec **le père de son fils**.

14. Question : Est ce que dans la famille, tout le monde a un couteau de poche? Réponse : Je sais que **mon père** possède un Laguiole.

15. D'après ce que je sais **la mère de Bernard** étaient à l'école normale en même temps que mon père ou ma mère.

16. Christophe a connu Laurence par l'intermédiaire de son copain Bruno car Laurence connaissait **la femme de bruno**.

17. Il me disait avoir appris ce qu'il était arrivé à **la concubine de DURAND**, il trouvait ça moche.

Si certains de ces exemples adhèrent à la structure de NOUVEL et al. (2015), c'est-à-dire une structure en *le* + substantif + individuante, ce n'est pas le cas de tous, comme l'usage de « mon père ». Néanmoins, nous rattachons ces cas à la catégorie des **descriptions définies**, étant donné que le substantif est ici caractérisé par l'adjectif possessif.

Les **procédures** comportent également des mentions de personnes beaucoup plus floues et incertaines, comme l'illustrent les exemples ci-dessous.

18. En début d'après-midi, je me trouvais dans la cour de mon domicile. J'étais occupé à bricoler. A un moment donné, que j'estime entre 15 heures et 16 heures, un homme qui était à vélo m'a appelé de la rue. Cet homme était de type européen, il paraissait propre, il était habillé d'une chemise et d'un pantalon. Il s'est exprimé en français. Il n'avait pas d'accent particulier.

19. Le conducteur est un homme d'environ 45-50 ans, de forte corpulence.

20. QUESTION : POUVEZ-VOUS NOUS DONNER UNE DESCRIPTION DE L'HOMME QUI L'ACCOMPAGNAIT ?

REPONSE : La description est la suivante : environs 40 ans, cheveux courts roux foncés, cheveux dégarnie sur le devant et au milieu, taille environs 1,75 — 1,85, 70 — 80 kg, corpulence moyenne, aucune lunette, aucune barbe, teint légèrement bronzé et tee-shirt à manches courtes vert foncé et un short à mi-cuisse beige. Il me semble qu'il avait une petite moustache mais je n'en suis pas sur. Je ne peux pas vous dire la couleur de ces yeux. Je ne l'ai pas entendu parler.

Il portait une montre à aiguille et le bracelet était vert foncé. C'est une montre de sport car elle avait la particularité d'avoir trois petits cadrans genre chronomètre et compte à rebours Je connais un peu les montres car mon père est employé dans une fabrique de montres « PATEK PHILIPPE »

21. QUESTION : Pouvez-vous me décrire l'individu en question ?

REPONSE : C'était un homme type maghrébin, assez maigre, environ 1m75, cheveux foncés courts, le visage plutôt en longueur. Il portait un jean bleu foncé, des chaussures de ville en cuir et une veste style anorak de couleur beige mais tirant vers le jaune. Il n'avait pas de gants et pas de bonnet.

22. Nous avons couru jusqu'à l'intersection suivante. A cet endroit sur le carrefour j'ai constaté la présence d'un individu qui m'a semblé bizarre. Il portait des vêtements de ville entièrement noirs et il avait un mouchoir blanc avec des carreaux violets dans la main. Il marchait vers moi c'est à dire en direction de CHATEAUBRIANT. J'ai trouvé d'autant plus son comportement étrange car il ne nous a même pas salué ou regardé, ce qu'on font généralement les gens en cet endroit. J'ai poursuivi mon chemin avec mes amies jusqu'à la route suivante qui est perpendiculaire au chemin. Nous avons ensuite fait demi tour sur le chemin et nous sommes donc repassées une seconde fois à la hauteur de l'individu en noir. Il n'a pas changé d'attitude et ne nous a pas adressé la parole. Je n'ai pas fait attention si il avait son mouchoir à la main. Je pense qu'il devait être environ 18 heures 45 à ce moment là. Cet homme était seul.

Question : Pouvez vous nous faire une description plus précise de cet individu ?

Réponse : Il s'agissait d'un homme de type européen, il avait cependant le teint mat, il devait avoir environ une cinquantaine d'année, il avait les cheveux bruns courts, il ne portait ni barbe ni moustache ni lunettes, il mesurait environ 1,70 mètre, de corpulence moyenne. Il portait un t-shirt noir à manches courtes et un pantalon style jean noir. Je n'ai pas fait attention à ses chaussures par contre elles étaient noires aussi.

Ces cinq mentions de personnes ne permettent pas d'accéder à leur identité. Elles sont constituées de descriptions subjectives à l'aide de traits variables et n'ont pas

de structure en particulier. En plus des critères anthropométriques type âge apparent, taille, corpulence, chevelure, couleur des yeux, etc., on relève aussi le mode de transport, l'habillement, la langue et l'accent, les accessoires. Ces caractéristiques sont parfois mentionnées pour leur absence : pas de barbe, pas de lunettes, la couleur des yeux est inconnue, etc. Les descriptions sont amorcées par un nom générique d'humain : on parle d'un « homme » ou d'un « individu ». Dans tous les exemples, le **témoin** adopte des stratégies de modulation :

- L'approximation : « j'estime », « environ », « assez », « plutôt »,
- L'usage de fourchettes temporelles, d'âge, de taille ou de poids, l'âge mentionné par dizaine,
- L'incapacité à citer une caractéristique : « je n'ai pas fait attention », « je ne l'ai pas entendu parler », « je n'en suis pas sûr »,

En plus de la modulation de la certitude, les descriptions comprennent aussi des éléments subjectifs relatifs à la perception de la personne et de la situation par le **témoin** : « [il] m'a semblé bizarre », « il ne nous a même pas salué » à l'exemple 22, « il paraissait » à l'exemple 18.

Face à ces exemples se pose la question de la délimitation textuelle de l'entité. Du point de vue de l'**enquête**, l'entité doit inclure le plus de critères de description possibles : on cherchera à obtenir le plus d'information à son sujet, ou bien les informations que l'on estime les plus significatives (selon l'appréciation de la personne conduisant l'**audition**). L'entité prend alors une forme linguistique qui ne correspond plus aux structures référentielles (**entités nommées** et **descriptions définies**) que nous connaissons déjà. Ceci pose la question de l'intégration des éléments dans l'entité : se limiter au nom générique qui l'amorce fait passer à côté des éléments caractéristiques et confond l'entité dans le reste des entités référencées par le même nom générique.

L'exemple 22 est particulièrement parlant. Le **témoin** a croisé la personne mentionnée à deux reprises, sa description est entrecoupée par le récit du cheminement du **témoin**. La description est relancée par l'**OPJ** en charge de l'**audition**. De même à l'exemple 20, on peut se demander s'il est utile de conserver les explications du **témoin** sur un accessoire (ici la montre), et si oui, comment procéder. Dans ces deux cas, la délimitation de la mention de l'entité pose question. Si l'on se limite au nom

générique qui initie la description, on néglige le reste des informations collectées qui sont pourtant tout l'intérêt et la matière de la description.

Si l'on replace la détection des **entités nommées** dans son objectif historique de constitution de base de données, ce type de mention de personne représente un défi. Il faudrait en effet tenter de proposer une modélisation de la description, c'est-à-dire déterminer les caractéristiques de base (les champs obligatoires de la base) et les caractéristiques supplémentaires (les champs facultatifs). Or, les descriptions ne sont pas stabilisées et l'on risque de manquer des informations dans les cas dépassant cette modélisation.

#### 1.1.4 Lieux

Plusieurs mentions de lieux figuraient dans les exemples précédemment cités. Le dictionnaire Larousse<sup>4</sup> en ligne propose la définition suivante pour « lieu » :

1. Situation spatiale de quelque chose, de quelqu'un permettant de le localiser, de déterminer une direction, une trajectoire : *Le lieu du rendez-vous n'est pas fixé.*
2. Endroit, localité, édifice, local, etc., considérés du point de vue de leur affectation ou de ce qui s'y passe : *Vous n'étiez pas sur votre lieu de travail.*

D'après cette définition, le concept de lieu relie dimension spatiale et usage. Dans le cas de l'**enquête**, le lieu délimite l'espace auquel on fait référence, et complète en cela les informations temporelles pour situer les faits évoqués.

Dans une perspective linguistique, HUYGHE (2009) souligne :

Lieu, endroit et place ont un vaste champ d'application référentielle, au sens où ils peuvent renvoyer à des segments du monde variés et hétérogènes.

Cette définition introduit deux autres concepts relatifs à l'espace et proches du lieu, l'endroit et la place, auxquels l'auteur ajoute également emplacement, zone, région, site, qu'il appelle des *noms généraux d'espace*.

Il existe une catégorie linguistique appliquée à la désignation des lieux : les toponymes. BOYER (2008) propose la définition suivante de la toponymie :

4. <https://www.larousse.fr/dictionnaires/francais/lieu/47076> [consulté le 24 octobre 2019]

Acte par lequel un nom propre est donné à un lieu.

Il apparaît d'après ces deux définitions que les lieux partagent avec les personnes la possibilité d'être désignés par des noms propres et par des noms généraux. Afin de recouvrir à la fois toponymes, noms généraux d'espace et d'autres désignations faisant référence à la situation spatiale, nous évoquerons pour notre étude des informations spatiales et géographiques plutôt que des lieux strictement, ce qui se justifie d'autant plus que les informations relatives à la situation dans l'espace peuvent, selon ce qui est rapporté, recouvrir différents niveaux et donc désigner différents types de lieux, d'endroits, etc. En fond, la question qui doit nous animer une nouvelle fois est celle de savoir qu'est-ce qu'une mention géographique intéressante pour l'analyse criminelle et quelle forme prennent-elles dans le texte des auditions. Comme dans le cas des informations temporelles, il n'y a pas de réponse toute faite à cette question, nous tâcherons d'y apporter un début de réponse en usant d'exemples tirés du corpus.

Le principe d'échelle que nous avons ébauché à propos des éléments temporels à la section 1.1.2 peut également s'appliquer aux éléments géographiques. Si l'on examine des informations biographiques ou des éléments distants dans le temps, l'échelle des informations spatiales sera plutôt de l'ordre des régions ou des villes (sous forme de toponymes). En revanche si l'on s'intéresse à un emploi du temps, son compte-rendu peut passer par la narration d'un itinéraire suivi au cours de la journée. Dans ce cas, les informations spatiales comprendront des voies, des adresses, des bâtiments, des lieux divers aux désignations diverses.

Par exemple, dans le cas d'une personne aperçue en train de cheminer avant de disparaître, on tentera de reconstituer son parcours au travers de plusieurs témoignages. Il sera utile de croiser les informations géographiques et temporelles et éventuellement d'autres sources comme le déclenchement de bornes téléphoniques. Dans ce cas, l'échelle de l'information géographique doit être très fine. Si l'on reprend le cas des faits sériels étalés sur plusieurs années, une trace de proximité géographique concomitante aux faits constitue déjà un indice solide. Par exemple, dans le cas de Francis Heaulme, tueur en série s'étant déplacé et ayant agi aux quatre coins de la France, les enquêteurs ont pu attester de sa présence dans les villes où les crimes étaient commis à l'aide de sources diverses (registres des amendes de la SNCF, des

foyers Emmaüs, des mairies dans lesquelles il bénéficiait de colis alimentaires, de registres d'hôpitaux dans lesquels il séjournait, ou encore des verbalisations de gendarmerie et de police pour vagabondage, bagarre ou ivresse manifeste<sup>5</sup>). Sans avoir à démontrer sa présence sur les lieux exacts du crime, sa présence à proximité au même moment représentait un indice concordant venant en renfort du reste de l'argumentation.

En somme, comme nous l'avons déjà vu pour d'autres entités, les informations géographiques et spatiales de la **procédure** concernent une vaste variété de concepts, et donc tout autant de formes linguistiques, que nous allons illustrer avec les exemples suivants.

1. Elle s'est mariée récemment et habite **Annecy**, c'était sa grande copine comme une soeur.
2. Sarah habite **près de CAEN** et Jessica habite **en Bretagne**.
3. Il y a environ sept ans, Clément travaillait **sur Lyon** dans une chocolaterie.
4. Je prends connaissance de l'objet de votre enquête suite à la personne décédé qui a été découverte sur un parcours sportif en **forêt de MONTMORENCY** dans la journée du samedi 12 mars 2010.
5. Je sais qu'il se promène **en forêt du côté de l'hôpital vers la route de Launaguet**.
6. La dernière fois que je l'ai vu, c'était il y a cinq ou six semaines **près du carrefour giratoire au niveau de la discothèque « Le Stanley » à FOIX**.

L'exemple 1 comporte une information spatiale précise exprimée par un toponyme, explicite et non-ambiguë, qui porte sur une information d'ordre biographique.

Dans l'exemple 2, qui apporte la même information, à savoir un lieu de résidence, l'information géographique est exprimée de manière moins précise : d'une

---

5. Voir : *Faites entrer l'accusé*, Francis Heaulme : le routard du crime, 2004 <https://www.youtube.com/watch?v=I66c5cN9HIQ>

part, on évoque la proximité d'une ville, et de l'autre part on mentionne une région. L'exemple 3, qui véhicule encore une fois la même information d'ordre biographique, utilise la structure *sur* + nom de ville, une structure décriée incorrecte par l'Académie Française<sup>6</sup> dont le sens semble osciller entre « dans la région de la ville en question » ou bien « dans la ville au sens large ». Dans ces trois cas, l'usage des toponymes désigne une zone plus large que dans le premier cas.

L'exemple 4 est le cas d'une information spatiale formée par la combinaison d'un nom de lieu et d'un toponyme. Cette structure nom générique + nom propre s'applique souvent à d'autres éléments naturels comme certains lacs (le lac d'Enghien, le lac d'Annecy, le lac du Bourget) mais pas tous (le lac Léman est parfois simplement désigné comme « le Léman »), les forêts et les bois (forêt de Sénart, forêt de Fontainebleau, bois de Vincennes). Il est intéressant de noter dans ce cas que la forêt de Montmorency se trouve en réalité à un peu plus de six kilomètres du centre de la ville de Montmorency. Cette structure peut donc avoir une influence sur la précision de l'information géographique rapportée.

L'exemple 5 est un exemple d'information géographique définie par plusieurs autres informations spatiales qui ne sont pas des toponymes : la mention d'un établissement et une voie. Leur combinaison cerne un secteur mais n'en marque pas de bornes fermes, l'information n'est donc pas très précise étant donné que « du côté de » peut être interprété de différentes manières.

L'exemple 6 est un exemple similaire, où le *témoin* mentionne une information géographique, « à Foix », et précise le lieu qu'il désigne par rapport à d'autres informations spatiales liées à un établissement et à une voie. Si l'information semble plus précise que dans l'exemple précédent, elle est toutefois également sujette à interprétation du fait du manque d'exactitude de « près de » et de « au niveau de ».

Dans leur diversité, le point commun de ces exemples est d'exprimer des localisations, c'est-à-dire un point géographique, emplacement d'une activité en particulier.

Comme autre type d'informations géographiques, nous avons repéré des exemples de cheminements et d'itinéraires.

---

6. Voir : <http://www.academie-francaise.fr/sur> [consulté le 28 octobre 2019]

7. A 17H00 je suis parti récupérer Léo à **la crèche du quartier des tournesols**. A la sortie de la crèche je suis passé au **garage Peugeot tout près**, j'ai parlé avec un vendeur j'y suis resté 10 minutes, le vendeur a pris mes coordonnées, je cherchais un véhicule 7-8 places. Je suis **rentré** avec Léo, il devait être 17H30.

8. Question : Quel chemin empruntez-vous exactement pour vous rendre chez vos parents depuis chez vous? Réponse : Je pars de chez moi, à l'hôpital je prends à droite en direction du garage PEUGEOT. Je descends la route de RENNES, jusqu'au Crédit-Mutuel puis je prends à gauche. Au panneau STOP, je prends à droite, puis la première sur la gauche, et là j'arrive en face d'un chemin qui rentre dans la forêt de VERRIERES. Ce chemin se trouve légèrement sur la gauche par rapport au petit parking où se garent les voitures.

L'exemple 7 exprime plusieurs déplacements du **témoin**, qui mentionne explicitement deux lieux, la crèche et le garage Peugeot. Le garage est mentionné relativement à la crèche, il est « tout près ». Néanmoins il faut déceler aussi une information géographique dans le verbe « rentrer », puisqu'il s'agit du point final de l'itinéraire évoqué et le lieu de la suite de l'emploi du temps. On note également que les informations géographiques sont complétées par des informations temporelles.

L'exemple 8 détaille un itinéraire routier, cette fois sans informations temporelles puisqu'il s'agit de décrire un itinéraire habituel. Plusieurs lieux sont évoqués, mais en réalité hormis le point de départ et d'arrivée, ces lieux servent de repères et ne sont pas à proprement parler le lieu de l'action décrite (« à l'hôpital », « au Crédit-Mutuel »). En plus des lieux, le **témoin** emploie des marqueurs spatiaux (« à droite », « en direction de », « légèrement sur la gauche », « en face de »), et des informations routières (un nom de voie et un panneau de signalisation). Dans cet extrait, les lieux sont des lieux de passage et non des lieux où se déroule une action en particulier. On relève que l'itinéraire est exprimé via la combinaison de lieux évoqués les uns par rapport aux autres.

Qu'il s'agisse d'une localisation ou d'un itinéraire, on constate que la mention d'information géographique dans plusieurs exemples se fait à l'aide de cette combinaison de plusieurs informations spatiales. Dans le cas des localisations, l'information peut être segmentée en éléments constitutifs :

[en forêt] [du côté de l'hôpital] [vers la route de Launaguet]

Chaque élément pouvant lui-même être découpé en un marqueur spatial (« en », « du côté de », « vers ») et un lieu. Cette segmentation permet de structurer les informations géographiques de localisation.

### 1.1.5 Véhicules

Tous les véhicules motorisés peuvent intéresser les investigations : un véhicule peut permettre de faire un lien avec une personne, qu'il s'agisse de deux roues, d'automobiles, d'utilitaires. Plus qu'un moyen de transport, la voiture personnelle est presque une extension du domicile dans laquelle il est possible d'effectuer une [perquisition](#) à la recherche d'objets et d'éléments utiles à l'[enquête](#).

La réglementation concernant les véhicules motorisés leur attribue un numéro unique d'immatriculation, visible à l'extérieur à l'avant et à l'arrière. Au premier abord, on pourrait donc imaginer qu'il faut s'appuyer sur ce numéro pour repérer les véhicules dans le texte des [auditions](#).

En réalité, la mention des véhicules est soumise aux mêmes contraintes d'expressions que les autres entités. La précision des propos, en lien avec la perception et la qualité des souvenirs des [témoins](#), entre de nouveau en jeu.

En effet un véhicule ne se désigne pas uniquement par un modèle ou une marque, et les [témoins](#) ne peuvent que rarement citer le numéro de plaque minéralogique entièrement, comme l'illustrent les exemples suivants :

1. Environ une semaine avant le meurtre de la dame, j'ai vu un véhicule stationné près du château d'eau en bordure de la route qui relie ORSAY à MASSY. En m'approchant de ce véhicule, j'ai vu que les deux portières avant étaient ouvertes et qu'un homme se tenait près de la portière passager avant. Je me suis arrêté à proximité de ce véhicule qui est de marque NISSAN, de couleur bleue foncée, assez haut, mais pas très long, de fabrication récente mais dont j'ignore le type et l'immatriculation.

2. J'ai un peu écouté ce qui se disait et j'ai compris que monsieur DAGENAIS Pierre de LACAPELLE-MARIVAL qui est un grand sportif aurait dit que

lorsqu'il s'entraînait en forêt de LACAPELLE-MARIVAL, un véhicule se serait dirigé vers lui à pleine vitesse [...] En ce qui concerne la voiture, elle serait bleue clair et l'immatriculation se terminerait par « 06 ».

3. Il s'agissait d'une camionnette assez longue, genre gros TRAFIC, elle était de couleur rouge, d'un modèle ancien.

4. Ce véhicule était de type break de couleur vert bouteille, il y avait des sièges à l'arrière. Je pense qu'il s'agissait d'une marque SUBARU, je pourrai reconnaître le modèle du véhicule si vous me présentez des photographies. J'ai bien vu que le véhicule était du département, il s'agissait bien d'un 59. Il me revient, je crois bien avoir lu sur le coffre la marque SUBARU. Il n'y a avait pas de barre de toit sur ce véhicule.

5. Question : Comment se déplace DURAND Patrick ? Réponse : Un vélo ou en mobylette. C'est un vélo de course ancien modèle, avec un guidon chromé, recourbé vers le bas. A la première vu on voit qu'il n'est pas récent. Il est de couleur vert militaire. Sa mobylette est de type moto de course type NSR, je cite ce modèle car je connais mais je ne sais pas si c'est cela exactement. C'est une mobylette 50 cm<sup>3</sup> qui ressemble à une 125 cm<sup>3</sup> avec carénage.

6. Je me rends tous les jours depuis mon domicile sur mon lieu de travail avec mon véhicule de service à savoir un véhicule de marque RENAULT, de type Clio immatriculé 000 AAA 00, de couleur bleue avec les petits logos du Conseil Général sur les portières et à l'arrière.

7. J'ai fait le tour de la maison et j'ai vu que la R5 blanche était présente mais que la Panda n'était pas là.

8. Dans votre véhicule de type super5, un mouchoir souillé de sang a également été retrouvé.

9. Je possède une Golf 4 bleue immatriculée en France.

La forme de ces mentions peut être mise en parallèle avec les mentions de personne que nous avons décrites en section 1.1.3. En effet, elles sont exprimées sur

des pans de textes pouvant aller de une phrase à un paragraphe, et font appel à des critères descriptifs tels que modèle, marque, couleur, numéro d'immatriculation, ancienneté de la fabrication, dimensions, signes distinctifs comme de la rouille, des autocollants, des sérigraphies, le nombre de portes, ou encore à des critères imprévisibles comme la présence de sièges. Comme l'exemple [voiture bleue] l'illustre, parfois le modèle et/ou la marque ne sont même pas cités.

Cet ensemble d'exemples peut se scinder en deux groupes : d'une part les véhicules que le **témoin** connaît (derniers exemples), et d'autre part les véhicules que les **témoins** ont vu ou dont ils ont entendu parler (premiers exemples).

Quand le **témoin** connaît le véhicule, il peut adopter différentes manières pour le mentionner. Dans l'exemple 6, le **témoin** précise le modèle, la marque, l'immatriculation, la couleur et les signes distinctifs de son véhicule à la demande de la personne qui conduit l'**audition**. La description est précise et permet d'identifier l'entité, elle est en quelque sorte d'ordre administratif.

Dans l'exemple 7, le récit concerne un emploi du temps. Le **témoin** se présente chez des amis et mentionne leur deux voitures, en utilisant seulement le nom des modèles et pour l'une sa couleur. On remarque dans l'exemple suivant que la même voiture est désignée sous une autre appellation. Le dernier exemple de ce type mentionne une voiture par son modèle, sa couleur, et le pays de son immatriculation. Hormis dans le cas où le **témoin** est invité à décrire sa voiture, les mentions de véhicules sont donc succinctes et servent soit à les désigner dans un récit plus large, soit à les décrire dans les grandes lignes.

En ce qui concerne les véhicules inconnus, les mentions sont plus longues et combinent le plus de critères de description possible. Tout comme les descriptions de personnes, la précision de ces descriptions est soumise à la perception, à l'attention et à la mémoire des **témoins**, et les critères peuvent être mentionnés pour leur absence (« pas de barre toit »). Les stratégies de modulation des propos sont aussi similaires aux descriptions de personnes, on exprime une subjectivité (« je pense », « j'ai vu », « je crois bien »), les propos sont nuancés par des adverbes (« assez », « pas très »). Les couleurs sont complétées par des adjectifs ou des noms (« foncé », « clair », « bouteille »).

Qu'ils soient connus ou inconnus des **témoins**, les véhicules sont mentionnés à

l'aide de noms génériques (« véhicule », « voiture », « camionnette », « vélo », « mobyette »), de noms de modèles (« trafic », « clio », « R5 », « panda », « super 5 », « Golf 4 »). On note l'usage fréquent de « type » (ou une fois de « genre ») pour spécifier la catégorie du véhicule (soit un modèle, soit un gabarit).

Les véhicules, à l'instar des autres entités que nous avons décrites, apparaissent dans les **auditions** sous des formes variées et l'on ne peut les réduire à un nom de marque ou de modèle. Les mentions de véhicules sont souvent subjectives et leur précision est influencée par l'incertitude des **témoins**. Les éléments de description font partie de l'entité, et comme dans le cas des mentions de personnes, si l'on remplace la détection automatique d'entité dans sa perspective historique de constitution de bases de données, la prise en compte de ces critères doit être examinée.

## 1.2 Synthèse : formes des entités dans les auditions et facteurs d'influence

Détailler les entités criminelles figurant dans les **auditions** de **témoignage**, ce que nous appelons les **entités criminelles textuelles**, a permis d'illustrer à travers des exemples la variété des formes qu'elles peuvent revêtir dans le texte, ainsi que leur place au sein des investigations. Cet inventaire n'est pas exhaustif, il est tout à fait possible de rencontrer d'autres formes linguistiques exprimant les mêmes entités.

Les entités figurent dans le texte sous la forme des deux concepts d'extraction d'information évoqués dans l'état de l'art, les **entités nommées** et les **descriptions définies**, toutefois ces deux concepts ne sont pas suffisants pour cerner l'ensemble des mentions faisant référence aux entités dans le texte. Ceci s'explique par le fait que la forme des entités est influencée par plusieurs facteurs. Il y a tout d'abord ce que nous appelons l'échelle d'intérêt : que ce soit au niveau du travail d'**analyse criminelle** en général, de l'**audition** ou de l'**enquête**, la personne en charge des investigations adopte un angle d'interrogation qui relève en général soit du biographique, soit de l'emploi du temps.

D'autre part, les **auditions** de **témoins** compilent des propos tenus par des personnes qui ne sont pas toujours certaines de ce qu'elles ont observé, qui ont parfois

mal vu ou qui ne se souviennent pas bien. Le **témoin** pouvant avoir du mal à se remémorer des faits même récents, la précision est également affectée par la qualité de ses souvenirs et de sa certitude ou incertitude.

Ainsi, les informations temporelles, les personnes, les informations géographiques et les véhicules apparaissent sous des formes peu normées et aux délimitations textuelles imprécises : des descriptions *indéfinies*. Dans cette catégorie, nous plaçons en particulier les descriptions rencontrées dans les exemples relatifs aux personnes et aux véhicules. S'il est délicat de stabiliser ces descriptions indéfinies sous un schéma lexico-syntaxique comme celui des **descriptions définies**, on peut néanmoins remarquer que ce modèle d'entité semble comporter au moins une mention d'un nom générique de la classe à laquelle elle fait référence (« personne », « homme », « individu », « camionnette », « voiture », « mobylette », etc.), faisant en quelque sorte office d'amorçage de l'entité. Dans le cas des informations géographiques, cet amorçage peut être un toponyme, c'est-à-dire un nom propre de la classe.

Les **auditions** s'inscrivant dans un processus de collecte d'information, elles doivent comporter le plus d'information de nature à identifier l'objet auquel il est fait référence, ce qui se répercute sur la longueur des descriptions exprimant des entités. Enfin, la description est le reflet de la perception subjective que le **témoin** a eu de l'entité et le fruit de sa mémoire, ce qui peut, comme nous l'avons vu, donner lieu à des descriptions basées sur des critères inattendus et rendant compte des impressions produites sur le **témoin**, tout cela en fonction de la qualité de ses souvenirs.

En somme, il apparaît que les **entités criminelles textuelles**, si elles recouvrent le concept d'entités nommées du **TAL** en comprenant aussi bien des **entités nommées** que des **descriptions définies**, le dépassent en incluant ce que nous appelons des descriptions indéfinies. Ces descriptions indéfinies découlent directement de la nature et de la fonction du texte abordé.

L'application des méthodes de **TAL** à la **procédure** judiciaire doit considérer les spécificités de l'expression des informations dans ce cadre. La prise en compte des descriptions indéfinies et de leurs éléments constituants permettrait d'opérer des rapprochements entre elles, ainsi qu'avec les **entités nommées** et les **descriptions**

définies dans l'objectif de construire une base de connaissance des entités de la *procédure*. En effet, détecter les *entités criminelles textuelles* est une première étape dans la construction de solutions de gestion automatique de l'information en analyse criminelle. Plutôt qu'un horizon indépassable, il s'agit d'un premier verrou que l'on doit chercher à faire sauter le plus largement possible, c'est-à-dire en ayant conscience dès maintenant de la variété des formes d'expression des entités.

Toutefois à l'heure actuelle et compte-tenu du matériau textuel disponible pour nos recherches, nous dirigeons la suite de nos travaux vers le test d'une approche de détection minimale orientée sur les *entités nommées*, c'est-à-dire sur des *entités criminelles textuelles* précises et possédant une structure modélisable. Leur détection serait déjà un progrès pour les *analystes criminels*, il est donc quoiqu'il en soit intéressant de se pencher sur leur détection.

## 2 Détection d'entités nommées criminelles

L'exploration des réalisations linguistiques qui correspondent aux entités criminelles a permis de mettre à jour leur grande diversité, aussi bien entre catégories d'entités qu'au sein de chaque catégorie. Cette double dimension de variété multiplie les formes à prendre en compte, ce qui représente une difficulté dans le développement de notre approche. Dans cette section, nous abordons la conception et le test d'une approche de détection des entités basée sur les ressources disponibles (*corpus* de *procédure* et ressources externes). Pour cela, nous revenons vers une « définition TAL » des entités, celle de segments référençant un concept du monde réel qui nous intéresse.

Pour développer cette approche, nous devons nous appuyer sur des données de *procédure*. Parmi celles que nous explorées jusqu'à présent, toutes sont au format PDF issu soit de la numérisation des documents papier, soit des fichiers traitement de texte originaux. Une seule présente une qualité d'OCR satisfaisante, condition *sine qua non* de l'application de traitements automatiques sur le texte. Cette *procédure* rassemble 370 *auditions* de *témoins*, pour un total de 594 631 mots<sup>7</sup>. Cet ensemble d'*auditions* nous servira donc de *corpus* de travail.

---

7. Statistique fournie par TXM.

## 2.1 Conception

### 2.1.1 Choix d'une approche

Comme nous l'avons expliqué au chapitre III, les approches de détection des entités nommées sont structurées en trois tendances : statistiques, symboliques et hybrides. En pratique, le choix d'une approche est guidé par les contraintes du projet (temporelles, humaines) et par les ressources disponibles (données de développement, ressources externes).

Pour donner un ordre d'idées, le travail d'anonymisation de documents cliniques de GROUIN (2013) repose sur un corpus de 21 749 documents médicaux (p. 131) qui lui permet de confronter les résultats d'approches symboliques et statistiques. Ce volume de données permet de réaliser des extractions et des échantillonnages, par exemple pour s'assurer que toutes les catégories d'entités traitées sont présentes dans le corpus ou encore pour renouveler les données de test. EHRMANN (2008, p. 201), quant à elle, exploite deux corpus pour la construction d'une ressource d'entités nommées : un corpus contenant l'ensemble des articles du journal *Le Monde* de 1992 à 1996 (2 830 180 phrases) et un corpus contenant articles et dépêches provenant de différentes sources et traitant tous de la crise en Côte d'Ivoire entre 2002 et 2003 (331 433 phrases).

Ces deux exemples sont donc largement mieux dotés en données. En comparaison, notre corpus de 370 textes compte environ 35 000 phrases<sup>8</sup>, une moyenne d'environ 95 phrases par document.

De manière générale, les campagnes d'annotation et de détection d'entités nommées reposent sur des corpus bien plus vastes que le notre et il est fréquent que les équipes mobilisent au minimum deux ou trois personnes. Le besoin de corpus pour la détection des entités nommées a conduit au développement spécifique de ressources qui sont distribuées sous diverses licences. Parmi les plus fréquemment utilisés dans la recherche française, on trouve notamment les corpus ESTER 1 et 2, ETAPE, QUAERO. L'avantage de ces ressources, au delà de leur volume, est d'être « prêtes à l'emploi », largement traitées et donc annotées, et déjà discutées dans la

---

8. Ce chiffre est obtenu par décompte de l'étiquette SENT, qui marque les ponctuations de fin de phrase dans le corpus étiqueté en parties du discours par TreeTagger dans TXM. Nous reviendrons sur l'étiquetage morpho-syntaxique du corpus dans le chapitre V.

littérature. Or comme nous l'avons abordé dans la première partie de ce chapitre, les **entités criminelles textuelles** revêtent des formes bien particulières, et de plus, il est avéré que le genre textuel est un paramètre qui influe sur la qualité des performances de systèmes de détection des **entités nommées** (NADEAU et SEKINE, 2007). Nous approfondirons cet aspect dans le dernier chapitre du manuscrit consacré au genre textuel des **auditions**, et faute de mieux, le développement d'une approche se fera à l'aide de notre petit **corpus**.

Les avantages des méthodes statistiques sont leur robustesse et la possibilité de traiter rapidement de grands volumes de données. En contrepartie, ces approches nécessitent des **corpus** de données annotées couvrant le plus de variantes possibles des cas que l'on cherche à annoter. Un autre inconvénient est l'aspect « boîte noire » : il est difficile de comprendre les choix opérés par le système, ce qui rend l'amélioration du système difficile (GROUIN, 2013, p. 76).

À propos des systèmes symboliques, NOUVEL et al. (2015, p. 92) indiquent qu'« [ils] prédominent pour les langues ou les typologies pour lesquelles il n'existe pas de **corpus** de données annoté de taille suffisante », ce qui est notre cas. Ils soulignent également que les systèmes symboliques permettent un meilleur contrôle des mécanismes, entraînant une meilleure précision des résultats mais générant également plus de silence. Cette observation s'accorde avec celle de MAUREL et al. (2011) qui remarquent que les approches symboliques sont les plus adaptées pour détecter les **entités nommées** si l'on a les moyens de développer des ressources linguistiques. GROUIN (2013, p. 63), dans le cas de l'anonymisation de documents cliniques, indique que les méthodes à base de règles sont optimales sur des documents présentant les caractéristiques suivantes :

Des documents de même origine (*un seul hôpital, un seul service au sein de cet hôpital*), issu d'un traitement de texte (*des documents correctement rédigés et bien formatés*), et dans une moindre mesure, des documents de même type (*lettres, comptes rendus, résultats d'examen, etc.*).

Compte tenu de notre situation, à savoir disposant d'un **petit corpus de bonne qualité**, rassemblant des documents de même origine et de même type, nous portons notre

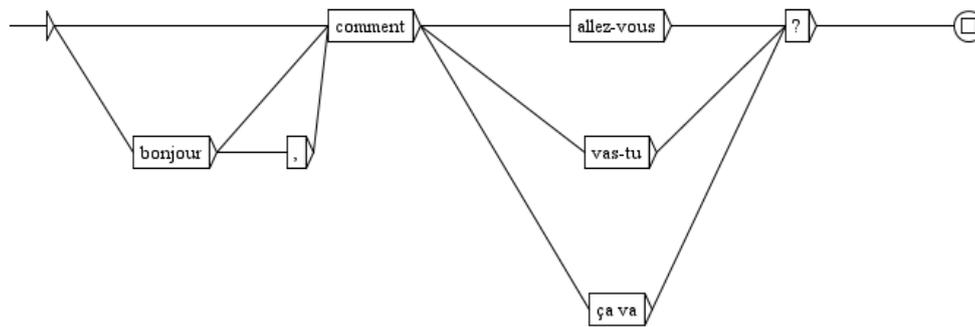


FIGURE 19 – Exemple de grammaire UNITEX

choix vers les approches symboliques. Nous nous intéresserons en particulier au mécanisme des grammaires locales.

### 2.1.2 Grammaires locales

La mise en œuvre d'approches symboliques pour le TAL exige soit de développer une chaîne de traitement de bout en bout (de l'entrée du texte à la sortie au format souhaité), soit de réutiliser une ou des solutions pré-existantes. La deuxième solution présente la plupart du temps l'avantage de ne pas avoir à réaliser les tâches de pré-traitement du texte comme le découpage en mots (tokenisation) et de plus facilite l'application des ressources externes.

UNITEX est un logiciel libre développé à l'université de Marne-la-Vallée qui permet le développement à l'aide d'une interface graphique de grammaires locales de type automates à états finis, disponible en téléchargement à l'adresse <https://unitexgramlab.org/fr>. Le mécanisme des grammaires locales a été décrit extensivement notamment par CONSTANT (2003) : il s'agit de définir des règles sous formes de graphes.

Le texte passe sous forme de chaîne de tokens (suites de caractères séparés par des espaces ou de la ponctuation) dans des « boîtes » de transition qui définissent des conditions. Si la suite de tokens examinée est acceptée par l'un des chemins de l'automate, elle est validée et donc considérée comme répondant aux contraintes que l'on a définies.

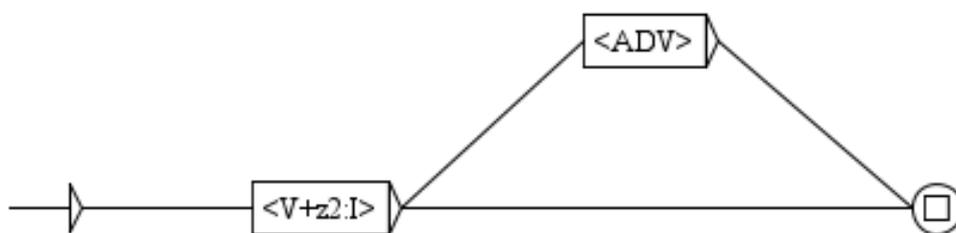


FIGURE 20 – Exemple de graphe utilisant des masques lexicaux

Le texte « entre » à gauche du graphe puis progresse dans le graphe à condition de valider les conditions précisées dans les boîtes. Chaque boîte correspond à un « état » de l'automate, que l'on appelle également « transition ». Si le texte est validé par toutes les boîtes du graphe, il ressort à droite et est considéré comme reconnu par l'automate.

L'exemple de la figure 19 est une grammaire qui reconnaît des suites du type « bonjour, comment ça va ? », « comment allez-vous ? ». « bonjour » ainsi que la présence de la virgule après sont facultatifs, en revanche « comment » et le point d'interrogation sont obligatoires (tous les chemins passent par leurs boîtes). « bonjour ça va ? » ne serait donc pas reconnu par cette grammaire.

Ici, le contenu de chaque transition est un mot ou une expression, pour permettre la lisibilité de l'exemple à un lecteur non averti. Préalablement à l'utilisation des graphes, UNITEX permet l'application de dictionnaires et de graphes de catégorisation morphe-syntaxique qui étiquettent les textes en entrée du logiciel. Ces étiquettes peuvent ensuite être utilisées dans les graphes (figure 20 : la première boîte correspond à la recherche d'un verbe à l'imparfait de niveau de langue spécialisé selon les termes du dictionnaire appliqué, suivi ou non d'un adverbe). De même, les graphes peuvent s'imbriquer : un graphe peut faire appel dans une boîte à un autre graphe. Dans ce cas, le texte doit être reconnu par le sous-graphe pour progresser dans le graphe principal.

Pour de plus amples détails sur le fonctionnement d'UNITEX, nous renvoyons à son manuel : PAUMIER (2016).

UNITEX est une solution clé en main pour le développement d'approches par

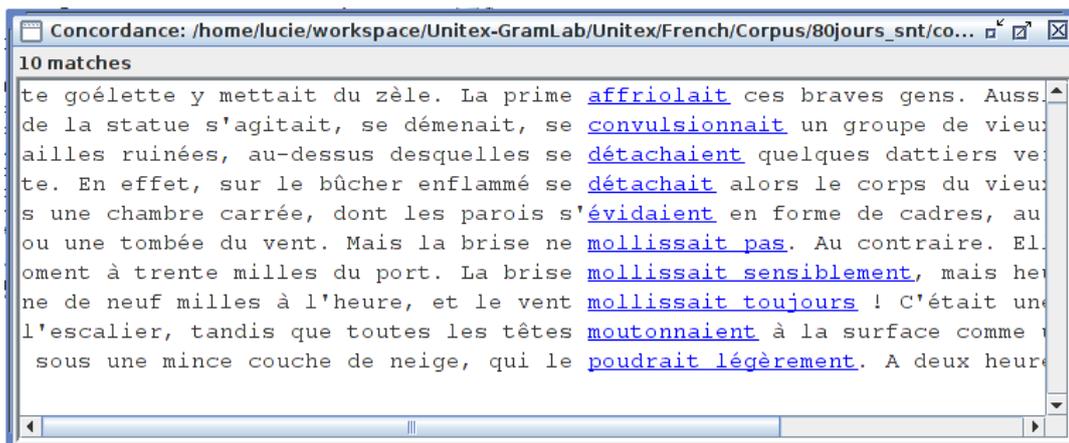


FIGURE 21 – Sorties du graphe de l'exemple 20 sur *Le tour du monde en 80 jours* de Jules Verne

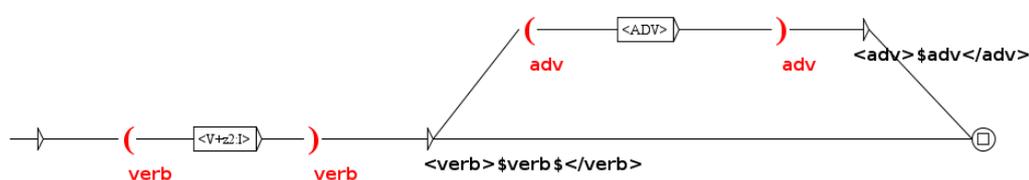


FIGURE 22 – Exemple de graphe transducteur

règles, et présente l'avantage de pouvoir combiner différentes stratégies de capture des motifs qui nous intéressent : usage d'un dictionnaire, description des contraintes, combinaison des approches. Étant donné que nous avons l'ambition de repérer différentes catégories d'entités sous différentes formes, la combinaison de stratégies apporte souplesse et adaptabilité.

Ce type de solution a déjà été éprouvé par MAUREL et al. (2011). Leur système repose sur la fonctionnalité de cascades d'UNITEX, qui déclenche des graphes les uns après les autres et dans un certain ordre, et sur un type de graphe particulier, les transducteurs. Les transducteurs sont des graphes qui produisent une sortie modifiée : lorsqu'une unité est reconnue, elle est encadrée par des balises, comme à l'exemple 22. La combinaison de cette fonctionnalité avec les cascades permet de réinvestir les sorties des premiers graphes de la cascade dans les graphes suivants. CasEN a été testé sur du texte journalistique, le corpus ESTER 2, et des entretiens sociologiques transcrits, le corpus ESLO 1.

La construction de grammaires par automates a été décrite par NOUVEL et al. (2015, p. 89) :

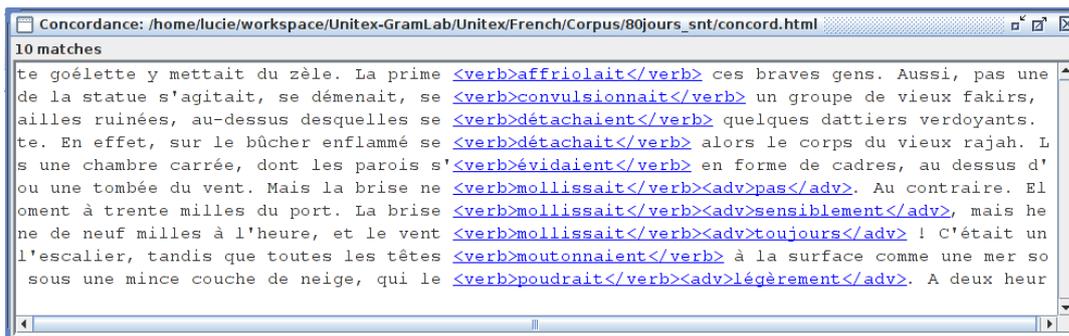


FIGURE 23 – Sortie du transducteur de la figure 22 appliqué sur le corpus 80 jours.

[...] L'enjeu est de contraindre correctement l'automate, afin qu'il reconnaisse toutes les expressions linguistiques souhaitées, et aucune autre. Ceci suppose de procéder par essais-erreurs, soit en examinant les résultats, soit en calculant des performances sur un jeu de données de développement.

Dans notre cas, nous avons opté pour l'examen des résultats.

### 2.1.3 Formes linguistiques prises en compte

Nous avons vu dans la section précédente que cinq types d'entités intéressent tout particulièrement l'analyse criminelle. Malheureusement, nous avons également constaté que les entités criminelles revêtent des formes linguistiques qui les éloignent des entités nommées et donc des stratégies classiques employées pour leur détection, c'est le cas notamment des personnes, des lieux et des véhicules. Le seul type d'entité à la forme stabilisée est les numéros de téléphone, entité qui ne fait pas intervenir de critères linguistiques dans sa formation.

Dans l'optique d'une progression par étape, il nous faut déterminer quelle entités sont détectables et sous quelles formes.

Nous avons constaté lors de la constitution de l'état de l'art que la reconnaissance d'entités nommées détecte principalement des éléments que l'on peut décrire sous une norme ou un format incluant des critères linguistiques. Ceci vaut aussi bien pour les approches statistiques, qui doivent décrire les entités pour les besoins de l'annotation du corpus d'entraînement, que pour les approches symboliques, pour lesquelles la description des entités en termes linguistiques est la base sur laquelle

reposera le système. Le processus de définition des entités préalable à leur détection peut donc, d'une certaine manière, être envisagé comme une phase de normalisation.

Or parmi les éléments que nous avons relevés qui correspondent à la mention d'une entité criminelle, tous ne sont pas « conformables » à un schéma prédéfini. C'est la conclusion de notre section précédente, dans laquelle nous avons relevé la mention d'entités criminelles par le biais de descriptions indéfinies.

Les **entités nommées** ne coïncident donc pas complètement avec toutes les formes prises par les entités criminelles dans le texte. Cela dit, on rencontre quand même des entités criminelles sous la forme d'**entités nommées**. Détecter les *entités nommées criminelles* pourrait être à la portée d'un système embryonnaire, serait déjà un apport pour l'**analyse criminelle**, et permettrait éventuellement de découvrir des particularités de la détection des entités nommées criminelles qui nous auraient échappé.

Nous avons constaté que les entités nommées criminelles peuvent être les noms de personnes, les noms de lieux, les dates. Le cas des véhicules est plus délicat, même lorsqu'elles apparaissent comme des **entités nommées**, cela peut être sous la forme d'un nom propre (nom de modèle ou de marque) ou sous la forme d'un nom générique de classe. Les trois types d'entité semblent pouvoir répondre à un « bornage » traduisible en règles de description applicables dans le cadre d'une approche symbolique.

**Personnes** Nous chercherons à détecter les noms de personnes lorsqu'ils apparaissent sous une forme permettant d'identifier un individu formellement, c'est-à-dire sous des combinaisons différentes des segments titre de civilité, nom, prénom (éléments basiques de l'identité) :

- « nom + prénom » (« Dupond Marie »),
- « prénom + nom » (« Marie Dupond »),
- « titre de civilité + nom » (« Mme Dupond »)

À ces formats de base, nous avons ajouté également à titre exploratoire :

- Les noms de couple : « le couple Dupond », « les époux Dupond »,
- La prise en compte du nom de jeune fille : « Marie Dupond née Martin »,

- La prise en compte du nom marital : « Marie Dupond épouse Martin »,
- La prise en compte du nom de divorcée : « Marie Dupond divorcée Martin ».

**Noms de lieux** L'examen des mentions textuelles d'entités criminelles de type lieux a révélé que ceux-ci pouvaient être composés de plusieurs éléments d'information spatiale, dont des noms propres de lieux. Nous tenterons donc de reconnaître les noms de communes et de villes françaises dans l'espoir que leur détection pourra servir de base pour la propagation de la détection des informations géographiques.

**Dates** Les dates ont été détectées également lorsqu'elles apparaissent sous une forme permettant d'identifier formellement un jour, c'est-à-dire en combinaison des segments jour, mois et année, y compris les formats de chiffres séparés par des barres obliques. Nous avons également cherché à repérer des dates « relatives », au format « jour de la semaine + numéro ». Certaines dates sont exprimées selon des formes propres à la [Gendarmerie nationale](#) : en toutes lettres, l'année précédant parfois le mois et le jour.

L'an deux mille sept, le vingt neuf septembre Nous soussigné (s) LUBSCK Thierry, Adjudant et GRENIER Stéphanie, Gendarme, en résidence à la Section de Recherches de LILLE, Officiers de Police judiciaire [...]

A DINAN, le trente octobre deux mil deux à seize heures quarante. Lecture faite par moi des renseignements d'état-civil et de la déclaration ci-dessus, j'y persiste et n'ai rien à y changer, à y ajouter ou à y retrancher.

## 2.2 Mise en œuvre

### 2.2.1 Pré-traitements

Les documents [PDF](#) des [auditions](#) comportent un fichier par [audition](#). Les fichiers sont nommés par une référence au sous-dossier dans lequel ils ont été classés, suivi d'un numéro, suivi du nom de la personne auditionnée et parfois d'une information complémentaire. Par exemple :

A2 35 - Audition de DUPONT Pierre.pdf

B5 12 - Audition de MARTIN Brigitte - amie de la victime.pdf

**Question** : Comment avez vous fait la connaissance de [redacted] ? -----  
**Réponse** : J'ai fait sa connaissance par l'intermédiaire de mon mari qui était un copain de [redacted], son ex mari. -----  
**Question** : Quand avez vous commencé à vous fréquenter ? -----  
**Réponse** : Je crois que cela date de 1978. C'était avant que nos deux couples se marient. -----

FIGURE 24 – Usage des tirets pour remplir les lignes dans les **procès-verbaux**.

Cette organisation est propre à cette **procédure** en particulier et n'est pas une norme.

La reconnaissance de caractère a été effectuée par le logiciel commercial Omnipage, que les **analystes criminels** utilisent au quotidien. Les sorties ont été enregistrées sous forme de documents texte brut, en conservant l'organisation selon laquelle un fichier correspond à une **audition**. Le nettoyage du texte a inclus la suppression des espaces surnuméraires, et la suppression des lignes de tirets insérées comme à la figure 24.

Nous n'avons pas distingué le corps de l'**audition** à proprement parler du texte d'en-tête et des informations d'état-civil qui se situent avant la transcription de l'échange pour deux raisons. D'une part, les éléments d'en-tête comportent les informations relatives au cadre de l'**enquête** : la situation géographique, la référence du **procès-verbal**, le nom du magistrat mandant, le numéro de commission rogatoire, la date de réalisation du **procès-verbal**, etc. Ces informations ne concernent pas directement le récit des faits mais apportent des informations qui pourraient s'avérer utiles pour de futur traitements (par exemple, pour la résolution des dates référentielles : si l'on rencontre « hier » dans le texte, il est nécessaire de connaître la date de l'**audition** pour retrouver la date à laquelle le locuteur fait référence). D'autre part, les éléments d'état-civil figurent soit sous la forme d'un tableau (figure 25), soit sous la forme d'un texte à la première personne (figure 26). Distinguer les différentes zones du document ne nous pas semblé prioritaire à ce stade des recherches, nous avons donc pris en compte l'ensemble du texte d'un seul tenant.

Pour repérer les entités nous avons développé des grammaires locales reposant, selon les entités, sur des stratégies différentes, conformément une fois de plus à ce qu'énoncent NOUVEL et al. (2015, p. 86) :

Dans un grand nombre de cas, les reconnaissances sont réalisées par prise en compte conjointe de multiples indices morphologiques et lexicaux, qui portent sur une partie de l'**entité nommée**. En voici quelques

Nous trouvant à la Brigade Territoriale de \_\_\_\_\_, entendons :

PERSONNE CONCERNÉE			
Nom		Prénom	
Nom usage			
Date naissance	Lieu naissance	Départ.	Pays
			F
Sexe	Situation Familiale	Nom marital	Nationalité
F	mariée		F
Filiation	Téléphone		Profession
De _____ et de _____	_____		_____
Adresse		Code postal	Ville
_____		_____	_____

qui déclare le \_\_\_\_\_ à quatorze heures et dix minutes : -----

FIGURE 25 – État-civil au format tableau

*Mission :*

Nous faisons comparaître devant nous, le témoin ci-après nommé et lui donnons connaissance des faits pour lesquels sa déposition est requise. ----

Je me nomme \_\_\_\_\_. Je suis né le \_\_\_\_\_ à \_\_\_\_\_ ( \_\_\_\_\_ ). Je suis fils de \_\_\_\_\_ et de \_\_\_\_\_. Je suis de nationalité française et je demeure 25 route de \_\_\_\_\_ à \_\_\_\_\_. J'exerce la profession d'\_\_\_\_\_. Mon numéro de téléphone est le \_\_\_\_\_.

Après avoir prêté serment de dire toute la vérité, rien que la vérité, le témoin, entendu séparément et hors la présence de la personne mise en examen, le témoin dépose ainsi qu'il suit :

Je suis le frère de \_\_\_\_\_ divorcée \_\_\_\_\_ . ----

FIGURE 26 – État-civil au format narratif

illustrations :

- *Personnes* : le premier mot est un prénom, le second un nom propre ;
- *Dates* : le premier et le dernier mot sont composés de chiffres, le second mot fait partie de la liste des noms de mois (*5 juillet 2012*) ;
- *Lieux* : contient *sur* ou *en* suivi d'un nom de cours d'eau (*Montlouis sur Loire*) ;
- etc.

Nous développerons nos grammaires sur un principe similaire, dont les détails sont présentés dans les sections suivantes.

### 2.2.2 Approche à base de lexique : cas des noms de lieux

Pour détecter les noms de lieux, étant donné que nous ne détectons que des noms de villes et de communes, nous avons décidé d'utiliser une liste pour faire office de dictionnaire.

Il existe notamment la base lexicale GeoNames<sup>9</sup>. Il s'agit d'une base rassemblant de nombreuses informations géographiques (plus de 7 millions d'entités et

9. <https://www.geonames.org/> [consulté le 05 novembre 2019].

```
Ayguesvives, .N+Loc  
Ayguetinte, .N+Loc  
Ayherre, .N+Loc  
Ayn, .N+Loc  
Aynac, .N+Loc  
Aynans, .N+Loc  
Ayrens, .N+Loc  
Aysse, .N+Loc
```

FIGURE 27 – Extrait du dictionnaire des communes

10 millions d'entrées lexicales, selon NOUVEL et al. (2015)). Le fichier d'informations françaises compte plus de 140 000 lignes. Malheureusement, ces informations sont très disparates et incluent aussi bien des toponymes que des bâtiments et des routes (on y trouve par exemple une liste des hôtels Campanile). La base Geonames est donc une ressource très riche mais qui semble peu adaptée au stade de nos recherches. Nous préférons employer une ressource de moindre ampleur à la mesure de nos moyens de contrôle et d'adaptation, comme la liste des noms de communes françaises libre de droits et disponible à l'adresse : <https://sql.sh/736-base-donnees-villes-francaises> [consulté le 31 octobre 2019]. Cette liste comporte 36 700 entrées à la date de consultation, ce qui concorde avec les données de l'INSEE (INSEE, 2018). Le fichier au format CSV a été remanié pour le conformer au format de dictionnaires attendu par UNITEX, le format .dic, dans lequel les entrées sont associées à des codes. Le fichier CSV d'origine étant composé de plusieurs colonnes rassemblant des informations comme le département, le code postal, la population, des données géographiques, etc., la colonne dite « nom réel », qui contient le nom officiel des communes avec accents et tirets, est extraite. À chaque entrée est ajouté le code « .N+Loc », qui sera l'étiquette attribuée aux mots du texte matchant une entrée du dictionnaire lors de son application.

Afin de faciliter la gestion de la liste, les entrées ont été triées alphabétiquement et dédoublonnées. En effet certaines villes en France partagent le même nom, cette distinction apparaît dans le fichier original, mais étant donné que nous n'exploitons que le nom et pas les autres informations du fichier, il n'est pas nécessaire de conserver plusieurs entrées pour un même nom.

Une fois la ressource constituée et appliquée, la détection des villes est ensuite très simplement déclenchée par un graphe à une boîte contenant le masque lexical

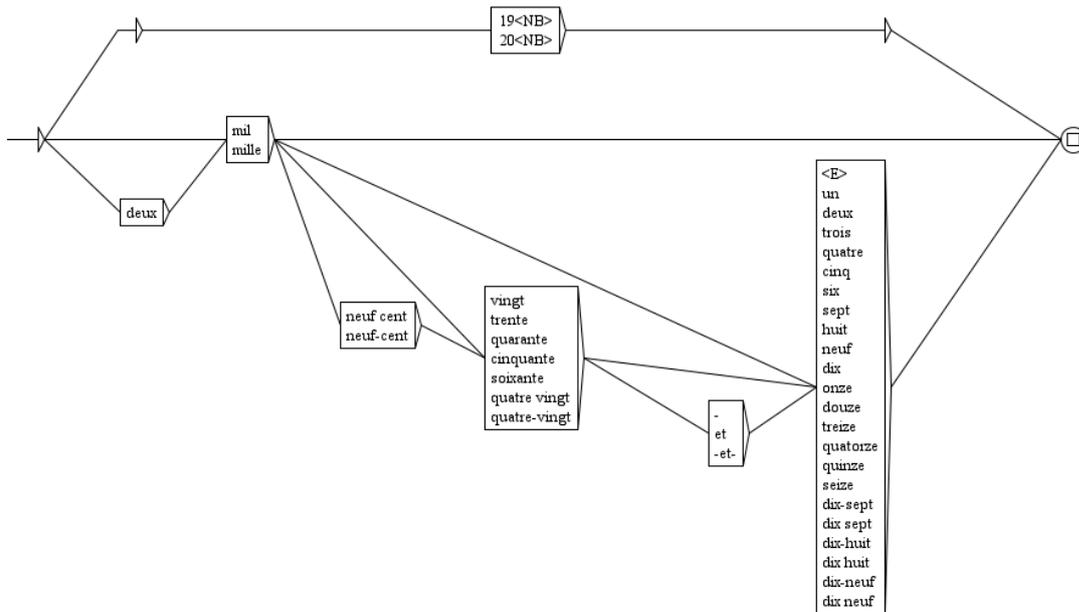


FIGURE 28 – Exemple de graphe détectant les années en chiffres et en toutes lettres.

<N+Loc>.

### 2.2.3 Approche à base de règles : cas des dates

En ce qui concerne les dates, nous avons opté pour l'intégration complète des contraintes dans les graphes. En effet, les différents formats de dates peuvent être circonscrits en un ensemble fini d'éléments agencés en différentes combinaisons.

Ces segments sont le jour de la semaine, le jour, le mois et l'année. Un sous-graphe pour chacun de ces segments a été créé puis inséré dans un graphe principal. Les sous-graphes consacrés à des valeurs numériques sont déclinés en chiffres et en toutes lettres. Nous distinguons quatre types de dates : les dates en toutes les lettres qui peuvent être de deux types (« 21 août 2014 » ou « deux mil neuf le vingt juillet »), les dates numériques (« 12/03/2003 ») et les dates relatives (« vendredi 18 »). Chaque type de date correspond à un chemin du graphe.

### 2.2.4 Approche mixte lexiques et règles : cas des noms de personnes

Les noms de personnes sont détectés en combinant l'application d'un dictionnaire et de règles descriptives. Le principe général est le suivant : si un prénom est détecté, on cherche un mot en majuscules situé immédiatement avant ou après.

```

Vicky, .N+Firstname
Victoire, .N+Firstname
Victor, .N+Firstname
Victor-Alexandre, .N+Firstname
Victor-Emmanuel, .N+Firstname
Victoria, .N+Firstname
Victoria-Lynn, .N+Firstname
Victorian, .N+Firstname
Victoriana, .N+Firstname

```

FIGURE 29 – Extrait du dictionnaire des prénoms

La détection des prénoms est donc basée sur l'utilisation d'un dictionnaire constitué de la liste des prénoms déclarés à l'état-civil en France entre 1900 et 2015. Cette liste, régulièrement mise à jour par l'INSEE, est disponible à l'adresse suivante : <https://www.insee.fr/fr/statistiques/2540004#consulter> [consulté le 31 octobre 2019]. Elle comporte le genre du prénom et le nombre d'attribution par année. En termes de retraitement, la colonne des prénoms a été extraite, chaque prénom a été passé de tout en majuscule à majuscule au début uniquement (de « CATHERINE » à « Catherine »), et les entrées ont été dédoublonnées. Certains prénoms ont dû être éliminés de la liste pour cause d'ambiguïté : les prénoms de moins de trois lettres, ainsi que le prénom « Elle ». Le dictionnaire a ensuite été constitué en ajoutant après chaque entrée le masque lexical <N+Firstname> (figure 29). Après l'application de cette ressource, les prénoms servent de déclencheur de la majorité des chemins de détection des noms de personnes (figure 30).

Cinq sous-graphes sont construits pour détecter chaque élément constituant un nom de personne. L'ensemble est rassemblé dans le graphe de l'exemple 30 (dans un souci de clarté, les boîtes de balisage ont été retirées de l'exemple).

Chaque boîte appelle un sous-graphe. Le sous-graphe *firstname* reconnaît les mots du dictionnaire des prénoms étiquetés <N+Firstname> lors du pré-traitement. Le sous-graphe *lastname* reconnaît les mots en majuscules. Le sous-graphe *persTitle* contient une liste de titres de civilités, grades, fonctions professionnelles construite par nos soins. Le sous-graphe *persNameCouple* reconnaît un prénom suivi de « et » puis d'un autre prénom. Cette forme a été privilégiée pour ce sous-graphe plutôt qu'une boucle sur le sous-graphe prénom, de cette manière le sous-graphe reconnaît *au maximum* deux prénoms. Le sous-graphe *marriedName* reconnaît une suite composée de « épouse » puis un mot en majuscule.

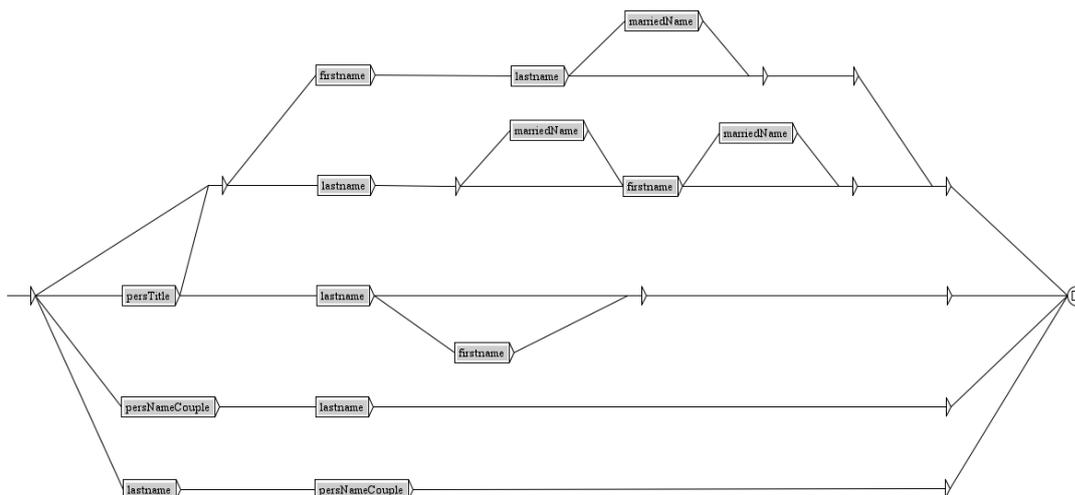


FIGURE 30 – Graphe général décrivant les noms de personnes purgé des boîtes de balisage dans un souci de lisibilité

Tag	Précision	Rappel	F-mesure	Hypothèse	Référence	Corrects
Weekday	100.0%	100.0%	100.0%	2	2	2
Day	96.7%	100.0%	98.3%	243	235	235
Month	96.7%	99.1%	97.9%	241	235	233
Year	96.8%	99.5%	98.1%	219	213	212
Date1	96.9%	100.0%	98.4%	159	154	154
Date2	100.0%	100.0%	100.0%	6	6	6
DateNum	96.1%	100.0%	98.0%	76	73	73
DateRel	100.0%	100.0%	100.0%	2	2	2

TABLE 3 – Performances pour les dates

Comme pour les dates, chaque chemin du graphe reconnaît un format de nom de personne en particulier. Cette distinction est conservée dans les sorties du graphe afin de permettre l'interprétation des résultats d'étiquetage.

### 2.3 Résultats et discussion

L'évaluation des résultats d'annotation produits par les graphes de reconnaissance a été opérée par comparaison avec un gold-standard composé de 10% du [corpus](#) (59 597 mots) annoté manuellement selon un guide d'annotation fourni en annexe [A](#). Nous avons choisi comme métrique d'évaluation la [précision](#), le [rappel](#) et la [F-mesure](#)<sup>10</sup>.

La colonne « Hypothèse » du [tableau 3](#) présente le décompte des entités dans le fichier annoté par les grammaires, et la colonne « Référence » le décompte des

10. Le détail des métriques a été exposé au [chapitre III](#)

Tag	Précision	Rappel	F-mesure	Hypothèse	Référence	Corrects
persName	83.0%	81.6%	82.3%	501	510	416
Firstname	97.7%	85.4%	91.1%	388	444	379
Lastname	95.7%	91.7%	93.7%	510	532	488
persTitle	58.8%	90.6%	71.3%	131	85	77
marriedName	100.0%	68.4%	81.2%	13	19	13
Couples	77.8%	31.8%	45.2%	9	22	7

TABLE 4 – Performances pour les noms de personnes

Tag	Précision	Rappel	F-mesure	Hypothèse	Référence	Corrects
locTown	88.4%	83.1%	85.7%	597	635	528

TABLE 5 – Performances pour les lieux

entités dans le fichier annoté manuellement. Au total, notre système a une **précision** de 91,2%, un **rappel** de 89,2%, et une **F-mesure** de 90,2%. Des trois entités détectées, les dates présentent les meilleures performances.

« Date1 » désigne les dates au format « 21 février 2015 », « Date2 » désigne les dates en toutes lettres, « DateNum » désigne les dates au format numérique entrecoupé de barres obliques, et « DateRel » désigne les dates relatives comme « Mercredi 22 ».

Dans le tableau 4 « persTitle » désigne les titres de civilité, « marriedName » les noms avec un nom d'épouse, et « Couples » les noms de couples (« Pierre et Marie Curie »).

Les performances de détection des noms de personnes sont en-dessous de celles des dates. Nous attribuons cela notamment à la variété de ce que nous avons inclus dans cette entité (noms de personnes, possibilité de nom d'épouse, et noms de couples) et aux fautes d'orthographe et coquilles présentes dans le texte.

Les performances de détection des noms de ville sont également relativement basses, compte-tenu du fait que notre approche consistait simplement à projeter un lexique sur le **corpus**. Nous avons donc décidé de les analyser plus spécifiquement : nous avons décompté 169 erreurs d'étiquetage et les avons catégorisées dans le tableau 6. Les erreurs les plus fréquentes sont provoquées par des ambiguïtés. Par exemple, « Orange » est étiqueté comme une ville alors qu'il s'agit de l'opérateur téléphonique, ou encore « Laguiole » alors que l'on parle d'un couteau. Certaines entités sont étiquetées comme des villes alors qu'il s'agit d'un nom de famille, ou

Catégorie d'erreur	Décompte	Proportion
Ambiguïté	56	33.1%
Orthographe	43	25,4%
Ressource	27	16%
Abréviation	17	10%
Étranger	9	5,3%
Nom de voie	9	5,3%
Autre	7	4,1%
Article	1	0,6%

TABLE 6 – Erreurs d'étiquetage des lieux

Tag	Précision	Rappel	F-mesure	Hypothèse	Référence	Corrects
locTown	76.8%	88,8%	82.4%	734	635	564

TABLE 7 – Performances pour les villes après modification de la ressource lexicale.

encore de simples noms communs : c'est le cas de « Rue » (commune de la Somme) ou de « Chemin » (commune du Jura). 16% des erreurs sont causées par un défaut de la ressource lexicale qui ne prend pas en compte l'absence de tirets dans les noms de villes composés. L'emploi abrégé de nom de villes composés cause 10% d'erreurs : par exemple, « Enghien » au lieu de « Enghien-les-Bains ». 5,3% des erreurs concernent des villes étrangères mentionnées dans le texte. 5,3% des erreurs concernent des noms de villes apparaissant dans des noms de voies : par exemple, « rue de Paris ». Enfin, dans un cas, une ville n'a pas été détectée car elle comporte un article : « Le Mans » apparaissant comme « au Mans » dans le texte.

À l'aide de ces observations, la ressource lexicale des noms de ville a été modifiée pour améliorer les erreurs catégorisées Ressource et Abréviation. Pour les villes avec un nom composé, nous avons extrait le premier segment du nom pour en faire une entrée à part entière, et nous avons ajouté ces mêmes noms composés sans tirets. Les résultats d'évaluation sont disponibles dans le tableau 7.

L'amélioration du [rappel](#) n'est pas suffisante pour contrebalancer la chute importante de la [précision](#) ce qui entraîne une baisse de la [F-mesure](#). Nous avons de nouveau analysé et catégorisé les erreurs (au nombre de 230). Les erreurs causées par les ambiguïtés représentent désormais 67,80% des erreurs totales (156 erreurs). L'extraction du premier segment des entités composées est à l'origine de cette hausse importante car il s'avère que de nombreux lieux de localités sont également des termes du langage courant. Par exemple la ville de Cette-Eygun (Pyrénées-Atlantiques),

dont l'entrée à été dédoublée en « Cette » et « Cette-Eygun », engendre des confusions lorsque le graphe rencontre le démonstratif « cette » en début de phrase. Il en va de même pour les villes d'Avant-Lès-Marcilly (Aube), Chef-du-Pont (Manche), Entre-Deux-Eaux (Vosges).

La détection des noms de villes reste donc à améliorer compte-tenu de ces résultats. L'une des pistes que nous envisageons est d'étudier les segments récurrents dans les noms de ville composés et de proposer une modélisation de leurs structures (« les-bains », « sur-mer », ainsi que les structures « nom-en-nom » et « nom-sur-cours d'eau »).

### 3 Synthèse

Nous avons cerné les manifestations linguistiques des **entités criminelles textuelles** dans le texte des **auditions de témoin** en vue du développement et du test d'une approche de détection automatique, ce qui nous a permis de mieux comprendre les objectifs propres à un tel système. L'approche développée est loin d'être complète et exploitable en l'état mais ouvre néanmoins la réflexion sur le sujet de l'extraction automatique d'information pour l'**analyse criminelle**. Nous livrons ici une série d'observations critiques sur l'expérience menée.

Vu la situation, avec peu de données et à l'aide d'une approche minimale, on obtient des résultats encourageants (l'approche fonctionne pour certaines formes d'entités et la mesure des résultats est plutôt bonne) qui, employés à bon escient, peuvent déjà être utiles aux **analystes criminels**, et sur lesquels on pourra compter pour la poursuite des recherches (par exemple dans le cadre d'une méthode hybride). Les graphes de détection des personnes et des dates sont fournis en annexe B. On signale toutefois qu'en termes de développement le système est au stade du prototype et ne peut être déployé tel quel.

Nous avons identifié des formes linguistiques prises par les entités qui ne correspondent pas aux objets traités habituellement en **extraction d'information**. Dans la partie de l'état de l'art consacrée à l'extraction d'entités criminelles dans du texte, nous avons souligné qu'aucune des références rencontrées ne décrivait les entités

prises en compte, ce qui à la lumière de ce que nous avons constaté dans le **corpus** apparaît comme un manque au niveau conceptuel qui ne permet pas d'apprécier la couverture des systèmes présentés.

Nous avons déterminé que l'incertitude et l'approximation dans les propos des **auditions** sont des facteurs qui influencent la forme des entités. Face à ce phénomène, on peut chercher à adapter les approches informatiques pour les extraire, mais le problème peut aussi être pris dans l'autre sens : l'**audition** de **témoin** étant une pratique linguistique se déroulant dans un cadre contraint, on pourrait envisager de travailler à la source du problème, c'est-à-dire lors de l'**audition**. Il s'agit là de la manifestation linguistique d'une préoccupation policière : comment amener un **témoin** à communiquer de manière la moins ambiguë possible sans pour autant influencer le contenu de ses propos ?

Ces formes inhabituelles, que nous avons baptisées *descriptions indéfinies*, par opposition au concept structuré de *description définie*, posent notamment un problème de délimitation textuelle. Les descriptions indéfinies sont l'expression de l'inconnu, or clarifier l'inconnu est le principe de l'**enquête**. Négliger leur prise en charge au même titre que les **entités nommées** de formes plus « classiques » revient à laisser de côté un point clé du traitement des **procédures**. Mais pour envisager d'adapter les approches de détection automatique à de tels exemples, il faut pouvoir compter sur un plus grand volume de données.

Cette remarque nous conduit à poser un regard critique sur les modalités des travaux réalisés, en particulier concernant le **corpus** : nous n'avons pu utiliser pour le développement qu'un seul **corpus** d'**auditions** de **témoins** tiré d'une même **procédure**, un **corpus** de taille réduite comparé aux pratiques de la recherche en **extraction d'information**. Ce **corpus**, relatif à un type d'affaire en particulier, a été produit dans une zone géographique particulière et n'est pas très étendu dans le temps. Les observations que nous avons réalisées à son sujet peuvent en conséquence manquer de perspective.

Ici se trouve la difficulté de travailler sur les dossiers de **procédure**. Les documents et les informations de la **procédure** sont couverts par le secret de l'**instruction** et seules les personnes concernées par l'**enquête** peuvent y avoir accès. Dans le cadre de nos travaux, un accord de confidentialité a été mis en place, mais cette

solution est un pis-aller qui ne peut suffire à assurer la poursuite des recherches sereinement.

Comme nous l'avons déjà évoqué rapidement dans ce chapitre, le domaine de l'extraction d'information offre d'autres possibilités au-delà de la détection des entités : la normalisation des entités peut ensuite conduire à la constitution de bases de connaissance, l'établissement de liens entre les entités au sein d'un même document et entre documents (par exemple : relier des informations temporelles entre [auditions](#), documents téléphoniques et documents bancaires). Dans cet objectif, la détection des entités et plus globalement la compréhension automatique du texte de la [procédure](#) est une première étape qui une fois résolue pourra ouvrir la voie d'autres applications d'aide à la gestion de l'information de la [procédure](#).

Cela dit, quelle que soit l'approche retenue pour la détection des entités (symbolique ou statistique), il est important de se rappeler que le principe est d'apprendre à un système à reconnaître des choses que l'on sait intéressantes d'avance. On *désigne* au système ce qu'il doit rapporter en le programmant soit par la rédaction de règles soit par la création d'exemples annotés, mais le système n'est pas capable d'innover et de dépasser les consignes qu'on lui a données. Ainsi, des éléments inattendus mais intéressants peuvent échapper à la détection. Dans cette configuration, il est utile de s'interroger sur l'introduction d'un biais dans la manipulation des documents de la [procédure](#). Le système de détection d'entités ou plus généralement le système de gestion des informations de la [procédure](#) ne doit pas devenir la lentille unique par laquelle l'[analyste criminel](#) perçoit le dossier. Il est indispensable de conserver une capacité de recul et un esprit critique sur un tel outil. C'est pourquoi nous préconisons l'intégration du système de détection des entités dans une infrastructure du type de celles employées en linguistique de [corpus](#) qui permettent le retour au texte à tout moment. Le repérage d'éléments en dehors du modèle est toujours possible et peut être assisté par un moteur de recherche textuelle. De cette façon, les analystes restent maîtres du travail d'[analyse criminelle](#) qui repose toujours sur leur expérience judiciaire et leur connaissance du terrain.

## 4 Conclusion

Dans ce chapitre consacré à la forme des entités criminelles dans les **auditions de témoin**, nous avons entamé un travail de recherche visant à leur détection automatique axé principalement sur l'aspect conceptuel. Les particularités de la matière textuelle considérée et les précautions nécessaires à l'introduction d'un système de gestion de l'information ont été mises au jour tout en démontrant que la situation se prête à l'application de méthodes déjà éprouvées, en l'occurrence les méthodes symboliques.

Néanmoins, l'approche reste minimale et ne couvre pas les singularités propres à l'expression de l'information dans les **auditions**, qui constituent le véritable levier d'innovation du cas étudié. Pour remédier à cela, il faudra encourager le rapprochement de l'analyse criminelle des acteurs de la recherche spécialisée en **extraction d'information** et **TAL** afin de construire des projets de recherche plus ambitieux et mieux équipés qui permettraient de traiter la problématique dans sa globalité, la question prioritaire demeurant la disponibilité des **corpus de procédure**, sans lesquels la recherche ne peut se faire. Dans cette perspective, il pourra être utile de s'inspirer de la recherche en traitement automatique de la langue médicale : ce champ de recherche est amené à manipuler des données de santé confidentielles comme des dossiers de patients dans le cadre de coopérations entre hôpitaux et laboratoires de recherche informatique.



## Chapitre V

# Le texte de la procédure judiciaire : l'exemple des auditions

Le chapitre [IV](#) a permis d'établir que les [auditions](#) de [témoins](#) présentent des formes d'entités spécifiques, influencées par leur contexte de production : l'[enquête](#), qui par définition cherche à clarifier l'inconnu, amène les [témoins](#) à s'exprimer de manière incertaine et imprécise. Nous chercherons dans ce chapitre à approfondir la description linguistique et textuelle du sous-corpus sur lequel nous avons focalisé nos recherches, celui des [auditions](#) de [témoins](#). Nous espérons que cela permettra, en apportant une compréhension approfondie des aspects linguistiques de la pratique de l'[audition](#), de poser le cadre du traitement automatique plus large de la [procédure](#).

En effet, les attentes sur l'apport des technologies numériques pour la société ne cessent de s'accroître et touchent de plus en plus de domaines différents. On appelle *LegalTech* le champ qui se consacre à produire des services numériques et technologiques pour la justice : par exemple, on espère que le passage au numérique accélérera le temps de traitement des affaires, ou encore que « l'intelligence artificielle » lissera les décisions de justice entre juridictions, pour une justice plus équitable. Sans présumer de la réalité factuelle et technique de telles attentes ou encore de leur impact réel, il est capital de penser ces usages et les modalités concrètes de leur développement. Comme bon nombre de ces technologies impliquent le traitement de données langagières, il est naturel que le [TAL](#) s'intéresse à ces problématiques et établisse un nouveau sous-domaine, le *LegalNLP*<sup>1</sup>. Dans cette perspective,

---

1. NLP : *natural language processing*, traitement automatique du langage naturel.

il est nécessaire d'étudier et de définir les données qui pourront être traitées.

Ce chapitre s'articule en deux parties : tout d'abord, nous étudions le texte des **auditions**, et nous procédons ensuite à une étude contrastive visant à organiser les textes de la justice en discours, genres et sous-genres, pour laquelle nous prendrons en compte d'autres types de textes impliqués de près ou de loin dans des processus de justice et de droit.

## 1 L'audition d'enquête : pratique et productions

Comme nous l'avons déjà évoqué, l'**enquête** et plus largement le processus judiciaire en France se fonde sur l'écrit. Cela signifie que dans le cas des **auditions** de **témoins**, l'**enquête** exploite les **procès-verbaux**, c'est-à-dire les documents écrits qui en sont tirés. Nous proposons dans cette partie d'étudier notre **corpus** d'**auditions** pour comprendre le contexte de sa production et la construction interne des **procès-verbaux** d'**audition**. Nous compléterons notre compréhension de la pratique à l'aide de l'expérience et de ce qui nous a été décrit par les gendarmes du **PJGN**.

Nous avons déjà abordé la forme du **procès-verbal** d'**audition** à la section 2.2.2 du chapitre **II**. Les remarques que nous formulons ne sont pas des règles absolues, et celles-ci sont valables pour les **auditions** pratiquées par la **Gendarmerie nationale** uniquement. Les pratiques de la Police peuvent être différentes.

### 1.1 Entendre, interroger, rapporter

La fonction de l'**audition** au cours d'une **enquête** est de recueillir les propos des protagonistes d'une affaire. Les échanges sont consignés par écrit dans un **procès-verbal**. Plusieurs types de personnes peuvent être entendues : victimes, **témoins**, mis en cause. Selon leur rôle dans les faits, le but de l'**audition** est différent. Les victimes et les **témoins** seront incités à se remémorer le plus de détails possibles, tandis qu'on tentera d'obtenir des aveux pour une personne mise en cause.

Dans une **audition**, qu'elle soit libre ou sous contrainte, les enquêteurs doivent assumer au moins trois rôles : écouter, diriger l'entretien, et prendre les propos en

notes. Ces trois rôles visent à assurer le recueil des informations et leur transmission dans la [procédure](#). Les enquêteurs doivent amener la personne entendue à délivrer le plus d'éléments possibles sans pour autant les influencer.

Comme le souligne VLAMYNCK (2011) dans son article sur les conditions de déroulement d'une [audition](#) ou d'un interrogatoire, paradoxalement la conduite d'[audition d'enquête](#) ne fait l'objet de formation ni à la [Gendarmerie nationale](#) ni dans la police, hormis pour les enquêteurs spécialisés dans l'[audition](#) de mineurs victimes. Il n'y a donc pas de formalisation officielle de la pratique mais des usages qui se transmettent par observation et formation par les pairs.

Les [auditions](#) sont généralement conduites par deux voire trois personnes. L'un d'entre eux mène la conversation en posant les questions et orientant la conversation, pendant que l'autre tape les échanges simultanément. Comme l'ont souligné BLANCHET et al. (2013), la difficulté de la conduite de l'[audition](#) réside dans le fait que l'un des personnels qui réalise l'[audition](#) se consacre à la retranscrire sur l'ordinateur. Cette tâche monopolisant alors son attention, il ne peut être une aide à part entière pour son collègue qui se charge de mener la conversation. Les auteurs notent que cette configuration incite celui qui mène l'[audition](#) à ralentir, à faire répéter ou bien à reformuler les propos.

Le [procès-verbal d'audition](#) est ce qui subsiste de l'échange qui s'est déroulé entre la personne auditionnée et les enquêteurs. Chaque page est signée par les deux parties, qui reconnaissent l'exactitude des propos rapportés et consentent donc à leur exploitation tels quels dans l'[enquête](#). Bien que dans certains cas, la loi exige l'enregistrement audiovisuel de l'[audition](#) (lorsque la personne entendue est mineure ou lorsqu'il s'agit d'une [audition de garde à vue](#)), le [procès-verbal](#) papier reste le document exploité en priorité pendant l'[enquête](#) et le procès. L'enregistrement peut être utilisé en cas de litige, si l'on veut avoir une meilleure vision sur le contenu, ou pour voir l'attitude de la personne auditionnée.

Plusieurs stratégies peuvent être adoptées pour mener l'[audition](#) : la conversation peut être spontanée, par exemple dans le cas d'un [témoin](#) se présentant de lui-même, ou bien les enquêteurs peuvent préparer une trame d'[audition](#) avec une série de questions prévues à l'avance (par exemple, dans le cas des [gardes à vue](#), où l'on sait précisément sur quels éléments on veut obtenir des éclaircissements). Quel que

soit le cas, il est important que les enquêteurs restent attentifs à l'attitude de la personne entendue pour orienter la poursuite de l'entretien.

Du point de vue de la production textuelle, nous avons constaté (et les exemples présentés dans le manuscrit l'illustrent) de nombreuses fautes d'orthographe, coquilles, insertions ou omissions de mots, y compris dans des entités (nous avons par exemple relevé des erreurs dans l'orthographe de prénoms ou de noms de famille ou des insertions de chiffres supplémentaires dans des numéros de téléphone).

Le **procès-verbal d'audition** s'organise en trois parties, chacune ayant des fonctions différentes : l'en-tête, les éléments d'état-civil, et le corps de l'**audition** à proprement parler (voir figure 2.2.2). Les deux premières parties servent à cadrer l'**audition** dans l'**enquête** et à collecter les informations de base sur la personne auditionnée.

C'est le corps de l'**audition** qui représente le plus d'intérêt pour une analyse linguistique, car il s'agit de la partie véritablement libre du document, aussi bien en termes de forme qu'en termes de contenu. Étant donné que nous avons abordé la question du contenu dans le chapitre IV, nous souhaitons maintenant aborder la description linguistique des **auditions** que nous avons explorées au cours de nos travaux.

## 1.2 Le discours de l'audition et le discours du procès-verbal

Bien que les **procès-verbaux d'audition** soient la retranscription de propos tenus à l'oral, nous n'avons rencontré dans le **corpus** que très peu de traces des répétitions et des reformulations<sup>2</sup> remarquées par BLANCHET et al. (2013), et aucune trace d'éléments habituellement caractéristiques du discours oral comme les pauses ou les hésitations.

En linguistique, l'étude de l'oral est un champ distinct de l'étude de l'écrit et qui a été amené à se pencher sur la question de la transcription. En effet, comme le soulignent DISTER et SIMON (2007) reprenant BLANCHE-BENVENISTE (2000), l'étude de l'oral doit passer par l'écrit, dans un paradoxe causé par la nature fugace des données orales même lorsqu'elles sont enregistrées. Cela a donné lieu à la création de conventions et de systèmes de transcriptions, afin de rendre les phénomènes propres

---

2. Pour les reformulations, nous avons recherché les marqueurs étudiés par ESHKOL-TARAVELLA et GRABAR (2017). Nous avons rencontré deux occurrences de *je veux dire*, trente-et-une de *c'est-à-dire*, aucune de *disons, en somme, en bref, autrement dit*.

à l'expression orale, la prosodie : intonation, élisions, disfluences verbales (pauses, répétitions, amorces). Les travaux de transcription doivent s'appuyer sur une théorie car les choix opérés par le transcripateur influent sur l'analyse qui sera faite des données (ROUBAUD, 2017).

Les objectifs de l'étude linguistique de la langue orale et de l'audition d'un témoin dans une enquête ne sont bien entendu pas les mêmes. Mais du point de vue de la description linguistique du texte des auditions, la mise en parallèle de ces pratiques nous amène à nous interroger sur la nature du texte des auditions. Nous supposons qu'il ne s'agit pas d'une transcription exacte de l'oral, puisque les éléments propres à ce mode d'expression ne sont pas rapportés, mais le texte produit trouve sa source dans une performance orale. Alors, comment appréhender les données langagières des procès-verbaux d'audition ? Des approches et théories de la linguistique de l'oral et de la linguistique de l'écrit, lesquelles faut-il privilégier ?

Pour répondre à cette interrogation, nous reprenons ce que RASTIER (2001, p. 83) appelle les *transpositions de l'expression* :

[...] Nous définirons la transposition comme le passage entre deux sémiotiques, deux langues, voire deux discours. [...] Parmi les transpositions de l'expression, on trouve par exemple la vocalisation d'un texte écrit, l'écriture d'un texte oral, ou, simplement, la copie d'un écrit. La transposition peut « perdre » des informations, mais en « rajouter » d'autres.

L'enquêteur qui prend la déposition par écrit transpose de l'oral à l'écrit et *réadapte* les propos, en les conformant au passage au modèle de l'audition que les membres de l'institution ont établi à force de pratique.

KOMTER (2006) va dans le sens de notre hypothèse : l'auteur étudie la construction de compte-rendus d'auditions de suspects au Pays-Bas en retranscrivant pauses, hésitations, répétitions, etc., puis cite en comparaison le texte qui apparaît finalement dans le document. Au lieu du dialogue tel qu'il a eu lieu, les propos de la personne entendue sont rapportés comme un monologue fluide et d'apparence spontanée. Il ne s'agit pas des propres mots de la personne entendue (« such a text [...] is not an exact representation of "the suspect's own words" ») mais un reflet de ce qui a été dit (« it is at least a reflection of what has been said in the interrogating room ») et

cela est suffisant du point de vue légal. Ce type d'altérations est considéré comme courant dans les pratiques de transposition de l'oral vers l'écrit dans les **auditions** policières (D'HONDT, 2019).

En France, l'article 429 du **Code de procédure pénale**<sup>3</sup> évoque l'auteur du **procès-verbal** :

Tout procès-verbal ou rapport n'a de valeur probante que s'il est régulier en la forme, si son auteur a agi dans l'exercice de ses fonctions et a rapporté sur une matière de sa compétence ce qu'il a vu, entendu ou constaté personnellement.

Dans le cas des **procès-verbaux d'audition** cette notion d'auteur mérite d'être précisée pour permettre l'interprétation et l'étude du texte produit.

Du point de vue de la loi, il est clair que l'enquêteur est l'auteur du **procès-verbal** puisque celui-ci exerce « dans le cadre de ses fonctions », mais c'est la personne entendue qui est principalement tenue d'assumer les propos consignés et leurs conséquences légales. Ce paradoxe est couvert par la signature du **témoin** apposée sur le **procès-verbal**.

Du point de vue linguistique, si l'on admet que le texte du **procès-verbal** n'est pas une transcription mot à mot de la conversation mais une transposition, et de ce fait que l'auteur du texte est celui qui l'a rédigé et non celui qui l'a prononcé, on peut en déduire qu'il s'agit également de l'enquêteur.

Dans les deux perspectives, le discours original du **témoin** n'est pas accessible au lecteur du **procès-verbal**, et donc au linguiste qui tenterait de l'analyser. Nous n'entrerons pas à ce sujet dans des considérations légales hors de nos compétences. En revanche, cela nous permet de distinguer deux situations d'énonciation différentes dans la pratique de l'**audition** de **témoin** correspondant aux deux stades de la transposition : il y a d'une part le discours oral du **témoin** tenu au cours de l'**audition**, dans lequel il est maître de ses mots, et d'autre part le texte du **procès-verbal** qui correspond à la retranscription des propos, dans lequel c'est l'enquêteur qui dispose de l'expression et qui constitue le discours qui fera foi. Ainsi, le discours du **témoin**

---

3. Légifrance : article 429 du **Code de procédure pénale** <https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006071154&idArticle=LEGIARTI000006576551> [consulté le 3 décembre 2019].

auditionné est doublement repris entre son énonciation et la rédaction du **procès-verbal** : en passant de l'oral au texte, le discours passe d'un locuteur initial à un auteur final, et chacun de ces transferts implique des adaptations.

Le statut de transposition du texte du **procès-verbal** établi, il émerge deux positions pour le linguiste afin d'analyser les **auditions** de **témoin** : soit au niveau de la conduite de l'**audition**, cas dans lequel il est possible de réaliser des observations directes sur le discours du **témoin**, soit au niveau du **procès-verbal**, notre cas, dans lequel les observations concerneront le résultat de la transposition, c'est-à-dire le texte de l'**audition**.

Cette posture nous paraît convenir d'autant mieux que le **procès-verbal** et son texte faisant foi, les investigations, décisions, etc., **analyse criminelle** comprise, exploitent la forme et le contenu véhiculés par le texte. Ainsi, c'est l'objet textuel **procès-verbal** d'**audition** en tant que tel, avec la situation d'énonciation qui est la sienne, qui nous intéresse et que nous souhaitons définir. En nous fondant là-dessus, nous adopterons pour l'étude linguistique des **auditions** de **témoin** une approche textuelle.

### 1.3 Le procès-verbal d'audition : régularité et narrativité

Dans cette partie, nous proposons un aperçu des particularités du texte des **procès-verbaux** d'**audition** de **témoin**.

D'après ce que nous avons observé, le texte de l'**audition** est consigné selon deux grandes tendances : soit comme un dialogue, où questions et réponses sont mises sur le papier, soit comme une déclaration d'un seul tenant du **témoin**. Encore une fois, il n'y a pas de règle établie mais des pratiques et des usages adoptés en fonction des situations. Par exemple VLAMYNCK (2011) indique que l'article 429 du **Code de procédure pénale**<sup>4</sup> oblige l'auteur du **procès-verbal** à écrire les questions, mais que leur oubli ne cause pas la nullité du document. Ceci illustre le décalage entre le cadre légal prescriptif et la réalité des pratiques.

Quelle que soit la structure, les propos du **témoin** sont invariablement rapportés à la première personne du singulier, et ceux des enquêteurs à la première personne

4. Légifrance : article 429 du **Code de procédure pénale** <https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006071154&idArticle=LEGIARTI000006576551> [consulté le 3 décembre 2019].

**Question** : Vous souvenez vous de votre emploi du temps pour la journée du ■■■ ?  
**Réponse** : Je fais tellement de chose avec les enfants. Je sais que j'ai passé la journée avec les enfants. Le matin comme tous les matins, je fais mon ménage, je prépare le repas de midi, je m'occupe des enfants. L'après midi, il se peut que l'on soit sorti avec les enfants, soit on est resté à la maison à regarder la télévision ou jouer à la playstation.  
**Question** : Nous vous informons qu'il faisait beau le ■■■.  
**Réponse** : Dans ce cas, s'il faisait beau on a du sortir ensemble. Je sais que ■■■ rentre

FIGURE 31 – Exemple de l'emploi des pronoms personnels

**Mention des enquêteurs** : ■■■ nous montre du doigt un chien photo n°7 et hésite avec la photo n°18 du panel cité.

FIGURE 32 – Exemple de mention des enquêteurs

du pluriel (figure 31). Ainsi, bien que l'audition soit une conversation entre une personne entendue et deux ou trois enquêteurs, les propos de ces derniers, lorsqu'ils sont consignés, ne leur sont pas attribués individuellement mais rassemblés sous la première personne du pluriel.

L'emploi des pronoms personnels, en construisant une illusion de discours direct, est une trace du caractère initialement oral de la matière de l'audition, et marque la dimension narrative des propos du témoin amené à faire le récit de son emploi du temps, de ses habitudes, de ses relations, etc.

On rencontre aussi parfois des mentions spéciales des enquêteurs comme aux figures 32 et 33. Leur but est d'apporter des précisions sur le déroulement de l'audition et des informations qui ne sont pas véhiculées par le langage mais qui pourrait néanmoins être pertinentes. Introduites par « Mention : », « Mentionnons » ou « Mention des enquêteurs : », leur usage n'est pas très fréquent : nous en avons relevé 39 dans notre corpus de 370 auditions. Les enquêteurs y emploient toujours la première personne du pluriel, mais le témoin est désigné à la troisième personne du singulier. On notera avec intérêt que dans l'exemple 33 où l'audition se poursuit avec une gendarme seule pour le confort du témoin, celle-ci est aussi désignée à la troisième personne du singulier dans la mention, mais c'est toujours la première personne du pluriel qui est employée dans le reste de l'audition.

Mention des enquêteurs : nous interrompons l'audition quelques instants. Le témoin se met à pleurer, elle est incapable de parler. Elle est visiblement très émue d'avoir à se souvenir de se qui s'est passé.  
 L'audition est continuée par le gendarme ■■■.

FIGURE 33 – Exemple de mention des enquêteurs. Ici l'audition est prise en charge par une gendarme.

Nous faisons comparaître devant nous, le témoin ci-après nommé et lui donnons connaissance des faits pour lesquels sa déposition est requise. ----

FIGURE 34 – Emploi de la première personne du pluriel dans une phrase réglementaire qui ouvre l'audition

A [REDACTED], le [REDACTED] à 16 heures 05, lecture faite par moi de la déposition ci-dessus, j'y persiste et n'ai rien à changer, à y ajouter ou à y retrancher.

FIGURE 35 – Emploi de la première personne du singulier dans une phrase réglementaire qui termine l'audition

L'usage des pronoms personnels a également cours dans des passages de « texte réglementaire », des passages de texte qui assurent la conformité légale du document. Les deux extraits des figures 34 et 35, sous ces formes précisément ou sous des variantes, entament et concluent les *auditions*, introduisant un flou sur la délimitation réelle de la conversation : n'étant pas énoncées à haute voix lors de l'*audition*, elles en précisent la situation dans le *procès-verbal*, au même titre que les mentions. Ainsi, les *procès-verbaux*, en plus d'être le fruit d'une double situation d'énonciation, consignent aussi deux récits imbriqués : le récit de l'*audition* réalisé par l'auteur du *procès-verbal*, et le récit du *témoin*.

Ces phrases réglementaires employées dans tous les *procès-verbaux* d'*auditions* semblent caractéristiques de leur régularité. Pour explorer cet aspect répétitif de l'expression dans les *auditions*, nous avons utilisé la fonction « segments répétés » du logiciel *Lexico5*, qui permet de repérer et de dénombrer les répétitions de suites de formes dans un *corpus* (LAFON et SALEM, 1983; LEBART et SALEM, 1994). Le logiciel a décompté près de 8 790 *segments répétés* dix fois ou plus allant de deux à onze *tokens*, un *token* étant une suite de caractères comprises entre des espaces ou des signes de ponctuation.

Dans le tableau 8, nous avons listé les segments de onze *tokens* apparaissant dix fois ou plus (les éléments identifiant des personnes, des lieux ou des dates ont été

---

Entendons la personne dénommée ci-dessus qui nous déclare :  
« « Je prends connaissance des raison qui motivent ma déposition ce jour devant vous. J'habite le même quartier que la personne qui est décédé [REDACTED]. -----

FIGURE 36 – Autre exemple d'ouverture d'une audition

anonymisés). Le logiciel en décompte 85. Par comparaison, le **corpus** de démonstration fourni avec le logiciel, le Père Duchêne, compte un seul segment de neuf **tokens** répétés dix fois pour environ 142 000 **tokens** au total, et dans le roman *Le tour du monde en 80 jours* de Jules Verne, le **segment répété** le plus long est de quatre **tokens** pour onze occurrences pour environ 72 000 **tokens** décomptés par le logiciel. Ces deux **corpus** n'ont bien entendu pas grand chose à voir avec des **auditions de témoin** dans une **enquête** criminelle mais ils permettent néanmoins d'avoir un ordre d'idée concernant la fréquence et la longueur des **segments répétés**.

Dans les **auditions de témoins**, la haute fréquence de **segments répétés** longs est le signe d'une grande régularité dans le texte : par exemple, le segment le plus fréquent apparaît 315 fois dans le **corpus**, c'est-à-dire dans 85% des **procès-verbaux**. L'examen du tableau montre que ces **segments répétés** sont bien des phrases « réglementaires », qui énoncent les circonstances de l'**audition**, les articles du code **procédure** pénale relatifs à la conduite d'**audition**, les éléments d'état-civil, etc. Seize d'entre elles comportent des formes de la première personne du pluriel. Sept comportent des formes de la première personne du singulier, et correspondent à peu près toutes à la première phrase utilisant cette personne dans les **auditions**, phrase par laquelle le **témoin** affirme connaître les raisons de son **audition**. En dehors de ces régularités, on repère aussi un ensemble de question « thématiques » à l'**enquête**, où l'on demande à la personne auditionnée si elle a rencontré des personnes en particulier, dont l'accoutrement ou l'attitude l'aurait interpellée. La fréquence de ces segments s'explique par le fait que dans cette affaire spécifiquement, tout un groupe de **témoins** a été interrogé selon une trame d'**audition** préparée à l'avance. Les questions sont donc rigoureusement les mêmes.

Par le biais des **segments répétés**, nous avons obtenu un aperçu du contenu structurant transversalement les **auditions**, aucun des segments ne présentant d'éléments particulier de réponse aux questions. On remarque que la régularité de l'expression touche les aspects réglementaires du texte mais aussi des passages utilisant les pronoms personnels. Il semble improbable que toutes les personnes auditionnées aient prononcé rigoureusement les mêmes phrases et que les enquêteurs aient posé exactement les mêmes questions. Ce constat renforce notre présomption selon laquelle le texte des **procès-verbaux** n'est pas une retranscription mot à mot,

Segment	Fréquence
et lui donnons connaissance des faits pour lesquels sa déposition est	315
après nommé et lui donnons connaissance des faits pour lesquels sa	309
Vu les articles 16 à 19 et 151 à 155 du	268
entendu séparément et hors la présence de la personne mise en	241
à 19 et 151 à 155 du Code de procédure pénale	228
Sexe Situation de Famille Date Naissance Code Postal et Commune Naissance	120
que nous avons effectuées en exécution de la commission rogatoire jointe	79
instruction près le tribunal de grande instance de XXX Vu l	60
de la compagnie de XX en résidence à la brigade de	53
GENDARMERIE NATIONALE ENQUETE Compagnie de XXX SUR COMMISSION ROGATOIRE Brigade de	52
du Code de procédure pénale Nous trouvant au bureau de notre	49
agit de la personne découverte sans vie dans la forêt ce	46
de Police Judiciaire Vu les articles 16 à 19 et 151	44
Officier de Police Judiciaire en résidence à Section de Recherches de	43
à 19 et 53 à 67 du Code de procédure pénale	41
Rapportons les opérations que nous avons effectuées en exécution de la	40
Renseignements sur la délégation Date Numéro Nom et fonction du Magistrat	40
Je prends connaissance du motif pour lequel ma déposition est requise	39
entendu séparément et hors la présence des mis en examen dépose	39
Vu les articles 16 à 19 et 53 à 67 du	39
Région de Gendarmerie XXX Section de Recherches de XXX BT XXX	38
Je prends connaissance des faits qui motivent cette déposition à savoir	37
GENDARMERIE NATIONALE Région de Gendarmerie XXX Section de Recherches de XXX	37
instruction au TGI de XXX Information ouverte contre Mis en examen	37
de Police Judiciaire GENDARMERIE NATIONALE Compagnie de XXX BRIGADE DE RECHERCHES	33
du Code de procédure pénale Nous trouvant au bureau de la	32
Nous vous montrons une série de 07 portraits robot numérotés de	32
des choses ou des gens qui ne collent pas avec le	31
Nous vous montrons une série de 06 photographies de couteaux référencée	31
Avez vous des connaissances qui vont dans cette forêt de XXX	30
Nous vous montrons une série de photographies représentant des chiens et	28
lequel en précisant avec quel type de personne il était et	28
XXX Vu les articles 16 à 19 et 151 à 155	27

Officiers de Police Judiciaire Vu les articles 16 à 19 et	27
des raisons qui motivent ma déposition ce jour devant vous à	26
Nous vous montrons une série de photographies représentant des modèles de	26
GENDARMERIE NATIONALE Compagnie de XXX BRIGADE DE RECHERCHES de XXX B	26
prends connaissance des raisons qui motivent ma déposition ce jour devant	26
Vu les articles 16 à 19 et 151 à 155	25
de XXX Vu les articles 16 à 19 et 151 à	24
Je prends connaissance des raisons qui motivent ma déposition ce jour	24
GENDARMERIE NATIONALE Compagnie de XXX BRIGADE DE RECHERCHES de XXX BT	24
rien à y changer à y ajouter ou à y retrancher	23
Nous trouvant à la brigade de XXX faisons comparaître devant nous	22
XXX Vu les articles 16 à 19 et 151 à 155	22
17 à 19 et 151 à 155 du Code de Procédure	22
Avez vous été dans cette forêt ce <jour> 00 XXX 0000	22
de XXX Vu les articles 16 à 19 et 151 à	21
Homicide volontaire sur la personne de XXX XXX cir ENQUETE Nous	21
XXX XXX XXX Vu les articles 16 à 19 et 151	21
des véhicules ou tout autres choses sur votre parcours de ce	20
ai rien à y changer à y ajouter ou à y	20
du Code de procédure pénale Nous trouvant à son domicile à	20
véhicules ou tout autres choses sur votre parcours de ce 00	20
de Police Judiciaire GENDARMERIE NATIONALE Région de Gendarmerie XXX Sec- tion de	19
0000 à XXX Mme XXX Vice présidente en charge à XXX	19
Le témoin Les Officiers de Police Judiciaire GENDARMERIE NATIONALE Compa- gnie de	19
V Année N°pièce Feuillet TÉMOIN Nmr dossier justice 03462 00907	19
Officier de Police Judiciaire en résidence à la brigade de XXX	19
Section de Recherches de XXX Vu les articles 16 à 19	19
0000 à XXX Madame XXX Vice présidente en charge à XXX	18
Officier de Police Judiciaire Vu les articles 16 à 19 et	18
vous montrons un exemplaire en couleur représentant une vue aérienne de	17
entre eux vous rappelle le véhicule vu en forêt de XXX	17
Officier de Police Judiciaire en résidence à BT XXX XXX XXX	17
Nous vous montrons un exemplaire en couleur représentant une vue aérienne	16

Les Officiers de Police Judiciaire GENDARMERIE NATIONALE Compagnie de XXX BRIGADE	16
Officier de Police Judiciaire GENDARMERIE NATIONALE Compagnie de XXX BRIGADE DE	16
Les Officiers de Police Judiciaire GENDARMERIE NATIONALE Région de Gendarmerie XXX	15
Je prends connaissance du motif pour lequel je suis entendu ce	14
La personne entendue Les Officiers de Police Judiciaire GENDARMERIE NATIONALE Compagnie	14
Allez vous toujours seule dans cette forêt ou êtes vous parfois	14
de Police Judiciaire en résidence à la brigade territoriale de XXX	13
La personne entendue Les Officiers de Police Judiciaire PV n°00907	13
Auriez vous couru la veille le <jour> 00 XXX 0000 dans	13
tracer au stylo rouge votre parcours du 00 XXX 0000 dans	13
Le témoin Les Officiers de Police Judiciaire GENDARMERIE NATIONALE Région de	12
Avez vous remarqué une personne dont la tenue vestimentaire ne vous	12
Est ce que vous rencontrez souvent les mêmes personnes durant votre	12
0000 à XXX XXX Vice présidente en charge à XXX de	11
Aviez vous votre téléphone portable avec vous ce <jour> matin 00	11
vous vu la photographie de la victime dans la presse et	10
Nous vous demandons de signer 04 exemplaire de cette vue aérienne	10
Avez vous déjà entendu parler ou vu un exhibitionniste dans cette	10
Pouvez vous sur quatre exemplaires identiques mais en noir et blanc	10

TABLE 8 – Segments répétés de onze mots et de fréquence supérieure à 10

et la régularité des formules conforte l'hypothèse d'un récit de l'audition en partie basé sur des formules toutes faites.

Au-delà de la régularité on repère aussi dans les segments répétés un aspect narratif, qui dans ce cas, comme nous l'avons expliqué, raconte la situation de l'audition (en somme, cette narration sert à attester que la conduite de l'audition s'est faite de manière réglementaire). Mais cet aspect narratif déborde des phrases toutes faites et domine dans les réponses des témoins, ce qui est logique car le but de l'audition

**QUESTION : Pouvez vous nous rappelez votre emploi du temps pour la journée du [REDACTED] ?**

**REPONSE :** Je me suis levée entre 09 et 10 heures. J'ai pris mon petit déjeuner. J'ai fait quelques tâches ménagères, je me suis occupé de [REDACTED] et j'ai déjeuné. J'ai quitté la maison vers 11h50, j'ai pris mon vélo et me suis rendu à mon travail. Auparavant, j'ai acheté un sandwich. Prenant mon travail à 12h30, je pense que je me trouvais sur mon lieu d'activité professionnelle dix minutes avant. --  
j'ai fini mon boulot à 20 heures. A mon retour vers 20h30, j'ai appris le décès de [REDACTED] par vos collègues qui se trouvaient à mon domicile. ---

FIGURE 37 – Extrait d'audition concernant un emploi du temps

**Question :** A quelle heure êtes vous allé courir aujourd'hui ? -----

**Réponse :** Je suis parti vers les 15 heures. J'ai pris ma voiture une peugeot 306 rouge foncé métallisé immatriculée [REDACTED], je me suis garé à proximité de [REDACTED]. Je suis sorti de la voiture et je me suis étiré. J'étais seul pour courir. J'ai emprunte le chemin le plus à gauche par rapport au parking. J'ai traversé la forêt, je me suis dirigé vers le pont. Normalement il me faut 6 à 7 minutes pour faire le trajet jusqu'au pont. Avant d'arriver au pont, une dame avec deux chien m'a demandé si j'avais un portable car elle avait découvert une femme couchée au sol dans un chemin. J'ai

FIGURE 38 – Autre extrait d'audition concernant un emploi du temps

est de recueillir un récit. Pour étayer cette hypothèse, nous avons consulté l'étiquetage morpho-syntaxique fourni par TreeTagger (SCHMID, 1994) dans le logiciel TXM (HEIDEN et al., 2010). Le tableau 9 présente les fréquences de chaque étiquette<sup>5</sup>.

On remarque que les pronoms personnels sont la troisième catégorie la plus fréquente après les noms et les prépositions et représentent environ de 9% des **parties du discours**. Pour conforter notre analyse, on peut remarquer aussi que parmi les temps des verbes, le présent, les participes passés et l'imparfait sont les trois plus fréquents. Or ces trois formes verbales sont caractéristiques du récit et de la narration. On les retrouve dans le tableau 8 des **segments répétés**, elles sont donc présentes dans les passages stabilisés de l'**audition**. En ce qui concerne les passages d'expression des **témoins**, les calculs statistiques ne sont pas d'une grande aide étant donné que ces passages ne sont pas aussi formalisés. Toutefois, nous avons déjà présenté dans le chapitre IV de nombreux exemples tirés du **corpus** qui présentaient des aspects narratifs, et pour les compléter nous proposons les exemples de la figure 37, où le **témoin** raconte sa journée, et de la figure 38, où le **témoin** raconte les circonstances précédant la découverte d'un corps. On y repère de nouveau l'emploi des pronoms personnels, ainsi que les temps repérés comme fréquents à l'aide des étiquettes morpho-syntaxiques.

5. Le détail des **parties du discours** est repris de la page de présentation du jeu d'étiquettes de TreeTagger pour le français : <https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html> [consulté le 5 décembre 2019]

Étiquette	Partie du discours correspondante	Fréquence
NOM	nom commun	97514
PRP	préposition	59555
PRO :PER	pronom personnel	55805
NAM	nom propre	44148
PUN	ponctuation	42132
VER :pres	verbe au présent	36584
SENT	ponctuation de fin de phrase	35032
DET :ART	article	34408
ADV	adverbe	30219
NUM	chiffre	24857
KON	conjonction	22636
VER :pper	participe passé	22292
ADJ	adjectif	15117
VER :impf	verbe à l'imparfait	12664
PRP :det	article contracté	11176
VER :infi	verbe à l'infinitif	10754
PRO :DEM	pronom démonstratif	8754
PRO :REL	pronom relatif	8485
DET :POS	déterminant possessif (ma, ta...)	7885
PRO :IND	pronom indéfini	2428
ABR	abréviation	2193
VER :ppre	participe présent	2061
VER :futu	verbe au futur	1980
PUN :cit	ponctuation de citation	1433
VER :cond	verbe au conditionnel	1185
INT	interjection	1180
VER :simp	verbe au passé simple	1039
VER :subi	verbe au subjonctif imparfait	520
VER :subp	verbe au subjonctif présent	476
PRO	pronom	53
PRO :POS	pronom possessif (mien, tien..)	47
SYM	symboles	19

TABLE 9 – Étiquettes morpho-syntaxiques et leurs fréquences

## 1.4 Conclusion

Nous avons souhaité dans cette section faire un point sur la nature du texte des **procès-verbaux** d'**audition**. Bien que nous n'ayons pas eu accès à des enregistrements, nous avons pu établir que le texte des **procès-verbaux** est une transposition des propos de la personne entendue et pas ses propres mots. La transposition est aperçue notamment à travers le caractère très régulier du texte que nous avons pu mettre en lumière en examinant les **segments répétés**. Une autre caractéristique des **auditions** est leur aspect narratif, que nous avons souligné par l'étude des **parties du discours**, en particulier les pronoms personnels et les temps des verbes. Narration et régularité s'inscrivent alors dans deux énonciations différentes : celle de l'**audition**, exprimée par le « nous » des enquêteurs, et celle des propos du **témoin** qui s'exprime par le « je ».

## 2 Étude textométrique pour une catégorisation en genres

La partie précédente a permis de caractériser le contenu du **procès-verbal** d'**audition** en tant que texte transposé de l'oral. Comme nous l'avons vu au chapitre II, les dossiers de **procédure** compilent toutes sortes de documents différents. De manière générale, le processus judiciaire et la justice produisent et se basent sur de nombreux documents textuels, pour lesquels l'intérêt des technologies va grandissant. Or, l'application de technologies informatiques au texte demande une bonne compréhension de leurs caractéristiques. Dans cette partie, nous souhaitons étudier différents types de textes en lien avec le processus judiciaire et la justice, y compris les **auditions** de **témoin**, afin d'en proposer une catégorisation en genres sur laquelle on pourra appuyer le développement de technologies.

### 2.1 *LegalNLP*, traitement automatique des langues et genre textuel

Nous partons du constat que le traitement automatique de la *langue juridique* ou de la *langue judiciaire* selon les cas est un nouveau sujet d'intérêt pour le domaine du **TAL** aussi bien dans l'industrie que dans la recherche universitaire. Il semble que ces dernières années a eu lieu une prise de conscience des besoins en technologie émanant du domaine du droit et de la quantité de données qui est générée. De cette

prise de conscience est née la *LegalTech*, un champ industriel et de recherche visant l'application des technologies numériques au service du droit, au sein de laquelle a germé le *LegalNLP*, qui se consacre spécifiquement aux aspects liant informatique, langue et droit.

La revue *Traitement Automatique des Langues* a consacré en 2017 un numéro spécial au traitement de la langue juridique<sup>6</sup>. Dans son introduction, NAZARENKO et WYNER (2017) indiquent :

La langue juridique est à entendre ici dans un sens large : elle est écrite ou orale ; elle recouvre la loi et les jugements aussi bien que les textes réglementaires comme les décrets, les règlements, les contrats ou les exigences.

Cette extrait montre que les coordinateurs du numéro connaissent la variété des productions qui peuvent tomber sous l'appellation de *langue juridique*. Le numéro compte deux contributions thématiques : l'une concerne des réglementations maritimes (EMANI et HARALAMBOUS, 2017), et l'autre le découpage automatique en phrases de décisions de justice américaines (SAVELKA et al., 2017). Si la première consacre une section à décrire en détails les documents traités (en termes de structure, de contenu, de volume, etc.), ce n'est pas le cas de l'autre contribution, qui évoque simplement 80 décisions issues de quatre domaines (cyber-crime, propriété intellectuelle, invalidité des anciens combattants, et cour suprême). L'absence de description des données et des corpus dans les travaux de traitement automatique de la langue est courante, les auteurs se bornant souvent à indiquer l'origine et la quantité de données avec laquelle ils travaillent.

Pourtant, BOMMIER-PINCEMIN (1999, p. 161), en s'appuyant sur RASTIER et al. (1994), déclare que la reconnaissance et la prise en compte des genres est une étape nécessaire à la réalisation d'une application qui opère des calculs sur des textes. Il a été démontré en effet que parmi les paramètres influençant la qualité des résultats produits par les traitements automatiques figure le genre textuel. NADEAU et SEKINE (2007) affirment que le genre textuel et le domaine ont été négligés dans la littérature concernant la reconnaissance d'entités nommées, et que

---

6. Appel à publications lisible ici : <https://tal-58-2.sciencesconf.org/> [consulté le 8 décembre 2019]

peu de systèmes de reconnaissance d'*entités nommées* ont été construits pour des domaines et des genres divers, alors que le transfert d'un système entre domaines représente un défi de taille. Dans leur article consacré au test d'un outil d'aide à l'élaboration de terminologies et d'ontologies intitulé CAMÉLÉON, basé sur des patrons lexico-syntaxiques supposés génériques, JACQUES et AUSSENAC-GILLES (2006) ont rapporté l'impact du genre textuel sur les performances de systèmes automatiques quelle qu'en soit la finalité. Le système a été testé sur plusieurs *corpus* : un guide de planification de réseau électrique, des articles scientifiques de la conférence Ingénierie des Connaissances, un manuel de géomorphologie, un manuel de spécifications logicielles dans le domaine de l'électricité, des articles de l'Encyclopaedia Universalis en géomorphologie, un manuel de parapente, plusieurs thèses en archéologie, des textes du domaine de la télécommunication. La variabilité des résultats produits par le système entraîne les auteurs d'une part à remettre en cause la généralité des patrons, et d'autre part à insister sur le besoin de caractérisation des textes et des genres textuels. Elles préconisent l'intégration de la question des genres textuels dans la recherche en TAL.

Qu'entend-on exactement par *genre d'un texte* ou *genre textuel*? RASTIER (2001, p. 299) en propose la définition suivante :

Genre : Programme de prescriptions (positives ou négatives) et de licences qui règlent la production et l'interprétation d'un texte. Tout texte relève d'un genre, et tout genre, d'un discours. Les genres n'appartiennent pas au système de la langue au sens strict, mais à d'autres normes sociales.

De même, cinq modalités de définition du genre sont énoncées par SWALES (1990) :

1. Le genre est une classe d'événements communicatifs,
2. Le critère principal qui transforme un ensemble d'événements communicatifs en un genre est un ensemble commun de buts communicatifs,
3. Les exemples de genres varient selon leur prototypicité,

4. La raison d'être d'un genre établit des contraintes quant au contenu, au positionnement et à la forme des contributions admissibles,
5. La nomenclature des genres d'une communauté discursive est une source importante d'information.<sup>7</sup>

Ces définitions partagent une vision du genre comme un ensemble de contraintes présidant à la production et à l'interprétation de productions langagières, contraintes découlant d'une pratique sociale qu'on appelle un discours et dans laquelle les genres s'inscrivent. Concernant le discours, RASTIER (2001, p. 298) le définit comme :

Ensemble d'usages linguistiques codifiés attaché à un type de pratique sociale.

Après ces éclaircissements, si l'on reconsidère l'extrait de l'introduction du numéro spécial de la revue TAL consacré à la langue juridique, l'inventaire détaillant ce qui est entendu comme relevant de la *langue juridique* peut être interprété comme décrivant un *discours*<sup>8</sup>.

SWALES (1990) considère que les étiquettes de genre telles que scientifique, médical, juridique ou médiatique font paraître le contenu des genres comme trompeusement homogène. Nous avons en effet constaté au chapitre II que le dossier de *procédure* uniquement compile des documents textuels hétérogènes, jouant un rôle dans un processus judiciaire mais de formes, de structures et de buts différents. La caractérisation de ces textes en termes de genre apparaît utile dans la perspective de leur interprétation linguistique mais aussi de leur traitement automatique.

Le sujet de la question des genres a déjà intéressé le domaine de la traduction judiciaire. MONJEAN-DECAUDIN (2012, p. 114) souligne l'hétérogénéité des documents que les traducteurs sont amenés à traduire pour la justice :

Les traducteurs sont amenés à traduire une palette infinie de documents, comme des rapports de toutes sortes (balistique, toxicologique, d'autopsie, d'expertises techniques, médicales, etc.) des comptes rendus d'accident, d'enquête policière, des transcriptions d'écoutes téléphoniques, des

---

7. Traduction par nos soins et assistée par le logiciel de traduction automatique en ligne DeepL : <https://www.deepl.com/translator> [consulté le 8 décembre 2019].

8. De même, NADEAU et SEKINE (2007) évoquent un *domain* mais *discourse* nous semblerait plus approprié.

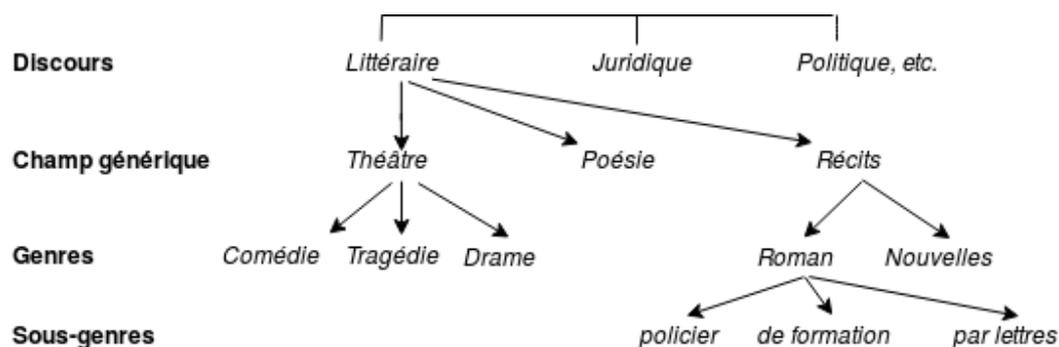


FIGURE 39 – Classification des discours en genres, schéma issu de MALRIEU et RASTIER (2001).

contrats commerciaux, des actes authentiques, des décisions de justice, des procès-verbaux de perquisition, mais plus simplement des extraits de compte bancaire, des cartes grises, etc. En établir une liste exhaustive s'avère impossible.

Très récemment, DIOT-PARVAZ AHMAD (2019) a produit une typologie des genres textuels du judiciaire en s'appuyant notamment sur sa pratique d'interprète judiciaire. Dans le cadre de cette pratique, comprendre les genres textuels permet d'aider le traducteur-interprète à comprendre le raisonnement du rédacteur (DIOT-PARVAZ AHMAD (2019, p. 120) citant DAMETTE (2013)).

Du côté de la linguistique textuelle, F. RASTIER évoque fréquemment un « discours juridique » sans toutefois entrer dans les détails de sa caractérisation. Par exemple, dans MALRIEU et RASTIER (2001) un « discours juridique » composé de « rapports, codes et lois » est comparé avec d'autres genres relevant des discours scientifique, littéraire et essayiste (figure 39). De même dans RASTIER (2001) où l'auteur mentionne souvent à titre d'exemple le « discours juridique », en parallèle du « discours littéraire », du « discours religieux », etc.

La caractérisation d'un genre textuel peut se faire selon différents critères : Citant BRANCA-ROSOFF (1999), JACQUES et AUSSENAC-GILLES (2006) rappellent que le genre textuel peut être caractérisé par des critères internes, les traits linguistiques, et externes, les pratiques sociales. Cette hypothèse semble assez largement admise depuis les travaux de BIBER (1988). Dans notre cas, étant donné que les pratiques sociales se rapportent à la justice et au droit, nous nous intéresserons plus aux critères internes.

Dans la littérature, nous avons rencontré de nombreux exemples de caractérisations des genres de tous domaines à l'aide de critères internes : BLASCO et CAPPEAU (2012) sur des interviews politiques orales et écrites, GLEDHILL et al. (2017) sur des directives européennes et leur transposition en droit français, TODIRASCU (2019) sur des sous-genres du genre journalistique (bulletin, interview, nécrologie, tribune, etc.). Tous s'appuient sur des **corpus**, et la plupart utilisent des critères morpho-syntaxiques, à l'exception de SCHNEDECKER (2017) qui propose une étude de l'usage des chaînes de référence. MALRIEU et RASTIER (2001) ayant établi la stabilité des **parties du discours** en tant que marqueurs du genre textuel, nous retenons nous aussi ces critères dans notre objectif de caractérisation du genre lié à la justice.

## 2.2 Corpus

Nous avons rassemblé un **corpus** composé de six sous-corpus de textes relevant au sens large du domaine de la justice, pouvant tous s'inscrire dans le cadre de l'appel du numéro thématique de la revue TAL cité précédemment :

- Trois codes : le code pénal, le code civil et le code du patrimoine ;
- Sept numéros du Journal Officiel de la République Française datant de l'automne 2018 ;
- Le « bloc de constitutionnalité » : il s'agit d'un ensemble composé du texte intégral de la Constitution de 1958, du préambule de 1946, de la Charte de l'environnement de 2005, et de la Déclaration des droits de l'homme et du citoyen de 1789 ;
- Un sous-corpus de 190 jugements de cour d'appel ;
- Un sous-corpus le 190 jugements de la Cour de cassation ;
- 370 **procès-verbaux d'auditions de témoins** réalisées au cours d'une **enquête** criminelle.

Les textes ont été collectés sur le site LegiFrance<sup>9</sup> pour les codes et les jugements, le site internet du Journal Officiel<sup>10</sup>, et le site du Conseil Constitutionnel<sup>11</sup>. Les **auditions** quant à elles sont celles qui ont déjà été évoquées dans ce manuscrit, mises à notre disposition dans le cadre de ce projet de recherche en lien avec le **PJGN**.

9. <https://www.legifrance.gouv.fr/> [consulté le 8 décembre 2019].

10. <https://www.journal-officiel.gouv.fr/> [consulté le 8 décembre 2019].

11. <https://www.conseil-constitutionnel.fr/> [consulté le 8 décembre 2019].

Sous-corpus	Nb. de textes	Nb. de mots	Moyenne
Codes	3	538 000	179 000
Journal Officiel	7	579 000	82 700
Bloc constitutionnel	4	15 000	5 000
Jurisprudence d'appel	190	524 000	2 760
Jurisprudence de Cassation	190	595 000	3 130
Auditions de témoins	370	577 000	1 560
<b>Total</b>	<b>764</b>	<b>2 828 000</b>	<b>3 700</b>

TABLE 10 – Statistiques du **corpus** (valeurs approximatives)

On obtient donc un **corpus** équilibré composé de cinq sous-corpus de cinq à six cent mille mots, et un sous-corpus d'environ quinze mille mots dont nous étudierons les propriétés morpho-syntaxiques à l'aide du logiciel de textométrie TXM (HEIDEN et al., 2010), qui intègre l'étiqueteur automatique en **parties du discours** TreeTagger (SCHMID, 1994).

### 2.3 L'analyse factorielle des correspondances

La méthode utilisée pour cette analyse repose sur une analyse en composante multivariée, dans notre cas une analyse factorielle des correspondances (AFC) sur les **parties du discours**. L'AFC est une méthode statistique d'analyse de données organisées en tableau et développée par BENZÉCRI (1973), qui permet de représenter des proximités entre lignes et colonnes, couramment utilisée en analyse de données textuelles. Outre TXM, les logiciels Lexico<sup>12</sup>, Iramuteq<sup>13</sup> ou encore Alceste<sup>14</sup> implémentent également l'AFC.

Pour expliquer le principe de l'AFC, nous reprenons les figures données en exemple sur le site d'analyse statistique STHDA<sup>15</sup>.

Les données examinées sont entrées dans un tableau de contingence. À la figure 40, le tableau consigne la répartition des tâches ménagères dans un couple hétérosexuel. Les lignes contiennent les tâches et les colonnes contiennent la répartition : soit la femme seule, soit l'homme seul, soit en alternance, soit conjointement.

12. <http://www.lexi-co.com/L5Presentation.html> [consulté le 9 décembre 2019].

13. <http://www.iramuteq.org/> [consulté le 9 décembre 2019].

14. <https://www.image-zafar.com/Logiciel.html> [consulté le 9 décembre 2019].

15. <http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique-74-afc-analyse-factorielle-des-correspondances-avec-r-l-essentiel/> [consulté le 9 décembre 2019].

	Wife	Alternating	Husband	Jointly
Laundry	156	14	2	4
Main_meal	124	20	5	4
Dinner	77	11	7	13
Breakfast	82	36	15	7
Tidying	53	11	1	57
Dishes	32	24	4	53
Shopping	33	23	9	55
Official	12	46	23	15
Driving	10	51	75	3
Finances	13	13	21	66
Insurance	8	1	53	77
Repairs	0	3	160	2
Holidays	0	1	6	153

FIGURE 40 – Tableau de contingence

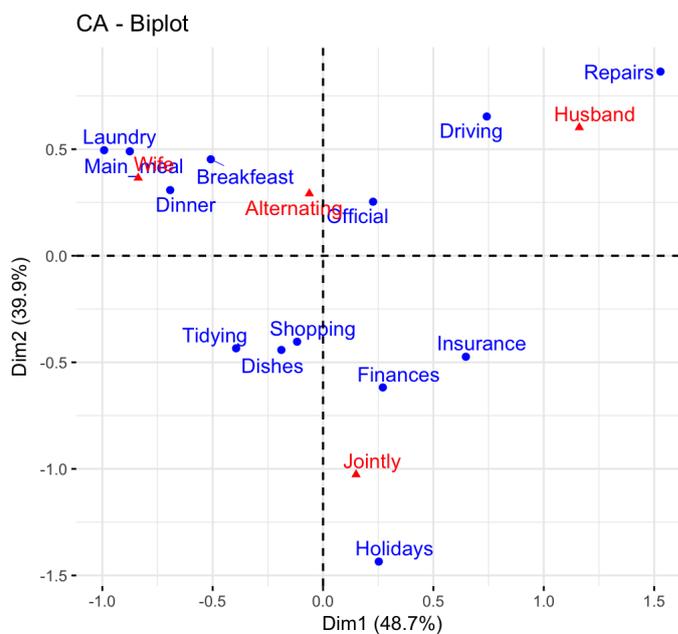


FIGURE 41 – Graphe d'AFC d'après les données du tableau 40.

Les résultats de l'AFC sont présentés à la figure 41. Les lignes figurent en bleu et les colonnes en rouge. La présentation sous forme de graphe permet de montrer la proximité des points, et donc leur corrélation. On remarque que les tâches très caractéristiques de l'homme sont les réparations et la conduite, tandis que celles associées fortement à la femme sont la lessive et les trois repas. On peut également souligner que la femme est caractérisée par le double de tâches par rapport à l'homme. Une seule tâche est fortement caractérisée par le partage des tâches, il s'agit de la planification des vacances. Un groupe de tâches est soit partagé soit effectué exclusivement par la femme, il s'agit du ménage, de la vaisselle et des courses, et un groupe de tâches est effectué soit exclusivement par l'homme soit partagé, il s'agit des finances et de la gestion des assurances.

Dans notre cas, le tableau de contingence consigne dans les lignes les **parties du discours**, et dans les colonnes les sous-corpus (annexe C).

## 2.4 Analyse

Le calcul de l'AFC sur le **corpus** constitué fournit le graphe présenté à la figure 42. Les points bleus représentent les **parties du discours** (les lignes) et les points orange les sous-corpus (les colonnes).

On peut observer sur ce graphe :

- Le rapprochement entre le bloc constitutionnel et les codes, on y fera désormais référence comme le « groupe législatif »,
- Le rapprochement entre les sous-corpus de jurisprudences, auquel on fera désormais référence comme le sous-corpus « jurisprudence ».

En revanche, le Journal Officiel ainsi que les **auditions** se distinguent des autres sous-corpus. On se retrouve donc avec quatre pôles pour la poursuite des analyses (entourés en rouge sur la figure 43), auxquels on peut associer à chacun un ensemble de **parties du discours** caractéristiques (en vert sur la figure 43). Il reste un groupe d'étiquettes à mi-chemin entre deux pôles, manifesté en bleu.

Afin de caractériser chaque sous-corpus, nous examinons les **parties du discours** qui en sont proches sur le graphe.



**Journal Officiel** Le sous-corpus Journal Officiel est caractérisé par les ponctuations, ponctuations de citation, les valeurs numériques et les symboles. Ceci peut être interprété comme caractéristique des références à d'autres textes (décrets, lois), ainsi que de la mention de dates et de valeurs numériques diverses (pourcentages, sommes d'argent). L'exemple ci-dessous en est représentatif :

Par arrêté du Premier ministre et du ministre de l'action et des comptes publics en date du **24 septembre 2018**, le taux mentionné au dernier alinéa de l'article **11** du décret no **99-945** du **16 novembre 1999** modifié portant statut particulier du corps des administrateurs civils et au **IV** de l'article **1er** du décret no **2005-1090** du **1er septembre 2005** relatif à l'avancement de grade dans le corps des administrations de l'Etat est fixé à **24 %** pour les promotions prononcées au titre des années **2019, 2020 et 2021**.

On note aussi la proximité des verbes au futur, marque de sa fonction d'annonce publique pour la diffusion de lois et décrets, mais également d'offres d'emplois publics, de modalités de concours de la fonction publique, de dates de réunions parlementaires, etc.

Art. 3. – Toute battue collective au cours de laquelle des chevrotines **seront** employées **devra** être inscrite sur un registre retiré auprès de la fédération départementale des chasseurs.

Le titulaire de l'emploi **assurera** la fonction d'adjoint au directeur de la direction de la législation fiscale.

**Jurisprudence** Les deux sous-corpus de jurisprudence, rassemblés en un seul pôle pour l'analyse, se caractérisent par les participes présent et passé, ce que l'on peut interpréter comme la trace de la construction du raisonnement et le rapport des faits investigués.

ARRÊT : Contradictoire, prononcé publiquement par mise à disposition de l'arrêt au greffe de la cour, les parties en **ayant été** préalablement avisées conformément à l'article 450 al 2 du CPC.

Il produit en outre un courrier adressé par l'agence régionale de santé, le

26 octobre 2012, **constatant** l'absence de traitement de l'eau de la piscine, ceci **présentant** un risque pour les baigneurs.

En application des dispositions de l'article 945-1 du code de procédure civile, l'affaire a été **débattue** le 29 Octobre 2013 à 14 H 00, en audience publique, les parties ne s'y **étant** pas opposées.

**Groupe législatif** Les codes et le bloc constitutionnel constituent le « pôle » le moins marqué par les **parties du discours**.

Les articles en sont la catégorie la plus caractéristique, mais il est plus intéressant de se pencher sur les verbes à l'infinitif : on remarque l'association des verbes à l'infinitif pour ce groupe. Néanmoins, nous en avons approfondi l'analyse en consultant leurs co-occurents les plus fréquents. On y trouve « peut », « peuvent », « pourra », « doit », « doivent » et « ne », ce que l'on peut interpréter comme relatif à la fonction coercitive de ces deux groupes de textes, précisant ce que chacun peut et doit (ou ne peut et ne doit faire) pour être en accord avec la loi :

Article 4 Toute personne **doit contribuer** à la réparation des dommages qu'elle cause à l'environnement, dans les conditions définies par la loi.

Toutefois, ces autorités **ne peuvent procéder** à la célébration du mariage entre un Français et un étranger que dans les pays qui sont désignés par décret.

On note également la proximité des pronoms indéfinis, dont la fonction référentielle prend une dimension exprimant l'universalité de l'application de la loi :

4. **Tout** homme persécuté en raison de son action en faveur de la liberté a droit d'asile sur les territoires de la République. 5. **Chacun** a le devoir de travailler et le droit d'obtenir un emploi. Article 221-5 Le fait d'attenter à la vie d'**autrui** par l'emploi ou l'administration de substances de nature à entraîner la mort constitue un empoisonnement.

**Auditions de témoins** Les **auditions de témoins** sont caractérisées par l'imparfait, ainsi que par différents pronoms et par les interjections. L'imparfait, temps du récit, est le signe de l'aspect narratif de ce sous-corpus. Les pronoms correspondent à

la pratique de l'**audition**, au cours de laquelle un **témoin** relate son expérience, expérience consignée à la première personne. Enfin, les interjections (essentiellement « oui » et « non ») sont une trace de l'aspect oral transcrit.

Les filles **étaient** maltraités, d'après ce que **je** sais il y **avait** peut-être de la violence et la mère **allait** travailler en laissant les filles enfermées à la maison.

Question : **Vous** est-il arrivé au cours de cette journée de boire de l'alcool ?

Réponse : **Oui**, j'ai bu deux verres de vin au repas de midi.

### Parties du discours partagées par le groupe législatif et les auditions de témoins

Un groupe de **parties du discours** (entouré en bleu sur la figure 43) est positionné entre le sous-corpus des **auditions** et le groupe législatif. On repère notamment en son sein la présence des verbes au conditionnel, qui s'explique différemment pour chacun des sous-corpus.

Pour le sous-corpus des **auditions**, l'emploi du conditionnel correspond au futur dans le passé. Cet usage est caractéristique du discours indirect et correspond à des propos tiers rapportés par le **témoin** entendu.

Il a dit qu'il n'avait pas le temps mais qu'il **viendrait** à sa pause.

Il m'a dit de ne pas inquiéter les parents, que dès qu'elle **rentrerait**, il me **rappellerait**.

Le conditionnel est également employé pour formuler des hypothèses sur les faits ou les personnes, le plus souvent dans un emploi d'irréel du passé

Il ne comprends pas, et s'il avait eu le moindre soupçon il en **aurait** parlé.

Si Christine lui avait annoncé son intention de ne plus se marier, Thierry se serait effondré et serait venu nous voir.

Dans le groupe législatif, l'emploi du conditionnel permet de prévoir les conditions d'une situation et donc les modalités d'application du texte.

Dans le cas où la déclaration **aurait** été omise ou **serait** erronée, la rectification de l'acte, en ce qui touche l'omission ou l'erreur, pourra être effectuée conformément à l'article 99-1.

Ceux de ces textes qui **interviendraient** après l'entrée en vigueur de la présente Constitution ne pourront être modifiés par décret que si le Conseil constitutionnel a déclaré qu'ils ont un caractère réglementaire en vertu de l'alinéa précédent.

## 2.5 Critique de l'étiquetage en parties du discours

L'analyse réalisée donne à voir le contraste marqué entre les différents sous-corpus de textes étudiés. La nature narrative de l'audition de **témoin** la dégage de l'ensemble du reste du **corpus**. Le journal officiel, également détaché du reste des autres sous-corpus, l'est probablement par la diversité des informations qu'il rapporte. Le bloc de constitutionnalité et les codes, textes dont la fonction est d'énoncer la norme, à des niveaux et dans des domaines différents, sont en revanche très proches.

Ainsi, l'usage des **parties du discours** s'est avéré pertinent, et plus spécifiquement encore, les temps des verbes sont apparus comme propres à chaque sous-corpus examiné. Toutefois, nous avons remarqué des problèmes d'étiquetage au cours de l'exploration des concordances, en particulier dans le cas des temps de verbes : certains étiquetages sont corrects, comme l'imparfait, alors que d'autres présentent de nombreuses erreurs. C'est le cas du passé simple et du subjonctif imparfait. Le passé simple est attribué à des noms propres, des abréviations, ainsi qu'à la conjonction « si » et à la préposition « de » employées en début de phrase, des tournures propres aux documents juridiques. Le subjonctif imparfait est attribué à des abréviations ou à des mots présentant des fautes de conjugaison et d'orthographe. Quoiqu'il en soit, l'impact sur le calcul de l'**AFC** est cependant limité, car ces erreurs d'étiquetage ne comptent que peu d'occurrences comparé au reste des **parties du discours** comme on peut le voir dans le tableau de contingence fourni en annexe C. D'autre part, l'étiquetage morpho-syntaxique de TreeTagger se fonde sur les **tokens**, ce qui pose problème pour l'interprétation des temps composés, dont l'auxiliaire et le participe sont annotés indépendamment l'un de l'autre.

En l'occurrence, ces erreurs d'étiquetage sur nos **corpus** démontrent que l'emploi d'outils automatiques qui ne sont pas adaptés à un genre en particulier peut être la

cause d'erreurs. C'est pourquoi il est nécessaire de rester vigilant sur les sorties produites par des traitements automatiques, afin de pouvoir identifier les erreurs et de les interpréter pour permettre leur correction. L'étiquetage automatique en **parties du discours** n'est pas une tâche fiable à 100% (MANNING, 2011). Dans notre cas, l'amélioration des performances passerait par l'entraînement spécifique de TreeTagger sur des données similaires à celles que nous étudions.

## 2.6 Conclusion

L'objectif de cette section était de comparer des textes concernant la justice et le droit afin de déterminer leurs similitudes et différences en termes linguistiques.

Après examen d'un **corpus** constitué *ad hoc* de textes liés à la loi, à son application, et aux **enquêtes** judiciaires, nous disposons de critères objectifs permettant leur distinction. Nous avons remarqué que les temps de verbes se sont révélés, parmi les autres **parties du discours**, comme particulièrement discriminants de chaque sous-ensemble de textes.

En fait de « texte légal », « genre juridique » ou de « discours juridique », l'étude textométrique a permis d'entrevoir une distinction entre documents *juridiques* et *judiciaires*. Le *Vocabulaire juridique* de CORNU (2016) précise à l'entrée « judiciaire » : « qui appartient à la justice, par opposition à législatif et administratif » ; et à « juridique » : « de droit, en droit ; qui a trait au droit », par opposition notamment « à une démarche non-normative ».

Selon WROBLEWSKI (1988), le discours juridique est « le discours dans lequel on formule le droit ou dans lequel on parle du droit ». Les textes juridiques sont donc des textes comportant des termes de la langue de spécialité du droit.

*Judiciaire* en revanche désigne les documents intégrés dans une **procédure** judiciaire, *juridique* et *judiciaire* étant tout de même très liés, comme l'explique MONJEAN-DECAUDIN (2012, p. 114) :

Lorsqu'un document est rattaché au déroulement d'une procédure judiciaire, il peut être qualifié de texte judiciaire, mais ce n'est pas pour autant qu'il appartient à la catégorie des textes juridiques. Sachant qu'est judiciaire ce qui « appartient à la justice, par opposition à législatif et

administratif, ce qui concerne la justice rendue par les tribunaux judiciaires »<sup>16</sup>, il en résulte que tout texte qui est rattaché à la justice ou qui intègre une procédure peut être qualifié de judiciaire. Cependant, seuls certains d'entre eux appartiennent à la catégorie des textes juridiques.

Dans cette configuration, les **procès-verbaux d'audition de témoin** sont des textes judiciaires non juridiques, du fait qu'ils sont rattachés à une **procédure** judiciaire mais qu'ils ne sont pas rédigés dans une langue qui formule le droit, tandis que les codes et le bloc constitutionnel sont des textes du genre juridique puisqu'ils entraînent des effets de droits.

Toutefois, il existe également des documents juridico-judiciaires : il s'agit de documents impliqués en **procédure** ayant des effets de droit, parmi lesquels DIOT-PARVAZ AHMAD (2019, p 168-169) mentionne les citations, les ordonnances, les jugements, etc. Dans le cadre de notre étude, cela semble correspondre au sous-corpus des jurisprudences.

Plus loin, S. MONJEAN-DECAUDIN ajoute :

Un document judiciaire peut être non juridique mais présenter une technicité terminologique autre que celle du droit.

Un bon exemple de ce type de cas est celui des documents d'expertise médicale. Un rapport d'autopsie est avant tout un document médical, avec ce que cela implique pour le vocabulaire, la structure du document, les tournures, mais du fait de son intégration dans la **procédure**, il acquiert une dimension judiciaire. L'interprétation que l'on fait de ce document, c'est-à-dire la lumière à laquelle on le reçoit, détermine en partie le genre duquel il relève.

Finalement, cette distinction donne sens à la catégorisation des documents entre documents réglementaires, qui documentent le déroulement de la **procédure** et des actes qui se déroulent dans son cadre, et documents d'information, qui apportent des informations relatives aux faits examinés que nous avons formulée au chapitre II.

Le cas des documents juridiques et judiciaires vient alors conforter PINCEMIN (2012a), qui préconise une approche herméneutique des genres prenant en compte non seulement les propriétés intrinsèques du texte mais également le point de vue

---

16. CORNU, 2016, p. 521

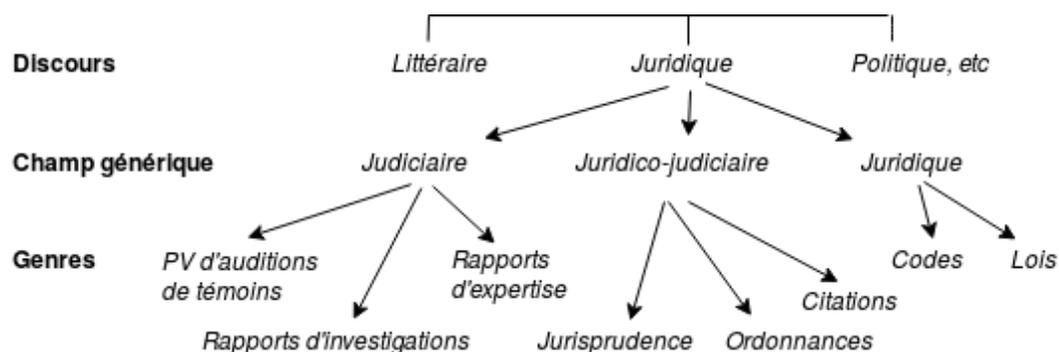


FIGURE 44 – Proposition d'organisation du discours légal en genres

que l'on choisit de leur appliquer, plutôt qu'une conception des genres qui cherche à lier par essence un texte à un genre.

Ainsi comme nous le présumons, sous les appellations de texte légal, genre juridique, langue juridique, etc., la réalité n'est pas homogène : il s'agirait plutôt d'un discours, qu'on appellerait le discours juridique, lui-même décomposé en champs génériques judiciaire, juridico-judiciaire et juridique, eux-mêmes décomposés en plusieurs sous-genres.

Pour synthétiser cette organisation, nous reprenons le schéma de la figure 39 issue de MALRIEU et RASTIER (2001) adapté à cette structure.

Nous avons dans cette section tenté de donner un aperçu de la diversité des textes impliqués dans la justice à l'aide de critères objectifs. Cela nous a permis au passage de proposer une organisation de ces textes en termes de genres et de discours. Vu l'intérêt de la technologie et de la recherche pour les données textuelles du droit et de la justice, nous estimons nécessaire de clarifier leur nature, et de démontrer que ces distinctions se constatent via des critères objectifs éprouvés pour caractériser le genre textuel.

### 3 Conclusion

Dans ce chapitre, nous avons souhaité examiner une dimension textuelle de la *procédure*, celles des *auditions* de *témoins*, puisqu'elles ont représenté notre matériau de travail jusqu'à présent. Pour cela, nous avons décrit la pratique et les productions des *auditions* de *témoins* réalisées dans le cadre d'une affaire criminelle. Cette description a permis de mettre en valeur les particularités du texte des

*procès-verbaux* d'*audition*, en particulier de souligner un certain flottement concernant l'auteur du document, ainsi que leur aspect narratif.

La deuxième partie, consacrée à une étude du genre textuel de documents impliqués dans le cadre de la justice et du droit, a permis d'illustrer concrètement via des critères objectifs les théories de la juritraductologie qui organisent les textes selon leur finalité juridique, juridico-judiciaire ou juridique.

Ainsi, la complexité des documents textuels a été démontrée aussi bien par l'exploration interne des *procès-verbaux* d'*auditions* de *témoin* que par leur mise en parallèle avec d'autres documents textuels judiciaires et juridiques.

Les observations formulées, dont l'utilité est déjà avérée dans le cadre de la pratique professionnelle de traducteur-interprète judiciaire, doivent également servir de cadre interprétatif et applicatif pour les futurs développements de technologies textuelles dédiées au genre légal.



## Conclusion générale

Cette thèse nous a menée à explorer les possibilités d'adaptation de méthodes d'[extraction d'information](#) et d'exploration textuelle dans le cadre d'une pratique d'[enquête](#) employée par la [Gendarmerie nationale](#), l'[analyse criminelle](#).

L'[analyse criminelle](#), dont l'objectif dans le contexte de la [police judiciaire](#) est de structurer les informations collectées dans une [enquête](#), souffre aujourd'hui de difficultés liées à la densité d'information non-structurée avec laquelle elle compose. Basée sur la reprise minutieuse du dossier de [procédure](#) judiciaire à la recherche d'information, elle vise justement à fournir une vue globale sur l'ensemble du dossier ou sur des aspects spécifiques, en s'appuyant notamment sur la création de schémas chronologiques ou relationnels. Pour cela, les [analystes criminels](#) relisent le dossier de [procédure](#) et en extraient l'information conceptualisée comme des entités criminelles, c'est-à-dire tout objet du monde réel impliqué dans les faits et mentionné dans les documents du dossier de [procédure](#). De nombreux facteurs y font obstacle : de la qualité de la numérisation des fichiers à l'absence d'organisation normalisée du dossier en passant par le référencement des pièces, il s'agit d'une problématique de gestion de l'information et de gestion documentaire à part entière, pour laquelle il n'existe à ce jour pas d'outil informatique dédié. Les [analystes criminels](#) font avec un logiciel propriétaire d'analyse de données pour l'élaboration des schémas, et avec des logiciels de bureautique classique pour parcourir les documents.

Étant donné que ce dossier de [procédure](#) est la base de travail des [analystes criminels](#), nous avons consacré le deuxième chapitre à sa description. Nous avons fait l'inventaire des pièces dans les [procédures](#) qui nous ont servi d'exemples, et avons découvert une grande variété de documents, divisés en deux grandes catégories : les pièces réglementaires, qui documentent le déroulement des [actes d'enquête](#), et les

documents d'information, qui rapportent des informations au sujet des faits examinés. Une telle variété dans les pièces de *procédure* demandait d'opérer un tri dans l'objectif de l'application de traitements automatiques. Leur hétérogénéité ne permet pas un traitement global du dossier, les types de pièces relevant de types documentaires différents : texte normé, texte libre, tableaux de données, photographies, images et vidéos. Après cet inventaire, nous avons considéré la notion de *corpus* du point de vue de la linguistique de *corpus*, afin de cerner le type de pièce le plus approprié pour nos recherches. En termes de volume dans la *procédure*, d'intérêt pour l'*analyse criminelle*, et d'enjeux linguistiques et informatiques, nous avons estimé que l'ensemble des *auditions* de *témoin* constituait le matériau le plus pertinent.

Dans l'état de l'art, nous avons passé en revue les apports théoriques et techniques des humanités numériques, de la linguistique de *corpus* et du *TAL* en particulier en ce qui concerne l'*extraction d'information* et la détection d'*entités nommées*. Ce panorama nous a aidée à préciser les contours conceptuels d'une solution informatique dédiée à la gestion de l'information de la *procédure* judiciaire dans l'objectif de l'*analyse criminelle*, et à définir le rôle de l'*analyste criminel* vis-à-vis d'une telle solution. Celui-ci restera le maître d'œuvre et de réflexion, les résultats obtenus par le logiciel étant une extraction brute qui doit être valorisée dans les directions fixées par les objectifs de l'analyse. Ainsi, il faut chercher à créer une synergie fructueuse entre l'intelligence et l'expérience du professionnel et les apports de la technologie. À l'issue de ce chapitre, nous avons comparé les notions d'*entité nommée* propre au *TAL* et celle d'entité criminelle propre à l'*analyse criminelle*. Nous en avons conclu qu'*entités nommées* et entités criminelles partagent un référentiel commun, les concepts, objets et êtres du monde réel, mais que leurs manifestations textuelles divergent, d'autant plus que les entités criminelles ne figurent pas uniquement dans le dossier sous forme textuelle. Par conséquent, nous avons redéfini l'objet de notre recherche comme des *entités criminelles textuelles*.

Le chapitre suivant a consisté à réaliser une étude de la manifestation textuelle de cinq types d'entités criminelles dans les *auditions* de *témoin* : les numéros de téléphone, les personnes, les informations temporelles, les informations spatiales, et les véhicules. Ce relevé a permis d'illustrer la diversité des formes linguistiques

des entités, qui ne se limitent pas à des mentions bien bornées et facilement identifiables de type « nom prénom » par exemple, mais qui peuvent parfois représenter des paragraphes entiers de description. Nous avons attribué des formes à la nature des textes dans lesquels elles apparaissent, puisque les *auditions* de *témoin* sont fortement teintées par l'incertitude des personnes entendues qui ne peuvent pas toujours nommer ce qu'elles ont vu. Par contraste avec le concept de *description définie*, qui formalise les structures référentielles en « le + substantif », nous avons baptisé ces mentions *descriptions indéfinies*. En conclusion, nous avons établi que les *entités criminelles textuelles* se manifestent sous trois formes : les *entités nommées*, les *descriptions définies*, et les descriptions indéfinies. Ainsi, si *entités nommées* et entités criminelles sont connectées, ces dernières dépassent les premières. La prise en compte des *entités criminelles textuelles* représente le défi principal de l'*extraction d'information* des *auditions* de *témoin*. Dans la deuxième partie de ce chapitre, nous avons développé une approche de détection des entités criminelles nommées à l'aide de graphes UNITEX. Une approche différente a dû être adoptée pour chaque type d'entité, combinant lexiques constitués sur des ressources publiques et règles de description. Les résultats d'annotation automatique ont été évalués par rapport à un standard annoté manuellement et mesurés à l'aide des métriques classiques de l'*extraction d'information*, la *précision*, le *rappel*, et la *F-mesure*. Les taux sont variables selon les types d'entités mais sont néanmoins encourageants considérant que notre approche est relativement rudimentaire et que nous disposions d'un *corpus* de petite taille en comparaison des volumes de données normalement employées pour ce genre de tâches.

Constatant via l'exploration des entités les particularités du texte des *auditions*, nous avons réalisé une étude de ce texte dans le dernier chapitre. À l'aide d'outils de linguistique de *corpus*, nous avons mis en lumière la nature à la fois narrative et normée des *procès-verbaux* d'*auditions*, qui, bien que rédigés aux premières personnes du singulier et du pluriel, ne sont pas pour autant des récits de première main des propos de la personne entendue. Nous avons établi que le *procès-verbal* d'*audition* contient en réalité deux récits imbriqués, celui des faits rapportés par le *témoin*, et celui de l'*audition*. Dans les deux cas, l'auteur du texte est la personne menant l'*audition* aussi bien du point de vue légal que du point de vue linguistique.

Dans la seconde partie du chapitre, nous avons procédé à une étude textométrique visant à comparer différents textes employés ou produits dans des processus légaux et de justice. Cet intérêt émanait de la constatation que le domaine légal attire l'attention des milieux technologiques et de la recherche en [TAL](#). Or, il est avéré que le genre textuel est un paramètre influençant les performances d'applications de [TAL](#). Nous avons alors souhaité démontrer par le biais de critères éprouvés que ces textes sont loin de fonctionner tous de la même manière. Les résultats de notre étude ont pu être reliés à des travaux de jurilinguistique ayant déjà établi une classification du genre textuel judiciaire.

Pour la poursuite des travaux, en particulier concernant les aspects informatiques dont l'approfondissement serait la suite logique de cette thèse, il sera nécessaire de trouver comment établir les conditions d'une collaboration productive entre chercheurs et forces de police, ménageant les contraintes légales des uns tout en satisfaisant les besoins en données des autres. Plusieurs fois lors de notre recherche a été évoquée l'idée de travailler sur des données anonymisées, en supposant que cela permettrait de mettre à disposition de plus grands corpus. Cette idée, bien qu'intéressante, ouvre en réalité un autre débat. D'une part, anonymiser des documents automatiquement est une tâche de [TAL](#) à part entière, dont la base est de repérer les éléments à anonymiser, c'est-à-dire... les noms de personnes, les lieux, les dates. Le problème tourne en rond. D'autre part, on peut s'interroger sur la pertinence de changer un prénom par un autre ou de le remplacer par « XXX » dans un dossier de [procédure](#) judiciaire, alors que les faits sont parfois eux-mêmes révélateurs du dossier : si l'on évoque un petit garçon retrouvé pieds et poings liés dans une rivière, qu'il s'appelle Thomas, Mehdi ou Alexandre dans le texte ne dupe personne. L'exemple est caricatural mais a l'avantage d'être parlant. Contrairement aux dossiers cliniques de patients, matière de prédilection de l'anonymisation et qui compilent des données personnelles mais qui n'intéressent a priori que lesdits patients et leurs médecins, les dossiers de [procédure](#) judiciaire relèvent parfois de la sphère publique et à ce titre font l'objet de l'attention médiatique et de la convoitise des journalistes, parfois même des années après les faits ou la clôture de la [procédure](#). Il sera donc très important de préserver la sécurité de la circulation des données.

Au commencement de cette thèse, la particularité des données nous a frappée d'emblée, tout comme la rareté (voire l'inexistence) des contributions en linguistique et en informatique appliquées à l'étude du dossier de [procédure](#) judiciaire<sup>17</sup>. Plutôt que d'embarquer directement dans une direction plaquant des approches de [TAL](#) et d'[extraction d'information](#) existantes, il nous a paru primordial d'emprunter un chemin de recherche qui passerait par la compréhension des enjeux de l'[analyse criminelle](#) et de sa matière de travail, afin de cerner exactement le besoin, un besoin que nous avons perçu comme étant celui de chercher à assister sans se substituer. Vu la singularité des objets que nous avons découverts, il apparaît que cette phase était effectivement indispensable, et a permis de délimiter les bases des recherches futures en [extraction d'information](#) du dossier de [procédure](#).

Finalement, la contribution des travaux, que bien entendu nous aurions souhaités opérationnels au terme de ces trois années, n'a fait qu'entr'ouvrir un champ de recherche nouveau dans lequel informatique et linguistique, déjà habituées à collaborer, doivent se conjuguer au service de l'[analyse criminelle](#). Nous sommes convaincue que chacune y trouvera de nouveaux objets, thèmes et perspectives de recherche.

---

17. Tout du moins pour la recherche francophone. La communauté de la *forensic linguistics*, largement plus développée dans les pays anglo-saxons, est active sur les thématiques touchant en général aux questions liant justice et linguistique, D. WRIGHT (Nottingham Trent University) affirmant que le but de cette discipline est d'améliorer la délivrance de la justice (« Forensic linguistics should aim at improving the delivery of justice. », conférence plénière de la table ronde annuelle de la German Society for Forensics Linguistics, septembre 2019, Graz). Toutefois, ses pratiques s'inscrivent dans des cadres juridiques et judiciaires différents, et il peut être difficile de les rattacher au cas français.



## Annexe A

# Guide pour l'annotation manuelle des entités

## Introduction

L'annotation porte sur trois types d'entités : noms propres de personnes, dates et lieux. Le but de ce guide d'annotation est de fixer les règles permettant d'annoter ces entités.

## Noms propres de personnes

Nous considérons qu'un nom propre de personne est composée au minimum d'un segment nom, auquel s'ajoute un segment prénom, et d'autres éléments optionnels.

### « prénom nom » ou « nom prénom »

Tout segment composé d'un prénom suivi d'un nom ou d'un nom suivi d'un prénom doit être annoté <persName>. La balise <persName> se décompose en balises <firstname> et <lastname>. Exemples :

```
<?xml version="1.0" encoding="UTF-8"?>
<names>
<persName><firstname>Pierre</firstname> <lastname>Dupont</lastname></persName>
<persName><lastname>Jeandin</lastname> <firstname>Marie</firstname></persName>
</names>
```

Le prénom peut être composé, avec ou sans tiret. Exemples :

```
<?xml version="1.0" encoding="UTF-8"?>
<persName><firstname>Jean-Marie</firstname> <lastname>Martin</lastname></persName>
<persName><firstname>Mai Lan</firstname> <lastname>Nguyen</lastname></persName>
```

Le nom peut être composé avec ou sans tiret, et comporter une ou des particules onomastiques. L'annotation doit inclure toutes les particules onomastiques et tous les segments composants le nom. Exemples :

```
<?xml version="1.0" encoding="UTF-8"?>
<persName><firstname>Louis</firstname> <lastname>de Broglie</lastname></persName>
<persName><firstname>Carla</firstname> <lastname>Bruni-Sarkozy</lastname></persName>
<persName><firstname>Louis</firstname> <lastname>de Funès de Galarza</lastname></persName>
<persName><lastname>de la Fressange</lastname> <firstname>Inès</firstname></persName>
<persName><lastname>ben Laden</lastname> <firstname>Oussama</firstname></persName>
<persName><firstname>Manuel</firstname> <lastname>dos Santos</lastname></persName>
<persName><firstname>Lucien</firstname> <lastname>van Impe</lastname></persName>
```

### « nom » seul

Lorsque le nom est seul, il est annoté <lastname> et <persName>. Exemples :

```
<?xml version="1.0" encoding="UTF-8"?>
<persName><lastname>Mercier</lastname></persName>
<persName><lastname>Duval</lastname></persName>
```

### « titre nom »

Dans les cas où le nom de personne est constitué uniquement d'un titre et d'un patronyme, l'annotation <persName> doit englober le titre et le nom. Si le segment est composé de « titre prénom nom » ou « titre nom prénom », seuls les segments « prénom nom » et « nom prénom » sont annotés, puisqu'ils permettent d'identifier une personne sans ambiguïté.

Ces titres comprennent : les titres civils (monsieur, madame, mademoiselle, veuve, etc.) les titres nobiliaires (marquis, baron, etc.), les fonctions (juge, procureur, notaire, professeur, président, etc.), les titres religieux (frère, sœur, monseigneur, etc),

et les grades (commissaire, général, commandant, etc.). Le titre est annoté `<persTitle>`. Exemples :

```
<?xml version="1.0" encoding="UTF-8"?>
Melle <persName><firstname>Mélanie</firstname> <lastname>Gaillard</lastname></persName>
<persName><persTitle>Madame</persTitle> <lastname>Martin</lastname></persName>
<persName><persTitle>M.</persTitle> <lastname>Dubois</lastname></persName>
<persName><persTitle>Général</persTitle> <lastname>Leclerc</lastname></persName>
<persName><persTitle>Procureur</persTitle> <lastname>Molins</lastname></persName>
```

## Noms incluant une initiale

Dans le cas des noms de personnes comportant une ou plusieurs initiales, celles-ci doivent être annotées dans une balise `<initial>`. Exemples :

```
<?xml version="1.0" encoding="UTF-8"?>
<persName><firstname>Philip</firstname> <initial>K.</initial> <lastname>Dick</lastname></persName>
<persName><firstname>John</firstname> <initial>Q.</initial> <lastname>Public</lastname></persName>
<persName><firstname>Cecil</firstname> <initial>B.</initial> <lastname>DeMille</lastname></persName>
<persName><firstname>Michael</firstname> <initial>C.</initial> <lastname>Hall</lastname></persName>
```

## Résumé

Balise globale :

— `<persName>`

Balise inférieure :

— `<persTitle>`

— `<firstname>`

— `<initial>`

— `<lastname>`

## Dates

Est considéré comme date tout segment mentionnant au minimum soit un jour et un mois, soit un jour de la semaine et un numéro. Comme dans la section 1, chaque

unité doit être annotée individuellement (<weekday>, <day>, <month>, <year>) au sein de l'annotation globale <date>.

Nous distinguons plusieurs formats de dates qui sont détaillés ci-dessous. Chaque type de date est inséré dans la balise <date> à l'aide de l'attribut « type », dont les valeurs peuvent être spelled1, spelled2, numeric, et rel (relative).

### En toutes lettres

Les dates en toutes lettres peuvent être écrites dans deux ordres différents : « chiffre mois année » ou « chiffre mois » (type 1) ou « année chiffre mois » (type 2). Afin de faciliter le traitement ultérieur de ces deux formats, l'ordre est spécifié en attribut de la balise <date> (« spelled1 » pour type 1 et « spelled2 » type 2). En plus des éléments énoncés, ce format de date peut comporter des éléments complémentaires (« de l'an », jour de la semaine). Ces éléments étant superflus pour la désambiguïsation de la date, ils ne sont pas inclus dans l'annotation. Exemples :

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
le mercredi <date type="spelled1"><day>14</day> <month>avril</month> <year>2003</year></date>
```

```
l'an <date type="spelled2"><year>deux mille-sept</year>, le <day>seize</day> <month>août</month></date>
```

### En chiffres

Les dates au format numérique sont composées de chiffres entrecoupés de barres obliques ou de points. Chaque élément est annoté. Exemples :

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
le <date type="numeric"><day>17</day>/<month>09</month>/<year>2015</year></date>
```

```
le <date type="numeric"><day>24</day>/<month>03</month>/<year>2014</year></date>
```

### Relatives

Il s'agit des dates composées uniquement d'un jour de la semaine et d'un chiffre. Il s'agit du seul cas où le jour de la semaine est inclus dans la structure <date>. Exemples :

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<date type="rel"><weekday>Lundi</weekday> <day>12</day></date>
```

```
<date type="rel"><weekday>mercredi</weekday> <day>24</day></date>
```

## Résumé

Balise globale :

- <date type="spelled1">
- <date type="spelled2">
- <date type="numeric">
- <date type="rel">

Balises inférieures :

- <weekday>
- <day>
- <month>
- <year>

## Lieux

Seules les villes sont annotées au sein de la balise <settlement>. Exemples :

```
<?xml version="1.0" encoding="UTF-8"?>
```

Je suis née à<settlement type="city">Toulouse</settlement>.

Il s'est rendu au Super U de <settlement type="city">Pontoise</settlement> pour faire des courses.



## **Annexe B**

# **Graphes Unitex**

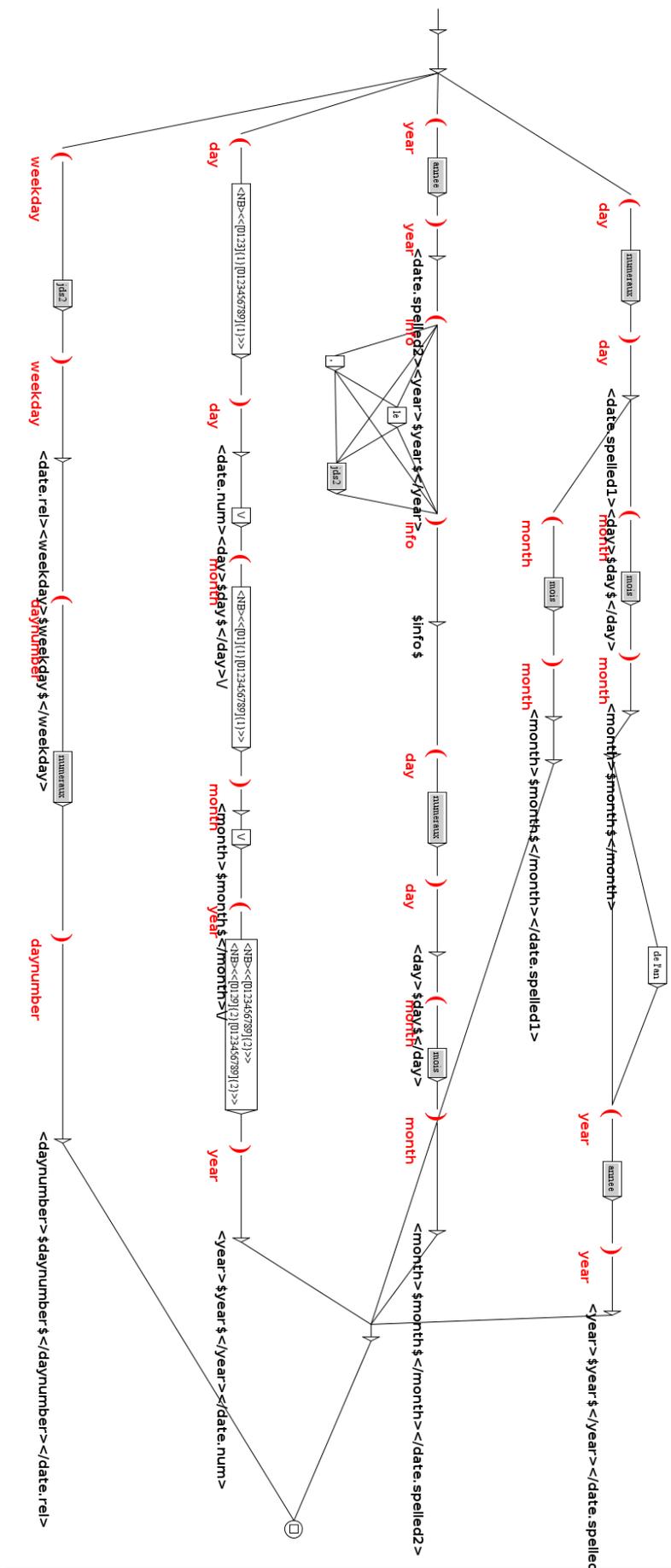


FIGURE 45 – Graphe global des dates

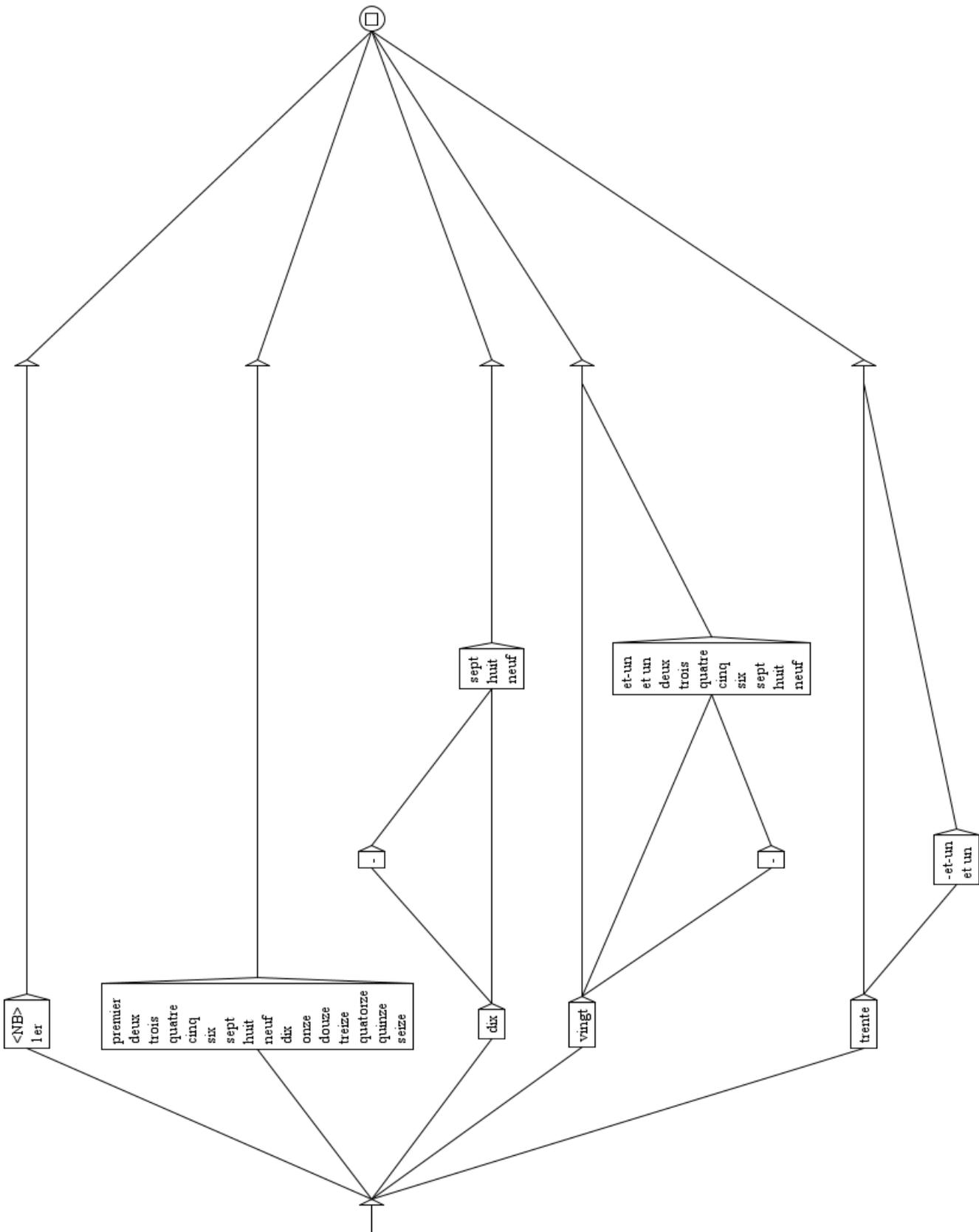


FIGURE 46 – Sous-graphes des numéros de jour

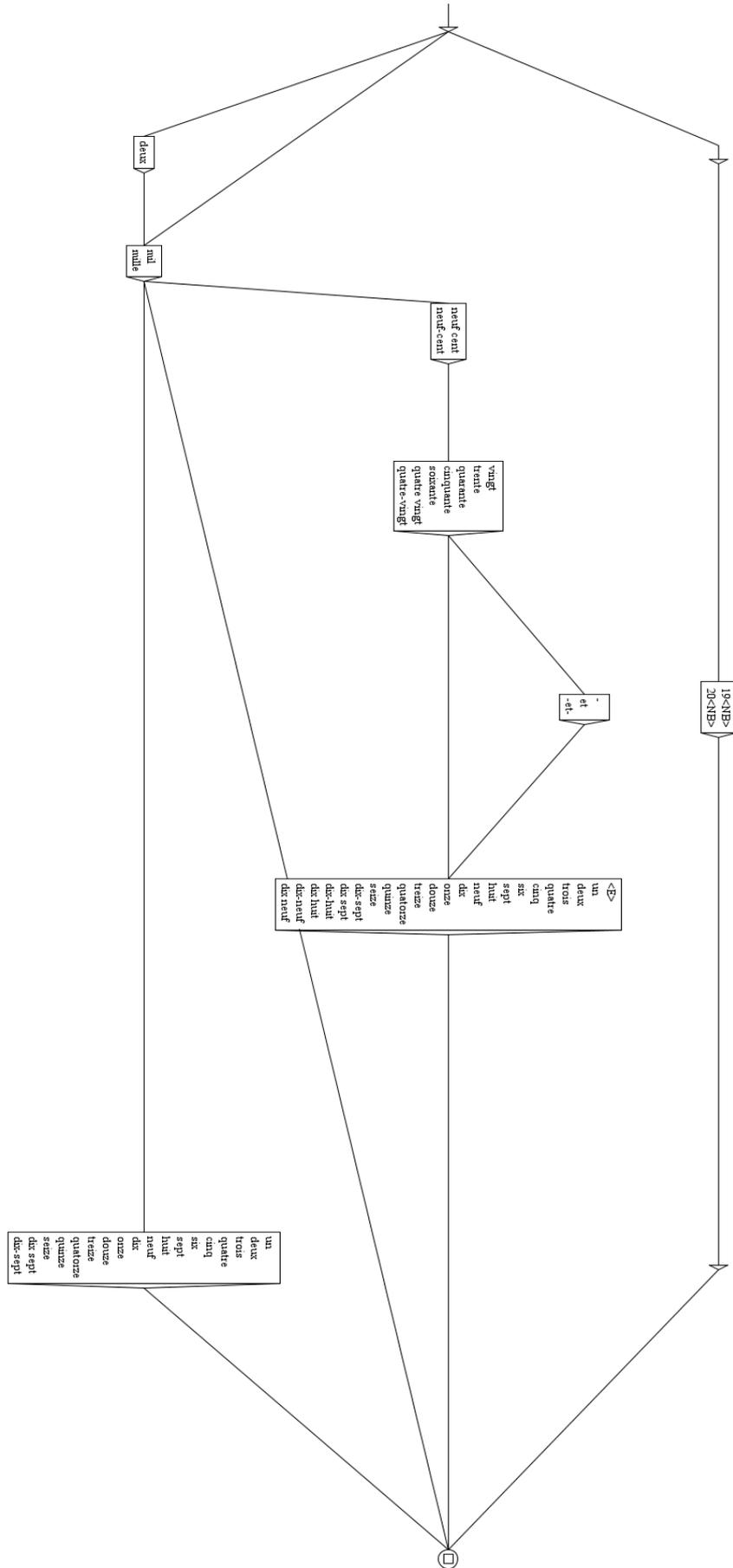


FIGURE 47 – Sous-graphe des années

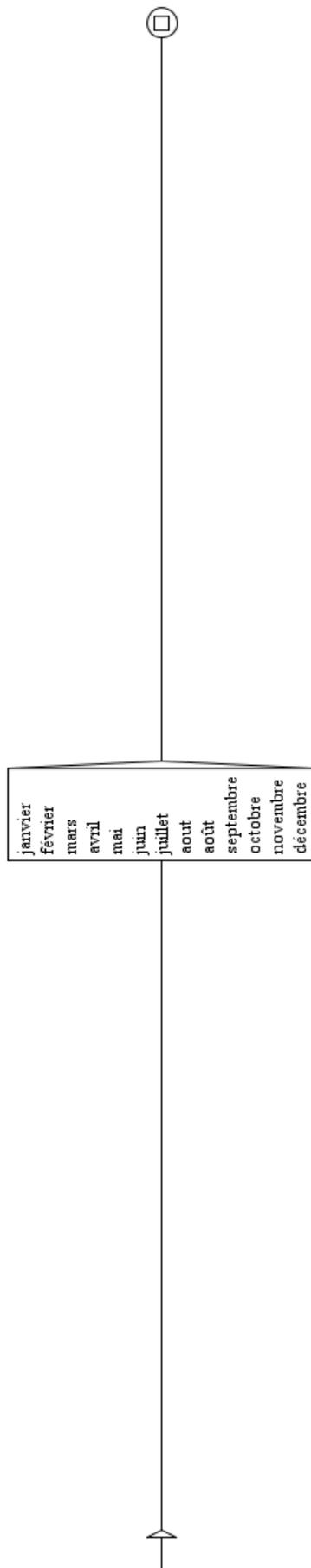


FIGURE 48 – Sous-graphe des mois

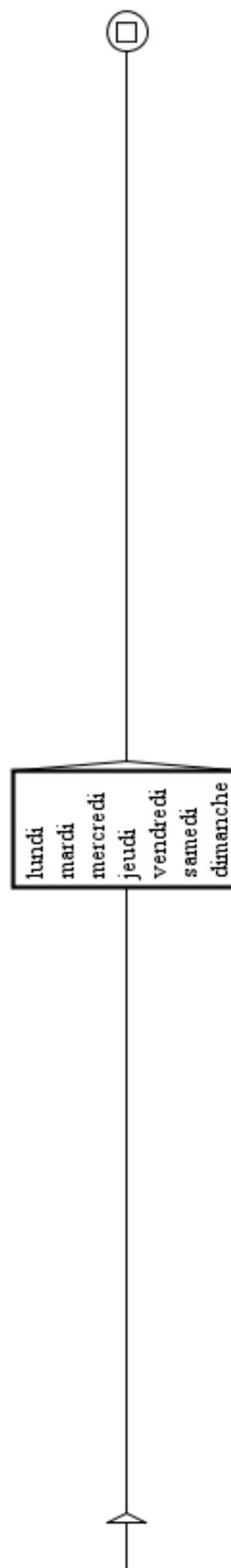


FIGURE 49 – Sous-graphe des jours de semaine

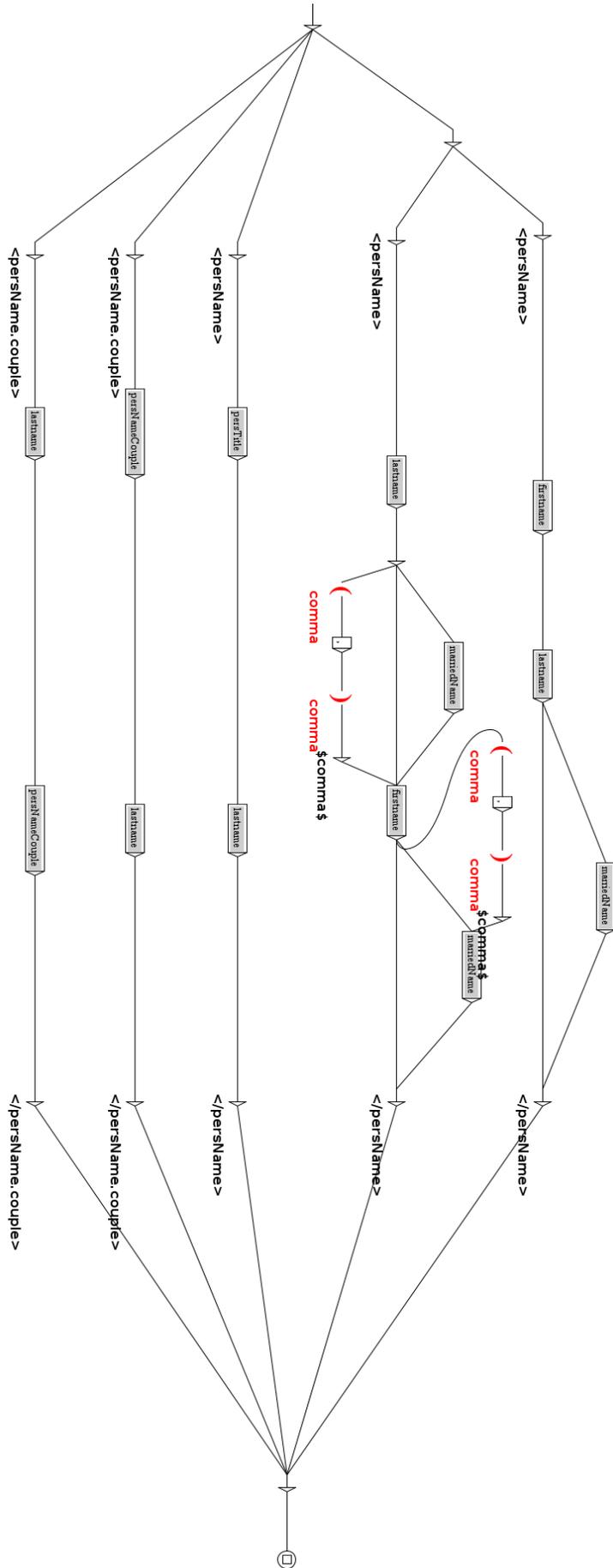


FIGURE 50 – Graphe général des noms de personnes

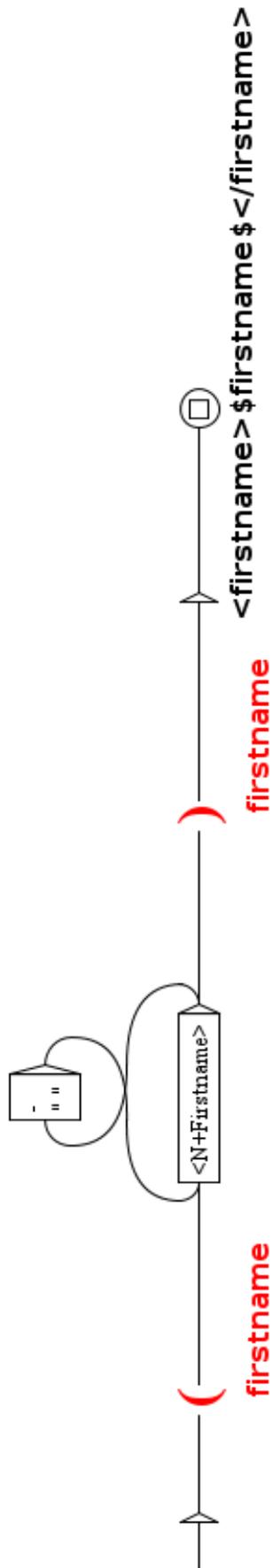


FIGURE 51 – Sous-graphe des prénoms

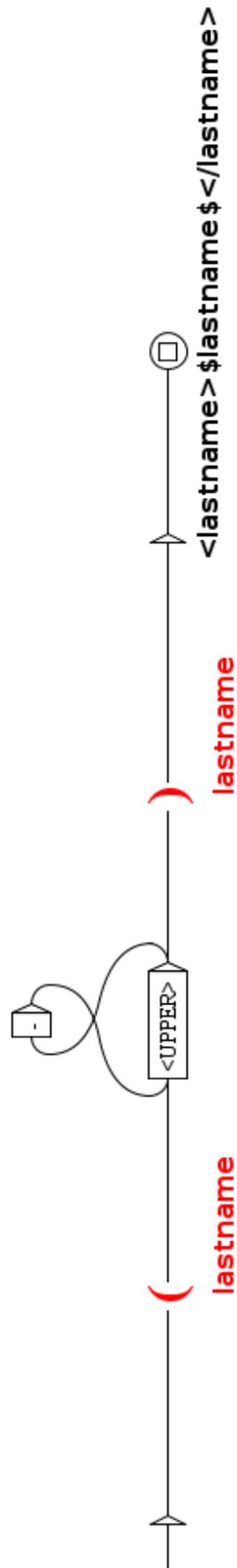


FIGURE 52 – Sous-graphe des noms de famille

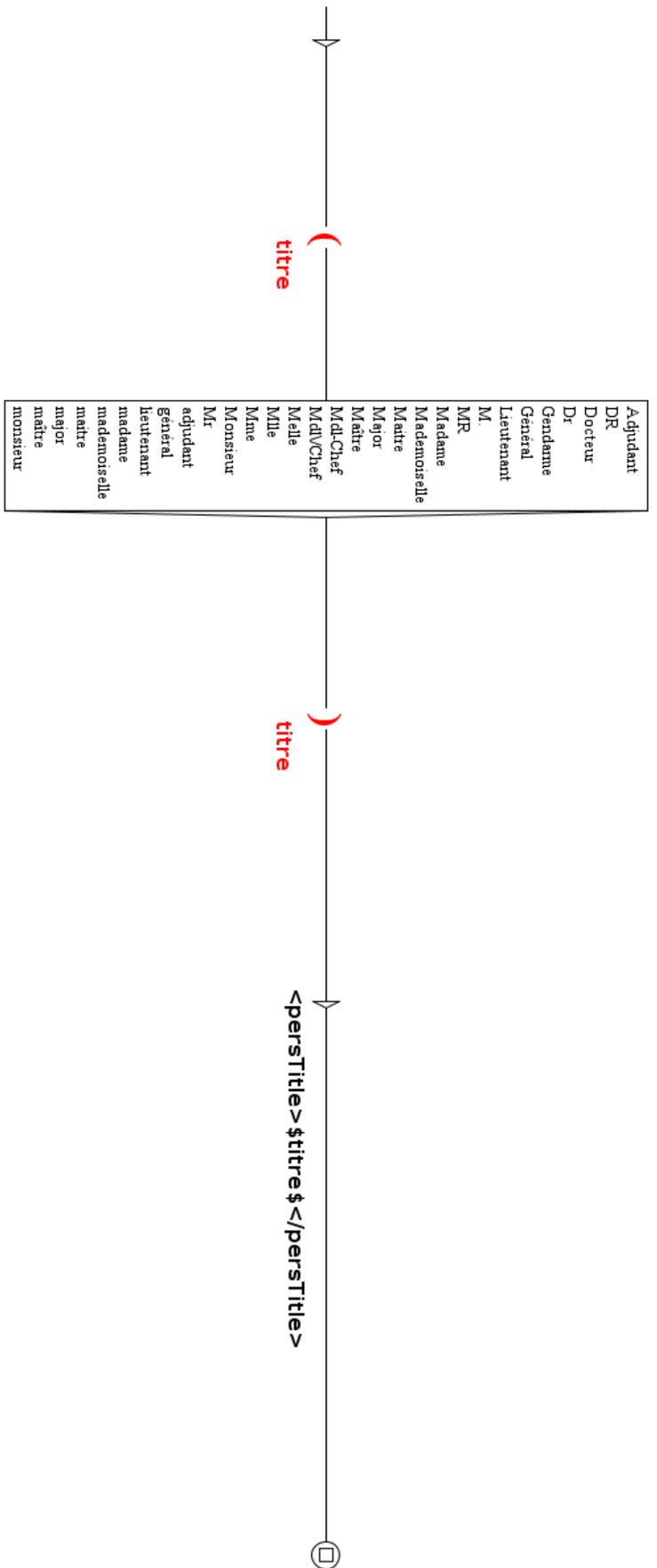


FIGURE 53 – Sous-graphe des titres de civilité

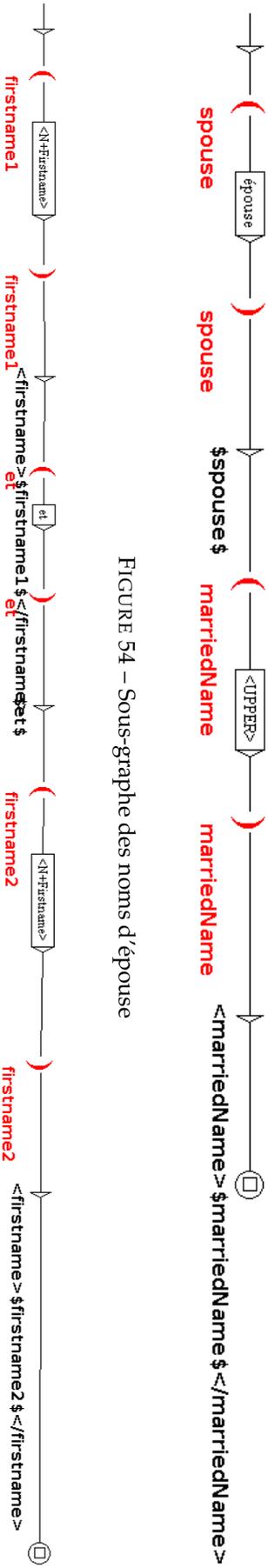


FIGURE 54 – Sous-graphe des noms d'épouse

FIGURE 55 – Sous-graphe des noms de couples

## Annexe C

# Tableau de contingence des parties du discours

Unités	Fréquence T2828392	a=577418	bc=15013	codes=538008	jo=579083	ja=523670	jc=595200
ABR	22891	5455	4	4104	8026	3002	2300
ADJ	150558	14571	951	32252	38713	27685	36386
ADV	80211	30160	340	12683	5803	13580	17645
DET:ART	212811	34277	1962	64686	32920	35103	43863
DET:POS	26430	7882	164	4333	2386	5247	6418
INT	1183	1180	0	0	2	0	1
KON	108132	22390	649	26131	14834	17868	26260
NAM	181954	45347	329	7278	59989	41871	27140
NOM	642007	93778	3693	135372	137858	122141	149165
NUM	135484	18731	445	17739	50266	24291	24012
PRO	74	53	0	2	0	6	13
PRO:DEM	26140	8708	121	4398	1847	4871	6195
PRO:IND	10278	2428	125	2533	1228	1980	1984
PRO:PER	81442	55425	254	10300	2037	6874	6552
PRO:POS	82	47	1	5	4	6	19
PRO:REL	25685	8463	132	5859	2249	3869	5113
PRP	345454	59057	2097	77677	61001	68134	77488
PRP:det	111914	11170	708	24360	28259	21302	26115
PUN	244088	38603	711	29440	68977	51555	54802
PUN:cit	11165	1424	4	861	3683	1406	3787
SENT	69474	31209	618	12770	14659	7423	2795
SYM	1472	12	0	32	722	410	296
VER:cond	3460	1172	6	283	138	867	994
VER:futu	7063	535	9	1110	2331	1888	1190
VER:impe	2	0	0	2	0	0	0
VER:impf	23374	12618	3	454	398	3704	6197
VER:infi	49314	10758	288	10856	5298	10071	12043
VER:ipper	123556	22338	582	25924	20093	25669	28950
VER:ppre	19703	2063	70	3148	3292	5155	5975
VER:pres	104492	36676	732	22467	10548	15246	18823
VER:simp	3000	321	0	246	612	983	838
VER:subi	765	76	0	25	258	227	179
VER:subp	4734	491	15	678	652	1236	1662

TABLE 11 – Tableau de contingences des parties du discours sur le corpus utilisé pour la caractérisation du genre judiciaire

# Glossaire

**acte d'enquête** Terme général désignant les techniques utilisées pour recueillir les éléments informationnels dans le cadre de l'enquête. [15](#), [16](#), [35](#), [50](#), [159](#), [182](#)

**ADN** Macromolécule biologique présente dans toutes les cellules et porteuse d'information génétique identifiante. La recherche de traces d'ADN et la gestion de bases de données de profils ADN sont des atouts importants de l'enquête. Voir RIBAUX (2014, p. 110). [3](#), [8](#), [38](#), [64](#)

**ADT** Analyse des données textuelles. Discipline des sciences sociales qui étudie les textes via des méthodes qualitatives et quantitatives. [45](#), [58](#), [61](#), [77](#)

**AFC** Analyse factorielle des correspondances. Méthode statistique d'analyse multivariée développée par BENZÉCRI (1973), basée sur un tableau de contingence. Dans le cadre de la linguistique de [corpus](#), l'AFC permet d'étudier les proximités et les divergences entre corpus. [45](#), [146](#), [148](#), [153](#)

**analyse criminelle** Discipline qui cherche à structurer l'information récoltée à propos de faits criminels et délictueux pour en faciliter la compréhension, dans un objectif d'action de sécurité ou de [police judiciaire](#). [1](#), [4–8](#), [10](#), [13–20](#), [22–25](#), [28](#), [35](#), [44](#), [45](#), [48](#), [50–54](#), [56](#), [57](#), [62–64](#), [72–76](#), [78–80](#), [94](#), [101](#), [109](#), [110](#), [120](#), [122](#), [131](#), [159](#), [160](#), [163](#), [181](#), [193](#)

**Analyst's Notebook** Logiciel d'analyse de données édité par la société IBM employé en [analyse criminelle](#) pour réaliser des schémas. [13](#)

**analyste criminel** Personnel formé et pratiquant l'[analyse criminelle](#). [1](#), [6](#), [7](#), [10](#), [13–25](#), [27](#), [45](#), [51](#), [53](#), [54](#), [61](#), [62](#), [64](#), [70](#), [72](#), [75–77](#), [79](#), [103](#), [112](#), [120](#), [122](#), [159](#), [160](#), [183](#)

**audition** Technique d'enquête consistant à entendre et recueillir sous forme de [procès-verbal](#) les propos d'une personne. Les auditions peuvent libres ou

- sous contrainte (dans le cadre d'une [garde à vue](#) ou d'une [mise en examen](#)).  
3, 9, 16, 21, 31, 33–35, 43, 44, 50–53, 62, 63, 72, 74, 81–83, 87, 88, 92, 94, 98, 100–103, 105, 111, 112, 120–123, 125–134, 137, 138, 140, 145, 148, 151, 152, 155–157, 160, 161, 187
- autopsie** Examen médical d'un cadavre réalisé par un [médecin légiste](#) en vue de déterminer les causes de la mort. 22, 38, 43
- CNFPJ** Centre National de Formation de la Police Judiciaire. Centre situé à Rosny-sous-Bois, dédié à la formation des officiers et sous-officiers de la [Gendarmerie nationale](#) ayant des activités de [police judiciaire](#). 17
- CNIL** Commission Nationale de l'Informatique et des Libertés. 17
- Code de procédure pénale** Ensemble des textes de lois régissant la conduite de la procédure pénale, consultable sur le site LégiFrance : <https://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006071154>. 2, 15, 16, 26, 27, 31, 33, 43, 48, 130, 131, 185, 186
- cold case** Le terme *cold case* désigne une affaire ancienne et non résolue. Ce terme popularisé par les séries télévisées américaines n'est toutefois pas complètement adapté à la situation française, dans laquelle tant qu'un dossier n'a pas été clos par un non-lieu il n'est jamais totalement inactif. Voir : Rencontre avec Me Corinne Herrmann, avocat des *cold cases*, *Droit Pénal* n°3, LexisNexis, mars 2019 <https://www.seban-associés.avocat.fr/wp-content/uploads/2019/03/Interview-Corinne-Herrmann.pdf> [consulté le 09 mai 2019]. 15, 20, 21, 26
- commission rogatoire** Acte juridique par lequel un juge charge un autre juge ou des services de police de réaliser des [actes d'enquête](#) à sa place. 16, 27, 33, 184
- corpus** Ensemble de textes rassemblés dans un objectif en particulier. Le corpus est une ressource de travail pour le [TAL](#), tandis que pour la linguistique il s'agit d'un objet d'étude dont la définition et les modalités de constitution font débat. 44–49, 51, 53, 55, 57–62, 66, 68, 69, 72, 74, 76, 77, 81, 94, 103–105, 108, 109, 117, 118, 121–123, 126, 128, 132–134, 138, 141, 142, 145, 146, 148, 153, 154, 160, 161, 181, 187, 189, 190, 193

**description définie** Structure référentielle conceptualisée par KLEIBER (1981), du type « le + nom » : « la mère de Pierre », « le voisin de Sylvain ». 63, 66, 67, 85, 89, 90, 92, 101, 102, 121, 161, 183

**DSAC** Département Sciences de l'Analyse Criminelle. Équipe d'une dizaine d'analystes criminels à la compétence nationale, rattachée au SCRC. 14, 18, 20, 21, 25

**enquête** Processus de collecte, d'organisation et d'interprétation d'éléments matériels (*traces*) et informationnels afin de résoudre des faits répréhensibles. 1–8, 15–17, 19, 21, 22, 27–29, 32–35, 43, 44, 49, 50, 64, 72–76, 83, 85, 92, 93, 98, 101, 112, 121, 125–129, 134, 145, 154, 159, 183

**enquête de flagrance** *Enquête* mise en place dans le cas de la constatation d'un *flagrant-délit*, c'est-à-dire la commission d'un délit ou crime en train de se commettre. Dans le cas de l'enquête de flagrance, les enquêteurs disposent de tous les droits de coercition et rendent compte de l'avancée de l'enquête au *procureur de la République*. 15, 16, 26, 33, 35

**enquête préliminaire** *Enquête* mise en place à l'initiative des services de *police judiciaire* ou du *procureur de la République* avant l'ouverture éventuelle d'une *information judiciaire*. 15, 26, 33

**entité criminelle textuelle** Mention textuelle d'une entité criminelle, sous forme de *description définie*, d'*entité nommée*, ou de description indéfinie. 81, 101–103, 105, 120, 160, 161

**entité nommée** Structure linguistique référant à un concept du monde réel, qu'il s'agisse de personnes, de lieux, de dates, de montants, etc. Les entités nommées sont souvent rattachées aux noms propres et aux *descriptions définies*, et peuvent être spécifiques à un domaine en particulier (par exemple, les noms de médicaments dans le domaine biomédical). 7, 53, 54, 63–68, 71–75, 78–80, 87, 88, 92, 93, 101–105, 109, 110, 112, 121, 141, 142, 160, 161, 183

**Europol** Agence européenne de police criminelle, basée à La Haye (Pays-Bas).

**extraction d'information** Sous-discipline du [TAL](#) qui cherche à extraire automatiquement des informations de texte en langue naturelle. [24](#), [45](#), [65](#), [72](#), [73](#), [75](#), [77–79](#), [120](#), [121](#), [123](#), [159–161](#), [163](#)

**F-mesure** Moyenne harmonique du [rappel](#) et de la [précision](#). [69](#), [70](#), [117–119](#), [161](#)

**flagrant-délit** Délit ou crime qui est en train ou qui vient d'être commis. [15](#), [183](#)

**garde à vue** Mesure de privation de liberté prise à l'encontre d'un suspect lors d'une enquête. Elle permet aux enquêteurs d'avoir le suspect à leur disposition pour pouvoir l'interroger et vérifier la véracité de ses déclarations. La durée de la garde à vue est limitée et le suspect a des droits liés à sa situation, dont celui d'être assisté par un avocat. [29](#), [34](#), [43](#), [50](#), [127](#), [182](#)

**Gendarmerie nationale** Force militaire qui partage avec la Police nationale le pouvoir de police en France. [1](#), [2](#), [10](#), [14](#), [15](#), [18](#), [28](#), [31](#), [111](#), [126](#), [127](#), [159](#), [182](#), [186](#)

**information judiciaire** Aussi appelée « instruction », l'information judiciaire est le processus de renseignement visant des faits de nature criminelle dirigé par le [juge d'instruction](#). [15](#), [16](#), [183](#), [184](#)

**instruction** Voir : [information judiciaire](#). [121](#)

**IRCGN** Institut de Recherche Criminelle de la Gendarmerie Nationale. Institut formant le Pôle judiciaire de la Gendarmerie nationale ([PJGN](#)) avec le [SCRC](#), situé à Pontoise depuis 2015. [14](#), [185](#), [187](#)

**juge d'instruction** Magistrat chargé de diriger les investigations dans le cadre de la [commission rogatoire](#). [16](#), [27](#), [33](#), [184](#)

**lemme** Forme canonique d'un mot (forme « du dictionnaire »). [58](#)

**mise en examen** Mesure de privation de liberté (contrôle judiciaire ou détention provisoire) ordonnée par le juge d'instruction sur la base d'indices forts et concordants laissant supputer l'implication comme auteur ou comme complice de la personne dans les faits. [33](#), [182](#)

**médecin légiste** Médecin chargé d'expertises médico-légales dont les conclusions sont des éléments de la procédure. [32](#), [182](#)

**OCR** *Optical Character Recognition*, ou reconnaissance optique de caractères en français. Technologie qui permet, en partant de fichiers image, de reconnaître les caractères qui y figurent et donc d'en extraire le texte dans un format exploitable informatiquement. [20](#), [21](#), [103](#)

**OPJ** Officier de police judiciaire. Dans le cas des services de police, il s'agit de fonctionnaires d'un certain grade dont les attributions comprennent de constater les infractions pénales et de mener certaines enquêtes. Voir : article 16 et suivants du [Code de procédure pénale](#) <https://www.legifrance.gouv.fr/affichCode.do?idSectionTA=LEGISCTA000006167412&cidTexte=LEGITEXT000006071154> [consulté le 16 décembre 2019]. [17](#), [33](#), [34](#), [92](#)

**partie du discours** Catégorie grammaticale d'un mot dans un texte. [58](#), [104](#), [138](#), [140](#), [145](#), [146](#), [148](#), [149](#), [151–154](#), [191](#)

**PDF** *Portable Document Format*. Format informatique de fichiers qui préserve la mise en page du document quel que soit la machine ou le logiciel utilisé pour l'ouvrir. [20](#), [21](#), [103](#), [111](#)

**perquisition** Technique d'enquête recherchant des preuves ou des éléments matériels dans un lieu privé (domicile, bureaux, véhicule, etc.). [35](#), [98](#)

**PJGN** Pôle Judiciaire de la Gendarmerie Nationale. Centre de gendarmerie consacré à la police judiciaire. Situé à Pontoise (95), il est formé par l'[IRCGN](#) qui traite les aspects relatifs à la police scientifique, le [SCRC](#), pour l'aspect renseignement, le Centre de lutte Contre les Criminalités Numériques (C3N), et l'observatoire central des systèmes de transports intelligents (OCSTI). [1](#), [14](#), [25](#), [64](#), [126](#), [145](#), [184](#), [187](#)

**PNIJ** Plate-forme Nationale des Interceptions Judiciaires. Structure permettant la communication entre enquêteurs et fournisseurs de services de télécommunications. [41](#)

**police judiciaire** Définie par les articles 12 et suivants du [Code de procédure pénale](#), la police judiciaire est une mission exercée en France par la [Gendarmerie nationale](#) et la Police nationale qui vise à la constatation des infractions pénales, d'en rassembler les preuves et d'en rechercher les auteurs. . [1](#), [2](#), [5](#), [14-16](#), [159](#), [181-183](#)

**portrait-robot** Technique d'enquête qui reconstitue les traits d'un individu d'après la description d'un [témoin](#). Autrefois pratiquée par des dessinateurs, des logiciels dédiés existent aujourd'hui. [43](#), [44](#)

**procureur de la République** Magistrat à la tête du parquet auprès duquel sont signalés les faits criminels constatés. Lors de l'enquête préliminaire et de l'enquête de flagrance, les enquêteurs rendent compte de leur travail au procureur. [15](#), [183](#)

**procès-verbal** Document produit par les enquêteurs rendant compte de mesures et d'actes relatifs à l'enquête. [3](#), [8](#), [10](#), [18](#), [29](#), [31](#), [32](#), [34](#), [35](#), [42](#), [48](#), [50](#), [112](#), [126-131](#), [133](#), [134](#), [140](#), [145](#), [155](#), [157](#), [161](#), [181](#), [190](#)

**procédure** 1) Ensemble des actions menées dans le cadre d'une action en justice.  
2) Le dossier rassemblant les pièces liées à la dite action en justice. [5](#), [7](#), [8](#), [13](#), [15-22](#), [24-29](#), [34](#), [35](#), [38](#), [42-45](#), [47-51](#), [55](#), [64](#), [72-77](#), [79](#), [80](#), [82](#), [86](#), [87](#), [90](#), [95](#), [102](#), [103](#), [112](#), [121-123](#), [125](#), [127](#), [134](#), [140](#), [143](#), [154-156](#), [159](#), [160](#), [162](#), [163](#), [189](#), [193](#)

**précision** Mesure d'évaluation de la pertinence des éléments reconnus dans une tâche de détection. [69](#), [70](#), [117-119](#), [161](#), [184](#)

**rappel** Mesure d'évaluation appréciant le nombre d'éléments pertinents reconnus par rapport au nombre total d'éléments. [69](#), [70](#), [117-119](#), [161](#), [184](#)

**réquisition** Ordre du magistrat ou des enquêteurs adressés à une personne, un expert ou une entreprise de fournir des services qui nécessitent une intervention tierce à l'enquête. [22](#), [29](#), [30](#), [38](#), [40](#), [41](#), [50](#), [189](#)

**scellé** Élément matériel potentiellement utile à l'enquête placé sous plomb afin de garantir sa préservation. [3](#), [28](#), [29](#), [38](#), [43](#)

**SCRC** Service Central de Renseignement Criminel. Ce centre est l'un des deux organismes qui forme avec l'IRCGN le PJGN. Sa mission concerne les aspects relatifs au renseignement dans le judiciaire géré par la Gendarmerie. 14, 18, 80, 183–185

**segment répété** Suite de **tokens** non séparés par des signes de ponctuation apparaissant au moins deux fois dans le **corpus** (LAFON et SALEM, 1983)!. 59, 133, 134, 137, 138, 140

**TAL** Traitement automatique des langues. Discipline de recherche alliant sciences du langage et informatique dans l'objectif de traiter automatiquement le langage dit « naturel » (par contraste avec les langages de programmation informatique). Le TAL est à l'origine de nombreuses technologies courantes : moteurs de recherche, traduction automatique, correction orthographique, transcription automatique, etc. 7, 45, 49, 51, 53, 63–67, 72, 73, 75–77, 79, 102, 106, 123, 125, 140, 142, 160, 162, 163, 182, 184

**tapissage** Le tapissage, ou parade d'identification, est une technique d'enquête qui consiste à placer une personne au sein d'un groupe d'autres personnes pour être identifiée par un ou des **témoins** placé(s) hors de la vue du groupe (par exemple, derrière une glace sans tain). 35, 43

**textométrie** Pratique de linguistique reposant sur un principe différentiel permettant de comparer des ensembles de textes à l'aide de logiciels dédiés. 55, 58, 61, 62

**TIC** Technicien en identification criminelle. Gendarmes amenés à intervenir sur les lieux de crimes et délits pour opérer les constatations, rechercher des indices et effectuer les prélèvements nécessaires. 31

**token** Suite de caractères comprise entre deux espaces ou un espace et un signe de ponctuation considérée comme un mot. 133, 134, 153, 187

**trace** Vestige matériel d'une activité. 3, 4, 38, 63, 64, 94, 183

**témoin** Personne détenant des informations relatives aux faits examinés dans l'enquête et qui les confie dans le cadre d'une **audition** plural. 3, 6, 8, 9, 15, 22, 31, 33–35, 43, 50, 52, 62, 63, 71, 74, 81–87, 89, 92, 96–98, 100–103, 120, 121, 123, 125–127, 129–134, 137, 138, 140, 145, 151–153, 155–157, 160, 161, 186, 187



# Table des figures

1	Modélisation des formes de raisonnement élémentaires issue de BARLATIER (2017, p.15) . . . . .	4
2	Exemple de schéma événementiel confrontant plusieurs témoignages et les faits. . . . .	11
3	Exemple de schéma relationnel simple . . . . .	12
4	Un dossier de procédure relatif à une disparition au format papier . . . . .	26
5	Exemple de réquisitions d'expert . . . . .	30
6	Exemple d'audition de témoin . . . . .	36
7	Procès-verbal de renseignement . . . . .	37
8	Extrait de rapport d'analyse présentant une technique . . . . .	39
9	Extrait de rapport d'analyse d'organe . . . . .	39
10	Extrait de rapport d'autopsie . . . . .	40
11	Extrait de conclusion de rapport d'autopsie . . . . .	41
12	Extrait de facture téléphonique détaillée . . . . .	41
13	Extrait de relevé bancaire. La mention « RET DAB » indiquant un retrait au distributeur automatique a été surlignée par l'enquêteur (dernière ligne), suivie de la localité du retrait (camouflée). . . . .	42
14	Exemple de concordance issue du corpus des vœux présidentiels sur le portail TXM . . . . .	59
15	Exemple de cooccurrence pour le mot « année » dans le corpus des vœux présidentiels (TXM) . . . . .	60
16	Segments répétés tirés d' <i>Alice au pays des Merveilles</i> obtenus à l'aide de Lexico5 . . . . .	60

17	Spécificités du mois d'août dans un sous-corpus de quatre mois de l'Est Républicain . . . . .	60
18	Spécificités du mois de juillet dans un sous-corpus de quatre mois de l'Est Républicain . . . . .	60
19	Exemple de grammaire UNITEX . . . . .	106
20	Exemple de graphe utilisant des masques lexicaux . . . . .	107
21	Sorties du graphe de l'exemple 20 sur <i>Le tour du monde en 80 jours</i> de Jules Verne . . . . .	108
22	Exemple de graphe transducteur . . . . .	108
23	Sortie du transducteur de la figure 22 appliqué sur le corpus 80 jours. . . . .	109
24	Usage des tirets pour remplir les lignes dans les procès-verbaux. . . . .	112
25	État-civil au format tableau . . . . .	113
26	État-civil au format narratif . . . . .	113
27	Extrait du dictionnaire des communes . . . . .	114
28	Exemple de graphe détectant les années en chiffres et en toutes lettres. . . . .	115
29	Extrait du dictionnaire des prénoms . . . . .	116
30	Graphe général décrivant les noms de personnes purgé des boîtes de balisage dans un souci de lisibilité . . . . .	117
31	Exemple de l'emploi des pronoms personnels . . . . .	132
32	Exemple de mention des enquêteurs . . . . .	132
33	Exemple de mention des enquêteurs. Ici l'audition est prise en charge par une gendarme. . . . .	132
34	Emploi de la première personne du pluriel dans une phrase réglementaire qui ouvre l'audition . . . . .	133
35	Emploi de la première personne du singulier dans une phrase réglementaire qui termine l'audition . . . . .	133
36	Autre exemple d'ouverture d'une audition . . . . .	133
37	Extrait d'audition concernant un emploi du temps . . . . .	138
38	Autre extrait d'audition concernant un emploi du temps . . . . .	138
39	Classification des discours en genres, schéma issu de MALRIEU et RASTIER (2001). . . . .	144

---

40	Tableau de contingence . . . . .	147
41	Graphe d'AFC d'après les données du tableau 40. . . . .	147
42	Analyse factorielle des correspondances sur les parties du discours . .	149
43	Rapprochements des sous-corpus et parties du discours caractéristiques	149
44	Proposition d'organisation du discours légal en genres . . . . .	156
45	Graphe global des dates . . . . .	172
46	Sous-graphe des numéros de jour . . . . .	173
47	Sous-graphe des années . . . . .	174
48	Sous-graphe des mois . . . . .	175
49	Sous-graphe des jours de semaine . . . . .	175
50	Graphe général des noms de personnes . . . . .	176
51	Sous-graphe des prénoms . . . . .	177
52	Sous-graphe des noms de famille . . . . .	177
53	Sous-graphe des titres de civilité . . . . .	178
54	Sous-graphe des noms d'épouse . . . . .	178
55	Sous-graphe des noms de couples . . . . .	178



# Liste des tableaux

1	Récapitulatif des documents rencontrés en procédure traitées en analyse criminelle. . . . .	28
2	Exemples de précision, rappel, F-mesure . . . . .	70
3	Performances pour les dates . . . . .	117
4	Performances pour les noms de personnes . . . . .	118
5	Performances pour les lieux . . . . .	118
6	Erreurs d'étiquetage des lieux . . . . .	119
7	Performances pour les villes après modification de la ressource lexicale.	119
8	Segments répétés de onze mots et de fréquence supérieure à 10 . . . . .	137
9	Étiquettes morpho-syntaxiques et leurs fréquences . . . . .	139
10	Statistiques du corpus (valeurs approximatives) . . . . .	146



# Bibliographie

- ARULANANDAM, Remy, Bastin Tony Roy SAVARIMUTHU et Maryam A. PURVIS (2014). « Extracting Crime Information from Online Newspaper Articles ». In : *Proceedings of the Second Australasian Web Conference - Volume 155*. Australian Computer Society, Inc., p. 31-38. ISBN : 978-1-921770-37-1. URL : <http://dl.acm.org/citation.cfm?id=2667702.2667706> (visité le 13/05/2019).
- ATILF (2018). *Corpus journalistique issu de l'Est Républicain*. ORTOLANG (Open Resources and TOols for LANGuage). URL : [https://www.ortolang.fr/market/corpora/est\\_republicain](https://www.ortolang.fr/market/corpora/est_republicain) (visité le 19/12/2019).
- BARLATIER, Jérôme (2017). « Management de l'enquête et ingénierie judiciaire, recherche relative à l'évaluation des processus d'investigation criminelle ». Thèse de doct. Université de Lausanne. DOI : [10.13140/RG.2.2.31577.42089](https://doi.org/10.13140/RG.2.2.31577.42089). URL : [https://serval.unil.ch/notice/serval:BIB\\_EFDAD4B7FB68](https://serval.unil.ch/notice/serval:BIB_EFDAD4B7FB68) (visité le 17/09/2019).
- BENZÉCRI, Jean-Paul (1973). *L'analyse des correspondances*. Dunod. Paris.
- BERRA, Aurélien (2015). « Pour une histoire des humanités numériques ». In : *Critique* 819-820.8, p. 613-626. ISSN : 0011-1600. URL : <http://www.cairn.info/revue-critique-2015-8-page-613.htm> (visité le 26/09/2019).
- BERRA, Aurélien et al. (2017). « Scaling up Arts and Humanities : The DARIAH Approach to Data and Services ». In : *DH 2017*. Montréal, Canada. URL : <https://hal.archives-ouvertes.fr/hal-02152986> (visité le 26/09/2019).
- BIBER, Douglas (1988). *Variation Across Speech and Writing*. Cambridge University Press. ISBN : 978-0-521-42556-8.
- BLANCHE-BENVENISTE, Claire (2000). « Transcription de l'oral et morphologie ». In : *Romania una et diversa. Philologische Studien für Theodor Berchem zum 65. Geburtstag*. Sous la dir. de M. GILLE et R. KIESLER. T. 1. Tübingen : Gunter Narr Verlag. ISBN : 978-3-8233-5211-2.

- BLANCHET, Alain et al. (2013). « Évaluation de la méthode Processus général de recueil des entretiens, auditions et interrogatoires ». In : *Revue de la Gendarmerie Nationale*, p. 116-121.
- BLASCO, Mylène et Paul CAPPEAU (2012). « Identifier et caractériser un genre : l'exemple des interviews politiques ». In : *Langages* n° 187.3, p. 27-40. ISSN : 0458-726X. URL : <https://www.cairn.info/revue-langages-2012-3-page-27.htm> (visité le 23/05/2019).
- BOMMIER-PINCEMIN, Bénédicte (1999). « Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents ». Thèse de doct. Université Paris IV. URL : [http://www.revue-texto.net/1996-2007/Inedits/Pincemin/Pincemin\\_these.html](http://www.revue-texto.net/1996-2007/Inedits/Pincemin/Pincemin_these.html) (visité le 01/07/2019).
- BOYER, Henri (2008). « Fonctionnements sociolinguistiques de la dénomination toponymique ». In : *Mots. Les langages du politique* 86, p. 9-21. ISSN : 0243-6450. DOI : 10.4000/mots.12962. URL : <http://journals.openedition.org/mots/12962> (visité le 28/10/2019).
- BRANCA-ROSOFF, Sonia (1999). « Des innovations et des fonctionnements de langue rapportés à des genres ». In : *Langage & société* 87.1, p. 115-129. DOI : 10.3406/lsoc.1999.2856. URL : [https://www.persee.fr/doc/lsoc\\_0181-4095\\_1999\\_num\\_87\\_1\\_2856](https://www.persee.fr/doc/lsoc_0181-4095_1999_num_87_1_2856) (visité le 26/03/2019).
- BRUNEL, Maité, Jacques PY et Céline LAUNAY (2013). « Cost and benefit of a new instruction for the cognitive interview : the open depth instruction ». In : *Psychology, Crime & Law* 19.10, p. 845-863. ISSN : 1068-316X. DOI : 10.1080/1068316X.2012.684058. URL : <https://doi.org/10.1080/1068316X.2012.684058> (visité le 20/05/2019).
- CADIOT, Pierre (1991). « A la hache ou avec la hache ? Représentation mentale, expérience située et donation du référent ». In : *Langue française* 91.1, p. 7-23. DOI : 10.3406/lfr.1991.6202. URL : [http://www.persee.fr/doc/lfr\\_0023-8368\\_1991\\_num\\_91\\_1\\_6202](http://www.persee.fr/doc/lfr_0023-8368_1991_num_91_1_6202) (visité le 18/12/2019).
- CARNAZ, Gonçalo et al. (2019). « An Automated System for Criminal Police Reports Analysis ». In : *Proceedings of the Tenth International Conference on Soft Computing and Pattern Recognition (SoCPaR 2018)*. Sous la dir. d'Ana Maria MADUREIRA et al.

- T. 942. Porto : Springer International Publishing, p. 360-369. DOI : [10.1007/978-3-030-17065-3\\_36](https://doi.org/10.1007/978-3-030-17065-3_36). URL : [http://link.springer.com/10.1007/978-3-030-17065-3\\_36](http://link.springer.com/10.1007/978-3-030-17065-3_36) (visité le 13/05/2019).
- CHAROLLES, Michel (2002). *La référence et les expressions référentielles en français*. Paris : Ophrys, p. 258. ISBN : 978-2-7080-1014-7.
- CHAU, Michael, Jennifer J XU et Hsinchun CHEN (2002). « Extracting Meaningful Entities from Police Narrative Reports ». In : *Proceedings of the 2002 Annual National Conference on Digital Government Research*, p. 1-5.
- CHEN, Peter Pin-Shan (1976). « The Entity-relationship Model — Toward a Unified View of Data ». In : *ACM Trans. Database Syst.* 1.1, p. 9-36. ISSN : 0362-5915. DOI : [10.1145/320434.320440](https://doi.org/10.1145/320434.320440). URL : <http://doi.acm.org/10.1145/320434.320440> (visité le 15/05/2019).
- CHINCHOR, Nancy et Patty ROBINSON (1998). « MUC-7 Named Entity Task Definition ». In : *Seventh Message Understanding Conference (MUC-7) : Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.
- CITTON, Yves (2015). « Humanités numériques. Une médiapolitique des savoirs encore à inventer ». In : *Multitudes* n° 59.2, p. 169-180. ISSN : 0292-0107. URL : <https://www.cairn.info/revue-multitudes-2015-2-page-169.html#> (visité le 17/05/2019).
- CONSTANT, Mathieu (2003). « Grammaires locales pour l'analyse automatique de textes : méthodes de construction et outils de gestion ». Thèse de doct. Université Paris-Est. URL : <https://tel.archives-ouvertes.fr/tel-00626252> (visité le 12/11/2019).
- CORNU, Gérard (2016). *Vocabulaire juridique*. 11<sup>e</sup> éd. Quadrige dicos poche. Paris : PUF. ISBN : 978-2-13-065205-2.
- CUSSON, Maurice et Gilbert CORDEAU (1994). « Le crime du point de vue de l'analyse stratégique ». In : *Traité de criminologie empirique*. Sous la dir. de Denis SZABO et Marc LEBLANC. Les Presses de l'Université de Montréal. Montréal, p. 91-112.
- D'HONDT, Sigurd (2019). « Humanity and Its Beneficiaries : Footing and Stance-Taking in an International Criminal Trial ». In : *Signs and Society* 7.3, p. 427-453. ISSN : 2326-4489. DOI : [10.1086/705279](https://doi.org/10.1086/705279). URL : <https://www.journals.uchicago.edu/doi/full/10.1086/705279> (visité le 03/12/2019).

- DACOS, Marin (2010). *Manifeste des Digital humanities*. Billet. URL : <https://tcp.hypotheses.org/318> (visité le 17/05/2019).
- DACOS, Marin et Pierre MOUNIER (2015). *Humanités numériques : État des lieux et positionnement de la recherche française dans le contexte international*. Rapport. Institut français, p. 89. URL : <https://hal.archives-ouvertes.fr/hal-01228945> (visité le 22/08/2019).
- DAMETTE, Éliane (2013). « Enseigner la traduction juridique : L'apport du français juridique, discipline passerelle entre droit, méthodologie juridique et linguistique ». In : *La traduction juridique : Points de vue didactiques et linguistiques*. Sous la dir. de Marion MEUNIER, Marion CHARRET-DEL BOVE et Éliane DAMETTE. Publications du CEL, p. 73-87.
- DIOT-PARVAZ AHMAD, Bénédicte (2019). « Production de ressources multilingues pour l'aide à la traduction du droit pénal en hindi, ourdou et français ». Thèse de doct. Paris : Institut National des Langues et Civilisation Orientales.
- DISTER, Anne et Anne-Catherine SIMON (2007). « La transcription synchronisée des corpus oraux. Un aller-retour entre théorie, méthodologie et traitement informatisé ». In : *Arena Romanistica* 1.1, p. 54-79. ISSN : 1890-4580. URL : <https://dial.uclouvain.be/pr/boreal/object/boreal:83571> (visité le 03/12/2019).
- DODDINGTON, George et al. (2004). « The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation ». In : *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal : European Language Resources Association (ELRA).
- DRIESEN, Christiane J. (2016). « L'interprétation juridique : surmonter une apparente complexité ». In : *Revue française de linguistique appliquée* Vol. XXI.1, p. 91-110. ISSN : 1386-1204. URL : <https://www.cairn.info/revue-francaise-de-linguistique-appliquee-2016-1-page-91.htm> (visité le 19/12/2019).
- DUBOIS, Danièle (1997). *Catégorisation et cognition : de la perception au discours*. Éditions Kimé. Paris. URL : <https://www.cairn.info/categorisation-et-cognition-de-la-perception-au-di-978284174101X.htm> (visité le 18/12/2019).
- EENSOO, Egle et Mathieu VALETTE (2015). « Une méthodologie de sémantique de corpus appliquée à des tâches de fouille d'opinion et d'analyse des sentiments :

- étude sur l'impact de marqueurs dialogiques et dialectiques dans l'expression de la subjectivité ». In : *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*. Caen : Association pour le Traitement Automatique des Langues, p. 107-118. URL : <https://hal-inalco.archives-ouvertes.fr/hal-01335116> (visité le 19/12/2019).
- EHRMANN, Maud (2008). « Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation ». Thèse de doct. Université Paris 7. URL : <https://hal.archives-ouvertes.fr/tel-01639190> (visité le 30/05/2019).
- EMANI, Cheikh Kafcah et Yannis HARALAMBOUS (2017). « Un système de questions-réponses dans le domaine légal : le cas des réglementations maritimes ». In : *Traitement Automatique des Langues* 58.2, p. 47-72. URL : <https://hal.archives-ouvertes.fr/hal-01814196> (visité le 06/12/2019).
- ERMINE, Jean-Louis, Mahmoud MORADI et Stéphane BRUNEL (2012). « Une chaîne de valeur de la connaissance ». In : *Management international* 16, p. 29. ISSN : 1206-1697, 1918-9222. DOI : [10.7202/1012391ar](https://doi.org/10.7202/1012391ar). URL : <http://id.erudit.org/iderudit/1012391ar> (visité le 21/06/2019).
- ESHKOL-TARAVELLA, Iris et Natalia GRABAR (sept. 2017). « Taxinomie dans les reformulations du point de vue de la linguistique de corpus ». In : *Syntaxe et sémantique* N° 18.1, p. 149-184. ISSN : 1623-6742. URL : <http://www.cairn.info/revue-syntaxe-et-semantique-2017-1-page-149.htm> (visité le 21/11/2019).
- ESHKOL-TARAVELLA, Iris et Anaïs LEFEUVRE-HALFTERMEYER (2017). « Linguistique de corpus : vues sur la constitution, l'analyse et l'outillage ». In : *Corela. Cognition, représentation, langage* HS-21. ISSN : 1638-5748. DOI : [10.4000/corela.4800](https://doi.org/10.4000/corela.4800). URL : <http://journals.openedition.org/corela/4800> (visité le 24/09/2019).
- FORT, Karen, Gilles ADDA et Kevin BRETONNEL COHEN (2011). « Amazon Mechanical Turk : Gold Mine or Coal Mine ? ». In : *Computational Linguistics*, p. 413-420. DOI : [10.1162/COLI\\_a\\_00057](https://doi.org/10.1162/COLI_a_00057). URL : <https://hal.archives-ouvertes.fr/hal-00569450> (visité le 11/10/2019).
- GALLIANO, Sylvain et al. (2006). « Corpus description of the ESTER evaluation campaign for the rich transcription of french broadcast news ». In : *In Proceedings of*

- the 5th international Conference on Language Resources and Evaluation (LREC 2006*, p. 315-320.
- GANASCIA, Jean-Gabriel (2015). « The Logic of the Big Data Turn in Digital Literary Studies ». In : *Frontiers in Digital Humanities 2*. ISSN : 2297-2668. DOI : [10.3389/fgdigh.2015.00007](https://doi.org/10.3389/fgdigh.2015.00007). URL : <https://www.frontiersin.org/articles/10.3389/fgdigh.2015.00007/full> (visité le 23/09/2019).
- GARRIC, Nathalie et Julien LONGHI (2012). « L'analyse de corpus face à l'hétérogénéité des données : d'une difficulté méthodologique à une nécessité épistémologique ». In : *Langages* n° 187.3, p. 3-11. ISSN : 0458-726X. URL : <https://www.cairn.info/revue-langages-2012-3-page-3.htm> (visité le 21/06/2019).
- GLEDHILL, Christopher, Stéphane PATIN et Maria ZIMINA (2017). « Lexicogrammaire et textométrie : identification et visualisation de schémas lexicogrammaticaux caractéristiques dans deux corpus juridiques comparables en français ». In : *Corpus* 17. ISSN : 1638-9808. URL : <http://journals.openedition.org/corpus/2868> (visité le 17/06/2019).
- GRISHMAN, Ralph (1997). « Information extraction : Techniques and challenges ». In : *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*. Sous la dir. de Maria Teresa PAZIENZA. Springer Berlin Heidelberg, p. 10-27. ISBN : 978-3-540-69548-6.
- GRISHMAN, Ralph et Beth SUNDHEIM (1996). « Message Understanding Conference-6 : A Brief History ». In : *COLING 1996 Volume 1 : The 16th International Conference on Computational Linguistics*. URL : <https://www.aclweb.org/anthology/C96-1079> (visité le 19/12/2019).
- GROUIN, Cyril (2013). « Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique ». Thèse de doct. URL : <https://tel.archives-ouvertes.fr/tel-00848672> (visité le 08/10/2019).
- HABERT, Benoît, Adeline NAZARENKO et André SALEM (1997). *Les linguistiques de corpus*. Paris : Armand Colin. URL : [http://lexicometrica.univ-paris3.fr/livre/les\\_linguistiques\\_de\\_corpus\\_1997/](http://lexicometrica.univ-paris3.fr/livre/les_linguistiques_de_corpus_1997/) (visité le 01/07/2019).

- HEIDEN, Serge, Jean-Philippe MAGUÉ et Bénédicte PINCEMIN (2010). « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement ». In : *JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data*. Rome : Edizioni Universitarie di Lettere Economia Diritto. URL : [http://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden\\_al\\_jadt2010.pdf](http://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf) (visité le 19/12/2019).
- HO ĐINH, Océane (2017). « Caractérisation différentielle de forums de discussion sur le VIH en vietnamien et en français ». Thèse de doct. Paris : Institut National des Langues et Civilisation Orientales. URL : <https://tel.archives-ouvertes.fr/tel-01964671> (visité le 19/12/2019).
- HUYGHE, Richard (2009). *Les noms généraux d'espace en français*. De Boeck Supérieur. ISBN : 978-2-8011-0042-4. DOI : 10.3917/dbu.huygh.2009.01. URL : <http://www.cairn.info/les-noms-generaux-d-espace-en-francais--9782801100424.htm> (visité le 25/10/2019).
- INSEE (2018). *Tableaux de l'économie française - Villes et communes de France*. Rapport. URL : <https://www.insee.fr/fr/statistiques/3303318?sommaire=3353488> (visité le 31/10/2019).
- JACQUES, Marie-Paule et Nathalie AUSSENAC-GILLES (2006). « Variabilité des performances des outils de TAL et genre textuel. Cas des patrons lexicosyntaxiques ». In : *Traitement Automatique des Langues* 47.1, p. 11-32. URL : [https://www.researchgate.net/publication/230776133\\_Variabilite\\_des\\_performances\\_des\\_outils\\_de\\_TAL\\_et\\_genre\\_textuel\\_Cas\\_des\\_patrons\\_lexico-syntaxiques](https://www.researchgate.net/publication/230776133_Variabilite_des_performances_des_outils_de_TAL_et_genre_textuel_Cas_des_patrons_lexico-syntaxiques) (visité le 19/12/2019).
- JESSUP, Leonard M. et Joseph S. VALACICH (2002). *Information Systems Today*. Prentice Hall Professional Technical Reference. ISBN : 978-0-13-009414-8.
- KIND, Stuart S. (1987). *The scientific investigation of crime*. Harrogate, England : Forensic Science Services. ISBN : 978-0-9512584-0-8.
- KLEIBER, Georges (1981). *Problèmes de référence : descriptions définies et noms propres*. Centre d'analyse syntaxique de l'Université de Metz.
- KOMTER, Martha L. (2006). « From talk to text : The interactional construction of a police record ». In : *Research on Language and Social Interaction* 39.3, p. 201-228. ISSN : 0835-1813. DOI : 10.1207/s15327973r1si3903\_2. URL : <https://>

- [research.vu.nl/en/publications/from-talk-to-text-the-interactive-construction-of-a-police-reco](http://research.vu.nl/en/publications/from-talk-to-text-the-interactive-construction-of-a-police-reco) (visité le 26/09/2019).
- KU, Chih Hao, Alicia IRIBERRI et GONDY LEROY (2008). « Crime Information Extraction from Police and Witness Narrative Reports ». In : *2008 IEEE Conference on Technologies for Homeland Security*. IEEE, p. 193-198. ISBN : 978-1-4244-1977-7. DOI : [10.1109/THS.2008.4534448](https://doi.org/10.1109/THS.2008.4534448). URL : <http://ieeexplore.ieee.org/document/4534448/> (visité le 30/05/2019).
- LAFON, Pierre et André SALEM (1983). « L'inventaire des segments répétés d'un texte ». In : *Mots. Les langages du politique* 6.1, p. 161-177. DOI : [10.3406/mots.1983.1101](https://doi.org/10.3406/mots.1983.1101). URL : [https://www.persee.fr/doc/mots\\_0243-6450\\_1983\\_num\\_6\\_1\\_1101](https://www.persee.fr/doc/mots_0243-6450_1983_num_6_1_1101) (visité le 04/12/2019).
- LAMY, Bertrand De (2012). « Chronique de droit pénal constitutionnel ». In : *Revue de science criminelle et de droit pénal comparé* N° 1.1, p. 217-237. ISSN : 0035-1733. URL : <https://www.cairn.info/revue-de-science-criminelle-et-de-droit-penal-compare-2018-1-page-163.htm> (visité le 19/12/2019).
- LE PEVEDIC, Solenn et Denis MAUREL (2016). « Retour sur les annotations des entités nommées dans les campagnes d'évaluation françaises et comparaison avec la TEI ». In : *Corela. Cognition, représentation, langage* 14-2. ISSN : 1638-5748. DOI : [10.4000/corela.4644](https://doi.org/10.4000/corela.4644). URL : <http://journals.openedition.org/corela/4644> (visité le 03/10/2019).
- LEBART, Ludovic et André SALEM (1994). *Statistique Textuelle*. Paris : Dunod. URL : <http://lexicometrica.univ-paris3.fr/livre/st94/st94-tdm.html> (visité le 04/07/2019).
- LÉON, Jacqueline (2005). « Claimed and Unclaimed Sources of Corpus Linguistics ». In : *Henry Sweet Society for the History of Linguistic Ideas Bulletin* 44.1, p. 36-50. ISSN : 0267-4971. DOI : [10.1080/02674971.2005.11745607](https://doi.org/10.1080/02674971.2005.11745607). URL : <http://www.tandfonline.com/doi/full/10.1080/02674971.2005.11745607> (visité le 24/09/2019).
- LIZUREY, Richard (2006). *Gendarmerie nationale : les soldats de la loi*. Paris : Presses Universitaires de France. ISBN : 978-2-13-055092-1. DOI : [10.3917/puf.lizu.2006.01](https://doi.org/10.3917/puf.lizu.2006.01). URL : <https://www.cairn.info/gendarmerie-nationale-les-soldats-de-la-loi--9782130550921.htm> (visité le 21/06/2019).

- MALRIEU, Denise et François RASTIER (2001). « Genres et variations morphosyntaxiques ». In : *Traitement Automatique des Langues* 42.2, p. 548-577. URL : [http://www.revue-texto.net/Inedits/Malrieu\\_Rastier/Malrieu-Rastier\\_Genres.html](http://www.revue-texto.net/Inedits/Malrieu_Rastier/Malrieu-Rastier_Genres.html) (visité le 13/02/2019).
- MANNING, Christopher D. (2011). « Part-of-speech Tagging from 97% to 100% : Is It Time for Some Linguistics? » In : *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I. CICLing'11*. Berlin, Heidelberg : Springer-Verlag, p. 171-189. ISBN : 978-3-642-19399-6. URL : <http://dl.acm.org/citation.cfm?id=1964799.1964816> (visité le 16/09/2019).
- MARGOT, Pierre (2014). « Traçologie : la trace, vecteur fondamental de la police scientifique ». In : *Revue internationale de criminologie et de police technique et scientifique* LXVII.1, p. 72-97. ISSN : 1424-4683. URL : [https://serval.unil.ch/fr/notice/serval:BIB\\_47CDEEB806E3](https://serval.unil.ch/fr/notice/serval:BIB_47CDEEB806E3) (visité le 19/12/2019).
- MARIN, Jean-Claude, Myriam QUEMENER et Alexandre GALLOIS (2003). *Analyse criminelle et analyse comportementale*. Rapport. Ministère de la Justice. URL : <https://www.vie-publique.fr/rapport/26000-analyse-criminelle-et-analyse-comportementale-rapport-du-groupe-de-tra> (visité le 19/02/2019).
- MAUREL, Denis et al. (2011). « Cascades de transducteurs autour de la reconnaissance des entités nommées ». In : *Traitement Automatique des Langues* 52.1, p. 69-96. URL : <https://hal.archives-ouvertes.fr/hal-00682805> (visité le 08/10/2019).
- MAYAFFRE, Damon (2002). « Les corpus réflexifs : entre architextualité et hypertextualité ». In : *Corpus* 1. ISSN : 1638-9808. URL : <http://journals.openedition.org/corpus/11> (visité le 21/06/2019).
- MONJEAN-DECAUDIN, Sylvie (2012). *La traduction du droit dans la procédure judiciaire*. Dalloz. Bibliothèque de la justice. ISBN : 978-2-247-11911-0. (Visité le 08/12/2019).
- MOUNIER, Pierre (2017). « Les Humanités numériques, gadget ou progrès? » In : *Revue du Crieur* N° 7.2, p. 144-159. ISSN : 2428-4068. URL : <http://www.cairn.info/revue-du-crieur-2017-2-page-144.htm> (visité le 26/09/2019).

- MUCCHIELLI, Laurent (2006). « L'élucidation des homicides : de l'enchantement technologique à l'analyse du travail des enquêteurs de police judiciaire ». In : *Deviance et Societe* 30.1, p. 91-119. ISSN : 0378-7931. URL : <http://www.cairn.info/revue-deviance-et-societe-2006-1-page-91.htm> (visité le 10/12/2019).
- NADEAU, David et Satoshi SEKINE (2007). « A survey of named entity recognition and classification ». In : *Linguisticae Investigationes* 30.1, p. 3-26. URL : [https://www.researchgate.net/publication/44062524\\_A\\_Survey\\_of\\_Named\\_Entity\\_Recognition\\_and\\_Classification](https://www.researchgate.net/publication/44062524_A_Survey_of_Named_Entity_Recognition_and_Classification) (visité le 19/12/2019).
- NAZARENKO, Adeline et Adam WYNER (2017). « Legal NLP Introduction ». In : *TAL* 58.2, p. 8-18. URL : <https://www.atala.org/content/introduction-1> (visité le 30/03/2019).
- NOUVEL, Damien (2012). « Reconnaissance des entités nommées par exploration de règles d'annotation ». Thèse de doct. Tours : Université François Rabelais. URL : <https://tel.archives-ouvertes.fr/tel-00788630> (visité le 03/10/2019).
- NOUVEL, Damien, Maud EHRMANN et Sophie ROSSET (2015). *Les entités nommées pour le traitement automatique des langues*. Science cognitive. Londres : ISTE. ISBN : 978-1-78406-104-3. URL : <https://iste-editions.fr/products/les-entites-nommees-pour-le-traitement-automatique-des-langues> (visité le 28/03/2019).
- PATIN, Stéphane, Maria ZIMINA et Serge FLEURY (2016). « Lecture Textométrique Différentielle (LTD) de textes législatifs comparables de l'Union européenne ». In : *International Conference on Statistical Analysis of Textual Data (JADT2016)*. Sous la dir. de Damon MAYAFFRE et al. T. 2. Proceedings of 13th International Conference on Statistical Analysis of Textual Data, 7-10 June 2016, Nice (France). Nice, France. URL : <https://hal-univ-diderot.archives-ouvertes.fr/hal-01351412> (visité le 25/09/2019).
- PAUMIER, Sébastien (2016). *Unitex 3.1 manuel d'utilisation*. URL : <https://unitexgramlab.org/releases/3.2rc/man/Unitex-GramLab-3.2rc-usermanual-fr.pdf> (visité le 17/06/2019).
- PINCEMIN, Bénédicte (1999). « Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative ». In : *Corpus et TAL : pour une réflexion méthodologique*. Sous la dir. d'Anne CONDAMINES, Marie-Paule PERY-WOODLEY et

- Cécile FABRE. Cargèse, p. 26-36. URL : [http://icar.univ-lyon2.fr/membres/bpincemin/biblio/pincemin\\_taln99.pdf](http://icar.univ-lyon2.fr/membres/bpincemin/biblio/pincemin_taln99.pdf) (visité le 08/07/2019).
- (2012a). « Hétérogénéité des corpus et textométrie ». In : *Langages* n° 187.3, p. 13-26. ISSN : 0458-726X. URL : <https://www.cairn.info/revue-langages-2012-3-page-13.htm> (visité le 14/06/2019).
- (2012b). « Sémantique interprétative et textométrie ». In : *Texto! Textes et Cultures* XVII.3, p. 1-21. URL : <https://halshs.archives-ouvertes.fr/halshs-00981180> (visité le 24/09/2019).
- (2018). « Sept logiciels de textométrie ». URL : <https://halshs.archives-ouvertes.fr/halshs-01843695> (visité le 24/09/2019).
- PINCEMIN, Bénédicte et Serge HEIDEN (2008). *Qu'est-ce que la textométrie? Présentation*. URL : <http://textometrie.ens-lyon.fr/spip.php?rubrique80> (visité le 24/09/2019).
- POIBEAU, Thierry (2014a). « La linguistique est-elle soluble dans la statistique ? » In : *Revue Sciences/Lettres* 2. ISSN : 2271-6246. DOI : 10.4000/rs1.402. URL : <http://journals.openedition.org/rs1/402> (visité le 24/09/2019).
- (2014b). « Le traitement automatique des langues pour les sciences sociales ». In : *Reseaux* n° 188.6, p. 25-51. ISSN : 0751-7971. URL : <http://www.cairn.info/revue-reseaux-2014-6-page-25.htm> (visité le 02/10/2019).
- RASTIER, François (2001). *Arts et sciences du texte*. Presses Universitaires de France. ISBN : 978-2-13-051932-4. DOI : 10.3917/puf.rast.2001.01. URL : <https://www.cairn.info/arts-et-sciences-du-texte--9782130519324.htm> (visité le 22/05/2019).
- (2004). *Enjeux épistémologiques de la linguistique de corpus*. URL : <http://www.revue-texto.net/index.php?id=543> (visité le 01/07/2019).
- RASTIER, François, Marc CAVAZZA et Anne ABEILLÉ (1994). *Sémantique pour l'analyse – De la linguistique à l'informatique*. Masson. Sciences cognitives. Paris.
- RIBAU, Olivier (2014). *Police scientifique : Le renseignement par la trace*. 1<sup>re</sup> éd. Lausanne : Presses polytechniques et universitaires romandes. ISBN : 978-2-88915-061-8.

- ROSSY, Quentin (2011). « Méthodes de visualisation en analyse criminelle : approche générale de conception des schémas relationnels et développement d'un catalogue de patterns ». Thèse de doct. Université de Lausanne. URL : [https://serval.unil.ch/notice/serval:BIB\\_1AC0D89CA5A4](https://serval.unil.ch/notice/serval:BIB_1AC0D89CA5A4) (visité le 05/04/2019).
- ROSSY, Quentin et al. (2019). « Le traitement de l'information dans l'enquête criminelle ». In : *Nouveau traité de sécurité. Sécurité intérieure et sécurité urbaine*. Montréal : Hurtubise, p. 576. ISBN : 978-2-89781-345-1.
- ROUBAUD, Marie-Noëlle (2017). « Le français écrit : transcription et édition. Le cas des textes scolaires ». In : *Corpus 16*. ISSN : 1638-9808. URL : <http://journals.openedition.org/corpus/2770> (visité le 03/12/2019).
- ROWLEY, Jennifer (2007). « The wisdom hierarchy : representations of the DIKW hierarchy ». In : *Journal of Information Science* 33.2, p. 163-180. ISSN : 0165-5515, 1741-6485. DOI : [10.1177/0165551506070706](https://doi.org/10.1177/0165551506070706). URL : <http://journals.sagepub.com/doi/10.1177/0165551506070706> (visité le 07/06/2019).
- SALEM, André (1986). « Segments répétés et analyse statistique des données textuelles ». In : *Histoire & Mesure* 1.2, p. 5-28. DOI : [10.3406/hism.1986.1518](https://doi.org/10.3406/hism.1986.1518). URL : [https://www.persee.fr/doc/hism\\_0982-1783\\_1986\\_num\\_1\\_2\\_1518](https://www.persee.fr/doc/hism_0982-1783_1986_num_1_2_1518) (visité le 14/02/2019).
- SAVELKA, Jaromir et al. (2017). « Sentence Boundary Detection in Adjudicatory Decisions in the United States ». In : *Traitement Automatique des Langues* 58, p. 21. URL : <https://www.atala.org/content/sentence-boundary-detection-adjudicatory-decisions-united-states> (visité le 19/12/2019).
- SCHMID, Helmut (1994). « Probabilistic Part-of-Speech Tagging Using Decision Trees ». In : *Proceedings of the International Conference on New Methods in Language Processing*. Manchester.
- SCHNEDECKER, Catherine (2017). « Les chaînes de référence : une configuration d'indices pour distinguer et identifier les genres textuels ». In : *Langue française* 195.3, p. 53-72. ISSN : 0023-8368. URL : <https://www.cairn.info/revue-langue-francaise-2017-3-page-53.htm> (visité le 19/12/2019).
- SCHRAAGEN, Marijn, Matthieu BRINKHUIS et Floris BEX (2017). « Evaluation of Named Entity Recognition in Dutch online criminal complaints ». In : 7, p. 3-16.

- SÉNAT (2009). *Le rôle de la police judiciaire dans l'instruction des affaires pénales*. Les documents de travail du Sénat LC 198, p. 35. URL : [https://www.senat.fr/lc/lc198/lc198\\_mono.html](https://www.senat.fr/lc/lc198/lc198_mono.html) (visité le 12/12/2019).
- SINCLAIR, John (1991). *Corpus, Concordance, Collocation*. Oxford University Press. ISBN : 978-0-19-437144-5.
- SOCIÉTÉ FRANÇAISE DE MÉDECINE LÉGALE (2009). *Modèle de rapport d'autopsie type*. Rapport. URL : [http://sfml-asso.fr/images/docs/Rapport\\_autopsie\\_.pdf](http://sfml-asso.fr/images/docs/Rapport_autopsie_.pdf) (visité le 14/06/2019).
- SÖZE-DUVAL, Keyser (2008). « Pour une textométrie opérationnelle ». URL : <http://www.tal.univ-paris3.fr/trameur/RTI6provisoire.doc> (visité le 19/12/2019).
- SWALES, John (1990). *Genre Analysis : English in Academic and Research Settings*. Cambridge University Press. ISBN : 978-0-521-33813-4.
- TAMBA, Irène (1983). « La composante référentielle dans « Un manteau de laine », « un manteau en laine » ». In : *Langue française* 57.1, p. 119-128. DOI : [10.3406/lfr.1983.5160](https://doi.org/10.3406/lfr.1983.5160). URL : [http://www.persee.fr/doc/lfr\\_0023-8368\\_1983\\_num\\_57\\_1\\_5160](http://www.persee.fr/doc/lfr_0023-8368_1983_num_57_1_5160) (visité le 18/12/2019).
- TAPASWI, Makarand, Martin BÄUML et Rainer STIEFELHAGEN (2014). « Story-Graphs : Visualizing Character Interactions as a Timeline ». In : *2014 IEEE Conference on Computer Vision and Pattern Recognition*, p. 827-834. DOI : [10.1109/CVPR.2014.111](https://doi.org/10.1109/CVPR.2014.111). URL : <https://ieeexplore.ieee.org/document/6909506> (visité le 19/12/2019).
- TODIRASCU, Amalia (2019). « Genre et classification automatique en TAL : le cas de genres journalistiques ». In : *Linx. Revue des linguistes de l'université Paris X Nanterre* 78. ISSN : 0246-8743. DOI : [10.4000/linx.3183](https://doi.org/10.4000/linx.3183). URL : <http://journals.openedition.org/linx/3183> (visité le 19/09/2019).
- VALETTE, Mathieu (2016). « Analyse statistique des données textuelles et traitement automatique des langues. Une étude comparée ». In : *International Conference on Statistical Analysis of Textual Data (JADT2016)*. Sous la dir. de D. MAYAFFRE et al. T. 2. Proceedings of 13th International Conference on Statistical Analysis of Textual Data, 7-10 June 2016, Nice (France). Nice, France, p. 697-706. URL : <https://hal-inalco.archives-ouvertes.fr/hal-01335084> (visité le 24/09/2019).

- VERKAMPT, Fanny (2013). « Comment entendre un enfant témoin lors d'une audition judiciaire ? » In : *Journal du droit des jeunes* 330.10, p. 11-22. ISSN : 2114-2068. URL : <https://www.cairn.info/revue-journal-du-droit-des-jeunes-2013-10-page-11.htm> (visité le 19/12/2019).
- VERKAMPT, Fanny, Magali GINET et Cindy COLOMB (2010). « L'Entretien Cognitif est-il efficace pour aider de très jeunes enfants à témoigner d'un Événement répété dans le temps ? » In : *L'Année psychologique* 110.4, p. 541-572. ISSN : 0003-5033. URL : <https://www.cairn.info/revue-l-annee-psychologique1-2010-4-page-541.htm><https://www-cairn-info.bibdocs.u-cergy.fr/revue-l-annee-psychologique1-2010-4-page-541.htm> (visité le 19/12/2019).
- VILAMOT, Bernard et al. (2010). « L'audition du mineur : la forme du discours ». In : *L'information psychiatrique* Volume 86.10, p. 859-867. ISSN : 0020-0204. URL : <https://www.cairn.info/revue-l-information-psychiatrique-2010-10-page-859.htm> (visité le 26/06/2019).
- VLAMYNCK, Hervé (2011). « Le questionnement policier ». In : *Les Cahiers de la Justice* 4.4, p. 57-68. ISSN : 1958-3702. URL : <https://www.cairn.info/revue-les-cahiers-de-la-justice-2011-4-page-57.htm> (visité le 01/07/2019).
- WISNIEWSKI, Guillaume (2007). « Apprentissage dans les espaces structurés : application à l'étiquetage de séquences et à la transformation automatique de documents ». Thèse de doct. Université Paris 6.
- WROBLEWSKI, Jerzy (1988). « Les langages juridiques : une typologie ». In : *Droit et Société* 8.1, p. 13-27. DOI : 10.3406/dreso.1988.983. URL : [https://www.persee.fr/doc/dreso\\_0769-3362\\_1988\\_num\\_8\\_1\\_983](https://www.persee.fr/doc/dreso_0769-3362_1988_num_8_1_983) (visité le 13/09/2019).