Traitement automatique des langues et linguistique de corpus pour la reconnaissance d'entités en analyse criminelle

Par **Lucie GIANOLA***

RÉSUMÉ

L'analyse criminelle est une discipline d'appui aux enquêtes fondée sur l'exploitation du dossier de procédure. Afin d'améliorer ses méthodes et d'accélérer le temps de traitement, l'analyse criminelle cherche à se doter de nouveaux outils informatiques. Nous présentons dans cet article les résultats d'une thèse de sciences du langage consacrée à la construction et au test d'une approche de détection des entités issue du traitement automatique des langues dans le texte des auditions de témoin. À l'occasion de ce développement, nous posons également les bases épistémologiques de l'utilisation d'outils d'extraction d'information et d'exploration textuelle dans le cadre de l'enquête criminelle.

Mots clés: analyse criminelle, linguistique de corpus, traitement automatique des langues, extraction d'information.

ABSTRACT

Criminal analysis is an investigative support discipline based on the processing of the judicial procedure file. To improve its methods and speed up response time, criminal analysis is looking to adopt new data-processing tools. We present here the results of a PhD thesis in linguistics devoted to the building and testing of an approach for detecting entities from natural language processing (NLP) in the text of witness interviews. We also set the epistemological foundations for the use of information extraction and textual exploration tools in criminal investigations.

Keywords: criminal analysis, corpus linguistics, natural language processing, information retrieval.

1. Problématique

L'analyse criminelle opérationnelle, en tant que pratique d'enquête basée sur l'exploitation du dossier de procédure, se heurte à des difficultés de gestion des documents et de gestion de l'information. À ce jour, cette pratique ne dispose pas d'outils informatiques dédiés simplifiant la navigation dans le texte et le repérage d'informations pertinentes. Lire le dossier, le comprendre et l'exploiter pour en produire une analyse est une tâche chronophage et fastidieuse.

* Ingénieure de recherche au Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, CNRS.

Or, le domaine du traitement automatique des langues (TAL) et celui de la linguistique outillée disposent de technologies et de solutions logicielles qui permettent d'appréhender méthodiquement de grands volumes de données textuelles. Établis depuis plusieurs dizaines d'années[1], leur développement s'est accompagné de réflexions conceptuelles et épistémologiques concernant toutes les étapes de leur mise en application, de la collecte et la préparation des données à l'interprétation des résultats produits. Nous présentons ici la synthèse d'une thèse de doctorat en sciences du langage soutenue en 2020 à l'Université de Cergy-Pontoise consacrée à l'application des méthodes du TAL et de la linguistique de corpus au cas de l'analyse criminelle. Ces travaux ont cherché à connecter le concept d'entité criminelle des sciences forensiques à celui d'entité nommée du TAL, dans l'objectif de proposer une approche de détection automatique des entités criminelles dans le texte des auditions de témoin.

2. Cadre et données de recherche

Nos recherches se situent dans le cadre d'un partenariat entre l'université de Cergy-Pontoise et le département sciences de l'analyse criminelle (DSAC) du Pôle Judiciaire de la Gendarmerie Nationale (PJGN). Composé d'une dizaine de gendarmes analystes criminels, le DSAC possède une compétence nationale qui lui permet d'être saisi sur l'ensemble du territoire français, et pratique essentiellement l'analyse criminelle opérationnelle[2].

Méthode. Si l'analyse criminelle peut être pratiquée de manière synchrone (au fil de la production des actes d'enquête, comme dans le cas des enquêtes en flagrant-délit) ou asynchrone avec les investigations, le DSAC travaille la plupart du temps de manière asynchrone, en particulier sur des affaires de type « cold cases ». Deux analystes lisent le dossier intégralement, extraient l'information qu'il contient, l'analysent, puis produisent des hypothèses de travail. Les hypothèses doivent être étayées par des faits référencés dans la procédure et sont soutenues par des schémas réalisés à l'aide du logiciel de représentation graphique Analyst's Notebook[3]. Le but est de « transformer » l'information brute difficilement accessible dans la masse en information exploitable pour l'enquête en créant une vue d'ensemble du dossier. La recherche et la structuration de l'information s'appuient sur le concept d'entité criminelle, que nous détaillerons plus bas.

Matière. Le dossier de procédure rassemble les documents produits et consignés à un certain stade de l'enquête. C'est une archive hétérogène, compilant une grande variété de documents issus de nombreuses sources différentes mais qui concourent tous à un même but, celui de la manifestation de la vérité. Du fait de sa position centrale et de sa compétence étendue, le DSAC peut être confronté à des dossiers anciens, volumineux, qui ont circulé entre

plusieurs acteurs de la justice et parfois même entre police et gendarmerie. En termes documentaires, cela implique que les documents, lorsqu'ils sont fournis au format numérique, sont la plupart du temps des scans au format PDF des documents papiers initiaux. Le format PDF n'étant pas un format exploitable numériquement tel quel, les fichiers doivent être pré-traités par un logiciel de reconnaissance optique de caractères (*optical character recognition*, OCR) qui extrait le texte des scans afin de rendre les recherches plein-texte[4] possibles. Malheureusement, plusieurs facteurs dégradent les performances de ce pré-traitement, comme la qualité initiale des documents papiers, l'ajout d'annotations manuscrites qui ne sont pas reconnues par le logicel d'OCR, ou encore lorsque les documents ont été photocopiés plusieurs fois avant d'être numérisés.

Organisation documentaire. L'organisation des documents au sein du dossier ne répond à aucune norme et dépend des pratiques de l'enquêteur initial ou du bureau du magistrat. 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. De même, l'organisation en sous-dossiers peut être thématique, chronologique, ou ne présenter aucune logique particulière. Autant de configurations qui compliquent la prise en main du dossier par l'analyste, qui ne peut pas compter sur un classement clair et établi pour se repérer dans les fichiers.

Le texte et le contenu. Les documents compilés dans la procédure sont très variés : procès-verbaux d'audition, de perquisition, de renseignement, de synthèse, rapports d'expertise divers (autopsie, toxicologie, balistique...), réquisitions, retours des réquisitions (factures téléphoniques détaillées, relevés bancaires), photos, vidéos, ainsi que toutes sortes de documents versés au dossier (avis de recherche, portraits-robots, plans, cartes, coupures de presse, schémas, documentation, lettres anonymes...). Certains sont des documents graphiques, d'autres consignent du texte plus ou moins normé, d'autres des tableaux de données.

Cette diversité, à laquelle s'ajoutent les volumes différents de chaque type de document (d'un seul document pour les rapports d'expertise à plusieurs centaines ou milliers pour les auditions), demande d'adapter les traitements : techniquement, il n'est pas possible de construire un logiciel dans lequel on verserait l'intégralité du dossier et qui traiterait tous les types de documents. Du point de vue des analystes criminels, l'intérêt des différents types de documents est fonction du type de cas traité (homicide, criminalité en réseau, disparition, etc.). Néanmoins dans la plupart des cas, les auditions de témoins sont des documents fondamentaux qui s'avèrent difficiles à traiter : en grand nombre, elles consignent sur une longueur variable d'une ou plusieurs pages du texte libre[5] riche en informations.

L'analyse criminelle est présentée comme une méthode structurée qui apporte de la maîtrise et qui clarifie les informations de l'enquête (Rossy, 2011, p. 7). Cette définition contraste avec la réalité du dossier qui ne répond à aucune

norme ou organisation établie et rassemble des informations hétérogènes aux formes diverses. Face à cette situation, c'est à l'outil de s'adapter plutôt qu'à la matière. Nous avons alors délimité nos recherches à un type de document : les auditions de témoins et de personnes gardées à vue. Ce choix s'explique par le fait que les auditions sont nombreuses, riches en information et présentent une forme difficile à synthétiser puisqu'elles sont sous forme de texte libre.

3. Le texte des auditions

Plusieurs dossiers traités par le DSAC ont été mis à notre disposition sous couvert d'une convention de confidentialité. Les dossiers que nous avons pu consulter concernaient tous des affaires d'atteintes aux personnes, homicides ou disparition, résolus et jugés, au format électronique et papier. Parmi les dossiers consultés, un seul était exploitable informatiquement en l'état, avec une qualité de reconnaissance optique de caractères ne nécessitant pas de corrections.

Ce dossier rassemble 370 auditions de témoins et de personnes gardées à vue (environ 600 000 mots[6]) qui constituent ce que nous appellerons désormais le corpus. En linguistique, un corpus est un ensemble de données langagières authentiques rassemblées selon une hypothèses qui le sous-tend (Sinclair (1991), Mayaffre (2002), Rastier (2004)), et comportant parfois des informations supplémentaires comme des étiquetages morpho-syntaxiques[7], ou une structuration interne en sous-corpus.

3.1 Description générale

Le procès-verbal d'audition est un écrit rendant compte d'un échange oral ayant eu lieu entre un acteur de l'enquête (officier de police judiciaire ou magistrat) et une personne détenant potentiellement des informations sur les faits (témoin ou mis en cause). Cet échange est transcrit soit par un officier de police judiciaire, soit par un greffier. Le document est constitué d'un en-tête contenant les éléments de contexte relatifs à l'enquête et au procès-verbal : date, lieu, objet de l'audition, etc., puis d'une partie rapportant les éléments d'état-civil de la personne entendue sous forme de texte ou de tableau, et enfin du corps de l'audition à proprement parler[8]. Le texte est rédigé soit sous la forme d'un dialogue avec une alternance question-réponse, dans laquelle le témoin s'exprime à la première personne du singulier et les officiers à la première personne du pluriel, soit sous la forme d'un récit du témoin d'un seul tenant à la première personne du singulier. L'emploi des pronoms personnels est une trace du caractère initialement oral de la matière de l'audition, et marque la dimension narrative des propos du témoin. Toutefois, il semble que le texte ne soit pas un compte-rendu exact des propos : on ne rencontre que très peu de marques de l'oral (répétitions, pauses, hésitations, reformulations). Pour approfondir ces aspects, nous explorons ci-dessous la régularité et la narrativité du texte.

3.2 Régularité & narrativité

La fonction « segments répétés » du logiciel de lexicométrie Lexico5[9], qui permet de repérer les répétitions de suites de formes dans un corpus (Lafon et Salem (1983), Lebart et Salem (1994)), a décompté près de 8790 segments longs de deux à onze tokens[10] répétés dix fois ou plus (échantillon en tableau 1, pour le tableau complet voir Gianola (2020, p. 135-137)). 85 segments sont longs de onze tokens et apparaissent dix fois ou plus. Pour se faire un ordre d'idée, le corpus de démonstration fourni avec le logiciel, le Père Duchêne (Salem, 1986), présente 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. Ces segments répétés sont, pour les plus fréquents (de l'ordre de plusieurs centaines à plusieurs dizaines), des phrases « réglementaires » énonçant les circonstances de l'audition, les articles du code de procédure pénale relatifs à la conduite d'audition, les éléments d'état-civil, etc. On repère aussi avec une moindre fréquence (plusieurs dizaines) 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. Cela s'explique du fait que dans cette affaire, plusieurs témoins ont été interrogés selon une trame d'audition préparée à l'avance.

L'étude des segments répétés donne l'aperçu d'un contenu structurant transversalement les auditions. La régularité de l'expression touche les aspects réglementaires du texte mais aussi des passages utilisant les pronoms personnels.

Tableau 1 : *Exemples de segments répétés*

Segment	Fréquence
et lui donnons connaissance des faits pour lesquels sa déposition est	315
Vu les articles 16 à 19 et 151 à 155 du	268
entendu séparément et hors la présence de la personne mise en	241
Je prends connaissance des faits pour lesquels ma déposition est requise	39
Nous vous montrons une série de 06 photographies de couteaux référencée	31
Avez vous remarqué une personne dont la tenue vestimentaire ne vous	12

Or il semble improbable que toutes les personnes auditionnées aient prononcé rigoureusement les mêmes phrases et que les enquêteurs aient posé mot à mot les mêmes questions.

L'étiquetage morpho-syntaxique fourni par TreeTagger (Schmid, 1994) dans le logiciel TXM (Heiden et al., 2010) indique qu'après les noms communs (16,4 %) et les prépositions (10 %), les pronoms personnels constituent la catégorie la plus fréquente, représentant environ 9 % des étiquettes morphosyntaxiques (qu'on appelle aussi parties du discours). En ce qui concerne les temps des verbes, le présent, les participes passés et l'imparfait sont les trois plus fréquents, en raison d'une double dimension narrative du

texte des auditions : d'une part, celle du récit de la conduite de l'audition, avec les phrases à la première personne du pluriel qui attestent de la conduite réglementaire de l'audition, et d'autre part celle du récit du témoin qui raconte son emploi du temps, ses habitudes, ce qu'il a vu, etc.

Ces deux aspects, régularité et narrativité, font apparaître le texte des auditions comme une retranscription orale conformée : on y retrouve l'aspect narratif de la conversation et sa structure dialogique, l'emploi des pronoms personnels contribuant à créer l'illusion d'un discours direct, mais la régularité de certains passages démontre l'hypothèse d'un récit en partie basé sur des formules toutes faites. D'après les analystes du DSAC, l'usage de ces formules est acquis « sur le tas » au cours de la carrière et au contact d'autres agents, la conduite d'audition ne faisant pas l'objet d'une formation spécifique[11].

Cette étude aide à cerner la nature du texte des procès-verbaux d'audition, qui doit être pris pour ce qu'il est : la trace écrite d'un échange oral produite par un acteur de l'enquête d'après les propos d'un protagoniste des faits. Cela permet également d'ancrer notre recherche dans une linguistique du texte malgré le caractère initialement oral de l'échange, et d'ouvrir la perspective à d'autres questions pour le futur, comme celle de l'auteur du texte de l'audition.

4. Linguistique, informatique et analyse criminelle

La linguistique et l'informatique, habituées à dialoguer dans le domaine du traitement automatique des langues, doivent ici s'adapter au cas de l'analyse criminelle, envisagée ici comme domaine d'application[12].

4.1 Entités criminelles

Le concept d'entité criminelle, en analyse criminelle et plus généralement dans le cadre de l'enquête, sert à désigner tout élément ou paramètre du monde réel impliqué ou mentionné 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] », et propose une revue de ce que le terme recouvre. 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. L'entité, replacée dans un cadre temporel et spatial, sert à délimiter ce dont on veut parler (Ribaux, 2014, p. 373) et formalise le processus : face à des faits criminels, l'enquêteur recherche des entités, les situe dans le cadre, et produit ensuite des liens entre elles. Les entités de l'enquête ne sont pas toujours clairement identifiées ou nommées, elles sont conceptualisées (Gianola, 2020, p. 64). Le concept d'entité criminelle n'est donc pas intrinsèquement lié aux documents produits par les investigations, pour les analystes criminels la conceptualisation des entités s'opère, à la lecture du dossier, via le texte. Le repérage automatique des entités criminelles faciliterait l'analyse criminelle dans les situations où la quantité de textes à traiter devient trop importante pour un traitement à la main.

4.2 Linguistique de corpus & Humanités Numériques

La linguistique de corpus est une branche des sciences du langage qui prend pour objet d'étude des ensembles de productions langagières dites « réelles », par contraste avec des pratiques d'étude de la langue employant des exemples construits pour les besoins de la démonstration. La linguistique de corpus, comme son nom l'indique, s'appuie sur de grands ensembles de données langagières (écrites ou orales) pour former des observations avec le souhait d'une validité statistique. Ces observations sont réalisées à l'aide de logiciels d'exploration textuelle comme ceux que nous avons déjà évoqués : TXM (Heiden et al., 2010), Lexico, Iramuteq[13], AntConc[14], Hyperbase[15]… (Pincemin, 2018).

Les humanités numériques (digital humanities), domaine apparu au milieu du XX^{e} siècle (Berra (2015), Citton (2015)), cherchent à exploiter les apports des technologies numériques aux sciences humaines et sociales. Bien que leur définition et leurs frontières soient sujettes à discussion, 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. Avec un caractère transdisciplinaire fort, elles cherchent à établir méthodes, dispositifs et perspectives heuristiques liés au numérique dans les sciences humaines et sociales (Dacos et Mounier (2015), Dacos (2010)).

Linguistique de corpus et humanités numériques apportent un cadre épistémologique et méthodologique pour le développement et l'usage de solutions informatiques au service de la réflexion humaine (Poibeau (2014a), Poibeau (2014b), Valette (2016)), un cadre que nous souhaitons réutiliser dans le cas de notre recherche. La dimension humaine et intellectuelle de l'analyse criminelle nécessite en effet de penser sa mise en application au-delà d'un simple progrès technologique.

4.3 Informatique : extraction d'entités nommées

L'extraction automatique d'entités nommées (*named-entities recognition* en anglais, ou NER) est une thématique de l'extraction d'information, une sous-discipline du TAL. L'objectif de l'extraction d'information est de repérer automatiquement des informations d'intérêt, la plupart du temps ce qu'on appelle des entités nommées, dans du texte en langue naturelle[16].

4.3.1 Définition

La revue de l'état de l'art au sujet du concept d'entité nommée laisse constater une absence de consensus dans sa définition. Ehrmann (2008, p. 168) fait la proposition suivante : « Étant donnés 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. » La notion clé est la référence : le lien entre une expression linguistique et un élément du monde (Nouvel, Ehrmann, & Rosset, 2015, p. 31)[17]. Il s'agit en somme, pour une liste définie d'informations d'intérêt correspondant à des objets du monde réel, de déterminer les formes linguistiques que ces objets peuvent prendre. Les entités

nommées peuvent être des personnes, des lieux, des dates, des montants, etc., ou dans des cas plus spécialisés comme celui du domaine biomédical, des noms de molécules, des dosages, des posologies. Le plus souvent, la recherche en extraction d'information concentre ses efforts sur des formes normées et facilement modélisables, par exemple pour les mentions de personnes, on considérera les segments « prénom + nom », « nom + prénom », « titre de civilité + nom », etc. Ces formats ne couvrent toutefois pas l'éventail possible des segments linguistiques faisant référence à un élément du monde réel : la description définie, structure en le + nom (« le président de la République », « le collègue de Pierre »), stabilisée par Kleiber (1981), peut aussi être considérée comme une entité nommée.

4.3.2 État de l'art des approches de détection d'entités nommées

Les méthodes de détection des entités nommées se décomposent essentiellement en trois approches :

— Les approches par règles, dites aussi symboliques : elles consistent à décrire via des grammaires ou des fichiers de règles les caractéristiques de ce que que l'on souhaite détecter (par exemple : « tous les mots qui se terminent par -er »). On peut également faire appel à des ressources linguistiques comme des lexiques projetés sur le texte.
— Les approches automatiques, qui consistent à fournir à un algorithme un corpus annoté comme corpus d'exemples d'après lequel l'algorithme apprend les éléments d'intérêt. Les approches neuronales *(deep learning)* construisent des représentations numériques sous formes de vecteurs et permettent une meilleure généralisation. Elles sont particulièrement en vogue aujourd'hui (Allauzen & Schütze, 2018).
— Les approches hybrides, qui combinent les approches par règles et automatiques, par exemple en utilisant un corpus annoté via une approche par règles pour alimenter un algorithme.

Grouin (2013) propose un état de l'art mettant en avant les avantages et les inconvénients de chaque approche. Les approches par règles, qui s'appuient sur des connaissances expertes, présentent l'avantage de pouvoir être déployées rapidement, d'être explicables et donc d'offrir une meilleure maîtrise de leur fonctionnement. Elles demandent en revanche une forte maintenance sur le long terme et sont peu généralisables. Les approches automatiques, plus adaptables et robustes, sont toutefois plus longues à déployer et coûteuses en ressources humaines, car la préparation des données d'exemple demande de longues et fastidieuses phases d'annotation. De plus, ces approches fonctionnent comme des boîtes noires : il n'est pas possible d'expliquer les erreurs ou les manquements. Enfin, les approches neuronales demandent des volumes de données d'entraînement très conséquentes, ainsi qu'une puissance de calcul et une bonne compréhension des mécanismes mathématiques qui les sous-tendent. Le choix d'une approche s'opère en fonction de la situation de recherche et des ressources à la fois en termes de données, de ressources humaines et de puissance de calcul.

Quelle que soit l'approche retenue, la recherche d'entités nommées commence par la détermination des catégories (quelles entités prendre en compte) et comment les annoter (les réalisations linguistiques correspondantes), comme l'expliquent Nouvel et al. (2015, p. 22).

4.3.3 Détection d'entités dans des textes liés au domaine criminel

Nous considérons cinq références dans la littérature ayant cherché à détecter des entités criminelles.

Arulanandam, Savarimuthu, et Purvis (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, Brinkhuis, et Bex (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, Beires Nogueira, Antunes, et Ferreira (2019) ont utilisé la suite d'apprentissage automatique OpenNLP[18] 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, Xu, et Chen (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, Iriberri, et Leroy (2008) détectent quinze catégories 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 »[19] construit sur des ressources linguistiques et encyclopédiques combiné à des modules GATE[20]. 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 en anglais.

Si cette revue sommaire illustre la diversité des entités qui peuvent intéresser l'extraction d'information à des fins de recherche criminelle et les écarts de couverture du sujet, aucune des contributions ne fournit d'exemple des entités recherchées et des formes qu'elles prennent dans les textes considérés. D'autre part, aucune ne porte sur des auditions de témoins, or il est avéré (Jacques et Aussenac-Gilles (2006), Nadeau et Sekine (2007)) que la prise en compte du genre textuel influence les performances des systèmes de TAL. Enfin, aucune contribution ne concerne le français.

4.4 Conclusion

L'enjeu de notre recherche se dessine comme la convergence des concepts, outils et méthodes de trois disciplines : au concept d'entité criminelle utile aux enquêtes, il faut adapter les méthodes de recherche d'information, tout en exploitant les apports d'une bonne connaissance linguistique des textes

exploités. Pour cela, nous devrons définir les entités d'intérêt et déterminer les formes linguistiques qu'elles prennent. La conception d'un outil d'extraction d'information et d'exploration textuelle pour l'analyse criminelle doit aussi, à terme, amener à l'élaboration de nouvelles méthodologies, inspirées notamment des humanités numériques, pour accompagner les utilisateurs finaux d'une telle solution. Il est important de souligner le découpage entre la recherche de l'information, tâche partiellement automatisable et sous certaines conditions, et le travail intellectuel de raisonnement mettant à profit les connaissances humaines et qui doit rester à la charge de l'analyste.

5. Les entités en analyse criminelle

Nous avons pu constater qu'une grande variété d'objets peut correspondre au concept d'entité criminelle. Cette variété ne permet pas d'anticiper l'importance d'une entité plutôt qu'une autre dans une affaire en particulier. Pour commencer, cinq types d'entités ont été retenus en concertation avec les analystes criminels du SCRC : les dates, lieux, personnes, véhicules et numéros de téléphone. En effet, si l'on admet une affaire criminelle comme le résultat d'une activité humaine inscrite dans le temps et l'espace, ces cinq entités faisant référence à des objets et paramètres du monde réel en donnent une description substantielle.

5.1 Catégories et réalisations linguistiques correspondantes

Nous proposons ci-dessous une définition de chacune des entités avec des exemples pseudonymisés : les éléments portant atteinte à la vie privée des personnes impliquées sont remplacés par d'autres éléments du même type (un prénom est remplacé par un autre prénom, un lieu par un autre lieu, une date par une autre date, etc.) Cette modification n'a été réalisée que pour les besoins de communication, notre travail a porté sur le texte authentique. Par ailleurs, tous les exemples cités dans la suite de cet article le sont tels qu'ils figurent dans le dossier : avec leurs coquilles, fautes d'orthographe, etc.

5.1.1 Numéros de téléphone

Les numéros de téléphone sont les entités les plus stables : ils se composent d'une suite arbitraire de chiffres configurés selon des normes nationales et séparés d'espaces ou de signes typographiques (point, tiret ou parfois barre oblique). Il est possible de détecter les numéros de téléphone à l'aide d'expressions régulières[21]. Les numéros de téléphone n'engageant pas de critères linguistiques, nous n'approfondissons pas notre propos à leur sujet.

5.1.2 Dates

Les dates, et plus largement les éléments temporels servent à reconstituer la chronologie des faits examinés, par exemple en comparant le déroulé en parallèle de plusieurs récits. Pour l'étude d'éléments biographiques, ce sont

les dates (jours), les mois, les périodes qui retiendront l'attention, alors que l'analyse d'un emploi du temps plus resserré s'intéressera aux horaires et aux périodes de la journée (heures, matin, midi, après-midi, soirée...).

— **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
— **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 [...]
— Savez-vous quel a été l'emploi du temps précis de Ghislaine pour la journée du **dix neuf septembre 2009** ?
— Philippe m'a parlé **jeudi 27/08/2017 un peu à près 17 h** alors qu'il s'apprêtait à aller chercher Léo à la crèche.
— Je l'ai eu au téléphone le **10 ou le 11 septembre 2009** avant le week-end.

Au-delà du mélange des formes (mélange chiffres et lettres, barres obliques, ordre des éléments différent), on note aussi dans le dernier exemple une imprécision dans les propos du témoin, qui ne se souvient pas de la date exacte qu'il évoque. Cet aspect est propre aux auditions et récurrent, puisqu'on ne peut exiger des souvenirs parfaits de la part des témoins.

Deux facteurs influencent la forme des informations recueillies : l'échelle, qui concerne la focalisation des éléments (biographiques ou relatifs à l'emploi du temps) et la précision, qui concerne l'exactitude (par exemple, « en fin de matinée » et « à 11 h 30 » désignent une même période temporelle mais pas avec la même précision). Précision et échelle ne sont pas en concurrence, des éléments biographiques peuvent être rapportés avec précision (par exemple, la date d'un mariage) et des éléments d'emploi du temps peuvent être imprécis (cf. l'exemple précédent comparant « en fin de matinée » et « à 11 h 30 »).

5.1.3 Lieux

Comme les éléments temporels, les lieux et informations géographiques sont influencés par la précision et l'échelle. Dans le cas de l'examen d'éléments biographiques, ce sont des lieux d'échelle large comme des villes, des régions, voire des pays qui seront rapportés. La présence d'une personne dans une région en particulier peut représenter un élément pertinent pour l'analyse de parcours de vie. Dans le cas de l'examen d'un emploi du temps plus précis, ce seront les adresses, les rues, les itinéraires qui constitueront une information pertinente pour les analystes.

— Elle s'est mariée récemment et habite **Annecy**, c'était sa grande copine comme une soeur.
— Sarah habite prés de **CAEN** et Jessica habite en **Bretagne**.
— Il y a environ sept ans, Clément travaillait sur **Lyon** dans une chocolaterie.
— 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.

— Je sais qu'il se promène **en forêt du côté de l'hôpital vers la route de Launaguet**.
— 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**.

Les exemples ci-dessus illustrent la complexité de la notion de lieu, qui s'applique aussi bien à des villes, des régions, des zones, mais aussi des pays ou des adresses,

5.1.4 Personnes

Les personnes sont des entités particulièrement importantes pour l'analyse criminelle puisqu'il s'agit des acteurs des faits examinés. Comme les entités vues précédemment, elles peuvent apparaître de différentes manières[22].

— QUESTION : **M^me^ WOERTH** vous a t elle parlé de problèmes quelle aurait pu avoir ?
— Question : Est ce que **M. PIÈGE** connaissait **Ghislaine WOERTH** ?

Dans ces deux exemples, la mention des personnes correspond à la définition des entités nommées. Les mentions sont claires, non-ambigës et correspondent aux structures attendues.

— Je me nomme **ROUSSEL Chantal épouse GATTIAT**. Je suis née le 05 mars 1960 à VESOUL.
— Lorsque le **couple KNOPE** habitait à Ivry-sur-Seine, nous allions souvent chez eux avec mes parents.
— Nous avons rencontrés **M. et Mme FERRAND** et nous avons emménagé dans la maison environ un mois après en novembre 2001.
— 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.

Ici, les mentions de personnes dévient du format structuré attendu par les entités nommées. Les noms d'épouse, de divorcée, de naissance apportent un complément d'information important. On note aussi des cas où le nom de famille est utilisé pour désigner un couple ou une famille et pas seulement un individu.

— 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.
— Le conducteur est un homme d'environ 45-50 ans, de forte corpulence.
— QUESTION : Pouvez-vous me décrire l'individu en question ? REPONSE : C'était un homme type maghrébin, assez maigre, environ 1 m 75, 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.

Enfin, ces derniers exemples illustrent des cas typiques des auditions de témoins où la mention d'une personne passe par un nom générique d'humain[23] (Cappeau & Schnedecker, 2018). Ce nom générique est complété d'une

description de longueur variable usant de critères variés et imprévisibles. Ces cas sont typiques aux auditions de témoins : la personne auditionnée, lorsqu'elle ne connaît pas le nom de l'individu qu'elle mentionne, doit en faire une description. Ces aspects rendent ces mentions difficiles à reconnaître dans la perspective de l'extraction d'information.

5.1.5 Véhicules

Les véhicules peuvent permettre, surtout lorsqu'ils sont identifiés par un numéro d'immatriculation, de remonter à une personne. Toutefois, la difficulté réside dans la variété de ce qui peut être considéré comme un véhicule (automobiles, deux roues motorisés, vélo, camionnettes, camions...), et du fait que les témoins sont très rarement en mesure de désigner un véhicule de façon non-ambiguë. Typiquement, ils ne peuvent pas citer le numéro d'immatriculation entièrement mais en livrent des bribes, accompagnés d'une description fondée sur les éléments qui auront retenu leur attention.

— 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.
— En ce qui concerne la voiture, elle serait bleue clair et l'immatriculation se terminerait par « 06 ».
— 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.
— Il s'agissait d'une camionnette assez longue, genre gros TRAFIC, elle était de couleur rouge, d'un modèle ancien.
— 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.

Comme dans le cas des personnes inconnues, les témoins emploient différents critères pour décrire les véhicules, mais aucun n'est « stabilisé » en ce qui serait un critère minimal distinctif : ni nom propre (marque ou modèle), ni nom générique (véhicule, camion, camionnette, etc.).

5.2 Synthèse

La revue des formes prises par les entités montre que celles-ci sont très diverses, à la fois pour chaque entité et entre types d'entités. Ce panorama nous a semblé indispensable pour saisir précisément le besoin de l'analyse criminelle, qui dans l'idéal ne peut se limiter au repérage des formes des entités nommées normées, comme le TAL a l'habitude de traiter. Les formes des entités rencontrées sont propres aux auditions de témoin, et nous avons identifié plusieurs facteurs d'influence, selon la focalisation des propos et la

qualité des souvenirs du témoin. Prendre en compte toutes les formes de mentions d'entités représente un défi de taille pour l'extraction d'information.

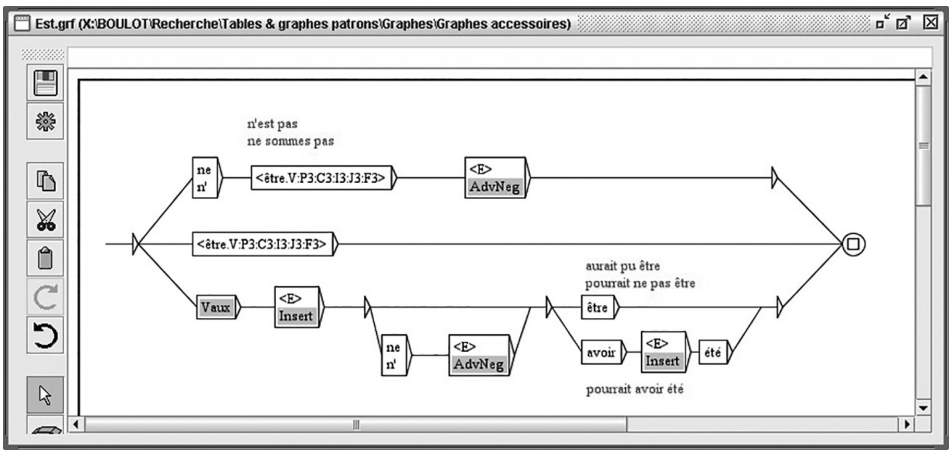
Dans la situation de recherche qui est la nôtre, c'est-à-dire disposant d'un petit corpus de 370 auditions pour environ 600 000 mots et d'une équipe réduite à une seule personne, nous avons dû choisir les formes prises en compte afin de proposer des résultats préliminaires. La reconnaissance d'entités nommées détectant principalement des éléments que l'on peut décrire sous une norme ou un format incluant des critères linguistiques, l'approche de détection s'est limitée aux dates, personnes et lieux sous des formes modélisées que nous décrivons à la section suivante.

6. Approche de détection automatique

Notre corpus est un très petit corpus en comparaison des corpus habituellement traités en TAL. Comme l'ont souligné Nouvel et al. (2015) ainsi que Maurel et al. (2011), dans une telle configuration, les approches symboliques sont privilégiées. C'est donc vers ce type d'approche que nous avons porté notre choix, à l'aide du logiciel de développement de grammaires locales UNITEX (Paumier, 2020).

Le mécanisme des grammaires locales de type automate à états finis a été décrit en détails par Constant (2003) : il s'agit de définir des règles sous forme de graphes. Le texte passe sous forme de suite de tokens 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 1).

Figure 1 : *Exemple de graphe (Paumier, 2020)*



Deux outils sont populaires dans la communauté du TAL pour le développement de grammaires locales : NooJ[24] et UNITEX[25], qui partagent une

base de fonctionnement commune. Nous avons fait le choix d'UNITEX pour des questions de praticité d'installation et de compatibilité de système d'exploitation. Toutefois les performances et possibilités des deux logiciels sont similaires. Le principe est donc de décrire les motifs que l'on recherche, soit via le formalisme proposé par le logiciel (les graphes), soit via l'application de ressources comme des lexiques. Les graphes développés sont disponibles en annexe de Gianola (2020).

Dates. Les dates ont été détectées à l'aide de règles de description entièrement intégrées dans UNITEX, les différents formats de dates pouvant être décrits via un ensemble fini d'éléments agencés en différentes combinaisons. Quatre formats de dates, décomposés en quatre segments (jour de la semaine, jour, mois, année) ont été détectés.

Villes. La détection des villes s'est faite via l'application d'un lexique. Nous avons choisi une liste des noms de communes françaises[26] composée d'environ 36 700 entrées, ce qui concorde avec les données de l'Institut National de la Statistique et des Études Économiques (INSEE, 2018). La liste est adaptée au format des dictionnaires UNITEX puis projetée telle quelle sur le texte.

Personnes. La détection des personnes combine l'utilisation d'un lexique et des règles de description. Le lexique est une liste de plus de 31 000 prénoms déclarés à l'état-civil en France entre 1900 et 2015[27]. Les prénoms sont détectés par la projection de la liste sur le texte, puis l'automate recherche un mot en majuscule le précédant ou le suivant (nom de famille), soit un titre de civilité le précédant (d'après une liste). Les noms d'épouses ou de divorcée ont également été pris en compte.

6.1 Évaluation des résultats

L'évaluation a été réalisée en comparant 10 % du corpus annotés manuellement aux sorties produites par les graphes. Les métriques employées sont les métriques classiques appliquées en recherche d'information : la précision, le rappel et la F-mesure.

La précision évalue la pertinence des entités reconnues :

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

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 :

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

Autrement dit : sur l'ensemble des entités du texte, combien sont correctement reconnues. La F-mesure est la moyenne pondérée combinant rappel et précision[28] :

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

La précision, le rappel et la F-mesure prennent la forme d'un indice inférieur ou égal à 1, 1 correspondant à un score parfait. Un système avec 0.9 de précision, 0.6 de rappel et donc 0.72 de F-mesure est le signe d'un système qui rapporte des informations en grande majorité correctes, mais qui en manque presque la moitié.

Les performances de notre système sont synthétisées dans le tableau 2[29]. Les dates sont les entités les mieux reconnues, il s'agit aussi des entités les plus faciles à modéliser. À l'inverse, les personnes, qui sont difficiles à modéliser sous forme de règles, sont les moins bien reconnues. Enfin, les lieux présentent parfois des ambiguïtés avec d'autres mots, ce qui dégrade leur reconnaissance. D'un point de vue global, les entités sont plutôt bien reconnues et les performances sont équilibrées entre précision et rappel, ce qui est un résultat encourageant pour la poursuite des recherches. Celles-ci devront viser une plus large couverture des ressources, par exemple en incluant des données géographiques étendues et non plus limitées à la France. De plus, il existe un biais dû au fait que le développement et le test des approches ont été réalisés sur les mêmes données. De nouvelles données sont nécessaires pour mettre le système à l'épreuve, et s'assurer que le système est capable de reconnaître une grande diversité de formes des entités criminelles.

Tableau 2 : *Résultats d'évaluation*

Entités	Précision	Rappel	F-mesure	Hypothèse	Référence	Corrects
Personnes	83 %	81,6 %	82,3 %	501	510	416
Dates	98,25 %	100 %	99,1 %	243	235	235
Lieux	88,4 %	83,1 %	85,7 %	597	635	528
Global	**89,88 %**	**88,23 %**	**89,03 %**	**1341**	**1380**	**1179**

7. Conclusion

L'engouement de l'époque pour les technologies de l'intelligence artificielle fait peser de lourdes attentes sur les développements applicables au cas de l'enquête judiciaire. Le fantasme d'une sorte de logiciel tout puissant, solution miracle dépassant les capacités humaines, revient régulièrement dans les médias lorsqu'une nouvelle technologie est testée dans une affaire[30]. Il faut tempérer ces attentes et conserver un regard critique à la fois en ce qui concerne la conception de ces solutions, leurs capacités réelles, et les modalités d'analyse de leurs résultats.

À travers l'étude du cas de l'analyse criminelle au PJGN, nous avons pu découvrir les spécificités d'un genre textuel et d'une pratique professionnelle qui a besoin de solutions de gestion de l'information textuelle adaptées. Cette première étape a permis d'établir la singularité du matériau à exploiter et de définir les ressources et conditions nécessaires au développement de ces solutions. Il faudra notamment envisager la mise à disposition de plus grandes quantités de données d'exemple afin de tester des approches statistiques ou de

prouver la pertinence de l'approche symbolique développée. Les annotations produites par les graphes peuvent aussi être réinvesties comme exemples pour alimenter un système statistique, à condition d'avoir de nouvelles données de test sur lesquelles appliquer les algorithmes. À « l'autre bout de l'entonnoir », nous espérons que nos travaux pourront aussi contribuer à la réflexion sur la conduite de l'audition. La disponibilité des données est donc un enjeu crucial sans lequel les recherches ne pourront se poursuivre. Nous estimons qu'il sera aussi nécessaire de définir un cadre de mise à disposition entre le monde académique et le monde de la justice qui garantisse à la fois le secret de l'instruction et de bonnes conditions de recherche. Il parait nécessaire pour cela d'intégrer les recherches en reconnaissance d'entités nommées pour l'analyse criminelle à des projets de recherche solides et de bonne envergure.

Bibliographie

Allauzen, A., & Schütze, H. (2018). Apprentissage profond pour le traitement automatique des langues. *Revue Traitement Automatique des Langues, 59*, 7-14. Repéré à https://www.atala.org/content/introduction-4 (21 décembre 2020).

Arulanandam, R., Savarimuthu, B. T. R., & Purvis, M. A. (2014). Extracting Crime Information from Online Newspaper Articles. Dans *Proceedings of the Second Australasian Web Conference - Volume 155* (pp. 31–38). Darlinghurst, Australia, Australia : Australian Computer Society, Inc. https://doi.org/10.5555/1123098.1123138

Banga, R., & Mehndiratta, P. (2017). Authorship attribution for textual data on online social networks. Dans *2017 Tenth International Conference on Contemporary Computing (IC3)* (pp. 1-7). https://doi.org/10.1109/IC3.2017.8284311

Berra, A. (2015). Pour une histoire des humanités numériques. *Critique, n° 819-820(8)*, 613-626. https://doi.org/10.3917/criti.819.0613

Cadiot, P. (1991). A la hache ou avec la hache ? Représentation mentale, expérience située et donation du réfèrent. *Langue française, 91*(1), 7-23. https://doi.org/10.3406/lfr.1991.6202

Cappeau, P., & Schnedecker, C. (2018). Du degré de généralité des noms d'humains (pluriels) gens, hommes, humains, individus, particuliers, personnes : différences distributionnelles, sémantiques et génériques. *Langue francaise, N° 198*(2), 65-82. https://doi.org/10.3917/lf.198.0065

Carnaz, G., Beires Nogueira, V., Antunes, M., & Ferreira, N. (2020). An Automated System for Criminal Police Reports Analysis. Dans A. M. Madureira, A. Abraham, N. Gandhi, C. Silva, & M. Antunes (Éds), *Proceedings of the Tenth International Conference on Soft Computing and Pattern Recognition (SoCPaR 2018)* (Vol. 942, pp. 360-369). Porto,Portugal : Springer International Publishing. https://doi.org/10.1007/978-3-030-17065-3_36

Chau, M., Xu, J. J., & Chen, H. (2002). Extracting Meaningful Entities from Police Narrative Reports. Dans *Proceedings of the 2002 Annual National Conference on Digital Government Research* (pp. 1–5). Digital Government Society of North America. Repéré à http://dl.acm.org/citation.cfm?id=1123098.1123138 (21 décembre 2020).

Citton, Y. (2015). Humanités numériques 3.0. *Multitudes, N° 59*(2), 169-180. https://doi.org/10.3917/mult.059.0169

Constant, M. (2003). *Grammaires locales pour l'analyse automatique de textes : méthodes de construction et outils de gestion.* Thèse de doctorat. Université Paris-Est. Repéré à https://tel.archives-ouvertes.fr/tel-00626252 (21 décembre 2020).

Cusson, M., & Cordeau, G. (1994). Le crime du point de vue de l'analyse stratégique. Dans *Traité de criminologie empirique* (Les Presses de l'Université de Montréal, pp. 91-112). Montréal.

Dacos, M. (2010). Manifeste des Digital humanities [Billet]. *THATCamp Paris*. Repéré à https://tcp.hypotheses.org/318 (21 décembre 2020).

Dacos, M., & Mounier, P. (2015). *Humanités numériques : État des lieux et positionnement de la recherche française dans le contexte international.* Institut français. Repéré à https://hal.archives-ouvertes.fr/hal-01228945 (21 décembre 2020).

Dubois, D. (1997). *Catégorisation et cognition : de la perception au discours* (Éditions Kimé). Paris. Repéré à http://www.cairn.info/categorisation-et-cognition-de-la-perception-au-di—978284174101X.htm (21 décembre 2020).

Ehrmann, M. (2008). *Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation.* Thèse de doctorat. Université Paris 7 – Denis Diderot. Repéré à https://hal.archives-ouvertes.fr/tel-01639190 (21 décembre 2020).

Gianola, L. (2020). *Aspects textuels de la procédure judiciaire exploitée en analyse criminelle et perspectives pour son traitement automatique.* Thèse de doctorat. Université de Cergy-Pontoise. Repéré à https://tel.archives-ouvertes.fr/tel-02522680 (21 décembre 2020).

Grouin, C. (2013). *Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique.* Thèse de doctorat. Université Paris 6. Repéré à https://tel.archives-ouvertes.fr/tel-00848672 (21 décembre 2020).

Heiden, S., Magué, J.-P., & Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. Dans *JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data* (p. 12 p.). Rome, Italie : ENS-Lyon. Repéré à http://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf (21 décembre 2020).

INSEE (2018). *Tableaux de l'économie française - Villes et communes de France.* Rapport. Repéré à https://www.insee.fr/fr/statistiques/3303318?sommaire=3353488 (21 décembre 2020).

Jacques, M.-P., & Aussenac-Gilles, N. (2006). Variabilité des performances des outils de TAL et genre textuel. Cas des patrons lexico-syntaxiques. *Traitement Automatique des Langues, 47*(1), 22. Repéré à https://www.atala.org/content/variabilit %C3 %A9-des-performances-des-outils-de-tal-et-genre-textuel (21 décembre 2020).

Juola, P. (2008). Authorship Attribution. *Foundations and Trends® in Information Retrieval, 1*(3), 233-334. https://doi.org/10.1561/1500000005

Kleiber, G. (1981). *Problèmes de référence: descriptions définies et noms propres.* Paris, 1981.

Ku, C. H., Iriberri, A., & Leroy, G. (2008). Crime Information Extraction from Police and Witness Narrative Reports. Dans *2008 IEEE Conference on Technologies for Homeland Security* (pp. 193-198). Waltham, MA, USA : IEEE. https://doi.org/10.1109/THS.2008.4534448

Lafon, P., & Salem, A. (1983). L'inventaire des segments répétés d'un texte. Mots. *Les langages du politique, 6*(1), 161-177. https://doi.org/10.3406/mots.1983.1101

Lebart, L., & Salem, A. (1994). *Statistique Textuelle.* Paris : Dunod. Repéré à http://lexicometrica.univ-paris3.fr/livre/st94/st94-tdm.html (21 décembre 2020)

Léon, J. (2015). *Histoire de l'Automatisation des Sciences du Langage.* Lyon : Ecole Normale Supérieure.

Locker, A. (2019). "Because the computer said so!" *Journal of Language Works -*

Sprogvidenskabeligt Studentertidsskrift, 4(1), 23-37. Repéré à https://tidsskrift.dk/lwo/article/view/115710 (21 décembre 2020).

Maurel, D., Friburger, N., Antoine, J.-Y., Eshkol, I., & Nouvel, D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement Automatique des Langues, 52*(1), 69-96. Repéré à https://www.atala.org/content/cascades-de-transducteurs-autour-de-la-reconnaissance-des-entit %C3 %A9s-nomm %C3 %A9es (21 décembre 2020).

Mayaffre, D. (2002). *Les corpus réflexifs : entre architextualité et hypertextualité. Corpus,* (1). https://doi.org/10.4000/corpus.11

Nadeau, D., & Sekine, S. (2007). *A survey of named entity recognition and classification. Linguisticae Investigationes, 30*(1), 3-26. https://doi.org/10.1075/li.30.1.03nad

Nouvel, D., Ehrmann, M., & Rosset, S. (2015). *Les entités nommées pour le traitement automatique des langues.* ISTE éditions. Repéré à https://iste-editions.fr/products/les-entites-nommees-pour-le-traitement-automatique-des-langues (21 décembre 2020).

Paumier, S. (2020). Unitex 3.2 manuel d'utilisation. Repéré à https://unitexgramlab.org/releases/3.2/man/Unitex-GramLab-3.2-usermanual-fr.pdf (21 décembre 2020).

Pincemin, B. (2018). Sept logiciels de textométrie. Repéré à ttps://halshs.archives-ouvertes.fr/halshs-01843695 (21 décembre 2020).

Poibeau, T. (2014a). La linguistique est-elle soluble dans la statistique ? *Revue Sciences/Lettres*, (2). https://doi.org/10.4000/rsl.402

Poibeau, T. (2014b). Le traitement automatique des langues pour les sciences sociales. *Reseaux, N° 188*(6), 25-51. https://doi.org/10.3917/res.188.0025

Rastier, F. (s.d.). Enjeux épistémologiques de la linguistique de corpus. Repéré à http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html (21 décembre 2020).

Ribaux, O. (2014). *Police scientifique : Le renseignement par la trace* (1re éd.). Lausanne : PPUR.

Rossy, Q. (2011). Méthodes de visualisation en analyse criminelle: approche générale de conception des schémas relationnels et développement d'un catalogue de patterns. Thèse de doctorat. Université de Lausanne, Faculté de droit et des sciences criminelles. Repéré à https://serval.unil.ch/notice/serval:BIB_1AC0D89CA5A4 (21 décembre 2020).

Salem, A. (1986). Segments répétés et analyse statistique des données textuelles. *Histoire & Mesure, 1*(2), 5-28. https://doi.org/10.3406/hism.1986.1518

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Dans *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK.

Schraagen, M., Brinkhuis, M., & Bex, F. (2017). Evaluation of Named Entity Recognition in Dutch online criminal complaints. C*omputational Linguistics in The Nederlands journal, 7*, 3-16. Repéré à http://dspace.library.uu.nl/handle/1874/356185 (21 décembre 2020).

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford University Press.

Valette, M. (2016). Analyse statistique des données textuelles et traitement automatique des langues. Une étude comparée. Dans D. Mayaffre, C. Poudat, L. Vanni, V. Magri, & P. Follette (Éds), *International Conference on Statistical Analysis of Textual Data (JADT2016)* (Vol. 2, pp. 697-706). Nice, France. Repéré à https://hal-inalco.archives-ouvertes.fr/hal-01335084 (21 décembre 2020).

Notes

1 Pour un historique détaillé du TAL et des sciences du langage, voir Léon (2015).

2 Par contraste avec l'analyse criminelle stratégique dont l'objectif est l'action de sécurité et l'anticipation de faits, l'analyse criminelle opérationnelle est pratiquée dans le cadre de la police judiciaire et de l'enquête (Cusson et Cordeau (1994) ; (Ribaux, 2014, p. xxvi)).

3 https ://www.ibm.com/fr-fr/marketplace/analysts-notebook (Consulté le 14 octobre 2020).

4 La recherche d'un mot ou d'un motif dans du texte.

5 Du texte non-structuré et rédigé sans contraintes de forme.

6 Statistique fournie par le logiciel TXM (Heiden, Magué, & Pincemin, 2010).

7 L'étiquetage morpho-syntaxique est une tâche consistant à identifier la catégorie morpho-syntaxique d'un mot (adjectif, nom, verbe, etc.) dans une phrase. Voir : http://www.technolangue.net/article.php3?id_article=296 (Consulté le 31 juillet 2020).

8 Pour un exemple, voir Gianola (2020, p. 36).

9 http://www.lexi-co.com/ (Consulté le 30 juillet 2020).

10 Un token est une suite de caractères comprise entre deux espaces ou signes de ponctuation.

11 Hormis en ce qui concerne l'audition de mineurs victimes.

12 Considérer les sciences criminelles comme domaine d'application peut faire l'objet d'un débat connecté 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.

13 http://www.iramuteq.org (Consulté le 3 août 2020).

[14] http://www.laurenceanthony.net/software/antconc/ (Consulté le 3 août 2020).

[15] http://ancilla.unice.fr/ (Consulté le 3 août 2020).

[16] Par contraste avec les « langages machine », les langages de programmation informatique.

[17] Nous nous tenons ici à la définition de la référence pour le TAL. Il faut néanmoins souligner que cette notion fait l'objet de nombreux débats en linguistique et plus particulièrement en analyse du discours. Voir : Kleiber (1981), Tamba (1983), Cadiot (1991), Dubois (1997).

[18] https://opennlp.apache.org/ (Consulté le 15 octobre 2020).

[19] « crime-specific lexicon »

[20] GATE https://gate.ac.uk/sale/tao/split.html (Consulté le 9 septembre 2020) est une plateforme de gestion textuelle permettant de réaliser de l'annotation et divers traitements linguistiques.

[21] Les expressions régulières sont des moyens de décrire minimalement un motif à reconnaître dans une chaîne de caractères.

[22] Nous rappelons que les exemples présentés ont été pseudonymisés, et qu'ils ne mentionnent donc pas de noms de personnes réelles.

[23] Exemples de noms génériques d'humains : personne, femme, bonhomme, vieillard, enfant, etc.

[24] http://www.nooj-association.org/index.html (Consulté le 9 octobre 2020).

[25] https://unitexgramlab.org/fr (Consulté le 9 octobre 2020).

[26] Disponible à l'adresse : https://sql.sh/736-base-donnees-villes-francaises (Consulté le 9 octobre 2020).

[27] Disponible à l'adresse : https://www.insee.fr/fr/statistiques/2540004#consulter (Consulté le 9 octobre 2020).

[28] Selon la finalité de l'application, on peut vouloir privilégier l'une ou l'autre des mesures.

[29] Pour un tableau détaillé des résultats, consulter Gianola (2020, p. 117-118).

[30] Voir à ce sujet deux épisodes de « rebondissements » dans l'enquête sur la mort de Grégory Villemin en France :

– En juin 2017, l'utilisation d'un logiciel appelé « Anacrim » dans l'interpellation de suspects après plusieurs années de point mort est mise en avant par les médias. Il n'existe pas de logiciel « Anacrim », ce terme étant en réalité un diminutif pour désigner la méthode de l'analyse criminelle. Quant au logiciel utilisé pour faire progresser l'enquête, il s'agit d'Analyst's Notebook. Voir : « L'intelligence artificielle s'insinue dans l'analyse criminelle », Le Monde, 17 juillet 2017, https://www.lemonde.fr/sciences/article/2017/07/17/l-intelligence-artificielle-s-%0Ainsinue-dans-l-analyse-criminelle_5161548_1650684.html (Consulté le 20 décembre 2020), et « Expliquez-nous... Les logiciels d'analyses criminelles Anacrim et Salvac », France Info, 30 mars 2018, https://www.francetvinfo.fr/replay-radio/expliquez-nous/expliquez-nous-les-logiciels-d-analyses-criminelles-anacrim-et-salvac_2658774.html (Consulté le 20 décembre 2020).

– En décembre 2020, on annonce les résultats d'une analyse des courriers envoyés par le corbeau via une méthode de stylométrie pratiquée par une entreprise suisse. La stylométrie est une approche d'analyse bien connue en linguistique basée sur des calculs statistiques. Plus fréquemment appliquée dans le domaine judiciaire chez les anglo-saxons (sous le nom de authorship analysis ou authorship attribution (Juola (2008), Banga et Mehndiratta (2017), Locker (2019)), elle reste toutefois dépendante de l'analyse humaine de ses résultats. Le fait que l'analyse ait été pratiquée par une entreprise privée doit également nous inciter à la prudence : l'entreprise ne dévoile pas sa méthodologie, qui reste donc opaque et ne permet pas d'évaluer les éventuels biais. Voir : « Affaire Grégory : "On n'obtiendra pas une réponse irréfutable et incontestable avec la stylométrie" », Sciences et Avenir, 19 décembre 2020, https://www.sciencesetavenir.fr/high-tech/affaire-gregory-on-n-obtiendra-pas-une-reponse-irrefutable-et-incontestable-avec-la-stylometrie_150279 (Consulté le 20 décembre 2020).