TAL et linguistique textuelle pour l'extraction d'information en analyse criminelle

Lucie Gianola
lucie.gianola@yahoo.fr

Université de Cergy-Pontoise
LIMSI - Université Paris-Sud

2 octobre 2020

Table des matières

Réflexions sur le positionnement épistémologique pour la recherche appliquée aux sciences criminelles

Fiction policière & Sciences criminelles

Attentes

- Laboratoire
- ADN
- Puissance technologique
- ...

Réalité

- Longueur
- Lourdeur
- Manque de moyen humains et financiers
- Pression médiatique
- ...

Forensic Science : Last Week Tonight with John Oliver (HBO, octobre 2017)

https://www.youtube.com/watch?v=ScmJvmzDcG0
(extrait présenté à partir de 5 minutes 30 environ)

L'effet "Les experts"

La réalité inspire la fiction
La fiction influence la réalité

L'effet "Les experts"

ANACRIM et l'affaire Grégory Villemin

Affaire du petit Grégory : «AnaCrim», le super logiciel qui a aidé les gendarmes

Comment fonctionne Anacrim ?

C'est d'abord un travail méthodique de longue haleine pour les analystes criminels. Il s'agit de relire pièces à pièces tous les procès-verbaux rédigés dans le cadre d'un dossier judiciaire et d'en retenir les éléments les plus utiles pour les enquêteurs. Dans chaque document, les analystes vont retranscrire minutieusement tous les éléments constatés par les enquêteurs sur le terrain ou les détails figurant dans une audition.

Cela peut être la précision d'une conversation, le lieu et l'heure où un témoin déclare avoir été etc. Toutes ces informations sont alors versées dans une importante base de données qui replace tous les éléments d'une enquête dans le temps et dans l'espace. L es gendarmes formés «AnaCrim», ne travaillent que sur une affaire à la fois ce qui permet de se concentrer sur toutes les données accumulées au fil du temps. Ils sont de «véritables experts du traitement de l'information criminelle.»

Peut-on s'en passer ?

C'est désormais une aide fondamentale pour les enquêteurs sur les dossiers d'homicides notamment. Il est devenu un outil indispensable. Le cerveau humain a ses limites et n'est pas toujours capable d'analyser de manière objective des faits parfois anodins. AnaCrim permet surtout de ne pas passer à côté d'une hypothèse de travail.

Colonel Didier Berger, Le Parisien, 14 juin 2017
https://frama.link/GdbWT7zs

L'effet "Les experts"

Durnal 2010 *Crime scene investigation (as seen on TV)*

Because of 'CSI' shows, some prosecutors contend that more jurors believe every crime scene yields forensic evidence that offers conclusive scientific proof of innocence or guilt, almost instantly.

[...] Shows such as CSI [...] create and thus perpetuate a vast number of forensic science myths. [...]. One forensic scientist estimates that 40% of the science on CSI does not exist. [...]

Fiabilité

- Empreintes digitales : Brandon Mayfield et les attentats de Madrid (Ribaux 2014, p. 30-31)
- ADN : le "fantôme" d'Heilbronn (Ribaux 2014, p. 115)

Positionnement du chercheur

Le chercheur appliquant ses compétences acquises par ailleurs dans le domaine des sciences criminelles doit se défaire de ses préjugés acquis par le biais de la fiction policière afin de garantir la qualité de ses analyses et travaux.

Analyse criminelle

Qu'est-ce qu'une enquête ?

Processus de collecte, d'organisation et d'interprétation d'**éléments matériels** et **informationnels** afin de résoudre des faits répréhensibles

Définition

Discipline d'appui aux enquêtes dont l'objectif est de retrouver, reprendre et synthétiser l'information dans un dossier de procédure judiciaire afin de proposer de nouvelles pistes d'investigation

Passer de ça...

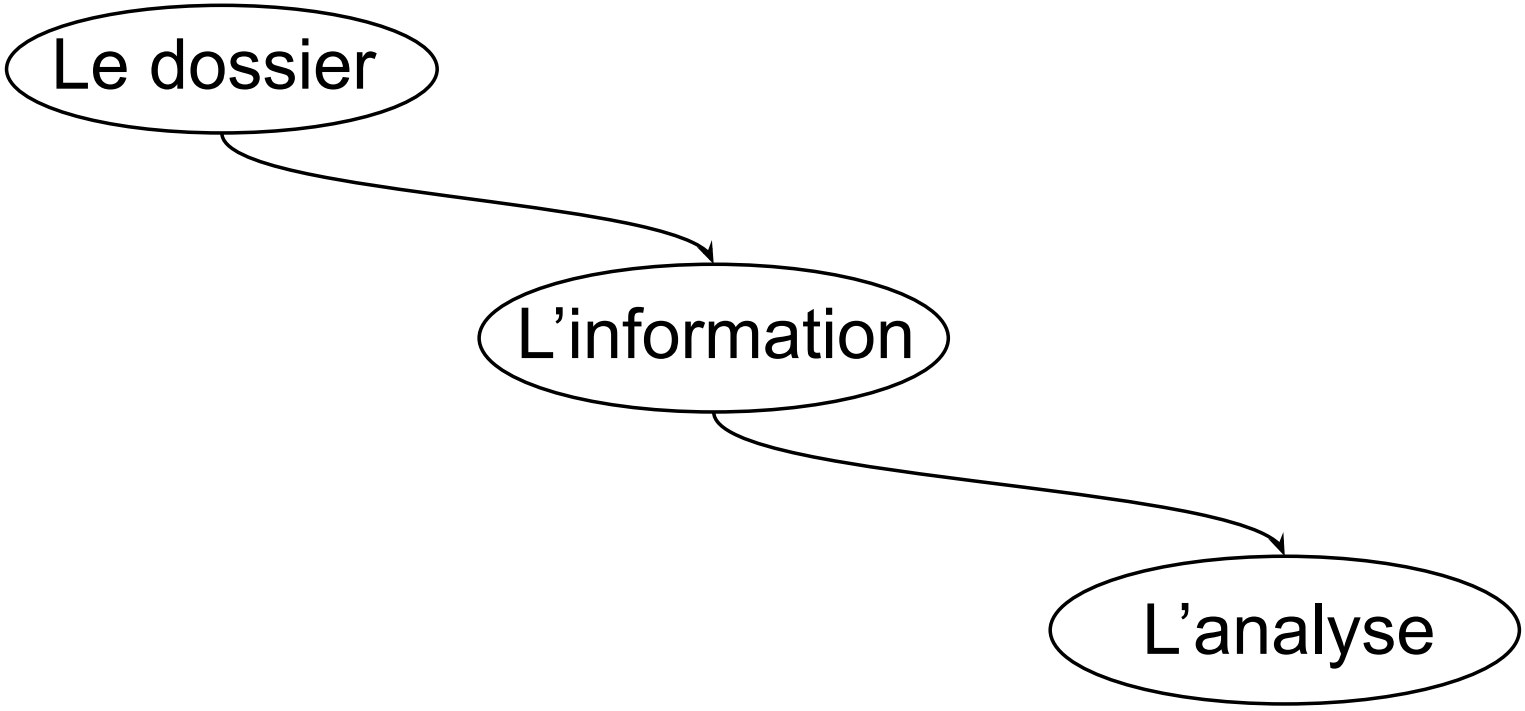

Esselte
DISPARITION 63
AUDITIONS
Esselte
DISPARITION 63
AUDITIONS
Esselte
ARCHIVES
D.63
PV 632 8/08
PV 866/10
Esselte
DISPARITION 63
- INVESTIGATIONS
- PERQUISITIONS

...à ça

Schéma chronologique.

Voir Gianola 2020, p. 11

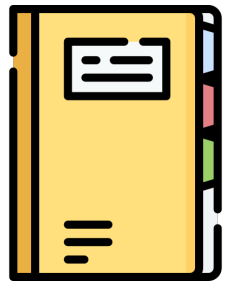
Principe de l'analyse criminelle



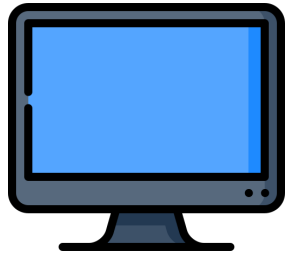
Problématique

Un humain

Un dossier



Des outils informatiques



Objectifs

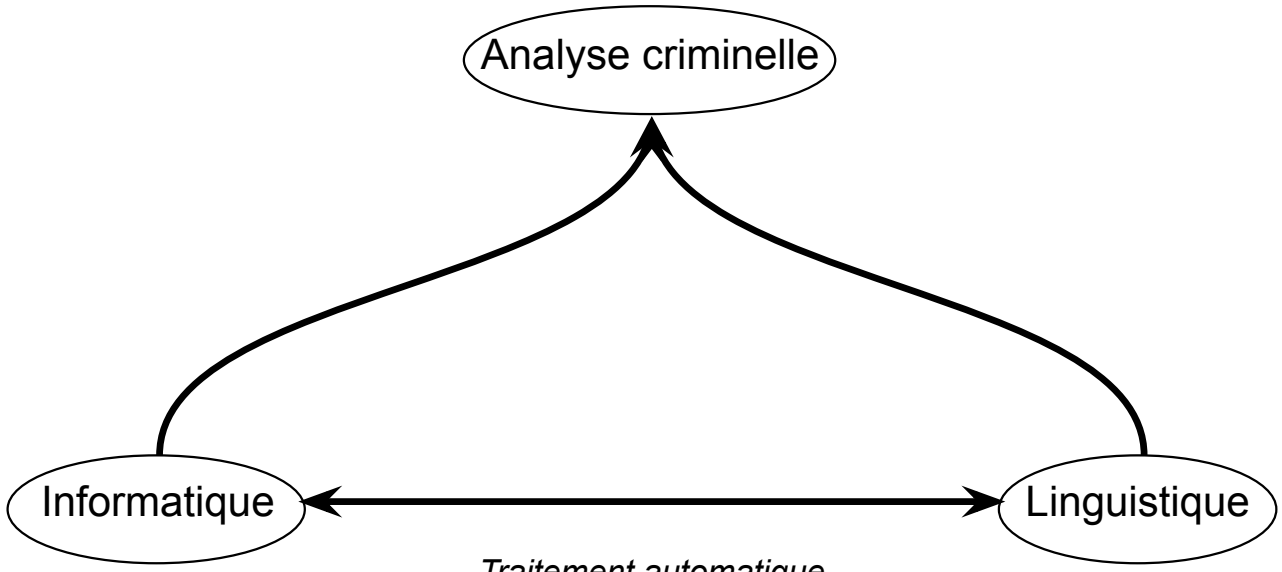
Informatique :

Repérer automatiquement certaines informations

Linguistique :

Concevoir un outil d'exploration *ad hoc*

Articulation épistémologique


Analyse criminelle
Informatique
Linguistique
Traitement automatique des langues (TAL)

Le dossier de procédure judiciaire

Documente et compile le déroulement de l'enquête

Deux types de documents :

- Réglementaires et procédure
- Informations

Documents réglementaires et de procédure

Garantissent la légalité et la conformité des opérations

- Réquisitions, bordereau d'envoi, inventaire de pièces, PV de notification de garde à vue, PV de saisie, certificats médicaux, etc.

Documents d'information

Convoient des informations sur les faits

- PV transport constatations mesures prises, PV d'actes d'enquête (investigations, renseignements, perquisition...), **PV d'audition de témoin et de garde à vue**, PV de synthèse, rapports d'expertises, retours de réquisitions (téléphoniques, bancaires, péages, assurances...), documents graphiques (photos, vidéos)

PV Transport constatations mesures prises

Relate l'arrivée des forces de l'ordre sur la scène de crime, décrit la scène et ses conditions (date, horaires, météo, topographie, corps du délit, personnes présentes, etc.) ainsi que les premières opérations menées

Voir Gianola 2020, p. 31-32

Rapports d'expertise

Synthèse d'analyses diverses ordonnées par le juge (toxicologiques, biologiques, balistiques, etc.)

Voir Gianola 2020, p. 39

Rapports d'autopsie

Description de l'état médical du cadavre dans le cas des homicides (y compris tenue vestimentaire)

Voir Gianola 2020, p. 40-41

Documents téléphoniques et bancaires

Extraits de comptes bancaires

Factures téléphoniques détaillées (”fadet”) comportant nature, durée, sens des communications, antenne activée, numéro de série du boîtier téléphonique utilisé

Voir Gianola 2020, p. 41-42

PV de renseignement

Relate des opérations menées par les enquêteurs : saisie, prélèvement, perquisition, porte à porte, battue, informations recueillies hors du cadre de l'audition, etc.

Voir Gianola 2020, p. 37

Les auditions de témoin

Voir Gianola 2020, p. 36

Les auditions de témoin

Compte-rendu écrit d'un échange oral entre un témoin et un enquêteur

- Volume, quantité d'information
- Expression en langage naturel
- Récit

Le genre des auditions

Régularité

85 segments de onze tokens répétés dix fois ou plus

Lafon and Salem 1983, Lebart and Salem 1994

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
agit de la personne découverte sans vie dans la forêt ce	46
Je prends connaissance du motif pour lequel ma déposition est requise	39
Nous vous montrons une série de 06 photographies de couteaux référencée	31
Avez vous des connaissances qui vont dans cette forêt de XXX	30

Narrativité

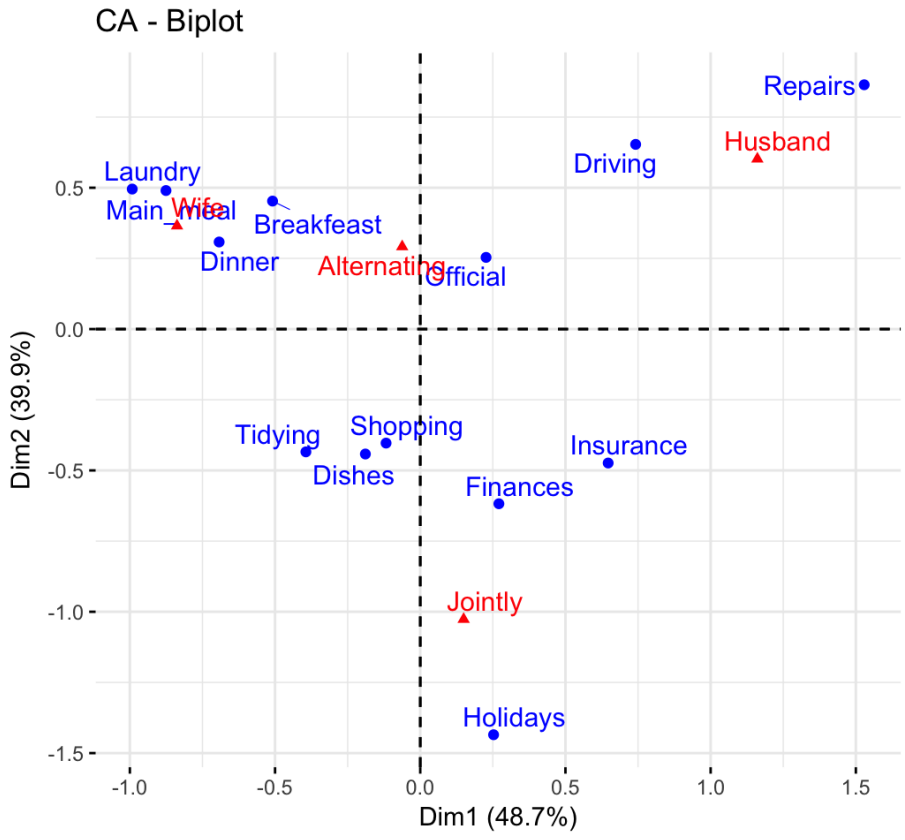
Étude des étiquettes morpho-syntaxiques
Schmid 1994

- Haute fréquence des pronoms personnels
- Temps de verbes : présent, participes passés et imparfait sont les plus fréquents

Éude contrastive des auditions (AFC)

Rappel sur l'AFC

	Wife	Alternating	Husband	Jointly
Laundry	156	14	2	4
Main_meal	124	20	5	4
Dinner	77	11	7	13
Breakfeast	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



Exemple tiré de 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/

Étude contrastive des auditions (AFC)

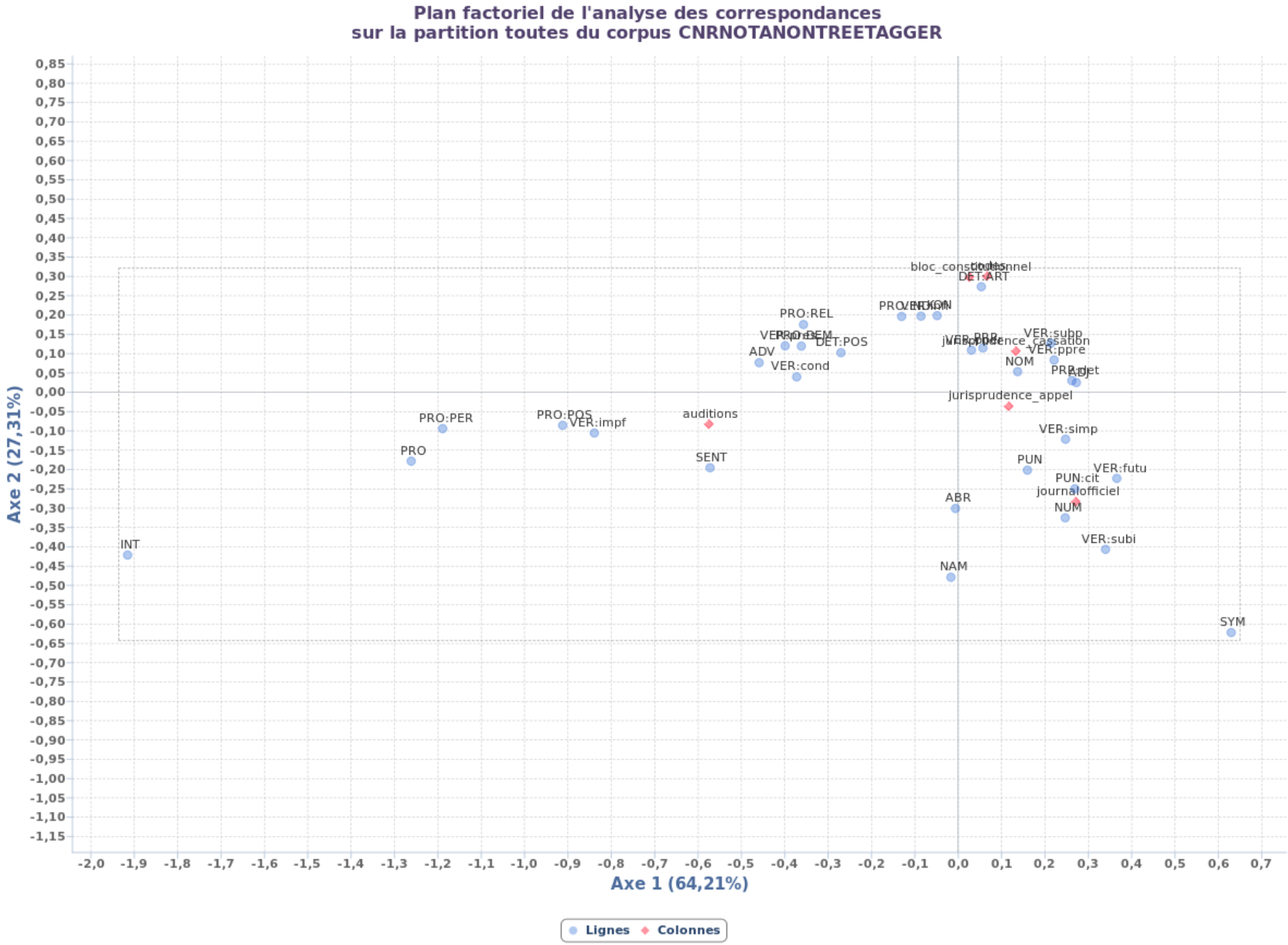
Corpus et statistiques

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
Total	764	**2 828 000**	3 700

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:pper	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

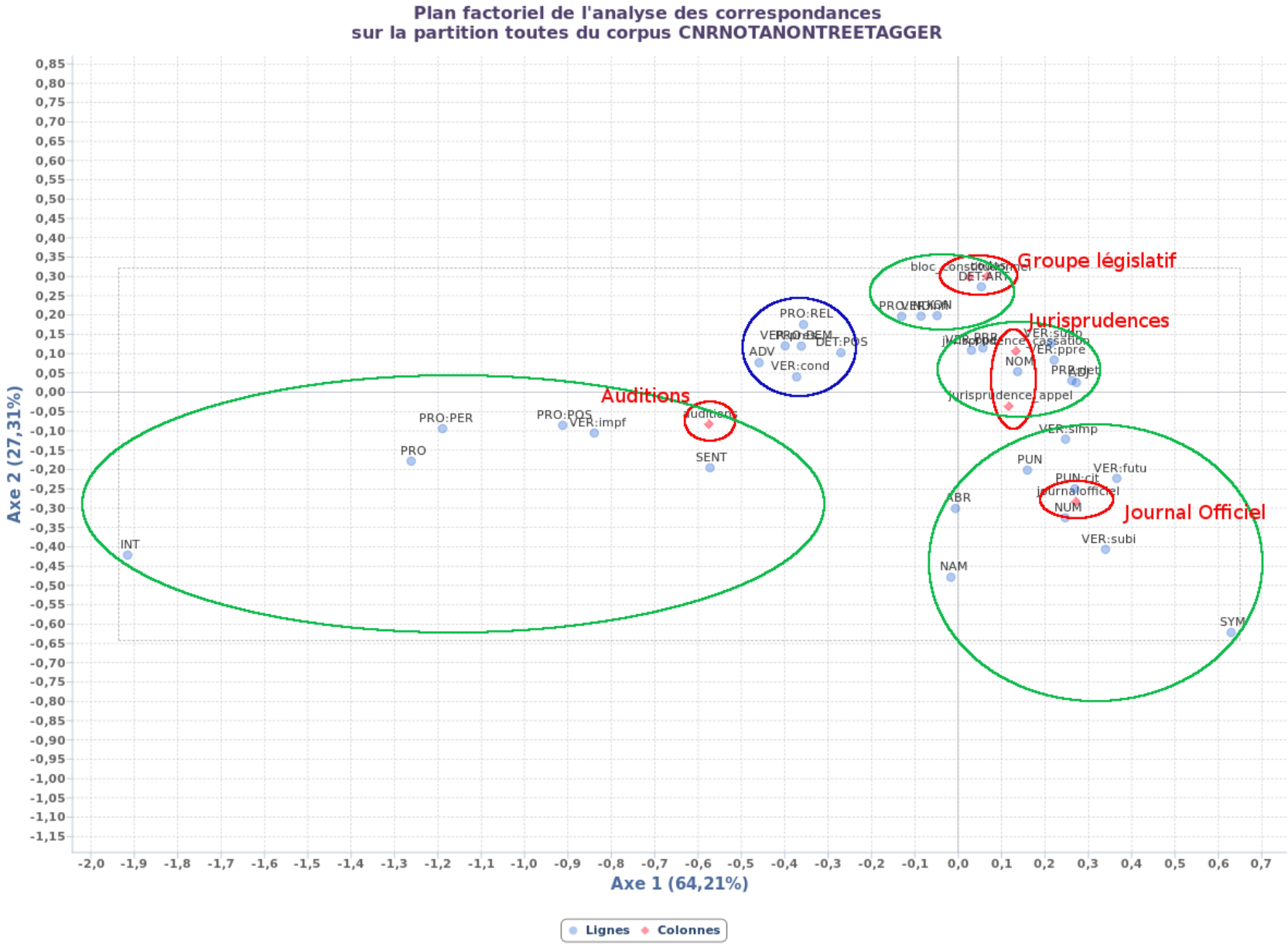
Étude contrastive des auditions (AFC)

TXM, Heiden & *al.* (2010)



Étude contrastive des auditions (AFC)

TXM, Heiden, Magué, and Pincemin 2010



Les entités

Entités criminelles

Un élément ou un paramètre du monde réel mentionné au cours de l'enquête

- Une personne
- Un lieu
- Une date
- Un numéro de téléphone
- Un véhicule
- Une organisation
- Une trace
- Un objet
- etc.

(Rossy 2011, Ribaux 2014)

Entités nommées

Absence de consensus sur la définition

Entités nommées	Descriptions définies
M. Dupont	Le collègue de Pierre
Jacques Durand	La présidente d'Estonie

Dates, lieux, événements, montants, quantités...

(Ehrmann 2008, Nouvel, Ehrmann, and Rosset 2015)

Du point de vue linguistique

Noms propres, descriptions définies, dénominations génériques

Marc Gianola, Solange, Mme Charron

le cousin de la voisine, le collègue de Pierre

un homme, une femme, un individu, un vieillard

(Kleiber 1981, Cappeau and Schnedecker 2018)

Les entités dans les auditions de témoin

Personnes

- Nous avons rencontrés M. et Mme FERRAND et nous avons emménagé dans la maison environ un mois après en novembre 2001.

- 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.

Les entités dans les auditions de témoin

Éléments temporels

- Je l'ai eu au téléphone le 10 ou le 11 septembre 2009 avant le week-end.
- QUESTION : Est ce que vous travailliez le jour de la mort de Dominique ?
- Cette conversation a eu lieu en début d'après-midi. Je dirais vers 14H30-15H00.
- 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 [...]

Les entités dans les auditions de témoin

Éléments géographiques

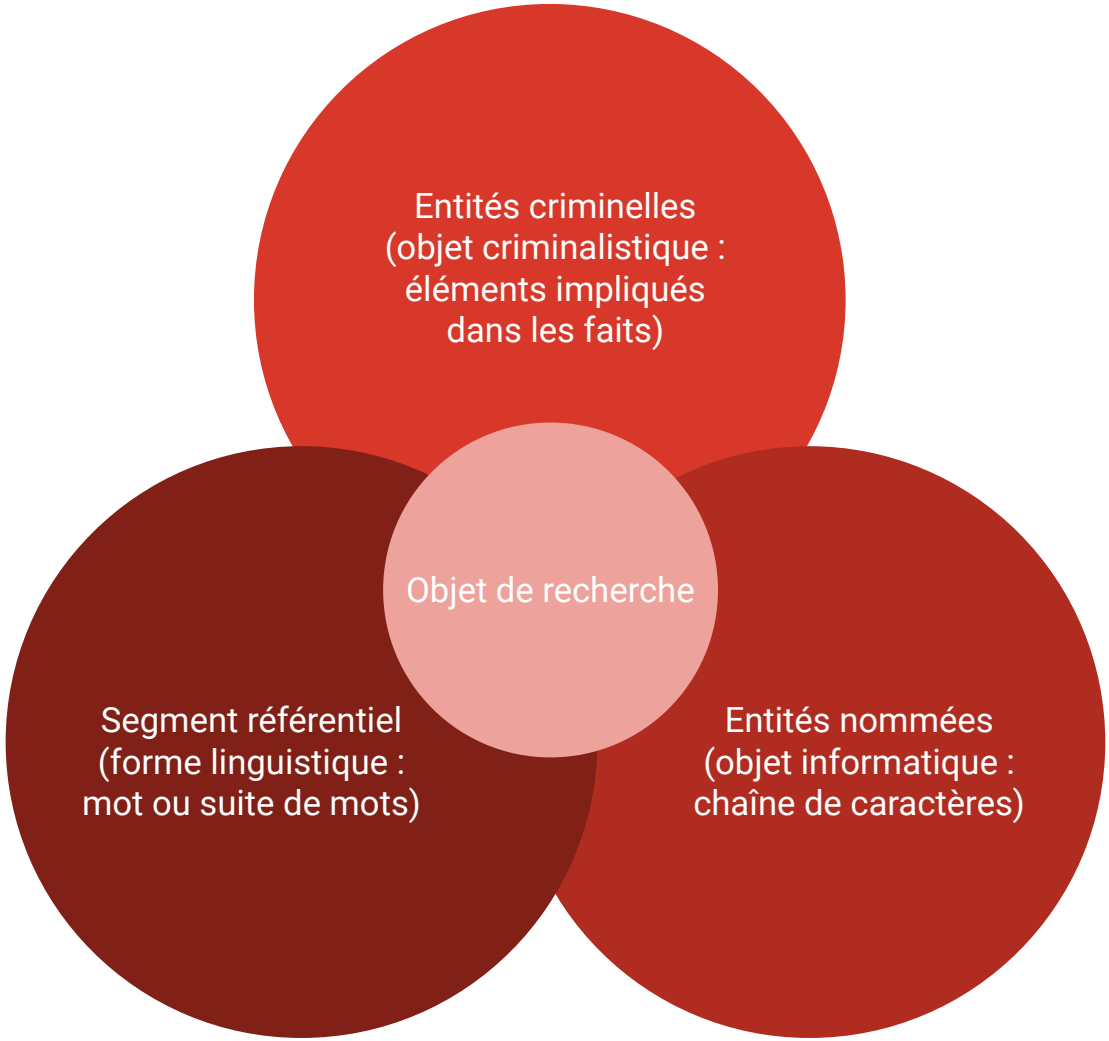
- Sarah habite prés de CAEN et Jessica habite en Bretagne.
- Je sais qu'il se promène en forêt du côté de l'hôpital vers la route de Launaguet.
- Je prends connaissance de l'objet de votre enquête suite à la personne décédé qui a été découverte sur un parcours sportif en forêt de MONTMORENCY dans la journée du samedi 12 mars 2010.

Les entités dans les auditions de témoin

Véhicules

- 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.
- Je me rends tous les jours 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 des 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.

Entre linguistique, informatique et analyse criminelle



Expérience de détection automatique d'entités criminelles

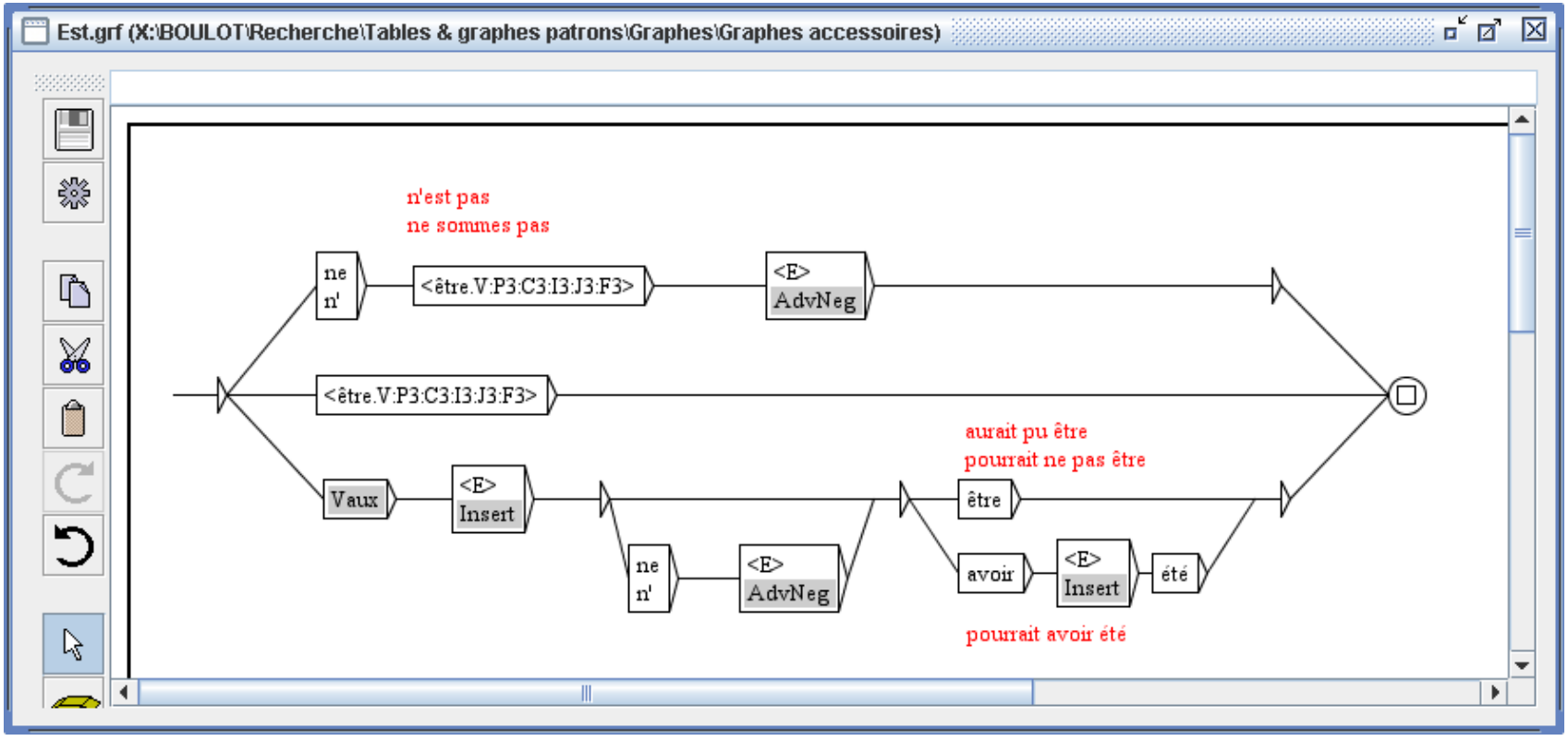
Situation de recherche

Un dossier exploitable compilant 370 auditions

Trois types d'entités retenus

Approche de détection

Grammaires locales



UNITEX (Paumier 2016)

Approche de détection

Adaptation du système de détection à chaque type d'entité

- Lieux : lexique
- Dates : règles
- Personnes : lexique et règles combinés

Résultats

Évaluation

Métriques d'évaluation : précision

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.

Évaluation

Métriques d'évaluation : rappel

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.

Évaluation

Métriques d'évaluation : F-mesure

La F-mesure combine rappel et précision :

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

La précision, le rappel et la F-mesure prennent la forme d'un indice inférieur ou égal à 1. Exemple :

Précision	Rappel	F-mesure
0.9	0.6	0.72

Ces chiffres témoignent d'un système qui reconnaît précisément les éléments mais qui ramène beaucoup de bruit.

Évaluation

10% du corpus annoté manuellement comme référence

Entités	Précision	Rappel	F-mesure	Hypothèse	Référence	Corrects
Personnes	83.0%	81.6%	82.3%	501	510	416
Dates	98.25%	100.0%	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**

Conclusion

Conclusion

Informatique :
Pertinence d'une approche minimale et peu coûteuse

Linguistique :
Exploration et description du genre textuel de l'audition

Limites de l'approche

Peu de données d'exemple

Introduction d'un biais ?

Perspectives

Perspectives

Élaboration d'un cadre de recherche adapté

Réinvestissement des résultats comme données d'entrée

Intégration de la problématique de l'analyse criminelle dans des programmes de recherche d'envergure

Bibliographie

Paul Cappeau and Catherine Schnedecker. "Du degré de généralité des noms d'humains (pluriels) gens, hommes, humains, individus, particuliers, personnes : différences distributionnelles, sémantiques et génériques". In: *Langue francaise* N 198.2 (June 2018), pp. 65–82. ISSN: 0023-8368. URL: http://www.cairn.info/revue-langue-francaise-2018-2-page-65.htm.

Evan W. Durnal. "Crime scene investigation (as seen on TV)". In: *Forensic Science International* 199.1 (June 2010), pp. 1–5. ISSN: 0379-0738. DOI: 10.1016/j.forsciint.2010.02.015. URL: http://www.sciencedirect.com/science/article/pii/S0379073810000678.

Maud Ehrmann. "Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation". PhD thesis. Université Paris 7, 2008. URL: https://hal.archives-ouvertes.fr/tel-01639190.

Lucie Gianola. "Aspects textuels de la procédure judiciaire exploitée en analyse criminelle et perspectives pour son traitement automatique". PhD thesis. Université de Cergy-Pontoise, Feb. 2020. URL: https://tel.archives-ouvertes.fr/tel-02522680.

Serge Heiden, Jean-Philippe Magué, and Bénédicte Pincemin. "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, 2010. URL: http://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf.

Georges Kleiber. *Problèmes de référence: descriptions définies et noms propres*. Centre d'analyse syntaxique de l'Université de Metz, 1981.

Pierre Lafon and André Salem. "L'inventaire des segments répétés d'un texte". In: *Mots. Les langages du politique* 6.1 (1983), pp. 161–177. DOI: 10.3406/mots.1983.1101. URL: https://www.persee.fr/doc/mots_0243-6450_1983_num_6_1_1101.

Ludovic Lebart and André Salem. *Statistique Textuelle*. Paris: Dunod, 1994. URL: http://lexicometrica.univ-paris3.fr/livre/st94/st94-tdm.html.

Damien Nouvel, Maud Ehrmann, and Sophie Rosset. *Les entités nommées pour le traitement automatique des langues*. Science cognitive. Londres: ISTE, 2015. ISBN: 978-1-78406-104-3. URL: https://iste-editions.fr/products/les-entites-nommees-pour-le-traitement-automatique-des-langues.

Sébastien Paumier. *Unitex 3.1 manuel d'utilisation*. 2016. URL: https://unitexgramlab.org/releases/3.2rc/man/Unitex-GramLab-3.2rc-usermanual-fr.pdf.

Olivier Ribaux. *Police scientifique : Le renseignement par la trace*. 1st ed. Lausanne: Presses polytechniques et universitaires romandes, 2014. ISBN: 978-2-88915-061-8.

Quentin Rossy. "Méthodes de visualisation en analyse criminelle: approche générale de conception des schémas relationnels et développement d'un catalogue de patterns". PhD thesis. Université de Lausanne, 2011. URL: https://serval.unil.ch/notice/serval:BIB_1AC0D89CA5A4.

Helmut Schmid. "Probabilistic Part-of-Speech Tagging Using Decision Trees". In: *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, 1994.