Analyse de données textuelles appliquée à des problématiques de sécurité et d'enquête judiciaire

Laura Ascone, Lucie Gianola

Analyse de données textuelles appliquée à des problématiques de sécurité et d'enquête judiciaire

Laura Ascone[1], Lucie Gianola[1]

[1]AGORA, Université de Cergy-Pontoise – laura.ascone@etu.u-cergy.fr, lucie.gianola@u-cergy.fr

Abstract

This presentation investigates two cases of textual analysis applied to security contexts:
- the analysis of the rhetorical strategies adopted in the Islamic State's official online magazines: Dabiq, published in English, and Dar al-Islam, published in French;
- the use of methods for named entities' automatic extraction, and the conception of a textual exploration software for criminal analysis.

Résumé

Nous présentons deux cas d'application de l'analyse de données textuelles dans des contextes liés à la sécurité :
- l'analyse des stratégies rhétoriques de propagande djihadistes à travers l'étude des revues Dabiq et Dar-al-Islam,
- l'utilisation de méthodes d'extraction automatique d'entités nommées et la conception d'un outil d'exploration textuelle pour l'analyse criminelle.

Keywords: analyse de données textuelles, radicalisation, analyse criminelle

1. Introduction

L'essor de préoccupations sécuritaires liées aux actes de terrorisme perpétrés à travers le monde depuis le début du XXIème siècle pousse les chercheurs, acteurs publics et sociaux à rechercher de nouveaux moyens d'analyse de ce phénomène. En France, les sciences humaines et sociales se saisissent de la question comme le démontre l'organisation de plusieurs journées d'études sur la question (« Nouvelles figures de la radicalisation », Toulouse, avril 2017, « Les SHS face à la menace », Cergy, septembre 2017, « Des sciences sociales en état d'urgence : islam et crise politique », Paris, décembre 2017).

Nous souhaitons présenter dans cet article deux sujets d'étude relatifs à ces préoccupations sécuritaires : une étude de la rhétorique de Daesh du point de vue du recours aux émotions dans les revues *Dabiq* (anglais) et *Dar al-Islam* (français), ainsi qu'une collaboration entre le Pôle Judiciaire de la Gendarmerie Nationale (PJGN) et l'Université de Cergy-Pontoise visant à fournir de nouveaux outils d'analyse textuelle des procédures judiciaires aux équipes d'analystes criminels.

Le phénomène de la radicalisation djihadiste a amené chercheurs et professionnels à examiner les raisons psychosociologiques qui sont à la base de l'adhésion à l'idéologie djihadiste (Khosrokhavar, 2014) ainsi que les stratégies adoptées par le groupe extrémiste pour diffuser les messages de propagande (Lombardi, 2015). Toutefois, bien qu'elles jouent un rôle crucial au sein de la propagande djihadiste, les stratégies rhétoriques qui visent à menacer ou à persuader les différents lecteurs restent inexplorées. La première partie de cette étude vise donc à présenter une analyse quanti-qualitative du schéma rhétorique et des émotions sur

lesquels se base la propagande djihadiste. Dans la continuité des travaux de Marchand (2014), les logiciels *Iramuteq* et *Tropes* ont permis d'étudier le corpus d'un point de vue quantitatif. Les résultats issus de cette analyse quantitative ont ensuite constitué le point de départ d'une analyse qualitative sur les extraits exprimant des émotions, afin d'examiner plus en détail les stratégies rhétoriques de la propagande djihadiste.

Le cas de l'analyse des procédures judiciaires nous confronte à une problématique typique d'extraction d'information passant par la reconnaissance automatique d'entités nommées : notre travail de recherche consiste notamment à concevoir les bases d'un outil de navigation textuelle *ad hoc*. Bien que les besoins des analystes criminels soient similaires à ceux d'autres domaines d'application (analyse de la voix du client, traitement automatique de la langue biomédicale, etc.), le contexte de l'enquête judiciaire pose de nouvelles contraintes de précision dans l'extraction et dans la mise à disposition des résultats à l'expert, c'est-à-dire à l'analyste criminel.

Le besoin social et institutionnel de nouvelles approches de documents d'origines variées dans les contextes judiciaires et sécuritaires nous permet de démontrer la pertinence de méthodes d'analyse de données textuelles déjà éprouvées dans ces deux cas d'étude.

2. Description de la rhétorique djihadiste : cas des revues *Dabiq* et *Dar al-Islam*

2.1. Corpus et méthodologie

Cette recherche a été menée sur les deux revues de Daech : *Dabiq*, publié en anglais, et *Dar al*-Islam, publié en français. *Dabiq* s'adresse aux sympathisants non arabophones de Daech, tandis que *Dar al-Islam*, qui n'est pas une traduction de *Dabiq*, s'adresse à un lectorat uniquement francophone. Cette distinction nous conduit à avancer l'hypothèse que les deux revues diffèrent dans leur contenu ainsi que dans la forme du message qu'elles portent. Toutefois, l'une et l'autre s'adressent à un lectorat qui a déjà adhéré à l'idéologie islamiste. Leur objectif n'est donc pas de persuader le lecteur de s'approcher de l'islamisme, mais de renforcer son adhésion et de l'amener à agir au nom de cette idéologie. Afin d'analyser les stratégies rhétoriques du discours jihadiste, une approche quanti-qualitative a été adoptée (Rastier, 2011). Plus particulièrement, cette approche itérative était constituée de quatre étapes. Une première analyse qualitative de l'idéologie djihadiste, du processus de radicalisation et des caractéristiques linguistiques du discours de haine a été essentielle à la compréhension du discours djihadiste ainsi qu'à l'avancement des premières hypothèses. La deuxième étape correspond à une analyse quantitative qui a permis de vérifier les hypothèses avancées : le corpus a donc été examiné avec le logiciel Tropes (Ghiglione *et al*, 1998), qui permet d'analyser un texte d'un point de vue sémantico-pragmatique à partir d'un lexique préétabli, et d'identifier les thèmes les plus récurrents dans le corpus ainsi que la manière dont ces thèmes sont liés l'un à l'autre. Afin d'analyser la manière dont le discours djihadiste arrive à persuader et menacer les différents lecteurs (Giro, 2014), une analyse qualitative a été menée sur les thèmes *sentiment*, pour le corpus français, et *feeling*, pour le corpus anglais (troisième étape). En d'autres termes, l'analyse quantitative a constitué le point de départ pour une étude qualitative, qui a donc été menée sur les énoncés exprimant des émotions et des sentiments (Caffi et Janney, 1994). Enfin, une dernière analyse quantitative a été menée avec le logiciel Iramuteq (Ratinaud et Marchand, 2012) qui, basé sur la méthode Reinart, permet, par exemple, de déterminer le sous- et suremploi de certains termes au sein des différents corpus (quatrième étape). La combinaison d'approches qualitatives et quantitatives a permis

d’examiner de discours djihadiste en relation avec le contexte dans lequel il a été produit (Valette et Rastier, 2006).

2.2. Résultats

L’analyse des énoncés exprimant des émotions et des sentiments dans les deux revues officielles de Daesh a permis de déterminer le schéma rhétorique sur lequel se construit la propagande djihadiste. Puisque l’objectif de *Dabiq* et de *Dar al-Islam* est de manipuler le comportement du lecteur, la propagande de Daech se fonde sur l’imposition d’obligations et d’interdictions. L’accord de récompenses ainsi que le sentiment de culpabilité visent à amener le lecteur à respecter ces indications. En revanche, tout musulman qui ne respecte pas ces indications, subira des conséquences négatives : il sera jugé d’apostat et il sera donc considéré comme un ennemi. On a ici la menace exprimée par Daech contre les musulmans. Les obligations sont exploitées également pour imposer au lecteur une action violente contre l’Occident, justifiée et alimentée par le sentiment de victimisation. Combattre l’ennemi est présenté comme une action héroïque et valorisante. En participant au combat contre l’Occident, le lecteur aura l’impression de devenir un héros qui lutte au nom d’une cause juste et noble (De Bonis 2015), et de voir ses faiblesses disparaître (Rumman, Suliman *et al* 2016). En outre, en citant des versets coraniques concernant la victoire des musulmans, l’auteur assure à son lecteur que la communauté musulmane aura la victoire sur l’ennemi ; l’extrait suivant en est un exemple : « *Allah par vos mains les châtiera, les couvrira d’ignominie, vous donnera la victoire sur eux et guérira les poitrines d’un peuple croyant* » (*Dar al-Islam*, n° 8). La victoire sur l’ennemi est perçue par les djihadistes comme persuasive. Toutefois, cet énoncé, perçu comme persuasif par les djihadistes, le sera comme menaçant par l’Occident. De même, le *djihad*, qui est interprété comme persuasif par les membres du groupe djihadiste puisqu’il permet d'accéder au Paradis, tend à être associé aux attentats terroristes et donc à être perçu comme menaçant par les occidentaux. Cette double interprétation rejoint la définition de Perelman et Olbrechts-Tyteca (1988), qui proposent d’« appeler persuasive une argumentation qui ne prétend valoir que pour un auditoire particulier » (p. 36).

Bien que *Dabiq* et *Dar al-Islam* présentent le même schéma rhétorique, leur contenu varie de manière conséquente. Cette étude a révélé, par exemple, que la revue française focalise son discours sur la figure de l’*autre* (*i.e.*, de l’ennemi). En revanche, la revue anglaise est focalisée sur la figure du musulman et, plus particulièrement, sur la conduite qu’un bon musulman devrait avoir.

3. Analyse textuelle des procédures judiciaires

Au sein d'une équipe d'enquête, le travail des analystes criminels consiste à lire et synthétiser les documents de procédures (auditions de témoins, données téléphoniques et bancaires, comptes-rendus d'expertise, etc.) afin de fournir aux enquêteurs et aux magistrats une vision plus globale des informations collectées, par le biais de schémas de représentation et de synthèses (Rossy 2011). Leur intervention est requise dans des affaires complexes comme les *cold cases* ou les affaires impliquant de larges réseaux, et permet de fournir de nouvelles pistes d'investigation pour les enquêteurs. À l'heure actuelle, les analystes s'appuient sur un logiciel de reconnaissance optique de caractères, des outils de bureautique classique (traitement de texte, tableur) ainsi que sur le logiciel de représentation graphique d'IBM Analyst's Notebook. Cet outillage ne les dispense pas d'une phase de lecture précise et chronophage de la procédure visant entre autres à repérer et extraire manuellement les

informations pertinentes pour l'enquête, regroupées en différents types d'entités qui une fois extraites sont agencées en représentation graphique (chronologique ou relationnelle).

3.1. Corpus de travail

Le corpus de travail mis à notre disposition par le PJGN est une procédure judiciaire complète jugée et résolue concernant un homicide. Le dossier, comme toute procédure judiciaire, rassemble une variété de documents : rapports d'expertise, procès-verbaux d'investigations, procès-verbaux d'auditions de témoins et de mis en cause, factures téléphoniques détaillées, données bancaires, planches photographiques, etc. Nous avons choisi de concentrer notre travail sur le sous-corpus composé des auditions de témoins et de personnes gardées à vue. Ce choix s'est fait lors de notre prise de connaissance du corpus et du domaine, les auditions représentant la masse d'information la plus dense et la plus difficilement accessible d'une procédure : le nombre des auditions (dans notre cas, 370 auditions pour environ 600 000 mots) et leur manque de structure gênent leur traitement avec des outils standards, contrairement par exemple aux données téléphoniques qui peuvent être intégrées telles quelles dans Analyst's Notebook ou à d'autres données collectées en gendarmerie sous forme de formulaires structurés.

3.2. Détection automatique d'entités nommées

La notion d'entité en analyse criminelle correspond à la notion d'entité nommée (EN) en extraction d'information : une unité linguistique monoréférentielle qui a la capacité de renvoyer à un référent unique (Nouvel & al, 2015). D'une manière générale, cinq types d'entités intéressent les analystes criminels : les personnes, les lieux, les dates et heures, les véhicules et les numéros de téléphone. Nous avons entrepris d'appliquer des techniques de détection d'EN éprouvées sur les documents de procédures judiciaires, tout en variant les approches de manière à répondre au mieux aux contraintes de chaque type d'entité. Deux fonctionnalités du logiciel UNITEX (Paumier, 2016) ont été mises en œuvres : l'édition de grammaires pour la détection des dates, l'utilisation d'un lexique pour la détection des villes, et la combinaison d'un lexique de prénoms et de règles pour les noms de personnes. Les numéros de téléphone quant à eux sont détectés à l'aide d'une expression régulière.

En l'état actuel des choses, nous sommes donc en mesure de détecter :

- Les dates normées : “le 10 janvier 2017”, “l'an deux mille dix-sept, le dix janvier”, “le 10/01/2017”
- Les noms et prénoms de personnes : “Blanche Rivière”, “Petit Noémie”, “Michel E. Dupont”
- Plus de 36000 villes figurant dans un lexique[1]

Le développement d’une approche de détection des véhicules, car leurs mentions dans le corpus combinent plusieurs types d’informations :

- genre de véhicule : moto, scooter, camionnette, voiture, etc.
- marque
- mention du modèle ou d’une forme (4X4, citadine, berline, break, etc.)
- couleurs et signes distinctifs (rouille, sérigraphie, année du modèle, etc.)

La délimitation de la mention d’un véhicule ne peut se résumer à la combinaison d’une marque et d’un modèle, comme le montrent les deux exemples suivants tirés du corpus :

[1]Disponible à l'adresse : http://sql.sh/736-base-donnees-villes-francaises (janvier 2018)

- *Il s'agit d'un petit modèle comme une TWINGO pour vous donner le volume. Il était de couleur orangé. Il est petit car il a un petit coffre.*
- *M. X. m'a cependant parlé d'un véhicule 4X4 conduit par un individu qui avait un fusil.*

La détection des véhicules nous amènera donc à envisager une approche de détection plus complexe que celles déjà mises en place.

3.3 Analyse de données textuelles et analyse criminelle, une même problématique ?

Si la détection automatique des entités nommées dans le contexte de l'analyse criminelle en gendarmerie constitue une tâche habituelle de TAL, on ne peut pas pour autant en circonscrire les apports potentiels à des aspects purement techniques. La méthodologie de travail de l'analyse criminelle repose sur l'interprétation humaine pour la production d'hypothèses, et en cela nous la rapprochons de l'analyse des données textuelles (ADT) telle que définie par (Ho-Dinh, 2017) : « Avec l'ADT, nous nous situons au contraire dans une perspective de construction des connaissances, par l'interprétation humaine des résultats obtenus grâce à des outils informatiques de calcul et de visualisation. La puissance informatique vient donc en assistance de l'exploration et la fouille des données. Cette différence fondamentale permet de produire des connaissances qualitatives sur les données et non seulement quantitatives. » La poursuite de nos travaux s'oriente donc non seulement vers l'amélioration des résultats de détection d'entités et l'introduction d'approches statistiques (TF-IDF, clustering de documents, etc) mais également vers le développement d'une interface d'exploration textuelle propre, prenant en compte les spécificités du genre textuel de la procédure judiciaire (tri du texte en fonction de sa nature : texte d'en-tête, informations d'état-civil), et permettant une navigation efficace entre entités détectées, mesures statistiques, et texte original. La méthodologie de l'analyse criminelle et les pratiques du métier pourraient être à revoir en conséquence, impliquant une phase de formation des analystes criminels aux méthodes textométriques.

4. Conclusion

Nous estimons avoir soulevé des perspectives théoriques et techniques pour l'analyse de données textuelles dans les domaines judiciaires et de la sécurité, relevant aussi bien de l'analyse de discours que du TAL et de la textométrie. Dans le cas de la propagande de Daesh, l'analyse et la compréhension du discours djihadiste pourraient contribuer à la formulation d'un contre-discours qui puisse faire face et contrer la propagande djihadiste. Concernant les pratiques d'analyse textuelles en analyse criminelle, nous espérons que la mise en place de techniques d'automatisation et d'un outil d'exploration textuelle permette de repenser la méthode d'accès à l'information en analyse criminelle et soit une première étape d'une réflexion plus large sur la collecte et la circulation de l'information et des documents dans le processus judiciaire. Ces deux cas d'études illustrent la pertinence d'approches de sciences humaines et sociales dans le contexte sécuritaire et judiciaire, qui a jusqu'à présent surtout eu recours à des expertises en sciences dites « dures » (médecine légale, biologie, chimie, informatique, etc.), regroupées sous l'appellation de « sciences forensiques ». Nous espérons que de telles contributions permettront de renforcer les liens et d'ouvrir la voie à d'autres projets associant institutions judiciaires et de défense et chercheurs en sciences humaines et sociales.

References

Caffi C., & Janney R. W. (1994). Toward a pragmatics of emotive communication. *Journal of pragmatics*, *22*(3), 325-373.

De Bonis M. (2015). La strategia della paura. *Limes*, *11*.

Ghiglione, R., Landré, A., Bromberg, M., & Molette, P. (1998). *L'analyse automatique des contenus*. Paris, Dunod.

Giro M. (2015). Parigi: il branco di lupi, lo Stato Islamico e quello che possiamo fare. *Limes*.

Ho Dinh O. (2017). *Caractérisation différentielle de forums de discussion sur le VIH en vietnamien et en français*. Thèse de doctorat, Inalco, Paris.

Marchand P. (2014). Analyse avec Iramuteq de dialogues en situation de négociation de crise : le cas Mohammed Mehra. *Actes des 12èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Paris, pp. 457-471.

Nouvel D., Erhmann M., Rosset S. (2015). *Les entités nommées pour le traitement automatique des langues*. ISTE Editions

Paumier S. (2016). Unitex 3.1 user manual, http://www-igm.univ-mlv.fr/ unitex

Perelman C., & Olbrechts-Tyteca L. (1988) (5e éd.). *Traité de l'argumentation.* Bruxelles : Edition de l'Université de Bruxelles.

Rastier F. (2011). *La mesure et le grain: sémantique de corpus*. Champion; diff. Slatkine.

Ratinaud P., Marchand P. (2012). Application de la méthode ALCESTE à de "gros" corpus et stabilité des "mondes lexicaux" : analyse du "CableGate" avec IraMuTeQ. *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Liège, 13-15 juin, p. 835-844.

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.

Rumman A., Suliman M. et al. (2016). *The Secret of Attraction: ISIS Propaganda and Recruitmenet*. Traduit par Ward, W. J. et al. Amman: Friedrich-Ebert-Stiftung.

Valette M., & Rastier F. (2006). Prévenir le racisme et la xénophobie: propositions de linguistes. *Langues modernes*, *100*(2), 68.