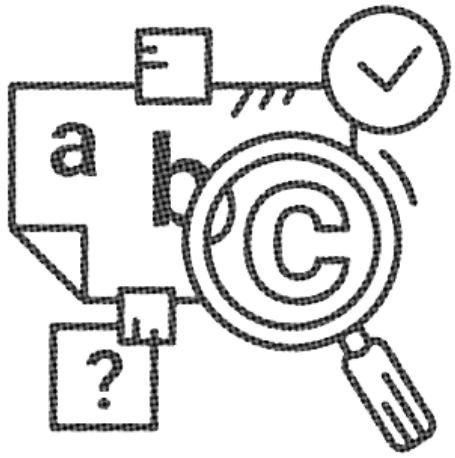
Un bref aperçu des méthodes modernes de la criminalistique linguistique pour déterminer les auteur·ice·s d'un texte. L'article suivant tente de donner un aperçu d'un point de vue non technique.

No Trace Project / Pas de trace, pas de procès. Un ensemble d'outils pour aider les anarchistes et autres rebelles à **comprendre** les capacités de leurs ennemis, **saper** les efforts de surveillance, et au final **agir** sans se faire attraper.

Selon votre contexte, la possession de certains documents peut être criminalisée ou attirer une attention indésirable. Faites attention aux brochures que vous imprimez et à l'endroit où vous les conservez.

Qui a écrit ça?



Qui a écrit ça ?

Texte d'origine en allemand

Wer schreibt denn da?

Zündlumpen #76

2020

zuendlappen.noblogs.org/post/2020/10/07/wer-schreibt-denn-da

Traduction et mise en page

No Trace Project

notrace.how/resources/fr/#qui-a-ecrit

qu'une autre. Celleux qui n'écrivent pas de communiqués évitent largement ce problème, mais sont tout de même concerné·e·s s'iels participent à des publications ou écrivent d'autres textes. Cellui qui camoufle des textes avant leur publication, par exemple en faisant réécrire et reformuler successivement des passages par plusieurs personnes, etc., court quand même le risque de développer des caractéristiques linguistiques et stylistiques exploitables ou de ne pas réussir à dissimuler des caractéristiques. Cellui qui pense pouvoir ignorer tout ça parce qu'il n'existe aucun échantillon de texte qui peut lui être attribué ou parce qu'iel est convaincu que la valeur juridique de la reconnaissance d'auteur·ice est trop fragile, risque qu'à l'avenir des échantillons de texte deviennent d'une manière ou d'une autre disponibles (par exemple parce qu'iel est reconnu·e coupable d'avoir écrit un texte) ou que la valeur juridique de la procédure évolue. Celleux qui pensent que la technologie n'est pas (encore) assez bonne peuvent être surpris·es par les développements futurs. Celleux qui utilisent des solutions techniques pour masquer leur qualité d'auteur·ice courent le risque de laisser de nouvelles caractéristiques et traces, et aussi de produire des communiqués mal écrits que personne ne veut lire de toute façon. Celleux qui n'écrivent jamais aucun texte, eh bien, n'écrivent jamais aucun texte.

Donc faites ce qui vous parle le plus, mais faites-le dès maintenant—si ce n'est déjà le cas—en gardant à l'esprit ces traces et cette sensation de malaise dans l'estomac qui, dit-on, a sauvé plus d'une personne d'une erreur d'inattention au moment crucial.

de tentatives pour déterminer des éléments comme le genre, mais ça semble généralement moins évident.

En revanche, il existe également des analyses plus quantitatives et statistiques qui examinent tout ce qui peut être mesuré de cette manière, de la fréquence des mots aux termes particuliers utilisés en passant par la structure syntaxique des phrases. Ces méthodes, connues sous le nom de stylométrie, sont parfois très controversées car il n'est pas possible de dire exactement ce qu'elles sont censées mesurer, mais elles donnent parfois des résultats étonnants, notamment en combinaison avec des techniques d'apprentissage automatique (*machine learning*). Je pense que ces approches sont donc surtout susceptibles d'être utilisées pour regrouper différents textes en fonction de leurs similitudes.

L'avantage évident de ces analyses quantitatives est qu'elles peuvent être réalisées en masse. Tous les textes disponibles ou numérisables peuvent être analysés de cette manière, des messages sur les réseaux sociaux aux livres. Bien que le succès de ces méthodes soit actuellement encore relativement modeste, et qu'il s'est souvent avéré que des textes supposés similaires le sont davantage par leur genre que par leur auteur·ice, si on part du principe que les styles d'écriture individuels pourraient correspondre à des modèles quantitatifs, ça signifie qu'une fois ces modèles connus, une attribution massive de textes à certain·e·s auteur·ice·s sera possible.

Et maintenant ?

Il y avait et il y a, bien sûr, diverses approches pour gérer cette situation, aucune n'étant meilleure ou pire

Sommaire

Introduction

Un bref aperçu des méthodes modernes de la criminalistique linguistique pour déterminer les auteur·ice·s d'un texte.

L'article suivant tente de donner un aperçu d'un point de vue non technique. Il existe quelques publications académiques sur ce sujet qui pourraient être examinées pour une meilleure analyse. Cependant, mon objectif principal ici est de soulever la question, et non de fournir un point de vue solide et concluant. Si vous en savez plus, publiez !

La plupart des gens qui commettent occasionnellement des délits et ont des démêlés avec la justice s'intéressent sans doute à la possibilité d'éviter de laisser des traces qui pourraient leur coûter cher à l'avenir, peut-être même après des années ou des décennies. Ne pas laisser d'empreintes digitales, de traces ADN, d'empreintes de chaussures ou de traces de fibres textiles ou au moins se débarrasser des vêtements après coup, éviter les caméras de surveillance, faire attention aux traces d'outils, éviter les enregistrements de toute sorte, détecter la surveillance, etc.—tout ça devrait être une préoccupation pour toute personne qui commet des délits de temps en temps et qui ne veut pas être identifiée. Mais qu'en est-il de ces traces qui n'apparaissent souvent qu'après la commission d'un délit, dans le désir d'expliquer son acte de manière anonyme ou même en utilisant un pseudonyme récurrent ? Lors de la rédaction et de la publication d'un communiqué ?

J'ai l'impression que souvent, aucune attention particulière n'est accordée à ces traces malgré un développement technologique rapide des capacités d'analyse. Ça peut être délibéré, être une négligence, ou être un

spécialiste du langage. Les fautes d'orthographe, les erreurs grammaticales, la ponctuation, mais aussi les fautes de frappe, l'orthographe nouvelle ou ancienne, les indications sur les particularités du clavier, etc., tout ça sert aux flics du langage à collecter des indices sur l'auteur·ice. Par exemple, si j'écris « muß » au lieu de « muss », ça peut être un indice que j'ai manqué certaines des réformes orthographiques les plus récentes quand j'étais à l'école. Si, en revanche, j'écris constamment des termes qui, selon les règles d'orthographe, utilisent « ß » et non « ss », ça pourrait signifier qu'il n'y a pas de « ß » sur mon clavier. Par exemple, si je parle de « dem Butter » [*au lieu de « die Butter »*], ça pourrait être une référence au fait que j'ai grandi en Bavière, etc. Mais peut-être aussi que je simule toutes ces choses dans le seul but d'induire en erreur les flics du langage. La plausibilité de mon profil d'erreur fait également partie d'une telle analyse. De même, l'analyse stylistique examine les particularités de mon style d'écriture. Quel type de termes j'utilise, ma structure de phrase présente-t-elle des schémas spécifiques, y a-t-il des termes particuliers qui se répètent d'un texte à l'autre, etc. Je pense que toute personne qui examine de plus près ses textes reconnaîtra certaines caractéristiques stylistiques qui lui sont propres.

De telles analyses qualitatives servent avant tout à établir le profil des auteur·ice·s. Il est certes possible de faire correspondre différents textes de cette manière, mais la véritable valeur de ces analyses réside dans la possibilité de déterminer des éléments tels que l'âge, le « niveau d'éducation », l'« appartenance à un milieu », les origines régionales, et parfois même des indications sur la profession/formation, etc. On entend aussi parler

suppose qu'un·e scientifique d'une certaine discipline est responsable d'une lettre, toutes les publications de cette discipline pourraient être utilisées comme échantillons de comparaison). Ça serait, par exemple, une explication (partielle) possible de ce qui a pu se passer avec Andrej Holm dans l'affaire contre le militante gruppe (mg), du moins si on suppose que le BKA n'a pas simplement tapé « gentrification » sur Google, donc je pense qu'il est tout à fait possible que de telles analyses soient effectuées.

Méthodes pour détecter des auteur·ice·s et établir des profils

Ceci dit, tout ça ne prend en compte que ce que le BKA prétend être capable de faire et pousse ces considérations jusqu'à certaines conclusions logiques. Mais comment fonctionne réellement la reconnaissance des auteur·ice·s ou l'établissement de profils ?

Qui n'a jamais eu peur que le prof d'allemand ne vous dénonce après qu'un poème moqueur sur un enseignant soit apparu dans les toilettes et que toute l'école se moque du fait que vous seul·e auriez pu écrire « aspirateur » [*Leerer*] au lieu de « professeur » [*Lehrer*]. Heureusement, toute la fac d'allemand a joué le jeu, adoptant le récit d'une faute d'orthographe et fermant les yeux sur le jeu de mots. La criminalistique linguistique semble exiger un peu de pratique, ou au moins une motivation criminologique, qui sait ? Quoi qu'il en soit, l'analyse d'erreurs, dont la plupart ont probablement entendu parler, était l'un des principaux outils d'analyse du BKA vers 2002, avec l'analyse de style, selon un article promotionnel de Christa Baldauf, flic

compromis entre des besoins divergents. Sans vouloir faire ici une suggestion générale sur la manière de traiter ces traces—après tout, chacun·e fera ce qu'iel lui semble le mieux—je voudrais présenter les méthodes avec lesquelles les autorités enquêtrices en Allemagne et ailleurs travaillent actuellement (probablement), ce qui semble possible en théorie et ce qui pourrait devenir possible à l'avenir.

Je devrais peut-être préciser à l'avance que tout ou du moins la plupart de ce que je présente ici est scientifiquement et juridiquement controversé. Et je m'intéresse moins à la validité juridique des analyses linguistiques—ou à leur validité scientifique—qu'au fait de savoir s'il semble plausible que ces recherches puissent contribuer à une opération de surveillance, car même si une piste n'est pas utile en soi devant un tribunal, elle peut toujours mener à d'autres pistes utiles.

Identification d'auteur·ice·s au BKA

Selon ses propres dires, l'Office fédéral de la police criminelle (BKA) dispose d'un département consacré à l'identification des auteur·ice·s de textes. L'accent est mis sur les textes liés à des actes criminels, comme les communiqués de revendication, mais aussi sur les « prises de position » des « milieux extrémistes de gauche », entre autres. Tous les textes collectés sont traités par des analyses linguistiques dans un « recueil de communiqués » et peuvent être comparés et parcourus avec le système d'information criminelle sur les textes (KISTE). Selon le BKA, les textes sont classés en fonction des caractéristiques biographiques

suivantes de leurs auteur·ice·s (présumé·e·s) : origine, âge, formation et profession.

Tous les nouveaux textes sont également comparés aux textes précédemment enregistrés pour déterminer si plusieurs textes peuvent avoir été écrits par la même personne.

Dans le cadre d'enquêtes spécifiques, les textes enregistrés peuvent aussi être comparés à des textes dont l'auteur·ice est connu·e, afin de déterminer s'ils ont été écrits par la même personne ou si ça peut être exclu.

Il s'agit des informations officielles du BKA concernant ce département. Qu'est-ce que ça veut dire en pratique ?

Je pense qu'on peut supposer qu'au moins tous les communiqués de revendication sont enregistrés dans cette base de données et analysés pour voir s'il existe d'autres communiqués de revendication par le(s) même(s) auteur·ice(·s). Le fait qu'ils enregistrent également les « prises de position » permet de tirer d'autres conclusions : ça semble au moins possible qu'en plus des textes ayant une pertinence pénale, ils stockent aussi d'autres textes qui sont censés provenir d'un milieu particulier. Par exemple, des textes provenant de journaux, des déclarations de groupes/organisations politiques, des appels, des articles de blog, etc. Dans le pire des cas, je suppose que tous les textes publiés sur des sites Internet d'« extrémistes de gauche » (après tout, il est assez facile de les dénicher), ainsi que les textes de publications papier qui semblent intéressants pour les autorités enquêtrices, seraient ajoutés à cette base de données.

Ça veut dire que pour chaque communiqué de revendication, le BKA disposerait d'un ensemble de textes

dont il présume qu'ils ont le même auteur·ice. Il peut s'agir d'autres revendications ou d'autres textes qui ont été ajoutés à la base de données. Outre le cas des délits commis en série, ça peut donner d'autres indices sur les coupables, comme des pseudonymes, des noms de groupe—ou, dans le pire des cas, des noms—sous lesquels l'auteur·ice d'une revendication peut avoir écrit d'autres textes, mais aussi, selon le texte, toutes sortes d'autres informations, dont souvent des indices sur le lieu de résidence et d'activité d'une personne, ses thèmes de prédilection, ses caractéristiques biographiques, son parcours éducatif, etc. Toutes ces informations peuvent au moins servir à réduire le cercle des suspects.

Ce qui n'est pas clair dans tout ça, ce sont les autres échantillons de comparaison que le BKA pourrait obtenir. Pour la plupart des gens, il existe certainement toute une série de textes auxquels les autorités enquêtrices ont (pourraient) avoir accès et qui pourraient être ajoutés à la base de données en cas de suspicion ou même à titre de précaution—si une personne est fichée avec une mention telle que « extrémiste de gauche violent », etc. Il peut s'agir de n'importe quel document portant votre nom, qu'il s'agisse d'une lettre adressée à une autorité ou d'une lettre à l'éditeur d'un journal. Je ne citerai ici intentionnellement que les sources les plus évidentes, histoire de ne pas donner par inadvertance une inspiration décisive aux autorités enquêtrices, mais je suis sûr que vous pouvez déterminer vous-même lesquels de vos textes pourraient être accessibles. Si les enquêteurs du BKA parviennent à réduire le cercle des suspects à une caractéristique spécifique, ça permet la comparaison avec des masses d'échantillons de textes disponibles (par exemple, si on