Stylométrie : analyse des caractères typographiques utilisés par un auteur

Le but est d'obtenir en un temps minimum, la meilleure signature possible d'un auteur à partir d'œuvres sélectionnées, à l'aide de l'informatique. Le sens du texte n'est pas étudié ; les algorithmes développés recensent des caractères typographiques correspondant à des critères choisis. Ensuite, il s'agit de prendre une œuvre au hasard dans la bibliothèque et de retrouver son auteur.

Je souhaitais faire un TIPE dans le domaine de l'informatique. En cherchant sur Internet, j'ai trouvé un article de Pierre Barthélémy sur LeMonde.fr proposant une analyse de texte de plusieurs auteurs assistée par ordinateur. J'ai choisi d'approfondir cette approche.

Positionnement thématique

INFORMATIQUE (Informatique pratique), INFORMATIQUE (Informatique Théorique), MATHEMATIQUES (Autres).

Mots-clés

Mots-Clés (en français)	**Mots-Clés** (en anglais)
Reconnaissance d'auteur	*Author Recognition*
Stylométrie	*Stylometry*
Khi-deux	*Chi-squared*
Traitement de données	*Data analysis*
Typographie	*Typography*

Bibliographie commentée

La reconnaissance d'auteur, appelée aussi « stylométrie », peut être définie comme une tentative de trouver des caractéristiques propres à une personne qui a écrit ou prononcé un discours. [1] La stylométrie inclut donc la typographie, ainsi que d'autres domaines tel que la reconnaissance vocale pour laquelle on peut utiliser plusieurs techniques d'attribution d'auteur pour savoir si la personne qui prononce le discours est différente de celle qui a écrit ce même discours.

Ce problème de reconnaissance n'est pas nouveau, puisqu'il existe depuis le Vieux Testament d'après Patrick Juola. Différentes méthodes d'authentification des documents ont été utilisées comme par exemple les cachets de cire pour signer les documents, ou la calligraphie des auteurs [5] : par exemple l'identification des lettres des soldats disparus pendant la Première Guerre Mondiale, regroupées sous forme de recueil comme Parole de Poilus de Jean-Pierre Guéno (Editions 84). Les habitudes ont changé avec l'imprimerie qui a imposé un formatage de la typographie ce qui rend plus difficile l'authentification des documents. De nos jours, de nouveaux moyens sont apparus comme les statistiques et l'informatique pour associer un auteur à un texte. L'avantage est que l'on peut ainsi obtenir une réponse en un temps raisonnable. Par exemple, en 2016, Pierre Barthélémy [3], un journaliste du journal Le Monde, s'interrogea sur la pertinence du critère de la ponctuation pour identifier l'auteur d'un texte. Cette question banale en 2016 était un besoin urgent aux États-

Unis vers les années 2000 pour une affaire de menace par lettre, l'affaire « Roy Van Wyck ». La cour de New Jersey avait demandé une expertise à James Fitzgerald, un «expert en stylistique », qui s'est appuyé sur un article scientifique de McMenamin (1993) pour réaliser son expertise. Malheureusement, selon la défense et le juge, les résultats n'étaient pas assez convaincants pour inculper Van Wyck. Ainsi l'affaire du New Jersey n'a pas pu aboutir en raison des doutes concernant la fiabilité de ces nouveaux tests. En effet, il n'existait à l'époque aucun standard pour certifier que quelqu'un était bel et bien un expert en analyse stylistique et rien ne permettait de vérifier la véracité de ses résultats. Cette affaire démontra le besoin d'une méthodologie fiable sur laquelle on peut compter dans des affaires importantes [2]. Plusieurs techniques aux Etats-Unis existaient à l'époque et étaient utilisées tels que l'analyse de la syntaxe, la ponctuation, la richesse du vocabulaire, la lisibilité, l'analyse de fond, les fautes d'orthographe, les erreurs de ponctuation et les fautes de grammaire. Chaski s'est alors proposée de tester à la suite de cette affaire chaque technique d'analyse séparément sur un corpus de documents écrits par 4 femmes ayant un niveau de vie similaire, en utilisant la méthode du Khi-deux [6] Elle conclut alors que seulement 2 techniques permettent de différencier les auteurs : l'analyse de la syntaxe et la ponctuation. Pour notre étude, nous nous concentrerons sur la technique d'analyse de la ponctuation.

Mais qu'en est-il d'une analyse sur un corpus plus important ? Est-ce que l'on peut encore considéré que la technique d'analyse de la ponctuation est satisfaisante ?

Problématique retenue

Il s'agira d'évaluer la validité de l'étude de la ponctuation sur un corpus plus important à l'aide du Khi-deux. Pour cela nous utiliserons des œuvres de la littérature française du domaine publique libre de droits.

Objectifs du TIPE

1. Mise en place d'un programme pour relever la ponctuation sur chaque œuvre par auteur.
2. Détermination des caractéristiques propres à chaque auteur du corpus.
3. Détermination de la variabilité de ces caractéristiques pour l'ensemble des œuvres d'un auteur donné.
4. Exploitation des données pour évaluer la pertinence de la technique d'analyse de la ponctuation à l'aide de la méthode du Khi-deux.

Références bibliographiques

[1] Patrick Juola : Authorship Attribution : *www.mathcs.duq.edu/~juola/papers.d/fnt-aa.pdf*

[2] Carole Elisabeth Chaski : Empirical Evaluations of Language-Based Author Identification Techniques : *https://www.iula.upf.edu/materials/050520spassova.pdf*

[3] Pierre Barthélémy : Peut-on reconnaître un écrivain à sa ponctuation ? : *http://passeurdesciences.blog.lemonde.fr/2016/04/04/peut-on-reconnaitre-un-ecrivain-a-sa-*

ponctuation/

[4] Efstathios Stamatatos : Survey About Authorship Methods : *http://www.icsd.aegean.gr/lecturers/stamatatos/papers/survey.pdf*

[5] Sara Shaheen, Alyn Rockwood and Bernard Ghanem : SAR Stroke Authorship Recognition : *https://ivul.kaust.edu.sa/Documents/Publications/2015/SAR%20Stroke%20Authorship%20Recognition.pdf*

[6] Julien Barnier : Tout ce que vous n'avez jamais voulu savoir sur le χ 2 sans jamais avoir eu envie de le demander : *https://alea.fr.eu.org/git/doc_khi2.git/blob_plain/HEAD:/khi2.pdf*