Amélioration de la Classification d'Auteur par l'Apprentissage du Style Linguistique Personnalisé en Discours Oral

Soumaya SABRY[1, 3] Gaël Dias[1] Mohamad Ghassany[2] Faten Chakchouk[2]
Mohammed Hasanuzzaman[3] (1) GREYC UMR6072, Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ, 14032, Caen, France. (2) EFREI Paris Panthéon-Assas Université, 94800 Villejuif, France. (3) Department of Computer Science, Munster Technological University, Cork, T12P928, Irlande.
gael.dias@unicaen.fr

RÉSUMÉ

Le style linguistique personnalisé (SLP) constitue une manifestation de la synchronisation interpersonnelle et suscite un intérêt croissant en raison de son potentiel à renforcer le caractère humain des systèmes de traitement automatique des langues. Toutefois, peu de méthodes permettent de modéliser efficacement le style individuel dans le contexte du langage oral. Dans cet article, nous mettons en évidence l'importance du SLP à travers son application à la classification d'auteur, en proposant une approche intégrant explicitement l'information stylistique afin d'améliorer les performances. Nous incorporons le SLP dans des modèles transformeurs pré-entraînés selon deux stratégies : l'affinage supervisé avec la fonction de perte d'entropie croisée et l'apprentissage contrastif via une architecture siamoise. Nos résultats montrent que l'intégration du style linguistique oral personnalisé améliore significativement la précision de classification par rapport aux modèles ne prenant pas en compte les indices stylistiques, surpassant également les performances humaines.

ABSTRACT

Style Matters : Improving Authorship Classification with Personalized Linguistic Style in Spoken Language

Personalized linguistic style (PLS) is one modality that manifests the phenomenon of interpersonal synchrony and has gained attention for its potential to enhance the human-like qualities of natural language processing systems. Despite this attention, there exist very few methodologies that effectively represent an individual's linguistic style in spoken language. In this paper, we highlight the importance of PLS through its application to the authorship classification, introducing a novel approach that explicitly integrates stylistic information to improve performance. We integrate PLS into pre-trained transformer-based models using two strategies : fine-tuning with cross-entropy loss and contrastive learning via Siamese architecture. We show that embedding personalized spoken style significantly improves classification compared to baselines that do not consider spoken stylistic features. These findings reinforce the hypothesis that individuals exhibit distinctive linguistic styles, which can be used to improve authorship classification. Overall, our models' performance surpasses both baselines and human annotation established in this study.

MOTS-CLÉS : Apprentissage de représentations, Apprentissage par transfert, Apprentissage profond, Traitement automatique du langage naturel, Style linguistique personnalisé.

KEYWORDS: Representation Learning ,Transfer Learning, Deep Learning, NLP, Personalized Linguistic Style .

1 Introduction

Dans les interactions sociales, les individus communiquent souvent sans avoir conscience des schémas subtils mais distinctifs qui structurent leurs échanges. Ces schémas, communément désignés sous le terme de synchronie, instaurent une adaptation mutuelle entre les interlocuteurs (Kleinbub, 2017; Lakin, 2013). Cette synchronisation favorise l'alignement spontané des comportements, renforçant ainsi l'empathie et l'alliance (Koole & Tschacher, 2016). Spécifiquement, la synchronie du style oral, ou *entrainment* (Weise *et al.*, 2019), englobe des caractéristiques linguistiques et paralinguistiques distinctives qui définissent les modes de communication et influencent la dynamique sociale en reflétant des traits de personnalité et des états émotionnels (Niederhoffer & Pennebaker, 2002). La manière de s'exprimer, au-delà du contenu littéral, joue un rôle central dans la formation des relations et le développement de l'attachement (Talia *et al.*, 2017), soulignant l'importance du style linguistique personnalisé (SLP) dans les relations interpersonnelles. La synchronie, une caractéristique fondamentale du dialogue, a été largement étudiée en interaction humain–robot (Hoegen *et al.*, 2019; Biancardi *et al.*, 2021; Spillner & Wenig, 2021), afin de concevoir des agents conversationnels capables d'imiter les schémas communicationnels humains. Les études antérieures se sont concentrées sur des indices tels que le pitch, la formalité ou l'extraversion (Lubold *et al.*, 2016; Spillner & Wenig, 2021). À l'inverse, notre étude est la première à cibler spécifiquement l'apprentissage de représentations individuelles du style oral. Nous proposons deux approches basées sur des transformeurs pour extraire la représentation vectorielle du style : (1) un affinage par classification permettant de séparer les auteurs dans un espace latent, (2) une architecture siamoise contrastive, rapprochant les phrases d'un même auteur et éloignant celles d'auteurs différents. Notre approche vise à capturer les nuances de la structure linguistique orale afin de modéliser le SLP. L'espace de représentation stylistique est ensuite appliqué à l'identification d'auteur, permettant la reconnaissance du locuteur à partir d'un minimum de texte, en évaluant à la fois des configurations de *linear probing* et d'affinage supervisé. Nos résultats montrent que l'intégration du SLP dans des modèles de langage préentraînés améliore la précision de classification par rapport aux méthodes de référence, dans le contexte de transcriptions de séries télévisées et de films, qui simulent la complexité des conversations réelles. Ces résultats préliminaires soutiennent l'hypothèse selon laquelle les individus présentent des styles linguistiques distincts, exploitables pour améliorer leur identification.

Les principales contributions de cet article sont les suivantes : (i) Introduction d'un modèle de plongements stylistiques pour des transcriptions orales, fondé sur deux approches : contrastive et classification. (ii) Démonstration de la capacité de ce modèle à apprendre des nuances stylistiques pertinentes pour l'identification d'auteur, via des protocoles de *linear probing* et d'affinage supervisé. (iii) Comparaison des performances du modèle proposé à celles de modèles d'identification d'auteur (ex., Luar (Rivera-Soto *et al.*, 2021)), de modèles sémantiques (ex., RoBERTa (Liu *et al.*, 2019), MPNet (Song *et al.*, 2020)) et de modèles de plongements stylistiques pour le texte écrit (ex., Lisa (Patel *et al.*, 2023), Struct-BERT (Wang *et al.*, 2020)) en contexte de données limitées. Les résultats de notre modèle dépassent significativement les performances des modèles comparés et surpassent même les capacités humaines.

2 État de l'Art

Définition du style linguistique. Le style linguistique est une composition multidimensionnelle qui reflète la manière dont les individus expriment leurs pensées, émotions et traits personnelles à travers la parole, le choix lexical et la syntaxe. La personnalisation stylistique trouve des applications dans divers domaines (ex., marketing, communication (Mukherjee *et al.*, 2024)). Elle peut s'opérer par la modulation de l'expression émotionnelle, la variation prosodique (ex., la hauteur vocale) ou encore l'adaptation linguistique alignée sur des traits de personnalité comme l'introversion–extraversion (Lubold *et al.*, 2016). Dans cet article, nous considérons le style linguistique comme une modalité de synchronie interpersonnelle. Des études antérieures montrent que les interlocuteurs adaptent leur langage, notamment via l'usage des *function-words* (Danescu-Niculescu-Mizil & Lee, 2011), de manière similaire à la synchronisation non verbale. Cette synchronie linguistique et comportementale favorise des interactions positives et renforce la dynamique relationnelle (Ireland *et al.*, 2011), tout en étant liée à l'empathie (Lord *et al.*, 2015). L'adaptation stylistique, appelée *entrainment* (Weise *et al.*, 2019), se manifeste aux niveaux acoustique, lexical et structurel. Nous nous concentrons sur l'*entrainment* structurel — l'alignement des structures syntaxiques et des règles linguistiques (Reitter & Moore, 2007) — en raison de sa pertinence pour le traitement automatique du langage, contrairement aux caractéristiques acoustiques relevant du traitement du signal. Alors que les choix lexicaux varient souvent selon le thème (Weise *et al.*, 2019), le style structurel demeure relativement stable à travers les sujets (Niederhoffer & Pennebaker, 2002). Le style d'écriture a été largement étudié (Ireland & Pennebaker, 2010), au contraire, notre étude porte spécifiquement sur le style oral, qui concerne non pas le contenu de la phrase mais sa forme d'expression. Le style propre à chaque individu comprend des éléments distinctifs tels que les *function-words*, considérés comme des marqueurs stylistiques révélant la singularité conversationnelle (Postmus, 2023). La manière de structurer les phrases et de sélectionner ces mots peut induire un effet d'alignement stylistique chez l'interlocuteur (Niederhoffer & Pennebaker, 2002).Cette focalisation sur les *function-words* et les structures syntaxiques constitue précisément la réponse au problème du démêlage (*disentanglement*) entre style et sémantique : contrairement aux choix lexicaux pleins, fortement corrélés au thème abordé, les *function-words* sont quasi-indépendants du contenu et reflètent davantage la manière de s'exprimer que ce qui est exprimé (Niederhoffer & Pennebaker, 2002). Ainsi, notre approche vise explicitement à capturer le style comme une empreinte structurelle stable, distincte de la dimension thématique ou sémantique du discours.

Études computationnelles du style. L'intégration de l'adaptation ou de la synchronie dans les interactions numériques suscite un intérêt croissant en interaction humain–robot, en raison de son potentiel d'amélioration de la qualité des échanges et de l'engagement (Biancardi *et al.*, 2021). Les études antérieures ont exploré l'adaptation stylistique, également appelée *linguistic style matching* (Postmus, 2023; Aafjes-van Doorn *et al.*, 2020) ou alignement (Branigan *et al.*, 2010). Ritschel *et al.* (2017) montrent qu'adapter le style linguistique d'un robot via la génération automatique de texte, notamment en modulant l'extraversion, accroît l'engagement utilisateur. De même, Spillner & Wenig (2021) démontrent que l'alignement lexical et structurel basé sur des règles dans les interactions chatbot–humain améliore l'engagement et le rapport interpersonnel. Des recherches plus récentes examinent la synchronie linguistique dans les agents conversationnels (Hoegen *et al.*, 2019; Aneja *et al.*, 2021). Par exemple, Hoegen *et al.* (2019) sélectionnent, parmi plusieurs réponses générées par un modèle de langage, celle correspondant le mieux au style de l'utilisateur, montrant que les participants réagissent plus positivement aux agents adoptant un style aligné. Par ailleurs, les travaux

sur l'apprentissage de représentations stylistiques intègrent l'information stylistique dans des espaces vectoriels, souvent structurés selon des dimensions interprétables. Le modèle Lisa (Patel *et al.*, 2023), par exemple, utilise des techniques de stylométrie par *prompting* pour générer des données synthétiques et apprendre des plongements stylistiques interprétables de 768 dimensions, chacune correspondant à une caractéristique stylistique prédéfinie. Toutefois, cette approche se concentre sur le texte écrit et contraint le style à des dimensions fixes. À l'inverse, notre étude vise l'extraction du style linguistique personnalisé à partir de données orales, afin de permettre, à terme, l'alignement stylistique dans des systèmes interactifs.

Le problème de l'identification d'auteur. L'attribution d'auteur constitue un défi fondamental aux applications variées, notamment en linguistique forensique, en analyse littéraire et en sécurité numérique (Stamatatos, 2009). Elle consiste à identifier l'auteur d'un texte ou à estimer si deux textes partagent le même auteur, généralement en projetant les documents dans un espace latent où les caractéristiques occupent des régions distinctes (Seroussi *et al.*, 2011). Parmi les avancées récentes, Luar (Rivera-Soto *et al.*, 2021) représente l'état de l'art en exploitant l'apprentissage contrastif pour capturer des schémas d'écriture distinctifs tout en restant invariant aux effets de thème. Les modèles de LLMs ont également influencé les recherches récentes (Huang *et al.*, 2025; Dubey, 2024). (Jafariakinabad & Hua, 2019) proposent un modèle intégrant des caractéristiques lexicales, syntaxiques et structurelles, tandis que Hu *et al.* (2020) introduisent DeepStyle, un modèle de plongements stylistiques pour textes courts combinant indices sémantiques et stylistiques. Ces recherches témoignent des progrès réalisés et des défis persistants en attribution d'auteur. Au-delà du texte écrit, Aggazzotti *et al.* (2024) proposent le premier cadre de vérification d'auteur pour le langage oral, via un benchmark fondé sur des conversations transcrites. En contrôlant les biais thématiques, ils montrent que la performance varie significativement selon le degré de contrôle du sujet, soulignant l'influence du style de transcription. Toutefois, leur étude n'explore pas les plongements stylistiques individuels. Ces études soulignent la nécessité de modèles sensibles au style et capables de généraliser à différentes modalités. Dans cette continuité, notre recherche examine le style linguistique personnalisé dans le langage oral, dans des conditions plus réalistes intégrant dynamiques conversationnelles, données limitées et bruit accru.

3 Méthodologie

Nous fondons notre approche sur (1) l'utilisation d'une architecture de type Transformer basée uniquement sur un encodeur, capable d'apprendre des représentations vectorielles capturant des caractéristiques stylistiques, et (2) son exploitation pour évaluer la performance des plongements stylistiques dans la tâche de la classification d'auteur en langage oral.

Jeu de données. L'acquisition d'un jeu de données approprié est complexe en raison de notre focalisation sur le texte oral. De nombreux corpus de dialogue, tels que DailyDialog (Li *et al.*, 2017) et EmpatheticDialogues (Rashkin *et al.*, 2018), ne fournissent pas de labels explicites d'auteur, ce qui les rend inadaptés à l'analyse du style linguistique personnalisé. Bien que des recherches antérieures (Tripto *et al.*, 2023) aient utilisé des données orales issues de monologues, d'entretiens et de dialogues multipartites, nous nous concentrons spécifiquement sur des échanges conversationnels naturels afin de mieux capturer la complexité des interactions réelles. D'autres jeux de données pertinents

(Aggazzotti *et al.*, 2024) ne sont pas publiquement accessibles, ce qui limite la reproductibilité. Pour répondre à ces contraintes, nous utilisons des scripts de séries télévisées et de films, qui reflètent les caractéristiques du langage oral tout en fournissant des dialogues annotés à grande échelle avec identification des personnages. Leur disponibilité publique garantit à la fois praticité et reproductibilité. Plus précisément, nous exploitons des jeux de données composés du dialogue de personnages provenant de deux séries télévisées, The Big Bang Theory [1], Friends [2], ainsi que du Cornell Movie Dialogs Corpus (Danescu-Niculescu-Mizil & Lee, 2011) [3] Ce choix est délibéré, car les dialogues scénarisés des séries et des films mettent souvent en évidence des schémas d'expression distinctifs propres aux personnages, rendant les différences stylistiques plus apparentes et offrant une simulation des dynamiques conversationnelles. Nous reconnaissons néanmoins que les dialogues scénarisés peuvent parfois accentuer ou caricaturer les traits stylistiques des personnages par rapport au langage oral authentique, ce qui pourrait rendre la tâche moins représentative d'un cas d'usage réel. Cette limitation devra être prise en compte lors de la généralisation des résultats à des contextes réels. L'utilisation de l'identification de personnages comme proxy du style repose sur l'hypothèse, soutenue empiriquement, que chaque individu présente un style linguistique stable et distinctif, notamment dans son usage des *function-words* (Pennebaker & King, 1999). Cette stabilité stylistique inter-individuelle constitue le fondement de notre approche.

Le prétraitement a consisté à supprimer les phrases très courtes telles que «okay» ou «fine», tandis que les dialogues plus longs contenant plusieurs phrases ont été segmentés en unités plus petites. Précisément, la segmentation a été effectuée en utilisant la ponctuation (ex., points, points d'exclamation). Cette étape a permis d'augmenter le nombre d'échantillons disponibles pour l'entraînement tout en garantissant la cohérence stylistique et la faisabilité computationnelle de chaque segment. Nous avons également supprimé les symboles non alphabétiques tels que les crochets, ceux-ci n'étant pas considérés comme des *function-words*, afin de nous concentrer exclusivement sur les caractéristiques textuelles. Nous n'avons conservé que les personnages disposant d'au moins 100 phrases, seuil choisi pragmatiquement afin de garantir une quantité suffisante de données pour une représentation stylistique pertinente, dans la lignée de travaux similaires appliquant des critères comparables (Aggazzotti *et al.*, 2024). La diversité thématique du corpus MovieClips, regroupant 919 personnages issus de films aux genres variés, constitue un argument empirique contre la confusion entre style et sémantique : un modèle capturant uniquement le contenu échouerait face à une telle hétérogénéité (Stamatatos, 2009).

Approche générale. Notre expérimentation mobilise deux modèles basés sur des Transformers afin de produire des plongements représentant des vecteurs de style propre à chaque personnage. Nous appuyons notre apprentissage sur RoBERTa (Liu *et al.*, 2019) et MPNet (Song *et al.*, 2020), deux modèles fondation de type encodeur seul, en raison de leurs performances. Nous privilégions des modèles conçus pour capturer les connexions linguistiques brutes et fondamentales, sans biais introduits par des optimisations orientées tâche qui pourraient masquer les caractéristiques stylistiques intrinsèques que nous cherchons à isoler. Les plongements de style sont appris selon deux stratégies : (1) une approche par classification et (2) un apprentissage contrastif reposant sur une architecture siamoise. Ces deux approches diffèrent fondamentalement dans leur optimisation : l'approche discriminante (classification) apprend à assigner directement une classe à chaque représentation en optimisant une frontière de décision, tandis que l'approche contrastive structure l'espace des représen-

1. https://bigbangtrans.wordpress.com/
2. https://edersoncorbari.github.io/friends/
3. https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html

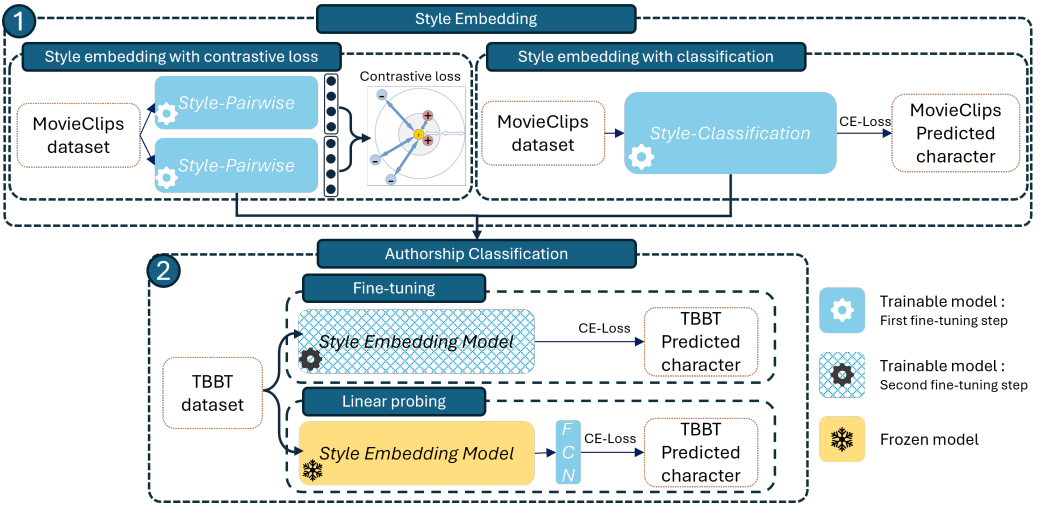

FIGURE 1 – Architecture Globale : (1) Apprentissage des plongements de style via le modèle siamois avec *contrastive loss* ou par classification avec *cross entropy* ; (2) transfert pour l'identification d'auteur à l'aide de l'affinage ou du *linear probing* sur la base du modèle gelé à partir de TBBT.

tations en rapprochant les exemples de même auteur et en éloignant ceux d'auteurs différents, sans supervision directe par classe. Les plongements stylistiques obtenus sont ensuite exploités pour la tâche de classification d'auteur selon deux protocoles d'évaluation, comme illustré en Figure 1.Les corpus Friends et MovieClips sont utilisés séparément comme données d'entraînement afin d'évaluer indépendamment la capacité de chaque corpus à apprendre des représentations stylistiques généralisables. TBBT est ensuite utilisé exclusivement comme corpus de test, permettant d'évaluer le transfert des représentations apprises vers un domaine non vu durant l'entraînement. Le premier protocole repose sur le *linear probing* : le modèle est gelé et seul un classifieur linéaire est entraîné sur TBBT, évaluant ainsi si les représentations stylistiques apprises sont directement exploitables sans adaptation supplémentaire. Le second protocole consiste en un affinage complet sur TBBT, permettant d'évaluer si le modèle peut affiner sa capacité à distinguer les styles lorsqu'il est exposé au domaine cible. Cette démarche s'inscrit dans la continuité de travaux montrant que l'adaptation spécifique au domaine des transcriptions orales améliore les performances sur différentes tâches linguistiques (Aggazzotti *et al.*, 2024).

Modèles de base. Nous évaluons notre approche en comparant nos modèles à deux catégories distinctes de *baselines* : des modèles intégrant des informations stylistiques et des modèles conventionnels ne prenant pas en compte le style. La première catégorie comprend Lisa (Patel *et al.*, 2023) et Struct-BERT (Wang *et al.*, 2020), conçus pour capturer des structures linguistiques profondes au-delà du sens superficiel. Lisa est spécifiquement entraînée pour extraire des variations stylistiques en exploitant de grands modèles de langage via des techniques avancées de prompting, ce qui en fait une référence solide pour évaluer la capacité de nos modèles à encoder des traits stylistiques individuels. Struct-BERT améliore BERT en intégrant l'ordre des mots et des informations structurelles lors du pré-entraînement, renforçant ainsi sa capacité à modéliser l'organisation syntaxique et les nuances stylistiques. Cette catégorie inclut également Luar (Rivera-Soto *et al.*, 2021), qui représente l'état de

l'art en identification d'auteur et démontre des performances supérieures, comme l'ont confirmé des analyses comparatives récentes (Aggazzotti *et al.*, 2024). Luar constitue ainsi un point de comparaison particulièrement exigeant pour notre approche d'encodage stylistique. La seconde catégorie regroupe des modèles n'intégrant pas explicitement la notion de style, tels que RoBERTa (Liu *et al.*, 2019) et MPNet (Song *et al.*, 2020), utilisés comme encodeurs de phrases à usage général. Ces modèles sont pré-entraînés sans objectif stylistique.

Plongements stylistiques contrastifs. Une première approche pour créer des plongements de style repose sur la fonction de perte contrastive implémentée via une architecture siamoise (Chicco, 2021). Nous adoptons en particulier la stratégie par paires proposée par Zhang *et al.* (2021). Nous affinons des modèles Sentence Transformers (Reimers, 2019) basés sur RoBERTa et MPNet, séparément, sur les jeux de données Friends et MovieClips afin d'analyser l'influence des différentes sources de données. L'objectif est de minimiser la distance entre les vecteurs de phrases stylistiquement similaires (même personnage) tout en maximisant la séparation entre phrases stylistiquement distinctes (personnages différents). La similarité cosinus est utilisée comme fonction de distance, avec une marge fixée à 0,5 (par défaut dans Sentence Transformers). Les modèles obtenus sont *Style-RoBERTa-Pairwise* et *Style-MPnet-Pairwise*.

Plongements stylistiques par classification. Dans notre seconde approche, nous reformulons l'apprentissage du style comme une tâche de classification. Nous affinons les modèles RoBERTa et MPNet sur les corpus Friends ou MovieClips à l'aide d'une fonction de perte d'entropie croisée, en les entraînant à prédire le locuteur de chaque phrase. Le modèle de plongement stylistique correspond alors au modèle fondation affiné dans ce cadre de classification. Les modèles résultants sont notés *Style-RoBERTa-Classification* et *Style-MPnet-Classification*.

Classification humaine. Afin d'évaluer la performance humaine en attribution d'auteur, nous avons sollicité deux experts très familiers avec la série *The Big Bang Theory*, l'ayant visionnée à plusieurs reprises et maîtrisant les spécificités stylistiques propres à chaque personnage. Ils ont volontairement annoté l'ensemble du jeu de test TBBT. Pour chaque des 1,041 répliques isolées, ils devaient leur attribuer l'un des 13 personnages principaux. L'annotation a nécessité environ 5 à 6 heures par expert, et la précision finale correspond à la moyenne de leurs prédictions [4].

4 Résultats et Discussion

Dans cette section, nous présentons une analyse comparative de l'efficacité des modèles d'attribution d'auteur dans des conditions expérimentales contrôlées, afin de garantir une évaluation équitable des performances. Les modèles sont évalués sur l'ensemble de test standardisé TBBT, comprenant 13 classes correspondant à des personnages distincts, ainsi que sur les partitions d'entraînement et de validation de TBBT [5]. Les *splits* sont donnés dans le Tableau 2 de l'annexe. Pour assurer la cohérence expérimentale, nous avons adopté des configurations d'hyperparamètres uniformes lorsque

4. Des travaux futurs viseront à élargir cette évaluation en augmentant le nombre de participants afin d'obtenir des annotations plus généralisables.

5. Notons que 20% d'entraînement est utilisé pour l'ensemble de validation, ce qui est identique pour toutes les exécutions.

cela était méthodologiquement pertinent : une taille de batch de 64 est utilisée pour l'affinage et les tâches de classification, tandis que les modèles entraînés avec une fonction de perte contrastive utilisent une taille de batch de 128. Le taux d'apprentissage dépend du modèle et varie de 1×10^{-6} à 1×10^{-4}. L'entraînement est effectué avec arrêt anticipé lorsque la perte de validation cesse de diminuer, indiquant une convergence. Lors de l'étape d'affinage, nous sélectionnons le point de contrôle (checkpoint) présentant les meilleures performances sur l'ensemble de validation, la tâche de classification montrant une tendance au sur-apprentissage sur les données d'entraînement. Les expériences de *linear probing* sont réalisées avec un taux d'apprentissage fixe de 1×10^{-4}. Le nombre d'époques varie selon les performances sur la validation, avec arrêt anticipé afin de limiter le sur-apprentissage. Comme précédemment, le modèle retenu est celui obtenant les meilleurs résultats sur l'ensemble de validation. Les performances sont évaluées à l'aide de deux métriques complémentaires : l'accuracy, qui mesure la proportion globale de prédictions correctes, et le F1-score (macro), qui équilibre précision et rappel, particulièrement pertinent dans le cas de jeux de données déséquilibrés tels que TBBT. Afin d'obtenir une estimation fiable, chaque modèle est entraîné avec cinq graines aléatoires différentes et les résultats sont moyennés (Colas *et al.*, 2018).

Nom du modèle	Linear Probing		Affinage	
	F1-score	Accuracy	F1-score	Accuracy
RoBERTa (Liu *et al.*, 2019)	0.06	0.16	0.20±0.04	0.28±0.02
MPNet (Song *et al.*, 2020)	0.06	0.17	0.15±0.02	0.25±0.01
Struct-BERT (Wang *et al.*, 2020)	0.07	0.17	0.19±0.01	0.27±0.02
Luar (Rivera-Soto *et al.*, 2021)	0.08	0.18	0.16±0.02	0.26±0.01
Lisa (Patel *et al.*, 2023)	0.05	0.14	0.09±0.02	0.20±0.02
Pré-entraînement sur Friends				
Style-RoBERTa-C.	0.09	0.19	0.20±0.01⋆	0.27±0.01⋆
Style-MPnet-C.	0.11	0.20	0.18±0.01⋆	0.26±0.01⋆
Style-RoBERTa-P.	0.10	0.19	0.22±0.01×	0.35±0.02 †
Style-MPnet-P.	0.14	0.20	0.22±0.01×	0.35±0.01 †
Pré-entraînement sur MovieClips				
Style-RoBERTa-C.	0.10	0.18	0.16±0.02	0.25±0.02
Style-MPnet-C.	0.15	0.22	0.15±0.04	0.24±0.03
Style-RoBERTa-P.	0.16	**0.23**	**0.23**±0.01×	0.36±0.01 †
Style-MPnet-P.	**0.17**	**0.23**	**0.23**±0.00×	**0.37±0.01** †
Human classif.	0.20±0.04	0.24±0.04	0.20±0.04	0.24±0.04

TABLE 1 – Performance de l'attribution d'auteur sur TBBT (5 exécutions pour l'affinage et 1 pour le *linear probing*). Les symboles indiquent la significativité statistique selon un test t apparié avec une $p - value < 0.05$: † indique des résultats significativement supérieurs à *tous* les modèles de base ; × indique une supériorité significative par rapport à tous les modèles de base *sauf RoBERTa* ; ⋆ indique une supériorité significative uniquement par rapport à deux modèles de base (MPNet et Lisa).

Le Tableau 1 montre que les modèles Style-X-Pairwise [6] ainsi que les modèles Style-X-Classification entraînés sur MovieClips surpassent systématiquement les modèles de référence dans la configuration de *linear probing*, confirmant que l'intégration d'informations stylistiques améliore les capacités des modèles. En configuration d'affinage, les modèles Style-X-Pairwise obtiennent les meilleures performances globales, tandis que les modèles Style-X-Classification présentent des résultats inférieurs à ceux de leurs modèles de base correspondants, sans que cette différence soit statistiquement significative. La comparaison des deux méthodologies met en évidence l'avantage de l'apprentissage

6. X désigne RoBERTa ou MPNet.

contrastif par rapport à l'approche par classification, en particulier dans la configuration de *linear probing*. Cet effet est particulièrement marqué avec RoBERTa, où le modèle contrastif surpasse le modèle de classification et obtient le meilleur score global. Cette observation est cohérente avec des travaux antérieurs (Schroff *et al.*, 2015), soulignant l'efficacité des pertes contrastives pour produire des représentations plus discriminantes. Le corpus MovieClips comprend plus de 900 personnages. Un tel nombre de classes accroît la complexité des frontières de décision, ce qui tend à dégrader les performances des modèles basés sur la classification (Liu & Wangperawong, 2019). Cette tendance se reflète dans leurs scores plus faibles en *linear probing* et en affinage. À l'inverse, l'apprentissage contrastif par paires se montre plus robuste face à cet espace d'étiquettes étendu. Néanmoins, les modèles de classification surpassent toujours l'ensemble des baselines en configuration de *linear probing*, ce qui constitue un résultat encourageant. Notre étude empirique indique que les représentations stylistiques issues de données orales contribuent de manière significative à l'amélioration des performances en attribution d'auteur, tandis que les modèles fondés uniquement sur des textes écrits ont un impact plus limité. Une analyse détaillée par classe est proposée dans le Tableau 3 de l'annexe, démontrant les capacités discriminantes de l'espace de représentation dans le cadre du *linear probing*.

Impact du corpus de pré-entraînement. Afin d'évaluer la robustesse de notre approche, nous avons appliqué l'ensemble du pipeline à différents jeux de données pour analyser l'effet du nombre de personnages sur les performances. Nous avons sélectionné le corpus *Friends*, qui partage des caractéristiques thématiques et conversationnelles avec TBBT, mais comprend un nombre significativement plus restreint de personnages que MovieClips. Le pipeline a été entièrement réentraîné et réévalué sur ce corpus (cf. Tableau 1). Les modèles entraînés sur Friends, bien qu'ils ne surpassent pas le modèle pairwise entraîné sur MovieClips, dépassent néanmoins toutes les baselines ainsi que leurs équivalents basés sur la classification. En particulier, le modèle de classification entraîné sur Friends obtient un meilleur score en affinage que son équivalent entraîné sur MovieClips et surpasse significativement Lisa et MPNet. Cela s'explique par le nombre plus restreint de personnages (6 principaux), facilitant l'apprentissage. Toutefois, le modèle basé sur une perte contrastive demeure le plus performant, en *linear probing* comme en affinage. Ces résultats démontrent la robustesse et l'adaptabilité de notre approche de plongements stylistiques [7].

Impact de la longueur du texte. Nous examinons également la difficulté d'identifier un auteur à partir d'une seule phrase. Ce choix expérimental s'appuie sur (Danescu-Niculescu-Mizil & Lee, 2011), qui montre que l'effet *chameleon* apparaît en réponse à la phrase précédente d'un interlocuteur. L'identification sur une seule réplique reflète ainsi des conditions d'interaction réelle. Nous avons étendu l'analyse à une configuration multi-phrases. La Figure 2 présente les résultats en *linear probing* sur TBBT, en utilisant un modèle préentraîné sur une seule phrase de MovieClips, puis testé sur des séquences de 1, 3, 5, 7, 10, 15 et 20 phrases consécutives par personnage, construites via une fenêtre glissante.Ces valeurs ont été choisies afin de couvrir différentes échelles de contexte avec des incréments réguliers, tout en limitant le coût computationnel lié à une évaluation exhaustive. Comme le montre la Figure 2, les performances convergent à 20 phrases, ce qui justifie l'arrêt de l'analyse à ce seuil. Les performances augmentent avec le nombre de phrases, ce qui est attendu : un contexte plus large apporte davantage de diversité lexicale et de régularités propres aux personnages. Dans les configurations courtes (une ou trois phrases), nos modèles préentraînés sur MovieClips, surpassent les baselines, montrant leur capacité à capturer des indices stylistiques même en contexte

7. Des travaux futurs étendront cette analyse à des corpus conversationnels plus diversifiés et réalistes, afin d'évaluer la robustesse en conditions réelles.

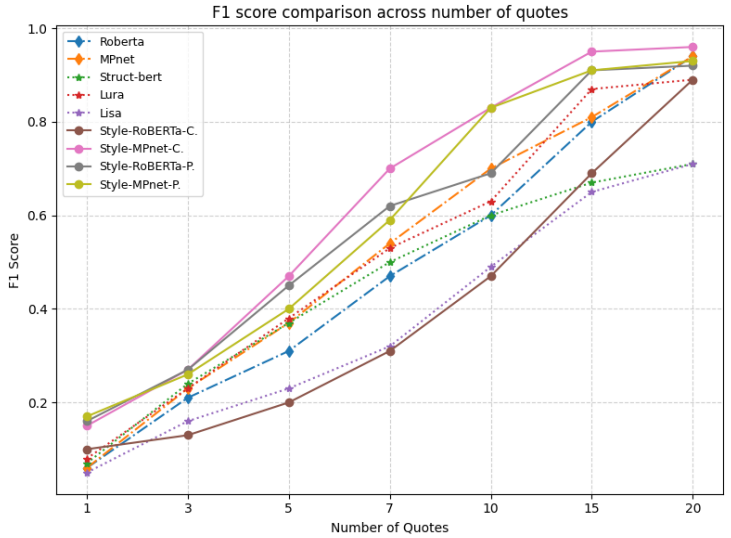

FIGURE 2 – Comparaison en *linear probing* entre nos modèles préentraînés sur MovieClips et les modèles de base sur TBBT dans une configuration multi-phrases. Le test comprend des échantillons composés de 1, 3, 5, 7, 10, 15 ou 20 phrases consécutives par personnage, tandis que l'entraînement est effectué sur des phrases uniques.

minimal. À mesure que le nombre de phrases augmente, les performances convergent entre les modèles, avec une amélioration globale entre sept et dix phrases, liée à la richesse structurelle et syntaxique accrue. Malgré cette convergence, nos modèles conservent un avantage, confirmant la contribution persistante de l'information stylistique. Au-delà de quinze phrases, les performances tendent à se stabiliser, suggérant une domination croissante des similarités sémantiques sur les variations stylistiques. Le modèle contrastif demeure toutefois plus stable et continue de capter des nuances fines. Enfin, le modèle Style-RoBERTa-Classification, initialement en retrait en configuration mono-phrase, progresse avec l'augmentation du contexte, indiquant une dépendance plus marquée aux indices sémantiques. Au final, ces résultats montrent que l'augmentation du contexte facilite la classification d'auteur, tout en confirmant la robustesse de nos modèles stylistiques via des variations de longueur d'entrée.

5 Conclusions

Cet article propose une nouvelle perspective sur le style linguistique personnalisé (SLP) dans le langage oral et démontre l'intérêt pour l'identification d'auteur. En nous appuyant sur des définitions linguistiques (Weise *et al.*, 2019) et computationnelles (Postmus, 2023), nous montrons que le style oral personnalisé existe et peut être exploité efficacement dans des tâches computationnelles.

Nous évaluons notre approche sur une tâche de classification d'auteur à partir des phrases issues de séries télévisées et de films. Bien qu'adaptés à une validation de principe, ces corpus restent limités pour l'apprentissage d'embeddings stylistiques à grande échelle, ce qui motive l'exploration future de

jeux de données plus diversifiés et étendus.

Nous proposons deux modèles fondés sur des transformeurs pour la représentation stylistique orale : une approche de classification et une d'apprentissage contrastif. Les résultats montrent que les modèles Style-X-Pairwise surpassent les baselines, y compris celles intégrant le style des textes écrits. Ces observations sont cohérentes avec la sociolinguistique, selon lesquelles les individus présentent des schémas distinctifs et inconscients (Niederhoffer & Pennebaker, 2002), traités de manière plus automatique et moins marquante que le contenu sémantique (Ireland & Pennebaker, 2010).

Au-delà de l'identification d'auteur, des recherches futures exploreront l'intégration stylistique pour des agents conversationnels capables de s'adapter dynamiquement aux préférences linguistiques des utilisateurs. Une telle adaptation peut être perçue comme empathique (Ireland & Pennebaker, 2010; Lord *et al.*, 2015) et favoriser l'alliance thérapeutique (Howick *et al.*, 2018).

Ces avancées ouvrent des perspectives en interaction humain–robot via des modèles adaptatifs de parole (Postmus, 2023), en validation computationnelle des théories de synchronie linguistique, en adaptation stylistique personnalisée pour les LLMs, ainsi qu'en amélioration des interactions thérapeutiques grâce à des agents conversationnels adaptatifs. Dans l'ensemble, cette étude contribue au développement de systèmes d'IA plus adaptatifs et plus proches du comportement humain en approfondissant l'étude du style linguistique et de son adaptation computationnelle.

Références

Aafjes-van Doorn K., Porcerelli J. & Müller-Frommeyer L. C. (2020). Language style matching in psychotherapy : An implicit aspect of alliance. *Journal of Counseling Psychology*, **67**(4), 509.

Aggazzotti C., Andrews N. & Smith E. A. (2024). Can authorship attribution models distinguish speakers in speech transcripts ? *Transactions of the Association for Computational Linguistics*, **12**, 875–891.

Aneja D., Hoegen R., McDuff D. & Czerwinski M. (2021). Understanding conversational and expressive style in a multimodal embodied conversational agent. In *Conference on Human Factors in Computing Systems (CHI)*, p. 1–10.

Biancardi B., Dermouche S. & Pelachaud C. (2021). Adaptation mechanisms in human–agent interaction : Effects on user's impressions and engagement. *Frontiers in Computer Science*, **3**, 696682.

Branigan H. P., Pickering M. J., Pearson J. & McLean J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, **42**(9), 2355–2368. How people talk to Robots and Computers, DOI : https ://doi.org/10.1016/j.pragma.2009.12.012.

Chicco D. (2021). Siamese neural networks : An overview. *Artificial neural networks*, p. 73–94.

Colas C., Sigaud O. & Oudeyer P.-Y. (2018). How many random seeds ? statistical power analysis in deep reinforcement learning experiments. *arXiv preprint arXiv :1806.08295*.

Danescu-Niculescu-Mizil C. & Lee L. (2011). Chameleons in imagined conversations : A new approach to understanding coordination of linguistic style in dialogs. *arXiv preprint arXiv :1106.3077*.

Dubey A. (2024). *Capturing Style Through Large Language Models-An Authorship Perspective*. Thèse de doctorat, Purdue University.

HOEGEN R., ANEJA D., MCDUFF D. & CZERWINSKI M. (2019). An end-to-end conversational style matching agent. In *19th ACM International Conference on Intelligent Virtual Agents (IVA)*, p. 111–118.

HOWICK J., BIZZARI V. & DAMBHA-MILLER H. (2018). Therapeutic empathy : what it is and what it isn't. *Journal of the Royal Society of Medicine*, **111**(7), 233–236.

HU Z., LEE R. K.-W., WANG L., LIM E.-P. & DAI B. (2020). Deepstyle : User style embedding for authorship attribution of short texts. In *4th Joint International Conference on Web and Big Data (APWeb-WAIM)*, p. 221–229 : Springer.

HUANG B., CHEN C. & SHU K. (2025). Authorship attribution in the era of llms : Problems, methodologies, and challenges. *ACM SIGKDD Explorations Newsletter*, **26**(2), 21–43.

IRELAND M. E. & PENNEBAKER J. W. (2010). Language style matching in writing : synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology*, **99**(3), 549.

IRELAND M. E., SLATCHER R. B., EASTWICK P. W., SCISSORS L. E., FINKEL E. J. & PENNEBAKER J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological science*, **22**(1), 39–44.

JAFARIAKINABAD F. & HUA K. A. (2019). Style-aware neural model with application in authorship attribution. In *18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, p. 325–328 : IEEE.

KLEINBUB J. R. (2017). State of the art of interpersonal physiology in psychotherapy : a systematic review. *Frontiers in psychology*, **8**, 2053.

KOOLE S. L. & TSCHACHER W. (2016). Synchrony in psychotherapy : A review and an integrative framework for the therapeutic alliance. *Frontiers in psychology*, **7**, 191242.

LAKIN J. L. (2013). Behavioral mimicry and interpersonal synchrony. *Nonverbal communication*, p. 539–576.

LI Y., SU H., SHEN X., LI W., CAO Z. & NIU S. (2017). Dailydialog : A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv :1710.03957*.

LIU X. & WANGPERAWONG A. (2019). Transfer learning robustness in multi-class categorization by fine-tuning pre-trained contextualized language models. *arXiv preprint arXiv :1909.03564*.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

LORD S. P., SHENG E., IMEL Z. E., BAER J. & ATKINS D. C. (2015). More than reflections : Empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior therapy*, **46**(3), 296–303.

LUBOLD N., WALKER E. & PON-BARRY H. (2016). Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, p. 255–262 : IEEE.

MUKHERJEE S., LANGO M., KASNER Z. & DUŠEK O. (2024). A survey of text style transfer : Applications and ethical implications. *arXiv preprint arXiv :2407.16737*.

NIEDERHOFFER K. G. & PENNEBAKER J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, **21**(4), 337–360.

PATEL A., RAO D., KOTHARY A., MCKEOWN K. & CALLISON-BURCH C. (2023). Learning interpretable style embeddings via prompting llms. *arXiv preprint arXiv :2305.12696*.

PENNEBAKER J. W. & KING L. A. (1999). Linguistic styles : language use as an individual difference. *Journal of personality and social psychology*, **77**(6), 1296.

POSTMUS T. (2023). A review on linguistic style matching.

RASHKIN H., SMITH E. M., LI M. & BOUREAU Y. (2018). I know the feeling : Learning to converse with empathy. *arXiv preprint arXiv :1811.00207*.

REIMERS N. (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv :1908.10084*.

REITTER D. & MOORE J. D. (2007). Predicting success in dialogue. In *45th Annual Meeting of the Association of Computational Linguistics (ACL)*, p. 808–815.

RITSCHEL H., BAUR T. & ANDRÉ E. (2017). Adapting a robot's linguistic style based on socially-aware reinforcement learning. In *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, p. 378–384 : IEEE.

RIVERA-SOTO R. A., MIANO O. E., ORDONEZ J., CHEN B. Y., KHAN A., BISHOP M. & ANDREWS N. (2021). Learning universal authorship representations. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 913–919. DOI : 10.18653/v1/2021.emnlp-main.70.

SCHROFF F., KALENICHENKO D. & PHILBIN J. (2015). Facenet : A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 815–823.

SEROUSSI Y., ZUKERMAN I. & BOHNERT F. (2011). Authorship attribution with latent dirichlet allocation. In *Proceedings of the fifteenth conference on computational natural language learning*, p. 181–189.

SONG K., TAN X., QIN T., LU J. & LIU T.-Y. (2020). Mpnet : Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, **33**, 16857–16867.

SPILLNER L. & WENIG N. (2021). Talk to me on my level–linguistic alignment for chatbots. In *23rd International Conference on Mobile Human-Computer Interaction (MOBHCI)*, p. 1–12.

STAMATATOS E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, **60**(3), 538–556.

TALIA A., MILLER-BOTTOME M. & DANIEL S. I. (2017). Assessing attachment in psychotherapy : validation of the patient attachment coding system (pacs). *Clinical Psychology & Psychotherapy*, **24**(1), 149–161.

TRIPTO N. I., UCHENDU A., LE T., SETZU M., GIANNOTTI F. & LEE D. (2023). Hansen : human and ai spoken text benchmark for authorship analysis. *arXiv preprint arXiv :2310.16746*.

WANG W., BI B., YAN M., WU C., XIA J., BAO Z., PENG L. & SI L. (2020). Structbert : Incorporating language structures into pre-training for deep language understanding. In *8th International Conference on Learning Representations (ICLR)*.

WEISE A., LEVITAN S. I., HIRSCHBERG J. & LEVITAN R. (2019). Individual differences in acoustic-prosodic entrainment in spoken dialogue. *Speech Communication*, **115**, 78–87.

ZHANG D., LI S.-W., XIAO W., ZHU H., NALLAPATI R., ARNOLD A. O. & XIANG B. (2021). Pairwise supervised contrastive learning of sentence representations. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 5786–5798. DOI : 10.18653/v1/2021.emnlp-main.467.

6 Annexe

Analyse par personnage Afin d'analyser plus finement le comportement des modèles, nous examinons les performances par personnage.Afin de mieux analyser les performances par personnage, nous avons présenté dans la table 2 une répartition détaillée des partitions du jeu de données TBBT, incluant le nombre de phrases et le nombre moyen de tokens par phrase pour chaque personnage. À noter que TBBT est utilisé comme jeu de données externe pour l'évaluation de la classification d'auteur. Il comprend trois niveaux distincts : trois personnages principaux avec le plus grand nombre de phrases ; des personnages principaux secondaires avec un nombre intermédiaire ; et des personnages secondaires avec significativement moins de phrases, afin d'évaluer plus finement les résultats suivants.

TABLE 2 – Répartition des phrases par personnage et longueur moyenne en tokens pour TBBT.

TBBT			
Name	**nb. quotes-Train**	**nb. quotes-Test**	**Avg. Token**
Sheldon	18749	110	19.1
Leonard	10704	110	16.0
Penny	8454	110	15.9
Howard	7275	110	16.6
Raj	6189	110	17.0
Amy	4090	110	16.7
Bernadette	2893	110	15.4
Stuart	778	110	16.0
Mary Cooper	272	48	20.6
Priya	179	32	14.5
Beverley	169	30	18.0
Emily	149	27	14.4
Arthur	131	24	18.1

Comme constaté dans le Tableau 3, les modèles de baseline rencontrent des difficultés à distinguer les personnages, y compris ceux disposant d'un nombre important de phrases, bien que Struct-BERT présente des résultats intéressants, soulignant l'importance de la structure linguistique. Cela est cohérent avec le fait que ces modèles n'ont pas été conçus spécifiquement pour différencier des identités individuelles dans un contexte oral.

Les modèles fondés sur des embeddings stylistiques issus de données orales montrent des améliorations notables, en particulier Style-MPnet-Pairwise, bien qu'ils demeurent limités pour les personnages faiblement représentés. Cette difficulté s'explique par la complexité intrinsèque de l'identification d'un locuteur à partir d'un nombre très restreint d'énoncés (un seul extrait étant utilisé pour la prédiction).

L'examen détaillé du tableau révèle un comportement contrasté : si les modèles de baseline améliorent progressivement leurs prédictions pour la majorité des personnages, notre meilleur modèle obtient des scores élevés pour les personnages abondamment représentés, tout en produisant des scores nuls pour ceux disposant de peu de données. Ce phénomène suggère une optimisation orientée vers les classes majoritaires dans un contexte de fort déséquilibre. Ce déséquilibre constitue un enjeu central pour des travaux futurs, nécessitant des stratégies de pondération plus équitables. Malgré cela, la qualité des modèles Style-X-Pairwise entraînés sur MovieClips est confirmée par leurs résultats en linear probing, où ils sont les seuls à maintenir des performances positives sur l'ensemble des classes.

Les résultats de l'annotation humaine, plus uniformément répartis entre 0,11 et 0,30, rejoignent les observations de la littérature linguistique (Ireland & Pennebaker, 2010), qui souligne la difficulté

Fine-Tuning									
Name	**Human Ann.**	**Roberta**	**Mpnet**	**Struct-BERT**	**Luar**	**Lisa**	**Style-Roberta-P. (MovieClips)**	**Style-Mpnet-P. (MovieClips)**	**AVG.**
Sheldon	0.30 ± 0.00	0.40 ± 0.02	0.39 ± 0.02	0.37 ± 0.06	0.41 ± 0.01	0.34 ± 0.03	0.43 ± 0.02	0.42 ± 0.01	**0.39 ± 0.02**
Leonard	0.22 ± 0.01	0.28 ± 0.02	0.24 ± 0.02	0.29 ± 0.02	0.30 ± 0.02	0.23 ± 0.02	0.34 ± 0.02	0.33 ± 0.01	**0.29 ± 0.02**
Penny	0.24 ± 0.04	0.32 ± 0.03	0.31 ± 0.01	0.31 ± 0.06	0.32 ± 0.04	0.25 ± 0.03	0.40 ± 0.02	0.42 ± 0.02	**0.33 ± 0.03**
Howard	0.21 ± 0.01	0.20 ± 0.04	0.24 ± 0.03	0.23 ± 0.05	0.23 ± 0.04	0.13 ± 0.03	0.43 ± 0.02	0.44 ± 0.02	**0.27 ± 0.03**
Raj	0.25 ± 0.09	0.31 ± 0.03	0.26 ± 0.03	0.32 ± 0.04	0.26 ± 0.03	0.19 ± 0.03	0.43 ± 0.01	0.44 ± 0.01	**0.32 ± 0.03**
Amy	0.21 ± 0.08	0.26 ± 0.04	0.19 ± 0.04	0.22 ± 0.06	0.22 ± 0.04	0.04 ± 0.06	0.47 ± 0.03	0.46 ± 0.02	**0.26 ± 0.04**
Bernadette	0.29 ± 0.06	0.27 ± 0.05	0.20 ± 0.04	0.24 ± 0.04	0.27 ± 0.01	0.04 ± 0.03	0.48 ± 0.04	0.47 ± 0.02	**0.28 ± 0.03**
Stuart	0.18 ± 0.04	0.04 ± 0.05	0.00 ± 0.00	0.04 ± 0.04	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	**0.01 ± 0.01**
M. Cooper	0.31 ± 0.02	0.15 ± 0.11	0.04 ± 0.06	0.28 ± 0.07	0.05 ± 0.07	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	**0.07 ± 0.04**
Priya	0.11 ± 0.08	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.03	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	**0.00 ± 0.00**
Beverley	0.13 ± 0.10	0.04 ± 0.03	0.02 ± 0.03	0.03 ± 0.05	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	**0.01 ± 0.02**
Emily	0.12 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.03 ± 0.04	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	**0.00 ± 0.01**
Arthur	0.11 ± 0.06	0.39 ± 0.22	0.07 ± 0.10	0.13 ± 0.12	0.02 ± 0.04	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	**0.09 ± 0.07**
Linear Probing									
Name	**Human Ann.**	**Roberta**	**Mpnet**	**Struct-BERT**	**Luar**	**Lisa**	**Style-Roberta-P. (MovieClips)**	**Style-Mpnet-P. (MovieClips)**	**AVG.**
Sheldon	0.30 ± 0.00	0.28	0.28	0.29	0.29	0.23	0.34	0.29	**0.29 ± 0.03**
Leonard	0.22 ± 0.01	0.24	0.21	0.21	0.25	0.18	0.28	0.21	**0.23 ± 0.03**
Penny	0.24 ± 0.04	0.19	0.25	0.25	0.27	0.19	0.23	0.29	**0.24 ± 0.04**
Howard	0.21 ± 0.01	0.02	0.09	0.07	0.09	0.02	0.13	0.25	**0.10 ± 0.08**
Raj	0.25 ± 0.09	0	0	0.06	0.09	0.03	0.23	0.27	**0.10 ± 0.11**
Amy	0.21 ± 0.08	0	0	0	0.02	0	0.22	0.17	**0.06 ± 0.09**
Bernadette	0.29 ± 0.06	0	0	0	0.02	0	0.26	0.25	**0.08 ± 0.12**
Stuart	0.18 ± 0.04	0	0	0	0	0	0.05	0.10	**0.02 ± 0.04**
M. Cooper	0.31 ± 0.02	0	0	0	0	0	0.20	0.12	**0.05 ± 0.08**
Priya	0.11 ± 0.08	0	0	0	0	0	0	0	**0.00 ± 0.00**
Beverley	0.13 ± 0.10	0	0	0	0	0	0.17	0.06	**0.03 ± 0.06**
Emily	0.12 ± 0.01	0	0	0	0	0	0	0	**0.00 ± 0.00**
Arthur	0.11 ± 0.06	0	0	0	0	0	0	0.14	**0.02 ± 0.05**

TABLE 3 – Scores F1 pour chaque personnage. Les scores moyens sont indiqués pour chaque personnage, hors annotation humaine. Les couleurs représentent les niveaux de performance : blanc pour des scores $\geq$ 0,30 (resp. 0,20 pour le linear probing), gris clair pour des scores dans]0,00 : 0,30[(resp.]0,00 :0,19[pour le linear probing), et gris foncé pour des scores de 0,00.

de la détection manuelle du style. Ces résultats mettent en évidence la dimension subjective de l'identification stylistique et confirment la complexité de la tâche.