Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Prédiction des pauses dans les données d'écriture en temps réel

Ioana-Madalina Silai[1] Kehina Manseri[2] Iris Eshkol-Taravella[3]
MODYCO, 200 Av. de la Republique, 92001 Nanterre, France
(1) madalina.silai@icloud.com, (2) manserikehina@gmail.com,
(3) ieshkolt@parisnanterre.fr

RÉSUMÉ

Cette étude explore la prédiction des pauses dans des données d'écriture enregistrées en temps réel. Deux hypothèses sont testées : (1) les pauses dépendent du contenu lexical des bursts, et (2) les catégories morpho-syntaxiques (POS) influencent leur distribution. Après prétraitement linguistique, plusieurs techniques de classification sont testées. CamemBERT atteint jusqu'à 90 % de précision en classification binaire, suggérant un lien fort entre structure linguistique et pauses.

ABSTRACT

Predicting pauses in real-time writing data

This study investigates pause prediction in real-time writing data. Two hypotheses are tested : (1) pauses are influenced by the lexical content of bursts, and (2) morpho-syntactic categories (POS) affect pause distribution. After linguistic preprocessing, several classification techniques are applied. CamemBERT reaches up to 90% accuracy in binary classification, suggesting a strong link between linguistic structure and pauses.

MOTS-CLÉS : séquences d'écriture, pauses, textualisation, écriture en temps réel, segmentation automatique, prediction automatique.

KEYWORDS: writing bursts, pauses, textualisation, keystroke logging, automatic segmentation, predictive modeling.

1 Introduction et contexte du travail

Le processus de textualisation, c'est-à-dire la structuration de la pensée en langage écrit, reste un objet d'étude central mais complexe, notamment en ce qui concerne les mécanismes cognitifs et linguistiques qui le sous-tendent (Cislaru & Olive, 2018; Flower & Hayes, 1981). L'écriture alterne entre phases de production, appelées bursts, et pauses, dont les plus longues (plus de 1,5–2s) signalent souvent des moments de planification ou de prise de décision, influencés par des contraintes linguistiques et événementielles (Matsuhashi, 1987; Schilperoord, 2002). Nous posons l'hypothèse que ces pauses longues ne sont pas aléatoires mais corrélées à des unités linguistiques et à des événements comme les révisions. Le but de cette étude est de modéliser les dynamiques du processus d'écriture en analysant les contraintes qui influencent l'apparition des pauses.

Le corpus analysé est issu du projet ANR Pro-TEXT. Il est composé de 33 textes rédigés par des étudiants en psychologie, enregistrés en temps réel avec Inputlog (Leijten & Van Waes, 2013). L'alternance temporellement linéaire entre pauses et production est interrompue régulièrement par des disfluences qui prennent la forme de révisions et marquent la non-linéarité spatiale du processus, se manifestant par des retours en arrière et des modifications du texte déjà produit. Ces disfluences constituent en moyenne 20 % des séquences selon des données antérieures (Cislaru & Olive, 2018). Nous distinguons ainsi trois types de bursts : production (P), révision (R) et révision de bord (RB)

Actes de l'atelier Traitement de données langagières dynamiques
par les outils et méthodes du TAL 2025 (DYN-TAL), pages 28–33.

pour une révision appliquée au burst immédiatement précedent. Deux critères sont testés : (1) lexico-sémantique, via les embeddings Word2Vec et CamemBERT (Mikolov *et al.*, 2013; Martin *et al.*, 2020), et (2) syntaxique, via les catégories POS produites par Spacy (Honnibal & Johnson, 2015). Ces données sont enrichies automatiquement par des variables linguistiques et temporelles issues de l'enregistrement (e.g. longueur du burst, nombre de suppressions, fréquence des mots (Hermit, 2016), chunks annotés par SEM (Dupont & Plancq, 2017)). Il est important de mentionner ici que l'annotation automatique a été possible seulement pour les mots complets présents dans les bursts. La Figure 1 montre l'état des données après ces pré-traitements réalisés [1]. À cause du fait qu'un burst peut contenir des actions differentes (suppression, écriture, ajout des lignes vides etc.), successives dans le temps, mais pas nécessairement dans l'espace, un seul burst peut être divisé sur plusieurs lignes. Le corpus final contient 4429 bursts, répartis en 80% pour l'entraînement et 20% pour l'évaluation.

n_burst	burstStart	burstDur	pauseDur	totalActions	totalChars	finalChars	totalDeletions	posStart	posEnd	docLen	categ	charBurst	POS	DepRel	Frequencies	Frequencies_in_text	Relative_frequencies_in_text
2	941.303	8.67	5.7	45	45	45	0	0	45	46	P	Dans le cadre de la médecine traditionnelle,	ADP,DET,NOUN,ADP,DET,NOUN,ADJ	nmod	4548946,3170,7225478,4134008,6524,785	1,6,1,21,12,12,1	0.00234,0.01402,0.00234,0.04907,0.02804,0.02804,0.00234
3	955.671	7.3	8.87	31	31	31	0	45	76	77	P	nous voici face à un dillemme.	PRON,VERB,NOUN,ADP,DET,NOUN	nsubj,obl:arg	1275361,71674,33918,3534420,3752833	3,1,3,10,3,1	0.00701,0.00234,0.00701,0.02336,0.00701,0.00234
4	971.839	21.78	8.49	100	97	94	3	76	128	171	P	Beaucoup critiqué cette dernière s'avère parfois ene	PROPN,VERB,DET,ADJ,ADV	nsubj	384,537137,83464,43775	1,1,5,3,4	0.00234,0.00234,0.01168,0.00701,0.00935
4	971.839	21.78	8.49	100	97	94	3	128	127	171	P	⌫					
4	971.839	21.78	8.49	100	97	94	3	127	170	171	P	effet fort efficace sur certaines points,	NOUN,ADV,ADJ,ADP	nmod	30167,50753,4301,801937	4,1,3,4	0.00935,0.00234,0.00701,0.00935
4	971.839	21.78	8.49	100	97	94	3	170	169	171	P	⌫					

FIGURE 1 – Structure du jeu de données

2 Prédiction de la pause à partir de mots qui l'entourent

2.1 Prédiction des valeurs continues : prédire la durée de la pause

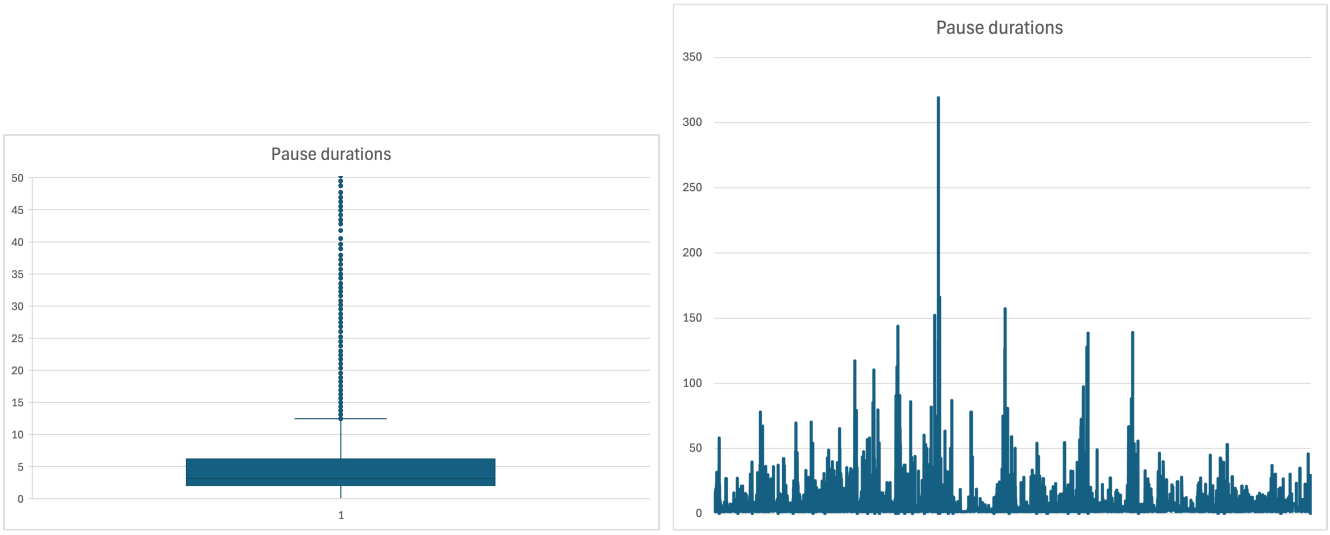
Les expériences ont débuté avec un objectif ambitieux : prédire la durée des pauses longues.Les données textuelles ont été vectorisées avec Word2Vec (Mikolov *et al.*, 2013), en moyennant les vecteurs des mots dans chaque unité de texte (burst, POS, chunks). Plusieurs modèles ont été testésde la régression linéaire à des approches adaptées aux données bruitées (Decision Tree, Random Forest, réseaux de neurones simples et RNN), ainsi qu'une version avec embeddings CamemBERT (Martin *et al.*, 2020), en lien avec notre première hypothèse.

Les premières expériences visaient à prédire une pause à partir du burst associé. Le corpus comprenait les informations linguistiques retenues (voir Figure 1) ainsi que les pauses non-significatives de chaque burst, car celle à prédire est la dernière du burst. La régression linéaire a donné un coefficient de détermination $R^2 = -0.6802$. Un score de $R^2 = 1$ indique un ajustement parfait, $R^2 = 0$ correspond à une prédiction égale à la moyenne, et une valeur négative indique que le modèle est moins précis que cette moyenne. L'arbre de décision, mieux adapté aux relations non linéaires, obtient un $R^2 = -1.07$ et une MSE (*mean squared error*) de 3.08 en validation croisée. Le modèle Random Forest, plus robuste aux variations aléatoires, atteint $R^2 = -0.07$ et une MSE de 1.59 : une amélioration notable mais encore insuffisante.

Face à ces résultats, nous avons réduit la variance des données et utilisé un RNN. La réduction de variance consistait à supprimer les pauses extrêmes à l'aide de l'IQR (borne inférieure = 5.075 s, borne supérieure = 14.005 s), et à réduire la dimension des vecteurs de 100 à 50. Figures 2a et 2b montrent la distribution des valeurs des pauses. Dans la figure 2a nous n'avons pas inclus les valeurs

1. Le nombre total de colonnes est 27, pour raisons de visualisation nous avons inclu ici que 18 colonnes.

les plus extrêmes, mais qui sont visibles dans la figure 2b.



(a) (b)

FIGURE 2 – Distribution des valeurs des pauses

Le RNN traite chaque séquence en conservant en mémoire les observations précédentes, ce qui est pertinent pour modéliser des relations temporelles. Plusieurs expériences ont été menées en faisant varier les hyperparamètres : nombre de couches (16 à 64), époch (10 à 100), batch size (16 à 256), pooling moyen, transformation logarithmique, attention, et dropout pour limiter le surapprentissage. Malgré tout, les performances sont restées faibles. Le meilleur score $R^2 = -0.0250$ est obtenu avec cinq couches, 100 époch, dropout de 0.5, et batch size de 8.
Un modèle séquentiel simple avec des embeddings CamemBERT et les caractéristiques linguistiques a aussi été testé. Le meilleur score $R^2 = -0.2389$ est obtenu avec 100 époch, 4 couches et un batch size de 8. De nombreux ajustements ont été faits, mais la faible corrélation entre les caractéristiques et la variable cible a limité les performances.
La prédiction reste donc difficulie, notamment en raison de la forte variabilité de la variable cible (cf. figure 2b) et du nombre limité d'exemples (4 429 bursts, dont certains très courts), ce qui limite la capacité des modèles à apprendre une tendance fiable.

2.2 Classification

Prédire précisément la durée des pauses étant difficile, nous avons reformulé le problème en une tâche de classification. Deux stratégies ont été explorées : regrouper les pauses en 5 ou 10 catégories de durée, ou simplement distinguer présence/absence de pause (classification binaire).

2.2.1 Classification en 5 et 10 catégories

Les pauses ont été réparties automatiquement en 5 ou 10 classes en fonction de la durée de pause, de façon que les differentes classes contiennent un nombre similaire de bursts. Nous avons d'abord utilisé un classificateur Random Forest, combiné à des embeddings CamemBERT, avec deux types d'entrée : burst avant la pause, ou pause avant le burst. Les meilleurs résultats sont obtenus dans le premier cas, avec une précision de 0,25 pour la classification en 5 catégories, et 0,13 pour celle en 10.

La suite des expériences s'est concentrée sur 5 catégories, en introduisant un réseau séquentiel entraîné sur des séquences textuelles étiquetées (ex : "cat_2 L'intention de l cat_5..."). Malgré un entraînement avec early stopping, dropout à 0,5, batch size 16, et 100 epochs, la précision reste faible (0,25, perte de 1,61).

Algorithme	Hyperparametres / autres details	Accuracy
Gradient Boosting classifier	learning_rate = 0.1 , pause après burst	0.17
Gradient Boosting classifier	learning_rate = 0.1 , pause avant burst	0.21
Random forest classifier	pause après burst	0.21
Random forest classifier	pause avant burst	0.24
Sequential neural network	dropout rate, early stopping, batch_size=32, 100 epochs - pauses avant burst	0.21
Sequential neural network	dropout rate, early stopping, batch_size=256, 100 epochs - pauses avant burst	0.23

TABLE 1 – Résultats de la classification en 5 catégories

Random Forest obtient la meilleure précision (0,2405), suivi du réseau neuronal avec batch size 256 (0,2328). Ces résultats suggèrent que les pauses précédant un burst sont plus informatives que celles qui le suivent, soulignant une dimension anticipatoire du processus d'écriture. Ce constat appuie l'idée d'un traitement incrémental du langage (Cislaru & Olive, 2018; Christiansen & Chater, 2016), en tension avec des modèles plus non-linéaires comme celui de Flower & Hayes (1981).

2.2.2 Classification binaire

Les résultats décevants obtenus avec CamemBERT pour la classification en cinq catégories nous ont conduit à explorer une classification binaire : pause présente / absente. Les textes ont été transformés pour annoter après chaque token par 0 (pas de pause) ou 1 (pause). Cette transformation a généré un déséquilibre (451 pauses vs. 2748 non-pauses), corrigé par suréchantillonnage de la catégorie pause et sous-échantillonnage de la catégorie non-pause, ainsi que par la pondération de la fonction de perte. En introudisant plusieurs copies des tokens suivis par pause tout en supprimant ceux qui ne le sont pas, nous obtenons ainsi 921 pauses et 5554 non-pauses. Avant ajustement des hyperparamètres, la précision était de 0,65 (pause) et 0,68 (non-pause). Après réglage (20 époques, taux $2e$-5), elle atteint 0,82, montrant que cette tâche est plus prometteuse que la classification à cinq classes. En revanche, la dépendence séquentielle est perdue ainsi que la notion de burst, car dans ce cas nous utilisons seulement les textes finaux dans lesquels nous ajoutons l'information sur la présence ou absence d'une pause.

3 Prédiction de la pause à partir des catégories POS des bursts

Nous testons ici si les pauses sont influencées par les séquences de catégories POS des bursts. Les mots sont remplacés par leurs POS (obtenus via SpaCy), puis vectorisés avec CamemBERT. En classification binaire, cette méthode atteint une précision de 0,90 (pause) et 0,92 (non-pause). On constate que les adjectifs sont les POS les plus fréquents avant une pause, tandis que les noms dominent après celle-ci. Une version randomisée des pauses (placées aléatoirement) donne des

résultats nettement inférieurs (0,31 pour pause), confirmant que le modèle capte une régularité liée aux structures morphosyntaxiques. Ces résultats suggèrent une corrélation entre pauses et structure grammaticale.

4 Conclusion et perspectives

Cette étude exploratoire visait à prédire les pauses dans des données d'écriture en temps réel. Deux hypothèses ont été testées : (1) l'influence des bursts sur les pauses, et (2) l'effet des POS. Les approches de régression se sont révélées peu concluantes (scores inférieurs à la moyenne). La classification en cinq catégories améliore légèrement la performance (précision de 0,25), mais reste limitée.

Bien que des forêts aléatoires aient été utilisées pour la classification en 5 et 10 catégories, des travaux futurs pourraient affiner l'analyse de l'importance des variables. Il serait notamment utile d'explorer la hiérarchie des décisions à l'intérieur des arbres, ou de recourir à des Partial Dependence Plots pour visualiser l'effet marginal de chaque variable. Ces approches offriraient une interprétation plus fine du rôle des variables dans la structuration des pauses.

Par ailleurs, la pertinence des classes définies par des seuils arbitraires pourrait être remise en question. Une alternative consisterait à prédire directement les trois types de bursts identifiés, en explorant par exemple s'ils peuvent être anticipés à partir de la durée des pauses (via k-moyennes, matrice de confusion, ou ANOVA). Cela permettrait de relier plus étroitement les caractéristiques temporelles aux types de production.

La classification binaire, notamment avec CamemBERT, atteint des performances bien supérieures (jusqu'à 90% de précision). Les POS s'avèrent particulièrement informatives, ce qui confirme une régularité morphosyntaxique et soutient l'hypothèse d'un processus d'écriture incrémental. Des pistes à explorer incluent la sélection automatique des variables, l'expérimentation avec d'autres modèles (RoBERTa, mBERT), et l'ajout d'informations syntaxiques (chunks, dépendances).

Enfin, ces travaux pourraient déboucher sur des applications concrètes, comme des outils d'aide à la rédaction ou des supports pédagogiques visant à améliorer la fluidité textuelle. La généralisabilité des résultats mériterait également d'être testée dans d'autres genres discursifs, tels que la rédaction académique ou la messagerie instantanée, afin de mieux cerner leur portée et leurs limites.

Remerciements

Nous remercions les relecteurs pour leurs commentaires pertinents, qui ont permis d'enrichir cette étude et de dégager de nouvelles pistes de recherche.

Références

CHRISTIANSEN M. H. & CHATER N. (2016). *Creating language*. The MIT Press.

CISLARU G. & OLIVE T. (2018). *Le processus de textualisation*. De Boeck Supérieur.

DUPONT Y. & PLANCQ C. (2017). Un 'etiqueteur en ligne du français. In *24e Conf'erence sur le Traitement Automatique des Langues Naturelles (TALN)*, p. 15–16.

FLOWER L. S. & HAYES J. R. (1981). A cognitive process theory of writing. *College Composition & Communication*.

HERMIT D. (2016). Frequencywords.

HONNIBAL M. & JOHNSON M. (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 1373–1378.

LEIJTEN M. & VAN WAES L. (2013). Keystroke logging in writing research : Using inputlog to analyze and visualize writing processes. *Written Communication*, **30**(3), 358–392. DOI : 10.1177/0741088313491692.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE , SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MATSUHASHI A. (1987). *Writing in Real Time : Modelling Production Processes*. Norwood, NJ : Ablex.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space.

SCHILPEROORD J. (2002). On the cognitive status of pauses in discourse production. In T. OLIVE & M. C. LEVY, Éds., *Contemporary Tools and Techniques for Studying Writing*, p. 61–87. Dordrecht : Kluwer Academic Publishers. DOI : 10.1007/978-94-010-0468-8_4.