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

Pré-traiter les données d'écriture en temps réel

Kehina Manseri[1] Amandine Jouvenel[2]
(1) MoDyCo, 200 Avenue de la République, 92 001 Nanterre, France
(2) Clesthia, 4 rue des Irlandais, 75 005 Paris, France
manserikehina@gmail.com, amandine.jouvenel@sorbonne-nouvelle.fr

RÉSUMÉ

Traiter les données d'écriture en temps réel est une tâche complexe, ces dernières combinant des informations spatiales et temporelles, et conservant les traces du processus d'écriture. Les outils actuels de traitement des données linguistiques - comme les tokenizeurs, les étiqueteurs morpho-syntaxiques ou les parseurs syntaxiques - ne sont pas conçus ni entraînés pour traiter ce type de corpus et de données à haute dimensionalité. Cela soulève donc la problématique du traitement automatique des données d'écriture dynamique. Le travail présenté ici constitue une série de premières expériences portant sur l'étiquetage morpho-syntaxique et le chunking de ces données. Il vise à annoter les données tout en prenant en compte les traces de l'écriture en temps réel, appelées ici disfluences.

ABSTRACT

Pre-processing real-time writing data

Pre-processing real-time writing data is a complex task, as it combines spatial and temporal information, and preserves traces of the writing process. Existing tools for linguistic data processing - such as tokenizers, part-of-speech taggers, and syntactic parsers - are not designed nor trained to handle this type of corpus and high dimensional data. This raises the challenge of processing dynamic writing data. The work presented here is a series of preliminary experiments focused on part-of-speech tagging and chunking of such data. Its goal is to annotate the data while accounting for traces of the real-time writing process, referred to here as disfluencies.

MOTS-CLÉS : Données d'écriture en temps réel, étiquetage morpho-syntaxique, chunking, pré-traitement, disfluences..

KEYWORDS: Real-time writing data, morpho-syntactic tagging, chunking, preprocessing, disfluencies..

1 Introduction

La recherche présentée porte sur les données d'écritures en temps réel et les disfluences qu'elles contiennent. Ces dernières sont de nature variée : suites de caractères saisies entre deux pauses (appelées ici « bursts »), pauses marquant une interruption de l'écriture, durée de production d'un burst ou d'une pause, etc. De plus, le processus d'écriture varie considérablement d'un scripteur à l'autre, selon son profil, son environnement ou encore le sujet traité. Les données produites se révèlent ainsi particulièrement hétérogènes. Le présent travail propose une démarche méthodologique visant à annoter ces données complexes et à mettre en lumière leur potentiel pour des recherches futures sur les comportements d'écriture. Cette recherche s'inscrit dans le cadre du projet ANR ProText.

Actes de l'atelier Traitement de données langagières dynamiques
par les outils et méthodes du TAL 2025 (DYN-TAL), pages 23–27.

2 Présentation et traitement des données

2.1 Recueil des données brutes

La recherche présentée s'appuie sur des données recueillies lors d'expériences menées auprès d'étudiants en licence de psychologie. Il leur avait été demandé de rédiger des textes courts (compris entre 5 et 20 lignes en moyenne) sur des sujets prédéfinis : les frais de scolarité, le tabagisme à l'université ou encore la pollution. Les productions ont été réalisées sur Microsoft Word, tandis que les mouvements des scripteurs étaient enregistrés à l'aide du logiciel InputLog (Leijten & Van Waes, 2013), sous Windows. Chaque scripteur était identifié à l'aide d'un code composé de la première lettre du corpus, d'un symbole « + » ou « - » indiquant son niveau d'expertise sur le sujet, suivi d'un numéro d'identification : P+S2, R-23, F+14, etc.

2.2 Des fichiers IDFX aux CSV

Le corpus initial est constitué de 56 fichiers IDFX dans lesquels ont été enregistrés les touches pressées lors de l'écriture, ainsi que les textes dans leur forme finale tels qu'ils ont été sauvegardés sur Word par les utilisateurs. Les fichiers IDFX bruts ont été convertis en un format tabulaire à l'aide d'un script Python : chaque ligne représentant un burst (Figure 1). L'identification de pauses a été fondée sur les informations temporelles à partir d'un seuil prédéfini de 1,5 seconde, reflétant la distinction entre les pauses causées par des contraintes mécaniques et celles causées par des processus cognitifs (Schilperoord, 2002). Ces pauses de plus de 1,5 seconde sont incluses dans les disfluences étudiées. Pour chaque burst, diverses caractéristiques ont été enregistrées, dont : l'ID du participant, le niveau de contrainte, l'ID et la durée du burst, la durée de la pause, la chaîne de caractères tapée. Des attributs spécifiques aux disfluences présentes au sein du burst ont été également intégrés : le nombre de caractères et d'espaces tapés, le nombre de caractères après suppression, le nombre de suppressions. La classification de chaque burst s'est basée sur les trois types suivants : P (production), R (révision) et RB (révision de bord) (Cislaru & Olive, 2018). Nous procédions ainsi ici à une première taxonomie de structures et d'opérations effectuées sur le clavier.

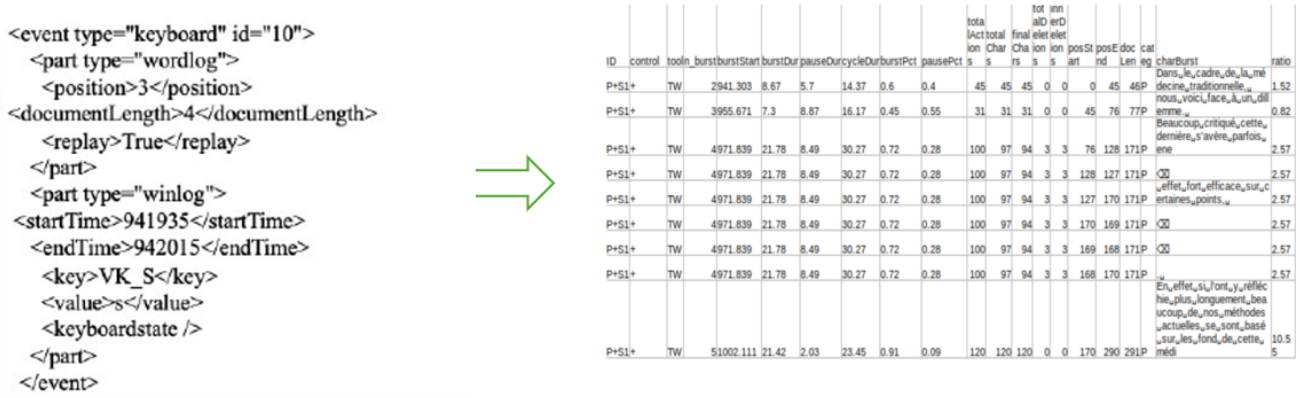

FIGURE 1 – Des fichiers IDFX aux fichiers CSV.

3 Annotation linguistique des données

Une fois nos données structurées de manière à identifier des pauses significatives et autres comportements d'écriture, nous avons cherché à les annoter linguistiquement à l'aide de méthodes permettant de contrer d'éventuelles limites spécifiques aux données d'écriture en temps réel.

3.1 Étiquetage morpho-syntaxique

Une première série d'expériences a porté sur l'étiquetage morpho-syntaxique des données. L'objectif était d'identifier les catégories grammaticales (Part-of-Speech, ou POS) des tokens impliqués dans des phénomènes de disfluence, tout en signalant explicitement ces derniers.

Pour ce faire, nous avons évalué trois étiqueteurs : Stanza (Qi *et al.*, 2020), Spacy et TreeTagger (Schmid, 1994). L'objectif était d'observer leur comportement face à des données contenant des fautes d'orthographe ou des unités tronquées. Stanza et Spacy tentent de deviner l'étiquette morpho-syntaxique en s'appuyant sur le contexte. À l'inverse, TreeTagger dispose d'une étiquette « UNKNOWN » qu'il attribue à un token non reconnu. Cependant, lorsqu'un token est intégré dans un contexte plus large, TreeTagger tente lui aussi de lui assigner une étiquette POS. Dans l'état actuel, aucun de ces trois outils ne s'avère pleinement adapté au traitement des données d'écriture dynamique. En effet, l'analyse de telles données nécessite un étiquetage homogène des tokens inconnus, essentiel pour une étude fiable des disfluences.

Pour chaque texte, nous avons alors construit un lexique à partir de l'extraction des tokens présents dans le texte sauvegardé dans sa forme finale, enregistré par le sujet à la fin de l'expérience. Lors de l'annotation, tout mot ne figurant pas dans ce lexique est étiqueté « UNKNOWN », tandis que les mots connus sont annotés à l'aide de Spacy, plus rapide que Stanza et TreeTagger. Cette méthode évite des étiquettes arbitraires pour les mots inconnus et garantit une reproductibilité de l'étiquetage, adaptable à tout nouveau texte pour lequel un lexique des mots utilisés est disponible.

L'analyse des mots annotés UNKNOWN a permis de repérer et classifier 5 types de disfluences : suppression d'un caractère à l'intérieur d'un mot, ajout d'un caractère à un mot précédent, ajout d'un espace à un mot précédent, suppression d'une chaîne via une ou plusieurs suppressions (delete et backspace), insertion d'un mot ou d'une chaîne entre deux mots.

Les résultats de l'étiquetage prenant en compte ces cinq formes de disfluences sont présentés dans la Figure 2. Les disfluences observées les plus fréquentes ont été les suppressions d'un caractère à l'intérieur d'un mot (57 %) et les insertions de chaînes (39 %).

3.2 Chunking

Une deuxième série d'expériences a porté sur le chunking des données, c'est-à-dire l'identification de la structure syntaxique superficielle d'un énoncé à travers la reconnaissance de ses constituants minimaux, sans spécification de la structure interne ni de la fonction grammaticale. Cette tâche, déjà reconnue comme pertinente pour l'analyse de données non linéaires telles que l'oral (Gadet, 1998), s'avère particulièrement adaptée aux spécificités de l'écriture en temps réel.

Étant donné que le chunking repose sur des séquences de tokens plus longues - un chunk pouvant regrouper plusieurs unités -, chaque texte a été reconstruit à partir des colonnes d'un fichier CSV, en respectant l'ordre chronologique des événements capturés dans les fichiers IDFX. Des annotations spécifiques ont ensuite été ajoutées au sein des bursts pour marquer les disfluences : les suppressions

Id	n_burst	Token_texte	POS Supposé	Pos réel	Lemme	Erreur	Details	Correction
P+S1	6	encà⌫ore	ADJ	Unknown	or	TRUE	Suppression de caractères à l'intérieur d'un mot	encore
P+S1	9	s	PRON	Unknown	s	TRUE	Lettre unique appartenant à "remède"	remèdes
P+S13	29	⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫	PROPN	BACKSPACE	⌫	TRUE	11 backspaces supprimant 's et autres'	plante
P+S19	81	⌦	AUX	DELETE	⌦	TRUE	1 delete (⌦) supprimant ' '	non
P-S10	26	qui	PRON	PRON	qui	TRUE	Mot inseré entre 'l'université' et 'comportera'	l'université qui comportera
P-S10	32	dans	ADP	ADP	dans	TRUE	Partie de la chaine ' dans la faculté' inserée entre 'fumoirs' et "	fumoirs dans la faculté

FIGURE 2 – Résultats de l'étiquetage morpho-syntaxique.

(~), les insertions d'un seul caractère pendant les révisions (<>), les insertions de chaînes de caractères dans les révisions ({}), les pauses (|). Ces textes annotés ont été soumis à l'outil SEM (Dupont & Plancq, 2017), qui réalise le chunking à partir d'un étiquetage morpho-syntaxique préalable. Dans un premier temps, les symboles visibles représentant les disfluences ont perturbé les performances de l'outil, ce dernier les interprétant comme des éléments textuels à part entière. Pour remédier à cela, les symboles ont été remplacés par des caractères invisibles, ce qui a permis d'améliorer significativement la précision du chunking, cette dernière passant de 0.65 à 0.95 dans les meilleurs cas.

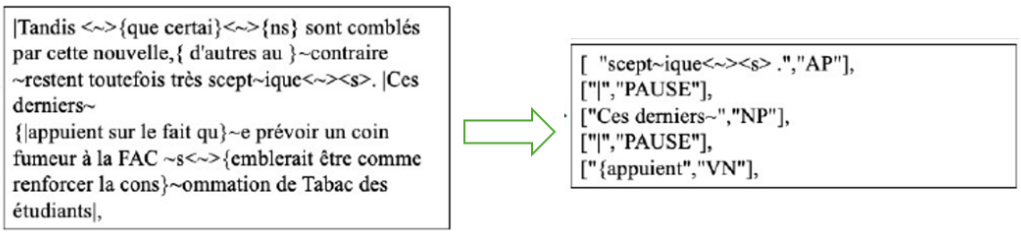

FIGURE 3 – Reconstruction et chunking des textes.

Les textes reconstruits présentaient ainsi une double annotation : les segments syntaxiques identifiés via le chunking et les types de disfluences rencontrées. Cette richesse d'annotation a permis d'obtenir des premiers résultats quantitatifs venant compléter les observations de la phase d'étiquetage morpho-syntaxique. Il a notamment été observé que 43% des pauses significatives coïncidaient avec une frontière de chunk, suggérant un lien étroit entre la planification cognitive et la structuration syntaxique. D'autres résultats exploratoires mettent en lumière des asymétries intéressantes : par exemple, 19% des adjectifs apparaissent avant une pause, contre seulement 1% après, ce qui pourrait indiquer une phase réflexive plus longue suivant un groupe adjectival. Enfin, l'analyse des caractères insérés lors de révisions a révélé une forte récurrence de graphèmes liés aux marqueurs morphologiques du genre et du temps — notamment les lettres *s* et *é*. Cela suggère que les révisions ne sont pas aléatoires, mais qu'elles visent des zones grammaticalement sensibles du texte.

Conclusion

L'ensemble des expériences menées dans cette étude, de l'identification des disfluences à l'étiquetage morpho-syntaxique, en passant par le chunking, a permis de construire une méthode robuste et

reproductible pour l'annotation de données d'écriture en temps réel. Cette démarche ouvre des perspectives intéressantes à plusieurs niveaux. Sur le plan linguistique, elle met en évidence certains phénomènes récurrents (pauses, révisions, ajustements morphologiques) pouvant servir à l'étude de comportements précis ou encore de profils présentant des troubles liés à l'écriture. Nous pourrions par exemple évaluer quantitativement les disfluences observées chez un public dyslexique afin d'organiser l'apprentissage autour des erreurs les plus courantes. De plus, ces données annotées pourraient être utilisées dans des projets d'apprentissage automatique visant à prédire certains comportements, dont les pauses ou les fautes d'orthographe, pouvant ainsi avoir des applications dans les secteurs de la santé ou de l'éducation.

Références

CISLARU G. & OLIVE T. (2018). *Le Processus de textualisation : Analyse des unités linguistiques de performance écrite*. Champs linguistiques. Louvain-la-Neuve : De Boeck.

DUPONT Y. & PLANCQ C. (2017). Un étiqueteur en ligne du français (an online tagger for French). In I. ESHKOL-TARAVELLA & J.-Y. ANTOINE, Éds., *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. Volume 3 - Démonstrations*, p. 15–16, Orléans, France : ATALA.

GADET F. (1998). Claire blanche-benveniste, approches de la langue parlée en français. paris : Ophrys, 1997, 164 pp. 2 7080 0830 7. *Journal of French Language Studies*, **8**(1), 116–118. DOI : 10.1017/S0959269500000600.

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

QI P., ZHANG Y., ZHANG Y., BOLTON J. & MANNING C. D. (2020). Stanza : A python natural language processing toolkit for many human languages. In A. CELIKYILMAZ & T.-H. WEN, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 101–108, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-demos.14.

SCHILPEROORD J. (2002). On the cognitive status of pauses in discourse production. In T. OLIVE & C. M. 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.

SCHMID H. (1994). *Probabilistic Part-of-Speech Tagging Using Decision Trees*. Thèse de doctorat, Universität Stuttgart. Institut für maschinelle Sprachverarbeitung.