Analysing the Dynamics of Textualisation through Writing Bursts and Pauses in French

MOTS-CLÉS : séquences d'écriture, pauses, textualisation, analyse en temps réel, processus cognitifs, segmentation en unités linguistiques, prediction automatique.

KEYWORDS: writing bursts, pauses, textualisation, keystroke logging, French writing, cognitive processes, chunking, predictive modeling.

Abstract

— Our research is part of a larger study which seeks to deepen our understanding of textualisation by analysing the temporal dynamics of writing bursts and pauses in French, with and without constraints. By examining how sequences of linguistic events unfold in real-time, we aim to uncover how cognitive processes shape writing behaviour.

— Our research suggests that the linear articulation of linguistic sequences during textual production is influenced by (i) multifaceted constraints affecting language performance, and (ii) complex relationships that extend beyond traditional syntactic boundaries.

— This study draws from multiple experiments conducted on various subject groups, focusing particularly on psychology students in their BA program. We instructed these participants to write short texts on diverse topics like student fees, smoking on campus, and environmental pollution. Students were free to move through the text, delete, add, or modify any section as needed.

— Their keystrokes and behaviours were logged using InputLog software (Leijten & Van Waes, 2013), producing IDFX files which serve as a comprehensive record of their writing actions.

— The corpus used for this study contains 56 IDFX files, out of which 33 were fully processed for analysis.

1 Data Collection and Processing Challenges

The use of keystroke logging through InputLog enabled the capture of spontaneous writing behaviours. Unlike traditional data where text is represented in its final form, our data reflect a real-time sequence of writing events. Real-time writing analysis provides an in-depth view of how writing is constructed, but also presents multiple challenges. Since the participants

were free to revise, navigate, and modify their texts without restrictions, reconstructing the writing process to accurately reflect the sequence of cognitive events was complex. The IDFX files, which recorded each keystroke, cursor movement, and pause, had to be meticulously processed to maintain an accurate and detailed sequence of the writing process.

To handle this complexity, we first reconstructed the original texts as they were produced and saved by the participants, preserving both the chronology and type of events. This process involved a key-by-key reconstruction and the use of time-stamps to mark every pause, be it a short or long one. This approach offers a comprehensive picture of writing behaviour but also underscores the difficulties in parsing and understanding bursts of writing, particularly when text modifications are interspersed throughout the document. A burst is defined as a sequence of words or characters produced fluently without interruption, typically framed by pauses. However, since participants often made use of key combinations (e.g., Ctrl+s) and arrow keys to navigate to different sections of their texts, it was challenging to link a writing burst directly to a contiguous section of the text.

Furthermore, the non-linear nature of the text meant that bursts were not spatially adjacent but could be temporally connected. This necessitated careful reconstruction to ensure that all movements, edits, and insertions were properly tracked in the context of the original writing sequence.

2 Data Processing and Annotation

The raw IDFX files were processed into tabular form using a Python script, with each row representing a distinct burst of writing and containing extensive information about that burst. The division into bursts was made based on a predetermined threshold of 1.5 seconds, as it is generally admitted that shorter pauses are due to mechanic constraints and longer ones to cognitive constraints (Schilperoord, 2002). For each burst, we captured key details, including the participant ID, constraint level, burst ID, temporal information expressed in seconds such as start and end time, burst duration, pause duration, the full burst cycle duration (typing + pause), the proportion of time spent on typing versus pausing, and the actual string of characters typed. Additionally, the table contained information pertaining to the string of characters that was produced such as the number of keys pressed, the number of letters and spaces, the number of characters in the burst after all the deletions, the total number of deletions and the total number of deletions in the current cycle. Finally, text length at the burst's end, and a category for each burst were also included. Bursts were categorised into three distinct types :

1. **Production (P)** - Character additions or deletions that immediately follow the previous burst.
2. **Edge Revision (ER)** - Modifications made to the previous burst, either adding or removing characters at its boundary.

3. **Revision (R)** - Modifications involving earlier bursts, not directly preceding the current burst, often involving multiple sections of text.

3 Chunking and Linguistic Annotation

The reconstructed texts were processed further for linguistic analysis. Symbols were introduced to mark different actions within the bursts, such as deletions (~), single character insertions in a revision (<>), string insertions in a revision ({ }), and pauses (|). An example of a reconstructed text, containing these symbols is found in Figure 1. This symbol-based representation maintained the structure of the original table but excluded some information, such as exact pause lengths, the occurrence of pauses preceding a deleted string of characters, as well as the link between the pause and the location of the associated revision. We are currently developing methods to retain and represent this additional information in a more integrated format.

|La médecine traditionnelle |peut être un atout dans no~s société~<s> |du fait qu'elle soit |exercée de génération en génération~, |selon les cultures~. |Cela dit ~|la médecine traditionnelle |peut ~{être un frein dans}~ la recherche{, } {|dans l'évolution}~{| d}~~{|e nouveaux traitements par exemple.}

FIGURE 1 – Exemple of a reconstructed text

The symbol-annotated texts were then chunked using SEM, a French text annotation tool (Dupont & Plancq, 2017). Initially, the presence of symbols disrupted the accuracy of the chunking, as SEM misinterpreted these markers as part of the text. To overcome this, we replaced visible symbols with invisible characters (e.g. zero-width space, zero-width joiner, zero-width non-joiner, word joiner, function application), which enhanced the chunking process while retaining the structural integrity of the data. The annotated output was then converted into a JSON format for ease of further processing. This format includes the chunk types and associated events, preserving the multilevel structure required for a detailed analysis of pauses, writing behaviours and linguistic chunks. All the processing steps, from the output of the InputLog software to the output of SEM, are illustrated in Figure 2.

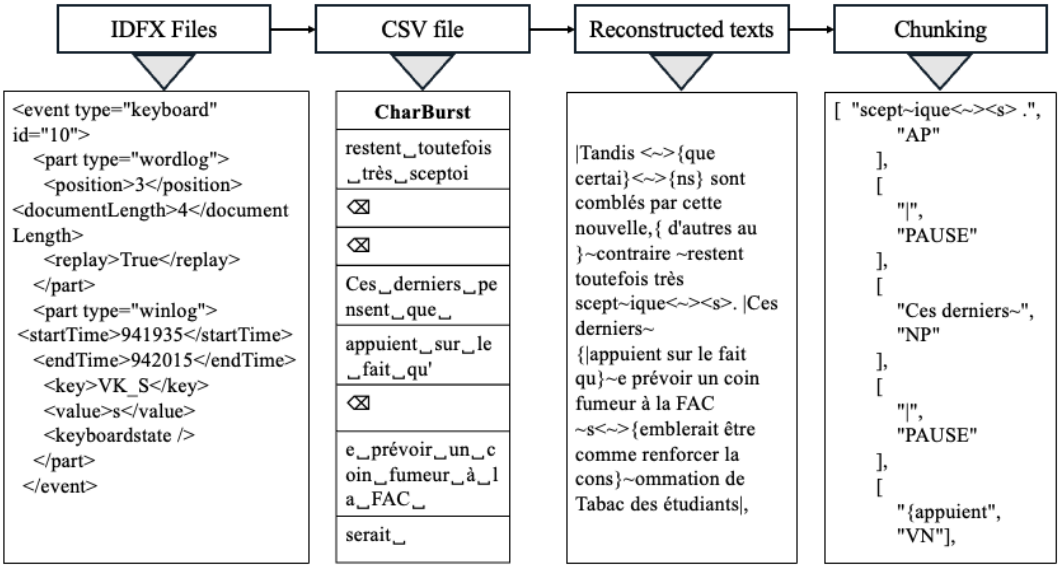

FIGURE 2 – Flow with Chunking [1]

4 Predictive Modeling of Pauses

To explore the potential for predicting pause lengths based on text and writing behaviours, we enriched our dataset with additional features. These included relative word frequency in individual texts, the absolute occurrences in the French language (Hermit, 2016) and in the text, part-of-speech tags using Spacy, and the pauses between each character typed within bursts. Textual data were vectorised using Word2Vec (Mikolov *et al.*, 2013), and multiple machine learning models were tested, including linear regression.

However, the complexity of the data and the presence of significant outliers reduced the performance of these models. As a result, we experimented with Recurrent Neural Networks (RNNs), although their effectiveness was limited due to the hierarchical nature of the data and the variability of writing sequences. A simpler neural network approach provided more manageable results, enabling iterative adjustments to the prediction script.

Ultimately, we transformed the prediction problem from a continuous one to a classification task by discretising pause lengths into five categories. This approach resulted in more accurate predictions, achieving an accuracy of 0.25 and a test loss of 1.61. Interestingly, predicting pauses preceding a burst produced better results than predicting pauses following it, suggesting that pauses are more influenced by the content and behaviour that follow rather than precede them. Ongoing work is focused on improving these results, with the

1. Due to space constraints, only one out of the 18 columns of the CSV file is represented here.

aim of exceeding 0.5 in predictive accuracy, and exploring the idea of a binary classification (significant versus non-significant pause). The full process, from the output of the InputLog software to the prediction algorithms, is illustrated in Figure 3.

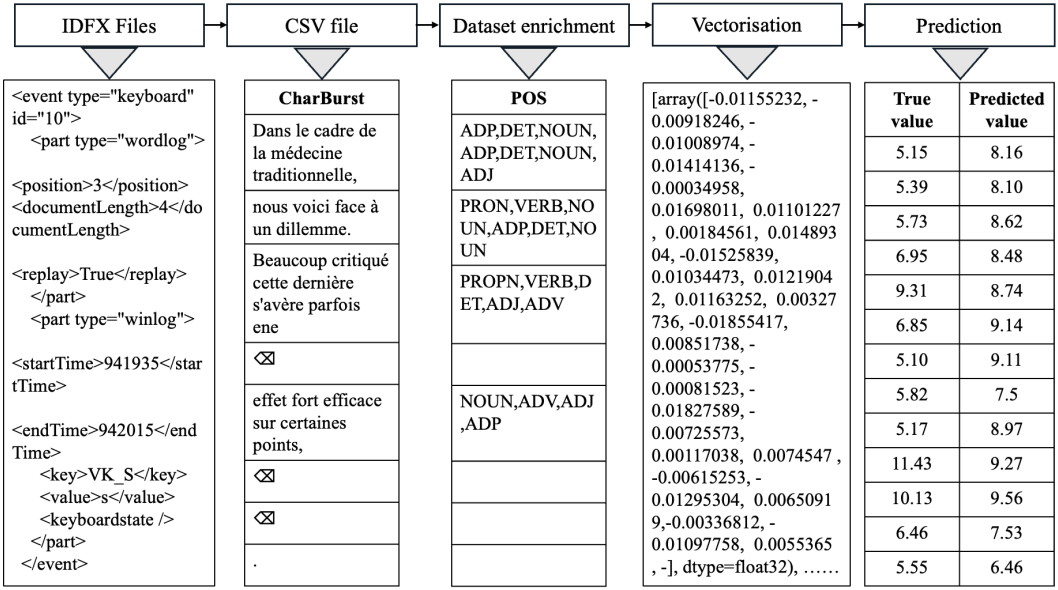

FIGURE 3 – Flow with Prediction [2]

5 Results : Chunking and Writing Behaviour Analysis

From the processed data in JSON format, we derived global metrics to describe the texts' features, such as the total number of pauses, chunk counts and lengths, overall text length, and the distribution of participants across constraint levels.

One initial hypothesis was that burst boundaries (marked by pauses) would align with chunk boundaries. Our findings showed a 43% overlap between pause and chunk boundaries, though further testing is needed to identify the significance of this result.

A secondary hypothesis posited a correlation between chunk types and pauses, aiming to identify language units likely to be found before or after significant pauses. For instance, results indicated that substantial pauses frequently occurred after adjectival chunks but were rarely found before them. More precisely, 19.6% of all adjectival chunks are located before

2. Due to space constraints, only one out of the 18 columns of the CSV file is represented here, as well as only one of added features for dataset enrichment and a small portion of a vector representing one word. We also did not represent all the predicted values, the ones here are the result of a Sequential Model neural network test, with 64 neurons.

the pause, and 0.9 % after the pause, while the rest are not directly located next to a pause. This pattern may point to a cognitive planning process occurring in post-adjective production. Additionally, our data revealed that the character most often added in revision bursts was “s,” implying that agreement errors, especially number agreement, were frequently corrected during the writing process.

6 Conclusion and Future Work

Our study offers valuable insights into the relationship between pauses, writing behaviours, and chunk boundaries, advancing our understanding of the cognitive mechanisms underlying textual production. By exploring both linguistic and behavioural aspects of writing, we have begun to identify patterns that link the nature of bursts and the pauses that surround them. The ongoing work aims to refine prediction models further and deepen the analysis of chunk-pause correlations. Improving predictive accuracy beyond 0.5, exploring other prediction strategies and expanding the generalisability of these findings to different writing contexts remain key objectives for future research.

Références

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.

HERMIT D. (2016). Frequencywords.

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.

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.