Textualisation Dynamics in French: A Study of Writing Bursts and Pauses

Kehina Manseri
manserikehina@gmail.com
University of Paris Nanterre

Ioana-Madalina Silai
madalina.silai@icloud.com
University of Paris Nanterre

Georgeta Cislaru
georgeta.cislaru@sorbonne-nouvelle.fr
Sorbonne Nouvelle University

Iris Eshkol-Taravella
ieshkolt@parisnanterre.fr
University of Paris Nanterre

Abstract

This study investigates the cognitive and linguistic processes underlying textualisation by analysing the temporal dynamics of writing bursts and pauses in French. Our key hypothesis is that long pauses during writing are influenced by cognitive parameters, shaped by linguistic and event constraints such as revisions. Using keystroke logging, we reconstructed real-time writing processes and analysed how pause durations relate to linguistic structures and revisions. Additionally, we developed predictive models to identify pauses based on text features, enriched with linguistic annotations. This work also emphasizes the methodological challenges of processing and analysing this unique dataset, given the limited research in natural language processing (NLP) on such data. Our findings highlight the relationship between linguistic units, cognitive planning, and temporal writing behaviours.

1 Introduction

The process of textualisation—the transformation of thoughts into structured written language—is central to writing research. However, understanding the cognitive and linguistic mechanisms underlying this process (see Cislaru and Olive (2018) for a summary) remains a complex challenge (Flower and Hayes, 1981). Writing involves multiple overlapping processes, such as planning, translating ideas into language, and revising, all of which interact dynamically. These processes are reflected in observable behaviours, including writing bursts (sequences of fluent text production) and pauses (interruptions in writing). Long pauses, in particular, are thought to signify moments of cognitive planning or linguistic decision-making, shaped by both linguistic constraints (e.g., sentence structure) and event-based constraints (e.g., revisions) (Matsuhashi, 1987), (Schilperoord, 2002).

Despite its importance, the study of textualisation has largely relied on analyses of final written products, neglecting the real-time dynamics of writing. Keystroke logging offers a powerful tool to address this gap, providing detailed records of writing processes, including keystrokes, pauses, and revisions ((Leijten and Van Waes, 2016). However, analysing and modelling such data pose significant methodological challenges, particularly given its non-linear and highly variable nature.

This study investigates the cognitive processes involved in writing by analysing the temporal dynamics of bursts and pauses in French. We hypothesize that long pauses during writing are determined by cognitive parameters, which are materialized through linguistic units and event constraints such as revisions. For this study, we examined chunks, defined by Abney (1991: 257) as "single content word surrounded by a constellation of function words, matching a fixed template". Our choice is determined by their identification in oral speech with "prosodic patterns", pauses being "most likely to fall between chunks" (ibid.) (Abney, 1991). One of our aims is therefore to check whether this association between chunks and pauses is also valid in writing. Our work also addresses the methodological challenge of processing and analysing keystroke data, which has received little attention in natural language processing (NLP) research.

Using keystroke logging data from psychology students tasked with writing short texts, we reconstructed

and segmented their writing processes to align linguistic units with pauses and revisions. We then developed predictive models to analyse pause dynamics, using both traditional and advanced machine learning techniques. In doing so, we aim to provide new insights into the interplay between cognitive planning, linguistic structures, and writing behaviour, while also contributing novel methodologies for processing complex writing datasets.

2 Data Collection and Processing Challenges

Keystroke logging with InputLog (Leijten and Van Waes, 2013) provided real-time data on spontaneous writing behaviors, offering a unique perspective on the construction process of text production. For this study, participants were allowed to freely revise, navigate, and edit their texts, generating a corpus of 56 IDFX files, with 33 fully processed for analysis. This decision introduced significant challenges for data processing. To accurately reconstruct the writing process, we preserved the chronological order and type of events by meticulously analyzing each keystroke, cursor movement, and pause recorded in the IDFX files.

To address these challenges, we developed a methodology to align keystroke events with linguistic structures. This reconstructive approach included using time-stamps to identify pauses of varying lengths, which helped delineate writing bursts—sequences of uninterrupted fluent production framed by pauses. However, participants frequently employed shortcuts (e.g., Ctrl+s) and navigation keys, making it challenging to associate bursts with specific text sections. Furthermore, the non-linear nature of text production often resulted in bursts that were temporally connected but not spatially contiguous.

Addressing these complexities required careful tracking of all movements, edits, and insertions to provide an accurate and comprehensive representation of the writing process, thus laying the groundwork for predictive modelling. Furthermore, this work contributes to NLP by addressing the methodological gap in processing keystroke data, which differs significantly from conventional text-based datasets.

3 Data Processing and Annotation

Raw IDFX files were converted into a tabular format using a Python script, with each row representing a distinct writing burst and its associated details. Bursts were identified based on a predefined pause threshold of 1.5 seconds, reflecting the distinction between pauses caused by mechanical constraints and those indicating cognitive processes (Schilperoord, 2002). For each burst, various features were recorded, including participant ID, constraint level, burst ID, start and end times, burst duration, pause duration, total cycle duration (typing and pause combined), typing-to-pausing time ratio, and the string of characters typed.

The dataset also captured specific attributes of the text produced, such as the number of keystrokes, letters, spaces, characters after deletions, total deletions, and deletions within the current cycle. Additional metrics included the text length at the end of the burst and a classification of the burst type.

Bursts were categorized into three distinct types (Cislaru and Olive, 2018):

1. **Production (P)**: Additions or deletions of characters that directly follow the previous burst.
2. **Edge Revision (ER)**: Changes made to the boundaries of the preceding burst, either adding or removing characters.
3. **Revision (R)**: Modifications affecting earlier bursts that are not immediately adjacent to the current one, often spanning multiple sections of text.

4 Chunking and Linguistic Annotation

One of the key hypotheses of this study is that long pauses during writing are influenced by cognitive planning demands, which are reflected in linguistic units such as chunks. These chunks are cohesive units of language based on parts of speech, such as nominal, adjectival, or verbal chunks, and they serve as critical markers for understanding the structure of text production. Chunks in this study were identified and annotated using SEM, a French linguistic annotation tool (Dupont and Plancq, 2017), which segments text into these syntactic units.

The reconstructed texts underwent further processing for linguistic analysis. Symbols were introduced to represent various actions within bursts, such as deletions (~), single-character insertions during revisions (<>), string insertions in revisions ({}), and pauses (|). An example of a reconstructed text with these symbols is provided in Figure 1. While this symbol-based representation preserved the structure of the original table, it omitted certain details, such as precise pause durations, pauses preceding deleted text, and the connection between pauses and their associated revisions. Methods are currently being developed to integrate this additional information into the dataset more effectively.

Next, the symbol-annotated texts were processed using SEM, a French text annotation tool (Dupont and Plancq, 2017). Initially, the symbols disrupted chunking accuracy as SEM interpreted them as part of the

text. To address this issue, the visible symbols were replaced with invisible characters (e.g., zero-width space, zero-width joiner, zero-width non-joiner, word joiner, and function application). This adjustment improved chunking accuracy while preserving the data's structural integrity.

The annotated data were then converted into a JSON format for streamlined further processing. This format captured chunk types and their associated events, maintaining the multilevel structure necessary for detailed analysis of pauses, writing behaviours, and linguistic chunks. The overall workflow, from InputLog output to SEM results, is illustrated in Figure 1.

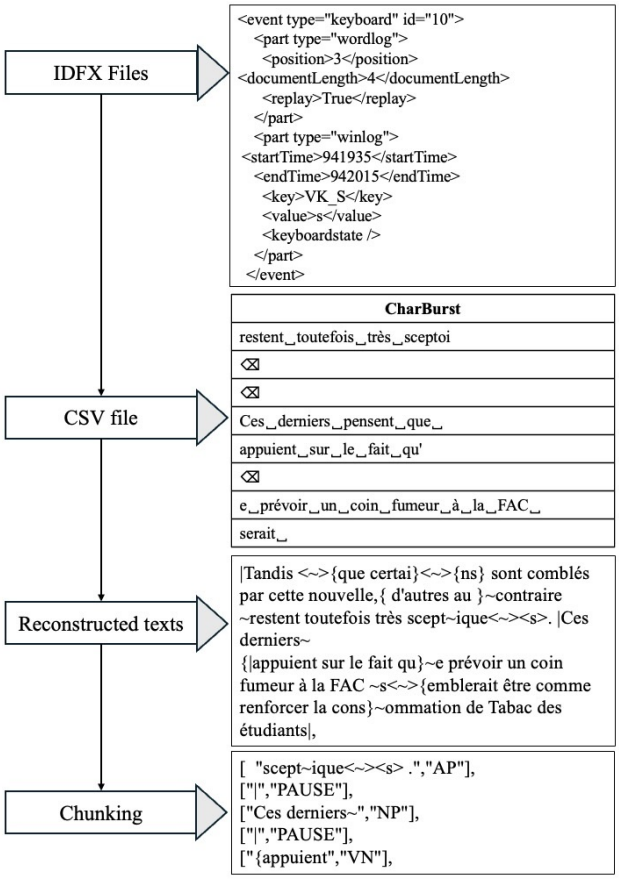

Figure 1: Flow with Chunking[1]

5 Predictive Modelling of Pauses

To investigate the potential for predicting pause durations based on text and writing behaviors, we enriched the dataset with additional features. These included relative word frequency within individual texts, absolute word occurrences in the French language (Hermit, 2016) and in the text, part-of-speech tags extracted using SpaCy, and the intervals between characters typed during bursts. Textual data were vectorized using Word2Vec embeddings (Mikolov et al., 2013), and various machine learning models, including linear regression, were evaluated.

The inherent complexity of the dataset, combined with the presence of substantial outliers, limited the effectiveness of traditional models. For instance, when attempting to predict exact pause durations, the model's accuracy dropped to approximately 0.11. To address this, we reframed the problem as a classification task by discretizing pause durations into five categories. These categories were defined to ensure an equal distribution of pauses across the dataset, reducing bias in the model. This adjustment improved performance, yielding an accuracy of 0.25. While still modest, this represents a significant improvement compared to earlier attempts at exact predictions.

Interestingly, predictions were more accurate when pairing the text of a burst with the pause that preceded it, rather than with the pause following it. This suggests that pauses are more influenced by the content and behavior immediately following the pause, rather than the burst immediately before the pause. This finding highlights the importance of context in modeling pause dynamics and warrants further exploration.

Additionally, we experimented with advanced models like CamemBERT (Martin et al., 2020) and RoBERTa (Liu et al., 2019), using them to tokenize the text and predict the presence of a pause (1) or no pause (0) between tokens. While further testing is required, these models showed promising results, achieving accuracy scores of around 90 percent in this binary classification task. The entire workflow, from InputLog outputs to predictive models, is shown in Figure 2.

6 Results: Chunking and Writing Behaviour Analysis

The processed JSON data allowed us to extract global metrics describing the texts, including the total number of pauses, chunk counts and lengths, overall text length, and the distribution of participants across different constraint levels.

One hypothesis was that the boundaries of writing bursts, marked by pauses, would align with chunk boundaries. Our analysis revealed a 43% overlap between pause and chunk boundaries, though further in-

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

[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.

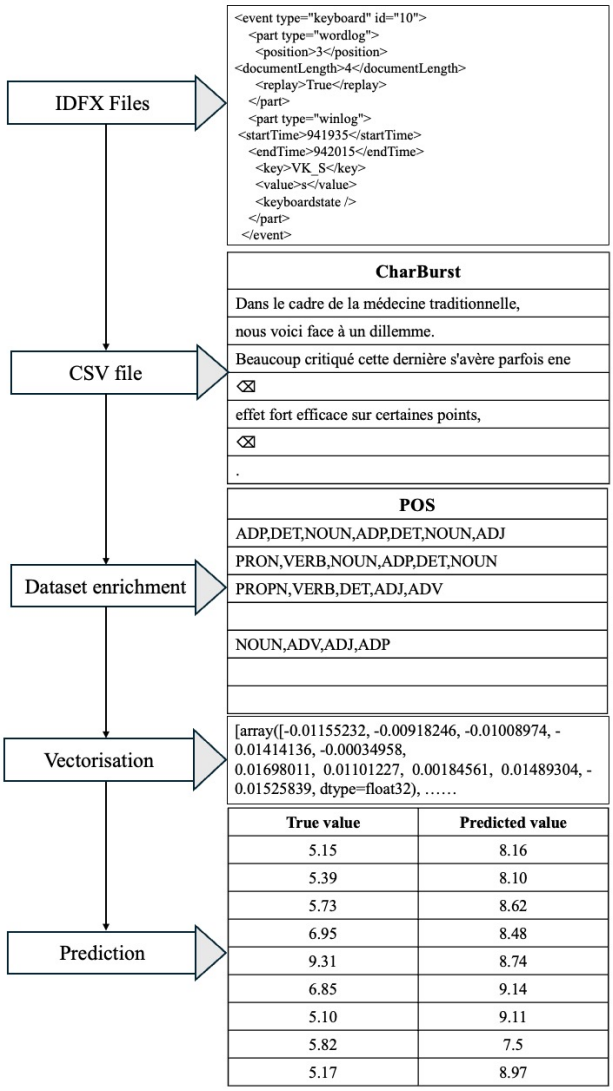

Figure 2: Flow with Prediction[2]

vestigation is needed to assess the significance of this finding.

A second hypothesis explored correlations between chunk types and pauses, aiming to identify linguistic units typically occurring before or after significant pauses. The results indicated that substantial pauses often followed adjectival chunks but rarely preceded them. Specifically, 19.6% of all adjectival chunks occurred before a pause, 0.9% followed a pause, and the remainder were not adjacent to any pause. This distribution suggests a cognitive planning process associated with post-adjective production.

Additionally, our data highlighted that the character most frequently added during revision bursts was "s," suggesting that corrections during writing often addressed agreement errors, particularly in number agreement.

7 Conclusion and Future Work

This study provides meaningful insights into the interplay between pauses, writing behaviours, and chunk boundaries, contributing to a deeper understanding of the cognitive processes involved in textual production. By examining both linguistic and behavioural dimensions of writing, we have begun to uncover patterns that connect the characteristics of bursts with the pauses framing them. In addition, the work emphasizes the methodological challenges of analysing keystroke data and proposes solutions to address the lack of NLP tools designed for this type of dataset.

Our findings underscore the importance of integrating linguistic and temporal features for understanding and predicting writing behaviours. By developing predictive models, we achieved promising results in pause classification, highlighting the potential for NLP techniques to model cognitive aspects of writing. Future work will refine these models further by incorporating additional linguistic features, such as syntactic complexity, phrase types, and lexical choices, as well as behavioural features like typing speed, revision patterns, and navigation behaviours. These enhancements aim to improve the generalizability of the models across different writing contexts and provide a deeper understanding of the interplay between cognitive processes and writing behaviou rs.

References

Abney, S. P. (1991). Parsing by chunks. In Berwick, R. C., Abney, S. P., and Tenny, C., editors, *Principle-Based Parsing*, volume 44 of *Studies in Linguistics and Philosophy*, pages 257–278. Springer, Dordrecht.

Cislaru, G. and Olive, T. (2018). *Le processus de textualisation: Analyse des unités linguistiques de performance écrite*. De Boeck Supérieur.

Dupont, Y. and Plancq, C. (2017). Un étiqueteur en ligne du français. In *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 15–16.

Flower, L. and Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, 32(4):365.

Hermit, D. (2016). Frequencywords.

Leijten, M. and Van Waes, L. (2013). Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication*, 30(3):358–392.

Leijten, M. and Van Waes, L. (2016). Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication*, 33(4):358–392.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, , Seddah, D., and 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.

Matsuhashi, A. (1987). *Writing in Real Time: Modeling Production Processes*. Ablex Publishing.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Schilperoord, J. (2002). On the cognitive status of pauses in discourse production. In Olive, T. and Levy, M. C., editors, *Contemporary Tools and Techniques for Studying Writing*, pages 61–87. Kluwer Academic Publishers, Dordrecht.