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

Predicting CEFR levels for learners of English with keylogging metrics, an exploratory study

Ahood Al Sawar[1] Erin Pacquetet[2] Cyriel Mallart[3] Andrew Simpkin[4] Nicolas Ballier[1]
(1) ALTAE, Université Paris Cité, 7 rue Thomas Mann, 75013 Paris, France
(2) SCIAM, 10 rue de Penthièvre, 75008 Paris, France
(3) INRIA, Campus de Beaulieu, 35042 Rennes, France
(4) School of Mathematical and Statistical Sciences, University of Galway, Irlande

ahoodswar2018@gmail.com, cyriel.mallart@inria.fr, erin.pacquetet@sciam.fr,
andrew.simpkin@nuigalway.ie, nicolas.ballier@u-paris.fr

Résumé

This paper describes a series of metrics developed to analyze writing strategies of learners of English and to possibly classify learner essays in relation to their level of English, expressed in terms of CEFR levels.

Abstract

Analyse exploratoire des traces numériques clavier pour la prédiction des niveaux d'apprenants
Cet article présente une typologie des métriques des traces numériques clavier en vue d'une analyse des stratégies d'écriture des différents profils d'apprenants appliquée à une tâche de prédiction du niveau CECRL.

Mots-clés : Traces numériques clavier, Jets textuels, corpus d'apprenants, DSprocessus d'écriture.

Keywords : Keystroke Logging, Textual Bursts, Learner Corpus, Writing Process.

1 Introduction

In this paper, we discuss the possibility of predicting CEFR levels on the basis of keylogging metrics. This set of metrics has been developed to characterize the behavior of learners, the type of writing processes by learners, to possibly associate those patterns to proficiency skills in English. The aim is to characterize learner writing behaviours and keystroke dynamics as captured by metrics based on bursts and linguistic units, as opposed to more biometry-centered metrics (Tappert & Villani, 2010) that operate at character level. The rest of the paper is structured as follows: Section 2 summarizes previous research on keylogging analysis of learner data for second language acquisition modeling. Section 3 presents our experimental design. We explain how keylogs can be used as one of the microservices analyzing learner data in a wider, bigger CMS based Moodle infrastructure. Section 4 presents our preliminary results on a prototype dataset. Section 5 discusses our results and concludes.

Actes de l'atelier Traitement de données langagières dynamiques
par les outils et méthodes du TAL 2025 (DYN-TAL), pages 7–11.

2 Previous Research on (learner) keylogging data

Within the Complexity Accuracy Fluency paradigm (Housen & Kuiken, 2009), many studies have addressed classification tasks based on spoken (Ballier et al., 2016) or written learner productions (Gaillat et al., 2020), where learner proficiency is usually expressed in terms of the CEFR reference framework (Council of Europe, 2001). To the best of our knowledge, no systematic review exists on the use of keylogs for L2 Second acquisition modeling. Previous research has mostly identified the location and the types of pauses as the main concepts or the main construct to operationalize planning, time management, and other skills.

Other studies followed Bayesian models to investigate writing processes and typing behaviour. (Conijn et al., 2019) suggested modeling writing and copy tasks with the use of Bayesian approach where she presented a process-based model of typographic error revisions which provided a systematic way to analyze how typists detect errors and correct them in real time. The Bayesian modelling showed a clear difference of writing patterns for the copy task and the writing task. In addition, (Roeser et al., 2024) examined typing disfluencies using Bayesian mixture models where they attempted to understand keystroke variability. Furthermore, (Pacquetet, 2024) computed fluency patterns using regression techniques and Bayesian modeling where she tried to fill the gap between linguistic properties and typing dynamics. She utilized analysis tools in R and Stan and the IKI dataset structure (interkey interval).

As a follow-up to previous investigations on learner writing (Gilquin, 2020; Gilquin, & Laporte, 2021) investigating constructions in learner essays and, more generally, the writing process in a corpus-based perspective (Gilquin, 2022).(Gilquin, 2024) examines the employment of keylogging to explore L2 writing processes. A key application examined in her study is the multiword unit (MWU) processing, where keylogging data shows that MWUs with pauses at their boundaries and not within them are probably processed systemically. (Garcés Manzanera, 2024) investigated the relationship between pauses and revision behaviour using keystroke logging software. He detected the writing process of 22 elementary students aged between 10-11 with A1-A2 CEFR levels while completing a task of picture description. His findings presented a negative correlation between pauses (P-bursts) frequency and text quality, while a positive correlation was found between revision bursts (R-bursts) with the quality of the text, which suggests the great importance of revision behaviour in boosting writing quality. While many studies have used keystroke logging in order to better comprehend writing behaviour, an integration between process-based metrics and CEFR-level classification is still underexplored. Most existing models do not systematically link dynamic typing features (e.g., burst size, backspaces, pause lengths) to English proficiency levels. We expected revision bursts to be of great importance for the CEFR classification, and our preliminary results support these expectations.

3 Material and methods

3.1 The dataset

The dataset was collected at the University of Rennes with 232 undergraduate English for Specific Purposes (ESP) students writing short essays about what they thought was the best invention in their scientific field. The CEFR levels were assessed by a set of four professional teachers with more than ten years of experience. The majority of essays were graded as B1-B2 (see for example the columns for the test set in Figure 3).

3.2 The processing pipeline and the typology of metrics

We implemented a pipeline to compute a set of metrics based on the final output (the typed text). Keylogs are collected online in .json format[1] via a JavaScript script embedded in an HTML text box. Each field captures a specific keyboard action (keypress or keyrelease), from which three metrics are extracted: the pressed key, action type, and timestamp. These metrics are converted to a tabular format where one row corresponds to one key press and release with corresponding timestamps. Using the extracted key names and timestamps, the final static text can be reconstructed and all keystroke metrics can be computed, such as the overall typing speed, the span of typing bursts, or the length and distribution of writing pauses. From the point of view of data format, our pipeline converts the keylogs from a .json format to a tabular format that is then stored in a .csv file. Our next stage uses a universal dependency annotation to try to capture the syntactic properties of the pauses. We followed the metrics suggested in (Pacquetet, 2024) to distinguish two types of bursts: Typing Bursts that correspond to sequences of active typing in between two inactive sequences (pauses) and Revision Bursts that correspond to sequences of active typing stopped by a revision (backspace).

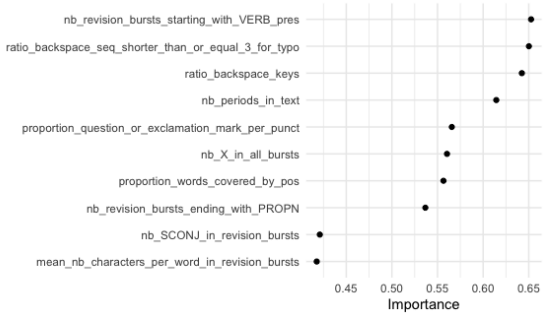
Since we annotated the data using Spacy for the universal dependency parsing, we operationalised metrics taking into account various types of constituent sizes and syntactic properties. We adopt a scope-based typology of metrics, which distinguishes between several types of domains, corresponding to linguistic constituents, from the whole text to individual character strings. Following a descending order in the presentation of our metrics, we first acknowledge metrics that are computed on the basis of the text (we call them 'text-based') and they are reported computed per text. For example, the metric 'total_nb_bursts_any_kind' computes the number of bursts (whether P-bursts or R-bursts) when writing a text. We then considered metrics that were computed at sentence level, such as "ratio_nb_rev_burst_per_sentence", which computes the number of revision bursts per sentence. We of course considered burst-based metrics such as "mean_time_revision_burst", the mean time for revision bursts. At word-level, we elaborated metrics like "mean_length_pauses_after_word", the mean duration of pauses after each word. Last, we identified character-based metrics, like the ratio of backspace keys (ratio_backspace_keys).

4 Preliminary Results

For exploratory purposes, we report our preliminary analysis on our dataset of 232 essays. We created a 80/20 training/test split ensuring proportional representation of the CEFR variable via stratification. We then used elastic net regularization (combining L1/LASSO and L2/ridge penalties) through GLMnet. The model optimizes the trade-off between bias and variance by minimizing a penalized cost function. Findings regarding pause patterns and CEFR levels revealed that long pre-burst pauses were linked with low proficiency levels, proposing a high cognitive load. Furthermore, logistic regression analysis of POS tagging indicates that the usage of collocations and formulaic expressions was a powerful proficiency predictor, aligning with (Pacquetet, 2024) findings, where advanced writers present better fluency in noun-verb sequence. Our Keystroke analysis provides preliminary results of systematic differences in typing behaviour across proficiency levels. Short pauses between sentences and phrases were exhibited by higher-proficiency participants, aligning with (Pacquetet, 2024) findings. In addition, regression models showed that higher-proficiency writers demonstrated more structured revision behaviour at

[1] https://github.com/taylor-arnold/keylog.js?tab=readme-ov-file

the sentence level, in comparison to lower-proficiency writers who were involved in within-word revisions and frequent backspacing. Since our preliminary analysis was performed on a limited sample, further work is suggested with a broader dataset and extending the sample size for additional validation. In our confusion matrix (Figure 3), the best predicted classes are unsurprisingly the most numerous ones.



	Truth				
Prediction	A1	A2	B1	B2	C1
A1	1	1	1	0	0
A2	2	6	4	1	0
B1	2	3	5	4	1
B2	0	0	7	4	2
C1	0	0	0	1	0

FIGURE 3 : Importance of features for the classification task (left) and confusion matrix (right)

5 Discussion and conclusion

The overall accuracy (0.356) of the classification of the learner levels is a bit disappointing, but the dataset is rather limited (232 essays) to build a completely reliable model. We plan to extend our methodology to bigger datasets like the KUPA-KEYS dataset and its 1,006 essays (Velentzas et al, 2024). The keylogging pipeline is integrated as a microservice for a moodle-based analysis of learner data as part of the A4LL project[2]. This infrastructure is meant to collect more data points and could be used to improve our modelling. The general aim of the project is to provide feedback to learners using a server on the HUMA-NUM infrastructure. For the time being, CEFR levels and feedback recommendations are based on the computation of complexity metrics but we aim to include the outputs of keylogging metrics to guide learners.

If we extrapolate the writing behaviour from the metrics, in this dataset, revision bursts seem to be of paramount importance for the CEFR classification. The importance of the metric ratio_backspace_seq_shorter_than_or_equal_3_for_typo suggests learners apparently edit more for typos than for revision. Some construction patterns could be detected by the fact that revision bursts tend to be observed for sequences beginning with verbs and ending with proper nouns. At keystroke level, backspace is the crucial 'behavioural metric' for learners, then comes the use of '?' and '!' in learner texts. In our data, word-based metrics and the size of p-bursts (captured by the number of keystrokes) did not seem to be relevant (even mean length of essays, when included in the model), but the size of r-bursts was, so we still believe that it is relevant to try to apprehend learner behaviour by varying the constituent scope of the metrics, which sounds like a promising avenue for research.

In this paper, we have illustrated the interest of investigating keylogging data with metrics. By varying the scope of the metrics with the size of the reference constituent, we manage a complementary approach between static computations and the dynamics of writing to characterise learner writing.

[2] The corresponding scripts are available on https://gitlab.huma-num.fr/lidile/a4ll_mlpipeline.

Acknowledgement

Part of the conceptualisation of the metrics presented in this research resulted from the King's College London / Université Paris Cité jointly funded DLLA project (Deep Learning for Language Assessment). The pipeline was implemented owing to the ANR-funded A4LL project (ANR-22-CE38-0015).

References

CONIJN, R., ROESER, J., and VAN ZAANEN, M. (2019). Understanding the keystroke log: the effect of writing task on keystroke features. Reading and Writing, 32(9), 2353–2374.

EUROPEAN COUNCIL (2001). *Common European Framework of Reference for Languages :Learning, teaching, assessment*. Cambridge : Cambridge University Press.

BALLIER N., MARTIN, P., & AMAND, M. (2016). Variabilité des syllabes réalisées par des apprenants de l`anglais (Analysing syllable variability in a French learner corpus of English) In, DANLOS L., & HAMON T., Éd., *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 1 : JEP*, p.732-740, Paris, France, AFCP - ATALA.

CONIJN R., VAN ZAANEN M., LEIJTEN M., & VAN WAES L. (2019). How to Typo? Building a Process-Based Model of Typographic Error Revisions *Journal of Writing Analytics*, 3, 69-95.

GAILLAT T., BALLIER N., SOUSA A., BOUYÉ M., SIMPKIN, A., STEARNS, B., & ZARROUK M. (2020). Un prototype en ligne pour la prédiction du niveau de compétence en anglais des productions écrites (A prototype for web-based prediction of English proficiency levels in writings) In BENZITOUN C., BRAUD C., HUBER L., LANGLOIS D., OUNI S., POGODALLA S., & SCHNEIDER S., Éd., *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 4 : Démonstrations et résumés d`articles internationaux*, p.30-33, Nancy, France, ATALA et AFCP.

GARCÉS MANZANERA, A. (2024). Language bursts and text quality in digital writing by young EFL learners *Journal of New Approaches in Educational Research*, 13.

GILQUIN G. (2020). In search of constructions in writing process data *Belgian Journal of Linguistics*, 34, 99-109.

GILQUIN, G. (2022). The Process Corpus of English in Education: Going beyond the written text *Research in Corpus Linguistics*, 10(1), 31-44.

GILQUIN, G. (2024). Keylogging and screencasting to help investigate L2 writing processes In, *Routledge Handbook of Technological Advances in Researching Language Learning*, p.285-296, London, Routledge.

GILQUIN, G., & LAPORTE S. (2021). The Use of Online Writing Tools by Learners of English: Evidence From a Process Corpus *International Journal of Lexicography*, 34(4), 472-492.

PACQUETET, E. (2024). *The Effect of Linguistic Properties on Typing Behaviors and Production Processes* PhD thesis, University of Buffalo, ProQuest LLC.

ROESER J., DE MAEYER S., LEIJTEN M., & VAN WAES L. (2024). Modelling typing disfluencies as finite mixture process *Reading and Writing*, 37(2), 359-384.

TAPPERT, C., VILLANI, M., & CHA, S. H. (2010). Keystroke biometric identification and authentication on long-text input. In *Behavioral biometrics for human identification: Intelligent applications,* 342-367. IGI Global Scientific Publishing.

VELENTZAS, G., CAINES, A., BORGO, R., PACQUETET, E., HAMILTON C., ARNOLD, T., ... & YANNAKOUDAKIS, H. (2024, May). Logging Keystrokes in Writing by English Learners. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024),* 10725-10746.