Investigating the role of revisions for learner level predictions

Ahood Al Sawar
ALTAE, Université Paris Cité
Place Paul Ricoeur
75013 PARIS, France
ahoodswar2018@gmail.com
0009-0007-3486-8421

Léo Nebel
Sorbonne Université
CNRS, LIP6, F-75005
EvidenceB
Paris, France
leo.nebel@lip6.fr
0000-0001-5859-265X

Nicolas Ballier
ALTAE, Université Paris Cité
Place Paul Ricoeur
75013 PARIS, France
nicolas.ballier@u-pariscite.fr
0000-0003-2179-1043

Abstract

This paper investigates the prediction of CEFR levels using revision features extracted from automatically identified revision episodes. We try to predict the different levels of learners from two keylogging datasets using the K-REV Python library developed as open source. The two different corpora are KUPA-KEYS collected by (Velentzas et al., 2024) and CELVA.Sp (Mallart et al., 2023). We extracted revision features representing different aspects of the revision behaviour. We applied an ordinal logistic regression, but the model's ability to predict CEFR levels was limited. Using KUPA-KEYS, it achieved 59% accuracy, while in CELVA.Sp, the performance was only 35%. Despite this, most prediction errors stayed within adjacent CEFR levels in both datasets. Cross-data comparison also showed several shared patterns in revision behaviour; features such as more revision episodes, more characters inserted per revision, more backspaces per sequence, and later revision timing were associated with higher CEFR levels. However, a higher proportion of writing time spent in revision was associated with lower CEFR levels in both datasets. These findings suggest that these revision-process features capture some information related to writing proficiency, yet are insufficient on their own for precise CEFR classification. In addition, this study highlights the need for interoperable input formats across keylogging datasets to overcome the practical challenges we encountered when applying automated revision analysis.

1 Introduction

Second-language (L2) writing assessment typically relies on the final submitted text. Although this approach reveals important insights into overall writing quality, it overlooks the writing process, during which many cognitive activities occur before the final version is completed. Advances in keystroke logging enable the real-time recording of writers' behaviours, such as pauses, revisions, deletions, and insertions. This writing process data enables the investigation and analysis of writing behaviours that the product alone cannot reveal (Leijten and Van Waes, 2013; Gilquin, 2022).

Revision behaviour, in particular, has received special attention because it reflects how writers monitor and self-correct their writing. Previous research has shown that revision patterns differ across proficiency levels (Xu, 2018; Al Sawar et al., 2025). However, analysing revisions on large datasets has been challenging. Revision has frequently been operationalised as the number of backspaces, but this ignores the difference between typos and meaningful edits. Conijn et al., 2022b tackled this by creating a classification system that describes revisions along nine properties, such as location, timing, and action. They found that the large majority of revisions were surface-level rather than changes to meaning or structure, highlighting the importance of distinguishing deep revisions from simple typo corrections. However, applying these annotations manually limited research to small samples. K-REV (Nebel, 2026), an open-source Python library, was developed to automate this process by extracting and characterising revisions from keystroke logs. This leads to the central question addressed in this paper: whether revision behaviour, automatically extracted from keystroke logs, can predict CEFR levels of English learners while account- ing for typing skill.

We address this by applying K-REV to two keystroke-logging datasets, KUPA-KEYS and CELVA.sp. Both datasets include copy and essay-writing tasks with keystroke logs, but they differ in participant populations and data collection settings. This enables us to examine whether the most predictive revision metrics generalise across different writing environments. Revision features are extracted from the revision episodes identified by the Python library, and then we feed them into an

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 65–77, November 19–20, 2026, University of Tartu Library

ordinal logistic regression model to predict CEFR levels.

This study therefore aims to provide a reproducible, K-REV-based operationalisation of revision behaviour and examine whether features of the revision process are associated with CEFR writing proficiency across two datasets.

Revision behaviour represents only one dimension of the writing process, making CEFR prediction based exclusively on revision features a challenging task. Previous work (Al Sawar et al., 2026) on the same learner keystroke datasets, KUPA-KEYS and CELVA.Sp, investigated CEFR prediction using a broader combination of writing-process and linguistic features extracted from keystroke logs and final texts. In contrast, the present study uses the K-REV Python library to identify revision episodes and extract revision-specific features. This narrower focus allows us to investigate the predictive value of revision behaviour independently of other writing-process and linguistic indicators.

The rest of the paper is structured as follows. Section 2 presents previous research on CEFR level predictions within the Complexity Accuracy Fluency paradigm. Section 3 details our methodology and the datasets we used in our experiments. We analyse our results in Section 4 and Section 5 discusses them.

2 Previous Research

Keylogging, or keystroke dynamics, has been first extensively used for user authentication (see Roy et al., 2022 for an overview). We focus in this section on educational contexts, where studies have also been conducted on emotional states: for instance, Rahman et al., 2025 developed a framework for cognitive-load detection distinguishing tasks such as typing password and transcription.

2.1 Learner corpus research based on keylogging

Keystroke logging records every keypress, pause, and deletion that writers produce, with their timing, which changed the focus from analysing only the final product to the writing process. This methodological approach was proposed by Leijten and Van Waes, 2013, where the Inputlog tool provided a standard way of logging, analysing, and visualising the writing process, and it is now widely used to investigate how learners write their texts in real time. This paved the way for building learner corpora, while collecting both keystroke logs and final texts.

A representative example of this is The Process Corpus of English in Education (PROCEED)(Gilquin, 2022). This corpus includes argumentative essays with their keystroke logs and screen recordings, collected from university students writing in both their first language (French) and their L2 (English). This type of corpus allows researchers to examine how texts are composed, and not just what is submitted. Bursts of typing, for instance, might indicate that a learner is typing a stored linguistic unit, and some constructions typed during drafting were later revised or abandoned before the final text (Gilquin, 2020, 2023, 2024). Although recent efforts have led to the development of learner keystroke corpora annotated for language proficiency, datasets that combine keystroke data with CEFR levels assigned by human raters remain relatively scarce. The KUPA-KEYS dataset (Velentzas et al., 2024) and the CELVA.Sp corpus (Mallart et al., 2023) represent important steps toward addressing this gap by supporting investigations into keystroke patterns associated with language proficiency. These resources provide the basis for the present study. Nevertheless, additional corpora containing both keystroke data and expert-assigned CEFR levels are needed to enable cross-corpus comparisons and assess the generalizability of findings.

2.2 Metrics-based analysis

Most of the process metrics used for analysis fall into two broad families, pause-based and revision-based measures. Pause-based metrics view typing interruptions as an indicator of cognitive effort. Common measures include the duration, frequency, and pause locations, whether before, between, or within words. Also, burst length (the text produced between two pauses), time on task, and start time (Van Waes and Leijten, 2015; Sinharay et al., 2019). These metrics vary with proficiency levels. Barkaoui, 2019 found that lower-proficiency L2 writers pause more frequently in comparison to proficient writers, who pause longer at the beginning of their writing, mainly for planning.

Revision-based metrics, which are the main focus of this study, quantify how writers alter text they have already produced. Previous research by Sommers, 1980 indicated that revision behaviour varies between experienced and novice writers. Ex-

perienced writers tend to revise by reorganising ideas and text structure, whereas novices often concentrate on small wording adjustments.

Building on these observations, Faigley and Witte, 1981 suggested a taxonomy that differentiates between surface changes, which preserve meaning, and meaning changes, which alter semantics or text structure. In keystroke-logging research, revision has often been reduced to simple counts, such as the number of deleted characters or backspaces (Sinharay et al., 2019). However, more recent work has used richer characterisations; for instance, Conijn et al., 2022b aimed to develop a comprehensive product- and process-oriented tagset for revisions so it could be used to analyse both product and writing-process data. These tagsets focus on external properties of revisions, including: trigger, orientation, evaluation, action, linguistic domain, spatial location, temporal location, duration, and sequencing. They used keystroke logging, eye tracking, screen replays, and NLP tools to show that features related to these properties can be extracted automatically and manually. Pacquetet, 2024 showed that the revision behaviour reflects higher-level writing processes rather than just simple error correction, where she distinguished between revisions and repairs, with repairs described as typing errors or spelling errors, while revision changes are those made to the text that change meaning or could be syntactic, semantic or form changes. Al Sawar et al., 2025 also found that revisions varied between lower- and higher-level writers, with the former making more within-word revisions and the latter making more structured, sentence-level revisions. This suggests that revision behaviour provides proficiency indicators that could serve as predictors.

Although revision behaviour varies with proficiency and keystroke features have been employed to predict proficiency levels (Al Sawar et al., 2025), to the best of our knowledge, no previous study has used automatically extracted revision metrics to predict CEFR levels. This study aims to address this gap.

Previous research motivated the revision-process feature selection in this study by showing that revision is a multidimensional process and that revision behaviour can vary across L2 proficiency. Barkaoui, 2016, for instance, showed that what and when learners revise vary across proficiency levels. Keystroke-logging research on the revision process has demonstrated that revision can be described from different dimensions, including when and for how long a revision is, and how a writer behaves while revising, in other words, their backspacing behaviour, and the amount of text inserted or deleted (Conijn et al., 2024, 2022b; Pacquetet, 2024). In particular, Pacquetet, 2024 showed that the number of backspaces alone is not necessarily equivalent to the amount of the actual textual change. Therefore, backspace sequences can indicate how a revision was executed physically; on the other hand, the amount of inserted or deleted text reflects the actual modification. She also argues that the time a writer takes before revising or editing can provide a temporal measure of editing intention and self-monitoring. Additionally, revision bursts have been associated with text quality in L2 writing (Garcés-Manzanera, 2024). Hence, the features selected in this study attempt to capture revision activity, temporal organisation, backspacing behaviour, and the extent of textual modification. We therefore test these features extracted from automatic revision episodes to see whether they can distinguish between CEFR levels.

2.3 CAF-based predictions of learner levels

As regards the prediction of learners, especially in the learner corpus research community, CEFR level prediction has essentially been carried out on full texts collected in corpora, mostly within the Complexity - Accuracy Fluency paradigm (Housen et al., 2012). In this paradigm, complexity has been more thoroughly investigated, lexical complexity, lexical sophistication and grammatical complexity have been used by automatically extracting features , using for instance tools like the Syntactic Complexity Analyzer (Lu, 2010), for example Lu (2017); Arnold et al. (2018), or suites of NLP tools such as TAASC (Kyle, 2016) or TAALES (Kyle et al., 2018) , like Gaillat et al. (2022) or other complexity metrics (Vajjala and Rama, 2018).

3 Material and Methods

3.1 Datasets

3.1.1 Dataset 1

The primary dataset in this study is the KUPA-KEYS dataset, collected by Velentzas et al., 2024. It contains keystroke-logging data, final texts, and metadata from over 1,000 participants, including both native and non-native speakers. Additionally, texts were rated for CEFR proficiency by three human evaluators and an automated marker, with B2

being the most common level, as shown in Figure 1. They reported that inter-rater agreement was assessed using three different measures: Gwet's AC, Spearman's rank correlation, and Root-Mean-Square Deviation (RMSD). The annotators showed good overall agreement (Gwet's AC = 0.854; significant Spearman correlations of 0.514/0.711). RMSD values indicated that the first annotator aligned most closely with the human average (0.622), while the second and third raters diverged most (2.586), reflecting different grading tendencies. The tasks in this dataset included a copy task of a 300-word excerpt from Steve Jobs' speech and writing an essay in response to one of many prompts. The essay task replicated prompts from the Write & Improve system, developed and maintained by Cambridge University Press & Assessment.

We calculated the average essay scores from the 4 raters, rounded this mean to the nearest integer, and then assigned the final CEFR level label. Thirteen writings were excluded because their keystroke data could not be reliably reconstructed and validated, and A2 was excluded because there were only four observations at this level. Of the remaining writings, 989 only had the necessary revision process features.

Although, C2 level has only 19 writings, stratified 10-fold cross-validation allowed representing this level in every test fold. Excluding C2 would have restricted the analysis on three levels (B1-C1) and prevented us from evaluating the model at the highest proficiency level.

Because the last batch of the dataset was collected around the time ChatGPT emerged, three annotators carried out a manual review using two visualization tools to identify potentially suspicious writing patterns. Inter-rater agreement was high in terms of raw agreement (97.32%). However, Fleiss' was low, = .091, p < .001, likely reflecting the strong imbalance in category prevalence. Because suspicious cases were rare, Gwet's AC1 was also examined and indicated very high overall agreement, AC1 = .982, 95% CI [.975, .989]. Only three essays were flagged by two annotators and none were flagged by all three, indicating limited agreement specifically on suspicious cases. This is why the three suspicious essays were not excluded from the final analysis.

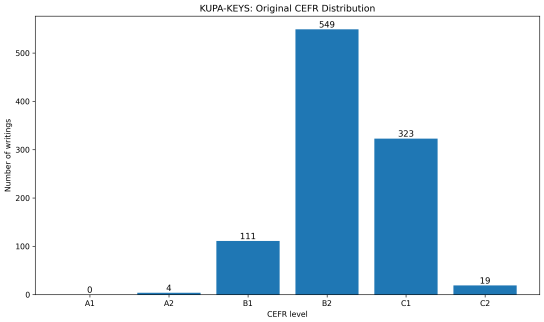


Figure 1: KUPA-KEYS CEFR class distribution

3.1.2 Dataset 2

The second dataset is CELVA.Sp collected by Mallart et al., 2023, and it also consists of keystroke logging data with final texts. It was collected as part of the Analytics for Language Learning (A4ALL) project.

The participants are French undergraduate learners of English as a second language, and they were asked to perform a copy task and an essay-writing task. CEFR levels were assigned by expert raters using the DIALANG test and the most represented levels are A2 and B1 (see Figure 2). CELVA.Sp included 238 writings, but 24 were excluded due to incompatible timestamp data for extracting time-dependent revision features. In addition, six writings were removed because they had not been assessed, and the only C2 writing was excluded as well. One more writing was also omitted because its reconstructed final text contained no content. This limited the modelling analysis to levels A1-C1, resulting in a total of 206 writings used.

The two datasets differ in their collection setting. While KUPA-KEYS was collected via the Prolific crowdsourcing platform, CELVA.Sp was collected in a more ecological setting, where the data were recorded during a normal classroom activity. CELVA.Sp participants wrote short argumentative essays and were a more homogeneous population than the KUPA-KEYS participants.

3.2 K-REV library

The K-REV library is the result of research on the characterisation of revision out of keystroke data (Conijn et al., 2024; Nebel et al., 2025b,a). It relies on a rule-based approach to extract the beginning and end of the revision events at a fine-grained level. A revision begins each time a deletion starts (first backspace), or if the cursor moves within the already written text (Conijn et al., 2024). When it

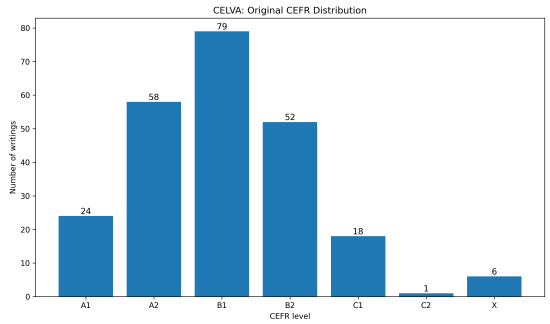


Figure 2: CELVA.Sp CEFR levels distribution

comes to the endpoints of revisions, the approach follows simple rules mimicking annotator reasoning (Nebel et al., 2025b). If a revision begins with a deletion, the algorithm looks for the next keystroke that is not a backspace; from there, the revision ends when the next stroke is a space, and the writer has typed as much as they deleted ($\pm$ 5 characters, which is a subjective threshold, shown effective on the dataset from Nebel et al. (2025b)'s study). If a revision begins with a cursor move, the revision ends when the cursor moves again. Such an approach was compared to traditional machine learning models on a binary classification task consisting of predicting, for each keystroke, if it was the end of the considered revision or not.

Once the revision is extracted, the library also allows the characterisation of the types of revision, based on existing taxonomies, mainly the distinction of Faigley and Witte (1981) between surface-oriented and meaning-oriented revisions. To do so, the library relies on LLMs, through a prompting approach, letting the user decide what model to use (Nebel et al., 2025a).

3.3 Data pre-processing

Applying K-REV requires a text list, a cursor list, and a timestamp list describing how each essay evolved during writing. In Nebel et al. (2025b), these lists were collected directly as the keylogger was designed for that. However, the KUPA-KEYS dataset logs are structured differently from the data K-REV was developed on. Instead of recording the state of the text at each moment, they record the events that occurred, such as key presses and releases, mouse clicks, insertions, and captures. Therefore, the writing process was reconstructed by processing these events in chronological order. During reconstruction, the current text and cursor position were continuously updated, while capture events were used to synchronise the reconstructed text and reduce the accumulation of reconstruction errors. After each key-down event, the current text, cursor position, and timestamp were stored, producing three aligned sequences (text list, cursor list, and time list), which were used as the input to K-REV see Figure 3. Regarding the CELVA.Sp dataset, since its raw data was in a different event format, we mapped the events to the structure used in the KUPA-KEYS dataset. Then, we applied the same reconstruction process to generate the text, cursor, and timestamp lists.

index	text_list	time_list	cursor_list
0		155998.09999999404	0
1	I	156113.39999999106	1
2	I	156588.5	2
3	I t	156829.39999999106	3
4	I th	156953.39999999106	4
5	I thi	157023.0	5
6	I thin	157228.0	6
7	I think	157309.39999999106	7
8	I think	157379.59999999404	8
9	I think t	157978.29999999702	9
10	I think th	158103.29999999702	10
11	I think tha	158335.70000000298	11
12	I think that	158694.89999999106	12
13	I think that	158888.5	13
14	I think that s	159105.0	14
15	I think that so	159223.89999999106	15
16	I think that som	159404.09999999404	16
17	I think that some	159475.29999999702	17
18	I think that somet	159709.09999999404	18
19	I think that someth	159784.20000000298	19
20	I think that somethi	159875.09999999404	20
21	I think that somethin	160055.20000000298	21
22	I think that something	160154.0	22
23	I think that something	160341.5	23
24	I think that something t	160545.59999999404	24
25	I think that something th	160734.79999999702	25
26	I think that something tha	160849.09999999404	26
27	I think that something that	161110.59999999404	27

Figure 3: Example of the text, time, and cursor lists required by K-REV.

3.4 Revision extraction with K-REV

The reconstructed text list, cursor list, and time list are then passed to the K-REV library, which identifies individual revision episodes within a writing session and returns, for each episode, its positions within the writing session together with a distinction between linear and embedded revisions.

3.5 Feature Extraction

Once this revision extraction is done, we then summarised features extracted from these episodes at the writing level. The final feature set was inspired by the KREV framework and by the revision process measures described by Conijn et al., 2022b and Pacquetet, 2024. The aim was to obtain a set of features that captures different aspects of revi-

sion behaviour, such as how much writers revise, when revision occurs, and the amount of deletion and insertion, pauses and timing around revisions, and the structure of backspace sequences (see Table 1). Regarding the backspace features, Conijn et al., 2022b provides a broader revision-event-level perspective (how many backspaces within a revision event), whereas Pacquetet, 2024 provides a more fine-grained sequence-level perspective on Backspace behaviour (How many consecutive backspaces). Extending Pacquetet (2024)'s use of backspace sequences, we also examine whether a sequence contains a pause of more than one second. The idea is that more proficient writers may maintain greater fluency even during revision, whereas less proficient writers may show greater temporal dispersion during editing. Thus, we wanted to test whether this temporal feature of backspace sequences could help distinguish proficiency levels. However, the feature itself is considered descriptive rather than proficiency-specific and it can be used in other analysis of revision behaviour such as examining whether temporal distribution during revision varies across writing tasks or task conditions. Also, because a fixed one-second threshold could be influenced by individual typing speed, we will conduct a sensitivity analysis controlling for baseline typing speed from the copy task. Limiting the number of predictors was particularly relevant given the small CELVA.sp sample; hence, to prevent possible overfitting, this also allows us to retain the same feature set across both datasets, enabling direct cross-dataset comparison. Consequently, each participant is represented by a single feature vector that summarises their revision behaviour, along with their CEFR level. All analyses were conducted in Python 3.12.0. The analysis code, software requirements, and instructions for reproducing the analyses are available in a public GitHub repository.[1]

3.6 Modelling

The target variable is CEFR levels, and the predictors are the revision-process features extracted from the revision episodes. Because the CEFR levels are ordinal, we applied an ordinal logistic regression model using LogisticIT from the *mord* package (Pedregosa-Izquierdo, 2015). As for missing feature values, they were imputed using the median and the features were standardised; these preprocessing steps were estimated from the training data within each 10-fold cross-validation to prevent data leakage. We evaluated the performance of the model using accuracy, balanced accuracy, MAE, and adjacent accuracy. Feature importance was analysed through model coefficients.

[1] https://github.com/AhoodSwar/krev-revision-cefr-levels-prediction

4 Results

The results are presented in three sections. First, we report model performance on KUPA-KEYS using the selected feature set, each feature's contribution to predicting CEFR levels, and its association with higher and lower proficiency levels. Then we report the corresponding results for the CELVA.Sp dataset. Finally, we compare the two datasets to identify similarities and differences in model performance and the associations between revision features and CEFR levels.

4.1 KUPA-KEYS Results

Table 2 summarises the ordinal logistic regression model performance compared to the majority-class baseline. The model achieved an overall accuracy of 0.599 and a balanced accuracy of 0.329, which is considered limited in terms of discriminating CEFR levels. The MAE is 0.417, while the adjacent accuracy reached 0.985, indicating that the prediction errors were wrong by no more than one level. These results suggest that our revision-based model outperformed the majority-class baseline model across all four evaluation metrics. These results suggest that revision-based features provide predictive information beyond the majority-class distribution, although low balanced accuracy indicates uneven performance across proficiency levels. The model explained approximately 20.52%, of the variance in KUPA-KEYS, based on the Mckelvey-Zavoina pseudo-R^2, indicating a modest performance in explaining a large proportion of the variance underlying CEFR levels.

Figure 4 shows the confusion matrix of the ordinal logistic regression based on the test sets from stratified 10-fold cross-validation. The model accurately categorized 592 out of 989 texts, with the majority falling into the most dominant B2 level. Most misclassifications were between neighbouring CEFR levels (e.g., B1 predicted as B2); however, large wrong predictions that misclassified levels by two or more proficiency levels were rare. This is reflected in high adjacent accuracy (0.985), which indicates that the model was mostly close to

Table 1: Revision-process features used for CEFR modelling.

Feature group	Feature	Description	Inspired by
Revision frequency and duration	Revision episodes per 100 words	Number of K-REV-defined revision episodes per 100 words of the final essay text.	K-REV
	Revision time / writing time	Proportion of the writing process spent in K-REV-defined revision episodes.	K-REV
Revision timing	Revision regularity: SD of intervals (s)	Variability in the temporal intervals between K-REV-defined revision episodes.	K-REV
	Median revision time from process start (s)	Median temporal location of revision episodes from the start of the writing process.	Conijn et al. (2022)
	Median pause before revision (s)	Median pause immediately preceding a revision episode.	Conijn et al. (2022)
Backspacing behaviour	Proportion of Backspace sequences > 3	Proportion of Backspace sequences containing more than three consecutive Backspaces.	Pacquetet (2024)
	Median Backspaces per revision	Median number of Backspaces per revision episode.	Conijn et al. (2022)
	Number of Backspace sequences	Number of physical consecutive Backspace sequences within K-REV-defined revision episodes.	Pacquetet (2024)
	Mean Backspaces per sequence	Mean number of Backspaces within a physical consecutive Backspace sequence.	Pacquetet (2024)
Backspace timing and edit latency	Backspace sequences $>$ 1,000-ms intervals	Number of consecutive Backspace sequences within K-REV-defined revision episodes containing at least one Backspace-associated press-to-press interval greater than 1,000 ms.	Pacquetet (2024)
	Time to edit within revisions (mean, ms)	Mean time between the last soon-to-be-erased character and the first Backspace within revision episodes.	Pacquetet (2024)
Extent of textual modification	Median characters deleted per revision	Median number of characters deleted per revision episode.	Conijn et al. (2022)
	Median characters inserted per revision	Median number of characters inserted per revision episode.	Conijn et al. (2022)

Table 2: Ordinal regression model performance for KUPA-KEYS (N = 989).

Metric	Majority Baseline	Revision-Based Model
Accuracy	0.549	**0.599**
Balanced Accuracy	0.250	**0.329**
Mean Absolute Error (MAE)	0.470	**0.417**
Adjacent Accuracy	0.981	**0.985**

the true proficiency level.

Figure 5 highlights the most important features, while Figure 6 presents the signed coefficients of the features. Positive coefficients suggest associations with higher CEFR levels; on the other hand, negative coefficients are associated with lower proficiency levels. The number of backspacing sequences showed the strongest positive association with higher levels. Followed by the time to edit

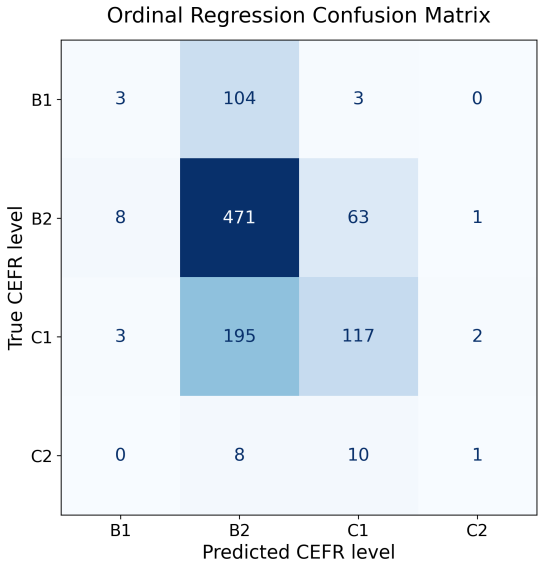


Figure 4: Confusion matrix of the Ordinal Logistic Regression model on the KUPA-KEYS dataset.

within revision, the average number of characters inserted and deleted per revision, mean backspaces per sequence, the proportion of backspace sequences longer than three keystrokes, and revision timing from the start of the writing process. The number of revision episodes per 100 words showed the strongest negative association with lower CEFR levels. This was followed by backspace sequences with intervals greater than 1,000 ms, backspace average per revision, pauses average before revision, and the proportion of writing time spent on revision. The second feature associated with lower levels in this dataset supports our hypothesis that less proficient writers tend to pause more during revision, unlike more skilled writers. However, further investigation is needed to consider typing speed.

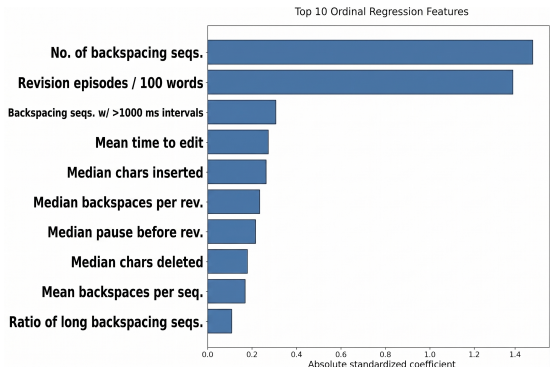


Figure 5: Top ten most influential K-REV features selected by the Ordinal Logistic Regression model for predicting CEFR proficiency levels in the KUPA-KEYS dataset.

Figure 7 illustrates the distributions of six revision features across CEFR levels. Despite significant overlap across several features, the number of backspacing sequences in general increased from B1 to C1; however, revision episodes per 100 words decreased at higher levels. Several features also contained outliers, suggesting substantial variation in revision behaviour within the same proficiency level. Overall, these distributions suggest that revision behaviour differs across proficiency levels. However, these patterns should be interpreted with caution, given the small number of C2 writings.

4.2 CELVA.Sp Results

4.2.1 Pre-processing of the CELVA.Sp dataset

Using the same modelling pipeline as for the KUPA-KEYS dataset, shown in Table 3, the ordinal logistic regression model achieved an accuracy of 0.350, suggesting a limited ability to discriminate exact CEFR levels. Balanced accuracy was 0.242, and the model also reached a MAE of 0.816, indicating, as with the KUPA-KEYS dataset, that the average prediction error was less than one CEFR level. Furthermore, the model attained an adjacent accuracy of 0.850, implying that predictions were either correct or differed by only one neighbouring proficiency level. Although precise classification is difficult due to the limited sample size, the model successfully captures the ordinal progression of writing proficiency (see Figure 8). Compared with the majority-class baseline model, our revision-based model outperformed it across all four evaluation metrics. Also, the model explained approximately 10.59% of the variance according to Mckelvey-Zavoina pseudo-R^2. This suggests that the model captured some proficiency-related information, yet a large proportion of the variation remained unexplained.

Table 3: Ordinal regression model performance for CELVA

Metric	Majority Baseline	Revision-Based Model
Accuracy	0.345	**0.350**
Balanced Accuracy	0.200	**0.242**
Mean Absolute Error (MAE)	0.845	**0.816**
Adjacent Accuracy	0.811	**0.850**

The key revision features identified by the ordinal logistic regression model are displayed in Figure 9. Figure 10 displays that the most influential predictors include revision episodes per 100 words, the number of backspacing sequences, the proportion of long backspacing sequences, the median number of characters inserted per revision, the average number of backspaces per sequence, and backspace sequences that contain intervals longer than one second. Higher CEFR levels were associated with the number of backspace sequences, more characters inserted per revision, the mean of backspaces per sequence. However, lower CEFR levels were associated with more revision episodes per 100 words, proportion of backspace sequences longer than three, backspaces sequences including more than one second interval, and a greater proportion of writing time spent in revision. Given the model's limited performance, interpret these coefficients cautiously.

The distribution of the top six features across proficiency levels in the CELVA.Sp dataset is demonstrated in Figure 11. The boxplots showed that

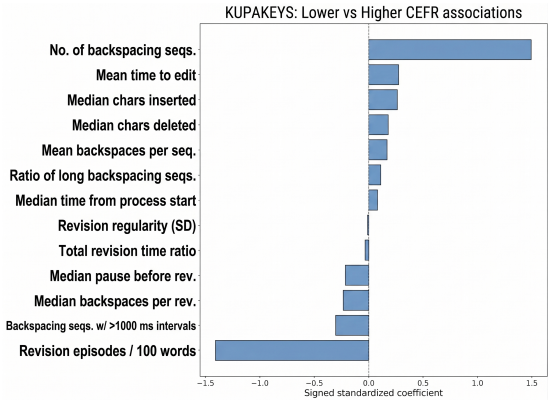


Figure 6: Signed coefficients from the final ordinal logistic regression model. Positive coefficients indicate associations with higher CEFR levels, whereas negative coefficients indicate associations with lower CEFR levels.

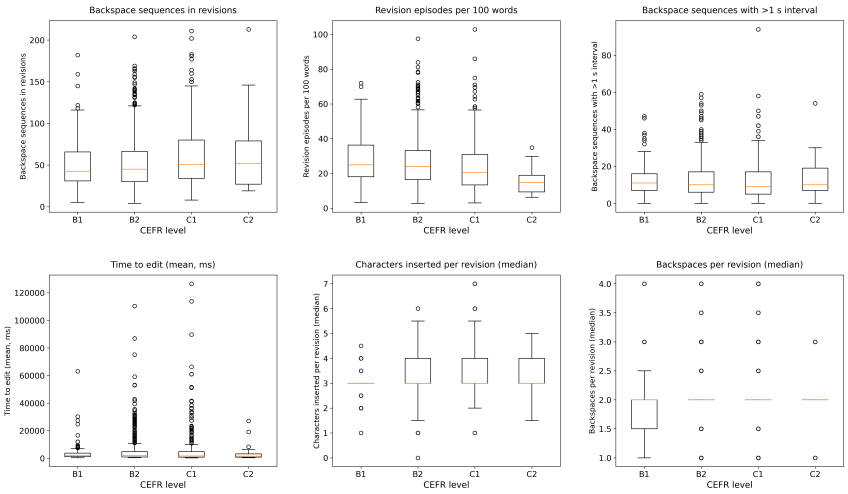


Figure 7: Distributions of the six revision-process features with the largest absolute standardised coefficients across CEFR levels in the KUPA-KEYS dataset.

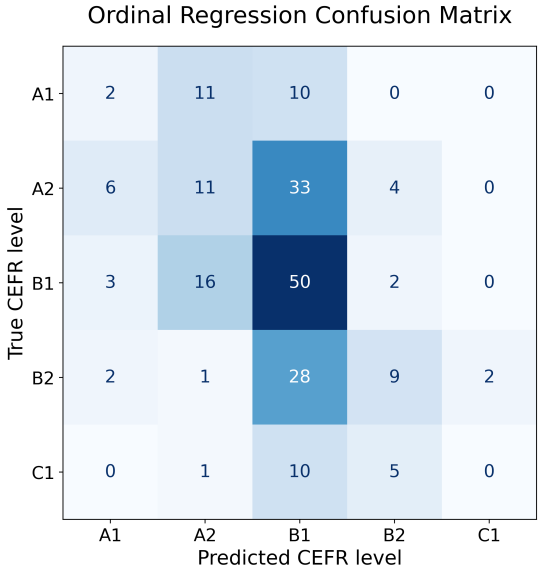


Figure 8: Confusion matrix of the Ordinal Logistic Regression model on the CELVA.Sp dataset.

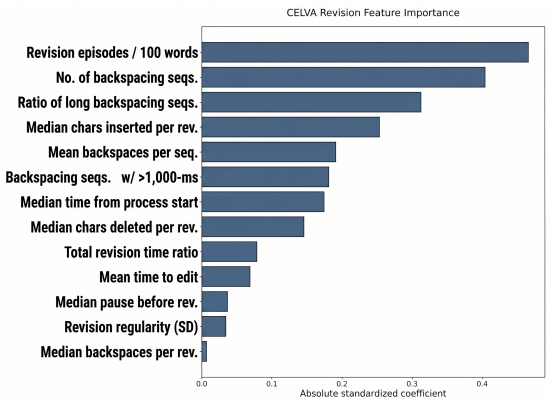


Figure 9: Top ten most influential K-REV features selected by the Ordinal Logistic Regression model for predicting CEFR proficiency levels in the CELVA.Sp dataset.

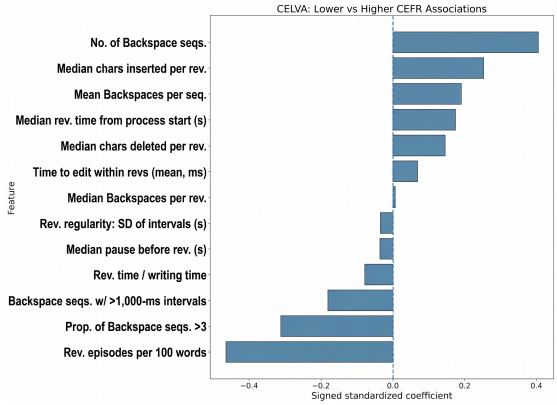


Figure 10: Signed regression coefficients of the top K-REV features in the Ordinal Logistic Regression model. Positive coefficients indicate an increase in the likelihood of higher CEFR levels, whereas negative coefficients indicate an association with lower proficiency levels.

revision episodes per 100 words decreased as the proficiency levels increased, while the number of backspacing sequences tended to increase. The other features showed smaller differences and substantial overlap across CEFR levels. Regarding the number of backspace sequences with intervals exceeding 1,000 ms, there was some overlap, but it was higher at higher CEFR levels which is inconsistent with our hypothesis. These results suggest that some revision behaviours vary with proficiency levels. However, this should also be interpreted with caution, due to the small size of this dataset and its imbalance.

4.3 Cross-corpus comparison

The ordinal logistic regression model performs better on the KUPA-KEYS dataset than on the CELVA.Sp dataset, with 59% accuracy versus 35%. This difference may be attributed to sample size variations. Hence, the better performance in KUPA-KEYS should be interpreted with caution because the two datasets differ in their sample size and CEFR distribution. The balanced accuracies in both datasets indicate that both models struggle with balanced predictions across all levels due to dataset imbalance. The adjacent accuracy using KUPA-KEYS is notably high at 98%, compared to 85% for CELVA.Sp. However, this implies that in both datasets, predictions were either correct or mainly off by only one level. Across both datasets, more backspacing sequences, more characters inserted and deleted per revision, more backspaces per sequence, later revision timing, and longer editing time were associated with higher CEFR levels. Also, more revision episodes per 100 words, a higher proportion of writing time spent revising, longer pauses before revision, and more backspacing sequences with intervals longer than one second were associated with lower CEFR levels. However, some features showed different patterns, such as the proportion of long backspacing sequences was associated with higher levels in KUPA-KEYS but it was associated with lower levels in CELVA.Sp. Median backspaces per revision showed the opposite pattern. Overall, several revision features showed similar patterns across the two datasets and captured some differences across CEFR levels, but they were limited in predicting the exact CEFR level.

4.4 Documentation and Interoperability

Applying K-REV to different keylogging datasets also highlighted the need for clear documentation and interoperable input formats. KUPA-KEYS and CELVA.sp were stored in different structures, requiring dataset-specific preprocessing before revision episodes could be extracted, as explained in the text reconstruction section. This could increase the effort required to apply the library to new keylogging datasets and make replication more difficult.

5 Discussion

Our findings suggest that higher-proficiency writers tend to revise less frequently relative to text length, but their edits are more extensive, with greater text insertion and deletion and more backspacing sequences. This may suggest that more proficient writers have stronger language skills and a wider vocabulary, allowing them to make larger changes while revising less often. In contrast, lower-proficiency writers tend to revise more frequently and spend more time revising. They also tend to pause and hesitate while making revisions. This may suggest that they could be having difficulty finding suitable replacements, leading to more frequent revisions and longer pauses during editing. Also, we found that keystroke-derived revision metrics are moderate indicators of CEFR levels, despite the limited performance of the models (Conijn et al., 2022a). However, our findings align with previous research showing that writing process features are associated with proficiency and writing quality. Barkaoui, 2016, for example,

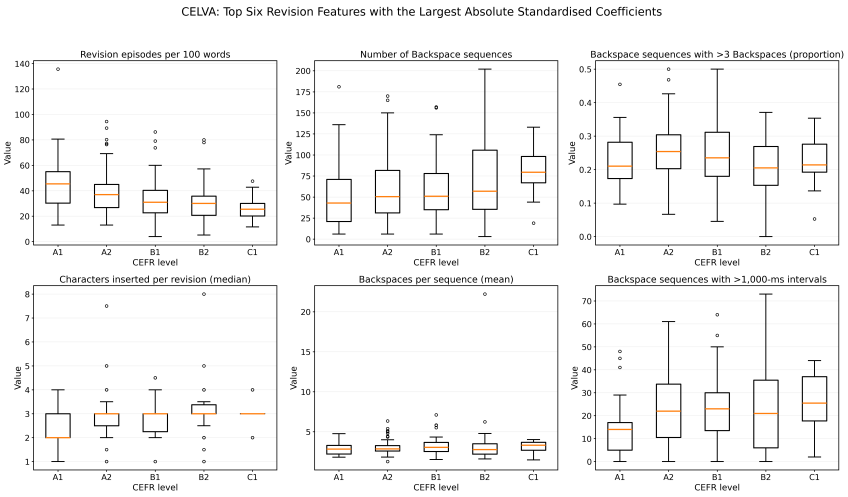


Figure 11: Distributions of the six revision-process features with the largest absolute standardised coefficients across CEFR levels in the CELVA.Sp dataset.

found that proficiency was related to learners' revision behaviour. More generally, previous studies have shown that process features extracted using keystroke data can provide information related to writing proficiency (Sinharay et al., 2019; Al Sawar et al., 2025). This study extends this by using automatically extracted revision episodes to predict CEFR levels across two datasets. However, the limited performance in predicting the exact CEFR levels and the differences between the two datasets we used suggest that revision behaviour should not be treated as a direct indicator of CEFR levels.

The cross-corpus evaluation provides valuable insights into the generalisability of our approach. By testing on two distinct datasets (KUPA-KEYS, a diverse crowdsourced corpus, and CELVA.sp, collected in an ecological and homogeneous setting), we identified several consistent trends. However, the difference in the resulting exact accuracies underlines a limitation in model transferability. This discrepancy, likely due to the variations in sample size and class distribution, suggests that, while the correlations between revision behaviours and proficiency seem to generalise well, the predictive thresholds remain highly context-dependent and require further validation on larger learner corpora.

From an educational perspective, these prediction models could significantly complement traditional product-oriented evaluations by introducing process-oriented insights. However, deploying such predictions for high-stakes summative assessment is premature. Because our models' performance is still moderate, and the causal mechanisms linking specific predictive features to distinct CEFR levels require more rigorous exploration, using them for grading could result in inaccurate evaluations. Instead, their true utility lies in formative assessment, where they can render some of the often-invisible cognitive mechanics of writing visible. By providing granular feedback on the writing process, these models can help teachers identify learners who struggle with fluency or over-rely on surface-level, local revisions rather than global planning. For students, visualising their own writing rhythms can foster metacognitive awareness and encourage better self-regulation (Vandermeulen et al., 2020). Furthermore, it would be highly valuable for future work to explore the early prediction of CEFR levels using real-time keystroke logs. Post-hoc predictions delivered after a task is completed are fundamentally less actionable for learners. Developing early detection models would allow Intelligent Tutoring Systems to provide timely feedback (Shute, 2008) while students are still actively engaged in the writing process.

6 Conclusion

In this paper, we have attempted to predict CEFR level of learners on the basis of revision types as systematically assigned by the K-REV library. Although our results are promising, another obvious candidate for CEFR level prediction on the basis of keylogging data is large language models; exploratory analysis is currently carried out (Gan and Adams, 2026).

Limitations

An obvious limitation of this study is the number of corpora we used to evaluate the benefits of the K-REV library. While we are currently testing a sub-component of the PROCEED corpus, we can only re-assert the (current) scarcity of CEFR-graded keylogging data for learners. We also acknowledge that C2 has a small sample size (n = 19) in the KUPA-KEYS dataset, which limits the reliability of predictions for this level. However, we retained C2 in the main analysis to examine revision-based prediction across the full B1–C2 proficiency range. To address this limitation, we also conducted a sensitivity analysis excluding C2 (N = 970). Although accuracy was slightly higher without C2 (0.612 vs. 0.599), the results are not directly comparable because the classification task changes from four levels to three. Future studies with larger C2 samples are needed to determine whether revision features can reliably distinguish advanced proficiency levels.

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: ChatGPT was used for language editing, including grammar and syntax correction, spell-checking, and improving the clarity and readability of the manuscript. Additionally, ChatGPT provided partial assistance with code development and refinement. All AI-assisted code was reviewed and validated by the authors, who retain full responsibility for the implementation, scientific content, and results.

References

Ahood Al Sawar, Thomas Gaillat, and Nicolas Ballier. 2026. Predicting CEFR writing levels from writing process and linguistic features: a cross-corpus comparison. In *Actes de l'atelier Intelligence Artificielle et ÉDUcation (IA EDU)*, pages 107–125, Nantes, France. ATALA.

Ahood Al Sawar, Erin Pacquetet, Cyriel Mallart, Andrew Simpkin, and Nicolas Ballier. 2025. Predicting CEFR levels for learners of English with keylogging metrics, an exploratory study. In *Actes de l'atelier Traitement de données langagières dynamiques par les outils et méthodes du TAL 2025 (DYN-TAL)*, pages 7–11, Marseille, France. ATALA & ARIA.

Taylor Arnold, Nicolas Ballier, Thomas Gaillat, and Paula Lissón. 2018. Predicting CEFR levels in learner English on the basis of metrics and full texts. *arXiv preprint arXiv:1806.11099*.

Khaled Barkaoui. 2016. What and when second-language learners revise when responding to timed writing tasks on the computer: The roles of task type, second language proficiency, and keyboarding skills. *The Modern Language Journal*, 100(1):320–340.

Khaled Barkaoui. 2019. What can L2 writers' pausing behavior tell us about their L2 writing processes? *Studies in Second Language Acquisition*, 41(3):529–554.

Rianne Conijn, Christine Cook, Menno Van Zaanen, and Luuk Van Waes. 2022a. Early prediction of writing quality using keystroke logging. *International Journal of Artificial Intelligence in Education*, 32(4):835–866.

Rianne Conijn, Emily Dux Speltz, and Evgeny Chukharev-Hudilainen. 2024. Automated extraction of revision events from keystroke data. *Reading and Writing*, 37(2):483–508.

Rianne Conijn, Emily Dux Speltz, Menno van Zaanen, Luuk Van Waes, and Evgeny Chukharev-Hudilainen. 2022b. A product- and process-oriented tagset for revisions in writing. *Written Communication*, 39(1):97–128.

Lester Faigley and Stephen Witte. 1981. Analyzing revision. *College Composition & Communication*, 32(4):400–414.

Thomas Gaillat, Andrew Simpkin, Nicolas Ballier, Bernardo Stearns, Annanda Sousa, Manon Bouyé, and Manel Zarrouk. 2022. Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. *ReCALL*, 34(2):130–146.

Qiao Gan and Benjamin Adams. 2026. From keystrokes to scores: Toward a multidimensional predictive model of writing evaluation by humans and large language models across linguistic, cognitive, and social dimensions. *Assessing Writing*, 69:101090.

Aitor Garcés-Manzanera. 2024. Language bursts and text quality in digital writing by young EFL learners. *Journal of New Approaches in Educational Research*, 13(1):12.

Gaëtanelle Gilquin. 2020. In search of constructions in writing process data. *Belgian Journal of Linguistics*, 34:99–109.

Gaëtanelle Gilquin. 2022. The process corpus of English in education: Going beyond the written text. *Research in Corpus Linguistics*, 10(1).

Gaëtanelle Gilquin. 2023. Causative constructions in process: How do they come into existence in learner writing? *Yearbook of the German Cognitive Linguistics Association*, 11(1):105–120.

Gaëtanelle Gilquin. 2024. Keylogging and screencasting to help investigate L2 writing processes. pages 285–296. Routledge.

Alex Housen, Folkert Kuiken, and Ineke Vedder. 2012. Complexity, accuracy and fluency. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, 32:1–20.

Kristopher Kyle. 2016. *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*. Ph.D. thesis, Georgia State University.

Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior research methods*, 50(3):1030–1046.

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

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Xiaofei Lu. 2017. Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing*, 34(4):493–511.

Cyrielle Mallart, Nicolas Ballier, Jia-Yu Li, Andrew Simpkin, Bernardo Stearns, Rémi Venant, and Thomas Gaillat. 2023. A New Learner Language Data Set for the Study of English for Specific Purposes at University. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 281–287, Portugal. NOVA CLUNL.

Léo Nebel. 2026. *Automation of Feedback on the Writing Process: Construction and Evaluation of Process-Oriented Feedback Based on the Granular Characterization of Revision Episodes*. Ph.D. thesis, Sorbonne Université.

Léo Nebel, François Bouchet, and Vanda Luengo. 2025a. Characterizing revision events in students' writing processes using llms. *Workshop - Writing And Literacy Instruction for Educational Data Mining*. Hal-05186500.

Léo Nebel, François Bouchet, Vanda Luengo, and Mathilde Couraud. 2025b. Towards automated characterization of revision events in student writing. In *Proceedings of the European Conference on Technology Enhanced Learning (ECTEL)*, pages 397–411, Cham. Springer Nature Switzerland.

Erin Pacquetet. 2024. *The Effect of Linguistic Properties on Typing Behaviors and Production Processes*. Ph.d. dissertation, University at Buffalo, State University of New York.

Fabian Pedregosa-Izquierdo. 2015. *Feature extraction and supervised learning on fMRI : from practice to theory*. Theses, Université Pierre et Marie Curie - Paris VI.

Mahmudur Rahman, Pranto Goswamee, Maria Sultana, and Kamruddin Nur. 2025. Typing under pressure: Estimating cognitive load from keystroke dynamics and textual patterns. In *2025 28th International Conference on Computer and Information Technology (ICCIT)*, pages 2521–2526. IEEE.

Soumen Roy, Jitesh Pradhan, Abhinav Kumar, Dibya Ranjan Das Adhikary, Utpal Roy, Devadatta Sinha, and Rajat Kumar Pal. 2022. A systematic literature review on latest keystroke dynamics based models. *IEEE Access*, 10:92192–92236.

Valerie J. Shute. 2008. Focus on formative feedback. *Review of Educational Research*, 78(1):153–189.

Sandip Sinharay, Mo Zhang, and Paul Deane. 2019. Prediction of essay scores from writing process and product features using data mining methods. *Applied Measurement in Education*, 32(2):116–137.

Nancy Sommers. 1980. Revision strategies of student writers and experienced adult writers. *College Composition and Communication*, 31(4):378–388.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 147–153.

Luuk Van Waes and Mariëlle Leijten. 2015. Fluency in writing: A multidimensional perspective on writing fluency applied to L1 and L2. *Computers and Composition*, 38:79–95.

Nina Vandermeulen, Mariëlle Leijten, and Luuk Van Waes. 2020. Reporting writing process feedback in the classroom: Using keystroke logging data to reflect on writing processes. *Journal of Writing Research*, 12(1):109–140.

Georgios Velentzas, Andrew Caines, Rita Borgo, Erin Pacquetet, Clive Hamilton, Taylor Arnold, Diane Nicholls, Paula Buttery, Thomas Gaillat, Nicolas Ballier, and Helen Yannakoudakis. 2024. 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)*, pages 10725–10746, Torino, Italia. ELRA and ICCL.

Chunyan Xu. 2018. Understanding online revisions in L2 writing: A computer keystroke-log perspective. *System*, 78:104–114.