Predicting CEFR writing levels from writing process and linguistic features: a cross-corpus comparison

Ahood Al Sawar [1] Thomas Gaillat [2] Nicolas Ballier[1]
(1) ALTAE, Université Paris Cité, rue Thomas Mann, 75013 Paris, France
(2) LIDILE, Université Rennes 2, Pl. Recteur Henri le Moal, 3500 Rennes, France
ahoodswar2018@gmail.com, thomas.gaillat@univ-rennes2.fr, nicolas.ballier@u-paris.fr

RESUME

Utilisation de métriques d'analyse pour les processus d'écriture

Cet article analyse un ensemble de métriques conçues pour caractériser le comportement des apprenants d'anglais, leurs processus d'écriture et les caractéristiques linguistiques de leurs productions écrites finales. Il explore également si les motifs/patrons observés dans ces métriques peuvent être utilisés pour prédire la maîtrise de l'anglais des apprenants en termes de niveaux du CECR (Cadre européen commun de référence pour les langues). Nous appliquons cette méthode à deux jeux de données, KUPA-KEYS et un corpus en cours de constitution dans le cadre d'un projet qui intègre l'analyse des traces numériques claviers dans une perspective de *learning analytics*.

ABSTRACT

This study evaluates two distinct metric sets: writing process and linguistic features, to profile the behaviour and proficiency of English language learners. It also explores whether patterns observed in these metrics can be used to predict the English proficiency of the learners within the Common European Framework of Reference for Languages (CEFR) levels. We apply this method to two datasets and discuss cross-corpus comparisons.

MOTS-CLES : Analytique de l'apprentissage, Traces numériques clavier, Métriques prédictives, Jets textuels, Corpus d'apprenants, processus d'écriture.

KEYWORDS : Learning Analytics, Keystroke Logging, Predictive Metrics, Textual Bursts, Learner Corpus, Writing Process.

1. Introduction

Advancements in keystroke logging technology have facilitated a deeper understanding of the learning experience and provided unprecedented insights into cognitive writing processes (Tian, & Cushing, 2025; Al Swar *et al*., 2025; Gilquin, 2024; Gilquin, 2022). Traditionally, second language (L2) writing research has prioritized final products and static features, often neglecting the dynamic cognitive processes involved in composition. (Crossley, 2020; Ballier et al., 2020; Gaillat et al., 2022).

Learning analytics has emerged to help us understand and enhance both educational processes and outcomes (Moreno-Marcos et al., 2020; Gašević et al., 2017; Gašević et al., 2015). (Paquette, & Bosch, 2020) have referred to the digital traces generated during writing as "invisible breadcrumbs," highlighting their hidden informative qualities. These traces serve as the basis for predictive models. Learning analytics differentiate between two types of learner metrics, namely, the process features and the product features (Tempelaar et al., 2019). Process features capture the behavioral dynamics of the writing task, such as pausing patterns, burst production, and revision sequences. On the other hand, product features focus on the final outcome and assessment scores. Furthermore, recent research has started exploring whether LLMs can use process-based features in addition to product features to offer learners more informative feedback (Zafar et al., 2025a; Zafar et al., 2025b). However, our study focuses on modeling CEFR levels using interpretable behavioral and linguistic predictors, which provides a clearer understanding of how writing process metrics relate to a learner's writing proficiency.

In this study, we explore the potential for predicting CEFR levels of English learners by analyzing writing processes and linguistic metrics, interpreting the findings within a learning analytics framework. These metrics aim to capture the typing and writing behaviors reflected in learners' texts, moving beyond an exclusive focus on final textual or linguistic features. This paper is organized as follows: Section 2 reviews prior research on automated modeling for second language prediction and learning analytics. Section 3 presents the experimental design along with the feature extraction pipeline. Section 4 reports the findings from the Elastic Net and Ordinal Logistic Regression analyses. Finally, Section 5 discusses the results in the context of learning analytics and provides educational interpretations of the predictive metrics.

2 Background on learning analytics and keystroke logging

2.1 Learning Analytics in Language Learning

This review explores how process and product features are used as predictive metrics and their interpretation in relation to pedagogical decision-making. Process features encompass the behavioural and dynamic aspects of learning, focusing on how learners interact with educational tools. These features include a variety of digital traces, such as temporal process indicators like time-on-task and session duration. Tempelaar et al. (2019) demonstrated, by combining proficiency scores with these temporal traces, valuable insights into learning gains. Specifically, lower-level learners showed higher predictive learning outcomes when they were highly engaged, whereas higher-level learners displayed the opposite trend. Pause patterns and revision behaviour offer valuable insights into the understanding of how L2 writers allocate cognitive activities when composing their texts. Gasevic et al. (2017) demonstrated that learning strategies can be identified through learning analytics. By analysing behavioural traces, such as resource usage patterns and the sequence of learning activities, we can uncover the underlying strategies and approaches students employ in their learning. According to Tempelaar et al. (2019), students who engage in deep learning approaches exhibit distinct behavioural patterns compared to those who adopt surface learning strategies.

Text-based metrics reflect the outcomes and proficiency levels achieved through the learning process. They showcase what learners have produced and serve as indicators of skill mastery and academic performance (Deane, 2014). The most straightforward product features are the assignment grades. (Moreno-Marcos et al., 2020) showed that the best predictors of students' performance were the features related to the exercises, doing even better than process features such as clickstream.

Many studies have found that phrasal features, syntactic complexity, lexical diversity, and syntactic complexity are significant in predicting writing proficiency levels, specifically, the complexity of

noun phrase and lexical sophistication (Arnold et al., 2018; Song, & Cai, 2023; Al Swar et al., 2025). Some studies showed that semantic context and the use of different sentence types or constructional variety are highly predictive in proficiency classification (Monteiro et al., 2023; Hwang, & Kim, 2023). On the other hand, other studies argue that cohesion and discourse features are not significant indicators of writing levels and do not clearly distinguish between proficiency levels (Crossley, & McNamara, 2011; Green, 2012).

2.2 Keystroke Logging for Writing and Language Learning

Research on L2 writing has mostly focused on the final learner submissions, which restricted analysis and overlooked the writing process. However, keystroke technology has bridged this gap and allowed researchers to analyse the real-time behaviour of L2 learners (Conijn et al., 2019; Roeser et al., 2024). Keystroke logging captures all keyboard activities during writing, such as pauses, deletions, and backspacing, offering deeper insights into the cognitive processes involved in composition. This type of data allowed researchers to explore questions that traditional methods could not address, such as how revision behavior varies across proficiency levels and how L2 learners distribute their time among different writing subprocesses. (Gilquin, 2024; Gilquin, 2022; Gilquin, 2020) demonstrated the importance of analysing L2 writing through process data, demonstrating how keystroke logs uncover behaviours not visible in the final version. (Roeser et al., 2024) showed that typing disfluencies can be modelled as both fluent and disfluent. (Velentzas et al., 2024) emphasised the crucial role of collecting large-scale keystroke data for examining English learner writing. To the best of our knowledge, no learning-analytics system like A4ALL exists that collects keystroke data along with the final texts, which aims to link learner proficiency with linguistic features and keystroke data, and provides diagnostic reports for educators and learners.

3. Moodle/A4LL data collection environment

Keylogging is captured within the A4LL system (Venant et al., 2026) as one of several microservices operationalised in its pipeline, along with other microservices such as the computation of textual metrics adapted from the TAASC, TAACO and TAALES tools (Crossley, 2020). The A4LL is compatible with the Moodle CMS via the LTI protocole. For the time being, keylogging data is not incorporated in the CEFR level prediction of the A4LL system, but data collection is underway, so that keylogs data may be incorporated in the future for level predictions. In the current architecture (Figure 1), students' texts are processed with linguistic metrics based on linguistic features and turned into visualisations using probabilistic models accounting for student CEFR predicted level and metrics likely to account for potential CEFR level upgrade. The metrics are grouped as part of a taxonomy of linguistic indicators which are aligned with sub-descriptors of the CEFR, allowing for specific predictive models. Visualisations include bar charts representing the distributions of the generic CEFR levels of the students' texts as well as a radar-chart representation of the CEFR levels per subdescriptor. All visualisations include a list of the main significant linguistic indicators taken into account by the model. These indicators are exploited for feedback generation.

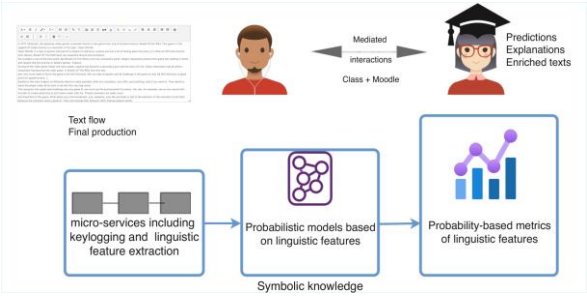

Figure 1 - The keylogging micro-service in the A4LL architecture

4. The datasets

4.1. Dataset 1: KUPA-KEYS

The dataset used in this study is the KUPA_KEYS dataset (Velentzas et al., 2024). This publicly available corpus contains a variety of data, including keystroke data, metadata, final texts, and text labels assigned with CEFR levels by three human raters and the Write & Improve automarker (W&I). To measure inter-rater agreement, they used three measures including Gwet's AC, Spearman's rank correlation, and Root-Mean-Square Deviation. Overall agreement was good, with Gwet's AC = 0.854, and Spearman correlations between raters ranged from 0.514 to 0.711 and were statistically significant. RMSD values demonstrated that the first annotator was closest to the human average (RMSD = 0.622), while the largest disagreement was between the second and third human assessors (RMSD = 2.586), indicating different grading tendencies. Final CEFR labels were calculated by averaging the four scores assigned by the three human annotators and the Write & Improve automarker. Then, the average score was rounded to the nearest integer on the 0-12 scale and mapped to the corresponding CEFR level. The dataset includes 1,006 participants who were asked to complete two tasks: a copy task and an essay writing task. Figure 2 shows the most frequently represented CEFR levels in the dataset were B1, B2, and C1, with B2 being the most dominant class.

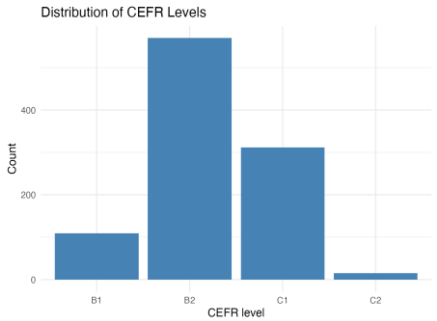

Figure 2 - CEFR levels distribution in the KUPA-KEYS dataset

4.2 Dataset 2:

The CELVA.Sp corpus (Mallart et al. 2023) was collected as part of the Analytics for Language Learning (A4LL) project. It includes texts written by French learners of English as a second language. In addition to these texts, the dataset contains the CEFR levels, which were assessed by expert raters using the DIALANG test, along with metadata about the students. This allows for a preliminary comparison between the two datasets.This study's analysis involved two subsets of the CELVA.Sp dataset, with the first subset including 56 essay texts, with B1 dominating the CEFR distribution as shown in Figure 3. Their skills were only monitored for written comprehension, but this sample is quite representative of the level of French undergraduates. Contrary to the KUPA-KEYS dataset, which was collected on Prolific, the data collection is much more ecological, keylogging data was collected as an activity which is very similar to weekly classes for English for specific purposes as taught at French universities.

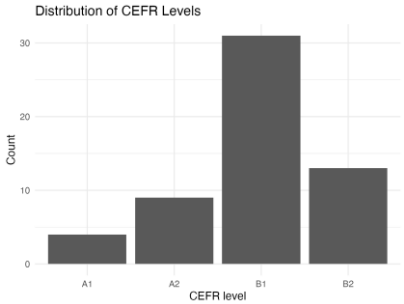

Figure 3 - CEFR levels distribution in the CELVA.Sp dataset

The second subset samples were collected during the 2024-2025 campaign, whereas the first subset was gathered during the 2023-2024 campaign. Consequently, both datasets are from CELVA.sp but were collected in different academic years. For clarity in this paper, we refer to the 2024-2025 dataset as CELVA.sp, and the 2023-2024 dataset as CELVA.sp2. The CELVA.sp2 dataset comprises 232 undergraduate students enrolled in English for Specific Purposes, who were asked to write short essays on their opinions about the best invention in their scientific field. It includes keylogging data from copying and writing tasks. After filtering the data, removing one observation at C1 level and any rows with missing CEFR labels, we retained 206 entries. The CEFR levels were evaluated by four experienced teachers, each with over ten years of expertise. The predominant levels were B1 and B2.

5. Feature sets used in the study

5.1 Process Features

This study analyzed writing process features derived from keystroke logs. We included process features from (Villani et al., 2006) that measure motor and temporal typing behaviors during long text compositions. These metrics reflect low-level typing actions, such as key hold durations and transition times between successive keys. Additionally, we utilised writing process features from (Sinharay et al., 2019), which encompass pause and burst metrics, like pauses before, after, and between words, burst length, start time, time on task, typing speed, as well as revision and editing metrics such as deleted characters, and long jumps.

5.2 Product Features

The set of product features used in this study is based on (Sinharay et al., 2019), capturing textual quality across lexical, syntactic, discourse, and error-related levels. These features include grammar, mechanics, vocabulary richness, word length, syntactic variation, and collocation with prepositions. Additionally, POS-derived linguistic features were added using UDPipe, encompassing proportions of main word classes such as nouns, verbs, and adjectives, as well as textual metrics like lexical density, average and median word length, and preposition ratio. Therefore, this study integrates both traditional essay product features and POS-based linguistic features extracted from the final written text.

In this study, we performed a classification task using two different feature selections. We utilized features from Sinharay et al., 2019, as well as POS features, both with and without the (Villani et al., 2006) features. These metrics can be viewed as learning analytics because they provide behavioral and linguistic insights into the writing process, including how learners write, pause, and revise. Instead of merely serving as prediction inputs, they should be regarded as signs of learner activity. Keystroke and timing features – such as typing speed, start time, pauses, and backspaces – can indicate fluency, hesitation, and self-correction. Burst-based metrics, like burst length, suggest a smoother production flow, acting as fluency markers. When typing speed is paired with a long burst length, it may suggest writing fluency and automatic text production.

Linguistic features derived from the final output, including POS distributions, reveal grammatical and lexical development. Consequently, these features can assist teachers in recognising patterns of fluency, revision, and potential conflicts in learner writing. Table 1 displays some features and their potential interpretation.

Feature	**Type**	**Description**	**Possible learning-analytics interpretation**
Typing speed	Keystroke	Number of characters or words produced per unit of time	Text production fluency/ automatisation
Start time	Temporal	Time before the first keypress after the writing prompt appears	Planning load / hesitation before production
Deleted characters	Revision	Number of characters removed during writing	Self-correction / local revision behaviour
Burst length	Burst-based	Average amount of text produced between pauses	Writing Fluency
Lexical density	Linguistic	Proportion of content words relative to total words	Reliance on content words versus function words

TABLE 1 : A summary of example features and their possible learning analytics interpretations

6. Methodology and modeling approach

6.1 Data Preparation

The data preparation phase involved the systematic transformation of raw keystroke logs and finalised essay texts into a structured format suitable for predictive modeling. First, the raw keystroke logs were cleaned and organized where key press and key release events were paired to compute temporal keystroke measures such as interkey intervals, hold time, and transition variables. Next, we extracted process and product features from keystroke logs and final outputs. After feature extraction, feature tables were merged with the target CEFR labels.

6.2 Modeling Approach

This study employed a hybrid method because the full feature set was unsuitable for ordinal logistic regression, as the model became unstable with a high-dimensional predictor space. To address this, Elastic Net was used to reduce the number of features, manage multicollinearity, and retain the most relevant variables prior to fitting the ordinal model. For illustration's sake of our pipeline, we prompted *ChatGPT.5* to generate Figure 4, which recaps the data preparation and modeling process.

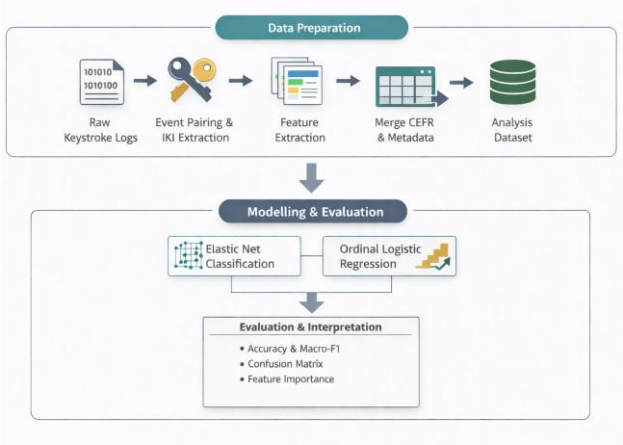

Figure 4 - A simplified overview of the data preparation and modelling pipeline

The dataset was split into stratified training and test sets with an 80/20 ratio by CEFR level, using a fixed random seed. Preprocessing involved ordering the CEFR outcomes, addressing missing values, and removing non-informative features. We used an Elastic Net multinomial logistic regression model to identify the key features, with model parameters optimised via cross-validation. These selected features were then utilised in an ordinal logistic regression model to predict CEFR levels.

7. Results across datasets and feature sets

We first report results with and without biometric features then report preliminary results of our comparison of two datasets.

7.1 Results on KUPA-KEYS with keystroke biometrics metrics

We first used keystroke biometrics metrics (Villani et al., 2006), process and product metrics from (Sinharay et al., 2019), and POS-linguistic features. These were combined to assess their impact on CEFR level prediction and to link them to learning analytics. The findings indicate that these features offered valuable predictive insights. However, due to the dataset's imbalance, performance varied across CEFR categories. As shown in Figure 5, the model predicted B2 level most accurately, with a high classification accuracy of 93.0%, a result likely attributable to the predominance of this proficiency level within the stratified dataset (see Figure 7 for the confusion matrix). Conversely, accuracy for C1 was considerably lower at 17.7%, and the model also performed poorly with B1 and C2, which were less represented. The overall accuracy is 58.5%,with macro-precision of 55.8%, macro-recall/balanced accuracy of 27.7%, and macro-F1 of 49.4%. The mean ordinal absolute error was 0.440 and the quadratic weighted kappa was 0.077.

The key predictive features in Figure 6 included home key use, insert key use, end key use, and typing speed. Additional features such as enter key use, start time, delete key use, and sentence-ending punctuation also emerged. Overall, these results suggest that features related to keyboard navigation, text execution, and writing behavior are highly valuable for prediction.
The ordinal regression coefficients in appendix A1 reveal how these features influence higher and lower levels of writing performance.Typing speed emerged as the primary positive predictor of advanced CEFR levels, suggesting that increased motor fluency is a hallmark of higher proficiency, followed by the use of the End key, the Enter key, and other punctuation marks. In contrast, the use of the Insert key exhibited a strong negative correlation, suggesting that frequent non-linear editing may indicate lower linguistic automaticity, with longer start times, sentence-ending punctuation, and Home key use also indicating lower proficiency. These findings suggest that faster, more efficient typing is associated with higher language proficiency, whereas hesitations, as indicated by longer start times and increased planning load, reflect lower proficiency.

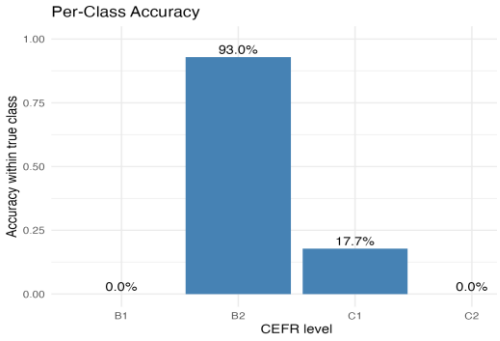

Figure 5 - Per-class accuracy with all sets of features

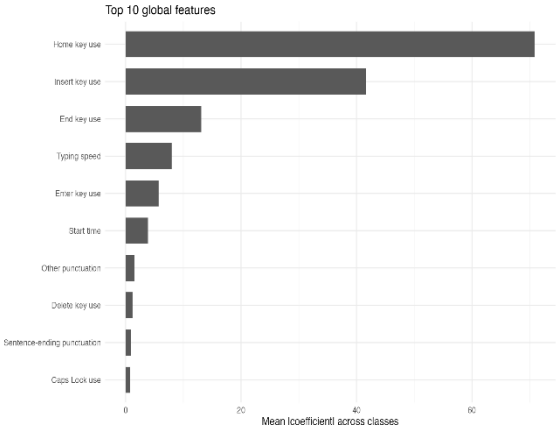

Figure 6 - Top 10 features

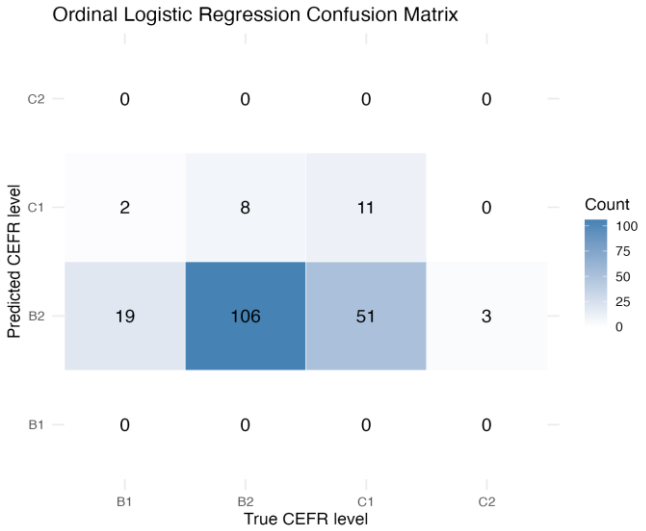

Figure 7 - KUPA-KEYS all features ordinal logistic regression confusion matrix

The second regression experiment includes (Sinharay et al., 2019) metrics and POS-linguistic features, excluding biometric features. The results show that this feature set of metrics continues to be valuable for prediction. However, similar to the previous experiment, the model's performance varied across CEFR levels, as expected. Figure 8 highlights this bias, caused by an imbalance in observations, which were more frequent at the B2 level with slightly lower accuracy than the first set of features, yet still high at 89.5%. The overall accuracy using this set of features is the same as the previous accuracy, 58%, with macro-precision of 55.0%, macro-recall/balanced accuracy of 28.4%, and macro-F1 of 52.2%. For this model, the MOAE was 0.435 and the QWK was 0.143.

The most predictive feature, as shown in Figure 9, was start time, which was significantly more important than other features. The second most influential was typing speed, followed by a group of POS-based variables, including the proportion of adjectives, auxiliaries, pronouns, collocation prepositions, adpositions, determiners, and deleted characters. These results suggest that both temporal processing features and linguistic distributional features (POS-based metrics) are crucial for predicting CEFR levels.

The coefficients from the ordinal regression, shown in Appendix A2, reveal the effect directions. Typing speed had the strongest positive coefficient, indicating that faster writing correlated with higher CEFR levels, and was the second most important feature in the first feature set. Other features like collocation preposition, proportion of adjectives, and lexical density also correlated with better writing performance, though their effects were weaker than typing speed. Conversely, start time exhibited the strongest negative coefficient, associated with lower CEFR levels. Other negatively associated features included the proportion of pronouns, determiners, deleted characters, and auxiliaries. These findings imply that longer pauses before writing and less variation in linguistic patterns are linked to lower writing proficiency as has been mentioned previously with the model results using the first set of features.

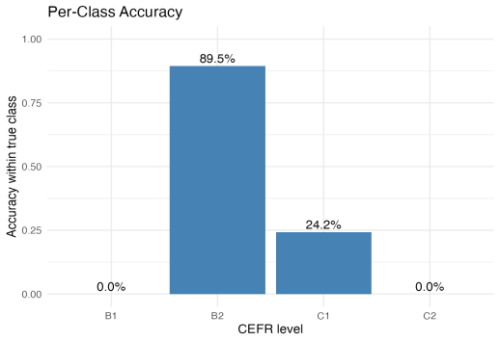

Figure 8 - Per-class accuracy with process features

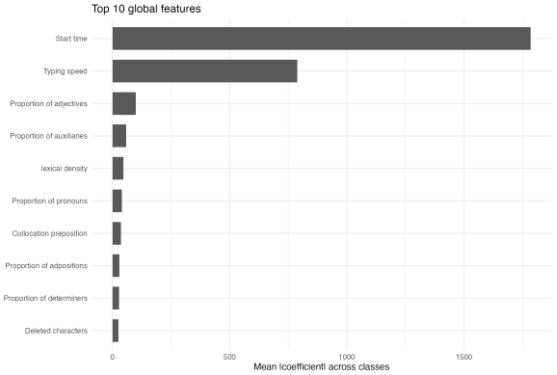

Figure 9 - Top 10 features

Despite the challenges posed by dataset imbalance, both models demonstrated consistent performance in identifying the B2 level compared to other proficiency categories. However, the importance and types of features varied significantly between the two settings. In the first set, the key metrics mainly involved keystroke navigation variables such as Home and End key usage. Conversely, the second set shifted focus from navigation behaviour to temporal writing signals and POS-based linguistic features. It is worth noting that both typing speed and start time appeared in the two feature sets, indicating their significance in predicting performance levels. Additionally, deleting behavior also appears in both sets, emphasizing the role of revision as a predictive indicator.

7.2 Preliminary Results on Dataset 2 and cross-dataset interoperability

In the CELVA.Sp dataset, the key features identified by the elastic net and ordinal models include behavioural keyboard metrics such as home key use, typing speed, delete key use, and enter/ctrl key use, and they were associated with higher proficiency levels (see Figure 11). The class-level accuracy was higher for B1 than B2 as shown in Figure 10, reflecting the dataset's different distributions. When biometric features were excluded, the model shifted its focus to linguistics and process-based variables, notably typing speed, the proportion of verbs, and TTR which were the most predictive features. Figure A3 in Appendix indicates that higher proficiency levels are associated with higher proportions of verbs and noun phrases, whereas typing speed and TTR have negative coefficients. Compared to the KUPAKEYS dataset, the overall pattern remains consistent: behavioural data dominate when biometric features are included, but excluding them shifts importance toward linguistic and process-based features. This is expected, given the nature of each feature set. In KUPAKEYS, with process-based and linguistic features, the model is more sensitive to POS distribution and lexical density, whereas in CELVA.Sp, it emphasises lexical diversity, syntactic structure, and sentence-level organisation, indicating a shared pattern with dataset-specific linguistic focuses. Including all features, the model reached an accuracy of 77.8%, 67.9% macro-precision, 50.0% macro-recall/balanced accuracy, and 71.2% macro-F, whereas excluding the initial set of features resulted in only 22.2%, 16.7% macro-precision, 33.3% macro-recall/balanced accuracy, and 50.0% macro-F1. This highlights the significant role of biometric features in predicting CEFR proficiency levels. For CELVA.sp with all features, the MOAE was 0.333 and the

QWK was -0.098, while without Villani et al. (2006) features, the MOAE was 0.333 and the QWK was 0.341. Given the small CELVA.sp test set, these ordinal metrics should be interpreted cautiously.

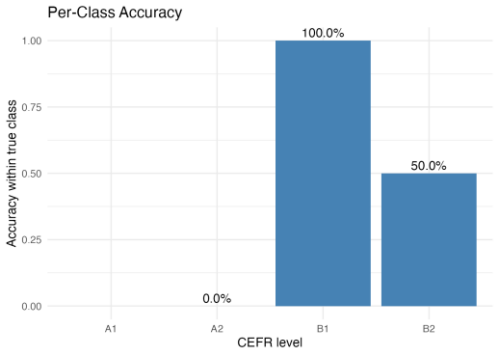

Figure 10 - CELVA.Sp dataset per-class accuracy with all sets of features

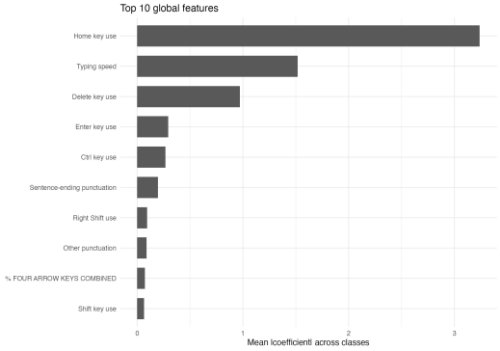

Figure 11- CELVA.Sp dataset top 10 features

The model's performance with the CELVA.sp2 dataset, using all features, reached 34.3% accuracy, 23.6% balanced accuracy, and 40.4% macro-F1. Figure 11 demonstrates how the model classified B1 most accurately at 64.3%. Excluding biometric features improved the metrics to 44.7% accuracy, 33.4% balanced accuracy, and 41.5% macro-F1, with better ordinal agreement as QWK increased from 0.169 to 0.327. Regarding feature importance, patterns were similar to previous datasets: when including all features, the model favored keystroke and navigation variables such as End key, Insert key, Home key use, and typing speed (see Figure 12). Without the behavioural features set, the model prioritised fluency and linguistic features like typing speed, adverb proportions, start time, and lexical density(see Figure 13). Although the CELVA.sp2 dataset yielded lower overall performance compared to CELVA.sp, it provides a more stable evaluation due to its larger sample size. In this dataset, the use of End key and typing speed showed positive association with higher CEFR levels, however, Home key use and Insert key use were associated with lower levels. Some feature effects were not fully consistent across datasets. From a learning analytics perspective, these differences indicate that keystroke features are useful for understanding learner behavior. However, their interpretation should be approached with caution, considering data collection settings, writing task context, and learner profiles.

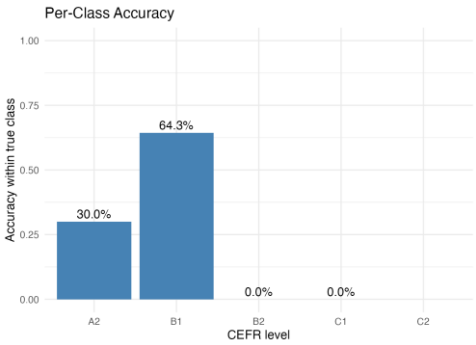

Figure 11 - CELVA.Sp2 dataset per-class accuracy with all sets of features

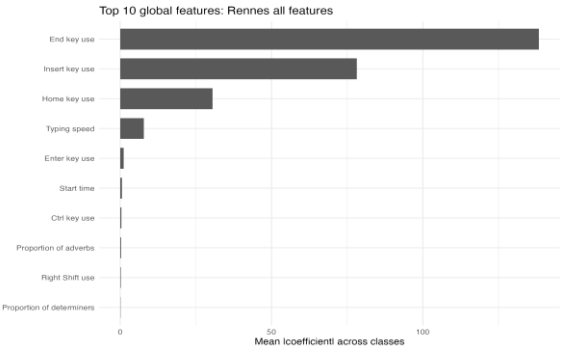

Figure 12- CELVA.Sp2 dataset top 10 features

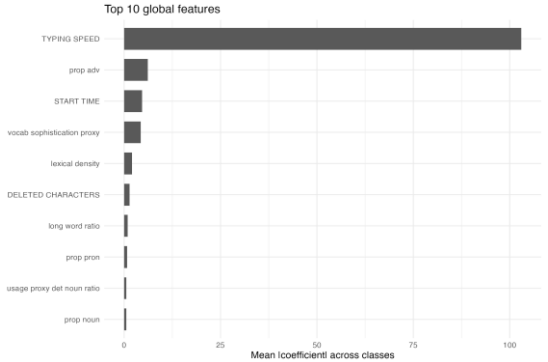

Figure 13- CELVA.Sp2 dataset top 10 Features without behavioural metrics

7.3 Cross-Dataset Comparison

One way to begin to assess the comparability of the two datasets is to question the potential discrepancy between the copy task and the essay writing task. This has been used in Bayesian modeling of keylog data (Conijn et al., 2019; Roeser et al, 2024). As a first step for this modelling, we tested the duration distribution of the two tasks across our two datasets. For this, we normalized and compared the probability density representation of the duration of the copy task, bearing in mind that the size of the copy task was different in the two cases. In CELVA.Sp, the copy-task consisted of 73 words, whereas KUPA-KEYS included 300 words. As shown in Figure 14, both datasets exhibit a similar pattern in copy-task distribution, with copy tasks being completed faster than essay tasks. The bimodal shape of the CELVA.sp copy distribution is due to its smaller size and to three slower typists who took twice as much time. In Appendix A.4, we plot the comparison of the two datasets without these three outliers. The different conditions in which the data was collected is reflected in the horizontal dimension for the CELVA.Sp, where participants had more time to produce the essay.

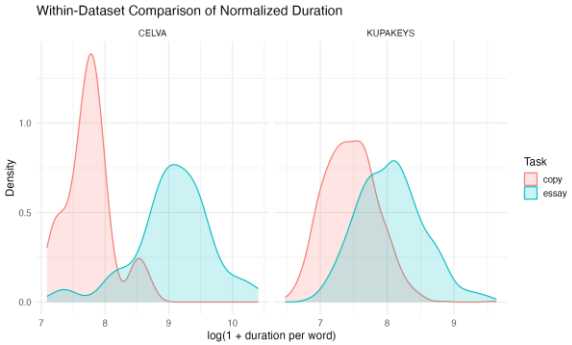

Figure 14 - Within-dataset probability density comparison

8. Discussion

The tested metric results indicate that CEFR prediction depends on keystroke behavior and revision features (such as navigation keys and deletion), real-time process timing features (like typing speed and start time), and POS-based linguistic features (including collocation use and lexical density). These findings align with previous research and demonstrate that keystroke logs reveal aspects of the writing process hidden from simply examining the final product (Gilquin, 2022). Features like

start time have been emphasized in (Al Swar et al., 2025), where the authors observed that long pre-burst pauses correlate with lower proficiency levels, possibly reflecting high cognitive load. Furthermore, they found that revision behavior is among the most predictive features for identifying CEFR levels. They noted that frequent revisions within words are associated with lower-level writers, which aligns with our findings that deleting characters indicates lower proficiency. (Pacquetet, 2025) found that collocations and phrasal expressions are produced with fewer pauses and disfluencies. Additionally, (Al Sawar et al., 2025) indicated that collections and formulaic expressions are strong predictors of learners' proficiency. Furthermore, (Gilquin, 2024) demonstrated that structurally complete units are processed holistically and are associated with higher text quality, supporting our findings. Overall, these findings not only offer strong predictive insights into learners' CEFR levels but also facilitate the interpretation of learning analytics based on the feature set. Temporal features (e.g., typing speed and start time) may indicate fluency and cognitive load, while revision-related features (such as deleted characters) are also significant. Additionally, the use of the home key and insert key can be seen as revision-related, as they are employed to move the cursor to the beginning of a line and to insert or overwrite text during editing. These preliminary findings should be interpreted as a pilot study exploring the feasibility of dataset interoperability and cross-corpus comparison. In the same way as several learner spoken corpora can be analysed for phonetic investigation of learner speech (Ballier & Martin, 2013) across several mother tongues and corpora, this pilot study calls for the need to develop corpus interoperability and explores transferable analyses for keylogging corpora. In spite of the differences in data-collection, data collection conditions and corpus size, we should find ways to analyse writing as a process with keylogging data, to enrich previous analyses based on finalised versions of the different texts (Kashefi et al., 2022). This is likely to foster the use of keylogging data for learning analytics, using revision detection methods (Nebel et al., 2025) or interfaces monitoring keylogging data as learners write (Nebel, 2026). In the current version of the A4LL system, keylogging data is captured and processed to generate keylog-based metrics but similar monitoring devices could be included in the A4LL dashboard (Venant et al., 2026) based on the KREV python library (Nebel, 2026).

9. Conclusion and future work

This study explores whether regression models can predict CEFR proficiency levels using keystroke-derived and POS-based features. Generally, the features provide valuable predictive insights, but due to data imbalance, performance varies across proficiency levels, with B2 being most prevalent in our data. The key features fall into three categories: temporal features, keystroke behavior and revision features, and POS-based features. These results highlight that learners' real-time behavior can reveal differences in CEFR levels, not just through final product analysis. However, despite their predictive value, class imbalance remains a significant limitation. From a learning analytics perspective, these metrics serve not only for prediction but also as indicators of learners' behavior during writing. Temporal variables can reflect fluency and planning effort, revision-related metrics may indicate editing behavior and changes made, and POS-based features might reveal learners' linguistic patterns. Therefore, this study extends beyond just assessing the final product to gain a deeper understanding of L2 writing and suggests that keylogging can be used for learning analytics. A promising avenue for future research involves integrating keystroke data into Large Language Models (LLMs) to provide more nuanced and motivational student feedback (Zafar et al., 2025).

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). Additionally, we acknowledge the use of NotebookLM for support in partially refining the wording and readability of this paper.

References

AL SAWAR A., PACQUETET E., MALLART C., SIMPKIN A.J., & BALLIER N. (2025). Predicting CEFR levels for learners of English with keylogging metrics, an exploratory study In, *Actes de 20e Conférence en Recherche d'Information et Applications (CORIA) 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN) 27ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL) Les 18e Rencontres Jeunes Chercheurs en RI (RJCRI) (CORIA-TALN'2025), Marseille (France)*.

ARNOLD T., BALLIER N., GAILLAT T., & LISSÓN P. (2018). Predicting CEFR levels in learner English on the basis of metrics and full texts In. *Conference paper presented at Conférence sur l'Apprentissage Automatique (CAp) 2018*, Rouen, France, INSA Rouen. *arXiv:1806.11099*.

BALLIER N. & MARTIN, P. (2013). Developing corpus interoperability for phonetic investigation of learner corpora. *Automatic treatment and analysis of learner corpus data,* Benjamins, 33-64.

BALLIER N., CANU S., PETITJEAN C., GASSO G., BALHANA C., ALEXOPOULOU T., & GAILLAT T. (2020). Machine learning for learner English: A plea for creating learner data challenges *International Journal of Learner Corpus Research*, 6(1), 72-103.

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

CONIJN R., VAN ZAANEN M., LEIJTEN M., & VAN WAES L. (2019b). How to Typo? Building a Process-Based Model of Typographic Error Revisions *The Journal of Writing Analytics*, 3(1), 69-95.

CROSSLEY S.A. (2020). Linguistic features in writing quality and development: An overview *Journal of Writing Research*, 11(3), 415-443.

CROSSLEY S.A., & MCNAMARA D.S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing *International Journal of Continuing Engineering Education and Life Long Learning*, 21(2-3), 170-191.

DEANE P. (2014). Using Writing Process and Product Features to Assess Writing Quality and Explore How Those Features Relate to Other Literacy Tasks *ETS Research Report Series*, 2014(1), 1-23.

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

GAILLAT T., SIMPKIN A., BALLIER N., STEARNS B., SOUSA A., BOUYÉ M., & ZARROUK M. (2022). Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach *ReCALL*, 34(2), 130-146.

GAŠEVIĆ D., DAWSON S., & SIEMENS G. (2015). TechTrends, 59(1), 64-71.

GAŠEVIĆ D., KOVANOVIĆ V., & JOKSIMOVIĆ S. (2017). Learning: Research and Practice, 3(1), 63-78.

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.
GREEN C. (2012). A Computational Investigation of Cohesion and Lexical Network Density in L2 Writing *English Language Teaching*, 5(8), 57-69.
HWANG H., & KIM H. (2023). Automatic Analysis of Constructional Diversity as a Predictor of EFL Students' Writing Proficiency *Applied Linguistics*, 44(1), 127-147.
KASHEFI, O., AFRIN, T., DALE, M., OLSHEFSKI, C., GODLEY, A., LITMAN, D., & HWA, R. (2022). ArgRewrite V.2: An annotated argumentative revisions corpus. *In Language Resources and Evaluation* 56.3, pp. 881–915.
MALLART C., BALLIER N., LI J. Y., SIMPKIN A., STEARNS B., VENANT R. & GAILLAT T. (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*, p. 281–287.
MONTEIRO K., CROSSLEY S., BOTARLEANU R.-M., & DASCĂLU M. (2023). L2 and L1 semantic context indices as automated measures of lexical sophistication *Language Testing*, 40(3), 576-606.
MORENO-MARCOS P.M., PONG T.-C., MUÑOZ-MERINO P.J., & DELGADO KLOOS C. (2020). Analysis of the Factors Influencing Learners' Performance Prediction With Learning Analytics *IEEE Access*, 8, 5264-5282.
NEBEL, L., BOUCHET, F., LUENGO, V., & COURAUD, M. (2025). Towards Automated Characterization of Revision Events in Student Writing. In *European Conference on Technology Enhanced Learning* (pp. 397-411). Cham: Springer Nature Switzerland.
NEBEL, L. (2026) *Automation of Feedback on the Writing Process Construction and evaluation of process-oriented feedback based on the granular characterization of revision episodes,* PhD thesis, Sorbonne Université.
PACQUETET, E. (2024). The Effect of Linguistic Properties on Typing Behaviors and Production Processes PhD thesis, University of Buffalo, ProQuest LLC.
PAQUETTE L., & BOSCH N. (2020). The Invisible Breadcrumbs of Digital Learning: How Learner Actions Inform Us of Their Experience In, *Handbook of Research on Digital Learning*, p.302-316, IGI Global Scientific Publishing.
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.
SINHARAY S., ZHANG M., & DEANE P. (2019). Applied Measurement in Education, 32(2), 116-137.
SONG K., & CAI Y. (2023). Predicting Chinese Efl Learners' Academic Writing Quality Through Indices of Linguistic Complexity Rochester, NY, Social Science Research Network.
TEMPELAAR D., RIENTIES B., & NGUYEN Q. (2019). Learning Engagement, Learning Outcomes and Learning Gains: Lessons from LA International Association for the Development of the Information Society.
TIAN Y., & CUSHING S.T. (2025). Exploring the application of keystroke logging techniques to research in second language (L2) writing *Research Methods in Applied Linguistics*, 4(1), 100179.
VELENTZAS G., CAINES A., BORGO R., PACQUETET E., HAMILTON C., ARNOLD T., NICHOLLS D., BUTTERY P., GAILLAT T., BALLIER N., & YANNAKOUDAKIS H. (2024). Logging Keystrokes in Writing by English Learners In, CALZOLARI N., KAN M.-Y., HOSTE V., LENCI A., SAKTI S., & XUE N., Éd., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p.10725-10746, Torino, Italia, ELRA and ICCL.
VENANT R., BALLIER N., MALLART C., SIMPKIN A., STEARNS B., LI J.-Y., & GAILLAT T. (2026) Un système de learning analytics linguistiques actionnables pour l'apprentissage de l'anglais en contexte universitaire, *Actes de la conférence TALN 2026*, atelier IA & ÉDUcation, Nantes.

VILLANI M., TAPPERT C., GIANG NGO, SIMONE J., FORT H.ST., & SUNG-HYUK CHA (2006). In, 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), p.39-39, New York, NY, USA. IEEE.
ZAFAR S., MINHAS S., ZAIDI S.A.H., NAEEM A., & ALI Z. (2025a). « I Wrote, I Paused, I Rewrote » Teaching LLMs to Read Between the Lines of Student Writing *arXiv:2506.08221*.
ZAFAR S., YOUSAF S., & MINHAS M.S. (2025b). « Can You See Me Think? » Grounding LLM Feedback in Keystrokes and Revision Patterns *arXiv:2508.13543*.

A Appendix

A.1 Ordinal logistic regression feature coefficients for KUPA-KEYS with behavioural metrics

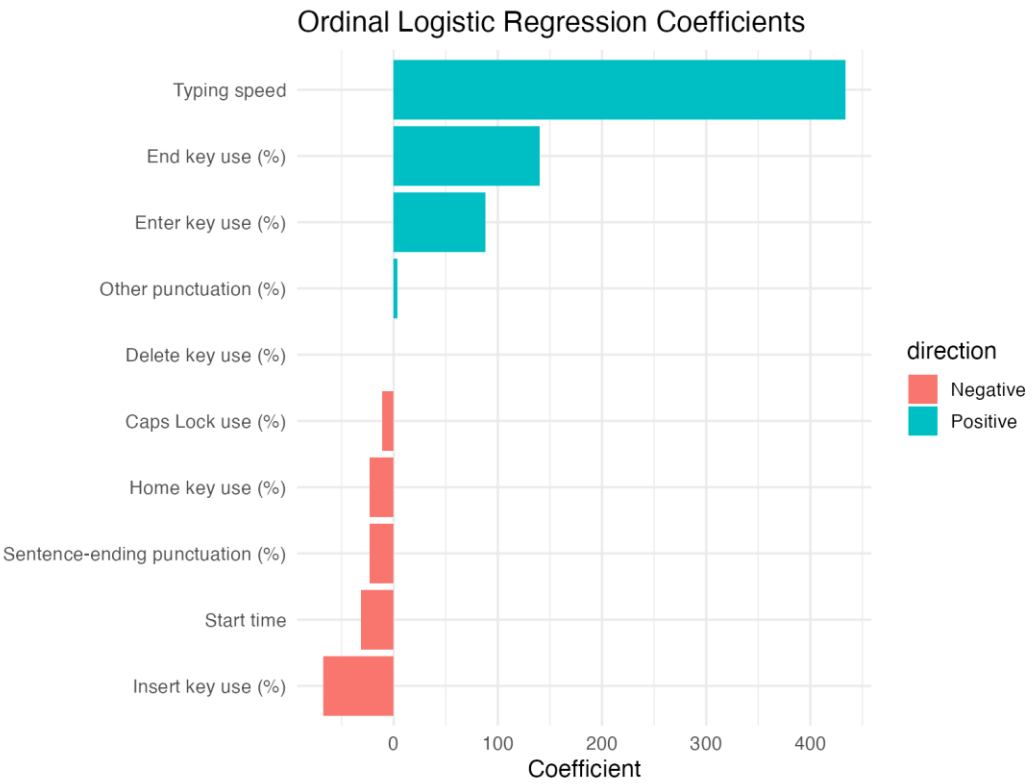

Figure A1 - Ordinal logistic regression feature coefficients with behavioral metrics

A.2 Ordinal logistic regression feature coefficients for KUPA-KEYS with linguistically oriented metrics

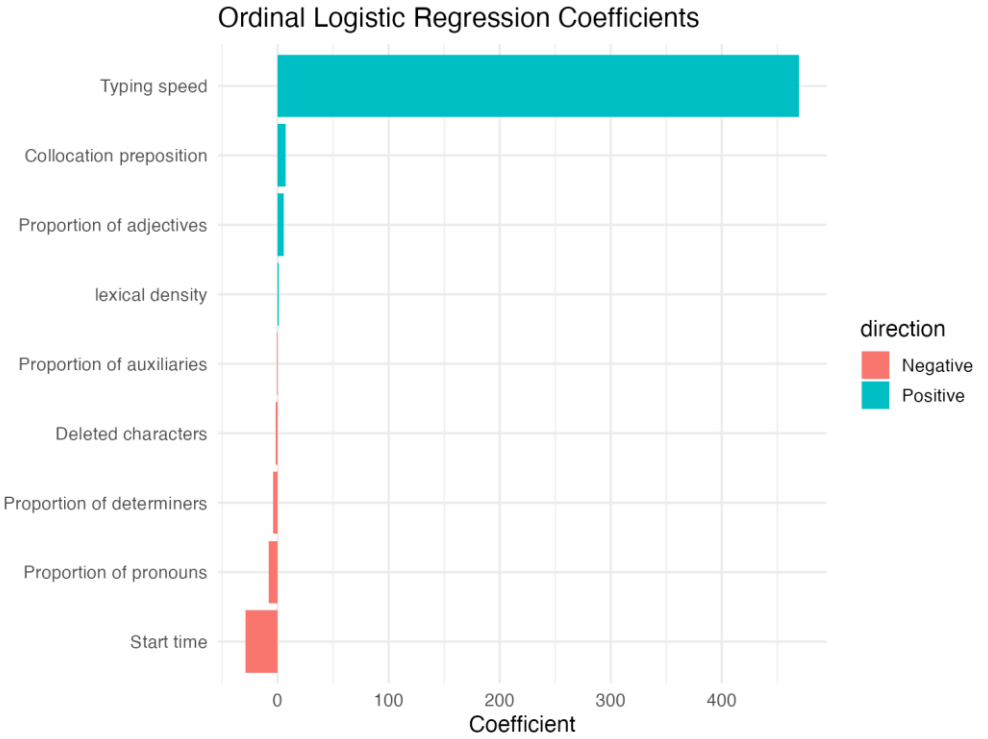

Figure A2 - Ordinal logistic regression feature coefficients with linguistic metrics

A.3 Ordinal logistic regression feature coefficients for CELVA.Sp with linguistically oriented metrics

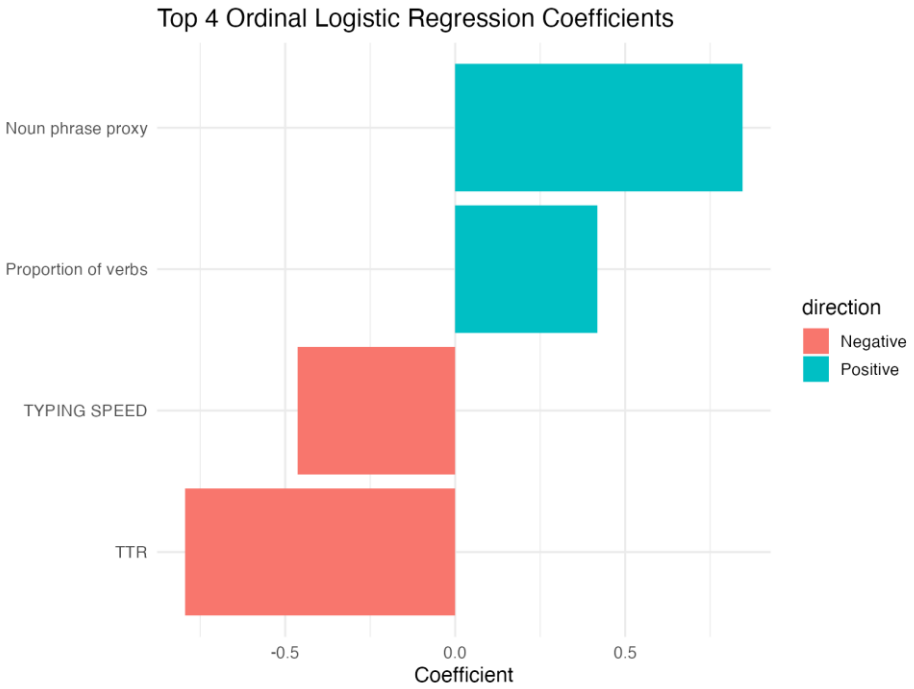

Figure A3 – CELVA.Sp Ordinal logistic regression feature coefficients with linguistic metrics

A.4 Ablation of the three copy task outliers in the CELVA.Sp dataset

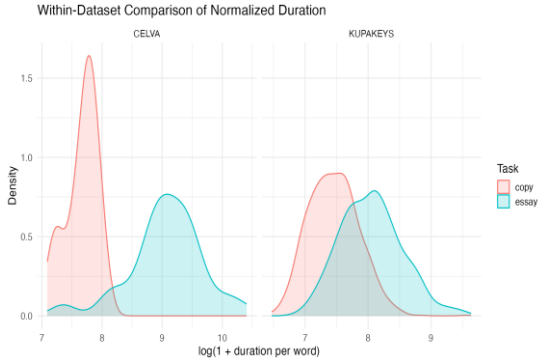

Figure A4- Within-dataset probability density comparison. We have taken off the three outliers for the copy task (while the remaining copy cases averaged 2.66 minutes and 2199 ms per word, these three cases averaged 6.21 minutes and 5172 ms per word for the same 72-word copy task). This figure was plotted after deleting these three outliers.