Understanding the Dynamics of Second Language Writing through Keystroke Logging and Complexity Contours

Elma Kerz[1], Fabio Pruneri[2], Daniel Wiechmann[3], Yu Qiao[1], Marcus Ströbel[1]
RWTH Aachen University[1], Harvard University[2], University of Amsterdam[3]
elma.kerz@ifaar.rwth.aachen.de, fabiopruneri@college.harvard.edu, d.wiechmann@uva.nl,
yu.qiao@rwth-aachen.de, marcus.stroebel@ifaar.rwth-aachen.de

Abstract

The purpose of this paper is twofold: [1] to introduce, to our knowledge, the largest available resource of keystroke logging (KSL) data generated by Etherpad (https://etherpad.org/), an open-source, web-based collaborative real-time editor, that captures the dynamics of second language (L2) production and [2] to relate the behavioral data from KSL to indices of syntactic and lexical complexity of the texts produced, obtained from a tool that implements a sliding window approach capturing the progression of complexity within a text. We present the procedures and measures developed to analyze a sample of 14,913,009 keystrokes in 3,454 texts produced by 512 university students (upper-intermediate to advanced L2 learners of English) (95,354 sentences and 18,32,027 words) aiming to achieve a better alignment between keystroke-logging measures and underlying cognitive processes, on the one hand, and L2 writing performance measures, on the other hand. The resource introduced in this paper is a reflection of increasing recognition of the urgent need to obtain ecologically valid data that have the potential to transform our current understanding of mechanisms underlying the development of literacy (reading and writing) skills.

Keywords: keystroke logging, ecologically valid data, contextualized data, language behavior data, second language writing, linguistic complexity contours, sliding window technique

1. Introduction

Recent years have witnessed increased efforts in obtaining ecologically valid data that provide a valuable resource for researching language production and comprehension as well as language development in both native and non-native speakers across multiple components of the linguistic systems. Two recent corpus resources in the area of reading are (1) the Ghent Eye-Tracking Corpus (GECO) (Cop et al., 2017), an eye-tracking corpus of monolingual and bilingual sentence reading containing data of reading 5,000 sentences from monolingual and bilingual English speakers and (2) the Provo Corpus, a large eye-tracking corpus with predictability norms for studying predictive processes in reading (Luke and Christianson, 2018). Such corpora are suitable for both exploratory purposes and more direct hypothesis testing providing an optimal basis for formulating the central assumptions and theoretical frameworks accounting for naturalistic reading processes in a meaningful context. These resources have the potential to inform current reading models and provide new insights into mechanisms and principles underlying reading and reading development more generally.

Research on writing, the second foundational literacy skill, may also profit from such a resource of ecologically valid data capturing the unfolding process of writing. One such tool that is well-suited to this purpose is keystroke logging (KSL) (Leijten and Van Waes, 2013). KSL is a methodological approach to automatically recording every keystroke – and, by extension, every cursor and mouse movement – an individual undertakes when writing on a computer to a logfile. Such recordings provide an unobtrusive record of the moment-by-moment creation of the text and support detailed analyses of the pauses, movements and revisions made during writing. A growing body of research has employed such KSL data to investigate various aspects of text, writing processes and writing development (for overviews, see, e.g., Van Waes, Leijten, Lindgren, & Wengelin, 2016). One line of research has successfully employed KSL to investigate how writing processes may be affected by varying degrees of cognitive load through manipulations of task complexity and input mode (Leitjen et al. 2010; Nottbusch 2010; Quinlan et al. 2012; Sahel et al. 2008; Van Waes et al. 2010). Another line of research has explored how KSL measures are associated with writing performance, operationalized either in terms of human ratings of text quality or through the measurement of the complexity of the writing samples (Alves et al., 2008; Medimorec and Risko, 2017; Zhang et al. 2016). For example, in their analysis of the complexity and quality of university students' written productions, Alves et al. (2008) found that the texts produced by fast typist were overall judged to be of higher quality and showed higher degrees of lexical complexity than those of slow typists. However, this previous work has been exclusively confined to global assessments of text complexity, i.e. representing text complexity in terms of a single score for a given indicator. More recently, an alternative approach to the assessment of text complexity has been put forward that implements a sliding window technique to track the progression of complexity within a text (see, Ströbel, 2014; Ströbel et al. 2016, 2018, to appear). In this approach, complexity scores are obtained for each sentence in a text allowing for a more local, higher-resolution assessment of complexity (see Section 4 for more details). This enables a better alignment between KSL measures and measures of complexity of the written products.

Writing research utilizing KSL has largely relied on specialized software, such as InputLog (http://www.inputlog.net/), designed to log and time stamp keystroke activity to reconstruct and describe writing processes on a computer that comes with ready made linguistic analysis (for a review, see Latif, 2009; Van Waes, Leijten, Wengelin, et al., 2012). In

```
{"changeset":"Z:1>0$","meta":
        {"author":"a.WTisTAb5lUQZeRLB",
        "timestamp":1555430466729}},
{"changeset":"Z:1>1*0+1$T","meta":
        {"author":"a.WTisTAb5lUQZeRLB",
        "timestamp":1555443863479}},
```

Figure 1: Example changeset output (JSON format) from a random participant gathered using the Etherpad text online text editor.

addition, this research has mainly been conducted in laboratory settings under highly controlled conditions with a relatively small number of participants.

Here we present a novel approach to investigating mechanisms underlying second language production based on a combined use of keystroke-logging measures obtained from an open-source, web-based real-time text editor, Etherpad (https://etherpad.org/), and complexity contours computed by a software that implements a sliding window technique to capture the progression of complexity across text.

2. Compilation of the Keystroke Logging Corpus of Second Language Writing in University Students

The keystroke logging data presented here were collected as part of a larger project aiming to advance our understanding of second language performance, proficiency and development using ecologically valid, dense longitudinal data from a large sample of students. The data currently consist of 3,434 text files and keystroke logs amounting to 1.82 million words of academic writing produced by 512 students at RWTH Aachen University over the course of one semester (approx. 12-14 weeks). University students produced a series of 'learning journals', i.e. reflective writing assignments on the main aspects covered in a lecture. Students were instructed to compose their texts using Etherpad 1.8.0-beta.1 (https://etherpad.org/), an open-source online text editor, which records every keystroke entered into the editing pad, including edits, deletes, copy/pastes, etc. and stores these data in a changelog file in compressed JSON format (see Figure 1). Next to information containing a unique user-ID (attribute: 'author') and information on when a given event occurred (attribute: 'timestamp'), the JSON files contain an attribute called 'changeset' whose value is a string, such as 'Z:z>1|2=m=b*0|1+1', that encodes information about the difference between two revisions of the document. Extracting the relevant keystroke logging data from the JSON files involved several steps: In a first step, a node.js script was used that parsed the JSON and wrote the changesets in separate lines into a new file. This file was then processed with a C++ script that utilized tools provided by the Changeset library (https://github.com/ether/etherpad-lite/wiki/Changeset-Library) to extract the information in the changesets and store it in a CSV file. The transformed output format resembles that of state-of-the-art keystroke-logging software such as InputLog (http://www.inputlog.net/) and is illustrated in Table 1.

s_id	u_id	file	char	op	pos	t
7	0	01894	t	+	687	5138
7	0	01894	H	+	687	502
7	0	01894		-	687	251
7	0	01894		-	687	252
7	0	01894	T	+	687	961
7	0	01894	h	+	688	253
7	0	01894	e	+	689	254
7	0	01894	s	+	690	172
7	0	01894	e	+	691	172
7	0	01894		+	692	172
7	0	01894	f	+	693	270
7	0	01894	o	+	694	270
7	0	01894	u	+	695	119
7	0	01894	r	+	696	119
7	0	01894		+	697	119
7	0	01894	c	+	698	120
7	0	01894	a	+	698	171
7	0	01894	t	+	698	171
7	0	01894	e	+	698	171
7	0	01894	g	+	698	501
7	0	01894	o	+	698	127
7	0	01894	r	+	698	128
7	0	01894	i	+	698	128
7	0	01894	e	+	698	128
7	0	01894	s	+	698	168

Table 1: Example of keystroke logging data after transformation. Every row represents one log event.: s_id = unique sentence ID; u_id = unique user ID; file = unique text ID; char = input character, op = operation (addition (+) or deletion (-), pos = unique text position, t = time since preceding keystroke (in ms)

Next we removed all contaminated data. Some of the students did not follow the instructions for the proper use of the Etherpad application, and instead of typing the text directly into the editor, copy-and-pasted text passages from other sources, thereby corrupting their keystroke log. Texts that contained any materials that were copy-and-pasted from another source were removed. From the remaining texts, all sentences that did not appear in the final version of the text were removed. Finally, we removed all text segments that comprised less than 20 characters, which included paragraph titles, annotations, and bullet points, characters outside the English alphabet as well as dates, titles, and author names. The preprocessed keystroke logging data amounted to a total of 3434 texts produced by 512 individuals (mean number of texts per user 6.73; SD = 3.64), dividing into 94,927 sentences containing 1,823,327 words, and a total of 14,849,245 keystrokes.

2.1. Keystroke Logging Measures Derived from the Corpus

The basic rationale underlying keystroke logging is that measures of writing fluency that can be derived from keystroke data can reveal traces of the underlying cognitive processes (see MacArthur, Graham, & Fitzgerald, 2008). Fluent writing processes are generally characterized by a high production rate (e.g. many words per minute), short pausing times and a low number of revisions (see, Leijten and van Waes, 2013). As in spoken production, pausing is seen to signal cognitive effort with the length and numbers of pauses between different text units (characters, words or sentences) being related to different morphological, grammatical, and discourse processes (cf. Wengelin, 2006, Nottbusch, Grimm, Weingarten, & Will, 2005; Spelman Miller, 2006). Revisions are seen to index a discrepancy between the writers' intentions and the text produced so far (Leijten, Van Waes, & Ransdell, 2010; Lindgren, Sullivan, & Spelman Miller, 2008) and can result from surface-related, grammatical, or content-related issues recognized by the writer during text production (Van Waes & Schellens, 2003). Writing fluency is thus a multi-faceted construct related to various aspects of the speed of 'production', 'revision', and 'pause behavior' (see, e.g., Van Waes and Lijten, 2015).[1] Table 2 presents the corpus-level descriptive statistics of a set of fifteen indicators of these aspects. The indicators in the 'production' group include the average number of words - defined as character strings surrounded by punctuation marks - and keystrokes produced within a text or a sentence as well as the average number of deleted characters per sentence. This group also includes 'inter-keystroke intervals' (IKI, aka 'interkey-transition times') defined as the time between two consecutive key-presses typically expressed in ms. Etherpad records keystrokes at regular intervals of 500ms. In case two or more key-presses were produced within that timespan, it was assumed that they occurred at regularly spaced intervals. The final four indicators in that group describe the average time in seconds spent on producing a sentence in total ($\text{SPT}_{\text{total}}$), only the actual lexical material of the sentence ($\text{SPT}_{\text{words}}$), only whitespaces and other punctuation marks ($\text{SPT}_{\text{non-words}}$), and on edits of the text (see below) (SPT_{edit}). The next group contains indicators of 'pause behavior', specifically the average number of pauses in a sentence. Pauses are defined based on three different IKIs thresholds: An IKI threshold of 2000ms is typically used in the literature (see Baaijen et al, 2012), although lower thresholds for pause behaviour have been considered more appropriate (Schilperoord, 2001). For purposes of comparison, we also report pause frequency for a threshold of 510ms, which was the median IKI score in our data, as well as for pauses of 10 seconds or longer. As the writing task was untimed and users could complete it in multiple sessions, meaning that IKIs can potentially be very high (hours or even days); to prevent the estimates of text production times to be distorted by such breaks during the writing process, all pauses greater than 10 seconds were reduced to that value (10000 ms). The last group of measures concerns 'revision' (edits), i.e. changes of the text at non-consecutive positions in the text. Apart from the average number of such edits within a text (Edits per text (global)), we also report the average number of edits per sentence, defined as the total number number of edits in the text normalized by the number of sentences in the text (Edits per sentence (global)), and the average number of edits of a specific sentence (Edits per sentence (local)). A visualization of the distribution of a representative measure from each group is presented in Figure 3. Arguably the best-known writing-fluency measure that can be derived from the measures in Table 2 is the number of words per minute (WPM) (see, e.g., Chenoweth and Hayes, 2001, Wolfe-Quintero et al., 1998). WPM is a product-based measure. From a process perspective, writing fluency can also be assessed in terms of the mean time between individual keystrokes, i.e. the mean IKI. Figure 4 presents information on the distribution of participant-averaged scores WPM and Mean IKI scores illustrating the between-subject variability in writing fluency in the dataset. The distribution of WPM scores was found to be right-skewed ($G = 1.07$). A Kolmogorov-Smirnov test for normality revealed that the distribution for writing fluency was non-normal ($d = 0.09, p < 0.0001$). A Spearman's rank correlation revealed a very high negative correlation between WPM and Mean IKI scores ($\rho = -0.92$, $p < 0.0001$). A visualization of the progression of inter-keystroke intervals in an example sentence from the dataset is presented in Figure 2.

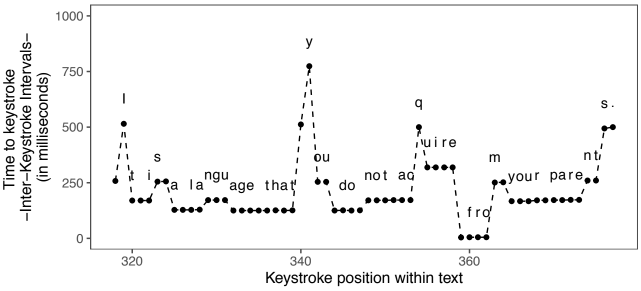

Figure 2: Unfolding of a single sentence. Inter-keystroke intervals (IKI), i.e. the time between two consecutive key-presses, across an example sentence.

Measure	M	SD
Production		
Words per text	530.41	280.95
Words per sentence	19.21	9.99
Keystrokes per text	4317.60	2379.09
Keystrokes per sentence	156.45	90.31
Deletions per sentence	16.16	18.96
Inter-keystroke interval	182.5	57.67
$\text{SPT}_{\text{total}}$ (sec)	76.26	54.06
$\text{SPT}_{\text{words}}$ (sec)	43.28	29.42
$\text{SPT}_{\text{non-words}}$ (sec)	21.67	18.84
SPT_{edit} (sec)	23.47	31.63
Pauses		
Pauses per sentence (510ms)	19.86	16.1
Pauses per sentence (2000ms)	5.72	5.23
Pauses per sentence (10000ms)	1.75	2.08
Revision		
Edits per text (global)	58.5	58.76
Edits per sentence (global)	2.30	1.57
Edits per sentence (local)	1.87	1.42

Table 2: Corpus-level descriptive statistics of keystroke measures/indicators of writing fluency. SPT = Sentence production time

[1] Based on the results of a principal components analysis of an original set of more than 200 viables gathered via InputLog, van Waes and Leijten (2015) proposed a fourth aspect, termed 'process variance', that concerns the variability of the production. This aspect is not covered in this study.

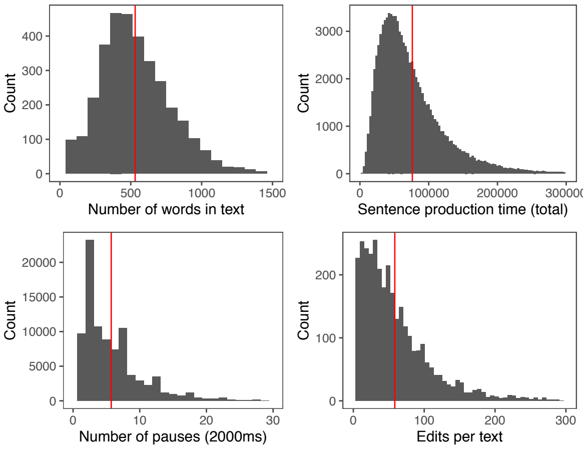

Figure 3: Distributions of the number of word in text (top left), total sentence production time (top right), number of pauses (2000ms) in text (bottom left) and number of edits per text (lower right). Red vertical lines represent the sample means of the respective measures.

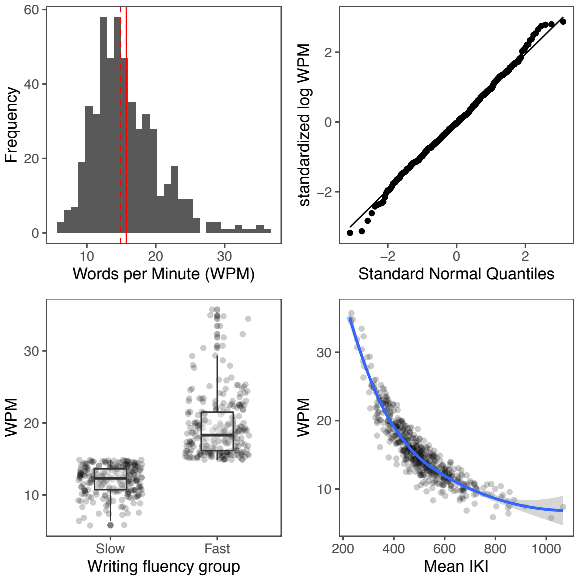

Figure 4: Visualization of the distribution of participant-averaged writing fluency scores (N=512): **(Top-left)** Histogram of writing fluency scores (Words per Minute). The Pearson coefficient of skewness G was 1.07, indicating that the distribution was right-skewed, such that the mean (solid line) was greater than the median (dashed line). **(Top-right)** Quantile-quantile plots of standardized log-transformed writing fluency scores (words per minute) against a standard normal distribution. A Kolmogorov-Smirnov test for normality revealed that the distribution for writing fluency was non-normal ($d = 0.09, p < 0.0001$) **(Bottom-left)** Distribution of WPM scores for slow and fast typist (grouping based on median-split; see Table 4). **(Bottom-right)** WPM scores against mean inter-keystroke intervals. A Spearman's rank correlation revealed a very high negative correlation between the two indices of writing fluency ($\rho = -0.92, p < 0.0001$)

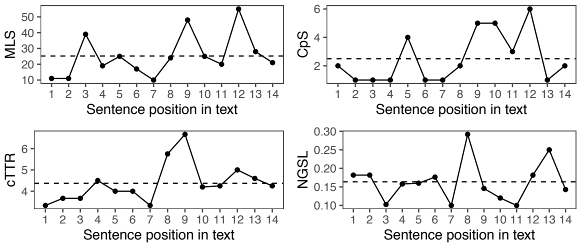

Figure 5: Progression of complexity across a random text from the present dataset for four selected measures: Mean Length of Sentence in words (MLS), Clauses per Sentence (CpS); corrected Type-Token Ratio (cTTR), and the number of words in sentence from the NGSL (NGLS). The plots show that - across measures - complexity is not uniformly distributed within a text but passes through a series of peaks and troughs.

3. Automatic assessment of linguistic complexity

The 3434 texts gathered were subsequently automatically analyzed with respect to their linguistic complexity using Complexity Contour Generator (CoCoGen), a computational tool that implements a sliding-window approach to track progression of complexity within a text (Ströbel, 2014; Ströbel et al., 2016). In this approach, a window of a user-defined size is moved across a text sentence-by-sentence, computing one complexity score per window for a given complexity measure (CM). The resulting series of measurements generated by CoCoGen captures the progression of linguistic complexity within a text for a given CM, which is referred to as a ‘complexity contour’. In its current version, CoCoGen supports a total of 107 measures of linguistic complexity for English. For the purposes of this paper, to illustrate the utility of combining keystroke logging with a high-resolution assessment of text complexity, we focussed on four widely used CMs that represent four distinct aspects of complexity (cf. Ströbel 2014 for details): (1) Mean Length of Sentence (MLS) (in words), a measure of length of production unit, (2) Clauses per Sentence (CpS), a measure of syntactic complexity, (3) the corrected type token ratio (cTTR), a measure of lexical diversity, and (4) number of words on the New General Service List (NGSL), a measure of lexical sophistication. An illustration of the derived complexity contours for a random text from the present data set for the four complexity measures is presented in Figure 5. The means and standard deviation for all measures at the corpus-level are presented in Table 3. Their distribution are shown in Figure 6.

4. Aligning Keystroke Logging with Linguistic Complexity

Our analysis of the relationship between keystroke measures and text complexity proceeded in two steps. In a first step, we aimed to determine whether individual differences in writing fluency are related to the linguistic complexity of the written products, and if so, whether the relationship can be observed for different aspects of complexity. To this end, participants were categorized as “slow” or “fast”

	Definition	M	SD
MLS	$\frac{N_{token}}{N_{sentence}}$	19.27	10.04
CpS	$\frac{N_{clause}}{N_{sentence}}$	1.9	1.25
cTTR	$\frac{N_{type}}{\text{floor}(\sqrt{N_{token}})}$	4.13	0.78
NGSL	$\frac{N_{token}-N_{NGSL}}{N_{token}}$	0.17	0.12

Table 3: Descriptive statistics (means and standard deviations) for the four complexity measures investigated. All scores based on measurements at the sentence level.

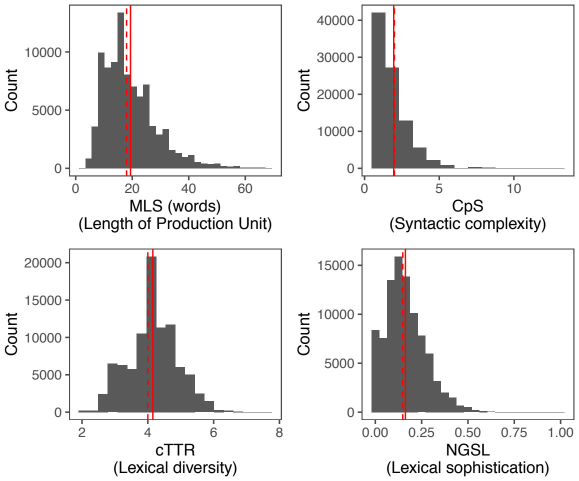

Figure 6: Distribution of sentence complexity scores (histograms) from four selected measures: Mean Length of Sentence in words (MLS), Clauses per Sentence (CpS); corrected Type-Token Ratio (cTTR), and the number of words in sentence from the NGSL (NGLS).

typists based on a median split on the WPM writing fluency variable and writing process scores (pause frequency and number of revisions per minute) and complexity scores (Mean Length of Sentence (MLS), Clauses per Sentence (CpS); corrected Type-Token Ratio (cTTR), words in sentence from the NGSL (NGLS)) were determined for each group. The means and standard deviations of these variables per group are presented in Table 4. A series of Wilcoxon Signed Rank tests were run to test for group mean differences in performance on these measures. It was found that fast typists produced on average about seven words per minutes more than slow typists ($V = 0, p(V) < 0.0001$) and exhibited inter-keystroke intervals that were about 175 ms shorter on average than those of their slow typing peers ($V = 32699, p(V) < 0.0001$). Fast typists also made 0.78 fewer pauses per minute ($V = 30011, p(V) < 0.0001$) and initiated roughly half as many text revisions per sentence than slow typists ($V = 29935, p(V) < 0.0001s$). The comparison of text complexity revealed that the written products of more fluent writers exhibited significantly higher syntactic complexity, as indicated by higher CpS scores ($V = 133318, p(V) = 0.01$), and higher degrees of lexical sophistication, as evinced by higher NGSL scores ($V = 19234, p(V) = 0.02$). The difference in lexical diversity between fast and slow typists, as measured by cTTR scores, was not significant ($V = 14426, p(V) = 0.09$). There were also no significant differences in average text length between the two groups ($V = 15207, p(V) = 0.29$) nor in mean sentence length ($V = 14795, p(V) = 0.16$). Although direct comparison is limited by the operationalization of linguistic complexity, the pattern of results is in line with the findings reported in previous studies indicating that fast typists produced texts of greater lexical diversity and lexical density than those of fast typists (Alves et al., 2008). The results are also consistent with the general finding in writing research that the text quality can suffer from difficulties in low-level motor execution (Connelly et al., 2005; Olive and Kellogg, 2002).

	Slow (N=256)		Fast (N=256)	
	M	SD	M	SD
Writing fluency				
WPM	12.06	1.97	19.42	4.16
Avg IKI	590.28	103.96	415.46	69.65
Avg text length	116.29	25.85	117.38	21.64
Pauses$_{2000\text{ms}}$/min	4.84	0.58	4.06	0.69
Edits				
Edits/text	83.43	57.41	41.87	35.99
Edits/sentence	3.08	1.65	1.51	1.01
Text complexity				
MLS	19.21	4.31	19.59	3.61
CpS	1.86	0.43	1.95	0.37
cTTR	4.12	0.29	4.15	0.24
NGSL	0.16	0.08	0.18	0.03

Table 4: Means and standard deviation of six indicators of writing fluency and four measures of sentence complexity for slow (left) and fast (writers).

In a second step, we fully exploited the high-resolution complexity measurement provided by CoCoGen by investigating the relationship between text production time and linguistic complexity at the sentence level. To determine how well particular aspects of the complexity of the sentences predict its production time, when controlling for all other predictors, we ran multiple linear mixed-effect regression models with by-subject adjustments to intercept and slopes. All models were specified with the maximal random-effects structure justified by the design (Barr et al., 2013) to account for individual differences in typing speed and to allow for the explanatory variables to have differential effects across participants. We first fitted a full model in which the behavioral response variable, log sentence production time, was regressed onto the main effects of (1) sentence length (MLS), (2) clauses per sentence (CpS), (3) corrected type token ratio (cTTR) and (4) words from the NGSL (NGSL), as well as all two-way interactions among the predictors. To reduce multicollinearity, all variables were standardized prior to being entered into the model. We then employed a stepwise bidirectional variable selection procedure based on Akaike's Information Criterion (AIC) to obtain the best-fitting (minimal adequate) model, i.e. only variables that decreased the AIC were retained. All models were fitted using the `lme4` package (Bates et al., 2015) in the statistical software `R` (R Core Team, 2017). Conditional and marginal coefficient of determination for Generalized mixed-effect models (R^2_{GLMM}) were computed using the `r.squaredGLMM`-

function from the R library MuMIn, which implements the method described in Nakagawa and Schielzeth (2013).

The best-fitting model included the main effects of all predictors as well as the three interactions terms involving the sentence length predictor (MLS). The conditional R^2_{GLMM}, i.e the variance explained by both fixed and random factors (the entire model), was found to be 65.9%. The marginal R^2_{GLMM}, i.e the variance explained by the fixed factors, was 46.6%. The regression coefficients (with 95% confidence intervals) of the best-fitting model (right) and a model that includes only the main effects of the complexity variables (left) are presented in Table 5. A visualization of the effects of the complexity predictors on sentence production times is provided in Figure 7. The results of the mixed-effects regression analysis revealed that, after controlling for sentence length, sentence production times were affected by both syntactic complexity (CpS) and lexical diversity (cTTR), and that the magnitude of the effects was mediated by the length of the sentence. Corroborating the findings obtained in step one, the results of the mixed-effects regression models indicate that increased cognitive effort associated with higher levels of linguistic complexity impacts writing fluency, supporting the claim made in previous research that drops in writing fluency may reflect increased cognitive processing load during writing resulting from the planning and execution of more complex lexical and syntactic structures (Schoonen et al., 2003).

	Dependent Variable: Log Sentence Production Time	
	Main effects model	Best-fitting model
(Intercept)	11.08^{***}	11.13^{***}
	(11.06, 11.11)	(11.11, 11.16)
Main effects		
MLS	0.41^{***}	0.46^{***}
	(0.40, 0.41)	(0.46, 0.47)
CpS	-0.02^{***}	-0.003
	$(-0.03, -0.02)$	$(-0.01, 0.001)$
NGSL	0.06^{***}	0.08^{***}
	(0.06, 0.07)	(0.07, 0.08)
CTTR	0.09^{***}	0.06^{***}
	(0.09, 0.10)	(0.06, 0.07)
Interactions		
MLS:CpS	–	-0.02^{***}
		$(-0.03, -0.02)$
MLS:NGSL	–	0.005^{**}
		(0.002, 0.01)
MLS:CTTR	–	-0.06^{***}
		$(-0.06, -0.06)$
Observations	94972	94972
AIC	98104.460	92457.650

Table 5: Regression coefficients (with 95% confidence intervals) of the best-fitting linear mixed-effects model (right) and a model that includes only the main effects of the complexity variables (left). Both models contained the maximal random effects structure, i.e. by-subject random intercepts and slopes for all predictors. (*p<0.05; **p<0.01; ***p<0.001).

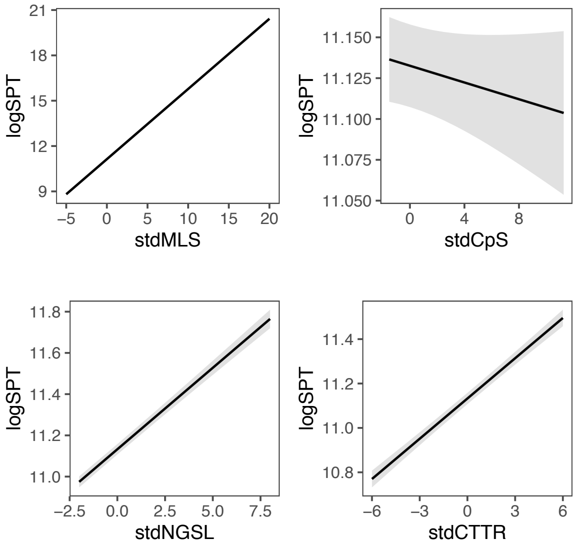

Figure 7: Marginal effects from best-fitting model of the four predictors. The plots indicate that, with the exception of CpS, all complexity measures exerted significant, independent effects on sentence production time, such that production time increased with increasing sentence complexity. Shaded areas represent 95% confidence intervals.

5. Conclusion and Future Directions

Recent years have witnessed increasing efforts to obtain ecologically valid data to advance our understanding of the mechanisms underlying reading and writing as well as the development of these two fundamental literacy skills. In this paper, we presented a new, large resource of keystroke logging that has the potential to inform theoretical models of second language (L2) writing and to evaluate the generalisability of such models. Here we demonstrated how the rich, behavioural information contained in keystroke logging data can be usefully combined with sentence-level assessments of text complexity to adequately align process and product analyses in L2 writing research.

In this paper, as a natural first step to demonstrate the utility of the new resource, we decided to focus our analysis on a small set of keystroke measures and complexity measures. In future work, we intend to include a wider range of indices of writing fluency from the basic units of keystroke logs (pauses, bursts and revisions). Regarding the assessment of the quality of the written product, we intend to incorporate a more comprehensive set of lexical and syntactic complexity measures, as well as recently introduced information-theoretic and n-gram measures. Another avenue of future research will pursue the potential of the dense longitudinal data as a window into understanding second language writing development. As specified in Section 2, the resource includes series of writing samples produced over a period of one semester (approx. 12-14 weeks), allowing for the study of developmental trajectories in second language writing through a growth curve modelling approach. The keystroke resource presented in this paper will be made publicly available for academic purposes upon signing an end user license agreement through the Open Science Framework (OSF) (https://osf.io/) and through the IRIS, a digital repository of data collection instruments for research into second language learning and teaching (https://www.iris-database.org/).

6. Bibliographical References

Alves, R. A., Castro, S. L., and Olive, T. (2008). Execution and pauses in writing narratives: Processing time, cognitive effort and typing skill. *International journal of psychology*, 43(6):969–979.

Baaijen, V. M., Galbraith, D., and De Glopper, K. (2012). Keystroke analysis: Reflections on procedures and measures. *Written Communication*, 29(3):246–277.

Chenoweth, N. A. and Hayes, J. R. (2001). Fluency in writing: Generating text in l1 and l2. *Written communication*, 18(1):80–98.

Connelly, V., Dockrell, J. E., and Barnett, J. (2005). The slow handwriting of undergraduate students constrains overall performance in exam essays. *Educational Psychology*, 25(1):99–107.

Cop, U., Dirix, N., Drieghe, D., and Duyck, W. (2017). Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49(2):602–615.

Latif, M. M. A. (2009). A state-of-the-art review of the real-time computer-aided study of the writing process. *International Journal of English Studies*, 8(1):29–50.

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

Leijten, M., Van Waes, L., and Ransdell, S. (2010). Correcting text production errors: Isolating the effects of writing mode from error span, input mode, and lexicality. *Written communication*, 27(2):189–227.

Luke, S. G. and Christianson, K. (2018). The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50(2):826–833.

MacArthur, C. A., Graham, S., and Fitzgerald, J. (2008). Handbook of writing research. Guilford Press.

Medimorec, S. and Risko, E. F. (2017). Pauses in written composition: on the importance of where writers pause. *Reading and Writing*, 30(6):1267–1285.

Nottbusch, G., Grimm, A., Weingarten, R., and Will, U. (2005). Syllabic sructures in typing: Evidence from deaf writers. *Reading and Writing*, 18(6):497–526.

Nottbusch, G. (2010). Grammatical planning, execution, and control in written sentence production. *Reading and Writing*, 23(7):777–801.

Olive, T. and Kellogg, R. T. (2002). Concurrent activation of high-and low-level production processes in written composition. *Memory & Cognition*, 30(4):594–600.

Quinlan, T., Loncke, M., Leijten, M., and Van Waes, L. (2012). Coordinating the cognitive processes of writing: The role of the monitor. *Written Communication*, 29(3):345–368.

Sahel, S., Nottbusch, G., Grimm, A., and Weingarten, R. (2008). Written production of german compounds: Effects of lexical frequency and semantic transparency. *Written Language & Literacy*, 11(2):211–227.

Schilperoord, J. (2001). On the cognitive status of pauses in discourse production. In *Contemporary tools and techniques for studying writing*. Springer, pp. 61–87.

Schoonen, R., Gelderen, A. v., Glopper, K. d., Hulstijn, J., Simis, A., Snellings, P., and Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language learning*, 53(1):165–202.

Spelman-Miller, K. (2006). Pausing, productivity and the processing of topic in online writing.

Ströbel, M., Kerz, E., Wiechmann, D., and Neumann, S. (2016). Cocogen-complexity contour generator: Automatic assessment of linguistic complexity using a sliding-window technique. In Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), pages 23–31.

Ströbel, M., Kerz, E., Wiechmann, D., and Qiao, Y. (2018). Text genre classification based on linguistic complexity contours using a recurrent neural network. In MRC@ IJCAI, pages 56–63.

Ströbel, M. (2014). Tracking complexity of l2 academic texts: A sliding-window approach. Master thesis. RWTH Aachen University.

Van Waes, L. and Leijten, M. (2015). Fluency in writing: A multidimensional perspective on writing fluency applied to l1 and l2. *Computers and Composition*, 38:79–95.

Van Waes, L. and Schellens, P. J. (2003). Writing profiles: The effect of the writing mode on pausing and revision patterns of experienced writers. *Journal of pragmatics*, 35(6):829–853.

Van Waes, L., Leijten, M., and Quinlan, T. (2010). Reading during sentence composing and error correction: A multilevel analysis of the influences of task complexity. *Reading and Writing*, 23(7):803–834.

Van Waes, L., Leijten, M., Wengelin, A., and Lindgren, E. (2012). Logging tools to study digital writing processes. *Past, present, and future contributions of cognitive writing research to cognitive psychology*, pages 507–533.

Wengelin, Å. (2006). Examining pauses in writing: Theory, methods and empirical data.

Wolfe-Quintero, K., Inagaki, S., and Kim, H.-Y. (1998). Second language development in writing: Measures of fluency, accuracy, and complexity. *University of Hawaii, Second language teaching and curriculum center*.

Zhang, M., Hao, J., Li, C., and Deane, P. (2016). Classification of writing patterns using keystroke logs. In *Quantitative psychology research*. Springer, pp. 299–314.