

**Ideas cascading into keystrokes – modelling writing hesitations as finite mixture
process**

Jens Roeser¹, Mark Torrance¹, Rianne Conijn², Gunn Helen Ofstad³, and & Evgeny
Chukharev⁴

¹ Department of Psychology
Nottingham Trent University
United Kingdom

² Artificial Intelligence Systems Institute
Eindhoven University of Technology
The Netherlands

³ Educational Sciences and Humanities
University of Stavanger
Norway

⁴ Department of English
Iowa State University
Iowa

Author Note

Correspondence concerning this article should be addressed to Jens Roeser, 50 Shakespeare St, Nottingham NG1 4FQ. E-mail: jens.roeser@ntu.ac.uk

Abstract

Classical serial models view the process of producing a text as a chain of discrete pauses and writing bursts. In contrast, parallel cascading models of writing assume that planning is not complete at production onset and operates in parallel with execution. These two theories make different assumptions about the underlying process that generates inter-keystroke intervals. Across six datasets comprising multi-sentence text composition we instantiated these assumptions as two sets of Bayesian mixed-effects models: (1) single-distribution models as assumed by the serial processing account, and (2) finite two-distribution mixture models as assumed by the parallel processing account. We analysed keystroke intervals at before-sentence, before word, and within word transitions. Model comparisons demonstrated strong evidence in favour of the statistical implementation of the cascading view across all datasets. When pausing occurred, sentence initial pauses were longer than word initial pauses which is consistent with the idea that larger linguistic edges are associated with higher level planning. However, we found – across populations – that keystroke latencies at word and even at sentence boundaries were often too rapid to permit planning of what was written next. Our results cannot be explained by the serial but are in line with the cascading, semi-parallel view of writing.

Keywords: Cascading semi-parallel models; serial models; finite mixture models; Bayesian models; text composition; keystrokes

Ideas cascading into keystrokes – modelling writing hesitations as finite mixture process

Translating ideas into language involves a sequence of processes starting with communicative intent, proceeding through generation of a message, retrieving appropriate lexical items, deciding with which part of the message to start the utterance which, depending on theoretical viewpoint (Bock & Ferreira, 2014; Wheeldon & Konopka, 2023), is determined by syntactic requirements of our language, then retrieving orthographic and / or phonological codes, and that are finally submitted to the motor programs that output speech, key or pen-strokes or signs (Olive, 2014; Van Galen, 1991; Wheeldon & Konopka, 2023).

The coordination of these processes in writing is arguably more demanding than in spoken language. This is because writing requires spelling retrieval which may or may not be mediated by phonology, and the motor planning and execution required to press keys or move a pen are later-developing and more effortful than those required to articulate language sounds in speech leading to a reduced output speed compared to speech (see for a recent review Kandel, 2023). Consequently when writing, more than in speech, we have to mentally buffer linguistic plans before we can express them in language. Written communication is also less time constrained: speaking underlies fluency requirements, because hesitating during speech has a communicative effect (e.g. Clark & Fox Tree, 2002). There is therefore stronger pressure to plan ahead in speech than in writing (Roeser et al., 2019; Torrance & Nottbusch, 2012). Indeed writers can pause when they wish – even in the middle of a word – and for how long they want (within reason) without compromising communication. Written text production – for example writing argumentative texts – is therefore a combination for relatively fluent production bursts followed by pauses (Chenoweth & Hayes, 2001).

The study reported in this paper compares statistical models that instantiate two different theories of how cognitive planning processes are coordinated in writing. Through the analysis of 6 different datasets from writers of different ages and competencies composing

multi-sentence texts, we compare these statistical models and therefore the cognitive model of how keystrokes in writing are being generated they are mapping on. Also, we evaluate competing predictions of these models by revisiting the often-reported finding that edges of larger linguistic units are associated with longer pauses (Conijn et al., 2019; Flower & Hayes, 1981; Matsushashi, 1981; Wengelin, 2006). In what follows we first summarise previous research that has explored pause patterns in spontaneous text production. We then introduce two possible cognitive frameworks in which these can be interpreted. Finally we argue that analysis of pause-patterns at linguistic edges are important to permit differentiation between these two theories.

We focus on data from spontaneous text production (see Gernsbacher & Givón, 1995) for the following three reasons: First, text production such as essay, argumentative, and narrative writing in response to a writing brief or topic statement is a natural environment for text production and have, therefore, real-world relevance (Dux Speltz et al., 2022; Dux Speltz & Chukharev-Hudilainen, 2021; Vandermeulen et al., 2023). Second, since the emergence of keystroke-logging software such as InputLog (Leijten & Van Waes, 2013) recording keystroke data during free text composition has become widely used among writing researchers and led to a rapidly expanding timecourse literature with a focus on the orchestration of subprocesses; for a recent special issue see Conijn and Torrance (2023).¹ These data are therefore important for contemporary theories of writing and freely available for statistical modelling and machine learning (Conijn et al., 2019; Conijn, 2020). Third, spontaneous text production – unlike picture description experiments (Damian & Stadthagen-Gonzalez, 2009; Meulemans et al., 2022; Nottbusch, 2010; Roeser et al., 2019) or copy tasks (Roeser et al., 2021; Van Waes et al., 2019, 2021) – content generation is not artificially constrained to individual sentence units (or less) by one or more images and

¹ Social Science Citation Index reports 40 journal papers that describe research exploring composition processes using keystroke logging methods in 2022, compared to 26 in 2017 to 2019, and 7 in 2014 to 2016.

linguistic form is not constrained to a sequence of words, respectively. Instead text composition provides data from both inter and intra-sentence transitions.²

Time intervals between adjacent keystrokes during typing vary as a function of various factors. Explaining the factors that influence the duration of inter-keystroke intervals (IKIs) requires a theory of how the various processes that transform intent into keypresses are structured. Processing – planning what to say and how to say it – may be serial. This means that the processing sequence operates one unit at a time. Processing of the next language unit can only start when the previous unit was fully processed from communicative intention to completion of the written output. Therefore longer IKIs represent periods when the writer is planning what to write next. The subsequent burst of fluent output is then the result this planning (Alves & Limpo, 2015; Hayes, 2012; Kaufer et al., 1986; Matsushashi, 1981; Schilperoord, 2002). This is particularly reflected in the model proposed by Flower and Hayes (1980) that characterises the writing process as a sequence of generating ideas, planning the text, translating the ideas into language, executing hand-writing or typing, and revising the texts (also Chenoweth & Hayes, 2001, 2003).

This serial account is consistent with the frequent finding that average IKI duration immediately before larger linguistic units compared to smaller linguistic units. For example, in spontaneous text the mean pause duration before a writer starts a sentence has typically been reported to be longer than before a mid-sentence word and these are, in turn, longer than between mid-word key presses (Deane & Zhang, 2015; Mohsen, 2021; Spelman Miller, 2006; Spelman Miller et al., 2008; Torrance, Rønneberg, et al., 2016; Wengelin, 2002; Xu & Qi, 2017). This is understood as being due to time needed to plan message and syntax for the next sentence, which is added to the time required to prepare the upcoming word, and the motor planning required to produce the first keystroke of that word (Baaijen et al., 2012;

² Also, pauses in text production do not reflect visual encoding of the stimulus (although possibly reflect looking back during writing which we will return to in the discussion).

Medimorec et al., 2017; Medimorec & Risko, 2017; Roeser et al., 2019). However this account is not consistent with what we know about writing. Consider these two observations:

First, writers and speakers can in principle plan utterances in full for one or multiple sentences before beginning to output words. However this is not what people typically do: utterances are not fully formed at production onset. Instead, syntax and lexical content and even details of the message itself are planned as the emergent result of a real-time process. For speech it is well known that the pre-planning duration depends – among other factors – on the need to avoid intra-sentence hesitation that are needed to plan upcoming words in parallel to production (e.g. Levelt & Meyer, 2000; Wheeldon & Konopka, 2023). For writing, there is evidence that the time to keystroke / pen onset increases when a sentence starts with a more complex sentence-initial phrases similar as in speech [Nottbusch (2010); dam09; Roeser et al. (2019)]. For example, in Roeser et al. (2019) participants described moving arrays of images in simple utterances such as *A and the B moved above the C* that either started with a conjoined noun phrase (i.e. *A and the B moved*) or a simple noun (i.e., *A moved above the B and the C*). Importantly while the overall complexity of the utterance (in terms of length, number of phrases and noun) was held constant, sentences that started with a conjoined noun phrase increased the time to typing onset (and speech onset). This is not what one would expect if writers plan sentences in full before production onset. In fact, eye-movement data reported in Torrance and Nottbusch (2012) and Roeser et al. (2019) demonstrate that writers do not plan the lexical form of the utterance beyond the first noun before production onset (similar to speech Griffin, 2001, 2003), although there was evidence that lexical pre-planning of the second noun (i.e. *the B*) was more likely for conjoined noun phrases, i.e. when the second noun was part of the same phrase as the first noun.

Second, even though the average sentence-initial IKI tends to be longer, they still tend to be very rapid. For example, Medimorec and Risko (2017) found that, in undergraduate students' writing on familiar topics 71% of sentence-initial IKIs were less than

1 seconds in duration. Also Rønneberg et al. (2022) reported that sentence-initial hesitations are rare, with over 50% of sentences preceded by very short pauses (around 430 msec) and a mean of 1.2 seconds for the remainder of pauses. For comparison, these short pauses are less than mean written naming response latency for single objects (Torrance et al., 2018) in a similar population. Longer pauses in the remainder were similar to response latencies for short sentences when describing arrays of images (Roeser et al., 2019), also in a similar population. Despite the fact that writers can in principle pause when they want to, written composition is often remarkably fluent. The fact that these very short sentence initial pauses exist, and in fact are quite common, is not consistent with the assumption that planning of the upcoming sentence is started and finished in the period between completion of the previous sentence and starting to write the next.

If utterances are not planned in full prior to writing onset, how is it possible that at least for reasonably competent writers producing composition often occurs remarkably fluently, with very few hesitations. The two examples above and similar findings point towards much of the mental activity associated with composition, including the relatively complex processing required to plan sentences, occurring as a result of a cascade of processes that operate partly in parallel and largely without executive control. Consistent with general trends in language processing theory (Bock & Ferreira, 2014; Chang et al., 2006; Dell et al., 2007), several researchers have argued that the processes associated with written production run at least in part in parallel (Bonin et al., 2012; Crump & Logan, 2010; Olive, 2014; Roux et al., 2013; Van Galen, 1991). Van Galen (1991), for example, argued for a cascade of modular processes, each responsible for a specific transformation (semantic, syntactic, and so forth), with processing occurring as soon as information from the immediately upstream process becomes available (Christiansen & Chater, 2016). Buffers provide transient storage at each processing level to accommodate unsynchronised output rates: when lower level processes lag behind higher level processes, buffers allow our fingers to catch up with thought and language related processes. In other words, when writers move from one sentence to

another without hesitation this is due to planning the next sentence, to some extent, while completing the output of the previous sentence.

This cascaded account of the composition process gives a rather different understanding of inter-keystroke intervals. IKIs result from one of two underlying mental processes. When upstream processes output at a rate equal to or faster than can be used for finger movement planning, then IKIs are determined just by the time needed for executing finger movements (i.e. just by processing below the last buffer in the processing cascade). However if one or more upstream processes provide output more slowly, then IKIs become dependent directly on upstream processing times and not on finger movement. IKIs therefore form two distributions, one associated with rapid, fluent output, and another that forms as a result of delays caused by some combination of processing at the semantic, syntactic, lexical, or orthographic levels (Roeser et al., 2021). The cascading account provides an explanation for why writers sometimes pause before onsetting a sentence (or word) but often are not. This is not because in some cases sentences are not planned, or planning was postponed until after sentence onset, but that this planning was completed in parallel with previous output.

In language comprehension, Vasishth and colleagues (Vasishth, Jäger, et al., 2017; Vasishth, Chopin, et al., 2017) demonstrated convincingly that some phenomena in online sentence comprehension observed in reading times are captured better by statistical models that assume that the underlying mental process that generates reading data is in fact a mixture of two process expressed as two distributions in the statistical implemental. Similar for keystroke data Roeser et al. (2021) demonstrated – using a similar approach as used in the present paper³ – how models that assume this mixture of two underlying data generating processes revealed two separate distributions when participants were copy-typing sentences or lists of words that are more pronounced for tasks with challenging orthographic sequences

³ For similar approaches see Almond et al. (2012), Baaijen et al. (2012), Chenu et al. (2014), Guo et al. (2018), Hall et al. (2022), Li (2021), and Van Waes et al. (2021).

but reduced down to just one distribution in a tapping task where participants merely pressed two keys repeatedly as quickly as possible (e.g. “dkdkdkdk”). As argued by the authors, copy-typing underlies a similar cascade of processes as free text composition. However composition tasks are different from copy tasks: in copy tasks but not in spontaneous text production, content generation and grammatical encoding is artificially constrained to particular sequences of letter bigrams, words or sentences, and copy-typing involves a combination of visual encoding of the target string and mental buffering prior to the activation of motor codes. Typing hesitations in the context of copy-typing may – to some extent – reflect not just production but also, and possibly largely, visual encoding. Text composition, in contrast, provides data from the free generation of words and full sentences that is not dependent on a visually presented text stimulus. For these reasons our research extends the analysis of keystroke data using a mixture model approach to data from spontaneous text production.

In the study that is presented in this paper we test using Bayesian hierarchical models of keystroke intervals – similar to Roeser et al. (2021) – whether the cascading view of writing generalises to contexts in which participants compose texts. On the basis of these statistical models of text composition, we appraise to what extent linguistic edges (sentence and word boundaries) are associated with different pausing behaviours as frequently suggested in the literature (Ailhaud et al., 2016; Ailhaud & Chenu, 2017; Chukharev-Hudilainen et al., 2019; Conijn et al., 2019; Mohsen, 2021; Torrance, Rønneberg, et al., 2016; Wengelin, 2002).

The hypotheses are as follows: if the preparation of the upcoming production unit happens entirely at the corresponding linguistic edge as predicted by the serial account, and not in parallel to production, key-transition intervals can statistically be modelled as a function of the associated transition location. In other words, we expect IKIs to be proportionally longer at before-sentence locations compared to before-word locations

compared to within-word locations. This is because transitions before larger linguistic units are associated with processing that involves higher levels of representation. The cascading view, in contrast, assumes that although IKIs at larger linguistic edges might be longer, it is in principle possible that writers do not (always) pause to plan their text but plan in parallel to writing. In other words, statistical models of the pause-and-burst view of written production do not capture that planning can operate, to some extent, in parallel to the output of the written production. Therefore, a statistical model that captures writing as a cascade of processes must account for the possibility that hesitations occur to some extent probabilistically across the entire text although some linguistic locations such as larger linguistic edges are associated with a higher pausing probability. We hypothesise that the statistical model of the parallel cascading view provides a better out-of-sample generalisation than the implementation of the serial model of writing in the context of unconstrained spontaneous text composition.

Methods

In a series of five Bayesian models, we (re)analysed keystroke data from six experiments in which participants composed texts. Full modelling details can be found in Appendix \ref{statistical-models}; we summarise the most important model properties in the following. The first three models map onto the serial account of writing hesitations and bridge between commonly used statistical models for the analysis of inter-keystroke intervals (IKI) as single distribution models and a principled single-distribution model of keystroke data. In particular we used a Gaussian distribution, and a log-Gaussian model, similar to statistical models that can frequently be found in the literature, and an unequal variance model that takes into consideration that intervals associated with larger linguistic edges may have a larger variability.

The remaining two models are implementations of the cascading view following the two-distributions mixture-model approach presented in Roeser et al. (2021). Importantly

these two models assume that IKIs result from a mixture of two processes: (1) uninhibited activation flow into motor programmes and (2) interruptions at higher levels cause delays in the information flow and therefore lead to longer lags between keystrokes. These two models assume that hesitations are not sufficiently determined by transition location – as in the serial account – but transition locations are also associated with different probabilities that a hesitation occurs. This view was implemented as a constrained model that assumes no keystroke-interval difference between transition locations for uninhibited activation flow and an unconstrained model which assumes that keystroke-transitions may vary between transition locations even when information flow from higher to lower levels is uninterrupted.

We reanalysed datasets from six experiments in which researchers elicited full texts in response to writing prompts. We discussed advantages with this approach in the introduction section. These datasets include samples from populations with various writing experience and languages (e.g. young / L2 writers, students) performing different writing tasks (e.g. essays, syntheses). Datasets from a variety of different populations of writers and writing tasks were deliberately chosen to challenge our modelling approach and to test to what extent pausing patterns generalise across writing contexts.

Data sets

TODO: Mark, can you check / add information for C2L1, CATO, GUNNEXP2. For GUNNEXP2 I’ve added a provisional citation that probably needs changing; we could just say “unpublished”. I could also just use the Escop presentation as reference (also for SPL2; maybe my SIG Writing talk). Was the GUNNEXP2 sample Norwegian? Should we include information like time limit or text length in the table?

Six datasets with keystroke data from free text production were used for analysis. An overview can be found in Table 1; descriptions below. Dataset used are C2L1 (Rønneberg et al., 2022), CATO (Torrance, Rønneberg, et al., 2016), GUNNEXP2 (Torrance & Ofstad, n.d.), SPL2 (Torrance et al., n.d.), LIFT published in Vandermeulen, Steendam, et al. (2020)

and described in Vandermeulen, De Maeyer, et al. (2020), PPlanTra data are published in Rossetti and Van Waes (2022b) and described in Rossetti and Van Waes (2022a). In C2L1 Norwegian 6th graders composed argumentative essays. In CATO upper secondary students in Norway with weak decoding skills (and control) composed expository texts either normally or with masked letters to prevent reading their unfolding text. In SPL2 undergraduate students produced argumentative essays in their L1 (English) or L2 (Spanish); order of language and two writing prompts were counterbalanced. The LIFT data contains pre-university students producing argumentative and informative text syntheses on four topics each. The PPlanTra data contains data from Master students in Business and Economics simplifying texts on sustainability before and after received either online instruction on how to apply plain language principles to sustainability content or a online instruction exclusively on the topic of sustainability. The GUNNEXP2 data [MARK, please add a sentence here to describe the production context].

For C2L1, CATO, and GUNNEXP2, keystroke data were captured using EyeWrite (Simpson & Torrance, 2007; Torrance, 2012). LIFT and PPlanTra data were captured using InputLog (Leijten & Van Waes, 2013; Van Waes et al., 2019, 2021) and SPL2 data were collected using CyWrite (Chukharev-Hudilainen et al., 2019).

Table 1*Datasets in brief.*

Dataset	Source	Keylogger	Writing task	N	Mean age	Sample	Country	Language
C2L1	Rønneberg et al. (2022)	EyeWrite	Argumentative essays	126	11.8	6th graders	Norway	Norwegian
CATO	Torrance et al. (2016)	EyeWrite	Expository texts	52	16.9	Secondary school students	Norway	Norwegian
GUNNEXP2	Torrance and Ofstad (n.d)	EyeWrite	?????	45	????	Undergraduate students	???? Norway	???? Norwegian
LIFT	Vandermeulen, Steendam, et al. (2020)	InputLog	Synthesis	658	17	Secondary school students	The Netherlands	Dutch
PLanTra	Rossetti and Van Waes (2022b)	InputLog	Text simplification	47	23	Master students	Belgium	English (L2)
SPL2	Torrance et al. (n.d.)	CyWrite	Argumentative essays	39	20.6	Undergraduate students	USA	English (L1) / Spanish (L2)

Table 2*Transition location classification.*

Transition type	Description	Example
Within word	Transitions between any letter	T [^] h [^] e c [^] a [^] t m [^] e [^] o [^] w [^] e [^] d. T [^] h [^] a [^] t[bsp][bsp]e [^] n i [^] t s [^] l [^] e [^] p [^] t.
Below word	Keypress after space followed by any letter	The [^] cat [^] meowed. That[bsp][bsp]en [^] it [^] slept.
Before sentence	Keypress following a space preceding any letter	The cat meowed. [^] That[bsp][bsp]en it slept.

Note. [^] marks transition location; [bsp] represents backspace.

Data extraction

From the keystroke data we extracted transition durations between adjacent keys at location that were found, by previous research, to be psycholinguistically meaningful (Chukharev-Hudilainen et al., 2019; De Smet et al., 2018; Torrance, Rønneberg, et al., 2016) and are detailed in Table 2. In particular we analysed the key-transitions that resulted in the insertion of a character that started a new sentence as before-sentence transition; transitions that started a new word other than those at the beginning of a sentence were treated as before-word transition; all transitions within a word but not the key-transition between the last letter of a word and the subsequent space or punctuation mark were treated as within-word transitions. At before-sentence locations, IKIs were timed to the shift keypress that resulted in the capitalization of the first key for most data sets (CATO, C2L1, SPL2, GUNNEXP2) but included the transition to the following sentence-initial letters in some data sets (PLanTra, LIFT); we return to this difference in the Results section. Transitions that occurred at the beginning of the text or the beginning of a paragraph were not treated as before-sentence transitions and were removed from the analysis. Also we removed transitions that were followed by an editing operation.

We removed participants that did not complete all conditions in studies with within-participant factors (reducing the number of participants to 343 for LIFT data, and 41 participants for PPlanTra data). We removed participants that produced less than 10 sentences (LIFT: 109 participants; PPlanTra: 3 participants; SPL2: 1 participant). We further removed keystroke intervals that were too short to represent a behavioural response (≤ 50 msecs) or too long (≥ 30 secs) and possibly involving other task-unrelated activities; percentages can be found in Table 3. From the remaining data we randomly sampled a maximum of 100 observations per participant, condition, and transition location (when more than 100 were available). This was done to reduce the computation time of the Bayesian models require to complete sampling. For the LIFT data set we reduced the number of participants to 100 because the total sample was substantially larger than for the other datasets. Because the LIFT data set included the largest number of writing tasks, we sampled 50 observations per condition, location and participant which would otherwise exceed the computational resources available to us. The number of keystroke data used in the analysis can be found in Table 3.

Bayesian statistical modelling

We reanalysed keystroke data – transition durations measured as the time between first and second keypress – from six datasets in a series of five Bayesian models. An overview of all models can be found in Table 4. More detailed model descriptions and motivation can be found in Appendix A.

Models M1, M2, and M3 are single distribution models – consistent with the serial account – include fixed effects for each combination of transition location and dataset-specific manipulations. In particular models M1 and M2 are consistent with standard models used in the literature and therefore serve as baseline models. For model M3 we relaxed the equal-variance assumption for transition locations, thus allowing transition locations on larger linguistic boundaries to assume a larger standard deviation; this is because slower

human behaviour is generally known to be associated with a larger variability (Schöner, 2002; Wagenmakers & Brown, 2007; Wing & Kristofferson, 1973); see Appendix A for rationale.

Models M4 and M5 are extensions of the unequal-variance single distribution model M3. Importantly models M4 and M5 assume that keystroke data result from a combination of two data generating processes – consistent with the cascading account – rather than one. These model are called finite mixture models in the literature (Gelman et al., 2014; MacLahlan & Peel, 2000; Roeser et al., 2021) and capture the assumption that processing at higher levels of activation leads to longer pauses which can in principle happen at transition before sentence, before words or even mid-word. In other words, instead of assuming that there is one process that shifts the distribution of transition durations for larger linguistic edges, we allow for the possibility that pauses at larger linguistic edges are more likely but not obligated by the cognitive system that generates keystrokes. This is achieved by

Table 3

Data reduction. Mean percentage of extreme data removed and the mean percentage of randomly sampled data included by transition location (‘–’ if all data were used). Standard error in parentheses.

Dataset	Extreme values in %		Randomly sampled data in %		
	≤ 50 msecs	≥ 30 secs	within word	before word	before sentence
C2L1	0.19 (0.1)	0.074 (0.063)	35 (2.6)	85 (1.8)	–
CATO	0.65 (0.15)	0.023 (0.016)	15 (0.95)	49 (2.2)	–
GUNNEXP2	2.2 (0.17)	0.0072 (0.0055)	6.2 (0.43)	23 (1.4)	–
LIFT	2.6 (0.16)	0.0048 (0.0038)	3.2 (0.22)	13 (0.86)	99 (0.15)
PLanTra	2.5 (0.41)	0.04 (0.031)	9.7 (0.64)	37 (1.9)	–
SPL2	2.3 (0.2)	0.028 (0.016)	5.7 (0.39)	23 (1.4)	–

Table 4

Overview of cognitive models of writing. All models included by-participant random effect.

Models	Description
Serial	
M1	Single Gaussian model with effects of transition location (and experiment-specific manipulation).
M2	Single distribution log-Gaussian model equivalent to M1.
M3	Equivalent to M2 but with different variance components for each transition location (unequal variance).
Cascading	
M4	Two-distributions mixture of a log-Gaussian for fluent transition intervals and another log-Gaussian that is wider than the previous distribution hesitant / disfluent key transitions; the distribution of hesitant keystroke intervals assumes – similar to M3 – different values for transition location but the distribution of fluent intervals has the same mean across all design cell. Mixing proportions capture the relative number of disfluent transitions for the distribution of hesitant transitions for each combination of transition location.
M5	Equivalent to M4 but distribution of fluent transitions is allowed to vary by transition location.

modelling key transitions as coming from a weighted mixture of two distributions associated with two different states, illustrated in equation 1:

1. Information from upstream mental processes can flow into keystrokes without interruption at intermediate levels. These fluent keystroke transitions are merely constrained by a person's ability to move their finger. The population estimate for fluent key transitions (i.e. typing speed) is captured by the β parameter.
2. Any interruptions at upstream levels of mental representation result in a delay in the information flow and therefore create a lag between keystroke intervals, for example when words or their spelling could not be retrieved in time. The population estimate of the hesitation slowdown is captured by δ and its frequency is captured by the mixing weight (proportion) θ .

$$\begin{aligned}
 \text{iki}_i &\sim \theta_{\text{location}[i], \text{participant}[i]} \cdot \text{LogN}(\beta + \delta_{\text{location}[i]} + u_{\text{participant}[i]}, \sigma_{e'_{\text{location}[i]}}^2) + \\
 &\quad (1 - \theta_{\text{location}[i], \text{participant}[i]}) \cdot \text{LogN}(\beta + u_{\text{participant}[i]}, \sigma_{e_{\text{location}[i]}}^2) \\
 \text{where: } u_{\text{participant}} &\sim \text{N}(0, \sigma_p^2) \\
 \text{constraint: } \delta, \sigma_e^2, \sigma_{e'}^2, \sigma_p^2 &> 0 \\
 \sigma_{e'}^2 &> \sigma_e^2 \\
 0 < \theta &< 1
 \end{aligned} \tag{1}$$

The first line of equation 1 represents the distribution of hesitant key transitions, and the second line represents the distribution of fluent – uninterrupted – key transitions. Each of these two distributions is associated with the mixing weight θ which is a proportion that is constrained to be larger than 0 and smaller than 1. θ is here parameterised to represent the probability that a key transition is associated with the distribution of hesitant transitions. This probability is inversely related to the mixing weight of the distribution of

fluent transitions by $1 - \theta$. In other words, a larger weight for either distribution inevitably means a lower weight for the other. We call this parameter the probability of hesitant transitions. When a parameter was assumed to vary by transition location (levels: before sentence, before word, within word) this was indicated as subscript; similar for participants. The hesitation probability was allowed to vary by participants constrained by a hyper-parameter for each transition location.

The constrained implementation M4 assume that the size of a fluent key-transition does not vary by transition location. In other words, the value that the β parameter is taking on is the same for transitions at before-sentence, before-word, and within-word locations. However, letter bigrams may be executed systematically faster than transitions between space and either a subsequent letter – as in before-word transitions – or shift – at before sentence locations. This is for two reasons: first, and importantly, highly frequent bi- and even trigrams are stored, retrieved and executed as one motor code (Behmer Jr. & Crump, 2016; Grudin & Larochelle, 1982; Ostry, 1983; Terzuolo & Viviani, 1980) because of the strong association between letters. This association is reducing the length of transition durations. It is less likely that the same association exists for space-letter and space-shift pairs. Second, before-sentence transition in two datasets (PLanTra, LIFT) are the sum of the transitions between space, shift and the following character key and therefore necessarily longer than other transition; we return to this in the results section. We therefore implemented an unconstrained mixture model M5 that assumes that the distribution over durations of fluent transitions – represented as β – may vary across transition locations.

All models were implemented in the Bayesian framework (Gelman et al., 2014; McElreath, 2016). The Bayesian framework is ideal for the estimation of parameter values. This is because Bayesian parameter estimates are expressed as probability distributions that capture the associated uncertainty (Farrell & Lewandowsky, 2018; Gelman et al., 2014; Lee & Wagenmakers, 2014). To achieve this, Bayesian models require the inclusion of prior

information, i.e. existing knowledge about parameter values. For small datasets priors influence the inferred parameter values; for larger datasets weakly informative and vague priors are quickly overcome by the data (i.e. automatic Ockham’s razor, Jefferys & Berger, 1992). In other words the choice of priors has less impact on the posterior, certainly for weakly informative priors. In the present analysis, we used weakly informative priors to aid model convergence by constraining the parameter space to plausible values (see e.g. Lambert, 2018; McElreath, 2016). Also, as the sample size of most of our datasets is large, weakly informative priors have no or a negligible effect on the posterior.

Stan code for mixture models was based on (Roeser et al., 2021; see also Vasishth, Chopin, et al., 2017; Vasishth, Jäger, et al., 2017) and can be found on OSF (ADD URL HERE); for a tutorial see <https://rpubs.com/jensroes/1000459>.⁴

Results

Model comparisons

To compare the out-of-sample predictive performance of our models we used leave-one-out cross-validation based on Pareto smoothed importance-sampling (Vehtari et al., 2015, 2017). Predictive performance was estimated as the sum of the expected log predictive density (\widehat{elpd}) and compared by the difference between models $\Delta\widehat{elpd}$. We also summarised this difference as normalised over its standard error $|\frac{\Delta\widehat{elpd}}{SE}|$, hence the z -score of the difference between models (Sivula et al., 2020). Similar to other cross-validation techniques, the advantage of leave-one-out cross-validation is that more complex models – models with more parameters – are penalised to prevent overfit.

⁴ The R (R Core Team, 2020) package `rstan` (Stan Development Team, 2018) was used to interface with the probabilistic programming language Stan (Carpenter et al., 2016) which was used to implement all models. Models were run with 20,000 iterations on 3 chains with a warm-up of 10,000 iterations and no thinning. Model convergence was confirmed by the Rubin-Gelman statistic ($\hat{R} = 1$) (Gelman & Rubin, 1992) and inspection of the Markov chain Monte Carlo chains. The predictive performance of our models was compared using leave-one-out cross-validation (Sivula et al., 2020; Vehtari et al., 2015, 2017).

Model differences are shown in Table 5. All datasets revealed the same pattern. Importantly, the constrained two-distributions mixture model rendered a higher predictive performance compared to the unequal variance single-distribution model; the difference $\widehat{\Delta elpd}$ between models ranged within $[18, 40]$ standard errors, indicating a substantially higher predictive performance for the two-distributions model. The unconstrained mixture model M5 rendered a higher predictive performance than the constrained model M4 which did not allow the distribution of fluent keystroke-intervals to vary by transition location; the difference in predictive performance was between $[7, 15]$ standard errors higher for the unconstrained model. Among the single distribution models we found higher predictive performance for the unequal variance model compared to the log-Gaussian model; the difference $\widehat{\Delta elpd}$ ranged within $[43, 68]$ standard errors. The weakest model was the single distribution Gaussian which had – compared the single distribution log-Gaussian – a lower predictive performance in a range of $[43, 69]$ standard errors.

Table 5

Model comparisons. Models were compared incrementally from the simplest to the most complex model. The comparison is shown along with the distribution type of the more complex model. A negative difference in $\Delta\widehat{elpd}$ indicates higher predictive performance for the more complex model. Standard error is shown in parentheses.

Comparison	Distribution	C2L1	CATO	GUNNEXP2	LIFT	PLanTra	SPL2
M4 – M5	Mixture of two log-Gaussians (unconstrained; M5)	-546 (51)	-607 (43)	-554 (38)	-500 (37)	-130 (18)	-564 (41)
M3 – M4	Mixture of two log-Gaussians (constrained; M4)	-1,637 (72)	-1,782 (75)	-2,427 (92)	-5,105 (127)	-2,384 (90)	-1,181 (64)
M2 – M3	Single log-Gaussian (unequal variance; M3)	-38,301 (823)	-38,972 (913)	-41,155 (653)	-76,961 (1,656)	-35,204 (556)	-30,430 (450)
M1 – M2	Single log-Gaussian (M2)	-37,517 (798)	-37,673 (881)	-38,843 (616)	-74,851 (1,624)	-33,823 (526)	-28,717 (418)

Note. $\Delta\widehat{elpd}$ = difference in predictive performance – estimated as expected log pointwise predictive density

We also tested to what extent model predictions fit observed data. Comparisons can be found in Appendix B and echo the findings in Table 5. Data predicted by the two-distribution mixture models are similar to observed data; data predicted by single distribution Gaussian models showed a poor fit to the data.

Transition location effect

As demonstrated in the model comparisons, the two-distributions mixture-models is a better statistical representation on the underlying process generating keystroke data than compared to any of the single-distribution models. Information about this process is captured by the posterior estimates of the models' parameter values. To reiterate, there are three conceptually important parameters: (1) the magnitude of the slowdown for transition hesitations (represented as δ in equation 1), (2) the probability of hesitant transitions (represented as θ in equation 1), and (3) the duration of fluent transitions (the typing speed, which was indicated as β in equation 1).

Figure 1 illustrates the mixture model approach on the basis of the posterior of the constrained implementation applied to the CATO data: for each transition durations we obtain two posterior probability distributions, one associated with fluent transitions between keys and another for key transitions where upstream difficulty result in hesitant transitions. The hesitation probability is indicated as the relative height of either distribution. The transition locations are characterized by the hesitation parameters: as we illustrate the constrained mixture model, the distributions of fluent transitions – indicated in grey – are constant across transition location but their height differs as the probability of fluent transition is the inverse of the height of hesitant transitions; i.e. $1 - \theta$. More interesting are the distributions associated with hesitations; indicated in yellow: Within word transitions have a negligibly small hesitation probability barely visible in the visualisation. Hesitations before words and sentences are both roughly equally likely than fluent transitions. In other words, half of the time participants did not pause before a word or sentence; we explore this

below. However when hesitations occurred, these are longer at before-sentence transitions compared to before word transitions.

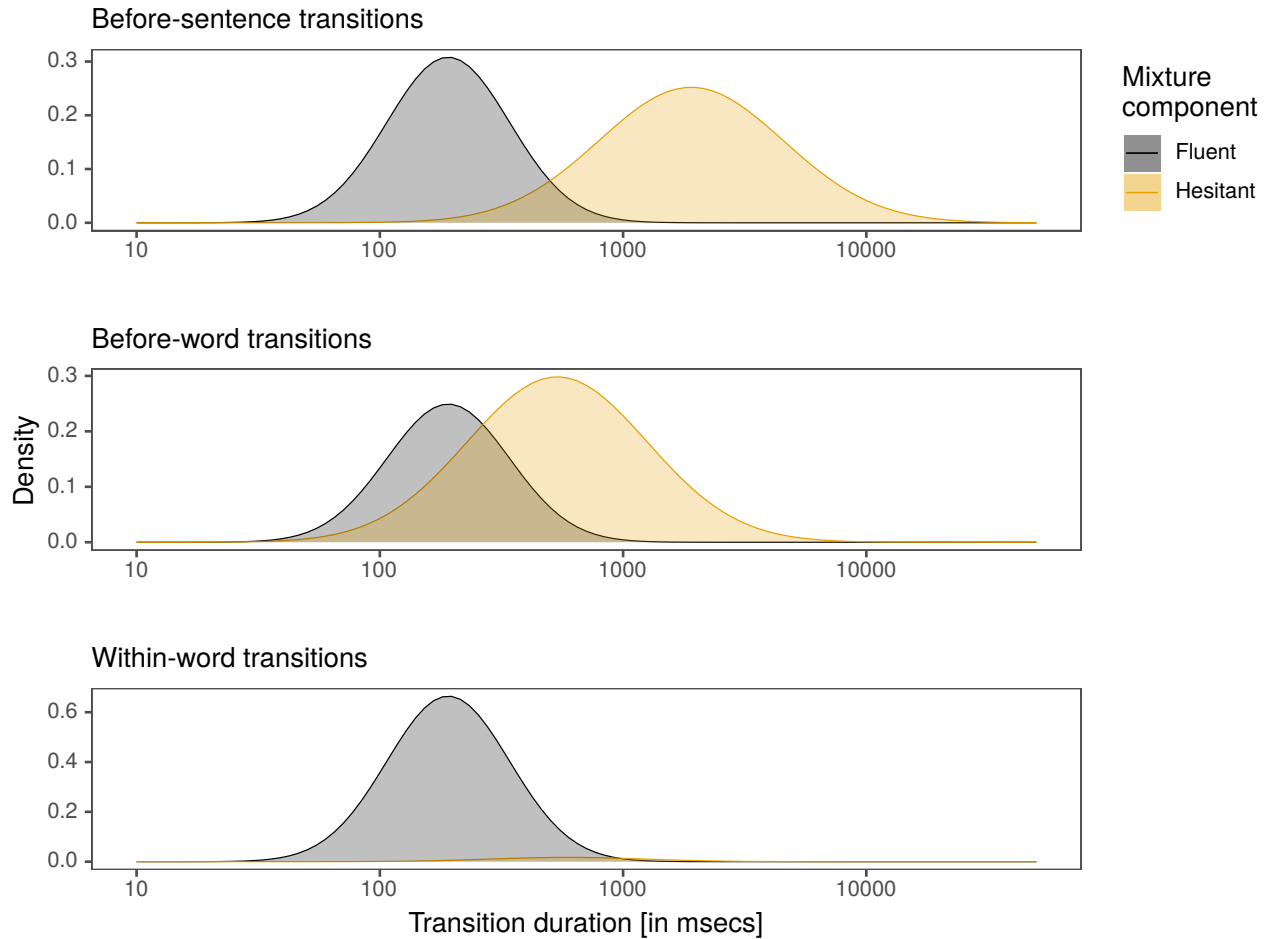


Figure 1

Mixture model plot for each transition location. Shown are the distributions for both mixture components (fluent and hesitant transitions) and their weighting. Posterior of CATO data.

We evaluate the posterior of the constrained mixture model; however we also summarise the posterior of the unconstrained mixture model below and the full analysis can be found in Appendix C. We demonstrated above that the predictive performance of the unconstrained mixture model was substantially higher than the predictive performance of the constrained mixture model. However, the constrained mixture model is the theoretically more parsimonious model of the cascading view; in other words, differences between transition locations for the component of fluent transitions were to some extent speculative

and unexpected. We therefore present the posterior parameter values for the constrained two-distributions mixture model in the following.

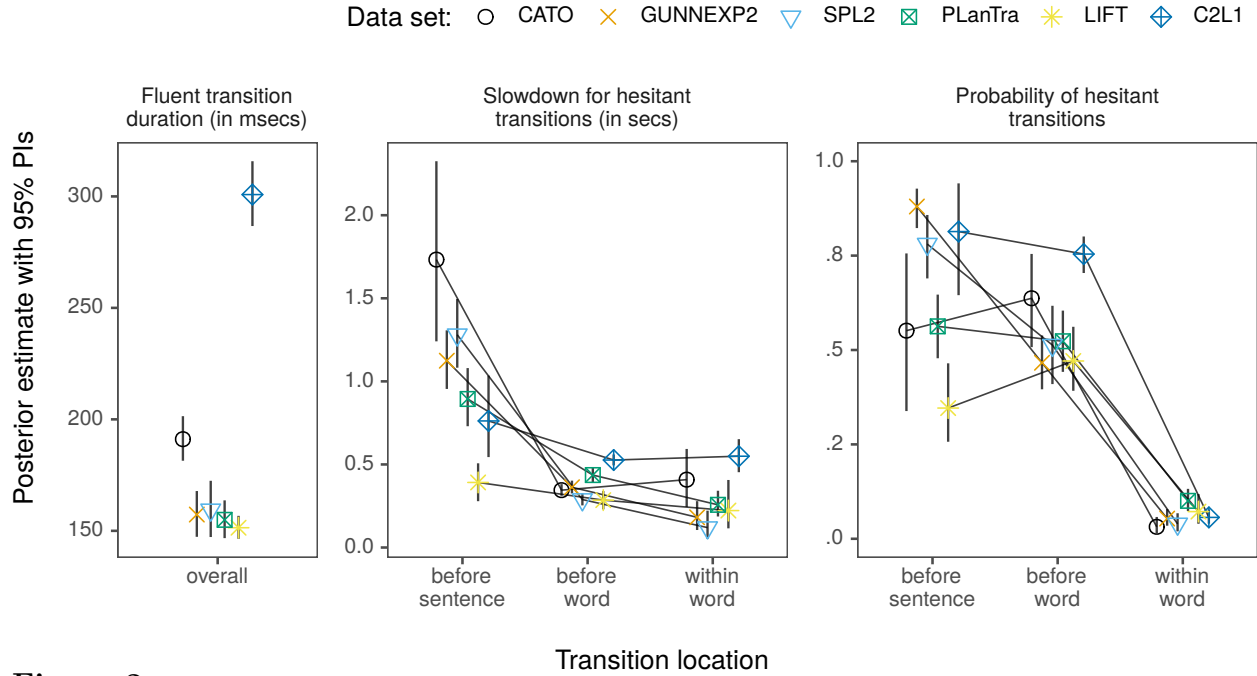
The mixture-model posterior parameter estimates for all datasets is visualised in Figure 2. Although models were fitted with all data set-specific conditions, we aggregated the posterior across conditions⁵, and excluded conditions that might conflate comparisons⁶. For posteriors of all conditions within data sets see Appendix H. The resulting posterior allows us to examine differences and similarities associated with transition locations for all dataset. In particular we can see how the linguistic location in the text is associated with hesitation duration and probability. The estimate for fluent transitions is shown by dataset, across transition location.

Figure 2 shows largely the same patterns (with caveats) for all datasets. We summarise the results of the pairwise comparisons for transitions locations; full results and parameter estimates can be found in Appendix K. We found that hesitations were more frequent – higher hesitation probability – before words than within words (BFs > 100 for all datasets) but only for half of the data sets also longer (GUNNEXP2: BF = 27; PJanTra: BF = 56; SPL2: BF = 9.1; negligible evidence for C2L1: BF = 0.1; CATO: BF = 0.2; LIFT: BF = 0.4). Hesitations were equally likely before sentences and words (C2L1 and CATO: BFs = 0.7; PJanTra: BF = 0.3; LIFT: BF = 1.6; except for GUNNEXP2 and SPL2, both BFs > 100) but longer (all BFs > 100, except for C2L1: BF = 1.11 and LIFT = 0.5).

To reiterate, for most datasets we found that pauses are before sentences are longer

⁵ We aggregated across pre-post test for the PJanTra dataset as well as genre and topic of the LIFT data set. We demonstrate in Appendix I and J, respectively, that there is negligible evidence for differences between these conditions.

⁶ We removed the masked writing condition in the GUNNEXP2 and CATO, the dyslexic group in the CATO data set, and L2 writing in the SPL2 data set. There was evidence for L2 writing effects (see Appendix F); evidence for marking effects was weak (see Appendix G).

**Figure 2**

Mixture model parameter estimates across studies. Distributions of parameter estimates are represented as posterior mean and 95% probability interval (PI). Estimates for the CATO dataset were calculated for the non-dyslexic group, unmasked condition; also the GUNNEXP2 estimates represent the unmasked condition; SPL2 estimates are for the L1 group.

but not necessarily more likely than before words (except for datasets SPL2 and GUNNEXP2). In fact, LIFT showed some indication that pauses were less likely before sentences than words. This is interesting because it is generally believed that pausing behaviour is associated with linguistic boundaries such that more and longer pauses are predicted for key transitions at larger linguistic edges. It is less clear in which contexts pre-sentence pausing is more likely than pausing before words.

Differences between datasets could, to some extent, be explained by the fact that two datasets (PLaTra, LIFT) defined before-sentence transitions as the sum of transitions between space, shift and the sentence-initial character while the remaining data sets measured before-sentence transitions as the difference between space key and shift. To

control for the use of complex key combinations, we addressed this possibility for the SPL2 dataset and showed that including the character following the shift key substantially increased the hesitation probability (BFs > 100) but it did not affect the duration of fluent or hesitant transitions (BFs < 0.3); see Appendix E. This conflicts the earlier finding that before-sentence hesitations were not more likely than before-word hesitations for datasets that did include the character following shift (PLanTra, LIFT). In other words, patterns in our results cannot be explained on the basis of how before-sentence key-transitions were operationalised.

Finally, the duration of fluent transitions was not distinguished by location in the constrained distribution model. Figure 2 highlights that two datasets (C2L1, CATO) show substantially longer fluent transitions. This is presumably reflecting that these data were sampled from was the youngest and least experienced population of writers in our data pool. Hesitation duration and probability followed the same patterns as the majority of datasets.

The predictive performance of the unconstrained model was higher compared to the constrained model (see Table 5). We, therefore, summarise briefly differences in the results for the transition-location effects; full results can be found in Appendix C. Higher predictive performance for the unconstrained model suggests that variability across transition locations for fluent transition durations; in particular fluent before-word transitions were longer compared to mid-word transitions (all BFs > 100) but not before-sentence compared to before-word transitions (all BFs < 0.08 , but SPL2 and GUNNEXP2: BFs > 100). Patterns for hesitation probability was generally similar in both the constrained and unconstrained model. However, for hesitation durations the unconstrained model showed evidence for longer durations at before-word locations compared to within-word location in all but one dataset (all BFs > 10 , except for LIFT: BF = 0.7); differences in hesitation duration for before-sentence transitions compared to before-word transitions was similar to the constrained model.

Discussion

The classical serial view (Flower & Hayes, 1980) characterizes writing as a sequence of planning and execution cycles that results in a writing process that consists of bursts interrupted by pauses. The cascading model of writing, in contrast, captures that planning, at least in competent writers can occur in parallel to production (Olive, 2014). We used Bayesian statistical models, in particular hierarchical single-distribution and two-distribution mixture models, to directly compare the serial and the cascading model of writing, respectively. On the basis of six datasets with key data from free text production tasks, we presented compelling and consistent statistical evidence in favour of the cascading view.

Patters observed in the parameter estimates are generally similar across data sets (with caveats). Under the cascading view of writing, key transitions that directly precede sentences or words are not necessarily associated with a pause because writers can plan in parallel to writing. On the basis of parameter estimates of the mixture model we tested to what extent larger linguistic edges are associated with the preparation of upcoming planning unit. We found that in the majority of datasets, pauses were as likely to occur before sentences and words, in both experienced and novice writers, while mid-word pauses are generally rare. Also, with few exceptions, pauses were typically longer before sentences than before words (which were longer than mid-word pauses). In other words, when writers do pause, the involvement of higher-level processing at sentence boundaries is typically leading to longer pauses. However, the pausing probability results can only be explained by the cascading view of writing, not by the serial view. This is because our results for the hesitation probability suggest that writers often do not pause before sentences; indeed, hesitation probabilities are largely identical before sentences and before words. This suggest that writers frequently plan the next word and even sentence in parallel to production which mitigates the need to pause to plan up-coming linguistic information.

Interestingly even the youngest writers in our data pool (C2L1) showed evidence of

parallel planning expressed in fluent transitions between words and even sentences.

Key-transitions were substantially longer compared to all other datasets with a larger slowdown before words. Yet the overall pausing probability showed a patterns comparable to most other data sets. In particular, a larger pause probability was observed for transitions before sentences / words compared to mid-word transitions. This is interesting because, for example, Olive (2014) described a cascading model that is operating in a serial fashion for inexperienced or struggling writers explained on the basis of task demands that reduce the ability to plan in parallel to processing which leads to a serialisation of planning units. However, the results presented in this paper suggest that pausing behaviour in young writers largely mirror results from more experienced writers but pauses were more likely before words and sentences but similarly long indicating a tendency to prepare utterances in planning units possibly smaller than the clause. The main difference for these young writers was that their writing speed (fluent transitions) was substantially slower compared to more experienced writers in the sample. Also for L2 writers (PlanTra) we observe pausing patterns that do not resemble a serialised writing process as indicated by a larger number of fluent transitions across word and sentence boundaries.

We observed two more differences between data sets – other than the patterns observed for C2L1 summarised above – that are worth highlighting. These can largely be explained on the basis of differences associated with the population (writing competence / experience) and text genre (see also Conijn et al., 2019) that these samples were obtained from: The first exception concern the GUNNEXP2 and SPL2 data sets: pauses were more likely to occur at before-sentence transitions compared to before-word transitions while in the remaining data sets hesitations were equally probably at before-word and before-sentence locations. **TODO: I don't have an explanation for this.** Second, pauses were not always longer before sentence compared to before words compared to mid-word pauses; there were two exceptions: (1) no difference was found for before-sentence compared to before-word transitions in C2L1 possibly due to a word-level planning strategy in this

sample of young writers as inexperienced writers tend to use a more localised strategy for text production. (2) pause durations did not vary by transition location in the LIFT dataset although the posterior showed the same general patterns as observed in most other datasets. The absence of a difference could be explained as specific to the synthesis writing task because there is no need for content creation [**REFERENCE?**] but the results for the PlanTra dataset rule out this possibility.

A possible concern with our result – substantially better predictive performance for two-distributions mixture models – is that, in principle, as the mixture model has more parameters it might always lead to a better fit. This concern is preempted by using cross-validation techniques for model comparison which is preventing model overfit. In addition we used a simulation approach to compare a single distribution model and a two distributions mixture model. We simulated two data sets that were either based on a single distribution or two-distributions random data generating process. This approach allows us to test the predictive performance of models in a context where we know the true underlying data generating process – single distribution vs a mixture of two distributions – and we can test to what extent these models can successfully uncover the true parameter values. Modelling details can be found in Appendix L. As might be expected, we found, that the two-distribution mixture was successful at uncovering the parameter values of the data generated with a two-distributions mixture models; we observed the same for the single-distribution model of the data generated with a single distribution process. Parameters were not successfully uncovered when we switched model type and dataset. Importantly though, cross-validation did not show a higher predictive performance for the mixture model, compared to the single distribution model, when applied to data based on a single distributions process but a substantially higher predictive performance for data simulated from a two-distributions mixture process. Therefore we can rule out the possibility that the overwhelming statistical support for the two-distributions mixture models can be explained in terms of overfitting.

Throughout this paper we assumed that hesitations are indicators of planning upcoming ideas and encoding of linguistic units. Alternatively at least some if not all pauses in the writing process might be the result of reading or looking back into previously produced text. There is limited research that used eye tracking to investigate regressive gaze behaviour during writing (Alamargot et al., 2010; Beers et al., 2010; Chukharev-Hudilainen et al., 2019; De Smet et al., 2018; Torrance et al., n.d.; Van Waes et al., 2010). There are at least four reasons why eye movements away from the inscription point might occur during writing: First, reading is necessary for revision (i.e. writers must check if writing goals have been met). Second, writers might use text to cue or reinstating ideas after interruption of writing process. Third, writers might look for or correct spelling or grammatical errors. We removed transitions that terminated in an editing operation so the pauses we detected are unlikely to reflect revisions of the previous produced text. Fourth, writers might look at the source they are writing from. For two datasets (PlanTra, LIFT) which involved writing from sources, we removed transitions that were made when the sources (rather than the writing document) was in focus.

Frequencies of lookbacks during writing follow a similar pattern as the one we observed for pausing: for example Torrance, Johansson, et al. (2016) reported that lookbacks appear with a frequency of 45% before sentences, 12% before words and 5% within words of which 36% were associated with sustained reading but mostly less patterned forward and backward saccades between words (“hopping,” see also Chukharev-Hudilainen et al., 2019). This is except, lookbacks are more frequent before sentences compared to words while we found that pause frequencies appear equally often before sentences and words. Finally even though reading is likely to explain some proportion of our results it is overall rare, certainly in samples of university students. This can be seen in the result of two datasets in our pool (CATO, GUNNEXP2): these experiments included a manipulation where the produced text was masked (or not). In a reanalysis we found no substantial differences in pausing behaviour for the masked condition; see Appendix G. This suggests that pausing patterns do

not change if reading previous text is not possible.

A practical advantage of the mixture-model approach (also Roeser et al., 2021) is that we are not required to stipulate interval thresholds as plausible candidate for lower pause bound to handle the complex distribution of latency from spontaneous text production. Such threshold are widely used by writing researchers, a strategy that was inherited from early research in speech production (for review see Rochester, 1973), with studies dating back to at least the mid 1990s (Foulin, 1998). There is one central problems with this approach: A threshold requires a definition of what passes as a pause (Van Waes et al., 2016; Wengelin, 2006), i.e. a pause criterion threshold often set to 2 secs (Chanquoy et al., 1996; Kaufer et al., 1986; Sullivan & Lindgren, 2002; Wengelin, 2002) or some other lower bound (Chukharev-Hudilainen, 2014; Connelly et al., 2012; Leijten & Van Waes, 2013). Even if these thresholds were adjusted relative to factors such as writing medium, experience of writer, and text location of pause (see e.g. Wengelin, 2006) they would be arbitrary. For example, a sentence-initial pause of 2 secs has a very different interpretation from a 2 secs pause that occurs before or within a word. Our current understanding of the processes that underlie text production does not provide a strong theoretical basis on which to make this decision. So while a 2 secs threshold undoubtedly captures an interesting distinction – processing that occurs the range zero to 2 secs is very likely to be qualitatively different from processing that takes more than 2 secs. However the same could be argued for any threshold between perhaps 250 msecs and 10 secs that a researcher might care to choose (Chenu et al., 2014). Mixture models provide a principled statistical framework that allows the researcher to model behavioural data that come from a combination of cognitive processes without imposing threshold values (see also Almond et al., 2012; Baaijen et al., 2012; Hall et al., 2022; Li, 2021).

Conclusion

In contrast to the serial pause-and-burst view, the cascading view emphasizes that writers do not necessarily pause to plan the upcoming language unit but plan in parallel to writing execution. Using the Bayesian framework we implemented models of both the serial model of writing and the cascading view. Model comparisons and the inspection of posterior parameter estimates supported with a cascading view of writing but provided strong evidence against the serial view. We found that key transitions across word and even sentences boundaries are frequently fluent and not interrupted but planning of upcoming linguistic information. This pattern was found to be largely consistent for populations with different levels of writing experience and languages (e.g. young / L2 writers, students) and writing tasks (e.g. essays, syntheses).

References

- Ailhaud, E., & Chenu, F. (2017). Variations of chronometric measures of written production depending on clause packaging. *CogniTextes*, 17. <https://doi.org/10.4000/cognitextes.992>
- Ailhaud, E., Chenu, F., & Jisa, H. (2016). A developmental perspective on the units of written French. In J. Perera, M. Aparici, E. Rosado, & N. Salas (Eds.), *Written and spoken language development across the lifespan: Essays in honour of Lilibiana Tolchinsky* (pp. 287–305). Springer. https://doi.org/10.1007/978-3-319-21136-7_17
- Alamargot, D., Plane, S., Lambert, E., & Chesnet, D. (2010). Using eye and pen movements to trace the development of writing expertise: Case studies of a 7th, 9th and 12th grader, graduate student, and professional writer. *Reading and Writing*, 23(7), 853–888.
- Almond, R., Deane, P., Quinlan, T., Wagner, M., & Sydorenko, T. (2012). *A preliminary analysis of keystroke log data from a timed writing task* (Research Report No. RR-12-23). Educational Testing Service.
- Alves, R. A., & Limpo, T. (2015). Progress in written language bursts, pauses, transcription, and written composition across schooling. *Scientific Studies of Reading*, 19(5), 374–391.
- Baaijen, V. M., Galbraith, D., & De Glopper, K. (2012). Keystroke analysis: Reflections on procedures and measures. *Written Communication*, 29(3), 246–277.
- Beers, S. F., Quinlan, T., & Harbaugh, A. G. (2010). Adolescent students' reading during writing behaviors and relationships with text quality: An eyetracking study. *Reading and Writing*, 23(7), 743–775.
- Behmer Jr., L. P., & Crump, M. J. C. (2016). Crunching big data with fingertips: How typists tune their performance toward the statistics of natural language. In M. N. Jones (Ed.), *Big data in cognitive science* (Vol. 1, pp. 329–345). Psychology Press. <https://doi.org/10.4324/9781315413570-27>
- Bock, J. K., & Ferreira, V. S. (2014). In M. Goldrick, V. S. Ferreira, & M. Miozzo (Eds.), *The Oxford Handbook of Language Production* (pp. 21–46). Oxford University Press.
- Bonin, P., Roux, S., Barry, C., & Canell, L. (2012). Evidence for a limited-cascading account

- of written word naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(6), 1741–1758. <https://doi.org/10.1037/a0028471>
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, 20.
- Chang, F., Dell, G. S., & Bock, J. K. (2006). Becoming syntactic. *Psychological Review*, 113(2), 234–272.
- Chanquoy, L., Foulin, J.-N., & Fayol, M. (1996). Writing in adults: A real-time approach. In G. Rijlaarsdam, H. Van den Bergh, & M. Couzijn (Eds.), *Theories, models and methodology in writing research* (pp. 36–44). Amsterdam University Press.
- Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication*, 18(1), 80–98.
- Chenoweth, N. A., & Hayes, J. R. (2003). The inner voice in writing. *Written Communication*, 20(1), 99–118.
- Chenu, F., Pellegrino, F., Jisa, H., & Fayol, M. (2014). Interword and intraword pause threshold in writing. *Frontiers in Psychology*, 5. <https://doi.org/10.3389/fpsyg.2014.00182>
- Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39, e62. <https://doi.org/10.1017/S0140525X1500031X>
- Chukharev-Hudilainen, E. (2014). Pauses in spontaneous written communication: A keystroke logging study. *Journal of Writing Research*, 6(1), 61–84.
- Chukharev-Hudilainen, E., Saricaoglu, A., Torrance, M., & Feng, H.-H. (2019). Combined deployable keystroke logging and eyetracking for investigating L2 writing fluency. *Studies in Second Language Acquisition*, 41(3), 583–604.
- Clark, H. H., & Fox Tree, J. E. (2002). Using *uh* and *um* in spontaneous speaking. *Cognition*, 84, 73–111.

- Conijn, R. (2020). *The keys to writing: A writing analytics approach to studying writing processes using keystroke logging* [PhD thesis]. Tilburg University.
- Conijn, R., Roeser, J., & van Zaanen, M. (2019). Understanding the keystroke log: The effect of writing task on keystroke features. *Reading and Writing*, 32(9), 2353–2374.
- Conijn, R., & Torrance, M. (Eds.). (2023). Timecourse method. *Reading and Writing*. <https://link.springer.com/collections/gedbaiibja>
- Connelly, V., Dockrell, J. E., Walter, K., & Critten, S. (2012). Predicting the quality of composition and written language bursts from oral language, spelling, and handwriting skills in children with and without specific language impairment. *Written Communication*, 29(3), 278–302.
- Crump, M. J. C., & Logan, G. D. (2010). Hierarchical control and skilled typing: Evidence for word-level control over the execution of individual keystrokes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(6), 1369–1380. <https://doi.org/10.1037/a0020696>
- Damian, M. F., & Stadthagen-Gonzalez, H. (2009). Advance planning of form properties in the written production of single and multiple words. *Language and Cognitive Processes*, 24(4), 555–579.
- De Smet, M. J. R., Leijten, M., & Van Waes, L. (2018). Exploring the process of reading during writing using eye tracking and keystroke logging. *Written Communication*, 35(4), 411–447.
- Deane, P., & Zhang, M. (2015). *Exploring the feasibility of using writing process features to assess text production skills* (RR-15-02). Educational Testing Service.
- Dell, G. S., Martin, N., & Schwartz, M. F. (2007). A case-series test of the interactive two-step model of lexical access: Predicting word repetition from picture naming. *Journal of Memory and Language*, 56(4), 490–520.
- Dux Speltz, E., & Chukharev-Hudilainen, E. (2021). The effect of automated fluency-focused feedback on text production. *Journal of Writing Research*, 13(2), 231–255.

- Dux Speltz, E., Roeser, J., & Chukharev-Hudilainen, E. (2022). Automating individualized, process-focused writing instruction: A design-based research study. *Frontiers in Communication: Emerging Technologies and Writing: Pedagogy and Research*, 7, 933878. <https://doi.org/10.3389/fcomm.2022.933878>
- Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.
- Flower, L. S., & Hayes, J. R. (1980). The dynamics of composing: Making plans and juggling constraints. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing: An interdisciplinary approach* (pp. 31–50). Lawrence Erlbaum.
- Flower, L. S., & Hayes, J. R. (1981). The pregnant pause: An inquiry into the nature of planning. *Research in the Teaching of English*, 229–243.
- Foulin, J.-N. (1998). To what extent does pause location predict pause duration in adults' and children's writing? *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, 17, 601–620.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Chapman; Hall/CRC.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472.
- Gernsbacher, M. A., & Givón, T. (1995). *Coherence in spontaneous text* (Vol. 31). John Benjamins Publishing.
- Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, 82(1), B1–B14.
- Griffin, Z. M. (2003). A reversed word length effect in coordinating the preparation and articulation of words in speaking. *Psychonomic Bulletin & Review*, 10(3), 603–609.
- Grudin, J. T., & Larochelle, S. (1982). *Digraph frequency effects in skilled typing* (ONR-8201). Office of Naval Research.
- Guo, H., Deane, P. D., van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic

- writing processes from keystroke logs. *Journal of Educational Measurement*, 55(2), 194–216.
- Hall, S., Baaijen, V. M., & Galbraith, D. (2022). Constructing theoretically informed measures of pause duration in experimentally manipulated writing. *Reading and Writing*, 1–29.
- Hayes, J. R. (2012). Evidence from language bursts, revision, and transcription for translation and its relation to other writing processes. In M. Fayol, D. Alamargot, & V. Berninger (Eds.), *Translation of thought to written text while composing: Advancing theory, knowledge, methods, and applications* (pp. 15–25). Psychology Press.
- Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, 80(1), 64–72.
- Kandel, S. (2023). Written production: The APOMI model of word writing: Anticipatory processing of orthographic and motor information. In *Language production* (pp. 209–232). Routledge.
- Kaufer, D. S., Hayes, J. R., & Flower, L. S. (1986). Composing written sentences. *Research in the Teaching of English*, 121–140.
- Lambert, B. (2018). *A student's guide to Bayesian statistics*. Sage.
- Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication*, 30(3), 358–392.
- Levelt, W. J. M., & Meyer, A. S. (2000). Word for word: Multiple lexical access in speech production. *European Journal of Cognitive Psychology*, 12(4), 433–452.
- Li, T. (2021). Identifying mixture components from large-scale keystroke log data. *Frontiers in Psychology*, 12, 628660.
- MacLahlan, G., & Peel, D. (2000). *Finite mixture models*. John & Sons.
- Matsushashi, A. (1981). Pausing and planning: The tempo of written discourse production.

- Research in the Teaching of English*, 113–134.
- McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in R and Stan*. CRC Press.
- Medimorec, S., & Risko, E. F. (2017). Pauses in written composition: On the importance of where writers pause. *Reading and Writing*, 30, 1267–1285.
- Medimorec, S., Young, T. P., & Risko, E. F. (2017). Disfluency effects on lexical selection. *Cognition*, 158, 28–32.
- Meulemans, C., De Maeyer, S., & Leijten, M. (2022). Generalizability of pause times in sentence production to distinguish between adult writers. *Behavior Research Methods*, 54(4), 1976–1988. <https://doi.org/10.3758/s13428-021-01707-1>
- Mohsen, M. (2021). Second language learners' pauses over different times intervals in L2 writing essays: Evidence from a keystroke logging program. *Psycholinguistics*, 30(1), 180–202. <https://doi.org/10.31470/2309-1797-2021-30-1-180-202>
- Nottbusch, G. (2010). Grammatical planning, execution, and control in written sentence production. *Reading and Writing*, 23(7), 777–801.
- Nottbusch, G., Weingarten, R., & Sahel, S. (2007). From written word to written sentence production. In M. Torrance, L. van Waes, & D. W. Galbraith (Eds.), *Writing and cognition: Research and applications* (Vol. 20, pp. 31–53). Elsevier.
- Olive, T. (2014). Toward a parallel and cascading model of the writing system: A review of research on writing processes coordination. *Journal of Writing Research*, 6(2), 173–194.
- Ostry, D. J. (1983). Determinants of interkey times in typing. In W. E. Cooper (Ed.), *Cognitive aspects of skilled typewriting* (pp. 225–246). Springer. https://doi.org/10.1007/978-1-4612-5470-6_9
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rochester, S. R. (1973). The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research*, 2, 51–81.

- Roeser, J., De Maeyer, S., Leijten, M., & Van Waes, L. (2021). Modelling typing disfluencies as finite mixture process. *Reading and Writing*, 1–26.
<https://doi.org/10.1007/s11145-021-10203-z>
- Roeser, J., Torrance, M., & Baguley, T. (2019). Advance planning in written and spoken sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(11), 1983–2009. <https://doi.org/10.1037/xlm0000685>
- Rønneberg, V., Torrance, M., Uppstad, P. H., & Johansson, C. (2022). The process-disruption hypothesis: How spelling and typing skill affects written composition process and product. *Psychological Research*, 86(7), 2239–2255.
- Rossetti, A., & Van Waes, L. (2022a). It’s not just a phase: Investigating text simplification in a second language from a process and product perspective. *Frontiers in Artificial Intelligence*, 5. <https://doi.org/10.3389/frai.2022.983008>
- Rossetti, A., & Van Waes, L. (2022b). *Text simplification in second language: Process and product data* [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.6720290>
- Roux, S., McKeeff, T. J., Grosjacques, G., Afonso, O., & Kandel, S. (2013). The interaction between central and peripheral processes in handwriting production. *Cognition*, 127(2), 235–241.
- Schilperoord, J. (2002). On the cognitive status of pauses in discourse production. In T. Olive & C. M. Levy (Eds.), *Contemporary tools and techniques for studying writing* (pp. 61–87). Springer.
- Schöner, G. (2002). Timing, clocks, and dynamical systems. *Brain and Cognition*, 48(1), 31–51.
- Simpson, S., & Torrance, M. (2007). *EyeWrite*. SR Research; Nottingham Trent University.
- Sivula, T., Magnusson, M., Matamoros, A. A., & Vehtari, A. (2020). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *arXiv Preprint arXiv:2008.10296*.
- Spelman Miller, K. (2006). Pausing, productivity and the processing of topic in online writing. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer key-stroke logging and*

- writing* (Vol. 18, pp. 131–156). Brill. https://doi.org/10.1163/9780080460932_009
- Spelman Miller, K., Lindgren, E., & Sullivan, K. P. (2008). The psycholinguistic dimension in second language writing: Opportunities for research and pedagogy using computer keystroke logging. *Tesol Quarterly*, 42(3), 433–454.
- Stan Development Team. (2018). *RStan: The R interface to Stan*. <https://mc-stan.org/>
- Sullivan, K. P. H., & Lindgren, E. (2002). Self-assessment in autonomous computer-aided second language writing. *ELT Journal*, 56(3), 258–266.
- Terzuolo, C. A., & Viviani, P. (1980). Determinants and characteristics of motor patterns used for typing. *Neuroscience*, 5(6), 1085–1103.
[https://doi.org/10.1016/0306-4522\(80\)90188-8](https://doi.org/10.1016/0306-4522(80)90188-8)
- Torrance, M. (2012). EyeWrite – a tool for recording writers’ eye movements. In M. Torrance, D. Alamargot, M. Castelló, F. Ganier, O. Kruse, A. Mangen, L. Tolchinsky, & L. V. Waes (Eds.), *Learning to write effectively: Current trends in European research* (pp. 355–357). Brill. https://doi.org/doi.org/10.1163/9781780529295_082
- Torrance, M., Johansson, R., Johansson, V., & Wengelin, Å. (2016). Reading during the composition of multi-sentence texts: An eye-movement study. *Psychological Research*, 80(5), 729–743.
- Torrance, M., & Nottbusch, G. (2012). Written production of single words and simple sentences. In V. W. Berninger (Ed.), *Past, present, and future contributions of cognitive writing research to cognitive psychology* (pp. 403–422). Psychology Press.
- Torrance, M., Nottbusch, G., Alves, R. A., Arfé, B., Chanquoy, L., Chukharev-Hudilainen, E., Dimakos, I., Fidalgo, R., Hyönä, J., Jóhannesson, Ó. I., et al. (2018). Timed written picture naming in 14 European languages. *Behavior Research Methods*, 50, 744–758.
- Torrance, M., & Ofstad, G. (n.d.). *Cascading processes in multi-sentence text production*.
- Torrance, M., Roeser, J., & Chukharev-Hudilainen, E. (n.d.). *Lookback in L1 and L2 writing: An eye movement study*.
- Torrance, M., Rønneberg, V., Johansson, C., & Uppstad, P. H. (2016). Adolescent weak

- decoders writing in a shallow orthography: Process and product. *Scientific Studies of Reading*, 20(5), 375–388.
- Van Galen, G. P. (1991). Handwriting: Issues for a psychomotor theory. *Human Movement Science*, 10(2), 165–191.
- Van Waes, L., Leijten, M., Lindgren, E., & Wengelin, Å. (2016). *Keystroke logging in writing research: Analyzing online writing processes*. 410–426.
- Van Waes, L., Leijten, M., Pauwaert, T., & Van Horenbeeck, E. (2019). A multilingual copy task: Measuring typing and motor skills in writing with inputlog. *Journal of Open Research Software*, 7(30), 1–8.
- Van Waes, L., Leijten, M., & 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., Roeser, J., Olive, T., & Grabowski, J. (2021). Measuring and assessing typing skills in writing research. *Journal of Writing Research*, 13(1), 107–153. <https://doi.org/10.17239/jowr-2021.13.01.04>
- Vandermeulen, N., De Maeyer, S., Van Steendam, E., Lesterhuis, M., Van den Bergh, H., & Rijlaarsdam, G. (2020). Mapping synthesis writing in various levels of Dutch upper-secondary education: A national baseline study on text quality, writing process and students' perspectives on writing. *Pedagogische Studiën: Tijdschrift Voor Onderwijskunde En Opvoedkunde*, 97(3), 187–236.
- Vandermeulen, N., Steendam, E. V., & Rijlaarsdam, G. (2020). *DATASET – baseline data LIFT synthesis writing project* [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.3893538>
- Vandermeulen, N., Van Steendam, E., De Maeyer, S., & Rijlaarsdam, G. (2023). Writing process feedback based on keystroke logging and comparison with exemplars: Effects on the quality and process of synthesis texts. *Written Communication*, 40(1), 90–144.
- Vasishth, S., Chopin, N., Ryder, R., & Nicenboim, B. (2017). Modelling dependency

- completion in sentence comprehension as a Bayesian hierarchical mixture process: A case study involving Chinese relative clauses. *ArXiv e-Prints*.
- Vasishth, S., Jäger, L. A., & Nicenboim, B. (2017). Feature overwriting as a finite mixture process: Evidence from comprehension data. *arXiv Preprint arXiv:1703.04081*.
- Vehtari, A., Gelman, A., & Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv Preprint arXiv:1507.02646*.
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432.
- Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, 114(3), 830–841. <https://doi.org/10.1037/0033-295X.114.3.830>
- Wengelin, Å. (2002). *Text production in adults with reading and writing difficulties* [PhD thesis]. Göteborg University.
- Wengelin, Å. (2006). Examining pauses in writing: Theory, methods and empirical data. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer keystroke logging and writing: Methods and applications* (Vol. 18, pp. 107–130). Elsevier.
- Wheeldon, L. R., & Konopka, A. (2023). *Grammatical encoding for speech production*. Cambridge University Press. <https://doi.org/10.1017/9781009264518>
- Wing, A. M., & Kristofferson, A. B. (1973). Response delays and the timing of discrete motor responses. *Perception & Psychophysics*, 14(1), 5–12.
- Xu, C., & Qi, Y. (2017). Analyzing pauses in computer-assisted EFL writing—a computer-keystroke-log perspective. *Journal of Educational Technology & Society*, 20(4), 24–34.

Appendix A

Statistical models

We are using the Bayesian framework (e.g. Farrell & Lewandowsky, 2018; Gelman et al., 2014; Lee & Wagenmakers, 2014) to implement 5 statistical models of writing and to evaluate which – the serial or the cascading view – captures more successfully the data that arise during keyboard typing. In other words, we use statistical models to map between the data and the theoretically assumed process that generates the data to then compare the predictive power of those models. The models presented in this section build on one another so that later model include assumptions of earlier models.

Serial model of writing

Single distribution Gaussian

Under the serial view, all planning must be completed prior to the production onset of the corresponding planning unit. The resulting IKI is sometimes faster or slower depending on, among others, psycholinguistic factors. For example, the interval before a word is shorter for easily retrievable high-frequency words, or longer for low frequency words, shorter for words with fewer graphemes, syllables, and morphemes. There are word-specific factors that influence the IKI that precedes a word but these are beyond the scope of our analysis. We will capture the variability associated with word-features by assuming that before-word IKIs can be described as coming from a distribution that is normal (Gaussian) with two parameters, an unknown mean μ , that describes the average IKI associated with word-level planning, and a standard deviation σ_e^2 , that captures the variability associated with factors that we did not further specify in the model. This can be expressed as $\text{iki}_{\text{before word}} \sim \log\mathcal{N}(\mu, \sigma_e^2)$. Of importance is the estimated posterior value of μ as this value captures that time it takes to mentally plan a word.

We can extend this simple model of word-planning to other linguistic location. We introduce earlier that larger linguistic edges are associated with planning on higher levels.

For example, at sentence boundaries, planning needs to happen for word-level properties – which was captured above as average IKI μ – but also higher linguistic planning such as clause-level meaning, and dependencies of the sentence-initial noun (Nottbusch et al., 2007; Roeser et al., 2019). Regression formulas can, and commonly are, used to decompose μ and capture that there is a change in the outcome variable associated with another factors. We can decompose μ as $\mu = \alpha + \beta \cdot x_{\text{sentence}[0,1]}$ so that when x_{sentence} takes on the value 0, the equation reduces to $\mu = \alpha$ which is the average IKI for word boundaries. However when x_{sentence} takes on the value 1, the average IKI for word boundaries α is incremented by a sentence-related slowdown of β msecs. Therefore the value of the β parameter represents the additional cognitive demand associated with sentence-initial planning. The application of such a statistical model to the data will then provide us with an estimate of the parameter value that can be used for statistical inference (e.g. whether there is a statistically meaningful difference between IKI associated with words and sentences).

We implemented the average keytransition associated with a given text locations as β_{location} in equation A1 where location is taking on an index for each transition location. Therefore, the model will return one β per transition location that capture the posterior distribution of average IKIs. Also the decomposition of μ allows us to address the fact that some writers are faster than other writers by introducing a parameter for what is typically called random intercepts $u_{\text{participant}}$. The random intercepts term $u_{\text{participant}}$ is constrained so that the value it takes on comes from a normal distribution with a mean of 0 and a standard deviation of σ_p^2 . This value is therefore the difference between the overall posterior estimate and estimate average IKI of a particular participant (i.e. a positive value indicates that a participant is slower than average, a negative value indicates that a participant is faster than average).

The final model is shown in equation A1 and represents a Gaussian mixed effects model

$$\begin{aligned}
\text{iki}_i &\sim \log\mathcal{N}(\mu_i, \sigma_e^2) \\
\text{where: } \mu_i &= \beta_{\text{location}[i]} + u_{\text{participant}[i]} \\
u_{\text{participant}[i]} &\sim \mathcal{N}(0, \sigma_p^2) \\
\text{constraint: } \sigma_e^2, \sigma_p^2 &> 0
\end{aligned} \tag{A1}$$

where i iterates over every key-transition $i \in 1 \dots N$ with N being the total number of keystroke intervals, and thus $\text{location}[i]$ is the transition location associated with the i^{th} keystroke and $\text{participant}[i]$ is the participant associated with the i^{th} keystroke.

Standard-deviation parameters were constrained to be positive because standard deviations can never be negative.

Single distribution log-Gaussian

The previous model assumes that the data-generating process is a Gaussian distribution. The next model is largely identical to the previous model but instead of assuming a Gaussian, we assume that the data come from a log-normal (log-Gaussian) distribution. There are, at least, two arguments for using a log-Gaussian distribution: (1) log-Gaussians are zero-bound; in contrast to Gaussians, a log-Gaussian does not allow negative values. IKIs, as the distance between two subsequent key-down events, must be positive. The lower bound is delimited by a person's ability to move their fingers and keyboard polling. (2) the log-scale is known to be a better match for data from human behaviour and motor responses. In particular, a Gaussian distribution assumes that units are scaled linearly. For example, a difference of 25 msecs is the same between 100 and 125 msecs as between 5 secs and 5,025 msecs. This does not necessarily map onto the psychological interpretation for short and long keyintervals. For example effects that result on the motor level within words (e.g. typing a high vs a low frequency bigram) are smaller than differences that are due to high levels of processing (retrieving a word in an L1 or L2). In other words, although an effect of 25 msecs is large in the context overall fast keyintervals, it is small in

the context of overall slow intervals. Log-Gaussian distributions are a natural way of translating a linear scale to an exponential scale so that a 25 msec difference on the lower end of the IKI scale (motor activity) is more meaningful than a 25 msec difference on the upper end of the IKI scale (retrieving words, planning sentences).

The model can be described as in equation A2 in which the distribution $\log\mathcal{N}()$ was replaced by $\log\mathcal{N}()$.

$$\begin{aligned}
 \text{iki}_i &\sim \log\mathcal{N}(\mu_i, \sigma_e^2) \\
 \text{where: } \mu_i &= \beta_{\text{location}[i]} + u_{\text{participant}[i]} \\
 u_{\text{participant}[i]} &\sim \mathcal{N}(0, \sigma_p^2) \\
 \text{constraint: } \sigma_e^2, \sigma_p^2 &> 0
 \end{aligned} \tag{A2}$$

Single distribution unequal-variance log-Gaussian

The third model representing the serial view of writing is a single distribution unequal-variance model that, except for the unequal-variance assumption is identical to the model presented in the previous section. The previous models modelled IKIs as a function of transition location so that the estimated average IKI depends on the position of an IKI in the text. The variance associated with the estimated IKIs for each transition location was assumed to be identical (equal variance). This assumption is however not in line with what we know about data from human behaviour. Longer latencies are known to be associated with a larger variances for both response-time data in particular (Wagenmakers & Brown, 2007) and human motor behaviour in general (Schöner, 2002; Wing & Kristofferson, 1973). For IKIs pauses at larger linguistic edges are plausibly associated with a larger variance because of the larger number of associated processes. Therefore, in equation A3, we introduce the assumption that standard deviation $\sigma_{e_{\text{location}}}^2$ varies by transition location.

$$\begin{aligned}
iki_i &\sim \log \mathcal{N}(\mu_i, \sigma_{e_{\text{location}[i]}}^2) \\
\text{where: } \mu_i &= \beta_{\text{location}[i]} + u_{\text{participant}[i]} \\
u_{\text{participant}[i]} &\sim \mathcal{N}(0, \sigma_p^2) \\
\text{constraint: } &\sigma_e^2, \sigma_p^2 > 0
\end{aligned} \tag{A3}$$

Parallel cascading model of writing

The following two models are extensions of the models introduced for the serial view. Crucially, the cascading view allows planning to happen in parallel to production. Therefore, we will reduce the constrain of the serial models that requires all planning to be completed before writing onset. This is done by assuming that IKIs come from a weighted combination of two distributions.

Two-distributions log-Gaussian mixture model (constrained)

This model extends the assumption of the previous model that processing that involves higher levels of activation lead to longer pauses. Instead of assuming that there is one process that is shifted for IKIs of larger linguistic edges, we introduce the assumption that pauses at larger linguistic edges are more likely but not obligatory. This is achieved by modelling IKIs as coming from a weighted mixture of two distributions associated with two different states:

1. Activation can flow into keystrokes without interruption. These fluent keystroke transitions are merely constrained by a person's ability to move their finger and will be captured by the β parameter in equation A4. Note that β is represented in both log-Gaussian distributions in equation A4 referring to the same unknown parameter.
2. Interruptions in the activation flow from higher to lower levels result in longer keystroke intervals, when words or their spelling could not be retrieved in time. The slowdown for these hesitant transitions is captured by δ in the first line of equation A4. The magnitude of the slowdown is associated with transition location. This is because

delays at larger linguistic units are likely to be associated with higher level planning. By constraining δ to be positive, it captures how much longer hesitant IKIs in addition to β .

$$\begin{aligned}
\text{iki}_i &\sim \theta_{\text{location}[i], \text{participant}[i]} \cdot \log \mathcal{N}(\beta + \delta_{\text{location}[i]} + u_{\text{participant}[i]}, \sigma_{e'_{\text{location}[i]}}^2) + \\
&\quad (1 - \theta_{\text{location}[i], \text{participant}[i]}) \cdot \log \mathcal{N}(\beta + u_{\text{participant}[i]}, \sigma_{e_{\text{location}[i]}}^2) \\
\text{where: } u_{\text{participant}[i]} &\sim \mathcal{N}(0, \sigma_p^2) \\
\text{constraint: } \delta, \sigma_e^2, \sigma_{e'}^2, \sigma_p^2 &> 0 \\
\sigma_{e'}^2 &> \sigma_e^2 \\
0 < \theta &< 1
\end{aligned} \tag{A4}$$

The first line of equation A4 represents the distribution of hesitant key transitions, and the second line represents fluent key transitions. Each of these two distributions is associated with the mixing weight θ which is a proportion that is constrained to be larger than 0 and smaller than 1. θ is here parameterised to represent the probability that an IKI is associated with the distribution of long IKIs. This probability is inversely related to the mixing weight of the distribution of short IKIs by $1 - \theta$. In other words, a larger weight of one distribution inevitably means a lower weight for the other distribution. The weights of both distributions must sum to 1. We will call this parameter the probability of hesitant transitions.

The probability of hesitant transitions is assumed to vary as a function of both transition location and participants. In line with the literature discussed in the introduction, we assume that pauses are more likely at larger linguistic edges. As pausing is subject to individual differences and writing style (skills), we also assume that some participants pause more at certain transition locations and other participants pause less. This is akin to what is known as a random by-participant slopes model.

Lastly, we carried over the unequal variance assumption and let the standard deviations $\sigma_{e'}$ and σ_e^2 vary by transition location. In addition we constrained the variances so that $\sigma_{e'}$ associated with the distribution of typing disfluencies is larger than the variance associated with fluent transitions σ_e (see Vasishth, Chopin, et al., 2017; Vasishth, Jäger, et al., 2017). This was achieved by introducing a parameter σ_{diff} . The consequence is that fluent transitions are assumed to come from a narrower distribution than hesitant transitions.

Two-distributions log-Gaussian mixture model (unconstrained)

The size of a fluent key-transition does not necessarily vary by transition location. In other words, the parameter β is the same for before-sentence, before-word, and within-word transitions. This is what we called a constrained model. However, letter bigrams and trigrams may be systematically executed faster than transitions between space and a letter (REFERENCE?) or complex keystrokes that comprise space and shift-letter combinations for upper case characters before sentences (we will address the latter possibility in the results section). This is because bigrams / trigrams might be stored, retrieved and executed as motor codes but not transitions to a space following a word or between a space or shift key press preceding a letter. Also, because of the necessarily larger number of within-word transitions, as opposed to before-word and sentence transitions, the posterior of the constrained model is dominated by within-word transition data. We therefore also implemented an unconstrained model that assumes that the size of fluent transitions varies across transition locations.

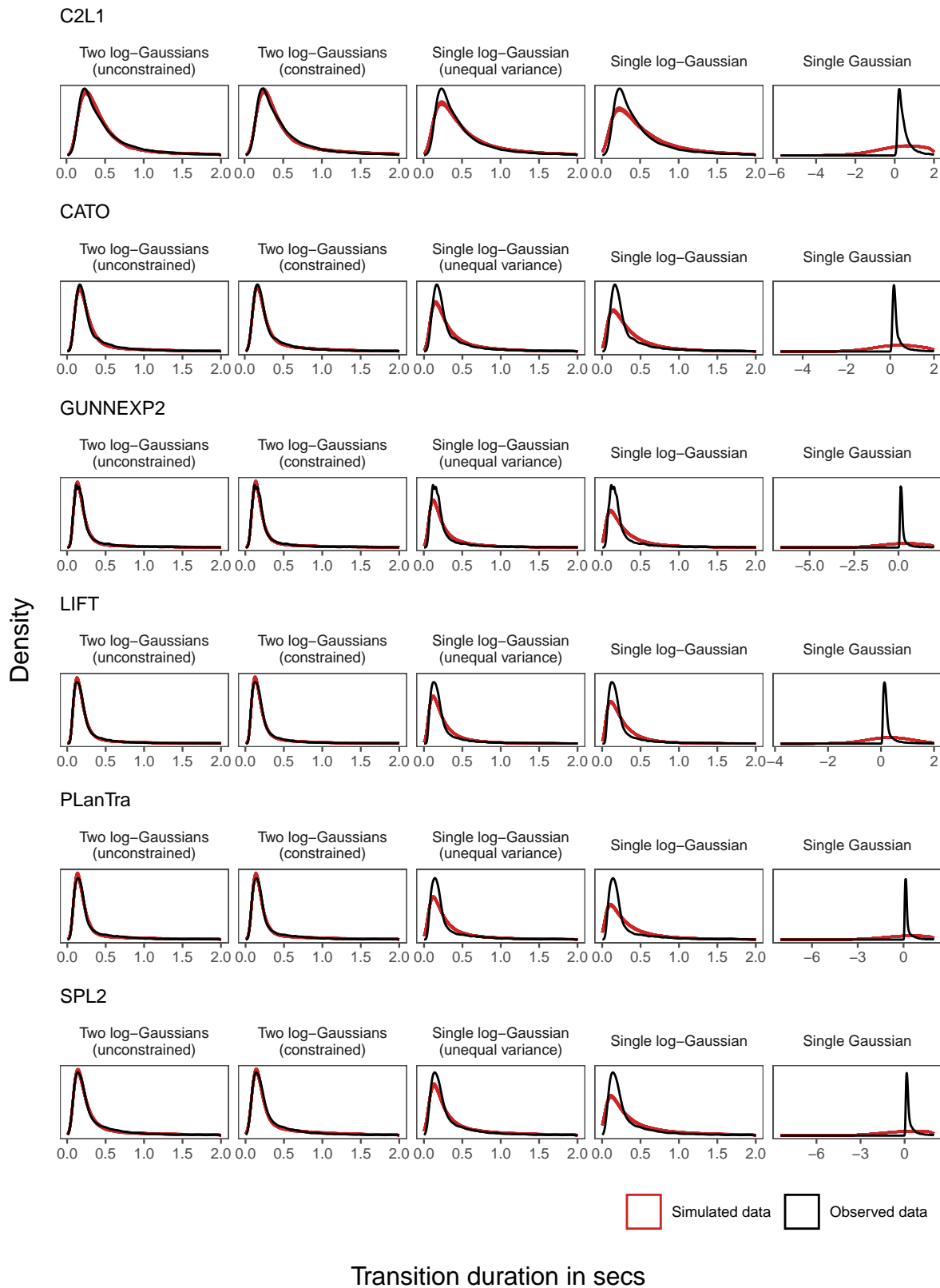
In this model we assume that β varies by transition-location as illustrated in equation A5.

$$\begin{aligned}
 \text{iki}_i &\sim \theta_{\text{location}[i], \text{participant}[i]} \cdot \log \mathcal{N}(\beta_{\text{location}[i]} + \delta_{\text{location}[i]} + u_{\text{participant}[i]}, \sigma_{e'_{\text{location}[i]}}^2) + \\
 &\quad (1 - \theta_{\text{location}[i], \text{participant}[i]}) \cdot \log \mathcal{N}(\beta_{\text{location}[i]} + u_{\text{participant}[i]}, \sigma_{e_{\text{location}[i]}}^2) \\
 \text{where: } &u_{\text{participant}[i]} \sim \mathcal{N}(0, \sigma_p^2) \\
 \text{constraint: } &\delta, \sigma_e^2, \sigma_{e'}^2, \sigma_p^2 > 0 \\
 &\sigma_{e'}^2 > \sigma_e^2 \\
 &0 < \theta < 1
 \end{aligned} \tag{A5}$$

Appendix B

Fit to data

Model fit was assessed by testing to what extent simulations obtained from the model fit the observed data. From each model we obtained 100 simulations (predictions). These can be found in Figure B1. The model predictions fitted the data substantially better for the mixture models. The worse fit was observed for the single distribution Gaussian model which predicted large amounts of negative keystroke intervals which are implausible. Using a log-normal distribution removed these negative keystroke intervals.

**Figure B1**

Fit to data with models (distribution type) above the plots. Comparison of 50 simulations to observed data summarised by model. For illustration the x-axis was truncated at 2 secs.

Appendix C

Transition location effect (unconstrained mixture model)

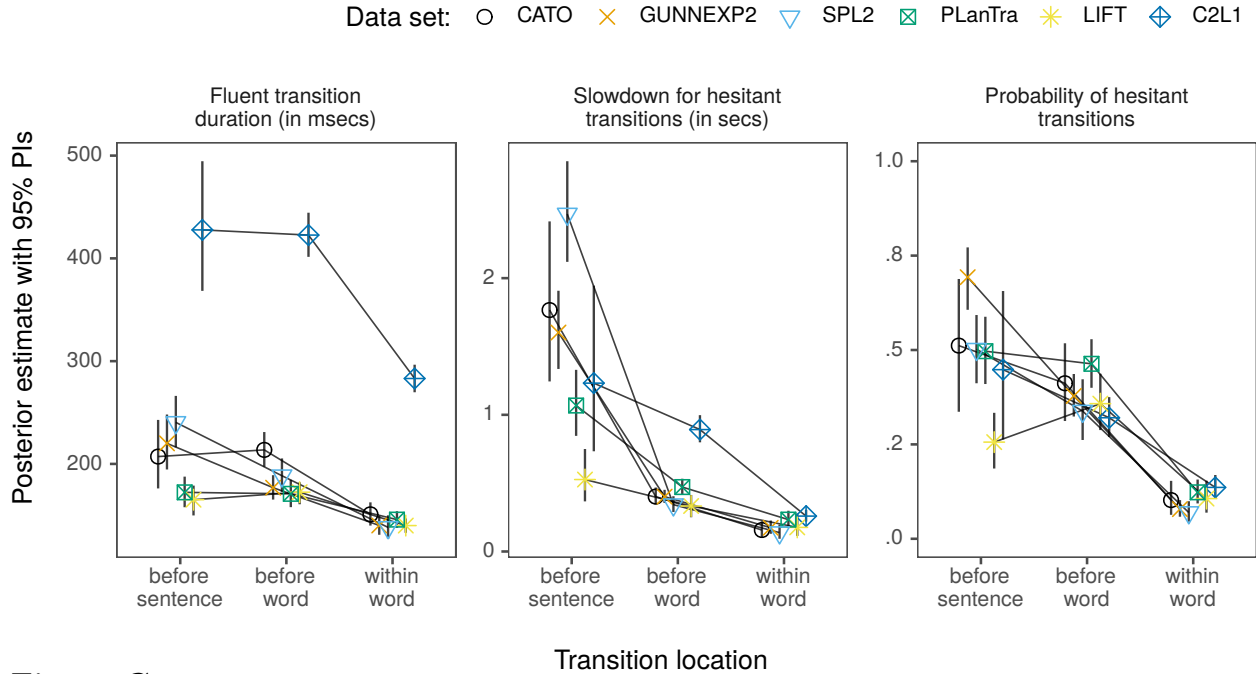
We present the mixture-model posterior estimates for each of the three parameter values in the three facets of Figure C1. Similar to the constrained mixture model, we aggregated the posterior across conditions⁷ and removed conditions that might conflate comparisons⁸. The resulting posterior allows us to examine differences between transition locations for each data set associated with each of the three mixture-model parameters.

In the following we evaluate differences between transition locations for all three mixture model parameters. Figure C1 shows largely the same patterns (with caveats) for keystroke interval estimates by transition location across data sets for all three mixture model parameters. We found that hesitations appear more frequently at before-word transitions than within words across data sets (all BFs > 100). Also hesitations are longer at before-sentence transitions compared to before-word transitions (all BFs > 100 ; except for C2L1: BF = 0.35 and LIFT = 2.14) and before-word transitions compared to within-word transitions (all BF > 10 ; except for LIFT: BF = 0.71). Further, we observed that fluent key-transitions are slower at before-word and before-sentence locations compared to within-word locations (all BFs > 100) but there is generally no evidence for a difference for fluent transitions for before-sentence transitions compared to before-word transitions (BFs < 0.08 ; except for SPL2: BF > 100 and GUNNEXP2: BF > 100). The full results of these pairwise comparisons can be found in Table C1.

Hesitation duration tends to be longer at before-sentence locations compared to before-word locations (except for data sets C2L1 and LIFT) and longer for before-word locations compared to within-word locations (except for the LIFT data set). However, for

⁷ We aggregated across pre-post test for the PLaTra dataset as well as genre and topic of the LIFT data set.

⁸ We removed the masked writing condition in the GUNNEXP2 and CATO, the dyslexic group in the CATO data set, and L2 writing in the SPL2 data set.

**Figure C1**

Mixture model parameter estimates across studies. Distributions of parameter estimates are represented as posterior mean and 95% probability interval (PI). Estimates for the CATO dataset were calculated for the non-dyslexic group, unmasked condition; also the GUNNEXP2 estimates represent the unmasked condition; SPL2 estimates are for the L1 group. Axes for transitions durations are log-scaled for visibility.

most data set there is negligible evidence for the idea that writers pause more frequently at before-sentence locations compared to before-word locations (except for data sets SPL2 and GUNNEXP2). This is interesting because it is generally believed that pausing behaviour is associated with syntactic edges such that more and longer pauses are predicted for key transitions at larger syntactic edges following the pattern before-sentence > before-word > within-word. In fact, the data set LIFT showed less pausing before sentences compared to before words. In brief, while keystroke transitions and pauses tend to be longer and more frequent at before-word locations compared to within-word transitions, it is not clear in which contexts keystroke transitions are before-sentence locations are slower, and their hesitations are longer and more frequent.

Outstanding is the overall substantially longer fluent transitions for the C2L1 data. This is presumably reflecting that the population that this sample is from was the youngest among our data sets presumably involving the least experienced writers in our data pool. Hesitation duration and frequencies were similar to the other data sets. There are some inconsistencies for fluent before-sentence transitions compared to before-word transitions. Some data sets show before-sentence slowdowns for fluent transition compared to words. These inconsistencies could, to some extent, be explained on the basis that data sets differ as to whether before-sentence transitions involve complex key combination involve the mean or sum of transitions between space, shift and / or the sentence-initial letters. In particular, some data include the character key following the shift key at before sentence location (PLanTra, LIFT) but others did not scope over the character following the shift key (CATO, C2L1, SPL2, GUNNEXP2). Notice though that for the data sets PLanTra and LIFT, there was no evidence for a consistent differences between before-sentence and before-word transitions (except for longer hesitations in the PLanTra dataset).

To address this finding, and inconsistency in how before-sentence transitions were timed, we test to what extent this difference affected the modelling results for the SPL2 data set. The results are shown in Appendix D. We found that including the character following the shift key substantially extends both the transition duration and the hesitation duration but not the hesitation frequency. However, this conflicts with the absence of differences in data sets that included the character following shift at before-sentence transitions (PLanTra, LIFT). In other words, it is unlikely that patterns in our results can be explained on the basis of how before-keystroke transitions were operationalised (complex key-combinations at before-sentence locations).

Table C1

Effect of transition location on keystroke intervals. Differences are shown on log scale (for transition durations) and logit scale for probability of hesitant transitions. 95% PIs in brackets.

Comparison	Fluent transitions		Hesitation slowdown		Hesitation probability	
	Est. [95% PIs]	BF	Est. [95% PIs]	BF	Est. [95% PIs]	BF
C2L1						
before sentence vs word	0.01 [-0.13, 0.15]	0.07	0.21 [-0.13, 0.54]	0.35	0.54 [-0.31, 1.43]	0.89
before vs within word	0.4 [0.38, 0.42]	> 100	0.49 [0.39, 0.58]	> 100	1.1 [0.76, 1.44]	> 100
CATO (non-dyslexic, unmasked)						
before sentence vs word	-0.03 [-0.18, 0.11]	0.08	1.19 [0.88, 1.49]	> 100	0.41 [-0.44, 1.27]	0.67
before vs within word	0.35 [0.31, 0.38]	> 100	0.35 [0.14, 0.54]	17.48	1.83 [1.19, 2.49]	> 100
GUNNEXP2 (unmasked)						
before sentence vs word	0.22 [0.12, 0.32]	> 100	0.93 [0.78, 1.07]	> 100	1.32 [0.86, 1.78]	> 100
before vs within word	0.23 [0.21, 0.25]	> 100	0.38 [0.21, 0.54]	> 100	1.97 [1.59, 2.37]	> 100
LIFT						
before sentence vs word	-0.04 [-0.13, 0.05]	0.07	0.35 [0.05, 0.72]	2.14	-0.49 [-1.03, 0]	1.63
before vs within word	0.21 [0.13, 0.27]	> 100	0.26 [-0.16, 0.55]	0.71	1.56 [1.02, 2.16]	> 100
PLanTra						
before sentence vs word	0.01 [-0.09, 0.11]	0.07	0.65 [0.46, 0.83]	> 100	0.13 [-0.25, 0.53]	0.25
before vs within word	0.16 [0.1, 0.21]	> 100	0.37 [0.2, 0.54]	> 100	1.83 [1.48, 2.17]	> 100
SPL2 (L1)						
before sentence vs word	0.24 [0.17, 0.31]	> 100	1.39 [1.25, 1.52]	> 100	0.69 [0.17, 1.19]	8.01

Table C1 continued

Comparison	Fluent transitions		Hesitation slowdown		Hesitation probability	
	Est. [95% PIs]	BF	Est. [95% PIs]	BF	Est. [95% PIs]	BF
before vs within word	0.31 [0.28, 0.34]	> 100	0.35 [0.14, 0.54]	16.25	1.94 [1.4, 2.5]	> 100

Note. PI = probability intervals. BF = evidence in favour of the alternative hypothesis over the null hypothesis.

Appendix D

Key-combination effect (unconstrained mixture model)

The analysed datasets differ to the extent that keystroke intervals at before sentence location sometimes did (PLanTra, LIFT) or did not (CATO, C2L1, SPL2, GUNNEXP2) scope over the character following the shift key. In other words, the pause before sentences summed across two key intervals in the PPlanTra and LIFT data, namely $_^{[shift]}C$ but only involved one key interval, namely $_^{[shift]}$ for the remaining datasets. Therefore, longer or more frequent pauses at before-sentence locations compared to before-word locations can be explained without reference to linguistic edges. Also there is a possibility that some inconsistencies in our findings can be explained on the basis of including the keystroke following shift.

Therefore we compared whether the different patterns can be explain on the basis of the additional keystroke involved in before-sentence transitions. We compared the SPL2 data including and excluding the keystroke after shift. Although we modelled all transition locations, we present only before-sentence transitions below as there was, as one would expect, no difference at word locations. The results of this comparison can be found in Table D1.

Table D1

Mixture model estimates for key transitions. Cell means are shown for transitions that do and do not involve the transition to the character following shift in msec for fluent key-transitions, the slowdown for long transitions and the probability of hesitant transitions. The difference for including the transition duration to the character after shift is shown on log scale (for transition durations) and logit scale for probability of hesitant transitions. 95% PIs in brackets.

Language	$\hat{\gamma}[\text{shift}]$	Difference	BF
Fluent transitions			
L1	159 [147, 172]	0.48 [0.34, 0.63]	> 100
L2	159 [147, 172]	0.41 [0.19, 0.63]	46.08
Hesitation duration			
L1	1,281 [1,082, 1,496]	-0.46 [-0.61, -0.3]	> 100
L2	1,802 [1,557, 2,073]	-0.34 [-0.51, -0.17]	> 100
Hesitation probability			
L1	.78 [.69, .86]	0.49 [-0.04, 1.03]	1.43
L2	.92 [.87, .96]	0.48 [-0.16, 1.13]	0.94

Note. PIs are probability intervals. BF is the evidence in favour of the alternative hypothesis over the null hypothesis.

Overall, fluent transition duration and the hesitation duration were affected by whether or not the sentence-initial transition include the character following shift. Fluent key transitions were substantially longer when including the interval following the shift key. The slowdown for hesitations was affected too but the difference is numerically small. There

was no conclusive evidence for an increased hesitation probability.

Appendix E

Key-combination effect (constrained mixture model)

The analysed datasets differ to the extent that keystroke intervals at before sentence location sometimes did (PLanTra, LIFT) or did not (CATO, C2L1, SPL2, GUNNEXP2) scope over the character following the shift key. In other words, the pause before sentences summed across two key intervals in the PPlanTra and LIFT data, namely $_{}^{\wedge}[\text{shift}]^{\wedge}\text{C}$ but only involved one key interval, namely $_{}^{\wedge}[\text{shift}]$ for the remaining datasets. Therefore, longer or more frequent pauses at before-sentence locations compared to before-word locations can be explained without reference to linguistic edges.

Therefore we compared whether the different patterns can be explain on the basis of the additional keystroke involved in before-sentence transitions. We compared the SPL2 data including and excluding the keystroke after shift. Although we modelled all transition locations, we present only before-sentence transitions below as there was, as one would expect, no difference at word locations. The results of this comparison can be found in Table E1.

Table E1

Mixture model estimates for key transitions immediately preceding a sentence. Cell means are shown for transitions that do and do not involve the transition to the character following shift in msec for the slowdown and the probability of hesitant transitions. The difference for including the transition duration to the character after shift is shown on log scale (for transition durations) and logit scale for probability of hesitant transitions. 95% PIs in brackets.

Language	$_{}^{\wedge}[\text{shift}] + \text{C}$	$_{}^{\wedge}[\text{shift}]$	Difference	BF
Fluent transitions				
overall	156 [145, 169]	159 [147, 172]	-0.02 [-0.13, 0.09]	0.06
Hesitation duration				

Table E1 continued

Language	$\hat{_}[shift] + C$	$\hat{_}[shift]$	Difference	BF
L1	1,167 [1,037, 1,305]	1,281 [1,082, 1,496]	-0.06 [-0.21, 0.09]	0.11
L2	2,019 [1,772, 2,292]	1,802 [1,557, 2,073]	0.12 [-0.02, 0.27]	0.29
Hesitation probability				
L1	1.00 [.99, 1.00]	.78 [.69, .86]	4.91 [3.68, 6.35]	> 100
L2	1.00 [.99, 1.00]	.92 [.87, .96]	3.4 [2.12, 4.85]	> 100

Note. PIs are probability intervals. BF is the evidence in favour of the alternative hypothesis over the null hypothesis.

The fluent transition duration and hesitation duration were not affected by whether or not the sentence-initial transition include the character following shift for neither language. However, we found strong evidence for an increased hesitation probability – in both – languages when the before-sentence key transition included the character following the shift key. Notably, the hesitation probability increased ceiling. It is not entirely surprised that the constrained mixture model identifies essentially all before-sentence transition as pauses: the constrained model does not distinguish between transition locations for the duration of fluent transitions. Therefore essentially all before-sentence transitions are identified as hesitation as they always include two keystrokes while all other transition location (which is the majority of the data) do not.

Appendix F

L2 effect (SPL2)

For the SPL2 data (only the `_^[shift]` sentence-transitions) we calculated the L2 effect (i.e. the difference between writing in L2 and L1). The results can be found in Table F1. The results show longer hesitation duration and more pauses across all transition locations when writing in L2. Only within-word transitions did show negligible evidence for an increased hesitation duration in L2.

Table F1

Mixture model estimates for language effect. Cell means are shown the slowdown for hesitant transitions when writing in L1 and L2 and the probability of hesitant transitions. The language difference is shown on log scale (for transition durations) and logit scale for probability of hesitant transitions. 95% PIs in brackets.

Transition location	L1	L2	Difference	BF
Hesitation duration				
before sentence	1,281 [1,082, 1,496]	1,802 [1,557, 2,073]	0.31 [0.17, 0.45]	> 100
before word	289 [254, 327]	560 [507, 617]	0.47 [0.39, 0.55]	> 100
within word	120 [53, 222]	202 [156, 258]	0.27 [-0.07, 0.55]	0.68
Hesitation probability				
before sentence	.78 [.69, .86]	.92 [.87, .96]	1.23 [0.52, 1.97]	> 100
before word	.51 [.41, .62]	.89 [.83, .93]	2.02 [1.4, 2.66]	> 100
within word	.04 [.02, .07]	.15 [.10, .22]	1.52 [0.78, 2.29]	> 100

Note. PIs are probability intervals. BF is the evidence in favour of the alternative hypothesis over the null hypothesis.

Appendix G

Masking effect (CATO, GUNNEXP2)

Studies associated with two datasets (CATO, GUNNEXP2) investigated to what extent masking the previously written text affects keystroke behaviour. The mixture model results for the effect of masking is shown in Table G1. There was some evidence for longer before-sentence hesitation durations when the text was masked in non-dyslexic writers. Evidence for all other comparisons was negligible.

Table G1

Mixture model estimates for masking effect. Cell means are shown for the masked and unmasked writing task in msec for the slowdown for long transitions and the probability of hesitant transitions. The effect for masking is shown on log scale (for transition durations) and logit scale for probability of hesitant transitions. 95% PIs in brackets.

Transition location	Dataset	Group	Unmasked	Masked	Difference	BF
Hesitation duration						
before sentence	CATO	dyslexic	2,231 [1,543, 3,111]	1,680 [1,239, 2,205]	-0.25 [-0.64, 0.13]	0.44
before sentence	CATO	non dyslexic	1,733 [1,241, 2,325]	1,502 [1,134, 1,942]	-0.12 [-0.47, 0.22]	0.23
before sentence	GUNNEXP2	non dyslexic	1,124 [954, 1,307]	1,418 [1,218, 1,634]	0.21 [0.05, 0.36]	2.59
before word	CATO	dyslexic	444 [405, 485]	407 [369, 447]	-0.06 [-0.13, 0.01]	0.14
before word	CATO	non dyslexic	345 [312, 381]	326 [295, 361]	-0.04 [-0.11, 0.04]	0.06
before word	GUNNEXP2	non dyslexic	363 [325, 403]	379 [341, 420]	0.03 [-0.05, 0.11]	0.05
within word	CATO	dyslexic	278 [166, 400]	177 [116, 252]	-0.24 [-0.52, 0.06]	0.57
within word	CATO	non dyslexic	409 [244, 593]	392 [230, 591]	-0.03 [-0.42, 0.38]	0.2
within word	GUNNEXP2	non dyslexic	182 [105, 279]	209 [135, 306]	0.08 [-0.24, 0.4]	0.19
Hesitation probability						
before sentence	CATO	dyslexic	.60 [.38, .79]	.80 [.60, .93]	0.2 [-0.06, 0.46]	0.45
before sentence	CATO	non dyslexic	.55 [.34, .76]	.68 [.47, .84]	0.13 [-0.16, 0.4]	0.21
before sentence	GUNNEXP2	non dyslexic	.88 [.82, .93]	.90 [.85, .94]	0.16 [-0.46, 0.79]	0.36
before word	CATO	dyslexic	.75 [.64, .85]	.72 [.59, .82]	-0.04 [-0.19, 0.12]	0.09
before word	CATO	non dyslexic	.64 [.51, .75]	.62 [.49, .74]	-0.02 [-0.19, 0.16]	0.09

Table G1 continued

Transition location	Dataset	Group	Unmasked	Masked	Difference	BF
before word	GUNNEXP2	non dyslexic	.47 [.40, .54]	.45 [.38, .52]	-0.06 [-0.46, 0.35]	0.21
within word	CATO	dyslexic	.07 [.03, .12]	.11 [.06, .19]	0.05 [-0.02, 0.13]	0.08
within word	CATO	non dyslexic	.03 [.02, .06]	.03 [.02, .06]	0 [-0.03, 0.03]	0.01
within word	GUNNEXP2	non dyslexic	.05 [.03, .08]	.04 [.03, .06]	-0.29 [-0.88, 0.28]	0.48

Note. PIs are probability intervals. BF is the evidence in favour of the alternative hypothesis over the null hypothesis.

Appendix H

Posterior parameter estimates

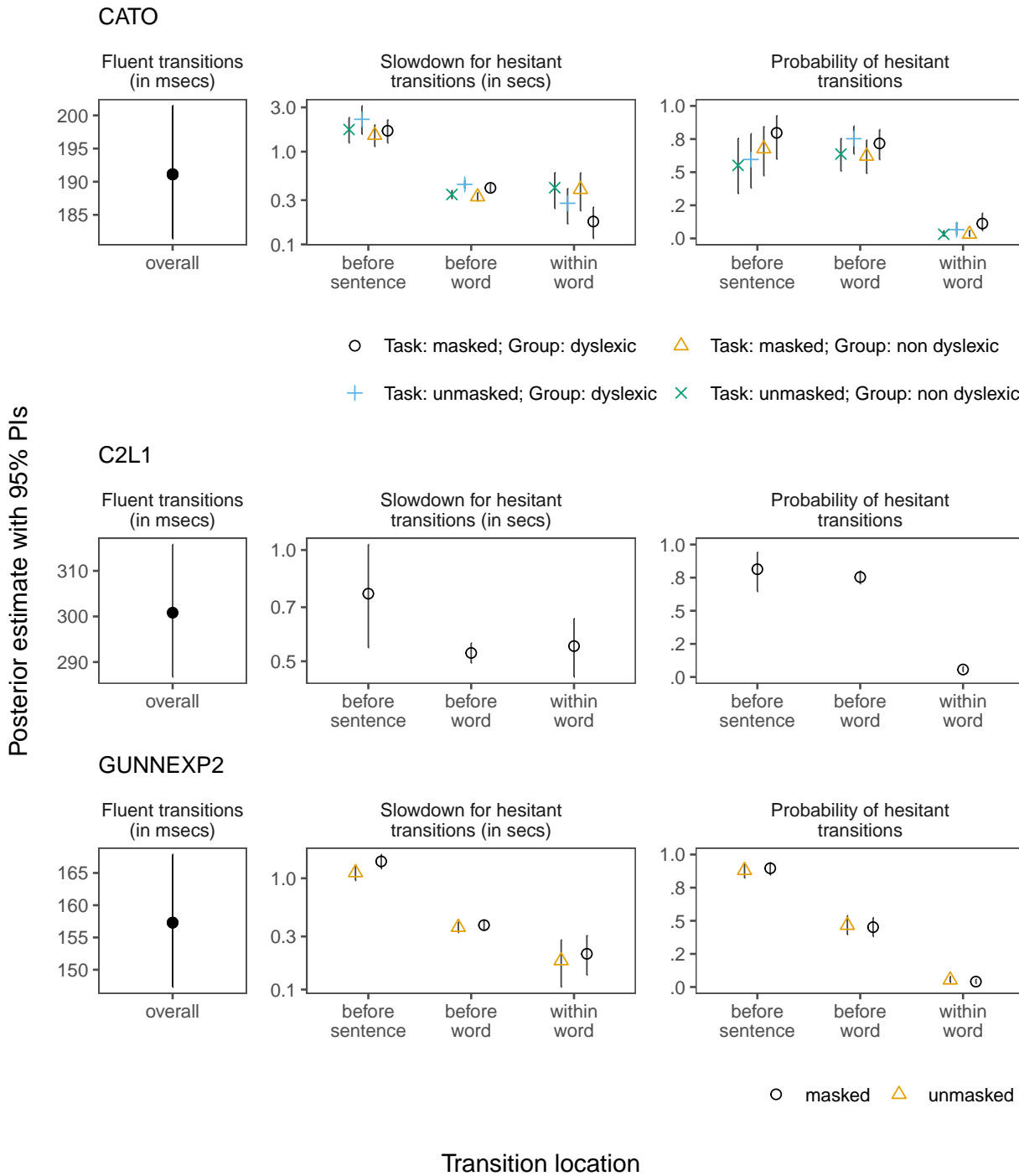
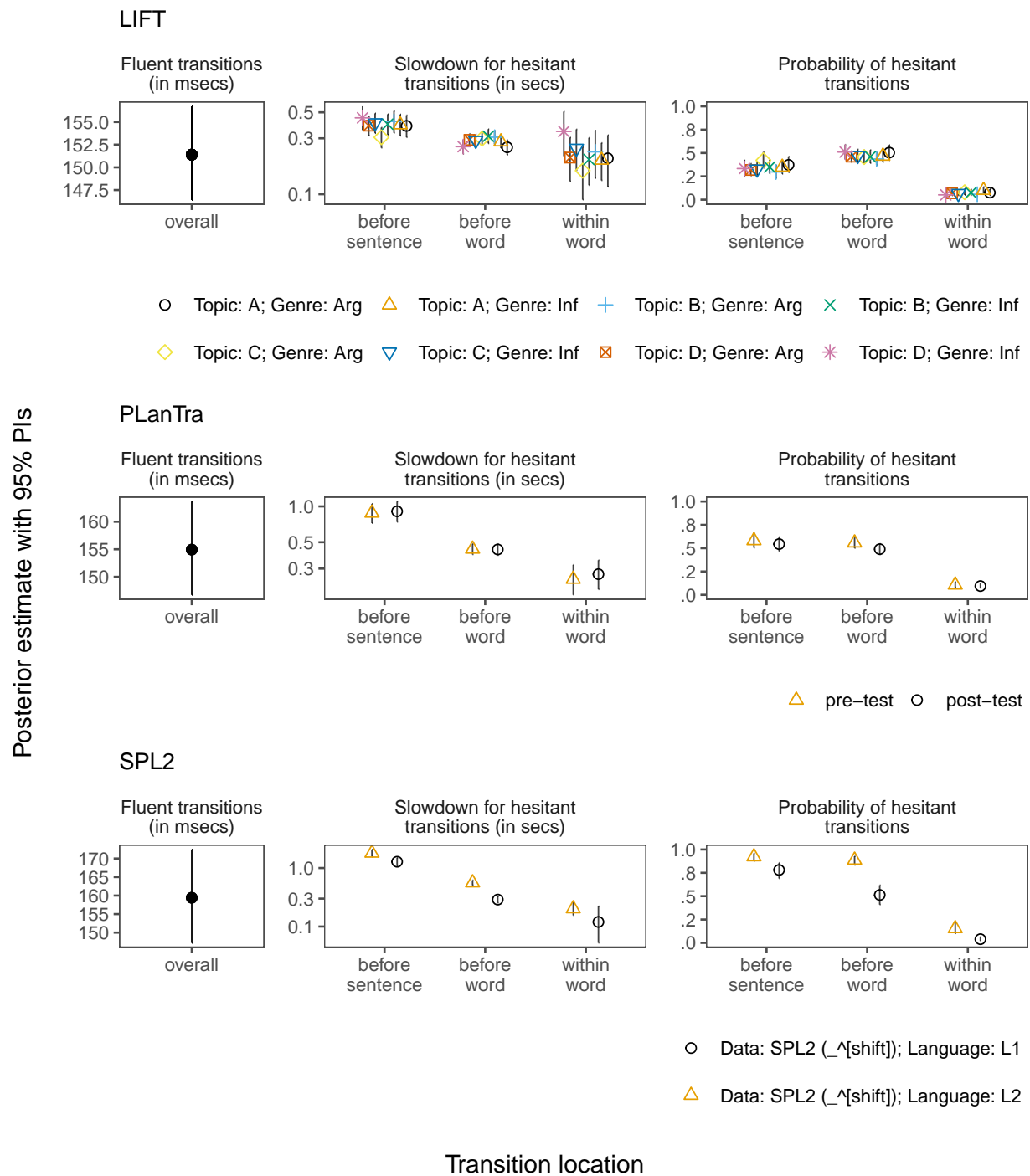


Figure H1

Posterior parameter distribution

**Figure H1** (cont.)*Posterior parameter distribution*

Appendix I

Pre-post test (PLanTra)

The pre-post test effect for the PLanTra dataset is reported in Table I1. Evidence for all comparisons was negligible.

Table I1

Mixture model estimates for post-test effect. Cell means are shown for the pre-test and post-test in msec for the slowdown for hesitant transitions and the probability of hesitant transitions. The effect for post-test is shown on log scale (for transition durations) and logit scale for probability of hesitant transitions. 95% PIs in brackets.

Transition location	Pre-test	Post-test	Difference	BF
Hesitation duration				
before sentence	879 [723, 1,052]	909 [740, 1,100]	-0.03 [-0.23, 0.17]	0.1
before word	438 [395, 485]	434 [389, 483]	0.01 [-0.08, 0.1]	0.05
within word	245 [180, 323]	269 [201, 353]	-0.06 [-0.3, 0.18]	0.14
Hesitation probability				
before sentence	.58 [.50, .66]	.54 [.47, .62]	0.15 [-0.26, 0.58]	0.27
before word	.56 [.50, .61]	.49 [.43, .55]	0.27 [-0.04, 0.59]	0.72
within word	.11 [.08, .14]	.09 [.07, .12]	0.13 [-0.27, 0.54]	0.26

Note. PIs are probability intervals. BF is the evidence in favour of the alternative hypothesis over the null hypothesis.

Appendix J

Genre effect (LIFT)

From the LIFT data, we compared the difference between genres, i.e. when writing an informative text as opposed to writing an argumentative text. The results are shown in Table J1. Cellmeans and differences were average across writing topic. Evidence for a genre comparisons was negligible.

Table J1

Mixture model estimates for genre effect. Cell means are shown for argumentative and informative texts in msec for the hesitation slowdown and the probability of hesitant transitions. The effect for genre is shown on log scale (for transition durations) and logit scale for probability of hesitant transitions. 95% PIs in brackets.

Transition location	Argumentative	Informative	Difference	BF
Hesitation duration				
before sentence	373 [301, 454]	411 [330, 504]	-0.07 [-0.33, 0.19]	0.16
before word	287 [250, 326]	284 [248, 323]	0.01 [-0.15, 0.16]	0.09
within word	199 [118, 306]	246 [155, 361]	-0.12 [-0.6, 0.36]	0.29
Hesitation probability				
before sentence	.35 [.28, .43]	.34 [.27, .42]	0.05 [-0.57, 0.73]	0.36
before word	.46 [.39, .54]	.48 [.41, .55]	-0.05 [-0.51, 0.44]	0.25
within word	.07 [.05, .10]	.07 [.05, .10]	-0.02 [-0.81, 0.76]	0.46

Note. PIs are probability intervals. BF is the evidence in favour of the alternative hypothesis over the null hypothesis.

Appendix K

Transition location effect (constrained mixture model)

Table K1

Effect of transition location on keystroke intervals. Differences are shown on log scale (for durations) and logit scale for hesitation probability. 95% PIs in brackets.

Comparison	Hesitation slowdown		Hesitation probability	
	Est. [95% PIs]	BF	Est. [95% PIs]	BF
C2L1				
before sentence vs word	0.24 [0.02, 0.48]	1.11	0.44 [-0.56, 1.66]	0.68
before vs within word	-0.03 [-0.14, 0.09]	0.07	3.95 [3.54, 4.38]	> 100
CATO (non-dyslexic, unmasked)				
before sentence vs word	1.27 [0.98, 1.55]	> 100	-0.36 [-1.41, 0.69]	0.67
before vs within word	-0.1 [-0.38, 0.21]	0.2	4.05 [3.2, 4.91]	> 100
GUNNEXP2 (unmasked)				
before sentence vs word	0.9 [0.76, 1.04]	> 100	2.15 [1.6, 2.75]	> 100
before vs within word	0.44 [0.17, 0.69]	26.78	2.76 [2.24, 3.28]	> 100
LIFT				
before sentence vs word	0.21 [-0.05, 0.48]	0.5	-0.53 [-1.04, 0.08]	1.61
before vs within word	0.17 [-0.32, 0.53]	0.41	2.48 [1.81, 3.22]	> 100
PLanTra				
before sentence vs word	0.57 [0.4, 0.74]	> 100	0.16 [-0.25, 0.57]	0.28
before vs within word	0.36 [0.16, 0.56]	55.53	2.3 [1.9, 2.7]	> 100
SPL2 (L1)				
before sentence vs word	1.17 [1.02, 1.3]	> 100	1.23 [0.59, 1.88]	> 100
before vs within word	0.48 [0.16, 0.76]	9.07	3.3 [2.56, 4.08]	> 100

Note. PI = probability intervals. BF = evidence in favour of the alternative hypothesis over the null hypothesis.

Appendix L

Simulation

A possible concern with these results – substantially better predictive performance for two-distribution mixture models – is that, in principle, as the mixture model has more parameters it might always lead to a better fit. We addressed this concern before by using cross-validation techniques for model comparison which is preventing overfitting models. To address this concern we repeated a comparison for a single distribution model and a two-distributions mixture model for two sets of simulated data that were simulated either with a single distribution or two-distributional process at heart. In other words, this allows us to test the predictive performance of our models in a context where we know the true underlying data generating process – single distribution vs two distributions – and we can test whether these models can successfully uncover the true parameter value.

In particular, the first data set was simulated using a two-distributions mixture model with two mixture components similar to the process described above (equations in A4 and A5). This process and the corresponding Bayesian model is summarised in equation L1.

$$\begin{aligned}
 y &\sim \theta \cdot \log\mathcal{N}(\beta + \delta, \sigma_1^2) + \\
 &\quad (1 - \theta) \cdot \log\mathcal{N}(\beta, \sigma_2^2) \\
 \text{constraint: } &\delta, \sigma_2^2, \sigma_1^2 > 0 \\
 &\sigma_1^2 > \sigma_2^2
 \end{aligned} \tag{L1}$$

This model is largely identical to the models before but reduced to its main parameters (but not mixed effects for participants). The model includes two log-normal distributions with a mixing proportion θ of which the distribution of shorter values has a mean of β and a standard deviation σ_2^2 ; the second distribution of longer values is constrained to have a mean that is larger by a factor of δ and has a larger standard deviation σ_1^2 .

The second data set was generated with a single distribution log-Gaussian process. The model and its corresponding Bayesian model is summarised in equation L2.

$$\begin{aligned} y &\sim \log\mathcal{N}(\beta, \sigma^2) \\ \text{constraint: } \sigma^2 &> 0 \end{aligned} \tag{L2}$$

Again, this model is a simplified version of the single distribution models used in the main analysis. The model assumes a log-Gaussian distribution with a mean parameter β and a standard deviation σ^2 .

The parameter values used for each of the two data simulations can be found in Table L1. The simulated data are visualised in Figure L1. Parameter values were chosen so that the simulated data are roughly similarly distributed to keystroke transitions.

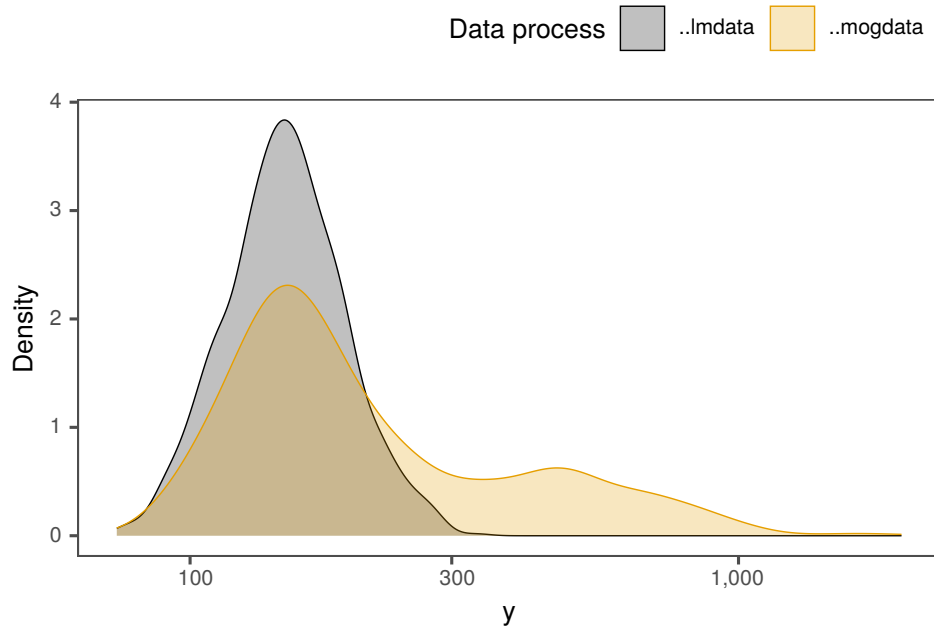


Figure L1

Data simulated with a two-distribution (grey) and a single-distribution (yellow) random data generating process. The x-axis showing the outcome y was log-scaled for visibility.

For each of these two data sets we simulated 1,000 observations. We run 2 models – a

Table L1

Uncovered parameter estimates with 95% probability interval (PI) and true parameter values for each simulated data set and by model and their respective parameters.

Parameter with true value	Estimate with 95% PI	
	Two-distributions model	Single-distributions model
Two-distributions data		
$\beta = 5$	5.02 [4.99, 5.04]	4.93 [4.76, 5.01]
$\delta = 1$	1.04 [0.92, 1.14]	0.13 [0.01, 0.3]
$\theta = .35$.33 [.28, .39]	.54 [.11, .92]
$\sigma_1^2 = 0.25$	0.25 [0.23, 0.27]	0.22 [0.16, 0.25]
$\sigma_2^2 = 0.5$	0.49 [0.42, 0.57]	0.25 [0.22, 0.3]
Single-distribution data		
$\beta = 5$	5.35 [5.32, 5.39]	5 [4.99, 5.02]
$\sigma = 0.25$	0.6 [0.57, 0.63]	0.25 [0.24, 0.26]

two-distribution mixture model and a single-distribution model – each for both data sets. Models were run with 3 chains, with each 6,000 iterations of which 3,000 were warmup. Estimates with 95% probability intervals are shown in Table L1. The parameters are shown by type of data generating process along with the true parameter values. Parameter value estimates are shown by type of Bayesian model. The results show that either of the two Bayesian models successfully uncovered the model parameters of the data with its corresponding data generating process, as shown in Table L1, but less so when the model was applied to data generated with the incorrect underlying process. Particularly the mixing proportion θ and the slowdown parameter δ were not uncovered at all by the single-distribution model.

We used LOO-CV to compare the predictive performance of the two models for each

data generating process. The model comparisons can be found for each data generating process in Table L2. The results rule out the possibility that the mixture model does always lead to higher predictive performance. Indeed, the mixture model showed a slightly lower predictive performance for the data that were generated with a single-distribution process. However, for the data generated with a two-distribution mixture process, the mixture model shows a substantially higher predictive performance. In fact, the ratio of $\widehat{\Delta elpd}$ and its standard error – a z -score approximation – reveals a value of 11.6 which is strong evidence in in favour of the mixture model compared to the single-distribution model (Sivula et al., 2020). In comparison, for the single-distribution data, the single-distribution model was negligibly better than the two-distribution mixture model which was evidence by a value of 0.77 for the ratio of $\widehat{\Delta elpd}$ and its standard error. Thus, as the difference $\widehat{\Delta elpd}$ between the model is negligible, the single distribution is preferred by the law of parsimony.

Table L2

Model comparisons by data set. The top row shows the models with the highest predictive performance for each data generating process.

Standard error is shown in parentheses.

Model	$\Delta\widehat{elpd}$	\widehat{elpd}
Data: Two-distribution mixture process		
Two-distribution mixture model	–	-6,068 (41)
Single-distribution model	-191.3 (16.5)	-6,259 (38)
Data: Single-distribution process		
Single-distribution model	–	-5,030 (24)
Two-distribution mixture model	-0.5 (0.7)	-5,030 (24)

Note. \widehat{elpd} = predictive performance indicated as expected log pointwise predictive density; $\Delta\widehat{elpd}$ = difference in predictive performance relative to the model with the highest predictive performance in the top row.