Modelling disfluencies in copy-typing as mixture process

Jens Roeser[1], Sven De Maeyer[2], Mark Torrance[1], Luuk Van Waes[3], & Mariëlle Leijten[3]

[1] Department of Psychology, Nottingham Trent University, United Kingdom
[2] Faculty of Social Sciences, University of Antwerp, Belgium
[3] Department of Management, University of Antwerp, Belgium

The analysis of keystroke timing data typically involves the calculation of summary statistics such as the mean inter-keystroke interval, pause frequencies etc. There are two fundamental problems with this: first, parametric descriptives for non-normal distributed data are going to be biased estimates and neglect information associated with variance; second, pauses and related process measures (e.g. bursts) are often defined using threshold value which are, in principle, arbitrary. We implemented a series of Bayesian models that aimed to address both issues by (a) providing reliable typing estimates and (b) statistically detecting process disfluencies. We tested these models on a random sample of 100 participants from the Dutch copy-task corpus. Our results illustrate how typing disfluencies can be statistically determined as a finite mixture process. The typing task was characterized by this mixture model as a combination of fluent and disfluent typing associated with parameters that capture the magnitude and the probability of disfluencies. Mixture models provide a principle method to detect disfluencies in keyboard typing data.

Introduction

Writing research has made extensive use of keystroke-logging to capture typing process data. In particular process disfluencies (loosely defined as relatively long intervals between subsequent keystrokes) are interesting to develop and understanding of the individuals writing progress. This is because language production is typically thought of as a cascade from the mental generation of a message, into grammatical processing and finally the generation and execution of motor codes that serve the transition of an idea. This can be found in theoretical models of speech (Bock & Ferreira, 2014), handwriting (Van Galen, 1991) and keyboard typing (Hayes, 2012). Disfluencies at the execution stage are therefore indicators of processing demands that arise on higher levels of mental representation (Christiansen & Chater, 2016; Olive, 2014); for example, when preplanning syntactic dependencies (Roeser, Torrance, & Baguley, 2019), retrieving a lexical entry for a word or its spelling (Torrance, Rønneberg, Johansson, & Uppstad, 2016). At present there is no prin-

cipled way of detecting keystroke lags that constitute a process disfluency. In this paper we present a series of statistical models that aimed at capturing the typing process and in particular process disfluencies.

Keystroke logs provide rich information about the typing process. From this log, researchers can calculate different process measures including measures of writing fluency (Chukharev-Hudilainen, Saricaoglu, Torrance, & Feng, 2019). To name a few researchers have performed data analysis on means, medians, standard deviations (SD) etc. of inter-keystroke intervals (the duration between two consecutive keystrokes), number of pauses or pause duration, within-word keystroke intervals and many other variables (for an overview see Conijn, Roeser, & van Zaanen, 2019). Conijn et al. (2019) suggested that these aggregates are sensitive to processing difficulty that arise on different levels of mental representation. However, using summary statistics for statistical inference comes with two central problems.

First, pause frequency, bursts and related measures (Alves & Limpo, 2015) requires a definition what constitutes a pause (Wengelin, 2006), i.e. a threshold value (Chukharev-Khudilaynen, 2014; Connelly, Dockrell, Walter, & Critten, 2012; Leijten & Van Waes, 2013) such as 2 secs (Kaufer, Hayes, & Flower, 1986). Such a threshold is to some extent arbitrary as normal and disfluent typing intervals are relative to task demands and the skills of the typist. For exam-

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

ple, when comparing the frequency of pauses larger than 2 secs for dyslexic and normal typists one might observe more pauses for the dyslexic group because a) 2 secs are indeed not unusual typing intervals for dyslexic typists or b) because pauses for normal typists are shorter than 2 secs and therefore unobserved given a 2 secs lower bound (Wengelin, 2001).

Second, parametric aggregates such the mean and the SD assume that, for the summary statistic to be representative for the sample, the data must come from a normal distribution. This is not the case as keystroke intervals are zero-bound and therefore right-skewed.[1] Hence, aggregates are biased estimates of the typing process and therefore may lead to incorrect inference based on biased parameter values. To prevent biased parameter values, i.e. to ensure normal distributed data, summary statistics involve data trimming (Hoaglin & Iglewicz, 1987) to remove data that were a priori considered outliers. Again, struggling typists might exceed trimming criteria more frequently than healthy typists. Also potentially important information about disfluencies and time course variation and therefore the writing process are being removed in the trimming and aggregation process.

A central methodological challenge with implications for writing research (Hayes, 2012; Kaufer et al., 1986; Wengelin, 2006) is the detect writing disfluencies. We addressed this problem by implementing statistical models that aim to capture the nature of the data generating process (i.e. keyboard typing). Crucially we want these models to provide reliable estimates to an individuals typing performance without subjecting the data to trimming criteria or threshold values.

Modelling typing process data

We implement statistical models that aim to capture these data. As a guiding principle, we aim to produce statistical models that are in line with the mental theoretical process that creates the data. The typing process data are the lags between subsequent keypresses (the inter-keystroke interval; IKI), for example, the lags c$^\wedge$a$^\wedge$t for the word *cat* where $^\wedge$ indicates the typing difference between pressing <c> and <a>. Further, our models aim to provide a systematic way of addressing process variations / disfluencies. For example when the time for pressing <a> is unusually large, our models should make notice of this.

Statistical models can be used to characterize an underlying data generating process as a function with parameter values. For example, if we assume that the process that we are interested in can be described as normal distribution, we need a value μ that indicates the centre of the distribution and a scale σ^2 (i.e. the variance) to draw this distribution. The values that parameters μ and / or σ^2 take on are unknowns that can vary across task demands, population, etc. This model can be written as $y \sim Normal(\mu, \sigma^2)$; the data y come from a

process that follows a normal distribution with an unknown mean μ and an unknown error variance σ^2.

We assume throughout that IKIs can be characterized as log-normal distributed because IKIs are zero-bound (Baayen, 2008), as in equation 1.

$$y_{ij} \sim LogNormal(\mu + u_i + w_j, \sigma_e^2) \qquad (1)$$

Equation 1 is an extension of the example above, which is generally referred to as linear mixed effects models (LMM) which has been used to model keystroke data (Quené & Van den Bergh, 2004; Van Waes, Leijten, & Quinlan, 2010; Van Waes, Leijten, Roeser, Olive, & Grabowski, 2020). The advantage of LMMs is that the parameters of interest, the mean μ, can be estimated while accounting for variance that is unknown and random. For example, some participants are faster typists than others. The variance associated with the is participant, expressed as u_i, can be assumed to be normal distributed around 0 with a between participants variance σ_u^2 with $i = 1, \ldots, I$, where I is the number of participants (see 2). We will come back to w_J below.

$$u_i \sim Normal(0, \sigma_u^2)$$
$$\sigma_u \sim Normal(0, 2.5) \qquad (2)$$
$$\text{constraint: } \sigma_u > 0$$

For model fit purposes, we parameterised the mean μ as non-centred with weakly regulating priors (Lambert, 2018) as shown in equation 3.

$$\mu = \mu_\alpha + \mu_\sigma * \mu_\eta$$
$$\mu_\alpha \sim Normal(5, 4)$$
$$\mu_\sigma \sim Normal(0, 10) \qquad (3)$$
$$\mu_\eta \sim Normal(0, 1)$$
$$\text{constraint: } \mu_\sigma > 0$$

For the unexplained variance σ_e^2 we used a uninformative half-Cauchy prior (equation 4; Gelman et al., 2014).

$$\sigma_e \sim Cauchy(0, 2.5)$$
$$\text{constraint: } \sigma_e > 0 \qquad (4)$$

We used weakly informative priors throughout to aid model convergence by reducing the parameter space to relatively plausible values. All models presented below share these properties as listed before but differ as to how (1) variation between IKIs and (2) disfluencies are addressed.

---

[1]In fact, the minimum size of keystroke intervals is determined by the time it takes to plan and execute the motor program.

A standard approach to model the variation between individual keystroke pairs (i.e. bigrams), as shown before in equation 1, is to assume that each bigram j with $j = 1, \ldots, J$, where I is the total number of bigrams, is independent of the other bigrams and associated with a unknown variance w_j that is distributed around 0 with a between bigram variance σ_w^2 as in equation 5 (see Van Waes, Leijten, Pauwaert, & Van Horenbeeck, 2019).

$$w_j \sim Normal(0, \sigma_w^2)$$
$$\sigma_w \sim Normal(0, 2.5) \qquad (5)$$
$$\text{constraint: } \sigma_w > 0$$

Further, we can extent this model by assuming that larger variations in typing differences for bigrams depend on the typing speed of each participant. For example, fast participants might show less variation between bigrams that slow participants. This assumption can be modelled by including by-participant slope adjustments for bigrams by introducing a variance-covariance matrix Σ_u; LKJ prior with $\nu = 2.0$ (Lewandowski, Kurowicka, & Joe, 2009).

Typing as autoregressive process

The standard analysis can model variation that is associated with particular bigrams but assumes that disfluencies are subject to random noise. Further, the standard analysis assumes that subsequent keystrokes are independent and thus exchangable. If we assume that subsequent keystrokes are autocorrelated (Eltahir, Salami, Ismail, & Lai, 2004), a disfluency can be characterised as a slowdown relative to a previous keystroke. In other words, we can predict an IKI_i with the previous keystroke IKI_{i-1}. This is called an autoregressive process. Autocorrelation is captured by the parameter ϕ that predicts an IKI y_j on the basis of the IKI preceding it y_{j-1}; see equation 6. ϕ_i was estimated separately for each participant with a pooled mean ϕ and error variance η^2. Note, to allow comparisons between this and all other models we had exclude the first bigram for each participant, i.e. y_{ij+1}.

$$y_{ij} \sim LogNormal(\mu + \phi_i * log(y_{ij-1}) + u_i, \sigma_e^2)$$
$$\text{where}$$
$$\phi_i \sim Normal(\phi, \eta^2)$$
$$\phi \sim Normal(0, 1) \qquad (6)$$
$$\eta \sim Cauchy(0, 1)$$
$$\text{constraint: } \eta > 0$$

Typing as (autoregressive) mixture process

Disfluencies during the typing process can be captured in finite mixture models. Mixture models assume that data come from a combination of distributions. For the present purpose we constrain the model to be finite. In other words, we fixed the number of underlying distributions to two, namely 2 log-Gaussian (normal) distributions, of which one represents fluent tying (shorter IKIs) and the other represents disfluencies (longer IKIs). This model can be summarised as in equation 7, following Vasishth, Chopin, Ryder, and Nicenboim (2017). The first and second line are the sum of two log-normal distributions of which the first distribution has a mixing proportion (weight) θ and the other distribution receives the remaining proportion $1 - \theta$. Both distributions have the same mean μ but the parameter δ that added to the first distribution and constrained to be positive. Thus, δ captures the magnitude of the disfluency. The mixing proportion θ_i, then, captures the probability of disfluent IKIs for each participant i.

$$y_{ij} \sim \theta_i \cdot LogNormal(\mu + \delta + u_i + w_j, \sigma_{e'}^2) +$$
$$(1 - \theta_i) \cdot LogNormal(\mu + u_i + w_j, \sigma_e^2)$$
$$\text{where} \qquad (7)$$
$$\delta \sim Normal(0, 1)$$
$$\text{constraint: } \delta > 0$$

The hyper-parameter θ captures the population disfluency probability (with an error variance τ^2) as shown in equation 8. The mixing proportion θ_i was transformed to range from 0 to 1 (inverse logit) where a value of 0 would indicate fluent typing and 1 indicates disfluency.

$$\theta_i = InvertedLogit(\theta_i)$$
$$\theta_i \sim Normal(\theta, \tau^2)$$
$$\theta \sim Normal(0, 1) \qquad (8)$$
$$\tau \sim Cauchy(0, 1)$$
$$\text{constraint: } \tau > 0$$

As 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), the variance $\sigma_{e'}^2$ associated with the distribution of typing disfluencies was constrained to be larger than the variance for normal typing σ_e^2 as shown in 9 (see Vasishth et al., 2017; Vasishth, Jäger, & Nicenboim, 2017).

$$\sigma_{e'} = \sigma + \sigma_{\text{diff}}$$
$$\sigma_e = \sigma - \sigma_{\text{diff}}$$
$$\sigma_{\text{diff}} \sim Normal(0, 1) \qquad (9)$$
$$\sigma \sim Cauchy(0, 2.5)$$
$$\text{constraint: } \sigma, \sigma_{\text{diff}}, \sigma_{e'}, \sigma_e > 0$$

Note that the mixture model, as well as the LMM, implies that subsequent keystroke intervals are independent. This might be the case for disfluencies but subsequent IKIs in fluent typing might involve autocorrelations. Therefore, we implemented another mixture model but replaced the bigram intercepts w_j, in the distribution that represents fluent typing in model 7, with an autoregressor $\phi_i * y_{ij-1}$, as in equation 6; random bigram intercepts were kept for the distribution of disfluent typing intervals.

Method

We applied these models to data from a subset of the Dutch copy-task corpus (Leijten & Van Waes, 2013; Van Waes et al., 2019; Van Waes et al., 2020) to test which model captures the typing process best. The copy-task corpus consists of keystroke data collected in Inputlog, a Javascript-based web application available on www.inputlog.net. In a set of different contexts participants have to produce keyboard typed responses. In this analysis we focus on the consonant task. Participants saw and copy-typed four blocks of six consonant sequences "tjxgfl pgkfkq dtdrgt npwdvf" that do not occur adjacently in Dutch. This task allows us to measure typing skills in a non-linguistic environment (Grabowski, Weinzierl, & Schmitt, 2010). Importantly for the present purpose, the task demands require participants to repeatedly look at the target string to copy it accurately. In other words, for this task we need to be able to disentangle fluent and disfluent IKIs. We used a random sample of 100 participants (78 females, 22 males) in the age range between 18 and 25 years (median age = 22 years) to keep the sample relatively homogeneous.

Results

Data overview

The raw data are visualized in Figure 1. Each grey line represents one participant and the coloured lines shown different measures of central tendency. This figure highlights that the consonants task is a combination of fluent and disfluent typing. This combination is expressed in the up-and-down pattern but also in the bimodality of the density function in the right margin. Further using central tendency descriptives highlight different properties of the distribution. The mean is sensitive to larger IKIs and is therefore not at the peak of the distribution while the mode and the median are relative close to the peak. All central tendencies vary depending on the position of the bigram in the string.

Model fit

All models were implemented as Bayesian models (see e.g. Gelman et al., 2014; Lambert, 2018; McElreath, 2016) in the
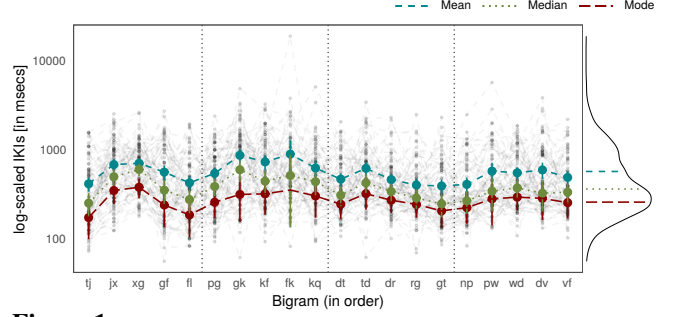
Figure 1

IKI data by participant over bigrams position shown in order. Shown are the raw data, each line representing one participant, and different measures of central tendency (with $2\times$ standard error [SE]). Vertical lines indicate the position of SPACES. The density distribution of IKI data is shown in the right margin.

probabilistic programming language Stan (Carpenter et al., 2016; Hoffman & Gelman, 2014; Stan Development Team, 2015a, 2015b). *R* and *Stan* code are available on GitHub. Models were fitted with 30,000 iterations (15,000 warm-up) on 3 Markov chain Monte Carlo chains. Convergence was tested via the Rubin-Gelman statistic (Gelman & Rubin, 1992), traceplots and cross-validation (Vehtari, Gelman, & Gabry, 2015, 2017).

The predictive performance of the models was established using leave-one-out cross-validation which penalizes models with more parameters (see Farrell & Lewandowsky, 2018; Lambert, 2018; Lee & Wagenmakers, 2014; McElreath, 2016). The out-of-sample predictive performance was determined via Pareto smoothed importance-sampling (Vehtari et al., 2015, 2017) and estimated as sum of the expected log predictive density ($\widehat{elpd}$). $\widehat{elpd}$ was used to compare the predictive quality of our models. Model comparisons can be found in Table 1. The mixture model M4 (see equation 7) revealed the highest predictive performance.

The second best performing model is the mixture model with the autoregressor *phi* for fluent typing. In other words, adding the autoregressor instead of random bigram intercepts for fluent typing did not improve the predictive performance of the model. In fact, the autoregressive model was found to be the model with the lowest predictive performance. Modelling bigrams as random intercepts (with and without by-participant slope adjustments) was found to have a higher predictive performance compared to the autoregession model (Van Waes et al., 2010, 2020).

Table 1

Model comparisons expressed as expected log predictive density ($\widehat{elpd}$). The bigrams column indicates how the models account for by-bigram variation. The top row shows the model with the highest predictive performance. Differences in predictive performance are shown as $\Delta\widehat{elpd}$. Standard errors (SE) are shown in brackets.

Model	Type	$\Delta\widehat{elpd}$	$\widehat{elpd}$	Bigrams
M4	MoG	0 (0)	-13857 (59)	Random intercepts
M5	AR + MoG	-14 (14)	-13871 (57)	Autoregressor φ or random intercepts
M2	LMM	-87 (24)	-13944 (53)	Random intercepts and slopes
M1	LMM	-153 (18)	-14010 (54)	Random intercepts
M3	AR	-223 (21)	-14080 (54)	Autoregressor φ

Note. LMM = linear mixed effects models; AR = Autoregressive model; MoG = Mixture of (log-)Gaussians

Parameter evaluation

The copy-typing process captured by the mixture model can be characterized with the posterior distributions of the model's parameter values. The process relevant parameters are illustrated in Figure 2. Firstly, IKIs, excluding disfluencies, are shown by participant in Figure 2A. The red line indicates the pooled overall parameter estimate for fluent typing $\hat{\beta}$=259 msecs centred around a 95% PI of [235, 286]. For each participant the probability of disfluent typing is shown in Figure 2B. The overall disfluency probability (in red) was $\hat{\theta}$=0.73 centred around 95% PI[0.63, 0.83]. In other words, for the consonant task we observe 73% disfluent typing and 27% fluent typing.

The y-axis in panel A and B are ordered by the average size of the respective values, thus the lines do not represent the same participants. In fact, Figure 2C suggests that the inferred latency for fluent typing and the probability to exhibit disfluencies are independent. Shown are the parameter estimates for each participant and the overall pooled estimates. In other words, fast as well as slow copy-typists can show low and high disfluency probabilities. Finally, the slowdown for disfluent typing is shown in Figure 2D. Disfluent typing in the consonant task is $\hat{\delta}$=297 msecs (95% PI[251, 347]) slower than fluent typing.[2] Overall, for the consonant task it is crucial to distinguish between fluent and disfluent typing because disfluencies are indeed more common. In other words, the task demands do not inhibit keystroke intervals throughout the task but a large proportion of produced bigrams. Mixture models can provide accurate estimates for fluent typing while account for disfluencies by modelling fluent and disfluent typing as a mixture process.

Discussion

Our aim was to provide a statistical model of inter-keystroke intervals that addresses process disfluencies in a principled manner. We compared a series of Bayesian models addressing this aim. Model comparisons showed that process disfluencies can be captured as a mixture process. This model allows us to extract reliable typing-interval estimates for fluent typing while accounting for process disfluencies by modelling fluent and disfluent typing as a combination of two distributions with a latent mixing ratio.

This model provides a) overall and by-participant keystroke intervals for fluent typing, b) overall and by-participant disfluency probabilities, and c) an parameter estimate that captures the overall slowdown for disfluencies. These parameter estimates are relevant on two levels. First, they allows us to characterize the writing task at hand. For example, we observed that copy-typing non-lexical strings of consonants shows indeed a larger proportion of disfluent compared to fluent typing. Second, by-participant parameter estimates all us to extract typing characteristics for individual typists. In particular, we extracted each participants fluent typing speed and the probability of disfluencies exhibited by each participant. The probability of disfluencies in this task can be understood as measures of memory span. If participants with a smaller memory span look more often to the target string, they will show a larger proportion of typing disfluencies. Taken together the overall and individual parameter estimates, we can determine whether an individual was a fast / slow typist or had unusually high / low probability to exhibit disfluencies compared to the population estimates.

---

[2]The size of the disfluency δ may vary by participant. For example, faster participants might show larger disfluencies magnitudes. We addressed this possibility by allowing the disfluency parameter δ to vary for each participant. This model revealed a negligible gain over model M4 ($\Delta\widehat{elpd}$ =-3, SE=2).
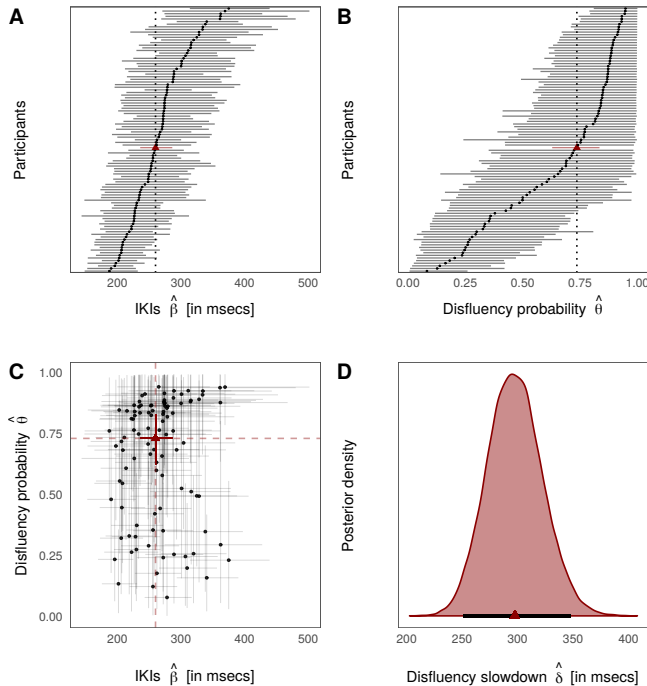
Figure 2

Posterior parameter values of the mixture model. Panel A shows by-participant IKIs and overall IKI value for fluent typing in red. Panel B shows by-participant disfluency probability (overall parameter value θ in red). Panel C shows fluent IKIs plotted against disfluency probability (red triangle indicates overall parameter value). Panel D shows the posterior distribution of the disfluency slowdown. All error bars are 95% probability intervals.

The central advantage of using mixture models to account for typing disfluencies is that we can by-pass the use of threshold values to define disfluencies. From the raw data it is not possible to know which data are disfluencies. Using threshold values ignores that some participants are generally slower typists and some tasks are more difficult. Mixture models allow us to capture disfluencies as a latent process in a principled way. This is important because the mixture model presented in this paper takes into account that a disfluency is relative to an individuals typing speed. Therefore, these models allow us to test predictions about typing disfluencies in certain population such as very young or old typists, language learners or individuals with genuine typing difficulty after account for individual differences in typing speed or vice versa. Further while we used copy-task data, other writing tasks will lead to other disfluency estimates. In other words, the presented model can be used to test hypothesis about psychological factors that might affect the ratio of disfluencies in the writing process. As an avenue for future research, mixture models as presented in this paper can be used for different types of writing tasks and particular populations.

Writing involves processing on various cognitive levels, from generating an idea, retrieving lexical entries to the encoding of spelling. As activation between these levels cascades from higher to lower levels of representation, a delay at any of these levels causes disfluencies. While we distinguished between fluent and disfluent typing in a binary way, processing difficulty on different levels of activation might be associated with a different disfluency magnitude and might be cumulative. If the size of the disfluency is assumed to depend on the inhibited process upstream or a combination of two processes, this can be implemented as additional mixture component (Baaijen, Galbraith, & Glopper, 2012) for address different types of disfluencies (Wengelin, 2001). In other words extensions of mixture models allow us to test different assumptions about the nature of the cascade of processes associated with writing in particular and language production in general.

References

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., & Glopper, K. de. (2012). Keystroke analysis: Reflections on procedures and measures. *Written Communication*, *29*(3), 246–277.

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bock, J. K., & Ferreira, V. S. (2014). Syntactically speaking. In M. Goldrick, V. S. Ferreira, & M. Miozzo (Eds.), *The Oxford Handbook of Language Production* (pp. 21–46). Oxford: Oxford University Press.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, *20*.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, 1–72. https://doi.org/http://dx.doi.org/10.1017/S0140525X1500031X

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.

Chukharev-Khudilaynen, E. (2014). Pauses in spontaneous written communication: A keystroke logging study. *Journal of Writing Research*, *6*(1), 61–84.

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.

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.

Eltahir, W. E., Salami, M., Ismail, A. F., & Lai, W. (2004). Dynamic keystroke analysis using AR model. In *IEEE international conference on industrial technology* (Vol. 3, pp. 1555–1560). IEEE.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

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.

Grabowski, J., Weinzierl, C., & Schmitt, M. (2010). Second and fourth graders' copying ability: From graphical to linguistic processing. *Journal of Research in Reading*, *33*(1), 39–53.

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* (pp. 15–25). New York, NY: Psychology Press.

Hoaglin, D. C., & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association*, *82*(400), 1147–1149.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Kaufer, D. S., Hayes, J. R., & Flower, L. (1986). Composing written sentences. *Research in the Teaching of English*, *20*(2), 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.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001.

McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in R and Stan*. CRC Press.

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.

Quené, H., & Van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, *43*(1-2), 103–121.

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.

Schöner, G. (2002). Timing, clocks, and dynamical systems. *Brain and Cognition*, *48*(1), 31–51.

Stan Development Team. (2015a). Stan: A C++ library for probability and sampling. http://mc-stan.org/.

Stan Development Team. (2015b). Stan modeling language user's guide and reference manual. http://mc-stan.org/.

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., 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 multi-level analysis of the influences of task complexity. *Reading and Writing*, *23*(7), 803–834. https://doi.org/10.1007/s11145-009-9190-x

Van Waes, L., Leijten, M., Roeser, J., Olive, T., & Grabowski, J. (2020). Designing a copy task to measure typing and motor skills in writing research. *Journal of Writing Research*.

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. (2001). Disfluencies in writing – Are they like in speaking? In *ISCA tutorial and research workshop (ITRW) on disfluency in spontaneous speech*.

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). Amsterdam: Elsevier.

Wing, A. M., & Kristofferson, A. B. (1973). Response delays and the timing of discrete motor responses. *Perception & Psychophysics*, *14*(1), 5–12.