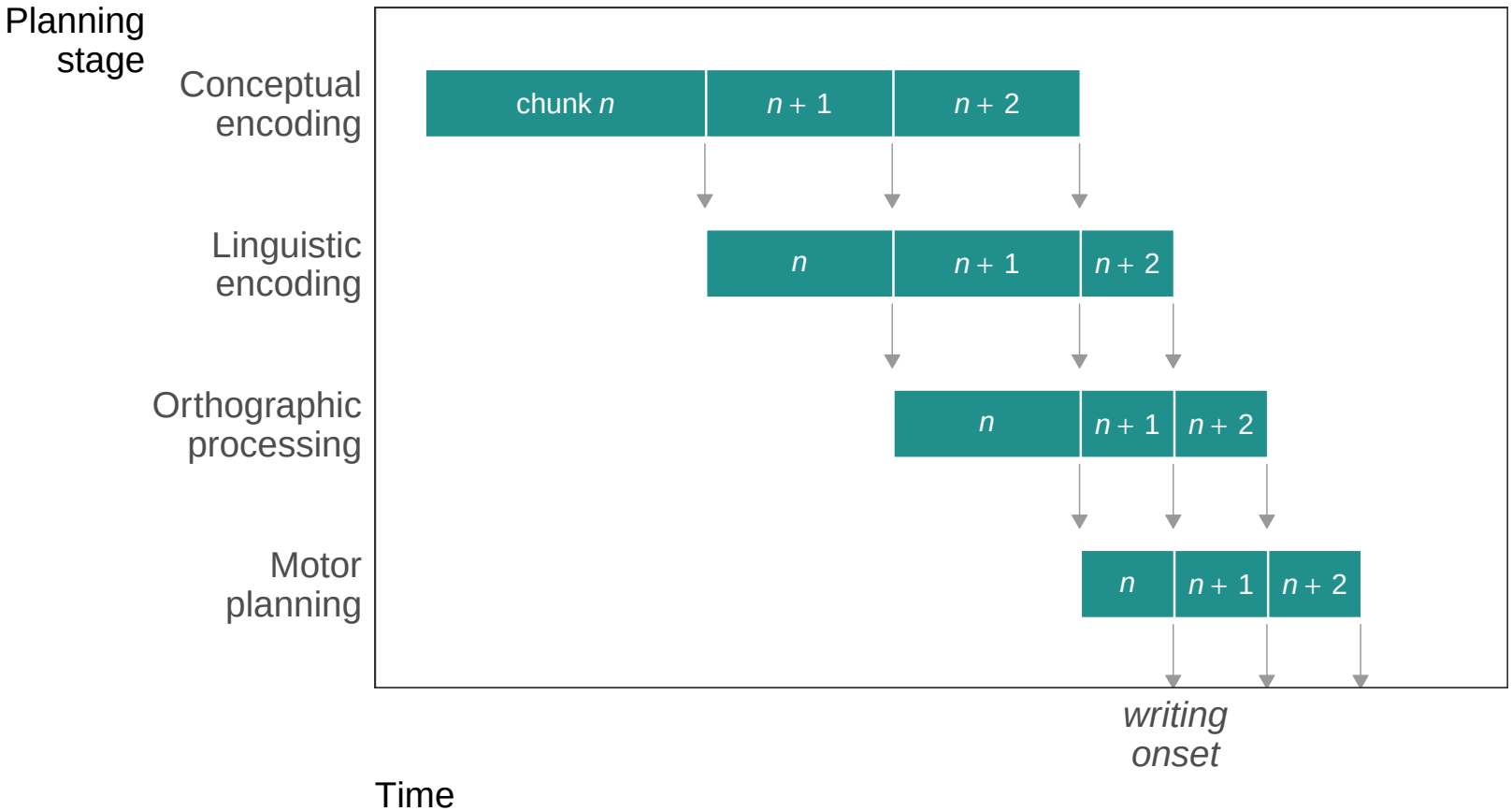
Analysing inter-key intervals: Beyond means, medians and pause frequencies

Jens Roeser
Sven De Maeyer Mark Torrance Mariëlle Leijten
Luuk Van Waes
jens.roeser@ntu.ac.uk

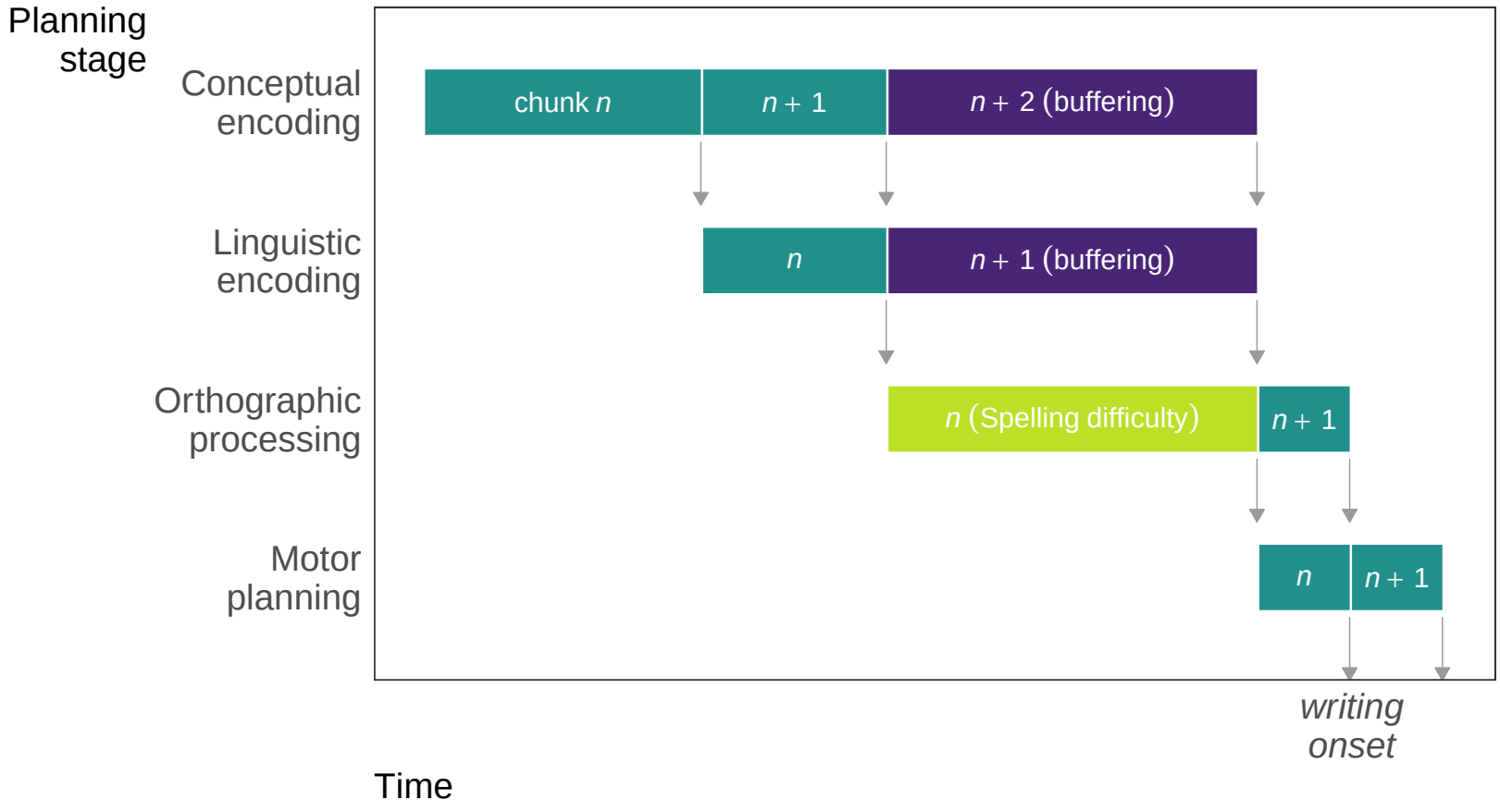
SIG 27 Conference
University of Antwerp

Dec 15, 2020

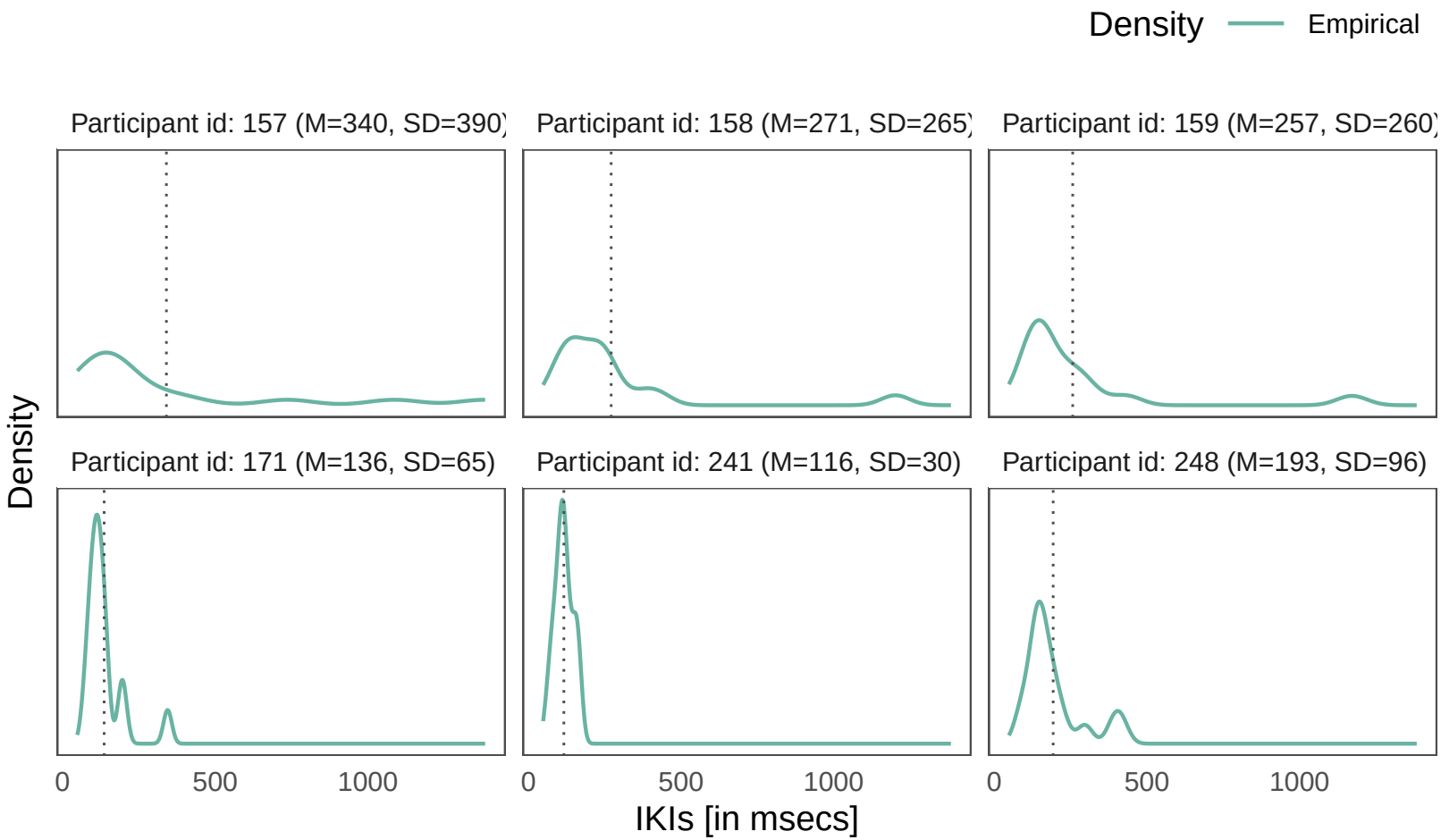
Planning cascade in writing



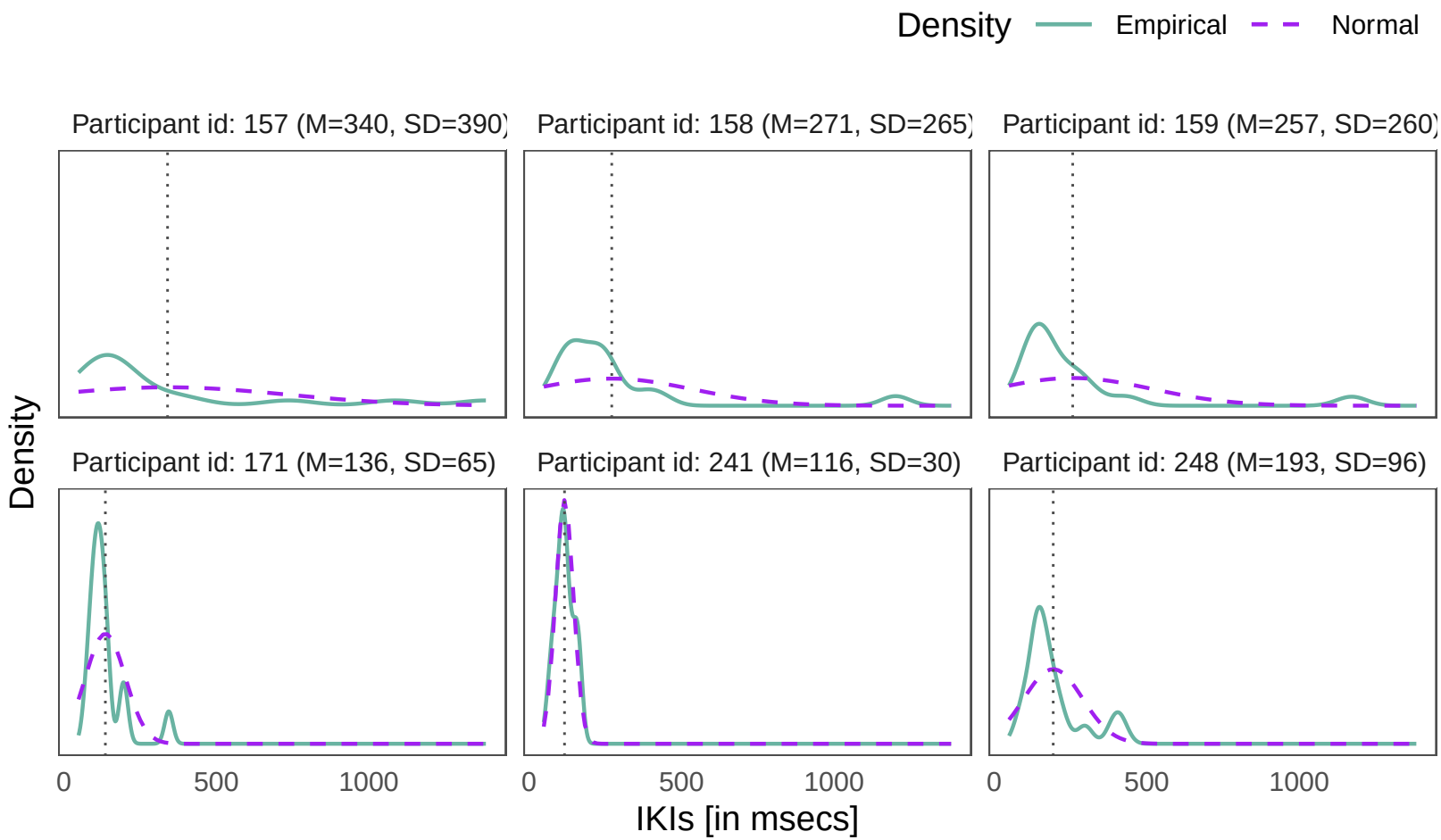
Planning cascade in writing



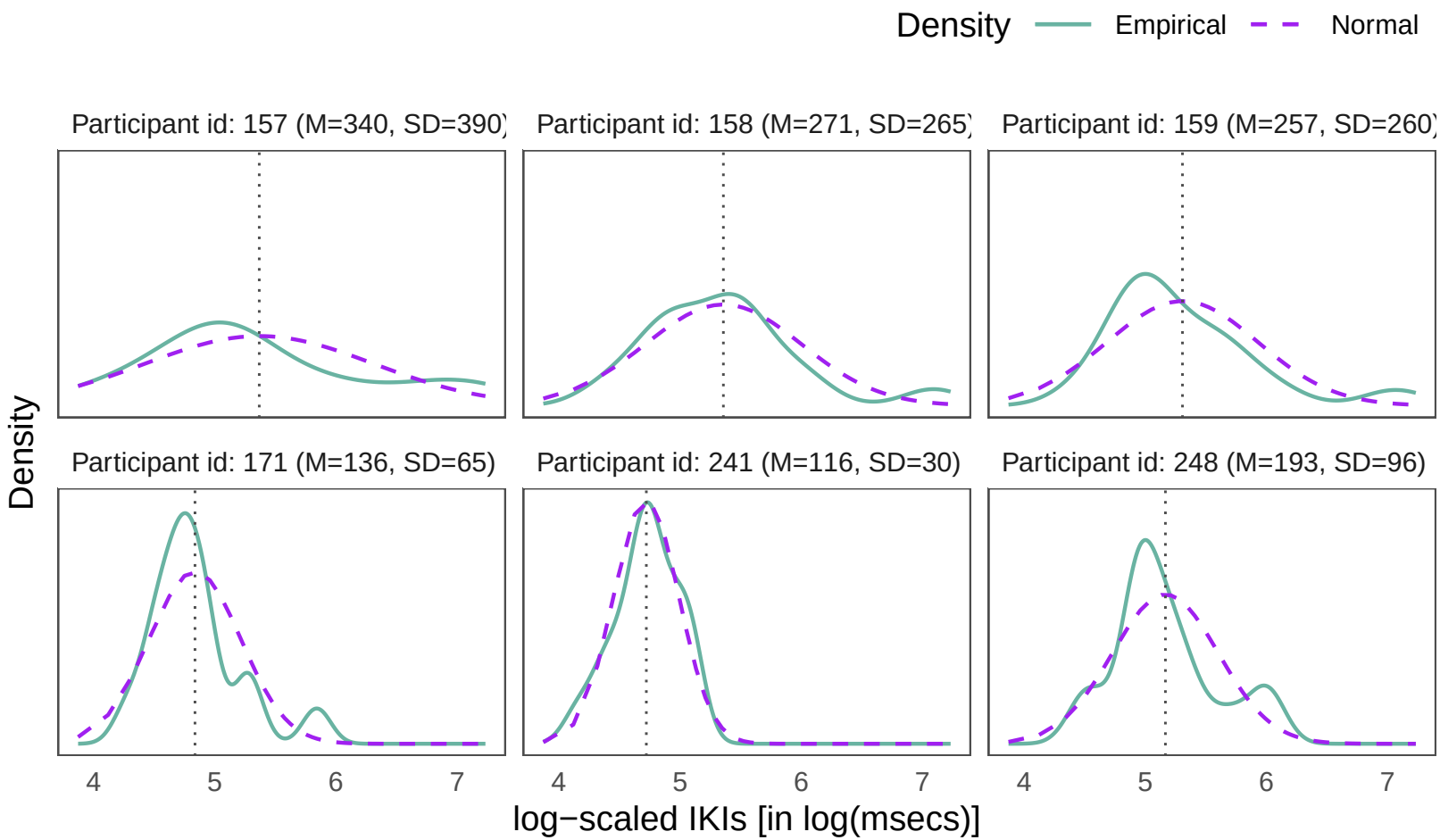
Keystroke transitions are not normal distributed



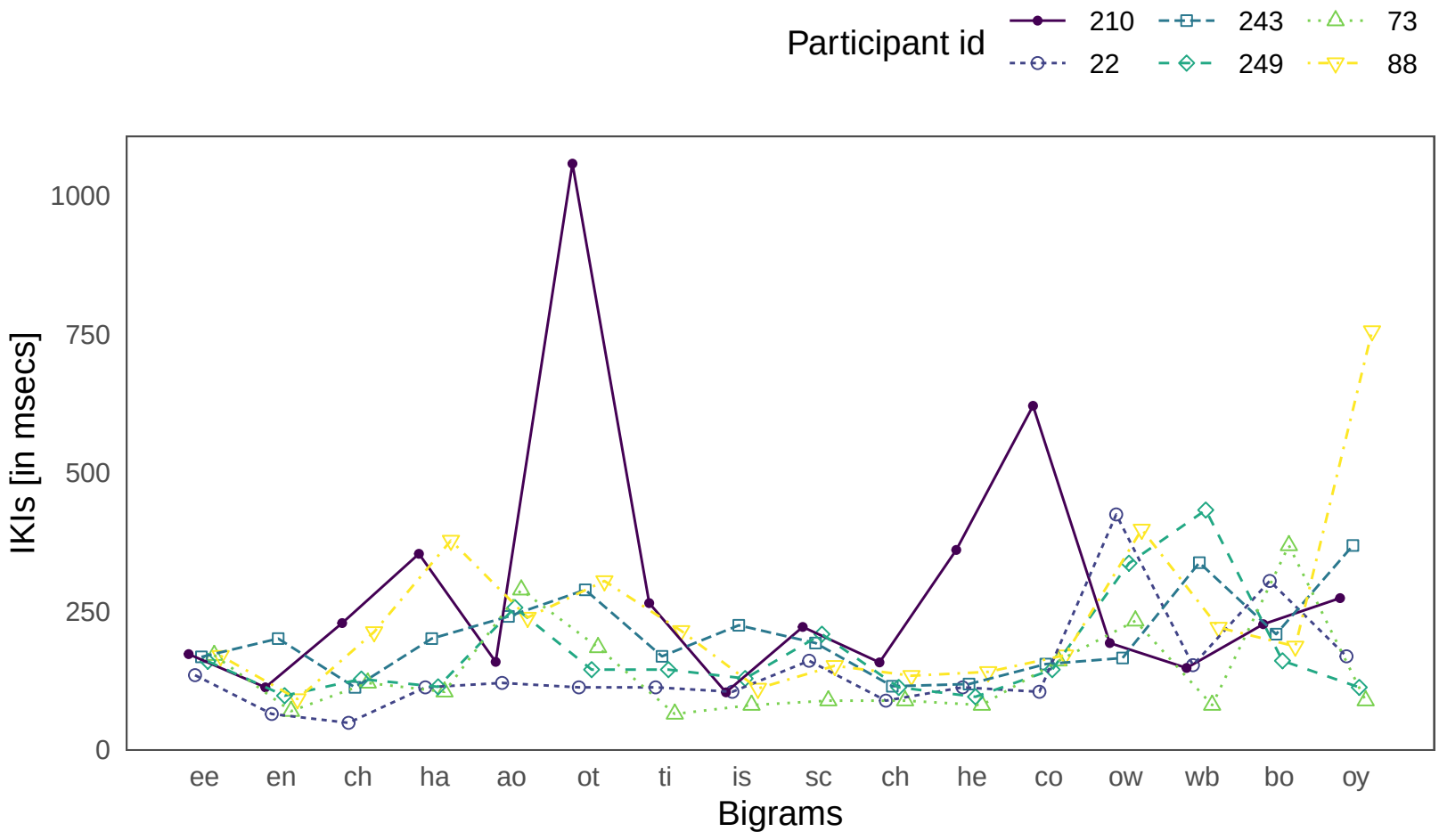
Keystroke transitions are not normal distributed



Keystroke transitions are not normal distributed



Long intervals are not bigram specific



Research focus

- How do we address the long right tail in statistical models without losing information, trimming data, imposing pause thresholds?
- log-Gaussian stable distribution (Guo et al., 2018)
- ex-Gaussian (Chukharev-Khudilaynen, 2014)
- Mixture of processes (Almond et al., 2012; Baaijen et al., 2012)
- Implementation of the copy-typing process as statistical models in Stan (Carpenter et al., 2016); code based on Sorensen et al. (2016) and Vasishth, Chopin et al. (2017); also Vasishth, Jäger et al. (2017).

LogNormal Mixed-Effects Model

$$y_{ij} \sim LogNormal(\mu_{ij}, \sigma_e^2)$$

LogNormal Mixed-Effects Model

$$y_{ij} \sim LogNormal(\mu_{ij}, \sigma_e^2)$$
$$\mu_{ij} = \alpha + u_i + w_j$$

LogNormal Mixed-Effects Model

$$y_{ij} \sim LogNormal(\mu_{ij}, \sigma_e^2)$$
$$\mu_{ij} = \alpha + u_i + w_j$$

- Average IKI α.
- Participants: $u_i \sim Normal(0, \sigma_u^2)$
- Bigrams: $w_j \sim Normal(0, \sigma_w^2)$
- Error variance σ_e^2

Extending mixed models to mixtures

$$y \sim LogNormal(\mu, \sigma_e^2)$$

Extending mixed models to mixtures

$$y \sim LogNormal(\mu, \sigma_e^2)$$

Extending mixed models to mixtures

$$y \sim \theta \cdot LogNormal(\mu, \sigma_e^2)$$
$$\theta = 1$$

Extending mixed models to mixtures

$$y \sim \theta \cdot LogNormal(\mu_1, \sigma^2_{e_1}) +$$
$$(1 - \theta) \cdot LogNormal(\mu_2, \sigma^2_{e_2})$$
$$\theta = \ ?$$

Mixture of Gaussians

$$y_{ij} \sim \theta_i \cdot LogNormal(\mu_{ij} + \delta, \sigma^2_{e'}) +$$
$$(1 - \theta_i) \cdot LogNormal(\mu_{ij}, \sigma^2_e)$$
$$\mu_{ij} = \alpha + u_i + w_j$$

- Disfluencies result in a slowdown δ.
- happening with a probability θ; differs by-participant i
- associated with larger variance $\sigma^2_{e'_k}$

Mixture of Gaussians

$$y_{ij} \sim \theta_i \cdot LogNormal(\mu_{ij} + \delta, \sigma^2_{e'}) +$$
$$(1 - \theta_i) \cdot LogNormal(\mu_{ij}, \sigma^2_e)$$
$$\mu_{ij} = \alpha + u_i + w_j$$

- Disfluencies result in a slowdown δ.
- happening with a probability θ; differs by-participant i
- associated with larger variance $\sigma^2_{e'_k}$

Mixture of Gaussians

$$y_{ij} \sim \theta_i \cdot \textit{LogNormal}(\mu_{ij} + \delta, \sigma^2_{e'}) + (1 - \theta_i) \cdot \textit{LogNormal}(\mu_{ij}, \sigma^2_e)$$

$$\mu_{ij} = \alpha + u_i + w_j$$

- Disfluencies result in a slowdown δ.
- happening with a probability θ; differs by-participant i
- associated with larger variance $\sigma^2_{e'_k}$

Mixture of Gaussians

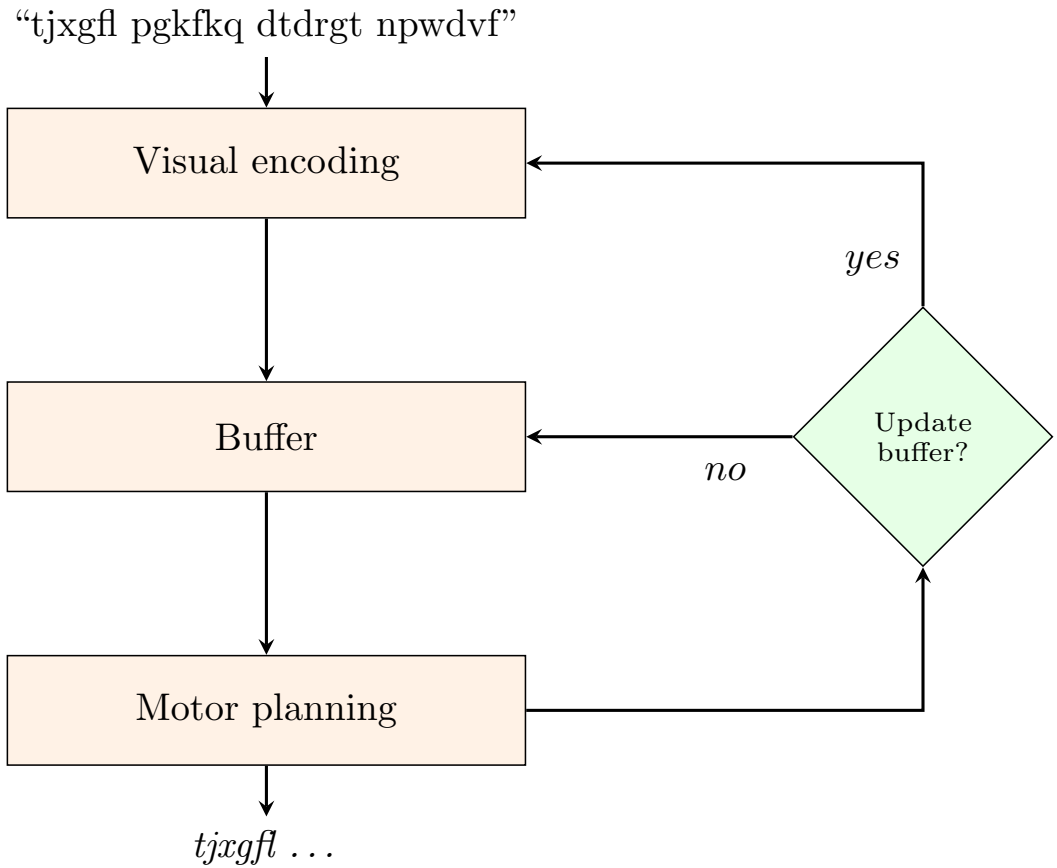
$$y_{ij} \sim \theta_i \cdot LogNormal(\mu_{ij} + \delta, \sigma^2_{e'}) +$$
$$(1 - \theta_i) \cdot LogNormal(\mu_{ij}, \sigma^2_e)$$
$$\mu_{ij} = \alpha + u_i + w_j$$

- Disfluencies result in a slowdown δ.
- happening with a probability θ; differs by-participant i
- associated with larger variance $\sigma^2_{e'_k}$

Method

- Bigram-keystroke interval data: Dutch subset of copy-task corpus (Van Waes et al., 2019; Van Waes et al., 2020)
- **Consonants task:** "tjxgfl pgkfkq dtdrgt npwdvf"
- **LF-bigrams task:** "een chaotische cowboy"
- Lexical and non-lexical copy-typing context
- Random sample of 250 ppts (18-25 years of age)

Copy-typing sequences of characters



References I

Almond, R., Deane, P., Quinlan, T., Wagner, M. & Sydorenko, T. (2012). *A preliminary analysis of keystroke log data from a timed writing task* (tech. rep. Research Report No. RR-12-23). Princeton, NJ, Educational Testing Service.

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

Carpenter, B., Gelman, A., Hoffman, M. D., 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*.

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

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.

Sorensen, T., Hohenstein, S. & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *Quantitative Methods for Psychology, 12*(3), 175–200.

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., Roeser, J., Olive, T. & Grabowski, J. (2020). Designing a copy task to measure typing and motor skills in writing research [submitted]. *Journal of Writing Research*.

References II

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*.