Can we **detect short-term fluency development** after 2 hours of chat with a dialogue system?

Serge Bibauw

Universidad Central del Ecuador · UCLouvain · KU Leuven

CALICO 2022

Seattle, June 4, 2022

Fluency development with a dialogue system

- Project: Effectiveness of dialogue systems/dialogue-based CALL
  - with Louis Escouflaire, Thomas François, Piet Desmet

- **Theoretical challenge:**
  - Developing fluency with written practice with dialogue system.

- **Methodological challenge:**
  - Precise measurement of fluency, to detect short-term gains.

Background

Utterance fluency and fluency development
Dialogue systems

Data & Methods

Dialogue-based CALL game
Speech test and semi-automatized analysis

Results & Discussion

Metrics of fluency ⇔ Proficiency
Longitudinal fluency development

Background

Utterance fluency and fluency development
Dialogue systems / Dialogue-based CALL

Data & Methods

Dialogue-based CALL game
Speech test and semi-automatized analysis

Results & Discussion

Metrics of fluency ⇔ Proficiency
Longitudinal fluency development

L2 fluency

(Segalowitz, 2010)

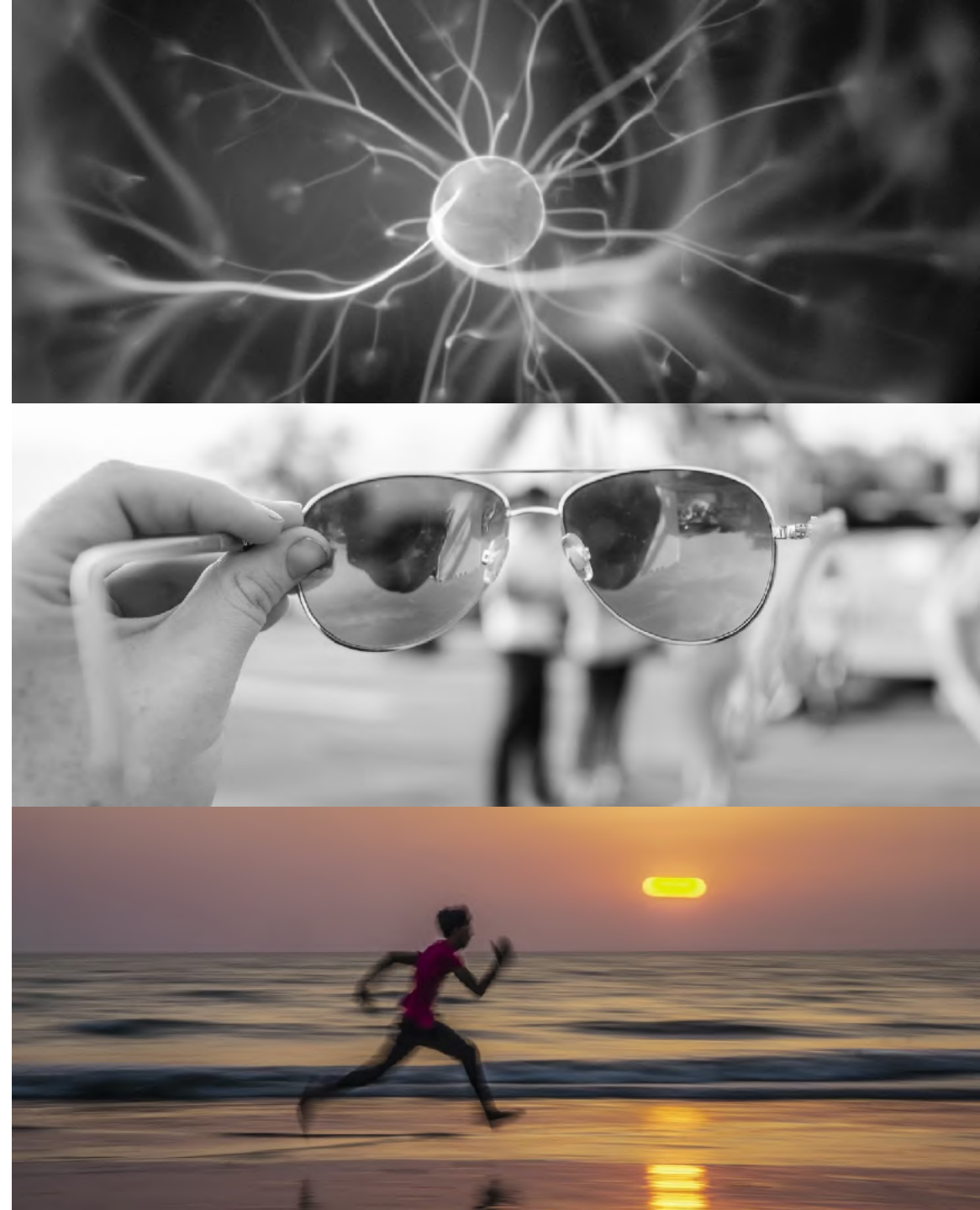- Cognitive fluency
- Perceived fluency
- Utterance fluency

Photos by Halacious, Erika Fletcher & Swapnil Dwivedi on Unsplash

L2 fluency

- ~~Cognitive fluency~~
- ~~Perceived fluency~~
- **Utterance fluency** (performance)
  - Speed fluency
  - Breakdown fluency
  - Repair fluency

Photos by Halacious, Erika Fletcher & SwapnIl Dwivedi on Unsplash

Utterance fluency & L2 proficiency

- Interested in relation to **L2 proficiency** (Tavakoli et al., 2020) for

  - Predicting speaking proficiency

  - Fast (initial) rating of learner/user

  - Detecting short-term development

    - ⇒ autonomous language learning apps

  - Evaluating effects of spontaneous/interactive output on fluency

    - ⇔ ouptut hypothesis/interactionist perspective

Fluency metrics to predict proficiency

- **Speed** fluency **Length/Time**

  - ✔ good differentiator between fluent/non-fluent
    (e.g., Bosker et al., 2013; Hilton, 2014; Götz, 2013; Kahng, 2014)

  - **Speech rate** ✔ (Detey et al., 2020)
    [# syllables / total time]

  - **Articulation rate:** ✔ (de Jong et al., 2020)
    [# syllables / phonation time]

  - **Syllable duration**: ✔ (Segalowitz et al., 2017; Saito et al., 2018)
    [phonation time / # syllables]

- **Speed/Breakdown fluency ⇒ Runs** Length/Pauses

  - **Length of runs** (= Syllable run): ✔ (Segalowitz et al., 2017)
    `[# syllables / # silent pauses]`

  - **Duration of runs** (= Phon. run): ✔ (de Jong & Bosker, 2013...)
    `[phonation time / # silent pauses]`

- **Breakdown fluency** Pauses/Time

  - **Duration of silent pauses**: ✘ (de Jong & Bosker, 2013; de Jong et al., 2015)
    `[total silent pausing time / # silent pauses]`

  - **Filled pauses rate**: ✘ (Cucchiarini et al., 2002; Segalowitz et al., 2017)
    `[# filled pauses / total time]`

  - Also: **Pause location**: Mid-/Final-clause pause ratio
    (discarded temporarily here for technical reasons)

- **Repair fluency**
  - False starts, corrections and repetitions
  - ✗ not good proficiency differentiator, nor predicitve of comm. adequacy or perceived fluency
    (Cucchiarini et al., 2002; Révèsz et al., 2016; Saito et al., 2018)

- Many other metrics…

Fluency development

- Most studies on fluency gains: **study abroad context**

  - Long-term: 3-24 months (O'Brien et al, 2007; Mora & Valls-Ferrer, 2012; Huensch & Tracy-Ventura, 2017)

  - "Short-term" = 5 weeks (Segalowitz et al, 2017)

  - Instructed: 3-4 months (Temple, 2005; Galante & Thomson, 2016)

  - ⇒ *Possible to measure shorter-term gains with precise metrics?*

- Fluency development < spoken interactions (Derwing et al., 2008)

  - Consensus: lack of speaking practice in classroom (Derwing, 2017)

  - ⇒ *Can technology help?*

Background

Utterance fluency and fluency development
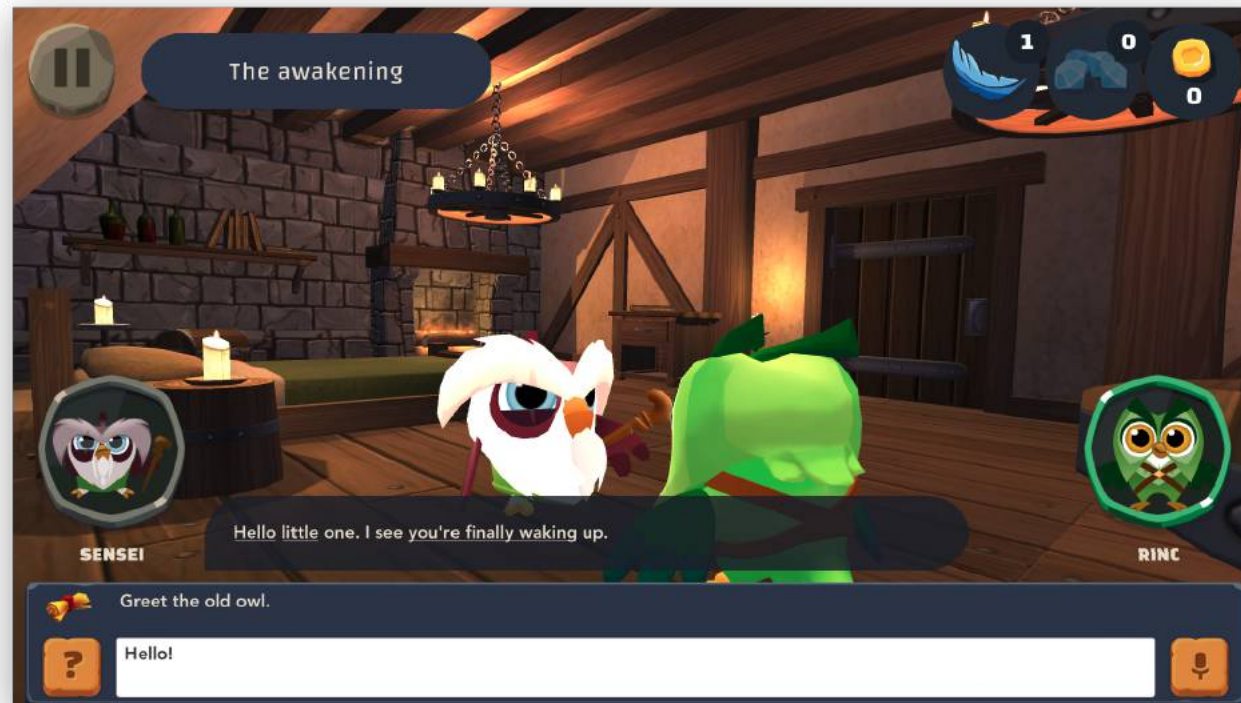**Dialogue systems / Dialogue-based CALL**

Data & Methods

Dialogue-based CALL game
Speech test and semi-automatized analysis

Results & Discussion

Metrics of fluency ⇔ Proficiency
Longitudinal fluency development

Dialogue-based CALL

- Chatbots, dialogue systems, conversational agents, talking robots, smart speakers…
  (Bibauw et al, 2019)

- Large potential: meaningful spontaneous practice (spoken/Written)

- Low-anxiety, fully controllable environment

Dialogue-based CALL → Fluency?

- We know very little... (Bibauw et al, 2022)
  - $d = 0.39$ in meta-analysis, but not significant
    - vs. $d = 0.58$ overall effect

Background

Utterance fluency and fluency development
Dialogue systems / Dialogue-based CALL

Data & Methods

Dialogue-based CALL game
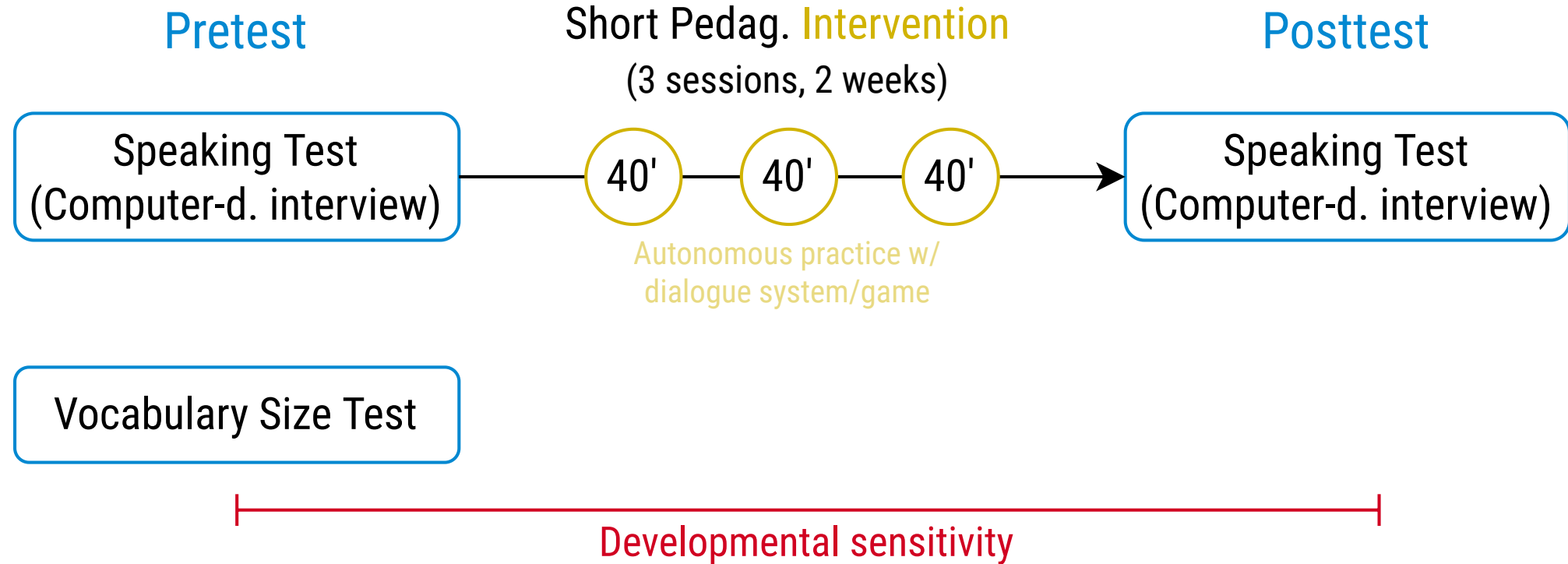Speech test and semi-automatized analysis

Results & Discussion

Metrics of fluency ⇔ Proficiency
Longitudinal fluency development

Research design
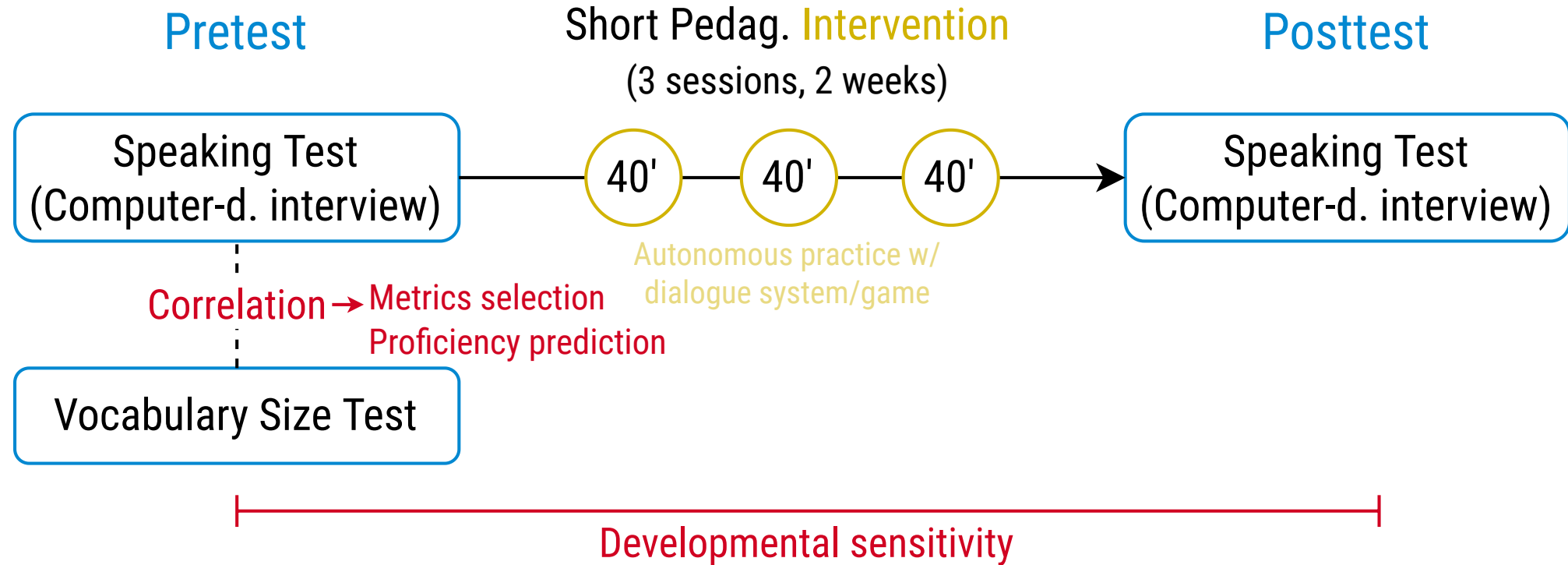
Pretest

Short Pedag. Intervention
(3 sessions, 2 weeks)

Posttest

Speaking Test
(Computer-d. interview)

40' 40' 40'

Autonomous practice w/
dialogue system/game

Speaking Test
(Computer-d. interview)

Vocabulary Size Test

Background

Utterance fluency and fluency development
Dialogue systems / Dialogue-based CALL

Data & Methods

Dialogue-based CALL game
Speech test and semi-automatized analysis

Results & Discussion

Metrics of fluency ⇔ Proficiency
Longitudinal fluency development

Intervention: Dialogue-based CALL game

- Interactive practice with a dialogue system
  - integrated into a video game
  - sponteanous written interaction (chat) + multimodal input
  - guided by microtask prompts

Participants

- $N = 164$
  (initially $N = 228$ but incomplete/problematic data)

- 4 schools, 11 classes

- 12-13 y.o. (2nd grade BE/8th grade US/Year 9 UK)

- L1: **Dutch**

- L2: **French** ~A1+→A2
  (but some outliers: up to B2 + heritage speakers)

Background

Utterance fluency and fluency development
Dialogue systems / Dialogue-based CALL

Data & Methods

Dialogue-based CALL game
Speech test and semi-automatized analysis

Results & Discussion

Metrics of fluency ⇔ Proficiency
Longitudinal fluency development

Computer-delivered speech test

- Autonomous simultaneous speaking test
  - Individual, in-class & simultaneous,
  - with headset, in front of indiv. computer
- 24 questions
  - from basic ("*How are you?*") to questions targeting specific communicative functions ("*Can you describe your French teacher?*")
- Oral question + written transcription
  - then automatically starts recording
  - 30 sec limit or "Next question" button

Automated speech analysis

- **Data**: >10 000 audio files (`WAV` , 2-30")
  - N=228 * 24 questions * pre+post
- **Transcription**: automated speech recognition (Google Cloud Speech-to-text)
  - Manual revision of transcriptions
- Manual annotation of filled pauses, L1/LF use, disfluencies…
- Automated detection of **silent pauses** & **phonation time**:
  - Praat Syllable Nuclei detection script (de Jong et al., 2020)
- Automated computation of **# syllables** from transcript
  - with different pruning alternatives
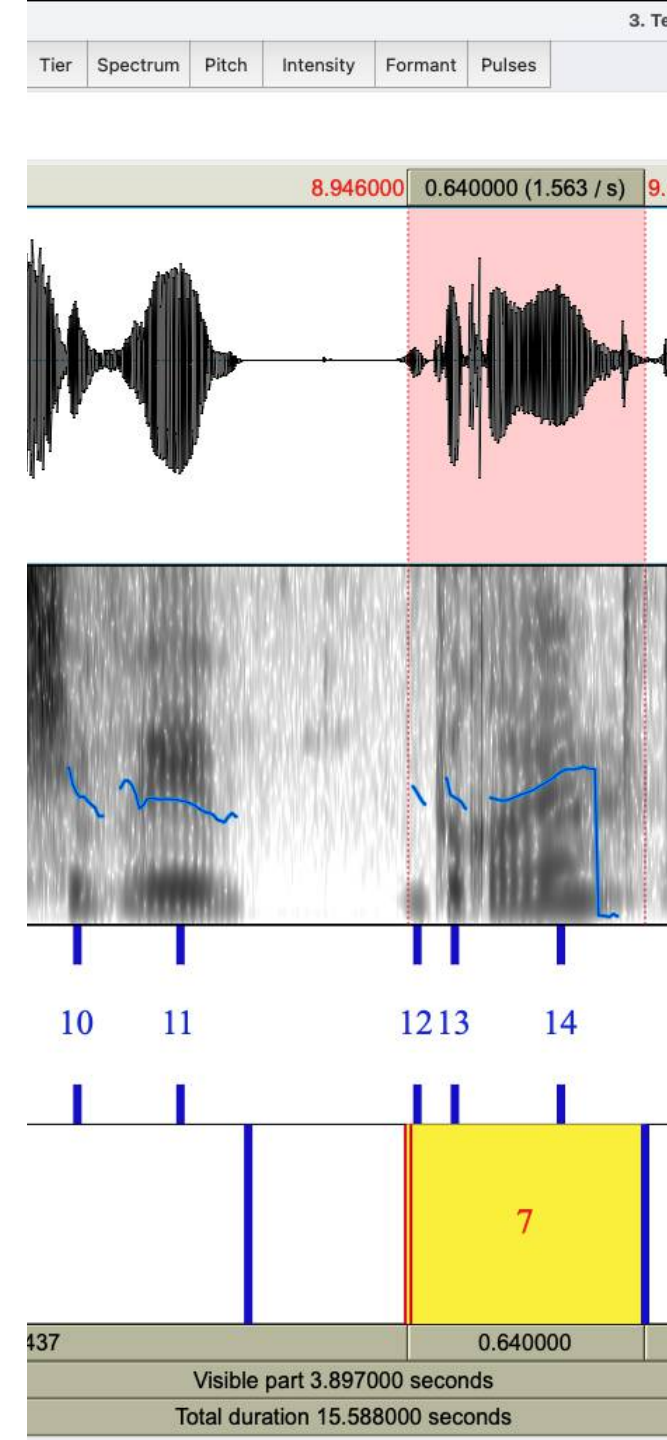
Validation of fluency metrics

- Internal consistency

- Comparison of metrics for proficiency (per-participant correlation)

  - **Vocabulary Size**

    - quick but reliable estimate of L2 proficiency
      (Noreillie et al., 2018; Milton, 2013)

  - **Vocabulary Size Test**

    - **productive** (gap-filling, with 1st letter + L1 translation given)

      - even better correlation with speaking proficiency
        (r = 0.77 in Koizumi, 2005; r = 0.79 in de Jong et al., 2012)

    - standardized & validated (Noreillie, 2019)

    - 30 words, 1K frequency band (A1)

VS1_6

Dans une démocratie, c'est le p_ _ _ _ _ (volk

VS1_7

Le général j_ _ _ (oordelen) qu'il n'est pas néc

VS1_8

Il a été condamné à une p_ _ _ _ (straf) de pri

VS1_9

La p_ _ _ _ (verovering) de la Bastille a été un

Background

Utterance fluency and fluency development
Dialogue systems / Dialogue-based CALL

Data & Methods

Dialogue-based CALL game
Speech test and semi-automatized analysis

Results & Discussion

Metrics of fluency ⇔ Proficiency
Longitudinal fluency development

Automated estimators vs. Manual annotation

Raw metrics	MAE (accur.)	RMSE	R^2 (consist.)	Cr. α (int.cons.)	r_{VS}
Nb of syllables (auto count, manual trscpt)	*"truth"*			.92	.373
↳ Google ASR transcript (auto count)	1.23	2.93	.874	.91	.370
↳ Syllable Nuclei Praat script (de Jong et al.)	4.25	7.60	.585	.88	.154

Pruning

Number of syllables Variant / Pruning	M	SD	Cr. α	r_{VS}	$r_{SR\text{-}VS}$
Unpruned (manual transcript)	13.4	5.44	.92	.373	.579
'Meant': – disfluencies (f.pauses, repet., self-corr., meta)	12.2	5.10	.92	**.443**	**.597**
'Meant', L2-only: – L1/lingua franca words	12.1	5.07	.93	**.459**	**.603**
'Meant', L2-only, – proper nouns	12.0	5.02	.93	**.473**	**.609**

- ⇒ Pruning improves the meaningfulness of length-based metrics

- ⇒ 'Harsher' pruning increases predictive power

Background

Utterance fluency and fluency development
Dialogue systems / Dialogue-based CALL

Data & Methods

Dialogue-based CALL game
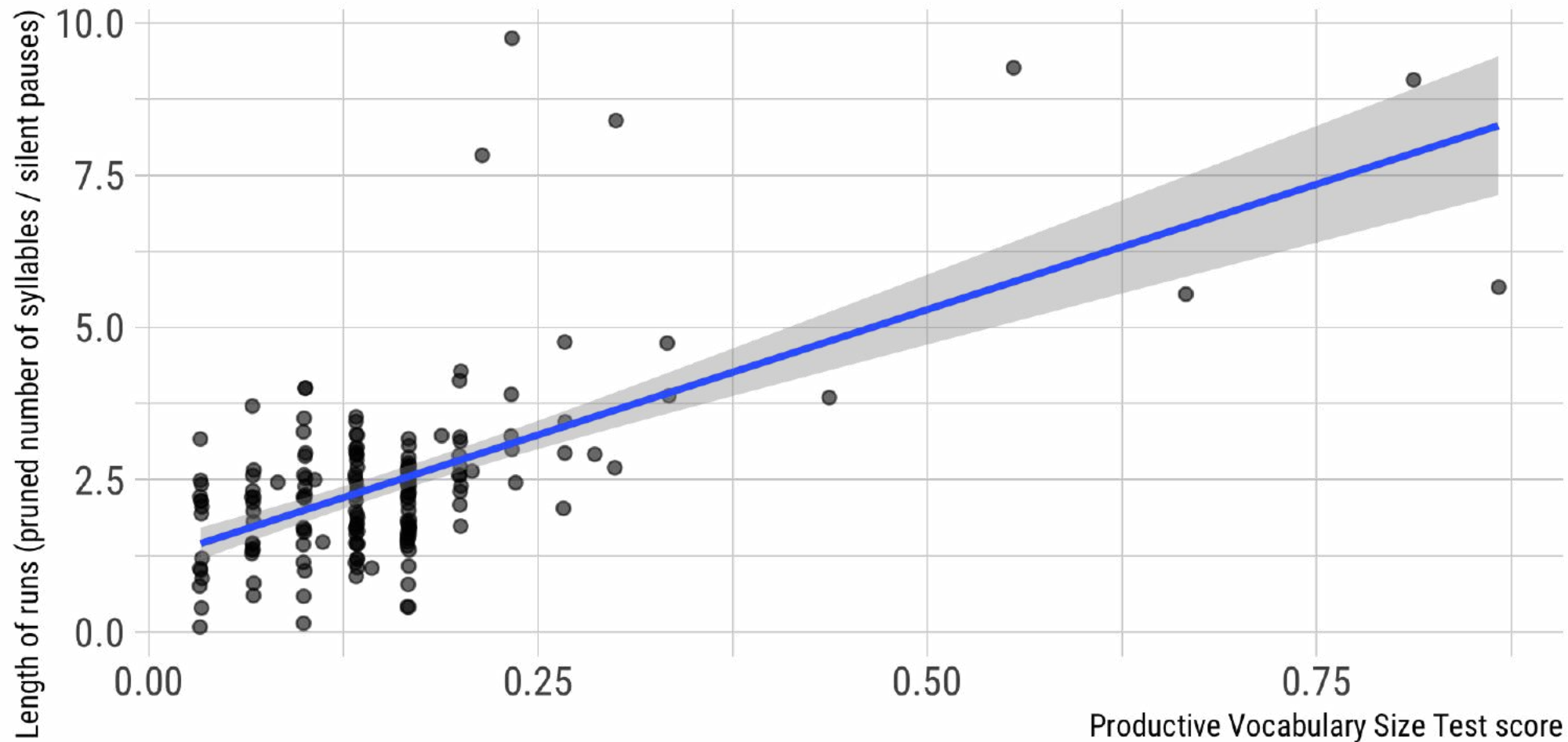Speech test and semi-automatized analysis

Results & Discussion

Metrics of fluency ⇔ Proficiency
Longitudinal fluency development

Best predictors of L2 proficiency

- Speech rate? Articulation rate?

- Length of runs? Duration of runs?

- Duration of silent pauses? Silent pauses rate?

- Speech-time ratio?

Length of runs is the best predictor of proficiency

r = 0.628, N = 164

Best predictors of L2 proficiency

- Length of runs (syll. runs), pruned*   .628
- Speech rate, pruned   .609
- Articulation rate, pruned   .524
- Syllable duration^{-1}, pruned   .473
- Number of syllables, pruned   .473
- Number of words, pruned   .463
- Silent pausing rate^{-1} .428
- Duration of runs (phon. runs)   .352
- Speech-time ratio   .305
- Pause duration^{-1} .197

Based on correlation with Vocabulary Size, Pearson's r

* Pruning: removed disfluencies, repetitions, meta-discourse, L1/LF words, proper nouns

Background

Utterance fluency and fluency development
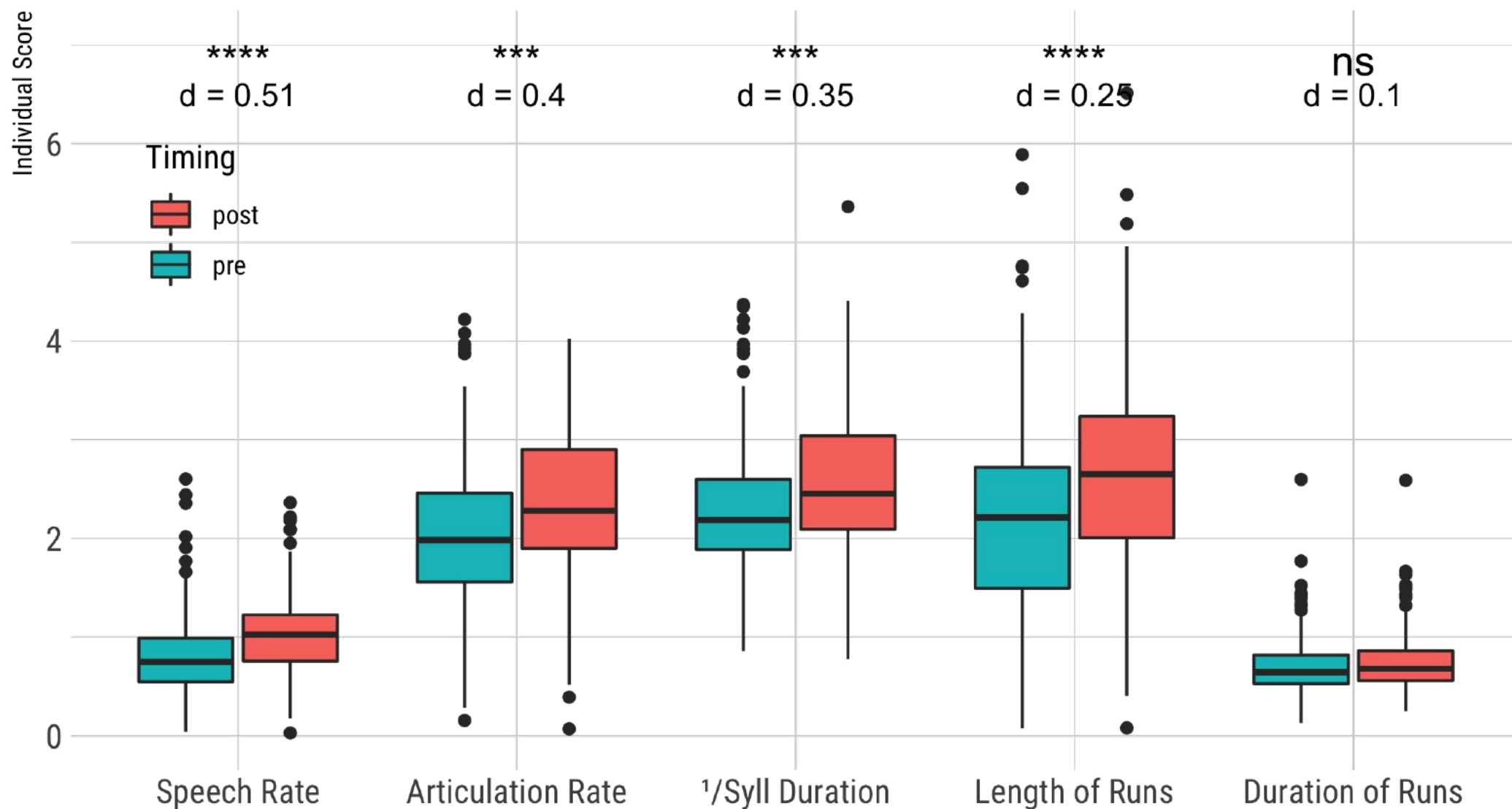Dialogue systems / Dialogue-based CALL

Data & Methods

Dialogue-based CALL game
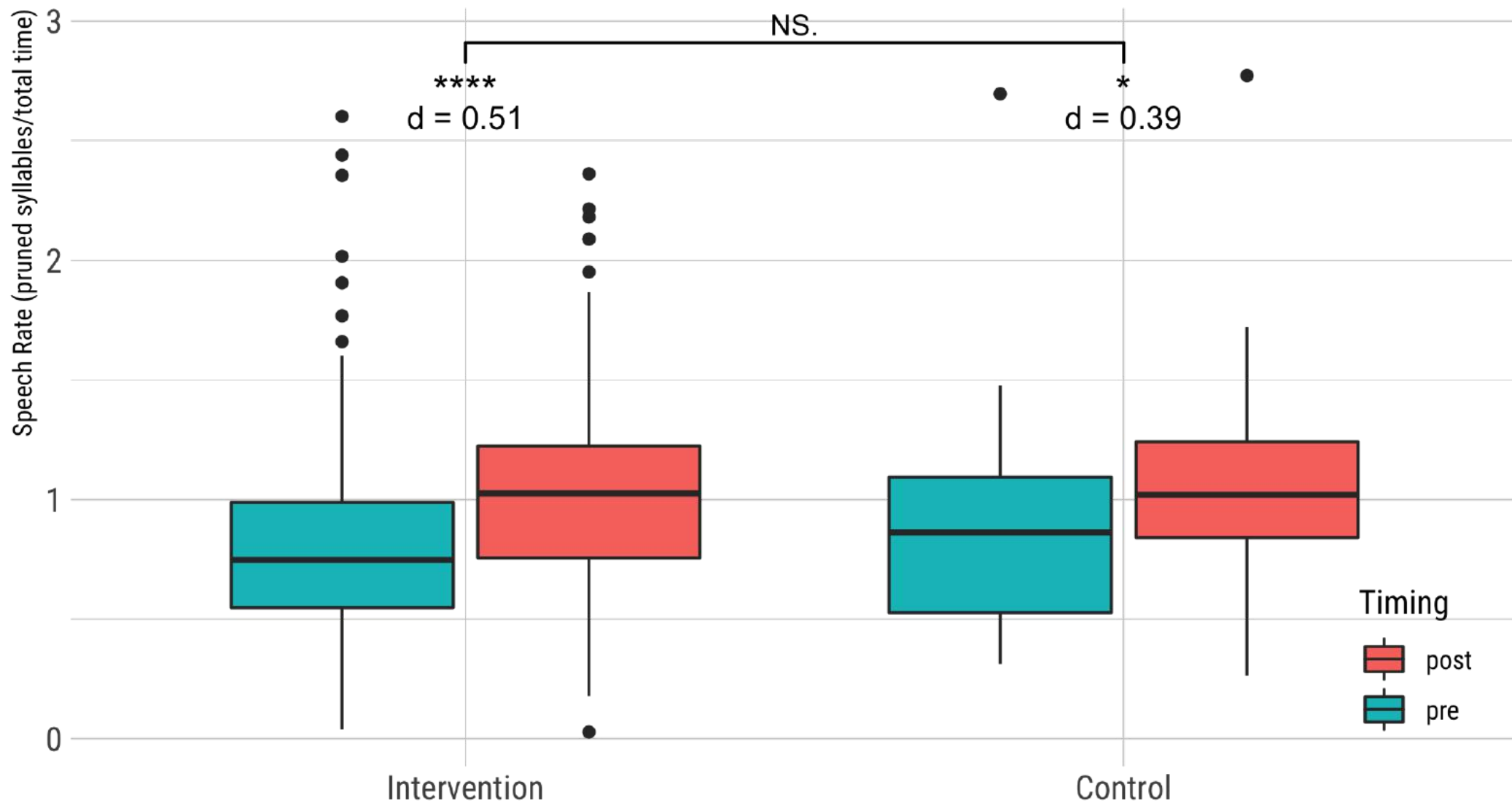Speech test and semi-automatized analysis

Results & Discussion

Metrics of fluency ⇔ Proficiency
Longitudinal fluency development

Developmental Sensitivity of selected Fluency Metrics

Significant, Medium Effect on Speech Rate (partial task repetition effect)

Conclusions

- Possible to measure very short-term gains!

- Automated metrics work.

- Harsh **pruning** improves predictive power.

- Best predictors of **L2 proficiency**:

  - Length of Runs > Speech Rate > Artic. Rate

- Best developmental **sensitivity**:

  - Speech Rate > Artic. Rate > Syll. Duration^{-1} > Length of Runs

- Dialogue-based CALL: large potential, but needs a slightly longer intervention

Questions, feedback & suggestions welcome!

Serge Bibauw
[sbibauw@uce.edu.ec]
[https://serge.bibauw.be]

Developmental Sensitivity of selected Fluency Metrics

Download the slides
[https://cutt.ly/calico22]

R scripts: e-mail me!