Automatizing L2 fluency measurement

Validity and developmental sensitivity of temporal fluency metrics variations

Serge Bibauw · Louis Escouflaire
Thomas François · Piet Desmet

U. Central Ecuador · UCLouvain · KU Leuven

AILA 2021 - Symp. Interdisciplinary approaches to L2 fluency
August 18, 2021

Automatizing L2 fluency measurement

Validity and developmental sensitivity of temporal fluency metrics variations

- **Why?**
  - for autonomous language learning apps,
    automatizing elicitation and measurement of fluency.
  - for dynamic, continuous, non-instrusive assessment.
- **How?**
  - autonomous speech test
    + automatized & semi-automatized fluency metrics
  - compare metrics and operationalizations,
  - validate against proficiency
  - compare developmental sensitivity




Background: Fluency ⇔ Proficiency

- Metrics of utterance fluency & proficiency

Data & Methods

- Computer-delivered speech test
- Semi-automatized analysis
- Vocabulary Size for validation

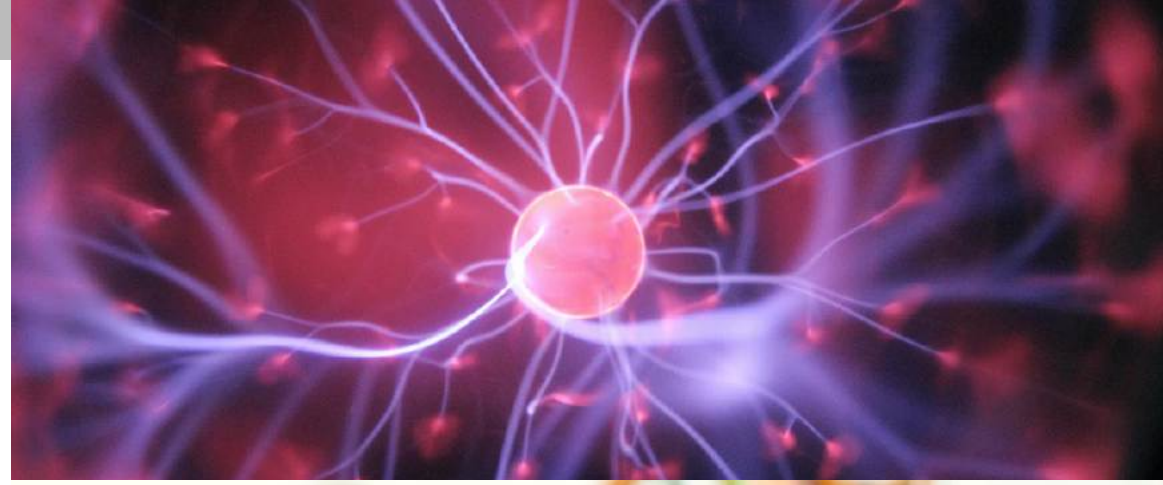
Results & Discussion

- Comparison of annotations & metrics
- Best predictors of L2 proficiency
- Developmental sensitivity

L2 fluency

(Segalowitz, 2010)

- Cognitive fluency
- Perceived fluency
- Utterance fluency

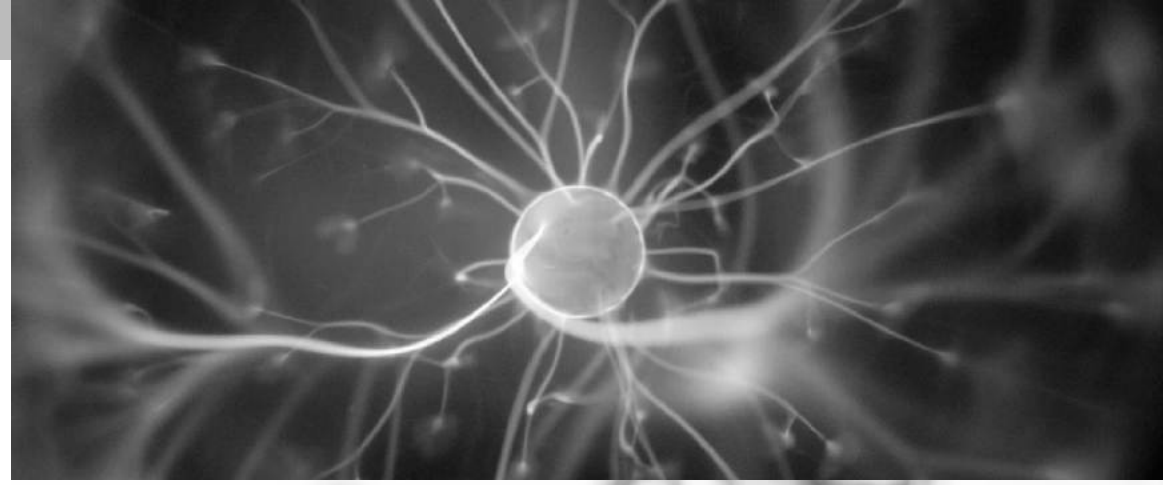


Photos by Halacious, Erika Fletcher & SwapnIl Dwivedi on Unsplash

L2 fluency

(Segalowitz, 2010)

- ~~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

- Often compared to Perceived fluency (Saito et al., 2018; Suzuki et al., 2021)
- Here, 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

Fluency metrics to predict proficiency

- **Speed fluency** Length/Time (1)
  - ✔ good differentiator between fluent/non-fluent, native/non-native (NS/NNS)
    (e.g., Bosker et al., 2013; Hilton, 2014; Götz, 2013; Kahng, 2014)
  - **Speech rate**
    `[# syllables / total time]`
    - ✔ stable, good predictor for automatization
      (Detey et al., 2020)
    - ? redundant with Syllable duration/Art. rate? (Segalowitz et al., 2017)

Speed fluency Length/Time (2)

- **Articulation rate**
  `[# syllables / phonation time]`
  - ✔ unconfounded by silent pauses (de Jong et al., 2020)
  - ? redundant with Syllable duration (Segalowitz et al., 2017)
- **Syllable duration**
  `[phonation time / # syllables]`
  - ✔ sig. differentiator across all fluency levels
    (Saito et al., 2018, but see Révèsz et al., 2016)
  - ✔ good predictor of perceived fluency (r = .67) (Saito et al., 2018)
  - ✔ selected as core fluency measure
  - slightly sensitive to short-term L2 learning gains
    (Segalowitz et al., 2017)

Speed/Breakdown fluency ⇒ Runs Length/Pauses

- **Length of runs** (= Syllable run)
  `[# syllables / # silent pauses]`
  - ✔ great differentiator between NS/NNS
  - ✔ selected as core fluency measure
  - ✔ sensitive to short-term L2 learning gains (Segalowitz et al., 2017)
- **Duration of runs** (= Phonation run) (see also **Silent pause rate**)
  `[phonation time / # silent pauses]`
  - ✔ great differentiator between NS/NNS, fluent/non-fluent (de Jong & Bosker, 2013; Bosker et al., 2013; Segalowitz et al., 2017)
  - ✔ selected as core fluency measure
  - ✔ sensitive to short-term L2 learning gains (Segalowitz et al., 2017)

Breakdown fluency Pauses/Time

- **Duration of silent pauses** ?
  `[total silent pausing time / # silent pauses]`
  - ✘ not good differentiator (de Jong & Bosker, 2013)
    explained mainly by speaking style (de Jong et al., 2015)
  - ✔ selected as core fluency measure, sensitive to short-term L2 learning gains (Segalowitz et al., 2017)
- **Filled pauses rate**
  `[# filled pauses / total time]`
  - ✘ not good differentiator, unrelated to other fluency metrics
    (Cucchiarini et al., 2002; Segalowitz et al., 2017)
- Also: **Pause location**: Mid-/Final-clause pause ratio
  (discarted temporarily here for technical reasons)

Repair fluency

- False starts, corrections and repetitions
- ✘ not good proficiency differentiator
  (Cucchiarini et al., 2002; Révèsz et al., 2016; Saito et al., 2018)
- ✘ not predictive of communicative adequacy (Révèsz et al., 2016)
- ✘ not predictive of perceived fluency (Saito et al., 2018)
- Many other metrics…

Background: Fluency ⇔ Proficiency

- Metrics of utterance fluency & proficiency

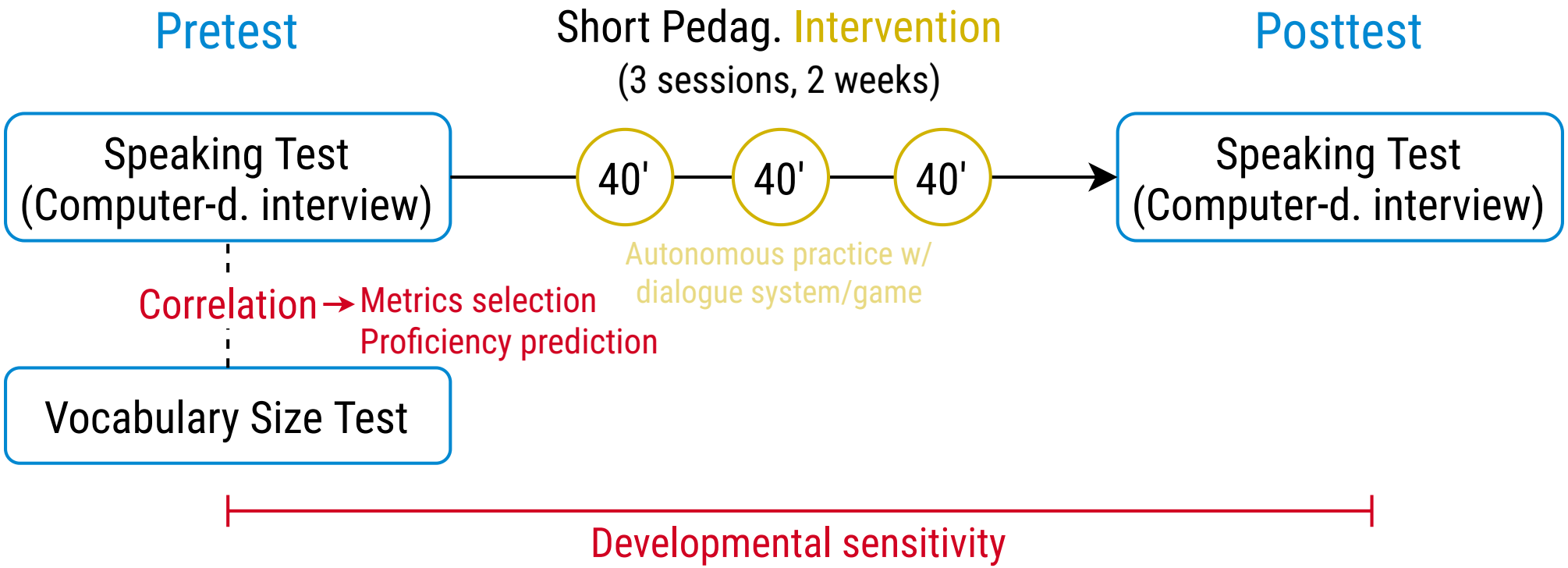
Data & Methods

- Computer-delivered speech test
- Semi-automatized analysis
- Vocabulary Size for validation

Results & Discussion

- Comparison of annotations & metrics
- Best predictors of L2 proficiency
- Developmental sensitivity

Research design



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)

Photo by Taylor Wilcox on Unsplash

Background: Fluency ⇔ Proficiency

- Metrics of utterance fluency & proficiency

Data & Methods

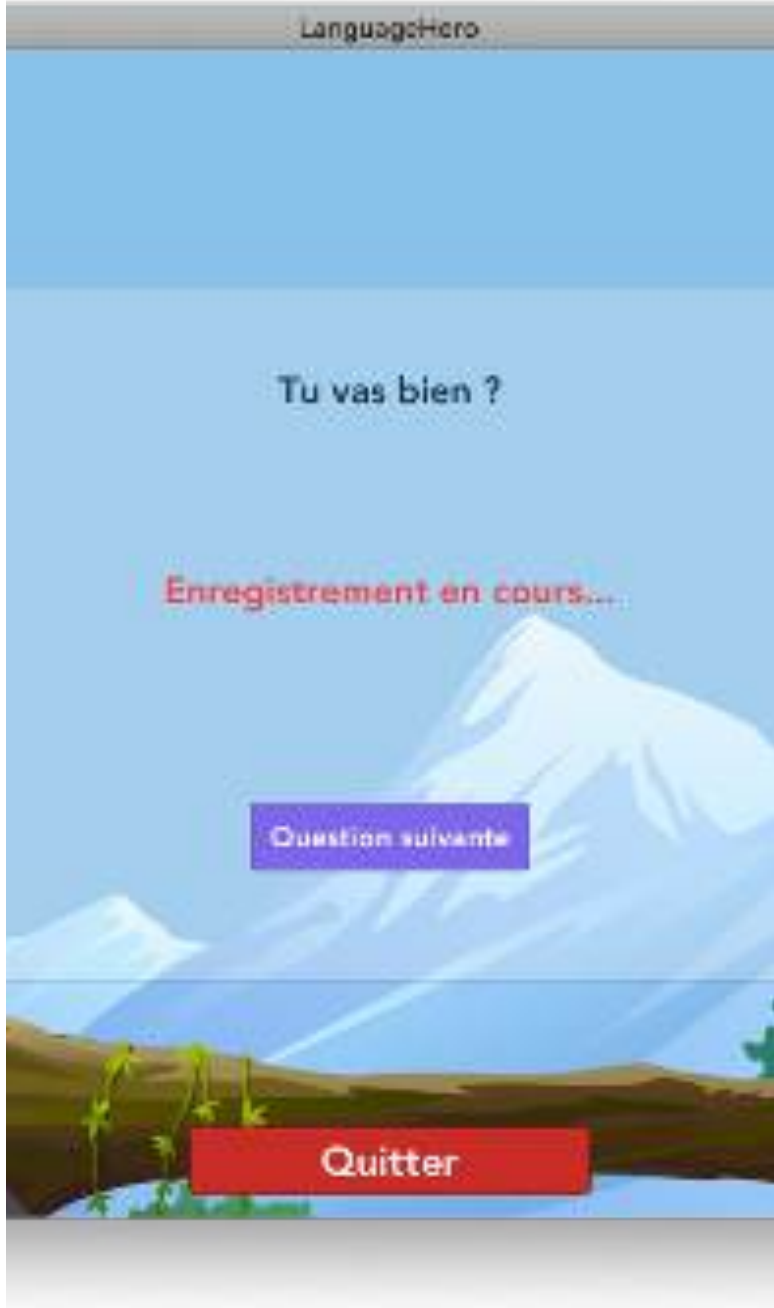
- **Computer-delivered speech test**
- Semi-automatized analysis
- Vocabulary Size for validation

Results & Discussion

- Comparison of annotations & metrics
- Best predictors of L2 proficiency
- Developmental sensitivity

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



Background: Fluency ⇔ Proficiency

- Metrics of utterance fluency & proficiency

Data & Methods

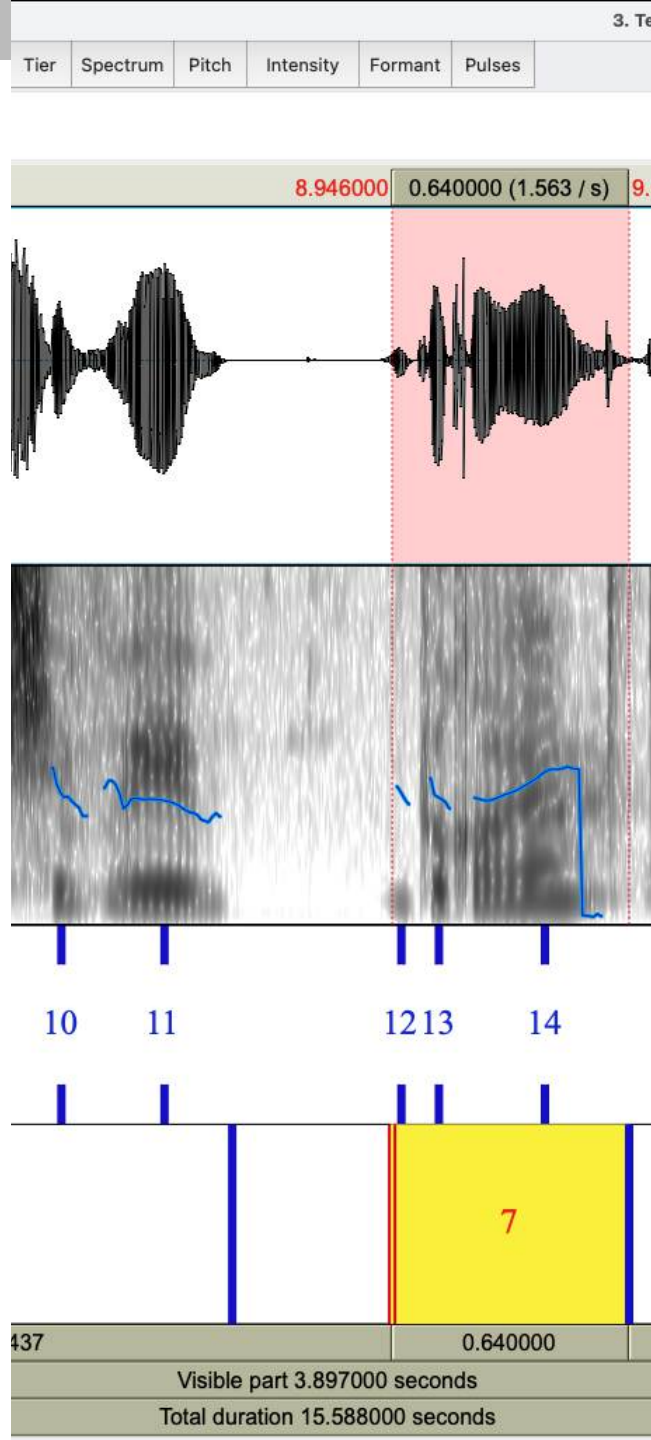
- Computer-delivered speech test
- **Semi-automatized analysis**
- Vocabulary Size for validation

Results & Discussion

- Comparison of annotations & metrics
- Best predictors of L2 proficiency
- Developmental sensitivity

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)
  - (Future: automated detection of filled pauses with new v3 script)
- Automated computation of **# syllables** from transcript
  - with different pruning alternatives



Background: Fluency ⇔ Proficiency

- Metrics of utterance fluency & proficiency

Data & Methods

- Computer-delivered speech test
- Semi-automatized analysis
- **Vocabulary Size for validation**

Results & Discussion

- Comparison of annotations & metrics
- Best predictors of L2 proficiency
- Developmental sensitivity

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)

Background: Fluency ⇔ Proficiency

- Metrics of utterance fluency & proficiency

Data & Methods

- Computer-delivered speech test
- Semi-automatized analysis
- Vocabulary Size for validation

Results & Discussion

- **Comparison of annotations & metrics**
- Best predictors of L2 proficiency
- Developmental sensitivity

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: Fluency ⇔ Proficiency

- Metrics of utterance fluency & proficiency

Data & Methods

- Computer-delivered speech test
- Semi-automatized analysis
- Vocabulary Size for validation

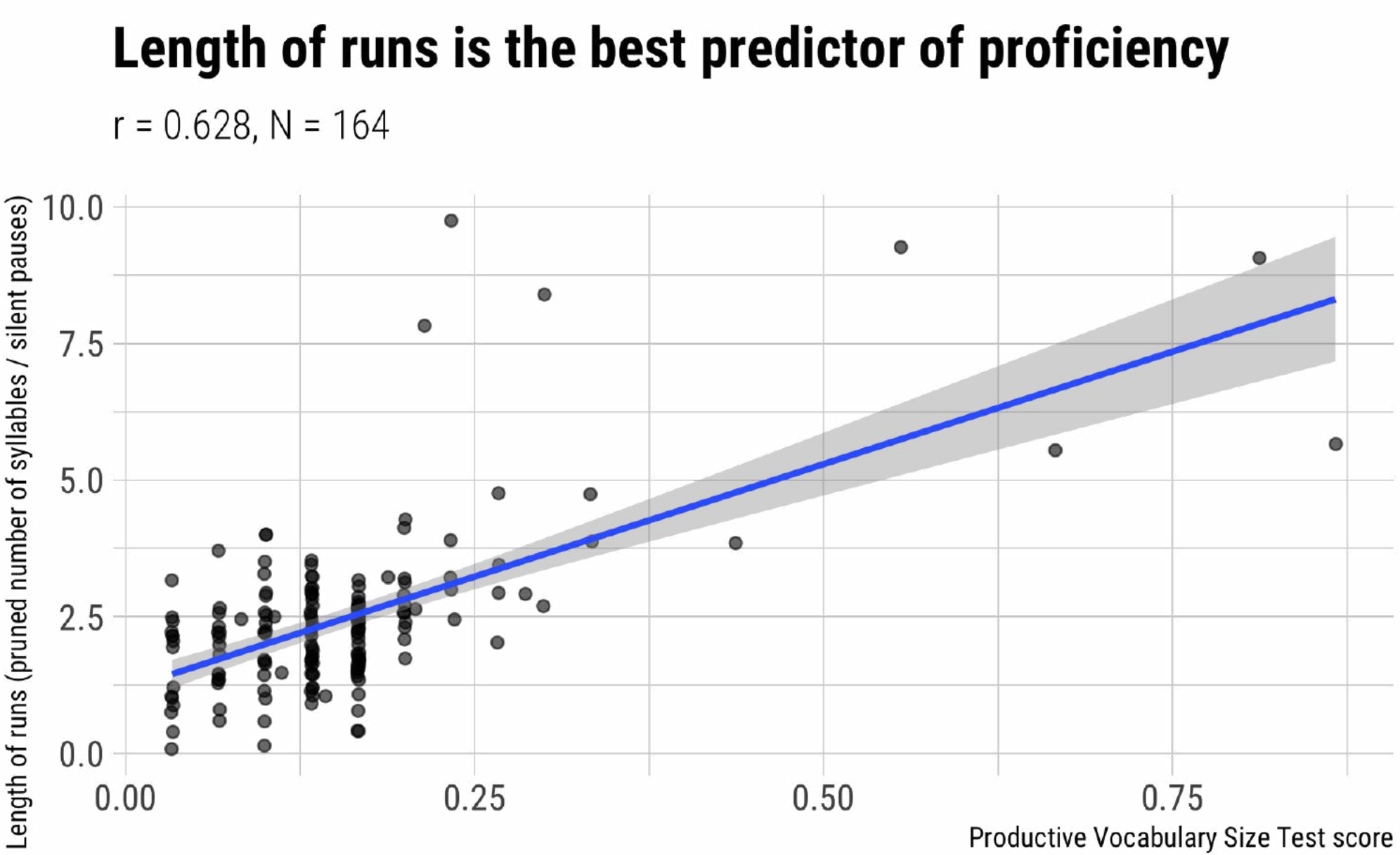
Results & Discussion

- Comparison of annotations & metrics
- **Best predictors of L2 proficiency**
- Developmental sensitivity

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



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

<SR: possibly due to lower quality auto `phonation time`

'Raw' metric suprisely useful for this type of speech

Based on correlation with Vocabulary Size, Pearson's r

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

Semi-auto vs. fully automated composite metrics

Metric	Semi-auto, pruned	Fully auto*, ASR-based count	Fully auto*, signal-based (deJong)	Fully auto signal alt.
Length of runs	**.628**	.588	.479	
Speech rate	**.609**	.585	.461	
Articulation rate	**.524**	.496	.392	.172
Syllable duration^{-1}	**.473**	.283	**.473**	.106
Number of syllables	**.473**	.370	.154	
Number of words	**.463**	.355	—	
Silent pausing rate^{-1}			.409	**.428**
Duration of runs			.338	**.352**
Speech-time ratio			.269	**.305**

* Fully automated metrics are not pruned

Background: Fluency ⇔ Proficiency

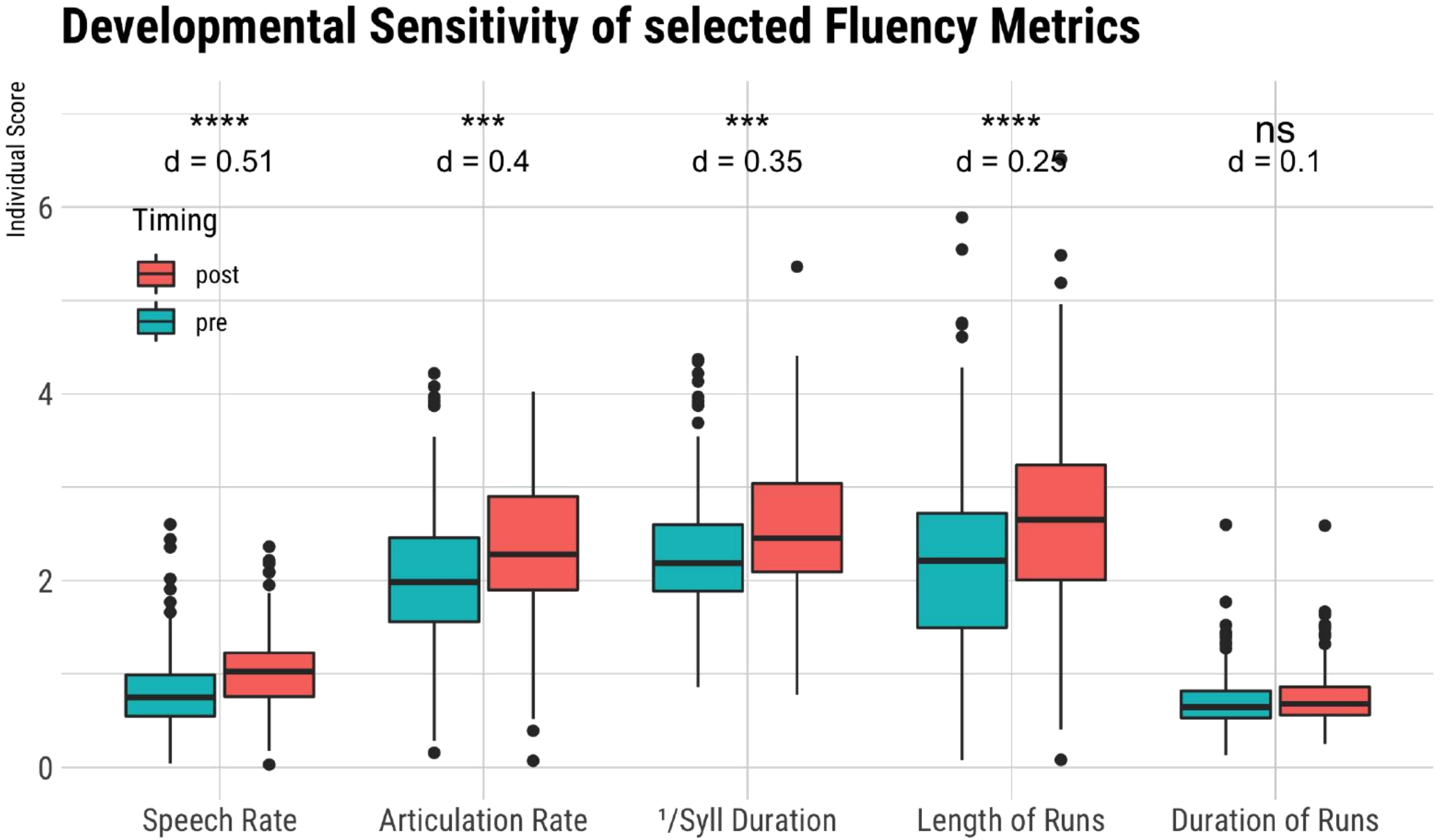
- Metrics of utterance fluency & proficiency

Data & Methods

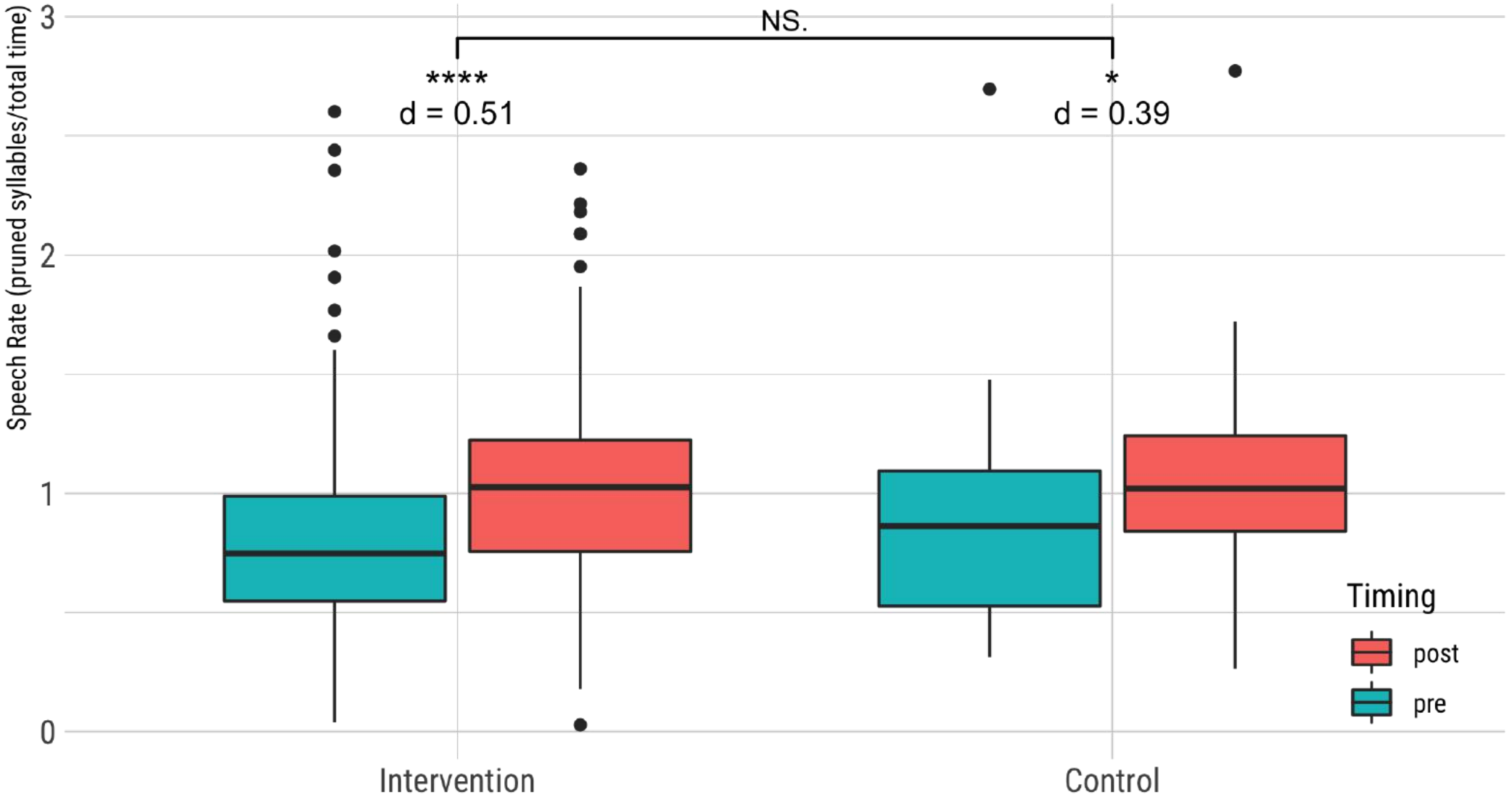
- Computer-delivered speech test
- Semi-automatized analysis
- Vocabulary Size for validation

Results & Discussion

- Comparison of annotations & metrics
- Best predictors of L2 proficiency
- **Developmental sensitivity**


Developmental Sensitivity of selected Fluency Metrics
Individual Score
6
4
2
0
Timing
post
pre
****
d = 0.51
***
d = 0.4
***
d = 0.35
****
d = 0.25
ns
d = 0.1
Speech Rate
Articulation Rate
1/Syll Duration
Length of Runs
Duration of Runs

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



Automatizing L2 Fluency Measurement

- **Automated metrics work!**
  - Fully automated only slightly less accurate than human transcript (max $diff_r = 0.04$)
  - ASR-based count of syllables more reliable than syllable nuclei detection (exc. Syll. dur.)
- Harsh **pruning** improves predictive power.
- Best predictors of **L2 proficiency**:
  - Length of Runs > Speech Rate > Artic. Rate > Syll. Duration^{-1} > #Syll. > Silent Pausing Rate^{-1}
- Best developmental **sensitivity**:
  - Speech Rate > Artic. Rate > Syll. Duration^{-1} > Length of Runs




Questions, feedback & suggestions welcome!

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

Louis Escouflaire

Thomas François

Piet Desmet

Download the slides
References & details
[https://cutt.ly/fluency]

R scripts: e-mail me!

Credits: Pictures from Unsplash and Author.