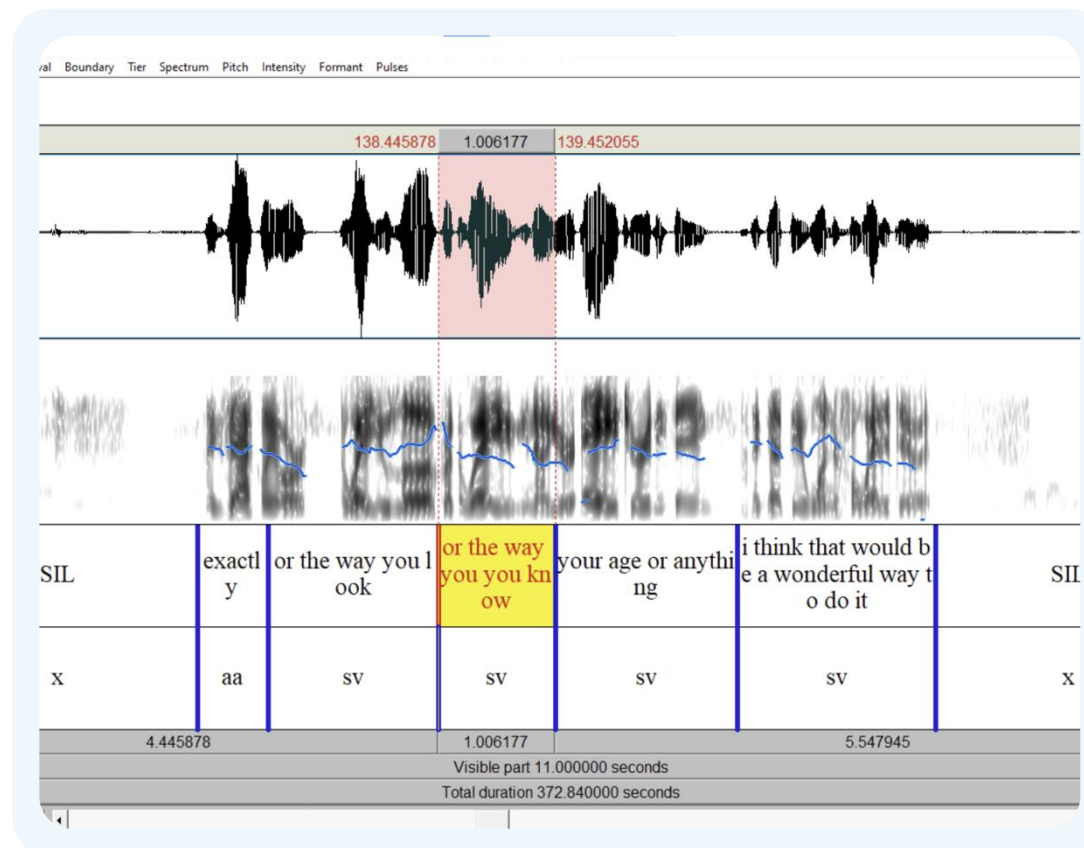


Disentangling Annotator Skill from Verifier Strictness

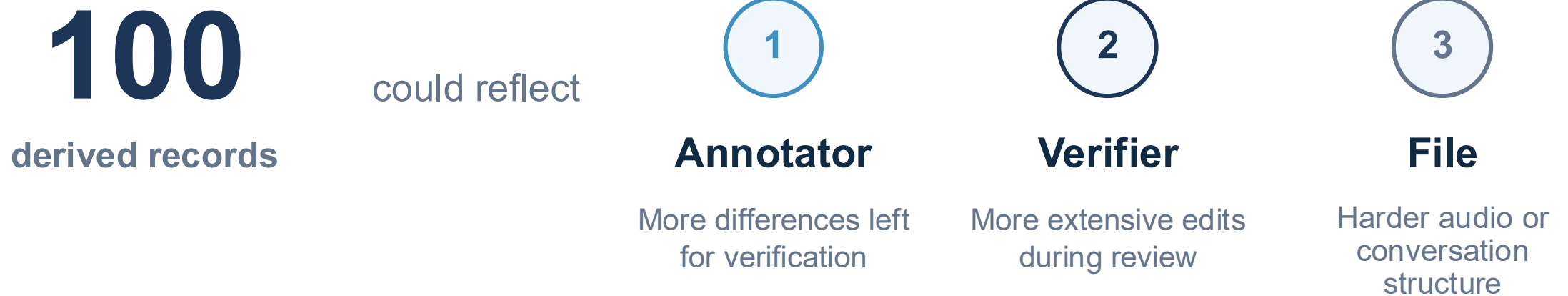
in Cross-Verified Dialogue Annotation

Zihao Tao · John A. Prado · Ignazio Steven LaManna
Ryan Puterbaugh · Mim Datta · Julia Hirschberg



What does a correction count measure?

One observed count can have three workflow sources

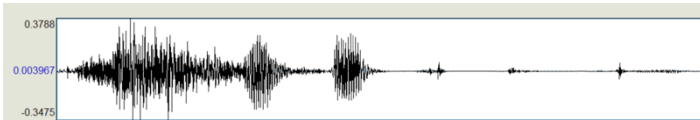


Observed record count $\neq$ annotator quality

Analysis records come from comparing two TextGrid versions

1 · Annotate

Realign SwDA intervals to the original audio



1,045

file sides

2 · Verify

Review and edit the submitted TextGrid against the audio

ANALYSIS UNIT
One algorithm-derived change record

105,005

derived records

Analysis · Compare versions

Match submitted and verified TextGrids

MATCH → CLASSIFY → COUNT
5 DERIVED CHANGE TYPES

5

derived change types

Crossed assignments separate annotator and verifier variation

Annot.	Verifier				
	T1	T2	T3	T4	T5
T1	—	164	63	13	43
T2	49	—	44	16	13
T3	38	38	—	46	10

Each identified annotator

was checked by all four other team members

537

complete cases
with both roles
identified

Off-diagonal overlap makes the two role effects separately estimable.

The model asks the same question about both roles

Outcome used in the paper

normalized correction rate = derived change records / non-silence intervals

Do annotators differ

after controlling for
who verified the file?

Do verifiers differ

after controlling for
who annotated the file?

Fixed effects annotator batch index · verifier batch index · log(non-silence intervals)

Random intercepts annotator · verifier · batch within annotator

The residual dominates, and the verifier estimate is zero

REML variance components · n = 537 complete cases

10.8%

Annotator identity

modest component

0.0%

Verifier identity

singular-fit boundary

80.6%

File-level residual

difficulty + measurement noise

8.7%

batch within annotator

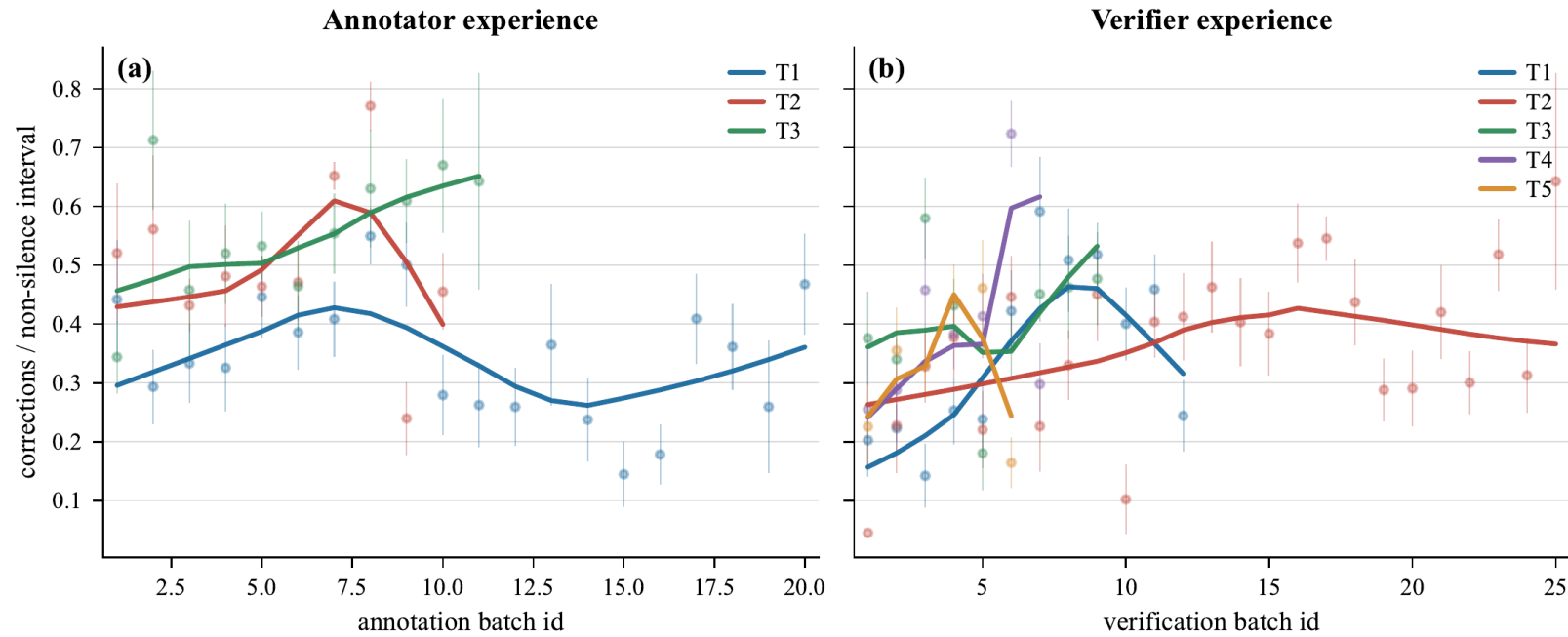


10.8% annotator **8.7% batch**

80.6% file residual

No detectable verifier variation
≠ no value from verification

No consistent learning trend appears across batches



Annotator batch

$p = 0.60$

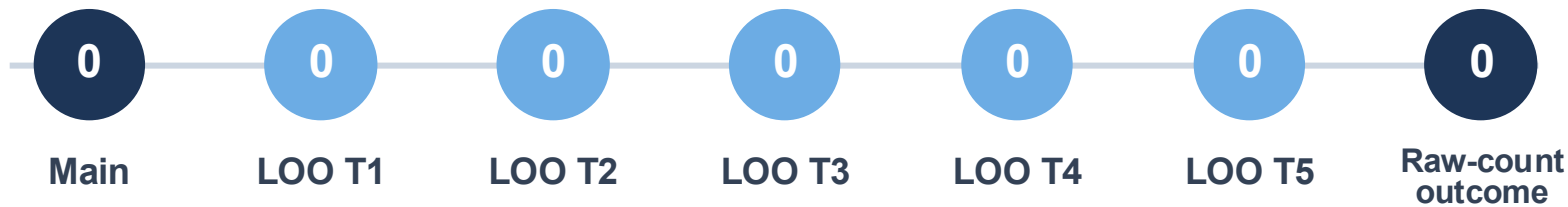
Verifier batch

$p = 0.06$

No consistent team-wide downward drift

All seven fits estimate verifier variance at zero

$\sigma^2_v = 0$ in all seven fits



Sensitivity limit

Only 5 verifier groups
Main fit: 537 complete cases
Small effects may remain undetected

No detectable verifier variation $\neq$ identical verifiers

Most derived change records involve boundary placement

Share of 105,005 derived change records



58.5%

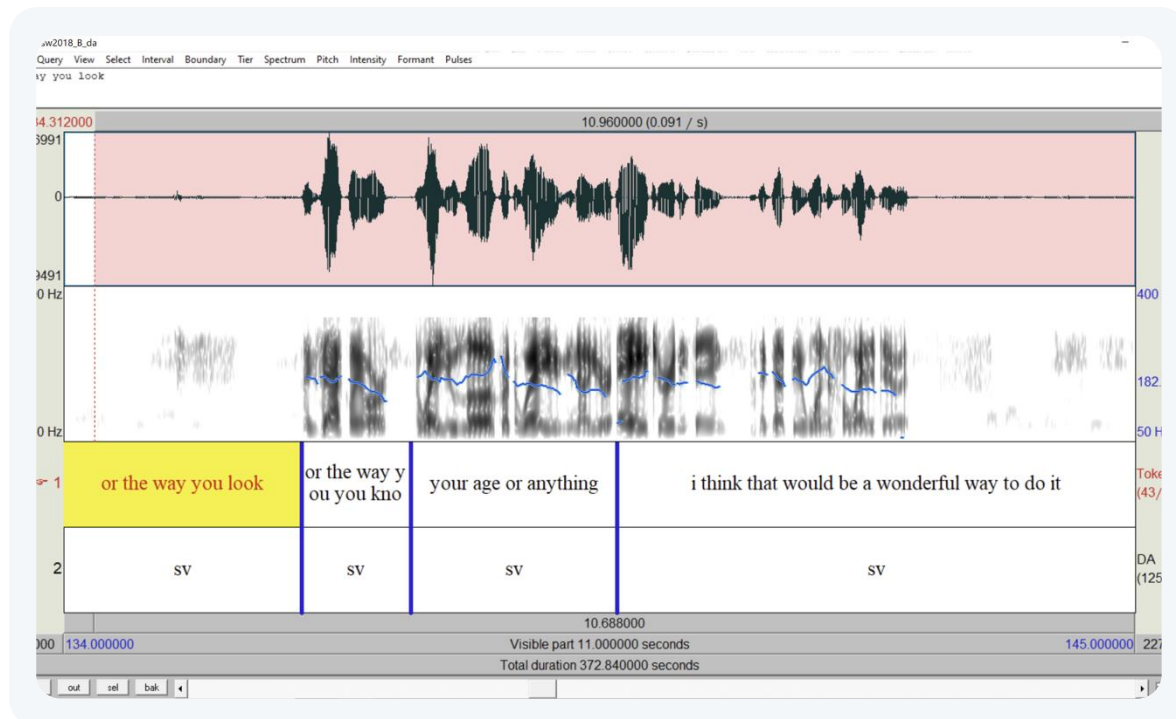
audio-anchored
boundary shifts

3.0%

pure dialogue-act
label changes

Boundary-shift records are audio-anchored, not pure label judgments.

Audio constrains boundaries, while label-related shares vary by annotator



Audio constrains where speech begins and ends

Semantic choices leave room to diverge

sd ↔ sv

b / aa / ny

Label-related share of derived records

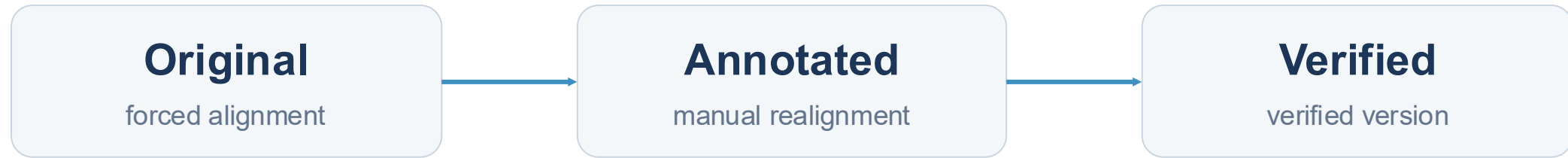
542 file sides with identified annotator



T3: 12.8% DA_CHANGE + 7.5% MODIFY

Descriptive shares, not skill scores

A version difference is not necessarily an annotator error



High-confidence same-text boundary matches · n = 5,108



Interpret the derived rate as a workflow measure, not an error rate.

Make verification auditable

1

Preserve evidence

Log edit events directly,
or derive typed records
from versioned files with
metadata.

2

Cross assignments

Use at least two verifiers
per annotator, with overlap.

3

Share data and methods

Share versions, derived records,
metadata, and comparison code.

For RASwDA: prioritize boundary-placement support.

Suggested by record composition; intervention not evaluated.

A correction count reflects the whole workflow

- **Crossed assignments separate the sources**

separate annotator, verifier, batch, and residual components of the derived rate.

- **10.8% vs. 0.0%**

Modest annotator variation; no detectable verifier-identity variation.

- **58.5% boundary-shift records**

Audio anchoring is consistent with limited room for verifier-specific differences.

Questions?

RASwDA realigns SwDA to the original audio

Original SwDA alignment

GMM-HMM forced alignment

Errors are common in quiet speech, overlap, noise, and short backchannels.

RASwDA workflow

Manual Praat realignment
+ verification of every file

Inherited DA labels remain unless
realignment makes them untenable.

LAW paper: corpus + downstream value | SIGDIAL paper: what the workflow teaches us

Coverage and crossed design

Annot.	T1	T2	T3	T4	T5	Unv.
T1	—	164	63	13	43	5
T2	49	—	44	16	13	0
T3	38	38	—	46	10	0

503 rows lack recoverable annotator identity and are excluded from the mixed-effects fit.

1,045 total file sides

542 identified annotator

1,025 identified verifier

537 complete cases

Full mixed-effects specification

$$y_i = \beta_0 + \beta_1 \text{ann_batch}_i + \beta_2 \text{ver_batch}_i + \beta_3 \log n_i^{\text{int}} \\ + u^A_{a(i)} + u^V_{v(i)} + u^B_{a(i), \text{ann_batch}_i} + \varepsilon_i$$

Effect	Estimate	95% CI	p
Annotation batch	-0.002	[-0.011, 0.007]	0.60
Verification batch	-0.004	[-0.009, 0.000]	0.06
log(non-silence intervals)	-0.008	[-0.090, 0.075]	0.86

REML fit with lme4 on N = 537 complete cases

Sensitivity checks and statistical limitations

Seven fits

Main complete case	N=537	$\sigma^2_v = 0$
Leave out John	N=335	$\sigma^2_v = 0$
Leave out Iggy	N=430	$\sigma^2_v = 0$
Leave out Ryan	N=462	$\sigma^2_v = 0$
Leave out Zihao	N=450	$\sigma^2_v = 0$
Leave out Mim	N=471	$\sigma^2_v = 0$
Raw-count response	N=537	$\sigma^2_v = 0$

Informative missingness

503 unattributed rows

Median length: 282 vs. 231 intervals

Median derived records: 51 vs. 91

Only 3 annotators and 5 verifiers

No annotator-profile covariates

No annotator × verifier interaction

Read the partition as evidence about this workflow, not a population estimate.

Derived action types and boundary provenance

Annot.	Boundary	DA	Insert	Delete	Modify	High-confidence derived provenance	
T1	64.9	0.5	27.8	5.9	0.9	50.3%	Further refinement
T2	61.1	1.3	23.1	12.9	1.5		
T3	41.4	12.8	23.8	14.4	7.5	32.2%	Original error unchanged
Corpus	58.5	3.0	26.1	9.4	2.9	16.2%	Moved away from final
Exact derived-record counts						1.35%	Neutral/equivalent
61,417 boundary · 27,425 insert · 9,922 delete · 3,144 DA · 3,097 modify							