A Shared English–Japanese Pipeline for Automatic Clause-Based Fluency Annotation: Staged Validation and Preliminary Evidence for Cross-Lingual Transportability

Author names and affiliations to be inserted

Revised draft, February 2026

Abstract

Automatic annotation of temporal speech features can scale second-language (L2) fluency research, but validity must be demonstrated at both component and measure levels. We developed a shared English–Japanese pipeline that computes nine clause-based utterance fluency (UF) measures from audio through five stages: ASR, filler-augmented forced alignment, disfluency detection, clause segmentation, and UF computation. Evaluation followed the same staged design in both languages. In this release, English has full staged validation (RQ1 component-level agreement, RQ2 pause-location agreement, and RQ3 concurrent validity), while Japanese validation is currently limited to measure-level concurrent validity (RQ3 only); Japanese component-level analyses (RQ1 and RQ2) are pending completion under the same protocol, and the cross-lingual component-level validation is therefore partial in this release. In English ($N = 40$), automatic clause boundaries showed almost perfect agreement with gold references (micro $F1 = .845$, $\kappa = .816$), and automatic MCP/ECP pause-location labels also showed almost perfect agreement with gold labels ($\kappa = .840$, accuracy $= .921$). Automatic measures correlated strongly with manual references for English ($N = 39$; mean Pearson $r = .936$, 95% CI $[.878, .965]$), supported by component-level agreement, and Japanese measure-level correlations were similarly high ($N = 40$; mean Pearson $r = .953$, 95% CI $[.916, .975]$), though Japanese component-level confirmation is pending. Relative to previously reported monologic L2 English systems, English component-level point estimates are numerically higher, though cross-study differences in corpora, gold standards, and evaluation methods preclude direct comparison. Across both languages, pause-duration measures were the most sensitive to residual pipeline error.

Keywords: L2 speech, utterance fluency, automatic annotation, clause segmentation, ASR, cross-lingual concurrent validity, English, Japanese

1 Introduction

1.1 Objective measurement of L2 oral fluency

Objective fluency measurement is central to second language (L2) speech research and assessment. Utterance fluency (UF), the temporal characteristics of speech production, is conventionally decomposed into three dimensions: speed fluency (SF), breakdown fluency (BDF), and repair fluency (RF) (Tavakoli and Skehan, 2005). SF reflects the pace and density of information delivery and is closely related to cognitive processing speed (Suzuki and Kormos, 2023). BDF captures pausing behavior and reflects disruptions in speech production processing; crucially, mid-clause pauses (MCPs) and end-clause pauses (ECPs) carry distinct psycholinguistic significance, as MCPs more strongly predict listener-based fluency judgments than ECPs (de Jong, 2016; Suzuki et al., 2021). RF is associated with disfluency phenomena such as repetitions, self-corrections, and false starts (Kormos, 2006).

A recent meta-analysis of the relationship between UF measures and perceived fluency (PF) ratings confirmed that speed and composite measures are most strongly associated with fluency judgments ($r = |.62–.76|$), followed by pause frequency ($r = .59$) and pause duration ($r = .46$), while RF measures show a weaker but significant association ($r = .20$) (Suzuki et al., 2021). Furthermore, the relationship is moderated by pause location: MCPs contribute more strongly to perceived fluency prediction ($r = .72$) than ECPs ($r = .48$). These findings underscore the importance of computing a comprehensive set of clause-based UF measures, including location-specific pause metrics, rather than relying on global temporal indicators alone.

A persistent methodological challenge concerns the unit of analysis for BDF. The clause is the standard unit for classifying pause location (Foster et al., 2000), yet operational definitions vary considerably across studies. Vercellotti and Hall (2024) argued for a broader clause framework in L2 research that encompasses not only finite clauses but also coordinated verb phrases with complements, nonfinite clauses with overt elements, and copula-less predicates. Their "verb + element" principle provides a principled and replicable standard: a verbal construction reaches clause status when the verb has a complement or adjunct. The present pipeline operationalizes this framework explicitly.

1.2 Automatic annotation of temporal features

Although objective UF measures are informative, producing them by hand requires substantial annotator time for pauses, disfluency words, and boundary coding, which constrains dataset size in practice. Early automation therefore focused on acoustic processing. de Jong and Wempe (2009) developed a Praat-based approach for detecting syllable nuclei and silent pauses, and de Jong et al. (2021) extended this line to filled-pause processing. Their approach relies on acoustic signatures such as segment duration and pitch/formant behavior, and reported English-corpus filled-pause detection accuracy in the mid-.80 range. Related acoustic systems have also been proposed for real-time fluency feedback (Rose, 2020).

These acoustic approaches are useful for detecting pause and filler events, but they are limited for clause-based fluency analysis. Without transcript-linked syntactic structure, they cannot reliably assign pause location relative to clause boundaries (e.g., MCP vs. ECP), and

they cannot directly represent repair phenomena needed for RF metrics. This limitation motivated transcription-aware ML pipelines. Chen and Yoon (2011) proposed NLP-based structural event detection, and a follow-up ASR-output evaluation reported clause-boundary detection at $F1 = .690$ (Chen and Yoon, 2012). Matsuura et al. (2022) further combined BERT-based disfluency pruning with dependency parsing, reporting substantial agreement for disfluency detection and pause-location classification ($\kappa = .674$ and $\kappa = .613$).

A remaining challenge is robustness across tasks and proficiency profiles. ASR and downstream NLP modules are sensitive to train-test mismatch, and errors can propagate across pipeline stages (Knill et al., 2018, 2019). In task-generalization work, Skidmore and Moore (2023) showed that BERT-based disfluency detection remained effective on an unseen learner corpus but declined from in-domain to cross-corpus testing ($F1 = 92.8$ on NICT-JLE test vs. 82.2 on KISTEC), with lower scores for less constrained activity types. Matsuura et al. (2025) likewise showed modality-conditioned variation, with generally stronger results in monologic than dialogic conditions and non-uniform agreement across measures and fluency bands (pause-location κ ranging from .596 to .749 across subsets). These findings motivate validation designs that report component-level and measure-level outcomes separately and stratify results by task and speaker profile.

1.3 Gaps and motivation

Three gaps motivate the present study. First, cross-lingual validation evidence for clause-based fluency annotation remains limited. Existing systems have been developed and tested mainly in English, leaving open whether a shared processing architecture can produce valid measures across typologically different languages. English and Japanese present contrasting challenges: different word orders (SVO vs. SOV), subject expression (overt vs. pro-drop), and clause-chaining structures (coordination vs. te-form chains), all of which bear on clause segmentation.

Second, prior validation studies have typically reported measure-level correlations without component-level checks on intermediate outputs. High end-to-end correlations can mask weak boundary or pause-label agreement, especially if errors cancel in aggregation. A staged evaluation design, from component-level agreement to measure-level validity, provides more diagnostic evidence for system quality.

Third, operational clause definitions in automated systems remain under-specified. Prior work has relied on implicit parser decisions or corpus-specific rules without explicit reference to a theoretical framework. The present pipeline operationalizes the Vercellotti & Hall clause framework at the script level, making clause-coding decisions inspectable and replicable.

1.4 The current study

To address these gaps, we developed a shared English–Japanese pipeline for automatic clause-based fluency annotation and evaluated it with a staged design. We define three research questions:

RQ1 Clause-segmentation agreement: To what extent do automatically generated clause boundaries agree with human-annotated clause boundaries?

RQ2 Pause-location agreement: To what extent do MCP/ECP pause labels from the automatic pipeline agree with gold-standard labels?

RQ3 Concurrent validity: To what extent do automatically computed UF measures correlate with manually referenced UF measures?

In the current report, all three English questions and Japanese RQ3 are complete. Japanese RQ1 and RQ2 use the same analysis framework but are pending completion for this release and will be reported in the final version. Table 1 summarizes the completion status.

Language	Question	Status
English	RQ1 (clause-segmentation agreement)	Complete
English	RQ2 (pause-location agreement)	Complete
English	RQ3 (concurrent validity)	Complete
Japanese	RQ1 (clause-segmentation agreement)	Pending
Japanese	RQ2 (pause-location agreement)	Pending
Japanese	RQ3 (concurrent validity)	Complete

Table 1: Completion status of the six research questions.

All pipeline scripts, trained models, gold annotations, and analysis outputs are publicly available as an open release bundle at `[URL_TO_BE_INSERTED]`.

2 Method

2.1 Data

2.1.1 English corpus

The English corpus consists of 40 speech files (20 ST1 picture narrative, 20 ST2 argumentative) selected from the ALLSSTAR corpus (Archive of L1 and L2 Scripted and Spontaneous Transcripts and Recordings). Speakers represent diverse L1 backgrounds, providing a range of proficiency levels and accent patterns. Each speech sample is approximately two minutes in duration. The selected files are documented in the release artifact `selected_files.json`.

2.1.2 Japanese corpus

The Japanese corpus consists of 40 speech files (20 ST1, 20 ST2) from an L2 Japanese speech dataset collected at a Japanese university. Speakers represent diverse L1 backgrounds; detailed demographic information (L1 distribution, proficiency levels) is not reported in this release for privacy reasons but is documented in the restricted metadata accompanying the corpus. The manual gold-standard clause annotations were independently produced by trained human coders following the same Vercellotti-adapted framework and finalized as the `manual_clauses_gold_v2` set. File identifiers and task assignments are documented in the release artifacts.

2.1.3 Gold-reference construction (English)

Gold-standard clause boundary annotations for the English corpus were constructed via an LLM-assisted workflow adapted from Morin and Marttinen Larsson (2025):

1. Two trained coders independently annotated clause boundaries on 10 randomly selected blind files.
2. Disagreements were adjudicated to produce a gold reference for these 10 files.
3. A large language model (Claude) was trained on the adjudicated data and evaluated on 5 locked test files, achieving micro $F1 = .929$ and $\kappa = .914$.
4. The model annotated 30 production files, and outputs were manually reviewed and accepted as gold by the first author.

Because 30 of the 40 gold files were initially annotated by an LLM and reviewed by a single author, shared systematic biases between the LLM-generated annotations and the NLP-based pipeline cannot be fully excluded; the resulting inter-rater reliability ($\kappa = .816$) should therefore be interpreted as an upper-bound estimate relative to a fully independent human gold standard. This yielded 40 gold files (20 ST1, 20 ST2). For RQ3 concurrent validity, one file (`ALL_139_M_PBR_ENG_ST1`) was excluded because the manual annotation included task-instruction preamble speech that was absent from the ASR-segmented output, causing a systematic alignment offset that would distort UF computation. This yielded a quality-filtered cohort of $N = 39$ (ST1 = 19, ST2 = 20).

2.2 Pipeline architecture

The shared release pipeline processes L2 speech through five stages:

$$\boxed{\text{Audio}} \rightarrow \boxed{\text{ASR + word alignment}} \rightarrow \boxed{\text{Disfluency-aware clause segmentation}} \rightarrow \boxed{\text{Neural filler scoring}} \rightarrow \boxed{\text{UF computation}}$$

Both languages follow the same high-level architecture, with language-specific adaptations at each stage. Note that disfluency detection (Stage 2) is embedded within the clause segmentation step (Stage 3): both clause segmenters load the disfluency detector as a preprocessing module before dependency parsing, rather than as a separate pipeline stage.[1] Three design constraints determined the implementation: (a) clause-based UF measures require precise word timing, not only transcript text; (b) one open stack must be usable in both English and Japanese within the same release; and (c) clause coding must be explicit, reproducible, and grounded in the Vercellotti & Hall framework.

[1]The Japanese evaluation corpus additionally required a span-blanking preprocessing step to align manual and ASR coverage of the speech sample. This step is corpus-specific and not part of the general pipeline.

2.2.1 Stage 1: ASR and word-time alignment

We employ Qwen3-ASR (1.7B parameters) as the open-source multilingual ASR front end, paired with Qwen3 Forced Aligner (0.6B parameters) for rough word-level timestamps. ASR is run with a disfluency-preserving prompt that encourages the model to transcribe fillers and hesitations rather than suppressing them, since accurate filler capture is critical for downstream pause metrics. To improve timestamp precision, we subsequently apply Montreal Forced Aligner (MFA) with high beam settings (beam = 100, retry beam = 400), set to their maximum practical values to ensure alignment convergence on disfluent L2 speech.

A key technical innovation is filler-augmented alignment: placeholder filler tokens (e.g., "uh" for English) are injected into inter-word gaps exceeding 400 ms before MFA runs. The number of injected fillers is determined by $k = \lfloor (gap - 0.35)/0.55 \rfloor + 1$, capped at 3 per gap; the gap threshold (350 ms) and filler interval (550 ms) were set based on typical L2 filler durations observed during development, and the cap of 3 prevents over-injection in long pauses. These placeholders are absorbed during forced alignment and discarded in the map-back step. This technique prevents MFA from distributing gap durations across adjacent word boundaries, which would otherwise compress pause onset and offset times and distort downstream pause measures. For English, the `english_us_arpa` acoustic model is used; for Japanese, the `japanese_mfa` model with deterministic filler token pools of varying length (short, medium, long Japanese fillers).

Japanese ASR includes an additional MeCab re-tokenization step (via `fugashi`) before MFA to split merged tokens and stabilize downstream token mapping.

2.2.2 Stage 2: Disfluency detection

Both clause segmenters load a shared neural disfluency detector before clause parsing. The model is a fine-tuned `XLM-RoBERTa-base` token classifier (two labels: fluent/disfluent), trained on a combination of real English disfluency data from Switchboard (Zayats et al., 2019) and synthetic English/Japanese disfluency data (88.5K training sentences total), following the synthetic data augmentation approach described by Kundu et al. (2022). Training parameters included 3 epochs, learning rate 2×10^{-5}, and batch size 16.

The detector serves a pruning function: identified disfluent tokens are suppressed before or during clause assembly. This standardizes syllable/mora counts for rate-based measures (Suzuki and Révész, 2023) and reduces noise in the dependency parse input for clause segmentation. The Japanese segmenter additionally applies deterministic post-rules (repeated-token collapse, elongated-form handling, split-repetition fixes) after neural prediction to improve robustness on conversational artifacts.

2.2.3 Stage 3: Clause segmentation

Both segmenters first apply sentence segmentation using wtpsplit (Segment any Text), a neural sentence boundary detector suitable for unpunctuated L2 speech. Sentence boundaries are enforced as clause boundaries. Each sentence is then dependency-parsed, and clause heads are identified and classified.

English. The English segmenter applies a rule layer over spaCy dependency parses (transformer model), explicitly operationalizing the "verb + element" clause logic of Vercellotti and Hall (2024). A token is considered a potential clause head if it carries a verb-like POS tag (VB*, VERB, AUX) and a clausal dependency relation. The "verb + element" check tests whether the verb has at least one qualifying dependent among complements (`obj`, `iobj`, `ccomp`, `xcomp`, `attr`), oblique arguments (`obl`), and adjuncts (`advmod`, `prep`). The segmenter classifies the following constructions:

- Independent clauses: ROOT-level finite verbs.
- Subordinate clauses: `advcl`, `ccomp`, `acl`, `acl:relcl`, `csubj` relations.
- Coordinated VPs: `conj` of a verb, but only if the conjunct has its own complement or adjunct (stricter than Vercellotti's inclusive approach for shared complements; this produces fewer clause boundaries than a fully Vercellotti-inclusive implementation, which may slightly affect MLR and clause-count-dependent measures).
- Nonfinite clauses: `xcomp` and participial constructions, with Vercellotti's "verb + element" requirement.
- Minor clauses: stance verbs (e.g., "I think", "I believe") taking a complement clause, tagged separately for analysis.
- Copula-less predicates: ROOT-level ADJ/NOUN without overt copula (common in L2 speech).

Clause spans are collected via subtree traversal (excluding other clause-head subtrees) and aligned back to TextGrid word intervals for timing.

Japanese. The Japanese segmenter uses GiNZA (`ja_ginza_electra`) with a Japanese Universal Dependencies rule layer adapted to the same theoretical principle. Predicate detection includes VERB, ADJ, and NOUN tokens; misparsed proper nouns tagged as verbs are filtered. A key language-specific adaptation concerns te-form verb chains, which are pervasive in Japanese speech. Following the Vercellotti "verb + element" rule, a te-form verb receives clause status only if it has its own complement or adjunct; bare te-forms are merged with the following clause. The complement check uses a strict set (`obj`, `iobj`, `obl`, `advmod`, `nmod`), deliberately excluding `nsubj` because shared and implied subjects are common in Japanese. The segmenter also handles explicit subordinators ("kara" 'because', "kedo" 'although', "node" 'because', "ba" 'if'), tari-form chains, and the clausal subject (`csubj`) construction.

Why explicit rules? Prior L2 fluency studies often under-specify operational clause decisions, making downstream pause and fluency measures difficult to compare across studies. We therefore expose all clause-coding decisions in inspectable, script-level rules grounded in Vercellotti and Hall (2024), improving reproducibility and enabling principled cross-lingual comparison.

2.2.4 Stage 4: Neural filler candidate scoring

Modern ASR systems often omit low-energy fillers and disfluencies from transcriptions. If undetected, these speech regions are over-counted as pure silence, biasing pause metrics. We

address this with a post-ASR acoustic filler detector that scores candidate filler regions within ASR-detected gaps.

The gap-only neural filler classifier uses a TC-ResNet8-style temporal CNN over per-clip normalized log-mel spectral features ($1 \times 64 \times 101$). The model was trained on the PodcastFillers dataset (Zhu et al., 2022) with binary labeling (positive: uh/um; negative: words/breath/laughter/music), achieving validation $F1 = .941$ and test $F1 = .933$. At inference, each ASR gap of sufficient length is scored against a threshold of 0.50; accepted filler candidates are used to split or suppress pause intervals before UF computation. The same English-trained model is applied to both languages. While filler-like vocalizations share some acoustic properties cross-linguistically (low-energy, quasi-periodic), Japanese fillers (e.g., "eto," "ano") differ phonetically from English fillers ("uh," "um"), and the cross-lingual generalization of this classifier has not been formally evaluated (see Limitation 13).

2.2.5 Stage 5: Utterance fluency computation

UF calculators compute nine measures from clause-segmented, filler-adjusted TextGrids (Table 2).[2] The classifier refinement step operates as follows: for each pause interval (≥ 250 ms, following de Jong and Bosker, 2013), any overlap with predicted filler-speech islands is subtracted, and the remaining silence segments are retained as pauses only if they individually meet the 250 ms threshold. Each resulting pause is then classified as MCP or ECP: a pause is labeled ECP if its onset falls within 150 ms of any clause offset; otherwise it is labeled MCP. The 150 ms onset window was chosen to accommodate the temporal imprecision of forced-alignment boundaries: MFA word boundaries typically carry alignment uncertainty on the order of 50–100 ms, and a 150 ms tolerance absorbs this uncertainty without reclassifying clearly mid-clause pauses. No formal sensitivity analysis across alternative thresholds has been conducted (see Limitation 12). Pauses that fall outside all clause boundaries (e.g., before the first clause or after the last) default to ECP.

English uses syllable-based normalization; syllable counts are derived from the phone tier when available, or estimated heuristically from orthographic form. Japanese uses mora-based normalization.

The current release computes nine speed and breakdown fluency measures; repair fluency (RF) measures are not included. The pipeline's disfluency detector can identify repetitions, self-corrections, and false starts, and the codebase supports RF computation. However, the manual reference annotations used for concurrent validity (RQ3) do not include disfluency-level labels, so RF measures could not be validated in this study. Extending the validation to include RF measures requires manual disfluency annotation of the reference corpus, which is a direction for future work (see Limitation 11).

[2]The release codebase uses the label "CAF calculator" in script file names (e.g., `caf_calculator.py`) for backward compatibility. In the paper we use "UF" because the pipeline computes only fluency measures; no complexity or accuracy measures are included in this release.

Type	Measure	Description
Speed	AR	Articulation rate: syllables (morae) per phonation time
Composite	SR	Speech rate: syllables (morae) per total speech time
	MLR	Mean length of run: mean syllables (morae) between pauses
Breakdown (freq.)	MCPR	Mid-clause pause ratio: MCPs per syllable (mora)
	ECPR	End-clause pause ratio: ECPs per syllable (mora)
	PR	Pause ratio: all pauses per syllable (mora)
Breakdown (dur.)	MCPD	Mean mid-clause pause duration (s)
	ECPD	Mean end-clause pause duration (s)
	MPD	Mean pause duration (s)

Table 2: Nine utterance fluency measures computed by the pipeline. The current release computes speed and breakdown fluency only; repair fluency measures are not included (see Limitation 11). Minimum pause threshold: 250 ms, following de Jong and Bosker (2013). English rates are syllable-normalized; Japanese rates are mora-normalized.

2.3 Evaluation methods

2.3.1 RQ1: Clause-segmentation agreement

Clause agreement was evaluated as per-word binary boundary classification after minimum-edit-distance alignment between canonical (manual transcript) and ASR token sequences, replicating the alignment logic of NIST's SCTK tool (Chen and Yoon, 2012). At each alignment position:

- **Correct/Substitution:** gold and auto boundary labels are compared directly at the aligned position.
- **Deletion** (manual-only word): auto label is set to 0 (pipeline penalized for missing word).
- **Insertion** (ASR-only word): gold label is set to 0.

This follows the strict speech-evaluation alignment logic of Chen and Yoon (2012) rather than longest-common-subsequence matching, which can produce optimistically inflated agreement by skipping substitution sites. Metrics include micro and macro precision, recall, $F1$, and Cohen's κ (Landis and Koch, 1977).

2.3.2 RQ2: Pause-location agreement

For each silent pause $\geq$ 250 ms in the automatic words tier, two MCP/ECP labels were obtained: (a) the automatic label from auto clause intervals, and (b) a gold label derived

by projecting gold clause boundaries onto auto word timing via the edit-distance alignment from RQ1. A pause was classified as ECP if its onset fell within 150 ms of any clause offset; otherwise MCP. Metrics include Cohen's κ, accuracy, and per-class precision, recall, and $F1$.

The strength of agreement in terms of Cohen's κ was interpreted following Landis and Koch (1977): $<$ .20 slight, .21–.40 fair, .41–.60 moderate, .61–.80 substantial, $>$ .80 almost perfect.

2.3.3 RQ3: Concurrent validity

For each of the nine UF measures, we computed Pearson r, Spearman ρ, ICC(2,1), and mean absolute error (MAE) between automatic and manual outputs. English uses the quality-filtered cohort ($N = 39$; see Section 2.1.3); Japanese uses $N = 40$.

2.3.4 Sample size considerations

For RQ1 and RQ2, the units of analysis are boundary positions ($N = 1{,}131$) and individual pause events ($N = 1{,}902$), respectively. According to power analyses for Cohen's κ, the minimum sample sizes to detect substantial agreement ($\kappa > .61$) with power = .80 ($1 - \beta$) are 151 for binary annotation tasks and 67 for two-category classification (Donner and Eliasziw, 1987). Both RQ1 and RQ2 exceed these thresholds. We report observation-level κ following the convention in the segmentation evaluation literature; Matsuura et al. (2025), for example, applied the same power calculation to $N = 14{,}460$ word positions and $N = 39{,}154$ pause events across 128 speakers without clustering adjustment. Boundary positions and pause events are nested within speakers, meaning speaker-level clustering may affect confidence interval width, but no correction is applied here, consistent with prior work. For RQ3 concurrent validity, the unit of analysis is the speech file ($N = 39$ for English, $N = 40$ for Japanese). For a concurrent validity study where the pipeline is intended to approximate manual measurement, the practically relevant threshold is $r \geq .80$ or higher; at this threshold (two-tailed $\alpha = .05$, power = .80), the minimum sample size is approximately $N = 8$, which both cohorts exceed. Using the field-norm threshold of $r \geq .60$ (Plonsky and Oswald, 2014), the minimum is approximately $N = 19$, also exceeded.

3 Results

3.1 English RQ1: Clause-segmentation agreement

Table 3 presents clause boundary agreement between automatic and gold-standard annotations across the 40 English files.

Overall $\kappa = .816$ indicates almost perfect agreement (Landis and Koch, 1977). Macro means were $F1 = .846$ ($SD = .093$) and $\kappa = .819$ ($SD = .102$). Per-file $F1$ ranged from .618 to 1.000 (median .859). Mean alignment WER between the manual transcript (human-transcribed words from ALLSSTAR) and the ASR output was .121 (12.1%). Precision and recall were well balanced, indicating no systematic over- or under-segmentation.

ST1 (picture narrative) yielded slightly higher agreement than ST2 (argumentative), consistent with lower ASR error rates on structured narrative speech (Knill et al., 2018). The

Subset	Files	Gold boundaries	Precision	Recall	$F1$ (micro)	κ (micro)
Overall	40	1,131	.848	.842	.845	.816
ST1	20	496	.882	.857	.869	.845
ST2	20	635	.822	.830	.826	.795

Table 3: English clause-boundary agreement (automatic vs. gold).

two files with lowest per-file agreement ($\kappa < .65$) both had WER $> .19$, confirming that ASR errors propagate to downstream clause boundary detection.

3.2 English RQ2: Pause-location agreement

Table 4 presents pause-location classification agreement.

Subset	Files	Pauses	κ	Accuracy	MCP $F1$	ECP $F1$
Overall	40	1,902	.840	.921	.929	.912
ST1	20	822	.873	.937	.941	.932
ST2	20	1,080	.815	.909	.920	.894

Table 4: English pause-location agreement (automatic vs. gold MCP/ECP labels).

Overall $\kappa = .840$ indicates almost perfect agreement. MCP classification ($F1 = .929$) slightly outperformed ECP ($F1 = .912$), as expected: mid-clause pauses are more clearly positioned within clause boundaries, while end-clause pauses near clause offsets are sensitive to boundary placement. Macro-level per-file statistics: mean accuracy = .921 ($SD = .052$), median = .927, range .789–1.000. Six files achieved perfect pause-location accuracy.

The pattern across tasks parallels RQ1: ST1 accuracy exceeds ST2. Notably, the file with the lowest RQ2 accuracy (.789) also had the lowest RQ1 agreement ($\kappa = .578$), confirming the error propagation chain from ASR errors through clause boundaries to pause classification, the fundamental cascading concern identified by Knill et al. (2018, 2019).

3.3 English RQ3: Concurrent validity ($N = 39$)

Table 5 presents overall concurrent validity for the quality-filtered English cohort.

Overall English correlations span .821–.988 (mean $r = .936$; 95% CI for the mean r via Fisher Z-transformation: [.878, .965]). All nine Pearson correlations are large ($r > .60$; Plonsky and Oswald, 2014), and all ICC values exceed .79, indicating good to excellent absolute agreement (Koo and Li, 2016). Confidence intervals for individual measures and ICC values are not reported here but should be considered when interpreting per-measure point estimates, particularly for task-level cells where $N = 19$–20 yields wide intervals.

By task, the picture narrative (ST1, $N = 19$) is most challenging for MCPD ($r = .713$, ICC = .669, MAE = 0.097), while the argumentative task (ST2, $N = 20$) remains strong across all measures (minimum $r = .910$ for MCPD). ST1 also shows relatively lower ECPR agreement ($r = .774$, ICC = .775). ST2 is uniformly strong, with the weakest measure being

Measure	Pearson r	Spearman ρ	ICC(2,1)	MAE
AR	.956	.911	.953	0.157
SR	.988	.978	.984	0.082
MLR	.971	.968	.965	0.454
MCPR	.966	.981	.962	0.011
ECPR	.864	.889	.866	0.008
PR	.961	.965	.958	0.014
MCPD	.821	.808	.799	0.090
ECPD	.938	.912	.935	0.115
MPD	.957	.908	.944	0.074

Table 5: English RQ3 overall concurrent validity (quality-filtered cohort, $N = 39$).

ECPR ($r = .935$). Per-task correlations should be interpreted with caution given limited per-cell sample sizes ($N = 19$–20); for instance, a Fisher Z-derived 95% CI for MCPD ST1 $r = .713$ is approximately $[.380, .882]$, and ICC values at these sample sizes carry similarly wide intervals. Table 6 presents the task-level breakdown.

	ST1 ($N = 19$)				ST2 ($N = 20$)			
Measure	r	ρ	ICC	MAE	r	ρ	ICC	MAE
AR	.958	.916	.958	0.153	.950	.860	.948	0.161
SR	.991	.981	.987	0.083	.985	.985	.983	0.082
MLR	.969	.965	.961	0.434	.972	.962	.968	0.472
MCPR	.966	.984	.965	0.013	.977	.974	.960	0.010
ECPR	.774	.781	.775	0.010	.935	.956	.935	0.006
PR	.954	.966	.952	0.015	.977	.956	.965	0.013
MCPD	.713	.741	.669	0.097	.910	.865	.886	0.084
ECPD	.916	.905	.904	0.141	.963	.920	.965	0.092
MPD	.934	.868	.920	0.093	.981	.958	.970	0.056

Table 6: English RQ3 concurrent validity by task.

3.4 Japanese RQ3: Concurrent validity ($N = 40$)

Table 7 presents overall concurrent validity for the Japanese cohort.

Japanese overall correlations span .903–.992 (mean $r = .953$; 95% CI $[.916, .975]$), slightly exceeding the English mean. All nine measures show large positive correlations. However, in the absence of Japanese component-level agreement data (RQ1 and RQ2, pending), the possibility that error cancellation contributes to these high correlations cannot be excluded.

By task, both ST1 and ST2 show consistently strong correlations (Table 8). The weakest task-specific value is ECPR in ST1 ($r = .872$), still well above the large-effect threshold. Unlike the English results, Japanese MCPD is uniformly strong across tasks ($r = .959$ in ST1, $r = .953$ in ST2), suggesting that the Japanese corpus may have a more stable MCP distribution.

Measure	Pearson r	Spearman ρ	ICC(2,1)	MAE
AR	.947	.935	.903	0.242
SR	.992	.988	.982	0.116
MLR	.991	.987	.975	0.460
MCPR	.912	.933	.913	0.009
ECPR	.903	.883	.898	0.008
PR	.984	.978	.978	0.009
MCPD	.955	.902	.878	0.151
ECPD	.942	.902	.884	0.231
MPD	.948	.876	.880	0.184

Table 7: Japanese RQ3 overall concurrent validity ($N = 40$).

	ST1 ($N = 20$)				ST2 ($N = 20$)			
Measure	r	ρ	ICC	MAE	r	ρ	ICC	MAE
AR	.965	.965	.922	0.227	.929	.911	.886	0.257
SR	.993	.989	.986	0.101	.993	.976	.979	0.132
MLR	.997	.987	.988	0.348	.986	.955	.961	0.572
MCPR	.895	.884	.898	0.009	.925	.925	.928	0.009
ECPR	.872	.838	.861	0.009	.921	.892	.921	0.008
PR	.987	.986	.982	0.007	.981	.927	.977	0.010
MCPD	.959	.863	.877	0.130	.953	.920	.882	0.173
ECPD	.932	.932	.865	0.250	.947	.880	.896	0.213
MPD	.957	.913	.889	0.174	.940	.872	.875	0.195

Table 8: Japanese RQ3 concurrent validity by task.

3.5 Japanese RQ1 and RQ2 (pending completion)

Japanese clause-boundary and pause-location agreement analyses follow the same metric suites as English but are pending completion in this release. These outcomes will be reported with full task-level breakdowns in the final manuscript version.

3.6 Comparison with prior systems

Table 9 places the current English results alongside published benchmarks.

While keeping in mind that cross-study comparison is constrained by methodological differences, the current pipeline shows numerically higher pause-location agreement than Matsuura et al. (2025) ($\kappa = .840$ vs. .626–.749 in their substantial-agreement subsets) and numerically higher clause boundary detection accuracy than Chen and Yoon (2012) ($F1 = .845$ vs. .690). These are descriptive cross-study comparisons based on published summary metrics, not statistical tests on harmonized raw data. Given at least five dimensions of non-equivalence (corpus, proficiency distribution, L1 background, gold-standard method, and clause definition breadth), these cross-study contrasts should be treated as context-setting rather than evidence of relative system quality. The observed numerical differences may reflect advances in ASR

System	Task	Pause loc. κ	Clause $F1$
Chen and Yoon (2012)	Monologue (L2 EN)	N/A	.690
Matsuura et al. (2025)	Monologue (L2 EN)	.626–.749	N/A
Matsuura et al. (2025)	Dialogue (L2 EN)	.596	N/A
Current pipeline	**Monologue (L2 EN)**	**.840**	**.845**

Table 9: Comparison with prior automatic annotation systems. These values are not directly comparable: cross-study comparisons are limited by differences in speaker populations, task types, L1 backgrounds, proficiency distributions, gold-standard construction methods (LLM-assisted vs. human-only), clause definitions, and segmentation conventions, and should be treated as indicative context rather than evidence of relative system quality.

quality (Qwen3 vs. Rev.ai/Wav2Vec2 forced alignment), MFA-based timestamp refinement, a modern dependency parser (spaCy transformer), and the filler-augmented alignment technique, but they could equally reflect differences in data difficulty or gold-standard construction.

4 Discussion

4.1 English component-level agreement in relation to prior benchmarks

English clause boundary agreement ($\kappa = .816$, $F1 = .845$) reached the "almost perfect" range (Landis and Koch, 1977) and was numerically higher than the clause boundary $F1$ reported by Chen and Yoon (2012) ($F1 = .690$). Pause-location classification ($\kappa = .840$) was also numerically higher than the .626–.749 range reported for most L2 English subsets by Matsuura et al. (2025). While direct comparison is limited by methodological differences, these should be interpreted as descriptive cross-study comparisons because prior reports provide summary outputs from different datasets, not harmonized raw annotations for unified testing. Additional caution is required because the current evaluation uses a stricter alignment method (edit-distance rather than longest-common-subsequence) and a broader clause definition (Vercellotti and Hall, 2024) that includes more boundary types than traditional definitions.

The component-level results also provide an important diagnostic that end-to-end correlations alone cannot offer. High measure-level correlations can arise even when intermediate boundary or pause-label agreement is moderate, if errors cancel in aggregation. By demonstrating strong agreement at both the boundary and pause-label levels, the present analysis reduces the risk that the strong RQ3 correlations are artifacts of error cancellation rather than genuine annotation quality.

4.2 Error propagation from ASR through clause boundaries to pause classification

A clear error propagation chain was observed across pipeline stages. Files with WER $> .20$ showed noticeably lower boundary agreement (mean $\kappa \approx .72$ vs. .84 for WER $\leq .20$),

confirming the findings of Knill et al. (2018, 2019) that ASR errors propagate to downstream NLP tasks. The file with the lowest clause boundary agreement also had the lowest pause-location accuracy, demonstrating that boundary errors cascade into pause classification.

Task type moderated agreement levels, with ST1 (picture narrative) consistently outperforming ST2 (argumentative) across both RQ1 and RQ2. This aligns with the expectation that structured narrative tasks produce more regular speech with lower ASR error rates, while argumentative speech introduces more complex syntax, longer utterances, and repair sequences that challenge both ASR and clause segmentation. Matsuura et al. (2025) similarly found weaker disfluency detection and pause-location agreement in their dialogic task compared to monologic tasks, attributing this to co-constructive discourse features. The present ST1/ST2 contrast, although less extreme (both tasks are monologic), follows the same pattern on the dimension of cognitive demand (Suzuki and Kormos, 2023).

4.3 Concurrent validity across languages and measures

Concurrent validity was strong in both languages: mean Pearson $r = .936$ for English and .953 for Japanese. Across both languages, speed and composite measures (SR, MLR, AR) and pause-frequency measures (MCPR, ECPR, PR) were the most stable, while pause-duration measures (MCPD, ECPD, MPD) were comparatively more sensitive. This pattern is consistent with prior reports that composite and speed measures correlate most strongly between automatic and manual systems (Matsuura et al., 2025; de Jong et al., 2021).

The relative weakness of MCPD, particularly in English ST1 ($r = .713$, ICC = .669), deserves interpretation. MCPD is computed as the mean duration of mid-clause pauses; in files with few MCPs, a single falsely detected or misaligned pause can substantially shift the mean. Mid-clause pauses are relatively rare events in picture-narrative speech (sparse counts), which may make mean duration estimates unstable in shorter speech samples and render MCPD more vulnerable to individual annotation errors. This explanation is consistent with the observation of Matsuura et al. (2025), who found that MCPD correlations varied across proficiency levels, with weaker agreement in high-fluency groups where MCP counts are smallest. However, the per-file MCP count distribution is not reported in the current study, and an alternative explanation—that systematic timing errors in MCP onset/offset detection drive the lower agreement—cannot be excluded. Quantifying the relationship between per-file MCP count and MCPD agreement is a direction for future investigation.

A related diagnostic is the divergence between Pearson r and Spearman ρ for pause-duration measures, which is most pronounced in Japanese: MPD ($r = .948$ vs. $\rho = .876$, gap = .072), MCPD ($r = .955$ vs. $\rho = .902$, gap = .053), and ECPD ($r = .942$ vs. $\rho = .902$, gap = .040). A similar pattern appears in English MPD ($r = .957$ vs. $\rho = .908$, gap = .049). Because Pearson r is sensitive to extreme values while Spearman ρ captures rank-order agreement, these divergences suggest that a few files with extreme pause durations may inflate Pearson correlations. ICC values, which account for both correlation and absolute agreement, are also notably lower than Pearson r for these measures (e.g., Japanese MCPD: $r = .955$, ICC = .878), indicating systematic mean-level bias in addition to correlation. These patterns reinforce the characterization of pause-duration measures as the most sensitive pipeline outputs.

Interestingly, Japanese MCPD showed uniformly strong Pearson correlations ($r = .959$

and .953 for ST1 and ST2), unlike the English pattern. This may reflect differences in pausing behavior between the two language cohorts, or alternatively, may result from the Japanese corpus having a more stable MCP distribution. Cross-lingual differences in MCP–ECPD sensitivity warrant further investigation with matched proficiency samples.

4.4 Cross-lingual transportability of the pipeline

The shared architecture yields strong concurrent-validity results in both English and Japanese, providing preliminary evidence for cross-lingual transportability with explicit language-specific adaptations. Three adaptations deserve highlighting. First, mora-based normalization in Japanese replaces syllable-based normalization in English, reflecting the fundamental prosodic unit difference. Second, the Japanese clause segmenter adapts the Vercellotti "verb + element" framework to handle te-form chains, a construction without an English parallel, by granting clause status only to te-form verbs with their own complement or adjunct. A related asymmetry is the exclusion of `nsubj` from the Japanese complement check (because Japanese is a pro-drop language where subjects are commonly omitted), whereas English implicitly counts subjects via the finite-verb requirement. This asymmetry could systematically produce fewer clause boundaries in Japanese relative to an equivalent English implementation, which should be considered when comparing clause-count-dependent measures across languages. Third, the Japanese pipeline includes MeCab re-tokenization and a robust map-back algorithm to handle tokenization mismatches between ASR and MFA, a challenge less prominent in English.

Despite these adaptations, the core architecture remains shared: the same ASR model, the same disfluency detector, the same filler classifier, and the same clause-coding principle. The Japanese RQ3 results indicate that this shared-architecture approach is promising for cross-lingual deployment, at least for the English–Japanese pair evaluated here. However, the cross-lingual transportability claim currently rests on measure-level evidence for both languages but component-level evidence for English only. Until Japanese RQ1 and RQ2 are complete, it remains possible that the Japanese pipeline achieves high measure-level correlations through different pathways than the English pipeline, including potential error cancellation. The staged evaluation design will become fully cross-lingual only with the completion of all six research questions.

4.5 Technical design contributions

Beyond the quantitative results, three design features of the pipeline address structural weaknesses in prior cascaded annotation systems.

Precise word-level timestamps via forced alignment. In a cascaded pipeline, all downstream annotations (clause boundaries, pause locations, and fluency measures) depend on accurate word-level timing. If timestamps are imprecise, errors propagate through every subsequent stage. Prior systems relied on ASR-internal timestamps or simple Wav2Vec2 forced alignment, which can be imprecise for L2 speech with long pauses and varied pronunciation. We address this by applying Montreal Forced Aligner with high beam settings after Qwen3-ASR, combined with filler-augmented alignment that injects placeholder tokens into gaps

before alignment. This prevents the aligner from distributing pause durations across adjacent word boundaries, yielding more accurate pause onset and offset times. The improvement in boundary and pause-location agreement over prior benchmarks (Table 9) likely reflects, in part, this timestamp quality improvement.

Multi-layer filler handling. Modern large-vocabulary ASR models (including Qwen3-ASR) tend to produce clean transcripts by suppressing fillers and hesitations. This is desirable for transcription accuracy but problematic for fluency annotation: unrecognized fillers result in their speech intervals being counted as silent pauses, inflating pause metrics. We therefore implement a three-part design. First, at the ASR stage, a disfluency-preserving prompt encourages the model to retain fillers in the transcript. Second, filler-augmented MFA alignment injects language-appropriate placeholder tokens into inter-word gaps before forced alignment; these placeholders prevent the aligner from distributing gap durations across adjacent word boundaries and are discarded in the map-back step. Third, a post-ASR neural filler classifier (trained on the PodcastFillers dataset) scores candidate speech regions within remaining ASR-detected gaps, suppressing or splitting pause intervals where filler speech is detected. This design provides redundancy, but its net filler-detection coverage cannot be directly verified in the current study because the manual reference annotations do not include an explicit filler-labeled tier (see Limitation 13).

Zero-shot cross-lingual disfluency detection. Disfluency-annotated speech corpora are concentrated in English (e.g., Switchboard), while Japanese and other languages lack comparable resources. We address this data scarcity by training the disfluency detector on a combination of real English data and synthetic disfluency data generated for both English and Japanese, following the data augmentation approach of Kundu et al. (2022). Since the underlying model (XLM-RoBERTa) supports more than 100 languages, this approach is potentially extensible to other languages without requiring language-specific annotated corpora, a practical advantage for scaling the pipeline beyond the English–Japanese pair validated here.

4.6 Implications for fluency research and assessment

The present pipeline can reduce annotation cost while preserving strong agreement and correlation on the evaluated cohorts. Practically, this makes clause-based UF analyses feasible at larger scales and supports transparent hybrid workflows where automatic annotation is combined with targeted manual checking. The staged evaluation design, component-level agreement followed by measure-level validity, provides a template that future automatic annotation studies can adopt, in line with recommendations to test systems under realistic error propagation constraints (Knill et al., 2018, 2019).

Methodologically, the explicit operationalization of the Vercellotti & Hall clause framework (Vercellotti and Hall, 2024) helps address a longstanding comparability problem in L2 fluency research. By making clause-coding rules inspectable at the script level, researchers can evaluate whether observed differences across studies reflect genuine linguistic phenomena or artifacts of inconsistent clause definitions. The cross-lingual adaptation of this framework,

particularly the principled treatment of Japanese te-form chains, indicates that the "verb + element" principle can extend beyond English.

For language assessment contexts, the strong correlations observed for SR, MLR, and PR (the measures most strongly associated with perceived fluency ratings; Suzuki et al., 2021) suggest potential applicability in automated scoring systems. However, the sensitivity of MCPD under certain conditions implies that assessments relying heavily on pause-duration measures should incorporate manual checking or targeted verification, at least until larger validation studies establish robustness across broader populations.

4.7 Future directions

Several directions merit investigation. First, the current clause segmenters use explicit, rule-based implementations of the Vercellotti & Hall framework. While this approach maximizes transparency and interpretability, because every boundary decision traces to an identifiable rule, it may fail on syntactically unusual L2 constructions that the rule set does not anticipate. A model-based approach (e.g., fine-tuning a sequence labeler on gold clause boundaries) could improve robustness to such cases, at the cost of reduced interpretability and the need for substantial annotated training data per language. Comparing rule-based and model-based clause segmentation on the same gold standard would clarify this trade-off.

Second, external-corpus validation is needed to establish generalizability beyond the cohorts and task types evaluated here. In particular, testing on dialogic speech, lower-proficiency learners, and L1 backgrounds beyond those in the current corpora would strengthen validity claims.

Third, the present study does not include perceived fluency (PF) ratings. Future work should evaluate whether the automatically computed UF measures predict listener-based fluency judgments at levels comparable to those reported by Matsuura et al. (2025), which would provide complementary construct validity evidence.

5 Limitations

Thirteen limitations should guide interpretation.

1. Japanese RQ1 (clause-boundary agreement) and RQ2 (pause-location agreement) are pending completion in this release. The complete six-question analysis will be reported in the final manuscript version.

2. Cohort sizes remain moderate (EN $N = 39$ for RQ3; JA $N = 40$), limiting subgroup precision and restricting the power of task-level comparisons. Following Donner and Eliasziw (1987), the minimum sample size to detect substantial agreement ($\kappa > .61$) with power = .80 ($1 - \beta$) varies by base rate, and some per-task cells may be underpowered for detecting moderate agreement differences.

3. English gold boundaries were constructed using an LLM-assisted stage (validated on locked test files at micro $F1 = .929$, $\kappa = .914$, then manually reviewed by the first author). Although this level of LLM accuracy is high and the workflow follows the methods-grade

standard of Morin and Marttinen Larsson (2025), residual model-shaped annotation bias relative to fully independent double-human coding remains possible. Because the manual review was conducted by a single author, anchoring bias (tendency to accept existing annotations) cannot be excluded, and the resulting inter-rater reliability may be inflated relative to a fully independent human gold standard. The Japanese gold standard was constructed through independent human annotation, meaning English and Japanese gold standards differ in construction method; cross-lingual comparisons of annotation quality are therefore not straightforward.

4. External-corpus generalization is not yet established. Current validity claims are specific to the corpora and cohorts evaluated; the pipeline has not been tested on other L1 backgrounds, proficiency levels, or task types (e.g., dialogic tasks, read-aloud conditions).

5. Unlike Matsuura et al. (2025), the present study does not include perceived fluency (PF) ratings. We therefore cannot evaluate the predictive power of automatic measures for listener-based fluency judgments, which is a complementary dimension of construct validity.

6. The disfluency detector was trained on Switchboard (L1 English conversational speech) and synthetic data. On held-out evaluation sets, the model achieves $F1 = .987$ (synthetic English), $F1 = .997$ (synthetic Japanese), and $F1 = .727$ on the DisfluencySpeech corpus (real L1 English; precision = .882, recall = .619). However, no evaluation on L2 speech has been conducted, and detection accuracy on learner speech—where disfluency patterns may differ from L1 norms—remains unknown. Although the detector serves a pruning function rather than producing a reported measure, errors propagate to clause segmentation: undetected disfluencies (e.g., content-level self-corrections) can introduce spurious tokens into the parse input, and false positives can remove fluent tokens. The Japanese pipeline applies additional deterministic post-rules (repeated-token collapse, elongated-form handling) to mitigate known failure modes, but these heuristics were not systematically evaluated. Quantifying disfluency detection accuracy on L2 speech, particularly for Japanese, is a priority for future validation.

7. Cross-language differences were described descriptively (e.g., mean r and task-level patterns) without formal between-language significance testing. In addition, benchmark comparisons rely on published summary metrics from different datasets rather than harmonized raw annotations. Claims about English–Japanese differences and cross-study superiority should therefore be interpreted as exploratory.

8. Learner demographic information (L1 background distribution, proficiency levels, age, and recording conditions) is not reported in detail for either corpus, limiting assessment of population representativeness and generalizability.

9. The MFA acoustic models (`english_us_arpa` and `japanese_mfa`) were trained on native speech. Their coverage of L2-accented speech is unvalidated; forced alignment errors in non-native productions, particularly for heavily accented speech, may propagate to pause duration estimates.

10. Word error rate (WER) was reported as an overall mean but was not stratified by proficiency level or L1 background. The relationship between ASR quality and UF measurement validity at different proficiency levels therefore remains unquantified.

11. The pipeline computes only speed and breakdown fluency measures in this release. Repair fluency (RF) measures—which capture repetitions, self-corrections, and false starts and are an established dimension of utterance fluency (Kormos, 2006; Suzuki et al., 2021)—are not reported. The pipeline's disfluency detector can identify disfluent tokens and the codebase supports RF computation, but the manual reference annotations used for concurrent validity do not include disfluency-level labels, precluding RF validation. Extending the validation to RF requires constructing disfluency-annotated reference data.

12. The 150 ms onset window used to classify pauses as ECP (end-of-clause) versus MCP (mid-clause) was set based on the temporal resolution of forced-alignment boundaries but has not been subjected to a formal sensitivity analysis. Different thresholds (e.g., 100 or 200 ms) could reclassify a non-trivial number of pauses and change MCPR, ECPR, MCPD, and ECPD values.

13. The neural filler classifier was trained on English podcast speech (PodcastFillers dataset). Its application to Japanese gaps assumes cross-lingual transfer of filler acoustics, which has not been empirically verified. Direct verification of net filler-detection accuracy was not possible in this release because the manual reference annotations do not include explicit filler labels or a dedicated filler tier. As a result, residual under-detection or over-detection of fillers may bias pause-duration measures, particularly in Japanese.

6 Conclusion

This study developed and evaluated a shared English–Japanese pipeline for automatic clause-based fluency annotation, using a staged design from component-level agreement to measure-level concurrent validity. In the current report, English component-level agreement reached almost perfect levels for both clause boundaries ($\kappa = .816$) and pause-location classification ($\kappa = .840$), with point estimates numerically above previously reported monologic L2 English values, although these cross-study contrasts are descriptive only and cannot establish relative superiority. Concurrent validity was strong across all nine measures for English (mean $r = .936$, 95% CI [.878, .965]), supported by full staged validation, and Japanese measure-level correlations were similarly high (mean $r = .953$, 95% CI [.916, .975]), providing preliminary evidence for cross-lingual transportability. Speed and composite measures showed the highest stability, while pause-duration measures showed the greatest sensitivity to pipeline error. Japanese component-level analyses (RQ1 and RQ2) remain pending; English and Japanese thus differ in validation completeness, and the cross-lingual validity argument will be complete only when Japanese component-level analyses confirm the staged pattern observed in English.

Three technical contributions underlie these results. First, filler-augmented forced alignment produces precise word-level timestamps that serve as the foundation for all downstream annotations and directly addresses error amplification in cascaded systems. Second, multi-layer filler handling at the ASR prompt, forced alignment, and post-ASR classifier stages

is designed to reduce systematic over-counting of silent pauses caused by filler suppression in ASR outputs, although direct filler-level verification is not yet possible with the current manual reference format. Third, zero-shot cross-lingual disfluency detection, trained on synthetic data augmentation, addresses the scarcity of annotated disfluency data in non-English languages.

Beyond these technical contributions, the pipeline advances L2 fluency methodology by providing a preliminary cross-lingual architecture for clause-based annotation, demonstrating a staged evaluation design that strengthens measure-level validity evidence, and explicitly operationalizing the Vercellotti & Hall clause framework for transparent and replicable clause coding. All pipeline code, trained models, and analysis scripts are openly available (`[URL_TO_BE_INSERTED]`) to support replication and extension. The immediate next step is completion of Japanese RQ1–RQ2, which will enable a full six-question cross-lingual analysis.

References

Chen, L., and S.-Y. Yoon. 2011. Detecting structural events for assessing non-native speech. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, 38–45.

Chen, L., and S.-Y. Yoon. 2012. Application of structural events detected on ASR outputs for automated speaking assessment. In *Proceedings of Interspeech 2012*, 767–770.

Chen, L., K. Zechner, S.-Y. Yoon, K. Evanini, X. Wang, A. Loukina, J. Tao, L. Davis, C. M. Lee, M. Ma, R. Mundkowsky, C. Lu, C. W. Leong, and B. Gyawali. 2018. Automated scoring of nonnative speech using the SpeechRater^{SM} v.5.0 engine. *ETS Research Report Series* 2018(1).

Coulange, S., T. Kato, S. Rossato, and M. Masperi. 2024. Enhancing language learners' comprehensibility through automated analysis of pause positions and syllable prominence. *Languages* 9(3): 78.

de Jong, N. H. 2016. Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *IRAL – International Review of Applied Linguistics in Language Teaching* 54(2): 113–132.

de Jong, N. H., and H. R. Bosker. 2013. Choosing a threshold for silent pauses to measure second language fluency. In *Proceedings of DiSS 2013*, 17–20.

de Jong, N. H., and T. Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods* 41(2): 385–390.

de Jong, N. H., J. Pacilly, and W. Heeren. 2021. PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education: Principles, Policy and Practice* 28(4): 456–476.

Donner, A., and M. Eliasziw. 1987. Sample size requirements for reliability studies. *Statistics in Medicine* 6(4): 441–448.

Foster, P., A. Tonkyn, and G. Wigglesworth. 2000. Measuring spoken language: A unit for all reasons. *Applied Linguistics* 21(3): 354–375.

Koo, T. K., and M. Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15(2): 155–163.

Knill, K. M., M. J. F. Gales, K. Kyriakopoulos, A. Malinin, A. Ragni, Y. Wang, and A. Caines. 2018. Impact of ASR performance on free speaking language assessment. In *Proceedings of Interspeech 2018*, 1641–1645.

Knill, K. M., M. J. F. Gales, P. P. Manakul, and A. P. Caines. 2019. Automatic grammatical error detection of non-native spoken learner English. In *Proceedings of ICASSP 2019*, 8127–8131.

Kormos, J. 2006. *Speech Production and Second Language Acquisition*. Lawrence Erlbaum Associates.

Kundu, R., P. Jyothi, and P. Bhattacharyya. 2022. Zero-shot disfluency detection for Indian languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, 4442–4454.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1): 159–174.

Matsuura, R., S. Suzuki, M. Saeki, T. Ogawa, and Y. Matsuyama. 2022. Refinement of utterance fluency feature extraction and automated scoring of L2 oral fluency with dialogic features. In *Proceedings of 2022 APSIPA ASC*, 1312–1320.

Matsuura, R., S. Suzuki, K. Takizawa, M. Saeki, and Y. Matsuyama. 2025. Gauging the validity of machine learning-based temporal feature annotation to measure fluency in speech automatically. *Research Methods in Applied Linguistics* 4: 100177.

Morin, C., and M. Marttinen Larsson. 2025. Large corpora and large language models: A replicable method for automating grammatical annotation. *Linguistics Vanguard* 11(1): 501–510.

Plonsky, L., and F. L. Oswald. 2014. How big is "big"? Interpreting effect sizes in L2 research. *Language Learning* 64(4): 878–912.

Rose, R. L. 2020. Fluidity: Real-time feedback on acoustic measures of second language speech fluency. In *Proceedings of the International Conference on Speech Prosody 2020*, 774–778.

Segalowitz, N. 2010. *Cognitive Bases of Second Language Fluency*. Routledge.

Skidmore, L., and R. K. Moore. 2023. BERT models for spoken learner English disfluency detection. In *Proceedings of SLaTE 2023*, 91–92.

Suzuki, Y., and K. Hanzawa. 2022. Massed task repetition is a double-edged sword for fluency development: An EFL classroom study. *Studies in Second Language Acquisition* 44(2): 536–561.

Suzuki, S., and J. Kormos. 2023. The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency. *Studies in Second Language Acquisition* 45(1): 38–64.

Suzuki, S., and A. Révész. 2023. Measuring speaking and writing fluency: A methodological synthesis focusing on automaticity. In Y. Suzuki (Ed.), *Practice and Automatization in Second Language Research* (pp. 235–264). Routledge.

Suzuki, S., J. Kormos, and T. Uchihara. 2021. The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *Modern Language Journal* 105(2): 435–463.

Tavakoli, P., and P. Skehan. 2005. Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and Task Performance in a Second Language* (pp. 239–273). John Benjamins.

Tavakoli, P., F. Nakatsuhara, and A. Hunter. 2020. Aspects of fluency across assessed levels of speaking proficiency. *Modern Language Journal* 104(1): 169–191.

Tavakoli, P., G. Kendon, S. Mazhurnaya, and A. Ziomek. 2023. Assessment of fluency in the Test of English for Educational Purposes. *Language Testing* 40(3): 607–629.

Vercellotti, M. L., and S. Hall. 2024. Coding all clauses in L2 data: A call for consistency. *Research Methods in Applied Linguistics* 3: 100132.

Zayats, V., T. Tran, R. Wright, C. Mansfield, and M. Ostendorf. 2019. Disfluencies and human speech transcription errors. In *Proceedings of Interspeech 2019*, 3088–3092.

Zhu, G., J.-P. Caceres, and J. Salamon. 2022. Filler word detection and classification: A dataset and benchmark. In *Proceedings of Interspeech 2022*, 3769–3773.