TalkTag:
Fine-Grained Morphosyntactic Error Annotation for Transcribed Speech

Shamira Venturini[1,2], Oliver Hennhöfer[2], Steffen Kinkel[2], Jannik Strötgen[2]

[1]Karlsruhe Institute of Technology, Germany
[2]Karlsruhe University of Applied Sciences, Germany
Correspondence: shamira.venturini@h-ka.de

Abstract

Fine-grained morphosyntactic error annotation is important in clinical and developmental language research, yet it is labour-intensive, expert-dependent, and difficult to scale. We present `TalkTag`, an LLM-based lightweight tool fine-tuned to automate CHAT-style error annotation in spoken-language transcripts. Developed under conditions of extreme data scarcity using children's narrative data, the system shows the feasibility of linguistic analysis in low-resource settings. Our evaluation demonstrates that `TalkTag` produces encouragingly precise annotation while effectively identifying instances where linguistic ambiguity makes automated tagging genuinely complex. In summary, with `TalkTag`, we provide a scalable alternative to manual error annotation and practically viable support for morphosyntactic error annotation.

1 Introduction

Language Sample Analysis has become an increasingly important method in clinical linguistics and developmental psycholinguistics (MacWhinney and Fromm, 2022). Drawing on naturalistic spoken interaction data, it supports the study of language development and impairment in context.

During the past decades, TalkBank [1] (MacWhinney et al., 2004) has substantially advanced the infrastructure for this line of research through large open spoken corpora, the CHAT transcription format, and the CLAN analysis tools (MacWhinney, 2000). CHAT provides a standard way to represent spoken-language transcripts, while CLAN (Computerized Language ANalysis) provides analysis programs for CHAT-formatted files. In this setting, morphosyntactic error codes are written inline on the transcript line, immediately after the form they describe, so that the annotation preserves both the child's production and the analyst's interpretation of the intended target. Although TalkBank includes tools for some aspects of automatic transcription and morphosyntactic annotation, to the best of our knowledge, it currently does not support automatic annotation of morphosyntactic errors.

However, fine-grained morphosyntactic error annotation is important because it provides evidence of grammatical development, impairment, and variation in typically developing (Moraleda-Sepúlveda and López-Resa, 2022), as well as populations such as children with developmental language disorders (Leonard and Deevy, 2020; Eadie et al., 2002; Rice and Wexler, 1996), Down syndrome (Witecy et al., 2023; Katsarou and Andreou, 2022; Penke, 2019), deaf and hard-of-hearing children with cochlear implant (Benassi et al., 2021; Golestani et al., 2018), and autism spectrum disorder (Huang and Finestack, 2020). Manually annotating these phenomena remains slow, labour-intensive, and dependent on expert knowledge. Unlike ordinary morphosyntactic tagging, this task often requires identifying a structured divergence between an erroneous production and an intended target, or recovering morphology that is absent from the surface string but obligatory in context. It therefore goes beyond simple sequence labelling, requiring structural linguistic reasoning rather than surface-level pattern matching.

Large Language Models (LLMs) offer a promising solution for this setup by integrating contextual modelling with schema-constrained generation (Devlin et al., 2019). Unlike assigning isolated tags token by token, LLMs can, in principle, evaluate an entire utterance holistically. This allows for the generation of well-formed inline annotations that capture syntactic dependencies and the underlying structural nature of the error.

[1] `http://talkbank.org`

Proceedings of the 20th Linguistic Annotation Workshop (LAW XX), pages 309–322

At the same time, annotated clinical and developmental language data are often scarce, access-restricted, and difficult to use for large-scale model development (Al-Marridi et al., 2026; Gagliardi and Maffia, 2024). We therefore introduce TalkTag[2], a tool for fine-grained morphosyntactic error annotation that follows the CHAT guidelines for word-level error coding. The tool was developed under extremely low-resource conditions, with very limited annotated data (Hedderich et al., 2021). To address this, we employed synthetic data augmentation to fine-tune a lightweight, open-weight LLM, effectively expanding the model's exposure to rare error patterns. In this initial prototype, we focus on children's narrative data, a domain selected not only for its availability but for its high density of developmentally salient morphosyntactic phenomena. These narratives provide a benchmark for evaluating fine-grained annotation capabilities in a complex, real-world linguistic context.

The main contribution of this paper is a lightweight tool for fine-grained morphosyntactic error annotation in clinical and developmental language research, together with its accompanying Python package, TalkTag. More specifically, we formulate CHAT morphosyntactic error coding as a constrained structured-generation task, adapt a small instruction-tuned model to this setting under severe data sparsity, and evaluate the resulting system using automatic scoring, blinded post-hoc adjudication, and human review on unseen corpus material.

The remainder of the paper is structured as follows. Section 2 defines the target annotation scheme and the subset of CHAT morphosyntactic error labels modelled in this study. Section 3 then situates the work in relation to prior research on clinical language annotation, error coding, and automatic linguistic analysis. Section 4 describes the model, training setup, data, and evaluation design. Finally, Section 5 reports the automatic and human-reviewed results. Section 6 discusses the main linguistic error patterns, the implications of the findings for annotation practice, and the limitations of the current system.

2 The Annotation Language

Within the TalkBank ecosystem, which provides infrastructure for the transcription and analysis of spoken interaction data, the MOR and GRASP programs (MacWhinney, 2012) support automatic morphosyntactic analysis of CHAT transcripts. MOR is a morphological analyser that assigns lexical and grammatical information to each token, producing the %mor tier with lemma, part-of-speech, and inflectional information. Building on this output, GRASP derives syntactic structure by assigning grammatical relations and dependency-based representations across the utterance on the %gra tier. Together, these tools make it possible to move from raw transcript text to a linguistically enriched representation of children's speech. However, their purpose is to recover morphological and syntactic structure rather than to identify, classify, or encode morphosyntactic errors explicitly.

The CHAT Transcription Guidelines provide a general system for marking word-level errors (MacWhinney, 2000, 2019). At the level of string form, the annotation tags have a simple and regular surface structure: first, they consist of a fixed bracketed frame following the relevant error, marked with an *. Next, a flat sequence of colon-separated fields indicates i) the error domain (phonological, semantic, neologistic, dysfluency, and morphological), ii) the error pattern (e.g., missing, superfluous, over-regularised, double-marked morphemes, unknown/known target, etc.) and iii) the morpheme or part-of-speech involved (e.g., past tense, perfective, plural, or third-person singular agreement morphemes; pronouns, prepositions, determiners as parts-of-speech). When the intended target is known, it can also be provided in brackets next to the error using the format [: target]. This is used when CLAN's MOR morphological analyser should analyse the target form instead of the produced one, whereas [:: target][3] can be used to preserve analysis of the produced real-word form while still recording the intended target.

The surface syntax of the labels is therefore relatively simple, yet their assignment is highly challenging: it often depends on contextual linguistic interpretation and, in many cases, on reconstructing an intended target form.

In this work, we focus on treating [* m:*] labels as mismatches between produced and target forms

[2] http://github.com/OliverHennhoefer/talk-tag

[3] Since the time of this study, TalkBank has updated the CHAT manual, replacing the [:: target] syntax with [= target]. While the model described here was trained on the earlier convention, the associated Python package includes post-processing to ensure compliance with the latest standard and offers options to toggle the visibility of reconstructions depending on user preference.

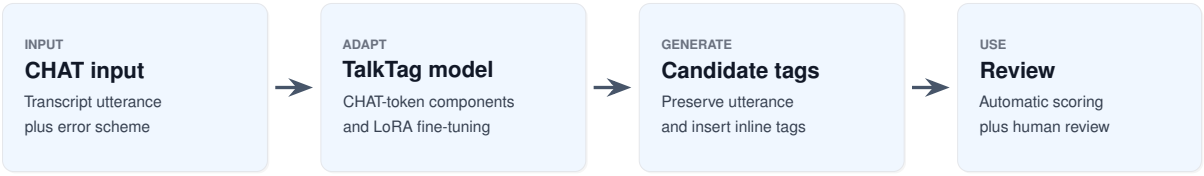

Figure 1: The TalkTag Workflow Pipeline.

under lexical identity (i.e., morpheme operations), and [* s:r:*] labels as substitutions involving the same lexical category (e.g., wrong preposition) or [* s:r:gc:*] wrong grammatical category (e.g., adjective for pronoun). Moreover, we use both reconstruction strategies, reserving [: target] for cases where the error produces a non-word form. Examples of morphological error annotations are:

- "*Yesterday I walk* [:: walked] [* m:0ed] *to school*", which marks a missing past tense morpheme.
- "*Yesterday I goed* [: went] [* m:=ed] *to school*", which marks an overregularised past tense morpheme resulting in a non-word form.

Examples of substitutional error annotations are:

- "*Yesterday me* [:: I] [* s:r:gc:pro] *walked to school*", which marks a wrong grammatical category of a pronoun.
- "*Yesterday I went in* [:: to] [* s:r:prep] *school*", which marks a wrong preposition.

The CHAT manual provides a standard inventory of error codes, but the framework is extendable: CHAT coding can be adapted to specific applications, and the error-coding system itself allows additional distinctions and combinations within that general format. The annotation scheme for morphological and substitutional errors is illustrated in Table 1 and Table 2, respectively. The complete label components inventory is illustrated in Table 11 in Appendix D.

Level 1	**Meaning**
* m:	morphosyntactic error
Level 2	**Meaning**
0	missing regular form
base:	base for irregular form
irr:	irregular for base form
sub:	past/perfective substitution
=	over-regularisation
+	superfluous marking
++	double marking
vsg:	irregular verb 3PS
vun:	irregular verb unmarked
allo	allomorphic errors
Level 3	**Meaning**
a	agreement error
i	irregular target
mor	target morpheme

Table 1: CHAT annotation scheme for morphological errors.

3 Related Work

Grammatical error detection. Work on grammatical error detection and grammatical error correction addresses linguistic errors more directly, but typically formulates the problem as one of edit detection or sentence-level correction. In this literature, annotation generally starts from a source sentence and a corrected target, from which error spans are identified and labelled with edit operations such as replacement, omission, insertion, or transposition. Schemes such as ERRANT (Korre and Pavlopoulos, 2020) add a further layer of linguistic classification, yielding a structured representation of each edit rather than a flat label. For example, ERRANT combines edit operations such as M (missing), R (replacement), and U (unnecessary) with linguistic error categories to produce fine-grained composite labels. However, the empirical basis of this literature is overwhelmingly written and learner-focused. As noted by Bryant et al. (2023), the main benchmark datasets are largely derived from L2 English essays, examinations, and learner-platform submissions, including FCE (Yannakoudakis et al., 2011), NUCLE (Dahlmeier et al., 2013), CoNLL-2013 and 2014 (Ng et al., 2013, 2014), Lang-8 (Mizumoto et al., 2012; Tajiri et al., 2012), JFLEG (Napoles et al., 2017), and W&I+LOCNESS (Bryant et al., 2019). Consequently, current annotation schemes and correction models are primarily optimized for

Level 1	Meaning
* s:	substitution error
Level 2	**Meaning**
r:	related lexical substitution
r:gc:	related grammatical substitution
Level 3	**Meaning**
POS	target morpheme or part-of-speech

Table 2: CHAT annotation scheme for substitutional errors.

sentence-level written L2 language, which limits direct transfer to spoken, interactional, or clinically atypical language.

Morphosyntactic error annotation of child language. More specific work on child language appears limited and idiosyncratic in terms of the annotation scheme and targeted granularity. Morley et al. (2013) first showed that even relatively coarse linguistic error codes could be sufficient for identifying neurodevelopmental disorders. Building on that result, Morley et al. (2014) developed a data-driven dependency-parser approach for detecting and labelling grammatical errors in SALT-annotated transcripts of children's speech. SALT (Miller et al., 2011) supports error coding through a relatively small default inventory of labels, including overgeneralization [EO:], pronoun error [EP:], other word-level error [EW:], extraneous word [EW], and utterance-level error [EU], although the active code set can be customised within the software. Morley et al. (2013) evaluated on the ENNI corpus from CHILDES and the NSR corpus from the SALT database, their system outperformed both Microsoft Word's grammar checker and a Naive Bayes baseline, while also showing that performance was sensitive to corpus-specific annotation practices and differences in label granularity. This work is therefore highly relevant to the present study, but it remains grounded in a relatively coarse inventory rather than a more fine-grained annotation scheme.

Earlier work by Hassanali and Liu (2011) explored a more fine-grained approach to grammatical error annotation in child language transcripts. Using 677 transcripts from the Paradise corpus (Paradise et al., 2005), they manually annotated ten error categories, with particular attention to verb-related errors, and compared rule-based parse-template methods with statistical classifiers for detecting six error types. Their results showed that statistical approaches generally outperformed rule-based ones. The study's main contribution was to show that automatic grammar checking could move beyond holistic measures of syntactic development, such as IPSyn (Sagae et al., 2005), providing a more differentiated profile of grammatical weaknesses. At the same time, it highlighted important limitations, including the difficulty of parsing spoken child language with disfluencies and incomplete utterances, ambiguity in assigning error categories, and the restricted coverage of systems built around a narrow set of constructions.

More recent work has also approached child grammar from other angles: Nikolaus et al. (2024) developed a context-sensitive scheme for annotating child utterances in caregiver conversation as grammatical, ungrammatical, or ambiguous, and trained Transformer-based models on 4,200 manually annotated CHILDES utterances. Their best models reached near-human agreement and were used to annotate a much larger corpus, confirming that grammaticality increases with age. Unlike work targeting explicit morphosyntactic error labelling, however, their focus was on broad utterance-level grammaticality in a conversational context.

Most recently, Gebauer et al. (2025) investigated grammatical error detection in spontaneous children's speech using German kidsTALC data (Rumberg et al., 2022), explicitly addressing both ASR errors and ambiguity in manual error labelling. They proposed a BERT-based recurrent model with iterative pseudo-labelling, showing significant improvements on both manual and automatically transcribed speech. This makes their study particularly relevant to the present work, since it tackles realistic spoken child-language data rather than written text alone. However, the task is still formulated as coarse binary error detection rather than fine-grained morphosyntactic error annotation, so it provides a close methodological precedent without addressing the richer annotation language used in this paper.

Taken together, the literature points to a clear research gap. While grammatical error classification and correction are well developed, this work is largely grounded in written learner-language data and does not transfer straightforwardly to clinical or developmental spoken-language settings. The more specific literature on child spoken-language error analysis is comparatively sparse, and the clos-

est prior systems either date back more than a decade or adopt different annotation scopes and levels of granularity. To our knowledge, there is currently no automated tool for fine-grained, CHAT-compatible morphosyntactic error annotation within the TalkBank ecosystem.

4 Methods

This section describes the study's methodological setup: the model architecture, training regime, data sources, and evaluation strategy for the morphosyntactic error annotation. The workflow pipeline is visually illustrated in Figure 1.

Model. The model is instruction-tuned base Meta-Llama-3.1-8B-Instruct (Grattafiori et al., 2024), loaded in 4-bit quantised form (bnb-4bit) for efficient fine-tuning. Training is implemented using the Unsloth framework, enabling parameter-efficient adaptation via LoRA while keeping the base model weights frozen. This setup allows us to fine-tune an 8B-parameter model under constrained hardware conditions.

The model is instruction-tuned to produce exactly one annotated utterance line, while preserving the original token order, spelling, punctuation, disfluencies, and CHAT symbols. The prompt constrains the model by specifying the structural conditions of the annotation language. The full prompt is provided in Appendix A.

Rather than treating CHAT morphosyntactic labels as a flat inventory of opaque output strings, we treat the annotation scheme as a structured symbolic language. Accordingly, we extend the tokeniser not with full label forms, but with reusable components that recur across the annotation system, including bracketed markers, domain indicators, and subtype fragments. The embedding matrix is resized to accommodate this augmented vocabulary. This reduces fragmentation of CHAT-specific sequences under the base tokeniser and encourages the model to compose licensed tags from meaningful subparts rather than retrieve them from a closed set. This design is motivated by the structure of the annotation scheme itself: CHAT error tags are compositionally organised, encoding contrasts such as domain, operation type, agreement sensitivity, and irregular morphology. The model is thus exposed to the building blocks of a small regular annotation language and must learn to generate well-formed tag combinations under task constraints.

Source	Total	0	1	2	≥3
Synthetic	5830	946	4771	109	4
Real	4585	4015	517	47	6
Total	10415	4961	5288	156	10

Table 3: Distribution of utterances by number of annotated errors in the full pre-split master dataset. Columns **0**, **1**, **2**, and **≥3** indicate the number of utterances containing zero, one, two, or three or more errors, respectively.

Training. Fine-tuning uses LoRA with rank 32, $\alpha = 64$, and dropout 0.05. Maximum sequence length is 384 tokens. Training uses a per-device batch size of 8 with gradient accumulation of 4, a warmup ratio of 0.03, and weight decay of 0.1. The model was fine-tuned on an NVIDIA A100-SXM4-80GB GPU. Training took approximately 47 minutes for 3 epochs and 723 optimiser steps.

Data. We use CHAT-formatted utterances drawn from a subset of the Edmonton Narrative Norms Instrument (ENNI) corpus (Schneider et al., 2006), available through TalkBank/CHILDES [4] (MacWhinney, 2000). We focus on narratives from children aged 4-5, since this developmental range provides a significant concentration of morphosyntactic phenomena, including overregularisation, agreement errors, tense marking, and clause linking (Cummings, 2023). The resulting real-data subset contains 4,585 utterances manually reviewed for the target annotation task. As shown in Table 3, this is an extremely low-resource setting (Hedderich et al., 2021) not only in corpus size but also in error density, since most real utterances are error-free and many labels have very limited support.

To mitigate this sparsity, we supplemented the real corpus with curated synthetic examples covering configurations described in the CHAT guidelines. These were generated from error-conditioned prompts and manually reviewed before inclusion. We targeted a minimum overall support of approximately 100 instances per label. We also retained clean utterances in both the real and synthetic portions of the data so that the model would learn when to leave an utterance unchanged and preserve valid CHAT formatting.

Evaluation. We evaluate the system at three complementary levels: automatic scoring on held-out

[4] https://talkbank.org/childes

data, followed by a blinded post-hoc review of disagreement cases, and human evaluation of unseen data annotation.

For automatic evaluation, we split both the real ENNI data into train and validation sets as the primary confirmatory benchmarks, as well as augmented data to be used as label-coverage diagnostics. To stabilise rare-label evaluation, we enforced minimum per-label support in the synthetic coverage splits of $N = 10$, without distorting the natural error distribution of the real set. This design allows rare-label behaviour to be measured under controlled support while preserving the full real training pool and avoiding aggressive downsampling.

Automatic evaluation combines line-level and label-level perspectives. We report exact match over the full annotated utterance line, but treat it as a secondary summary measure, since the high proportion of clean utterances inflates this score. Our main evaluation metrics focus on the error labels themselves: micro-F1, macro-F1, and per-label precision/recall/F1. Per-label results are further divided into confirmatory and exploratory subsets using minimum-support thresholds, so that claims about individual labels are not based on extremely small counts. Reconstruction targets are reported, but they are not treated as the primary object of evaluation.

Since manual reference annotation may be incomplete or underspecified, automatic scoring can over-penalise outputs that are linguistically plausible but do not match the gold line exactly. We therefore conduct blinded post-hoc adjudication on disagreement cases from the final model. For each reviewed utterance, the reviewer sees the original input and candidate annotations without knowing whether a candidate comes from the gold reference or from the model. Each candidate is assigned one of four labels (`correct`, `incorrect`, `ambiguous`, or `unsure`) and particularly informative cases are flagged for qualitative analysis. Source identity is stored separately and revealed only after review is complete. These judgments are then merged with the hidden source labels to estimate how often apparent automatic errors reflect genuine model failures, as opposed to ambiguity or omission in the reference annotation. We finally report a conservative updated exact match estimate for the label-bearing subset.

To assess generalisation to unseen data from the same corpus, we run the final model on the remainder of the ENNI corpus, comprising 13,637 utterances not used for training or in-domain evaluation. We manually review all 854 utterances for which the model produced an error annotation, a random sample of 2,200 clean unannotated utterances to estimate the frequency of missed errors, and 91 unannotated but modified utterances. The total reviewed sample is therefore 3,145 utterances. We apply the same coding labels as in the post-hoc review of the automatic evaluation.

5 Results

This section reports the results of `TalkTag` across three complementary stages of evaluation: automatic scoring on held-out training data, blinded post-hoc adjudication, and human review on test inference data. The main question is not whether the system can replace expert annotation, but whether it can provide a reliable first pass that surfaces plausible morphosyntactic error candidates for review.

5.1 Automatic Model Evaluation

As shown in Table 4, on the primary test split, the final model achieves 93.6% exact match, 86.0% micro-precision, 75.5% micro-recall, and 80.4% micro-F1. Validation performance is slightly higher, with 95.4% exact match and 89.0% micro-F1. On the synthetic support split, the model reaches 93.2% micro-F1. Because the real test split is dominated by clean utterances (600/687), we also report results restricted to the 87 utterances containing at least one gold error label. On this subset, the model achieves 66.7% exact match, 91.4% precision, 75.5% recall, and 82.7% micro-F1. The strongest per-label results on the real test data are obtained for overregularised past morphology (F1 89.4, N = 26), missing third-person singular marking (F1 86.3, N = 25), and allomorphic errors (F1 100.0, N = 14), while pronoun substitution is lower at F1 77.8 (N = 10). This pattern suggests a high-precision annotation aid: proposed labels are usually plausible, but the lower recall indicates that some linguistically valid errors remain difficult to recover automatically, especially when the intended target depends on wider discourse context.

5.2 Post-hoc Review

Aware that the manual gold annotations may themselves contain errors or omissions, we carried out blinded post-hoc adjudication on all 44 disagreement utterances in the test set: 20 reviewed disagreements were judged acceptable for both model

Split	N	EM	P	R	F1
Val.	687	95.4%	92.0%	86.2%	89.0%
Test	687	93.6%	86.0%	75.5%	80.4%
Labels	87	66.7%	91.4%	75.5%	**82.7%**
Post-hoc	87	**82.8%**			

Table 4: Automatic evaluation results. EM denotes full-line exact match; P, R, and F1 are micro-averaged over error tags. Labels restrict evaluation to the 87 tagged utterances in Test; Post-hoc refers to the increased score after manual review of disagreement cases.

and reference, and in 5 further cases the model output was preferred, yielding a conservative post-hoc acceptable-output rate of 82.8% (72/87) error-bearing utterances alone.

All the reported results on the real test set are displayed in Table 4, while synthetic support split (Table 7), label-wise results (Table 9), and post-hoc details (Table 8) can be found in the Appendix.

5.3 Human Evaluation of Model Outputs

Results of the model's annotation of the unseen portion of the ENNI corpus are summarised in Table 5. At the utterance level, 93.4% of reviewed outputs were judged acceptable and 6.6% incorrect. At the label level, 83.7% of reviewed label-bearing cases were judged correct. Among the 146 unaccepted label judgments, 94 were false positives, 31 were incorrect labels, and 20 were false negatives. In the audit of 2,200 model-clean utterances, 7 probable missed errors were identified.

At the label level, agreement-related labels were the most frequent in the reviewed predictions. Grouping the three main agreement labels – missing third-person singular marking (e.g., *he go* for *goes*), irregular unmarked for singular (e.g., *he are/were* for *is/was*), and irregular singular for unmarked (e.g., *they is/was* for *are/were*) – yields 408 reviewed instances in total. The next most frequent labels were allomorphic errors (105), overregularised past forms (84), and pronoun substitutions (70). Among past-related labels, the most frequent pattern was substitution of the base form for an irregular past form (e.g., *go* for *went*). Accuracy for these frequent labels was 87.3% for the agreement group as a whole, 92.4% for allomorphic errors, 78.6% for overregularised past forms, 71.4% for pronoun substitutions, and 82.1% for base-for-irregular past substitutions. Detailed per-label results are reported in Table 10 in the Appendix.

Measure	Count	Percent
Reviewed utterances (raw)	3145	100.0%
Out-of-scope exclusions	22	0.7%
Official reviewed total	3123	–
Utterance-level acceptable	2917	93.4%
Utterance-level incorrect	206	6.6%
Label-level reviewed cases	894	–
Label-level correct	748	83.7%
Label-level incorrect	146	16.3%
Incorrect label	31	3.5%
False negative	20	2.2%
False positive	94	10.5%
Audited model-clean utterances	2200	–
Probable missed errors	7	0.32%

Table 5: Human-reviewed evaluation on unseen ENNI data. Percentages for utterance-level outcomes are computed over the reviewed total ($N = 3123$). Label-level percentages are computed over reviewed label-bearing cases ($N = 894$). The clean-audit miss rate is computed over the audited model-clean sample ($N = 2200$).

Among the qualitatively reviewed cases, the clearest and most recurrent linguistic patterns causing confusion involved uninflected verb forms, especially the interaction between tense and agreement and cases of invariant verbs (irregular verbs whose past form is identical to the base form). We focus on this pattern here because it directly addresses a marker of developmental language-disordered speech: preference for uninflected verb forms.

We found 16/17 cases in which the model erroneously assigned the agreement label to uninflected verbs following a third-person singular subject, even though an obligatory context for licensing the past tense was present. Of these, 8 cases involved an invariant verb. We found five additional cases in which invariant verbs were overtly overregularised (e.g., *hurted*, *costed*, *putted*); in these cases, the model did not converge on a single analysis but alternated between overregularisation and double-marking annotation.

6 Discussion

Taken together, the results suggest that a useful tool for fine-grained morphosyntactic error annotation can be developed even under conditions of extreme data sparsity and limited computational resources. Although `TalkTag`'s scores on the full test set are partly inflated by the large number of error-free utterances, performance remains encouraging on genuinely error-bearing cases. In addition, the blinded post-hoc review shows that a non-trivial subset of apparent disagreements reflects linguistic

ambiguity or underspecification in the reference annotation rather than straightforward model failure. This is important both for metric interpretation and for practical use: full-line disagreement might not always correspond to a linguistically unacceptable output.

Clinical and developmental relevance. The results are especially encouraging for agreement-related errors, overregularised past forms and pronoun substitutions, which were among the most frequent and best-supported categories in the reviewed data. These categories are also linguistically meaningful. Tense and agreement morphology, as well as difficulties with pronouns, are central to the study of developmental language disorder and autism spectrum disorder (Leonard and Deevy, 2020; Eadie et al., 2002; Rice et al., 1998; Wexler et al., 1998; Rice and Wexler, 1996). Overregularisation is a well-established feature of typical language development and remains informative when it persists or occurs at elevated rates in atypical development (Moraleda-Sepúlveda and López-Resa, 2022; Marcus et al., 1992). From this perspective, the model's relative success on these labels is encouraging not only in engineering terms but also because it aligns with clinically and developmentally relevant dimensions of child language.

Agreement bias in tense-agreement ambiguities. At the same time, the qualitative review revealed a clear and recurrent failure mode involving bare verb forms following third-person singular subjects. In these cases, the model often preferred agreement-based analyses over missing past-tense interpretations, including in some contexts where the surrounding discourse licensed a past reading. This pattern was especially pronounced for zero-change verbs such as *hurt* and *put*, whose past forms are identical to their base forms. More generally, zero-change verbs formed a persistent challenge in the reviewed sample, and when they were overtly overregularised (e.g., *hurted*, *costed*, *putted*), the model alternated between overregularisation and double-marking analyses rather than converging on a single label decision.

These patterns are informative because they suggest that the model's errors are not simply random, but partly structured by a preference for locally recoverable agreement analyses over broader tense interpretation. One plausible explanation is that agreement errors are both more locally diagnosable and more strongly represented in the training signal, whereas missing past-tense interpretations often require integration of wider temporal and discourse context. At the same time, these cases also highlight a genuine property of the task itself: in spontaneous non-standard speech, morphosyntactic annotation is often difficult precisely because sparse morphology, discourse context, and lexical irregularity interact.

Generalisation and pre-annotation utility. The broader review on unseen ENNI material seems to reinforce this picture. Although overall acceptability remained high, the error profile on this larger sample differed somewhat from that of the held-out test set: agreement-related labels again dominated the reviewed predictions, but the main source of degradation was over-annotation rather than omission. The audit of model-clean utterances nevertheless suggests that silent misses remain comparatively infrequent. This pattern supports a cautious interpretation of the tool as one that is more useful for surfacing plausible candidate errors than for producing final annotations without review.

From a practical perspective, the present results suggest that the tool is already useful as a pre-annotation aid, even where its outputs still require human correction. Prior work on machine-assisted annotation has shown that automatic pre-annotation can reduce annotation effort, improve consistency, and increase annotation speed without harming final quality (Lingren et al., 2014; Fort and Sagot, 2010; Mikulová et al., 2022). The contribution of the present system is therefore not that it eliminates the need for expert review, but that it provides a linguistically informed first-pass annotation over a large volume of CHAT material. This is especially valuable in a domain where fully manual annotation is slow, costly, and itself subject to ambiguity and inconsistency.

Scope of the present prototype. At the same time, the present findings should be interpreted within the scope of the current prototype. The model was developed and evaluated on children's narrative data from a single corpus family, and the synthetic support split serves only as secondary evidence of label-space coverage under controlled conditions rather than as a substitute for naturalistic evaluation. Moreover, some reviewed errors were tied to pre-existing CHAT annotations or discourse-formatting cues rather than to morphosyntactic analysis alone.

Limitations

The present study should be understood as a prototype rather than a complete solution to CHAT-style error annotation. The model was developed under conditions of extreme data scarcity and trained on a reduced subset of the CHAT error inventory, focusing on selected morphosyntactic and closely related substitution labels. As a result, the current system does not yet cover the full range of CHAT-compatible error phenomena, and support for some rare or irregular patterns remains limited. This is particularly relevant for error types that were only sparsely represented in the available training data. Importantly, this is not simply an artefact of the present experiment: some of these phenomena are genuinely infrequent in naturalistic corpora, which makes it inherently difficult to obtain enough real examples for robust learning.

The current annotation scope is also intentionally narrow. For example, forms such as *they going to the shop* may plausibly be analysed not as cases of superfluous progressive marking, but as instances of missing auxiliary. However, the present model was not trained to annotate missing parts of speech, and such cases therefore fall outside the annotation scope of the current system. Within that restricted label space, we treated the model's superfluous-progressive analysis as acceptable, since it captures a genuine deviation while avoiding unsupported labels. More generally, the prototype is better understood as a first step toward fine-grained morphosyntactic error annotation than as a holistic grammar annotation tool.

These constraints also make it important to understand what the model is learning under conditions of scarcity. The present study does not disentangle the mechanisms by which the model produces its annotations. The fine-tuning was designed to encourage recovery of an intended target form while generating the final inline label, making the approach loosely related to an analysis-by-synthesis perspective. At the same time, the model was also allowed to generate CHAT labels compositionally rather than retrieve them from a fixed inventory of whole forms. The current results, therefore, do not establish whether the observed gains arise from implicit target-form reconstruction, from better formatting control, or from a learned mapping between linguistic error patterns and the internal structure of the annotation language. Future work will investigate these possibilities more directly, including whether improvements are concentrated in error types that require target-form recovery and whether performance depends on latent correction-like inference.

A further limitation concerns generalisation. Although the model was evaluated both on held-out data and on new unseen material, all real-data evaluation was conducted within the ENNI corpus. This provides a useful first test of robustness, but it does not establish how well the model transfers across other CHILDES corpora, elicitation settings, age ranges, or clinical populations. Future work should therefore test the system on additional child-language corpora, as well as on more clearly out-of-distribution material such as second-language learner data and adult's clinical speech.

These limitations point to the main conditions under which the prototype should be used: as a human-in-the-loop pre-annotation aid rather than as a fully automatic replacement for expert judgement. Broader annotation coverage, greater robustness to authentic CHAT markup, and wider evaluation across corpora and populations will be needed before the system can be treated as a more general annotation tool.

Acknowledgments

This work was supported by the Ministry of Science, Research and the Arts of Baden-Württemberg (MWK) and KIT's Accessibility through AI-based Assistive Technology (KATE) Graduate School.

Generative AI Assistance Declaration During the preparation of this work, the author(s) used ChatGPT to rephrase, proofread or summarise text content. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Data Availability Statement The data used in this study are not redistributed with our code release. They are hosted by TalkBank/CHILDES and should be obtained directly from the official source under the applicable TalkBank access and licensing rules.

References

Abeer Z. Al-Marridi, Samawiyah M. Ulde, Ahmed Bensaid, and Tariq A. Khwaileh. 2026. Speech and language disorders: A systematic review of corpora and future directions. *Applied Corpus Linguistics*, 6(1):100186.

Erika Benassi, Sonia Boria, Maria Teresa Berghenti, Michela Camia, Maristella Scorza, and Giuseppe Cossu. 2021. Morpho-syntactic deficit in children with cochlear implant: Consequence of hearing loss or concomitant impairment to the language system? *International Journal of Environmental Research and Public Health*, 18(18):9475.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, pages 1–59.

Louise Cummings. 2023. Communication disorders: A complex population in healthcare. *Language and Health*, 1(2):12–19.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

P. A. Eadie, M. E. Fey, J. M. Douglas, and C. L. Parsons. 2002. Profiles of grammatical morphology and sentence imitation in children with specific language impairment and down syndrome. *Journal of Speech, Language, and Hearing Research*, 45(4):720–732.

Karën Fort and Benoît Sagot. 2010. Influence of pre-annotation on POS-tagged corpus development. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 56–63, Uppsala, Sweden. Association for Computational Linguistics.

Gloria Gagliardi and Marta Maffia. 2024. Language resources for clinical linguistics: introduction to the special issue. *Language Resources and Evaluation*, 58(3):859–863.

Christopher Gebauer, Lars Rumberg, Lars Köhn, Hanna Ehlert, Edith Beaulac, and Jörn Ostermann. 2025. Grammatical Error Detection on Spontaneous Children's Speech Using Iterative Pseudo Labeling. In *Interspeech 2025*, pages 2865–2869.

Samane Dehghani Golestani, Nahid Jalilevand, and Mohammad Kamali. 2018. A comparison of morpho-syntactic abilities in deaf children with cochlear implant and 5-year-old normal-hearing children. *International Journal of Pediatric Otorhinolaryngology*, 110:27–30.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv preprint*.

Khairun-nisa Hassanali and Yang Liu. 2011. Measuring language development in early childhood education: A case study of grammar checking in child language transcripts. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 87–95, Portland, Oregon. Association for Computational Linguistics.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Timothy Huang and Lizbeth Finestack. 2020. Comparing morphosyntactic profiles of children with developmental language disorder or language disorder associated with autism spectrum disorder. *American Journal of Speech-Language Pathology*, 29(2):714–731.

Dimitra Katsarou and Georgia Andreou. 2022. Morphosyntactic abilities in young children with down syndrome: Evidence from the greek language. *International Journal of Language & Communication Disorders*, 57(5):937–947.

Katerina Korre and John Pavlopoulos. 2020. ERRANT: Assessing and improving grammatical error type classification. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 85–89, Online. International Committee on Computational Linguistics.

Laurence B. Leonard and Patricia Deevy. 2020. Retrieval practice and word learning in children with specific language impairment and their typically developing peers. *Journal of Speech, Language, and Hearing Research*, 63(10):3252–3262.

Todd Lingren, Louise Deleger, Katalin Molnar, Haijun Zhai, Jareen Meinzen-Derr, Megan Kaiser, Laura Stoutenborough, Qi Li, and Imre Solti. 2014. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing

gold standard development for clinical named entity recognition in clinical trial announcements. *Journal of the American Medical Informatics Association*, 21(3):406–413.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Brian MacWhinney. 2012. Morphosyntactic analysis of the CHILDES and TalkBank corpora. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2375–2380, Istanbul, Turkey. European Language Resources Association (ELRA).

Brian MacWhinney. 2019. Chat manual.

Brian MacWhinney, Steven Bird, Christopher Cieri, and Craig Martell. 2004. Talkbank: Building an open unified multimodal database of communicative interaction. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Brian MacWhinney and Davida Fromm. 2022. Language sample analysis with talkbank: An update and review. *Frontiers in Communication*, 7.

Gary F. Marcus, Steven Pinker, Michael Ullman, Michelle Hollander, T. John Rosen, and Fei Xu. 1992. Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, 57(4):1–182.

Marie Mikulová, Milan Straka, Jan Štěpánek, Barbora Štěpánková, and Jan Hajic. 2022. Quality and efficiency of manual annotation: Pre-annotation bias. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2909–2918, Marseille, France. European Language Resources Association.

Jon F. Miller, Karen Andriacchi, and Ann Nockerts. 2011. *Assessing Language Production Using SALT Software: A Clinician's Guide to Language Sample Analysis*. SALT Software, LLC.

Tomoya Mizumoto, Yuta Hayashibe, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2012. The effect of learner corpus size in grammatical error correction of ESL writings. In *Proceedings of COLING 2012: Posters*, pages 863–872, Mumbai, India. The COLING 2012 Organizing Committee.

Esther Moraleda-Sepúlveda and Patricia López-Resa. 2022. Morphological difficulties in people with developmental language disorder. *Children*, 9(2):125.

Eric Morley, Anna Eva Hallin, and Brian Roark. 2014. Data driven grammatical error detection in transcripts of children's speech. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 980–989, Doha, Qatar. Association for Computational Linguistics.

Eric Morley, Brian Roark, and Jan P. H. Santen. 2013. The utility of manual and automatic linguistic error codes for identifying neurodevelopmental disorders. In *BEA@NAACL-HLT*.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria. Association for Computational Linguistics.

Mitja Nikolaus, Abhishek Agrawal, Petros Kaklamanis, Alex Warstadt, and Abdellah Fourtassi. 2024. Automatic annotation of grammaticality in child-caregiver conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1832–1844, Torino, Italia. ELRA and ICCL.

Jack L. Paradise, Thomas F. Campbell, Christine A. Dollaghan, Heidi M. Feldman, Beverly S. Bernard, D. Kathleen Colborn, Howard E. Rockette, Janine E. Janosky, Dayna L. Pitcairn, Marcia Kurs-Lasky, Diane L. Sabo, and Clyde G. Smith. 2005. Developmental outcomes after early or delayed insertion of tympanostomy tubes. *New England Journal of Medicine*, 353(6):576–586.

Martina Penke. 2019. Regular and irregular inflection in down syndrome – new evidence from german. *Cortex*, 116:192–208.

Mabel L. Rice and Kenneth Wexler. 1996. Toward tense as a clinical marker of specific language impairment in english-speaking children. *Journal of Speech and Hearing Research*, 39(6):1239–1257.

Mabel L. Rice, Kenneth Wexler, and Scott Hershberger. 1998. Tense over time: The longitudinal course of tense acquisition in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, 41(6):1412–1431.

Lars Rumberg, Christopher Gebauer, Hanna Ehlert, Maren Wallbaum, Lena Bornholt, Jörn Ostermann, and Ulrike Lüdtke. 2022. kidstalc: A corpus of

3- to 11-year-old german children's connected natural speech. In *Interspeech 2022*, interspeech_2022, pages 5160–5164. ISCA.

Kenji Sagae, Alon Lavie, and Brian MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 197–204, Ann Arbor, Michigan. Association for Computational Linguistics.

Phyllis Schneider, Denyse Hayward, and Rita Vis Dubé. 2006. Edmonton narrative norms instrument.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 198–202, Jeju Island, Korea. Association for Computational Linguistics.

Kenneth Wexler, Carson T. Schütze, and Mabel Rice. 1998. Subject case in children with SLI and unaffected controls: Evidence for the Agr/Tns omission model. *Language Acquisition*, 7(2–4):317–344.

Bernadette Witecy, Eva Wimmer, Isabel Neitzel, and Martina Penke. 2023. Morphosyntactic development in german-speaking individuals with down syndrome–longitudinal data. *Frontiers in Psychology*, 14.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

A Annotation Prompt

Role. You are a TalkBank CHAT annotator for morphosyntactic error coding.
Task. Annotate the input utterance by inserting valid CHAT error tags inline.
Output requirements.

1. Preserve original token order, spelling, casing, punctuation, disfluencies, and CHAT symbols.
2. Do not rewrite, paraphrase, or correct the utterance.
3. Insert only error tags inline, following the error token.
4. If no target error is present, return the utterance unchanged.
5. Write the correct target form as `[: target]` when the incorrect morpheme yields a nonword, and as `[:: target]` when the error is an attested word.
6. Build each CHAT error tag compositionally from licensed scheme parts rather than relying on a memorised whole-label form.
7. Use `m:*` only for same-lexeme morphological contrasts and `s:*` only for substitutional contrasts.
8. Use `:a` only for agreement-sensitive labels that license it.
9. Use `:i` only where an irregular-sensitive label licenses it.
10. Output only licensed CHAT tags; do not invent unattested or unsupported combinations.
11. Output exactly one annotated utterance line and nothing else.

B Training Label Inventory

Label	Label	Label	Label
`[* m:++ed:i]`	`[* m:+ing]`	`[* m:base:er]`	`[* m:vsg:a]`
`[* m:++ed]`	`[* m:+s:a]`	`[* m:base:est]`	`[* m:vun:a]`
`[* m:++en:i]`	`[* m:+s]`	`[* m:base:s]`	`[* s:r:der]`
`[* m:++s:i]`	`[* m:0's]`	`[* m:irr:ed]`	`[* s:r:gc:det]`
`[* m:++s]`	`[* m:03s:a]`	`[* m:irr:en]`	`[* s:r:gc:pro]`
`[* m:+3s:a]`	`[* m:0ed]`	`[* m:irr:s]`	`[* s:r:prep]`
`[* m:+3s]`	`[* m:0ing]`	`[* m:sub:ed]`	
`[* m:+ed:i]`	`[* m:0s:a]`	`[* m:sub:en]`	
`[* m:+ed]`	`[* m:=ed]`	`[* m:base:ed]`	
`[* m:+en]`	`[* m:=en]`	`[* m:base:en]`	
`[* m:=s]`	`[* m:allo]`		

Table 6: Full training-set label inventory for the final confirmatory model package. The table lists all CHAT morphosyntactic error labels seen during training, irrespective of whether later evaluation support for a label comes from real or synthetic data.

C Additional Evaluation Results

Split	N	EM	P	R	F1
Val.	370	82.4%	91.9%	88.7%	90.3%
Test	370	82.7%	94.5%	92.0%	93.2%

Table 7: Automatic evaluation on the synthetic support splits. Exact denotes full-line exact match against the gold annotated utterance; P, R, and F1 are micro-averaged over CHAT error tags.

Subset	N	Prev. exact	Reviewed	Both acc.	Model pref.	Post-hoc acc.
Test real (full)	687	643 (93.6%)	44	20	5	97.2%
Test real (labelled)	87	58 (66.7%)	29	11	3	82.8%

Table 8: Post-hoc adjudication results on `test_real`. Previous exact gives the number of exact automatic matches before manual review. Post-hoc acceptable counts exact matches plus reviewed disagreement cases judged acceptable for both outputs or preferred for the model.

test_real					test_coverage				
Label	**P**	**R**	**F1**	**N**	**Label**	**P**	**R**	**F1**	**N**
[* m:allo]	100.0	100.0	100.0	14	[* m:=en]	100.0	100.0	100.0	12
[* m:vun:a]	100.0	100.0	100.0	2	[* m:irr:en]	100.0	100.0	100.0	11
[* s:r:der]	100.0	100.0	100.0	1	[* m:=ed]	100.0	100.0	100.0	10
[* m:=ed]	100.0	80.8	89.4	26	[* m:++en:i]	100.0	100.0	100.0	10
[* m:03s:a]	84.6	88.0	86.3	25	[* m:+en]	100.0	100.0	100.0	10
[* s:r:gc:pro]	87.5	70.0	77.8	10	[* m:0's]	100.0	100.0	100.0	10
[* m:base:ed]	75.0	75.0	75.0	4	[* m:base:en]	100.0	100.0	100.0	10
[* m:++ed:i]	100.0	50.0	66.7	2	[* m:base:er]	100.0	100.0	100.0	10
[* m:++ed]	50.0	100.0	66.7	1	[* m:base:est]	100.0	100.0	100.0	10
[* m:vsg:a]	50.0	50.0	50.0	2	[* m:irr:ed]	100.0	100.0	100.0	10
[* s:r:prep]	50.0	33.3	40.0	3	[* m:sub:ed]	100.0	100.0	100.0	10
[* m:+ed]	0.0	0.0	0.0	2	[* m:sub:en]	100.0	100.0	100.0	10
[* m:0ing]	0.0	0.0	0.0	2	[* m:++s]	90.9	100.0	95.2	10
[* s:r:gc:det]	0.0	0.0	0.0	2	[* m:+s:a]	90.9	100.0	95.2	10
[* m:0ed]	0.0	0.0	0.0	1	[* m:irr:s]	90.9	100.0	95.2	10
[* m:0s:a]	0.0	0.0	0.0	1	[* m:+ing]	90.0	90.0	90.0	10
					[* m:+3s]	81.2	100.0	89.7	13
					[* m:=s]	100.0	80.0	88.9	10
					[* m:base:s]	100.0	80.0	88.9	10
					[* m:++s:i]	88.9	80.0	84.2	10
					[* m:+3s:a]	100.0	60.0	75.0	10
					[* m:03s:a]	80.0	66.7	72.7	12

Table 9: Per-label automatic evaluation grouped by evaluation source. Values are percentages except for support (N). Real rows are taken from `test_real`; synthetic rows are labels absent from `test_real` and therefore reported on `test_coverage`.

Label	**Correct**	**N**	**Label**	**Correct**	**N**
[* m:03s:a]	88.7%	328	[* m:sub:en]	100.0%	6
[* m:allo]	92.4%	105	[* s:r:prep]	83.3%	6
[* m:=ed]	78.6%	84	[* m:+ed]	100.0%	5
[* s:r:gc:pro]	71.4%	70	[* m:0s:a]	75.0%	4
[* m:vun:a]	68.2%	44	[* m:0's]	50.0%	4
[* m:vsg:a]	97.2%	36	[* m:irr:s]	50.0%	4
[* m:base:ed]	82.1%	28	[* m:=en]	33.3%	3
[* s:r:der]	23.8%	21	[* m:++s:i]	50.0%	2
[* m:irr:ed]	86.7%	15	[* m:+3s:a]	50.0%	2
[* m:++ed:i]	76.9%	13	[* m:++est]	100.0%	1
[* m:++ed]	61.5%	13	[* m:+en]	100.0%	1
[* m:0ing]	100.0%	11	[* m:base:en]	100.0%	1
[* m:+3s]	54.5%	11	[* m:irr:en]	100.0%	1
[* m:0ed]	80.0%	10	[* m:=ed:i]	0.0%	1
[* s:r:gc:det]	50.0%	10	[* m:=ing]	0.0%	1
[* m:sub:ed]	100.0%	9	[* m:base:der]	0.0%	1
[* m:+ing]	100.0%	8			

Table 10: Per-label human-reviewed exactness on the rest of the ENNI corpus, sorted by support. Because ENNI does not have exhaustive gold annotation, this table does not report recall or F1.

D Expanded CHAT Scheme Reference

Level 1		Level 2		Level 3	
Code	**Meaning**	**Code**	**Meaning**	**Code**	**Meaning**
[* m:]	morphosyntactic error	0	missing	-ing	progressive
[* s:]	substitution error	base:	bare form	-3s	3SG
		sub:	substitution	-ed	past
		irr:	irregular	-en	perfective
		=	over-regularisation	-s	plural
		+	superfluous	's	possessive
		++	double marking	-s'	plural possessive
		vsg:	irregular verb 3SG	-er	comparative
		vun	irregular verb unmarked	-est	superlative
		allo	allomorph	a	agreement
		s:r:	related lexical substitution	i	irregular
		s:r:gc:	related grammatical substitution	POS	target POS

Table 11: Expanded reference table for CHAT-style error-label components used in this study.