Delayed *Wh*-Question Development in Children with Hearing Loss: Evidence for Morphosyntactic Vulnerability from Corpus-Based NLP and LLM Analyses

Tong Wu
Leiden University
t.wu.7@umail.leidenuniv.nl

Abstract

This study provides corpus-based evidence that English-speaking children with hearing loss (CHL) show both quantitative and qualitative delays in *wh*-question development compared to typically developing (TD) peers. Using Natural Language Processing (NLP)/Large Language Model (LLM) based methods and two clinical subcorpora from CHILDES, we analyzed child utterances across several syntactic dimensions: frequency, lexical diversity, structural completeness, clausal embedding, *wh*-fronting, and utterance length. CHL produced significantly fewer *wh*-questions, used a narrower range of *wh*-types, showed lower rates of embedding, and more structural incompleteness. These differences were most evident in syntactically complex forms, such as embedded and canonical fronted *wh*-questions. The results support input-sensitive and usage-based accounts of syntactic development and highlight the need for enriched linguistic input in supporting CHL's grammatical growth. Importantly, these group differences persisted when controlling for overall language development as indexed by mean length of utterance (MLU) in words, indicating that CHL's difficulties with *wh*-questions are not reducible to general grammatical delay. Methodologically, the study combines dependency-parsing-based analyses with exploratory LLM evaluation to assess the feasibility and limits of automated approaches to spontaneous child language. NLP-based analyses were more stable for formally defined syntactic features, while GPT-based analysis showed mixed performance, performing better on global structural judgments than on fine-grained syntactic diagnostics.

1 Introduction

1.1 Background

Syntax plays a crucial role in human languages, since it determines how words are organized, and information is packaged in the combination system (Jackendoff, 2003). Given its central role in language organization, syntax also serves as a key developmental marker in first language acquisition. The transition from telegraphic speech to productive syntax reflects the child's internalization of grammatical rules (Brown, 1973). In clinical application of language development, difficulties with syntax are widely recognized as reliable markers of language impairment (Leonard, 2017).

Hearing loss (HL) is considered one of the most prevalent developmental disorders (Lieu et al., 2020). Although HL primarily affects phonology, it also disrupts syntax, even when vocabulary is comparable. Children with hearing loss (CHL) often produce fewer complex syntactic constructions, such as subordinate clauses and *wh*-questions, compared to their typically developing (TD) peers (Werfel, 2017). Additionally, limited diversity in CHL children can signal delayed or atypical syntactic development (Klieve et al., 2023). Notably, these syntactic delays persist even when vocabulary size is controlled for, suggesting a specific disruption in grammatical development.

Among the various syntactic constructions, *wh*-questions have been widely recognized as particularly informative indicators of grammatical development. *Wh*-questions are structurally complex constructions that require advanced syntactic operations, such as *wh*-movement to CP (complementizer Phrase) and subject-auxiliary inversion (Adger, 2003; Radford, 2004). Their development is prolonged: early forms like *what* and *where* emerge first, while complex forms like *why* and *how* appear later, making *wh*-questions a sensitive indicator of syntactic maturity (Ervin-Tripp, 1970; Tyack and Ingram, 1977; Bloom et al., 1982; Rowland et al., 2003). For CHL, complex *wh*-constructions, such as embedded *wh*-questions and developmentally late *wh*-types like *why* and *how*, are frequently delayed or inaccurate, reflecting broader difficulties with complex syntax (Werfel

Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026), pages 291–304

et al., 2021; Klieve et al., 2023).

(1) a. TD child: how I get the plate out?
(Ambrose 02jw_36, age = 3;0)
b. HL child: how do?
(Ambrose lr24_36, age = 3;0)

Recent Speech Language Pathology (SLP) research has shifted from categorical deficit accounts toward viewing language development in deaf and hard-of-hearing children as attenuated and variably robust under conditions of reduced language access, with particular vulnerability in complex grammatical constructions (Goodwin et al., 2022; Coppola and Walker, 2025). Natural Language Processing (NLP) and Large Language Models (LLMs) are also introduced in this line of research (Lammert et al., 2025).

In this study, we use CHILDES (Child Language Data Exchange System) data together with NLP- and LLM-based analyses of *wh*-questions to empirically investigate how such developmental constraints are manifested both in the quantity and in the structural organization of interrogative syntax. In addition, methodological reflection on the basic NLP tools and LLMs used in this study provides a window for further computational research on language impairment data. Unlike previous task-based elicitation studies, this work provides a spontaneous, corpus-based comparison across multiple syntactic dimensions, addressing a gap in the naturalistic syntactic profiling of CHL.

1.2 Research Questions

We ask whether CHL and TD children differ in both the quantitative and the qualitative properties of *wh*-question production in naturalistic spoken English across developmental stages (Section 4). At the quantitative level, we examine group differences in the frequency, utterance length, and diversity of wh-forms (RQ1). At the qualitative level, we investigate structural properties of *wh*-questions, including syntactic completeness and complexity as reflected in *wh*-fronting, subject–auxiliary inversion, embedding, and formulaic responses (RQ2). We further examine how these quantitative and qualitative patterns are modulated by age in CHL compared to TD children (RQ3).

Finally, building on the empirical patterns identified in these analyses, we assess the extent to which different NLP paradigms, specifically dependency-parsing-based methods and LLMs, differ in their effectiveness at characterizing *wh*-question use in child language data (RQ4: Section 5).

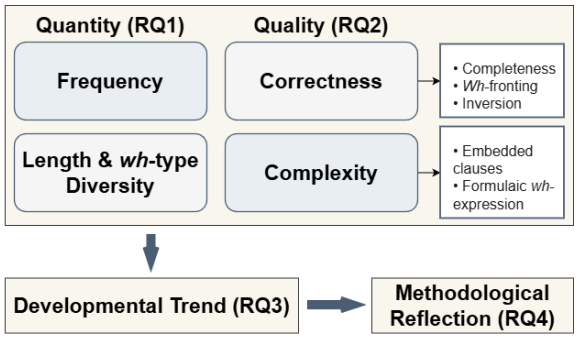

Figure 1: Analytical dimensions.

This paper is organized into six sections. Section 2 provides a brief overview of previous linguistic research related to hearing loss and the acquisition of *wh*-questions. Section 3 introduces the data sources and outlines the analytical pipeline. Section 4 presents the analysis results, accompanied by visualizations and discussion grounded in theories of language acquisition and clinical linguistics. Section 5 discusses the methodological implications of the findings for computational approaches to language impairment research. Section 6 summarizes the main findings of the study. Finally, a limitations section concludes the paper.

2 Inspiration from Previous Studies

CHL have consistently shown difficulties in producing complex syntactic constructions. Previous studies using structured tasks have reported low accuracy and frequency in syntactic constructions such as *wh*-complement clauses, including both finite and nonfinite forms (Werfel et al., 2021; Klieve et al., 2023). These difficulties persist even when controlling for vocabulary size or Mean Length of Utterance (MLU), suggesting a specific deficit in grammatical development. Klieve et al. (2023) identified interrogative complements as valuable diagnostic markers. Their acquisition requires coordination of *wh*-movement and clause embedding, which may be especially demanding for CHL due to combined cognitive and auditory load (Koehlinger et al., 2013).

Most existing research has relied on elicited production through narrative tasks. While these methods offer control, such elicitation methods may not fully capture the variability of spontaneous speech. In contrast, corpus-based analyses offer ecologically valid data over time. A combined approach

may thus be essential for a fuller picture of CHL's syntactic development.

Recent corpus-based and computational linguistics studies have shown growing interest in the analysis of clinical populations, including individuals with autism (Ferguson et al., 2009). More recently, a growing body of work has explored the use of NLP and LLM-based methods to provide structured and scalable analyses of disordered language, particularly in spontaneous speech data, with the aim of supporting clinically relevant insights (Jang et al., 2025).

However, research focusing on CHL remains comparatively scarce, and existing studies also emphasize that computational approaches are best treated as analytic aids rather than fully automated solutions, given persistent challenges related to data sparsity and the non-canonical nature of child speech (Lammert et al., 2025). By analyzing *wh*-questions in spontaneous speech, the present study contributes to corpus-based clinical linguistics while addressing an empirical gap in our understanding of syntactic development in CHL.

3 Method

3.1 Data

In SLP, corpus-based research is largely centered on the CHILDES database, now part of the broader TalkBank project (MacWhinney, 2014). Originally developed in the 1980s to archive child language data, CHILDES introduced widely used standards for transcription and analysis (Ferguson et al., 2009).

Spontaneous speech data of this study are extracted from *Clinical English Nicholas corpus* (Nicholas and Geers, 1997) and *Clinical English Ambrose corpus* (Ambrose, 2016), part of the Clinical-Eng subcollection in the CHILDES database. These two corpora include transcripts of the conversation between English-speaking children and their parents, encompassing both TD individuals and CHL, with the two groups approximately matched by age. Parent-child interactions in Nicholas corpus were recorded when children were 12, 18, 24, 30, 36, 42, 48, and 54 months of age, while Ambrose corpus examined children at roughly 13.5, 18, 22.5, 27, and 36 months.

To maintain comparability, we focus exclusively on fully transcribed child-produced utterances as marked by the **CHI:* tier. Non-verbal annotations and utterances by other participants were excluded. The corpora contain a total of 64,814 child utterances in the full transcripts, including 42,398 from TD children and 22,416 from CHL. These counts refer to all child utterances and precede the extraction of *wh*-related utterances for analysis.

3.2 Procedure

The analytical workflow was designed to balance linguistic interpretability with scalability for large-scale child language corpora. All analyses were implemented in Python (v3.11) using spaCy (v3.8.7) for syntactic annotation, together with standard data-science libraries (pandas, numpy, scipy, statsmodels). The procedure consists of four stages: preprocessing, automated linguistic analysis, statistical modeling, and validation. We applied a dependency-parsing-based automated annotation pipeline to all 64,814 utterances.

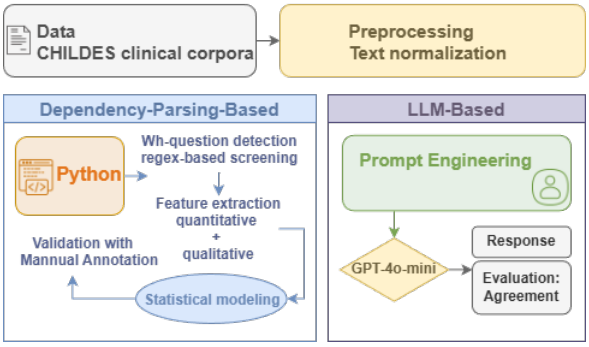

Figure 2: Overview of the analytical workflow.

Preprocessing. Annotations, such as bracketed repairs, event tags, non-speech markers, were removed using rule-based normalization to ensure comparability across files while preserving surface word order. Each transcript was associated with child age in months, corpus source, and group.

Identification of *wh*-questions. Target *wh*-forms included *what*, *who*, *where*, *when*, *why*, *how*, and *which*. Utterances containing at least one of these forms were identified via regular-expression matching. Frequencies were normalized per 100 child utterances to control for differences in corpus size and recording density.

Quantitative measures of *wh*-question use. To characterize the quantity of *wh*-question production, we extracted two file-level measures: lexical diversity (Rowland et al., 2003), defined as the number of distinct *wh*-types per transcript, and utterance length (Bloom et al., 1982; Rice, 2004;

Klieve et al., 2023), measured as the mean number of tokens per *wh*-question.

Qualitative structural measures. Qualitative properties of *wh*-questions were assessed along three dimensions targeting syntactic complexity, structural well-formedness, and productivity.

- **Embedded clauses.** An utterance was classified as embedded if it contained a matrix predicate from a predefined verb set (Diessel, 2004), e.g., *think, know, say, ask, wonder*, licensing a clausal dependent (*ccomp/xcomp/advcl*) under spaCy dependency parsing, whose subtree contained a *wh*-word. Formulaic *wh*-templates were excluded.
- **Syntactic completeness and inversion.** *Wh*-questions were automatically classified as complete, fragment, missing subject, or missing verb, based on the presence of a finite predicate and a grammatical subject (Radford, 2004). Subject–auxiliary inversion was detected using rule-based criteria sensitive to auxiliary position, subject *wh*-questions, and embedded contexts. Importantly, inversion was evaluated only for structurally complete, non-subject *wh*-questions in matrix clauses; fragments, subject *wh*-questions, and embedded *wh*-questions were excluded from the inversion analysis (Rowland et al., 2005).
- **Formulaic *wh*-expressions.** Formulaic expressions were defined as recurrent fixed or semi-formulaic *wh*-frames with contracted auxiliaries like *what's*, *who's* (Brown, 1973; Rowland et al., 2005; Roeper and De Villiers, 2011), and detected using predefined lexical patterns. Discourse markers were excluded when occurring outside a *wh*-clause.

Statistical analysis. All statistical analyses reported in Sections 4 are based on the dependency-parsing-based NLP automated annotations of the full corpus. Analyses were conducted at two levels. File-level continuous measures including normalized frequency, lexical diversity, length, fronting ratio, embedding ratio were compared between TD and HL children using Welch's independent-samples *t*-tests. Utterance-level categorical outcomes like syntactic completeness and inversion were evaluated using chi-square tests (Baayen, 2008; Johnson, 2011).

To model developmental trends, we fitted linear regression models with age, group, and their interaction as predictors. Significant age $\times$ group interactions were interpreted as evidence for divergent developmental patterns rather than uniform delay. Statistical significance was evaluated at $\alpha = .05$.

Validation. To assess the reliability of the automated analyses, we conducted a stratified manual validation on a balanced sample of 200 *wh*-utterances (100 TD, 100 HL), stratified by corpus source, syntactic completeness, and embedding status. Each utterance was manually annotated along multiple dimensions, including *wh*-question status, syntactic completeness, *wh*-fronting, and subject–auxiliary inversion. Manual annotations were compared against automated labels using Cohen's κ, accuracy, and confusion matrices for each syntactic dimension (Artstein and Poesio, 2008).

Model-based Extension. In addition, we conducted an exploratory analysis using a GPT-based large language model for *wh*-question detection (Gilardi et al., 2023). This analysis was restricted to a balanced sample of 200 utterances drawn from the validation step. We employed OpenAI's `gpt-4o-mini` model (Chat Completions API), a lightweight instruction-following model chosen for its computational efficiency and suitability for systematic prompt-based evaluation without task-specific fine-tuning or model adaptation (OpenAI, 2024).

The model was used to annotate each utterance along six syntactic dimensions using a fixed, linguistically explicit zero-shot prompt (system + user message), with deterministic decoding settings (temperature = 0, max_tokens = 300; other parameters left at defaults). Outputs were constrained to JSON-only responses with a predefined label schema (Wei et al., 2021; Schulhoff et al., 2024; Liu et al., 2023) (see Figure 3). Each child utterance was presented in isolation, and the prompt emphasized surface-form cues such as the presence of *wh*-forms and interrogative function, while explicitly noting that child speech may include fragments, disfluencies, or non-canonical word order.

This analysis is exploratory and intended as a methodological probe to estimate an upper-bound reference for *wh*-question detection and to assess LLM generalization to structurally irregular child language.

4 Empirical Results and Discussion

Across all syntactic dimensions we examine, CHL experience both quantitative and qualitative delays in acquiring *wh*-questions compared to TD peers (See Figure 4). CHL produced fewer and

Prompt template

You are an expert child language researcher analyzing spontaneous speech from young children (12–54 months).

Task. Annotate 6 dimensions for each wh-utterance. Judge based on SURFACE FORM only - do NOT correct or normalize child errors.

DIMENSION 3: INVERSION

- 1 (Inversion present): the auxiliary precedes the subject in a matrix wh-question e.g., *What is that? Where did you go? Why can't I?*
- 0 (No inversion): no auxiliary is present, or the subject precedes the auxiliary e.g., *What you doing? Where mommy is? Why he crying?*
- NA (Inversion not applicable): inversion is structurally irrelevant in the following cases:
  - Subject *wh*-questions, e.g., *Who came? What happened?*
  - Embedded *wh*-clauses, e.g., *I know what it is*
  - Fragments lacking clausal structure, e.g., *What? Why?*
  - Copula-less forms with no auxiliary available, e.g., *What that?*

Critical exclusions. Inversion was evaluated only for complete, matrix, non-subject wh-questions; subject wh-questions, embedded wh-clauses, and fragments were excluded from inversion judgment.

Output format. Return a JSON object only (no additional text) with a single field: `{"inversion": 1, 0, or "NA"}`. Code `1` if subject–auxiliary inversion is present, `0` if inversion is absent, and `"NA"` if inversion is structurally inapplicable.

IMPORTANT REMINDERS.

- Judge based on SURFACE FORM only
- Do NOT correct or normalize child errors
- Child speech is often fragmentary - that's expected
- When uncertain, be conservative (0 or "NA")

Figure 3: A representative excerpt from the prompt used to identify inversion in child speech.

structurally simpler *wh*-questions across key dimensions: frequency, lexical diversity, syntactic completeness, *wh*-fronting, question length, and formulaic responses (all $p < .001$). These results reflect persistent challenges in CHL's wh-related syntactic development.

Developmental analyses showed steady age-related growth in TD children, while CHL exhibited markedly flatter trends. These findings highlight the syntactic vulnerabilities linked to reduced perceptual access and support input-sensitive, usage-based models of syntax acquisition (Tomasello, 2003; Ibbotson, 2013; Moeller and Tomblin, 2015), while challenging purely maturational or parameter-setting accounts.

4.1 Quantitative Differences in *Wh*-Question Use

As illustrated in Table 1, TD children produced substantially more *wh*-questions and exhibited greater lexical type diversity in *wh*-word use per 100 utterances than CHL across both two corpora. This indicates a clear quantitative delay in interrogative use by CHL. On average, TD children used a broader range of *wh*-types, including *why*, a form typically acquired at later developmental stages (Tyack and Ingram, 1977; Bloom et al., 1982). In contrast, CHL productions showed a highly skewed distribution, being dominated by *what*, with *who*, *where*, *why* and *how* occurring only sporadically and at very low frequencies. Additionally, TD children produced longer *wh*-questions, while CHL utterances were markedly shorter and minimally elaborated. These quantitative gaps are consistent across datasets and age ranges, revealing an apparent group difference in interrogative production (Rice, 2004).

(2) a. TD child: where did the cows room?
(Ambrose 02jw_27, age = 2;3)
b. HL child: where sit?
(Ambrose 76am_27, age = 2;3)

(3) a. TD child: why won't it let me stick it on?
(Nicholas nh48m_brewster, age = 3;11)
b. HL child: why?
(Nicholas hi54f_paloma, age = 4;5)

4.2 Qualitative Differences in Syntactic Structure

Beyond quantitative differences, *wh*-questions produced by TD and CHL children also differ qualitatively in their structural properties.

At the level of structural correctness, TD children are significantly more likely to produce structurally complete *wh*-fronting questions; instead, CHL productions are dominated by fragments and incomplete forms.

With respect to syntactic complexity, group differences are most evident in the use of embedded structures. Although embedded *wh*-questions are rare overall, consistent with their late developmental emergence (Thornton, 1990; Diessel, 2004), TD children produced them at substantially higher rates than CHL children. While embedded *wh*-questions were attested in the TD data, only limited instances

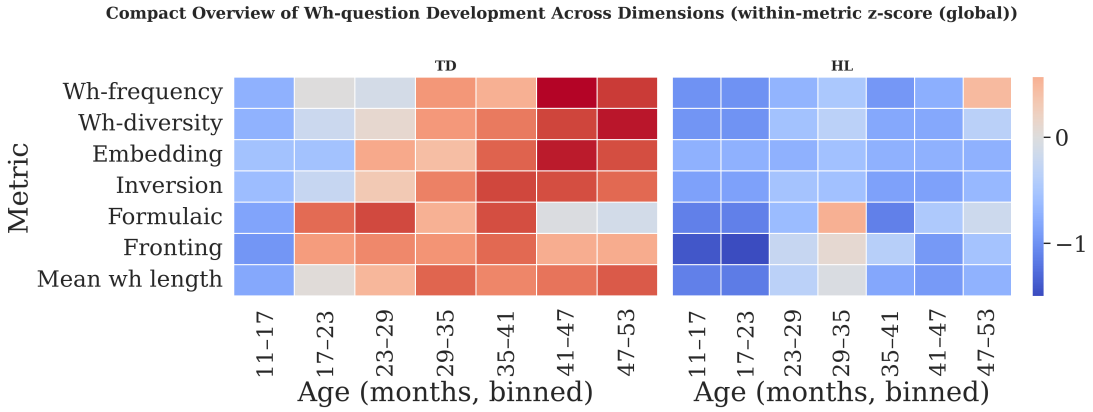

Figure 4: Developmental trends across syntactic dimensions. Rows correspond to syntactic measures, columns to age (in months), and color intensity reflects normalized frequency or proportion of each dimension.

Dimension	**TD**	**HL**	**Test**	p
Quantity				
Frequency	3.20	0.60	$t = 7.66$	$< .001$
Diversity	1.87	0.40	$t = 9.55$	$< .001$
Length	3.79	1.12	$t = 9.15$	$< .001$
Quality: Correctness				
Completeness	—	—	$\chi^2(3) = 87.42$	$< .001$
Complete	.53	.24	—	
Fragment	.39	.59	—	
Fronting	.46	.18	$t = 7.04$	$< .001$
Inversion	.31	.03	$t = 9.22$	$< .001$
Quality: Complexity				
Embedding	.027	.001	$t = 4.11$	$< .001$
Formulaic	.058	.020	$t = 2.80$	.005

Table 1: Group differences in *wh*-question production across syntactic dimensions. Completeness: χ^2 test for overall distribution (complete, fragment, missing-verb, missing-subject). Quality subdivided into Correctness and Complexity. TD = mean values for typically developing children; HL = mean values for children with hearing loss.

were found in the CHL sample. This asymmetry suggests a potential delay in the availability of clausal embedding in children with hearing loss.

(4) a. TD child: I know where that piece can go.
(Nicholas nh36m_calhoun, age = 3;0)
b. TD child: I don't know how.
(Nicholas nh36f_delfina, age = 3;0)
c. HL child: I know where is she.
(Ambrose 76am_36, age = 3;0)

At the same time, the extremely low frequency of embedded constructions poses challenges for heuristic and dependency-parsing detection methods, suggesting that more sensitive or context-aware approaches may be required to fully capture early instances of embedding in spontaneous child speech. We discuss the methodological implications of these issues in the next section below.

Another dimension of syntactic complexity concerns the use of formulaic responses as example (5). Interestingly, these formulaic patterns are more frequent in TD than in CHL productions. This pattern runs counter to previous findings suggesting that children with hearing loss tend to rely on formulaic substitutes in other syntactic constructions (Volpato and Vernice, 2014).

(5) a. TD child: Where's I don't see where this goes.
(Nicholas nh36m_nelek, age = 3;0)
b. HL child: Where's those girls?
(Ambrose 76am_27, age = 2;3)

Consistent with these structural patterns, TD children also produced longer *wh*-questions on average, while CHL utterances remained short and minimally elaborated.

4.3 Developmental Trends by Age

Developmental analyses revealed systematic group and age interactions across all examined dimensions (all interaction $p < .01$). TD children exhibit steep age-related gains in both the quantity and structural quality of *wh*-questions, while CHL show flatter developmental trends across measures.

To distinguish whether group differences in *wh*-question development reflect a general language delay or domain-specific deficits beyond overall grammatical ability, we additionally employed MLU as

a control variable, a widely used index of general language development in child language research (Brown, 1973; Rice et al., 2006; Flipsen and Kangas, 2014). We computed MLU in words (MLU-w) for each child and session as the ratio of valid words to utterances, excluding fillers, unintelligible segments, and CHAT formatting markers.

To enable within-MLU comparisons, we divided the sample into three MLU-w bins based on tertile cutoffs:

- **Low MLU**: MLU-w < 1.16
- **Mid MLU**: $1.16 \leq$ MLU-w < 2.23
- **High MLU**: MLU-w ≥ 2.23

As shown in Figure 5, MLU increased robustly with age in TD children ($r = 0.789$), while HL children showed a substantially weaker age–MLU association ($r = 0.279$), indicating greater developmental variability. The MLU-controlled analyses (Figure 6) show that group differences persist across most *wh*-question dimensions even when overall language ability is held constant. In particular, TD children outperformed HL children in *wh*-question frequency, inversion, and embedded clause production across MLU levels, with embedded *wh*-clauses observed only in the TD High-MLU group. By contrast, fronting showed the smallest group difference, suggesting relative robustness to reduced input.

These findings suggest that the *wh*-question difficulties observed in CHL are not solely attributable to general language delay, but reflect domain-specific patterns consistent with usage-based accounts of language acquisition (Tomasello, 2003; Ambridge and Lieven, 2011).

4.4 Interim Summary

Consistent with previous SLP research (Rice, 2004; Schick et al., 2007; Moeller and Tomblin, 2015; Werfel et al., 2021), CHL showed delayed development across multiple dimensions of *wh*-question use, including reduced *wh*-type diversity, lower rates of embedding and canonical *wh*-fronting, and shorter, less elaborated utterances. Importantly, these group differences persist when controlling for MLU, indicating that *wh*-question production remains selectively vulnerable beyond general language delay.

5 Methodological Implications

Given that the annotation pipeline relies on off-the-shelf dependency-parsing-based NLP tools not tailored to child language, we assess their reliability by comparing them against manual annotations with a LLM-based trial. To assess the potential of large language models as an alternative annotation strategy, we evaluated GPT-4o-mini against manual annotations on the same validation subset. Agreement was assessed by directly comparing manual annotations with two automatic annotation methods: a rule-based system and a GPT-based annotator. Results are summarized in Table 2, which reports agreement metrics for both methods across all syntactic dimensions.

Structural properties such as syntactic completeness, inversion, and embedding show strong agreement with manual annotation, particularly for the rule-based system (κ = .67–.96; accuracy = .90–.98), indicating that these properties are reliably captured by surface-based heuristics and dependency-parsing-based criteria. GPT-based annotations also achieve high agreement for embedding (κ = .77, accuracy = .98), suggesting that this dimension is relatively robust across annotation strategies.

At the same time, Example (6) illustrates a limitation of the embedding detection procedure. Although the utterance contains an embedded temporal clause introduced by *when*, it was classified as non-embedded because the operational criteria target verb-selected complements. This example highlights that adverbial embedding in spontaneous child speech is difficult to exhaustively capture using verb-based heuristics alone.

(6) TD child: I'll go check when we get home.
(Nicholas nh54f_raizel, age = 4;6)

For *wh*-detection and fronting, Cohen's κ values approach zero despite relatively high classification accuracy. This pattern reflects the properties of the manually validated subset, which was obtained via stratified sampling rather than mirroring the natural corpus distribution. Under such sampling conditions, κ becomes sensitive to marginal label distributions and may underestimate agreement for low-prevalence categories, even when accuracy remains high. We therefore interpret agreement on these dimensions descriptively, in relation to error profiles and relative performance across annotation methods, rather than relying on κ as an inferential measure.

Taken together, the two automatic annotation methods exhibit complementary strengths. Rule-based annotation performs particularly well on

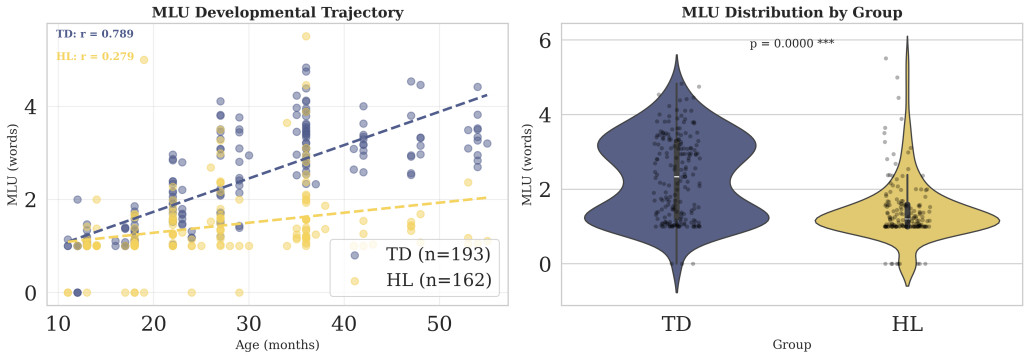

Figure 5: MLU developmental trajectories. Left: Scatter plot showing MLU as a function of age, with regression lines for TD ($r = 0.789$) and HL ($r = 0.279$) groups. Right: Violin plots showing MLU distribution by group ($p < 0.0001$).

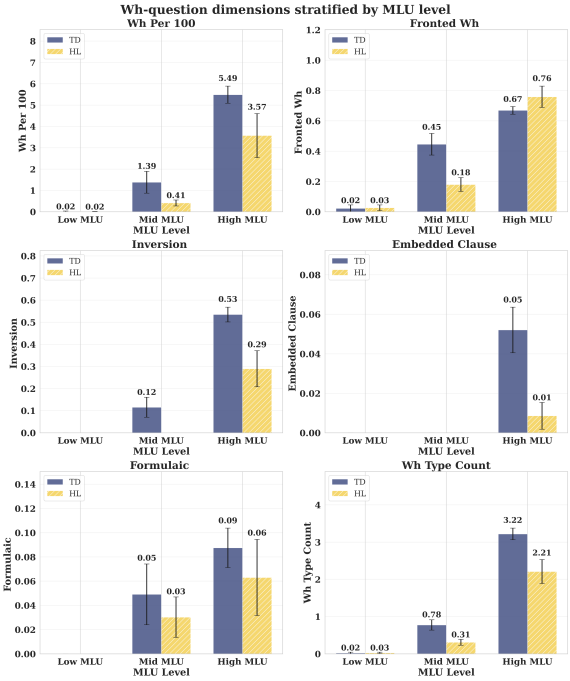

Figure 6: *Wh*-question dimensions stratified by MLU level. Error bars indicate standard errors. Group differences persist across most dimensions even when controlling for overall language ability.

structurally well-defined properties, while GPT-based annotation provides added robustness for dimensions that are less amenable to explicit rule formulation. This complementarity motivates their combined use in the analysis pipeline.

A further challenge for automatic annotation concerns the distinction between *wh*-questions and relative clauses like example (7), which can be difficult to disambiguate in fragmentary child speech.

Dimension	Manual vs Rule-based		Manual vs GPT	
	κ	Acc.	κ	Acc.
Wh-detection	0.00	0.94	0.42	0.91
Completeness	0.96	0.98	0.65	0.83
Fronting	0.00	0.68	0.30	0.81
Inversion	0.79	0.90	0.22	0.54
Embedding	0.67	0.96	0.77	0.98
Formulaic	0.51	0.89	0.42	0.81

Table 2: Agreement between manual annotation and two automatic annotation methods across syntactic dimensions. Acc. = accuracy.

(7) TD child: I can wake her up when it morning time.
(Ambrose 01dm_36, age = 3;0)

Formulaicity shows substantially lower agreement and accuracy. This pattern reflects the limited effectiveness of pattern-based detection for identifying formulaic expressions, which are often highly variable, context-dependent, and only partially characterized by surface templates. As a result, formulaicity is less amenable to strict binary classification based on fixed patterns and instead requires more explicit operational definitions as well as complementary qualitative validation.

Consistent with previous work showing strong performance of rule-based and supervised models on formally defined linguistic features (Lammert et al., 2025), GPT does not exhibit uniformly high agreement across dimensions in our data, but instead shows a strongly dimension-dependent profile.

Agreement is relatively high for dimensions involving overt clausal structure and global syntactic properties. GPT achieves good agreement for embedding (κ = 0.770, accuracy = 0.975) and moderate agreement for syntactic completeness (κ

= 0.648, accuracy = 0.830). *Wh*-detection also shows moderate agreement (κ = 0.424, accuracy = 0.910), indicating sensitivity to interrogative intent expressed through surface and contextual cues. For example, GPT correctly annotates Example (8) as a non-*wh*-question (0), whereas the dependency-parsing-based system misclassifies it as a *wh*-question rather than a relative clause.

(8) TD child: wha(t) I don't know.
(Nicholas nh48m_bronson, age = 3;11)

In contrast, GPT performs substantially worse on diagnostics that require precise identification of local word order and auxiliary position. Agreement is low for inversion (κ = 0.218, accuracy = 0.540), and Cohen's κ is not reported for fronting due to extreme class imbalance.

Finally, we observe small but non-zero differences across repeated runs using the same GPT prompt, reflecting inherent stochasticity in generation rather than instability of the annotation procedure (Achiam et al., 2023; Holtzman et al., 2019; Dodge et al., 2020). The use of generative models for clinical data and child language corpora therefore requires particular caution, given the sensitivity of these populations and the potential consequences of annotation error.

Overall, the results do not support a global assessment of GPT annotation quality. Rather, they point to a selective profile in which GPT is reliable for certain structurally recoverable dimensions but less suitable for formally explicit syntactic diagnostics. This contrast underscores a complementary division of labor: rule-based and parser-driven methods remain better suited for precise syntactic analysis, while GPT-based annotation offers a useful exploratory perspective on global structural properties in spontaneous child language.

6 Conclusion

This study demonstrates that children with hearing loss (CHL) exhibit both quantitative and qualitative delays in *wh*-question acquisition relative to typically developing (TD) peers. CHL produced fewer *wh*-questions, showed reduced *wh*-type diversity, and experienced persistent difficulties with structurally complex interrogatives, particularly embedding and canonical *wh*-fronting. These findings align with input-sensitive and usage-based accounts that view complex syntax as especially vulnerable under conditions of reduced language access, underscoring the importance of early, enriched syntactic input for supporting grammatical development in CHL. Importantly, these group differences persist when controlling for MLU, indicating selective vulnerability of *wh*-question development beyond general language delay.

Methodologically, by combining dependency-parsing-based NLP analyses with an exploratory GPT-based evaluation, this study highlights both the promise and the limitations of automated approaches to spontaneous child language. While such methods support scalable analysis, they require careful validation and theoretical grounding, especially for fine-grained syntactic phenomena. Overall, the results emphasize the value of integrating corpus-based methods, clinical perspectives, and computational tools in the study of language development under atypical input conditions.

Limitations

This study has several limitations. First, the analysis is based on two English CHILDES clinical corpora. While these datasets provide rich longitudinal data, they may not capture the full variability of children with hearing loss across different linguistic, educational, and socioeconomic backgrounds. The findings should therefore be interpreted with caution beyond similar populations. Second, the identification of syntactic complexity relies on operational definitions that simplify naturally variable child speech. For example, clausal embedding was approximated using a fixed set of embedding-triggering verbs, which may miss less canonical or emerging structures in early production. Third, dependency parsing and rule-based heuristics are challenged by the non-canonical nature of spontaneous child speech, including fragments, repairs, and ellipsis. Although validation suggests that such errors are limited, some misclassifications are unavoidable. Finally, the GPT-based analysis is exploratory and was conducted without task-specific training. Its performance should not be interpreted as a fully reliable annotation standard, but rather as an initial probe into the potential and limits of large language models for child language data.

Ethical considerations

This study analyzes previously collected and publicly available child language data from the CHILDES database. All data were anonymized at the source, and no new data collection or interac-

tion with human participants was involved.

Acknowledgements

We thank Carole Tiberius for her helpful comments and feedback on this paper. We thank the anonymous reviewers for their careful reading of the manuscript and for their insightful and constructive comments. Their feedback has been invaluable in helping us clarify the analysis, improve the presentation, and strengthen the overall quality of this paper.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

David Adger. 2003. *Core syntax: A minimalist approach*. Oxford University Press.

Ben Ambridge and Elena VM Lieven. 2011. *Child language acquisition: Contrasting theoretical approaches*. Cambridge University Press.

Sophie E Ambrose. 2016. Gesture use in 14-month-old toddlers with hearing loss and their mothers' responses. *American journal of speech-language pathology*, 25(4):519–531.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Harald Baayen. 2008. *Analyzing Linguistic Data: A practical introduction to statistics using R*.

Lois Bloom, Susan Merkin, and Janet Wootten. 1982. Wh-questions: Linguistic factors that contribute to the sequence of acquisition. *Child development*, pages 1084–1092.

Roger Brown. 1973. A first language: The early stages.

Marie Coppola and Kristin Walker. 2025. Early language access and steam education: Keys to optimal outcomes for deaf and hard of hearing students. *Education Sciences*, 15(7):915.

Jill De Villiers. 2007. The interface of language and theory of mind. *Lingua*, 117(11):1858–1878.

Holger Diessel. 2004. *The acquisition of complex sentences*, volume 105. Cambridge University Press.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Susan Ervin-Tripp. 1970. Discourse agreement: How children answer questions. *Cognition and the development of language*.

Alison Ferguson, Hugh Craig, and Elizabeth Spencer. 2009. Exploring the potential for corpus-based research in speech-language pathology. In *Proceedings of the HCSNet Workshop on Designing the Australian National Corpus. Somerville, MA: Cascadilla Proceedings*, pages 30–36.

Peter Flipsen and Kathleen Kangas. 2014. Mean length of utterance (mlu) in children with cochlear implants. *Volta Review*, 114:135–155.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Corina Goodwin, Emily Carrigan, Kristin Walker, and Marie Coppola. 2022. Language not auditory experience is related to parent-reported executive functioning in preschool-aged deaf and hard-of-hearing children. *Child Development*, 93(1):209–224.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Paul Ibbotson. 2013. The scope of usage-based theory. *Frontiers in psychology*, 4:255.

Ray Jackendoff. 2003. Précis of foundations of language: Brain, meaning, grammar, evolution. *Behavioral and brain sciences*, 26(6):651–665.

Wonkyung Jang, Diane Horm, Kyong-Ah Kwon, Kun Lu, Ryan Kasak, and Ji Hwan Park. 2025. Leveraging natural language processing to deepen understanding of parent–child interaction processes and language development. *Family Relations*, 74(3):1146–1173.

Keith Johnson. 2011. *Quantitative methods in linguistics*. John Wiley & Sons.

Sharon Klieve, Patricia Eadie, Lorraine Graham, and Suze Leitão. 2023. Complex language use in children with hearing loss: A scoping review. *Journal of Speech, Language, and Hearing Research*, 66(2):688–719.

Keegan M Koehlinger, Amanda J Owen Van Horne, and Mary Pat Moeller. 2013. Grammatical outcomes of 3-and 6-year-old children who are hard of hearing. *Journal of Speech, Language, and Hearing Research*, 56(5):1701–1714.

Jessica M Lammert, Angela C Roberts, Ken McRae, Laura J Batterink, and Blake E Butler. 2025. Early identification of language disorders using natural language processing and machine learning: Challenges and emerging approaches. *Journal of Speech, Language, and Hearing Research*, 68(2):705–718.

Laurence B Leonard. 2017. *Children with specific language impairment*. MIT press.

Judith EC Lieu, Margaret Kenna, Samantha Anne, and Lisa Davidson. 2020. Hearing loss in children: a review. *Jama*, 324(21):2195–2205.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.

Brian MacWhinney. 2014. *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press.

Jason Merchant. 2001. *The syntax of silence: Sluicing, islands, and the theory of ellipsis*. Oxford University Press.

Mary Pat Moeller and J Bruce Tomblin. 2015. An introduction to the outcomes of children with hearing loss study. *Ear and hearing*, 36:4S–13S.

Johanna G Nicholas and Ann E Geers. 1997. Communication of oral deaf and normally hearing children at 36 months of age. *Journal of Speech, Language, and Hearing Research*, 40(6):1314–1327.

OpenAI. 2024. Gpt-4o mini. https://platform.openai.com/docs/models/gpt-4o-mini. Accessed 2025.

Andrew Radford. 2004. *English syntax: An introduction*. Cambridge University Press.

Mabel L Rice. 2004. Growth models of developmental language disorders. In *Developmental language disorders*, pages 214–247. Psychology Press.

Mabel L Rice, Sean M Redmond, and Lesa Hoffman. 2006. Mean length of utterance in children with specific language impairment and in younger control children shows concurrent validity and stable and parallel growth trajectories.

Tom Roeper and Jill De Villiers. 2011. The acquisition path for wh-questions. In *Handbook of generative approaches to language acquisition*, pages 189–246. Springer.

John Robert Ross. 1967. Constraints on variables in syntax.

Caroline F Rowland, Julian M Pine, Elena VM Lieven, and Anna L Theakston. 2003. Determinants of acquisition order in wh-questions: Re-evaluating the role of caregiver speech. *Journal of child language*, 30(3):609–635.

Caroline F Rowland, Julian M Pine, Elena VM Lieven, and Anna L Theakston. 2005. The incidence of error in young children's wh-questions.

Brenda Schick, Peter De Villiers, Jill De Villiers, and Robert Hoffmeister. 2007. Language and theory of mind: A study of deaf children. *Child development*, 78(2):376–396.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, and 1 others. 2024. The prompt report: a systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*.

Rosalind Jean Thornton. 1990. *Adventures in long-distance moving: The acquisition of complex wh-questions*. University of Connecticut.

Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Cambridge, MA.

Dorothy Tyack and David Ingram. 1977. Children's production and comprehension of questions. *Journal of child language*, 4(2):211–224.

Francesca Volpato and Mirta Vernice. 2014. The production of relative clauses by italian cochlear-implanted and hearing children. *Lingua*, 139:39–67.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Krystal L Werfel. 2017. Emergent literacy skills in preschool children with hearing loss who use spoken language: Initial findings from the early language and literacy acquisition (ella) study. *Language, Speech, and Hearing Services in Schools*, 48(4):249–259.

Krystal L Werfel, Gabriella Reynolds, Sarah Hudgins, Marissa Castaldo, and Emily A Lund. 2021. The production of complex syntax in spontaneous language by 4-year-old children with hearing loss. *American Journal of Speech-Language Pathology*, 30(2):609–621.

A Formulaic *wh*-Patterns

To operationalize formulaic *wh*-expressions in child speech, we defined a set of recurrent fixed or semi-fixed patterns motivated by classic acquisition research (e.g., Brown, 1973) and corpus-based analyses of *wh*-questions (Rowland et al., 2005). These patterns were treated as item-based constructions rather than fully productive *wh*-dependencies.

All formulaic patterns were additionally restricted to utterances of four tokens or fewer, ensuring that longer and syntactically elaborated *wh*-questions were not misclassified as formulaic.

Category	Representative examples
Copula-based frames	*what's that*, *where's it*, *who's that*
Bare *wh* + NP / pronoun	*what that*, *where you*, *who that*
Bare / discourse-like *why*	*why?*, *why not*
Phonologically reduced forms	*whazzat*, *wassat*, *what dat*
Contracted auxiliary frames	*what'd you*, *where'd you*, *why's he*

Table 3: Categories of formulaic *wh*-patterns used for classification.

B Full Prompt Template for GPT-based Annotation

This appendix provides the full prompt used for the GPT-based annotation of *wh*-questions in spontaneous child speech. The prompt was applied uniformly to all utterances in the LLM-based evaluation and was designed to be linguistically explicit while remaining surface-form based. Figure 3 in the main text presents a representative excerpt of this prompt.

System prompt.

> You are an expert child language researcher analyzing spontaneous speech from young children (12–54 months).
>
> Task: Annotate 6 dimensions for each wh-utterance. Judge based on SURFACE FORM only – do NOT correct or normalize child errors.

DIMENSION 0: IS_WH_QUESTION

Determine whether the utterance functions as a wh-question.

- **1 (Yes)**: The utterance contains a wh-word (what, who, where, when, why, how, which) and functions interrogatively, e.g., *What that?*, *Where mommy go?*, *Why not?*
- **0 (No)**:
  - Wh-word used as a relative pronoun: *the man who came*
  - Wh-word in a declarative clause: *I know what it is*
  - No wh-word present

DIMENSION 1: COMPLETENESS

Classify syntactic completeness based on surface form.

- **complete**: Overt subject and verb present, e.g., *What is that?*, *Where did you go?*, *Who came?*
- **missing_subject**: Verb present, no overt subject, e.g., *What doing?*, *Where going?*
- **missing_verb**: Subject or nominal present, no verb, e.g., *What that?*, *Where mommy?*
- **fragment**: Wh-word alone or minimal fragment, e.g., *What?*, *Why?*

DIMENSION 2: FRONTING

Is the wh-word clause-initial?

- **1 (Fronted)**: *What is that?*, *Where you going?*
- **0 (Not fronted)**: *You want what?*, *Mommy go where?*
- **NA**: Fragments with no clausal structure, e.g., *What?*

DIMENSION 3: INVERSION

Is subject–auxiliary inversion present in a matrix wh-question?

- **1 (Inversion present)**: Auxiliary precedes the subject, e.g., *What is that?*, *Where did you go?*
- **0 (No inversion)**: No auxiliary or subject precedes auxiliary, e.g., *What you doing?*, *Where mommy is?*
- **NA (Not applicable)**:
  - Subject wh-questions: *Who came?*, *What happened?*
  - Embedded wh-clauses: *I know what it is*
  - Fragments: *What?*
  - Copula-less forms: *What that?*

Critical exclusions. Inversion was evaluated only for complete, matrix, non-subject wh-questions. Subject wh-questions, embedded wh-clauses, and fragments were excluded.

DIMENSION 4: EMBEDDING

Is the wh-clause embedded under a matrix verb (e.g., *know, think, ask, tell, wonder*) and realized as a complete clausal complement?

- **1 (Embedded)**: *I know what you said*, *She asked who came*
- **0 (Not embedded)**:
  - Formulaic wh-templates (short, routinized frames)
  - Discourse markers: *You know what?*
  - Main clause wh-questions
  - Fragments or truncated forms

DIMENSION 5: FORMULAIC

Is the utterance a short, routinized wh-expression (4 words) learned as an unanalyzed chunk (Rowland et al., 2005)?

- **1 (Formulaic)**: *What's that*, *Where's it*, *Why not*, contracted auxiliary frames

- **0 (Productive)**: Longer or structurally elaborated wh-questions

Output format.

Return a JSON object only (no additional text) with the following fields:

```
{
  "is_wh_question": 0 | 1,
  "completeness": "complete" |
  "fragment" | "missing_verb" |
  "missing_subject",
  "formulaic": 0 | 1,
  "inversion": 0 | 1 | "NA",
  "embedding": 0 | 1,
  "fronting": 0 | 1 | "NA"
}
```

IMPORTANT REMINDERS.

- Judge based on SURFACE FORM only
- Do NOT correct or normalize child errors
- Child speech is often fragmentary
- When uncertain, be conservative (0 or "NA")
- Formulaic expressions are never coded as embedded

C Annotation Guidelines for Embedded Wh-Questions

This appendix provides detailed annotation guidelines for the *embedding* dimension, with particular attention to ambiguous cases involving truncated or elliptical structures.

C.1 Definition of Embedded Wh-Questions

An embedded wh-question is defined as a wh-clause that serves as the complement of a matrix verb. Following standard criteria in child language acquisition research (De Villiers, 2007; Rowland et al., 2005), we require the presence of three components:

1. A matrix verb (e.g., *know, think, wonder, ask, tell, see*)
2. A wh-word (e.g., *what, where, who, why, how*)
3. Additional clausal content following the *wh*-word

Matrix verbs. Embedded *wh*-questions were identified as *wh*-clauses occurring as complements of a predefined set of matrix verbs: *think, know, wonder, say, believe, tell, ask, see, understand, remember, forget, show*.

C.2 Treatment of Sluicing and Truncated Forms

A critical annotation decision concerns utterances with sluicing or truncation, such as *I don't know how*.

From a syntactic-theoretic perspective, such structures can be analyzed as involving ellipsis of the TP under identity (Ross, 1967; Merchant, 2001). However, following conventions in child language acquisition research, we adopt a more conservative criterion and classify such truncated forms as **non-embedded** (coded as 0). This decision is motivated by several considerations:

1. **Productive knowledge criterion**: The presence of a complete wh-clause provides stronger evidence that the child has productive knowledge of embedded structures, rather than having acquired a formulaic chunk (Rowland et al., 2005).
2. **Methodological consistency**: Studies examining the development of complex syntax in children typically require overt realization of the embedded clause (De Villiers, 2007; Tyack and Ingram, 1977).
3. **Conservative estimation**: This approach may underestimate children's syntactic competence but avoids overattribution of complex grammatical knowledge.

D CLAN vs. NLP/LLM Treatment of Repair Annotations

In addition to model-internal annotation variability, a manual comparison revealed a single-utterance discrepancy between CLAN-based extraction and our NLP/LLM pipeline. Specifically, CLAN excluded an utterance containing the *wh*-form *what* due to a replacement annotation (*who's*), whereas our pipeline retained the utterance based on the presence of an overt *wh*-form on the child tier (CHI).

This discrepancy reflects a principled difference in annotation policy rather than an error. While CLAN prioritizes the repaired target form, our NLP/LLM-based approach operates strictly on the child's surface production, preserving non-canonical and self-repaired *wh*-forms that are developmentally informative in spontaneous speech. Importantly, this discrepancy is isolated and does

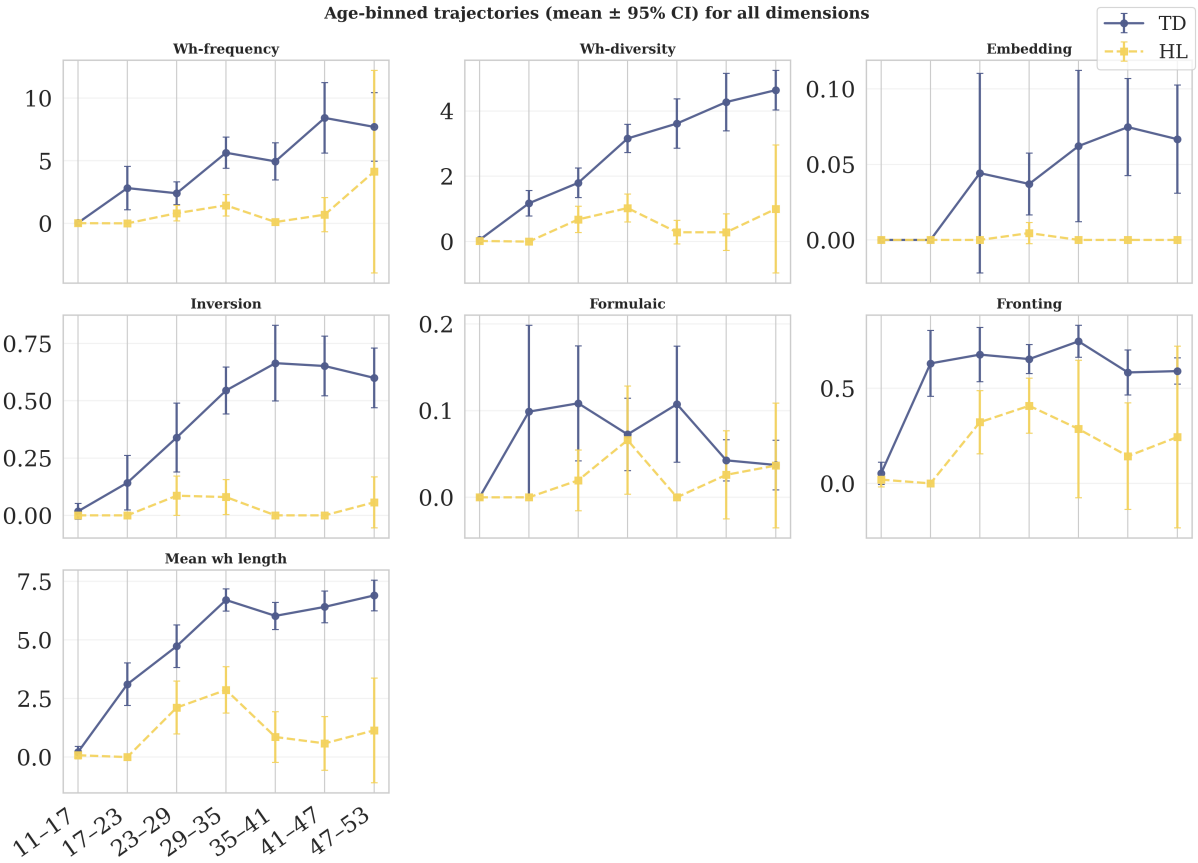

Figure 7: Age-binned developmental trajectories (mean ± 95% CI) for all *wh*-question dimensions, shown separately for typically developing (TD) children and children with hearing loss (HL). Age bins correspond to those used in the regression analyses reported in Section 4.3.

not affect the overall quantitative patterns reported in the main analysis.

(9) what [: who's] that boy ?
(Ambrose 02JW_27)

E Chronological age-binned Developmental Trajectories

This appendix provides age-binned developmental trajectories for all *wh*-question dimensions examined in the study, shown separately for TD children and CHL. The figure reports mean values with 95% confidence intervals for each age bin.

Age bins correspond to those used in the regression analyses reported in Section 4.3 (11–17, 17–23, 23–29, 29–35, 35–41, 41–47, and 47–53 months). The dimensions visualized include normalized *wh*-question frequency, *wh*-type diversity, embedding, inversion, formulaicity, fronting, and mean *wh*-question length.

This figure is intended to provide a comprehensive visualization of developmental patterns across dimensions. Statistical inference in the main text is based on regression models with age, group, and their interaction as predictors, rather than on visual inspection of these trajectories.