The OSCAR PUBLISHING Services
American Journal of Philological Sciences

VOLUME Vol.06 Issue03 2026
PAGE NO. 250-259
DOI 10.37547/ajps/Volume06Issue03-55

Methodological Foundations for Modeling A Longitudinal Corpus of Uzbek Child Language Within the Childes/Talkbank Infrastructure and Multi-Tier Linguistic Annotation

Zarnigor Ubaydullayeva
PhD Researcher (DSc Track), Department of Linguistics, Fergana State University, Fergana, Uzbekistan

Received: 02 February 2026; **Accepted:** 06 March 2026; **Published:** 31 March 2026

Abstract: This article establishes a comprehensive theoretical and methodological framework for designing, constructing, and curating a longitudinal corpus of spontaneous Uzbek child speech within the international CHILDES/TalkBank infrastructure, supported by multi-tier linguistic annotation. Although empirical inquiries into Uzbek developmental linguistics have addressed sociolinguistic, lexical-semantic, pragmatic, and preliminary audio-corpus dimensions, the global CHILDES repository as of 2026 lacks a dedicated, standardized longitudinal dataset for Uzbek. The primary objective of this study is to substantiate the typological, computational, and developmental necessity of creating such a resource and to delineate its core architectural and methodological principles.

Employing a conceptual-analytical approach, this study investigates the operationalization of the CHILDES/TalkBank ecosystem, the CHAT (Codes for the Human Analysis of Talk) transcription standard, CLAN (Computerized Language ANalysis) computational modules, Universal Dependencies (UD)-compliant morphosyntactic representation, and FAIR (Findable, Accessible, Interoperable, Reusable) open-data management principles as applied to spontaneous Uzbek child language. A comparative analysis incorporates the Russian longitudinal RusLan-M corpus, the Kazakh Balausa1 corpus, Turkish CHILDES historical datasets, and the modern Uzbek Universal Dependencies treebank.

The analysis formulates five indispensable pillars for constructing the corpus: (1) systematic repeated sampling over extended developmental windows; (2) high-fidelity naturalistic audiovisual recording; (3) standardized CHAT-compliant transcription; (4) a hybrid morphosyntactic annotation scheme adapted to the agglutinative structure of Uzbek; and (5) rigorous research ethics, bioethical anonymization protocols, and comprehensive metadata profiling. Establishing this standardized longitudinal corpus provides the indispensable empirical infrastructure required to systematically track individual developmental trajectories, the acquisition of inflectional categories, and cross-linguistic developmental dynamics, thereby integrating Uzbek developmental linguistics into global corpus-based scholarship.

Keywords: Child language acquisition, language ontogeny, longitudinal corpus, CHILDES, TalkBank, CHAT format, CLAN tools, linguistic annotation, Universal Dependencies, agglutinative morphology, FAIR principles, corpus linguistics.

Introduction: Investigating child language acquisition constitutes a fundamental paradigm for uncovering both cross-linguistic universals and language-specific mechanisms that govern human language ontogeny. Traditional investigations in developmental linguistics have historically relied on single-subject diary studies,

small-scale cross-sectional samples, or short-term synchronous comparisons across distinct age brackets. While these methodologies provide descriptive insights into isolated lexical or syntactic phenomena, they inherently obscure the fine-grained, dynamic trajectory of child linguistic development over extended periods. They fall short of identifying the precise onset, regression phases (such as U-shaped developmental curves), stabilization, and eventual productivity of morphosyntactic markers, as well as the immediate impact of child-directed environmental speech (input) [Berman and Slobin 1994; MacWhinney 2000]. Overcoming these empirical constraints necessitates a systematic longitudinal corpus design centered on continuous, dense observations of the same subjects over time.

The rapid evolution of corpus linguistics and open-access research infrastructures has transformed child language research from private, non-replicable personal archives into open, machine-readable, verifiable, and computationally exploitable datasets. Within this domain, the Child Language Data Exchange System (CHILDES), operating as a core component of the broader TalkBank infrastructure, serves as the international gold standard. It unifies high-resolution audiovisual recordings, synchronized transcripts, hierarchical metadata, and automated computational tools within a single interoperable environment. The integration of the standardized CHAT (Codes for the Human Analysis of Talk) transcription format and the specialized CLAN (Computerized Language ANalysis) software has substantially enhanced cross-linguistic comparability and methodological rigor [MacWhinney 2019; MacWhinney and Fromm 2022]. At present, CHILDES encompasses more than 50,000 transcripts, extensive cross-sectional data sets, and dozens of high-density longitudinal case studies across dozens of world languages [MacWhinney and Fromm 2022].

The inclusion of diverse linguistic typologies within CHILDES is essential to safeguard developmental linguistic theory from over-generalizing empirical models that are disproportionately grounded in isolating or fusional Indo-European languages (predominantly English). For instance, empirical research on agglutinative Turkish child language by Aksu-Koç and Slobin demonstrated that the linear, transparent concatenation of morphemes fosters early and distinct acquisition pathways for case, possession, and verb inflection paradigms compared to fusional languages [Aksu-Koç and Slobin 1985]. Furthermore, dense corpus-based usage-based studies have revealed that early multi-word utterances predominantly represent low-scope structural variants of previously acquired schemas rather than abstract syntactic transformations [Lieven, Salomo, and Tomasello 2009]. Within the Turkic family, the recently developed Kazakh Balausa1 corpus demonstrates that agglutinative languages can be effectively adapted to the CHILDES framework through time-aligned media and multi-tiered morphosyntactic annotation compatible with the Universal Dependencies scheme [TalkBank 2026b].

In Uzbek linguistics, child language has garnered scholarly attention through foundational socio-psycholinguistic investigations into child vocabulary [Qurbonova 2009], subsequent inquiries into spontaneous communicative pragmatics, and emergent proposals concerning the development of spoken audio corpora and initial CHILDES-based applications [Masharipova and Qutlimuratova 2024; Abduraxmonova 2025]. In addition, the adaptation of the Multilingual Assessment Instrument for Narratives (MAIN) to Uzbek underscores an evolving institutional competence in standardized, assessment-driven developmental methodologies [Abdumazhitova and Amanov 2020; Gagarina et al. 2019]. Despite these advancements, existing national initiatives do not fulfill the requirements of an open, dense, time-aligned, multi-layered longitudinal corpus capable of supporting cross-linguistic, computational comparative analysis.

Crucially, the global CHILDES index does not currently list an official, standardized repository for the Uzbek language [TalkBank 2026a]. This absence signifies a substantial typological gap: Uzbek exhibits an exceptionally rich, regular agglutinative morphology characterized by strict SOV constituent ordering, vowel harmony loss (distinguishing it from several Turkic counterparts), and an extensive inventory of bound morphemes representing case, person, number, aspect, mood, voice, and modality. Investigating these features within a dedicated longitudinal corpus is essential for defining both common Turkic and language-specific developmental pathways.

Consequently, the primary objective of this article is to substantiate the theoretical, typological, computational, and methodological imperatives for developing an Uzbek child speech longitudinal corpus within the CHILDES/TalkBank infrastructure, and to delineate its fundamental architectural principles.

To realize this objective, the study addresses the following research tasks:

1. Substantiate the developmental and empirical advantages of a high-density longitudinal corpus design over traditional cross-sectional methodologies for tracking morphological and syntactic ontogeny;

2. Analyze the current landscape of Uzbek child language studies to pinpoint existing methodological gaps relative to international corpus standardization protocols;

3. Formulate the linguistic guidelines for CHAT-compliant transcription of spontaneous Uzbek child-caregiver interactions and specify morphosyntactic annotation strategies aligned with Universal Dependencies (UD);

4. Systematize metadata architecture, bioethical anonymization protocols, research ethics standards, and open-science data curation in compliance with the FAIR principles.

Scientific Novelty: This study provides the first end-to-end methodological blueprint for capturing naturalistic, spontaneous Uzbek child speech via dense audiovisual recordings, transcribing the output using standardized CHAT coding, and integrating multi-tier morphosyntactic annotation tailored specifically to the structural characteristics of an agglutinative Turkic language.

METHODS

This investigation is conceptual, architectural, and analytical in design. The material base comprises international methodological handbooks, structural corpus descriptions from CHILDES/TalkBank published between 2000 and 2026, technical documentation on Universal Dependencies annotation schemes, and empirical literature examining Uzbek language structure and child language acquisition in Turkic languages.

The literature and technical specifications were curated according to three rigorous criteria:

1. Spontaneity and Dynamism: Direct relevance to naturalistic, spontaneous child-caregiver verbal interactions and continuous longitudinal tracking;

2. Standardization and Computability: Strict adherence to machine-readable transcription standards, inter-annotator validation protocols, and multi-tier morphological parsing;

3. Typological Relevance: Compatibility with the structural properties of agglutinative morphology and computational frameworks suited to under-resourced or low-resource languages.

The analytical methodology combines comparative typological analysis, corpus-analytic modeling, content analysis, and architectural systems design. Three key analogical resources within CHILDES served as methodological benchmarks: The Russian RusLan-M corpus was analyzed to evaluate longitudinal dense sampling protocols, parent-recorded video integration, and advanced multi-level privacy anonymization strategies [Lelik et al. 2025]. The Kazakh Balausa1 corpus provided a direct typological model for integrating audio-text time-alignment with Universal Dependencies-based morphological parsing for an agglutinative Turkic language [TalkBank 2026b]. The Aksu Turkish dataset provided historical and cross-linguistic reference points for the emergence of inflectional affixes in Turkic ontogeny [Aksu-Koç and Slobin 1985].

In addition, computational viability was assessed by referencing the first complete Universal Dependencies treebank for modern Uzbek (comprising 500 sentences and 5,850 tokens), published by Akhundjanova and Talamo, which provides the syntactic and morphological baseline for formal computational representations [Akhundjanova and Talamo 2025].

The scope of this article is intentionally architectural: it does not report finalized empirical findings from an assembled corpus, but defines the conceptual framework, resolves foundational methodological challenges, and sets forth technical standards prior to full-scale data acquisition.

RESULTS AND ANALYSIS

1. Theoretical and Empirical Superiority of Longitudinal Corpus Design in Language Ontogeny

Determining the precise mechanisms of language

acquisition relies critically on study design. Cross-sectional sampling captures snapshot observations from disparate age cohorts, identifying aggregate developmental averages; however, it obscures within-subject variability, transitional developmental phases, and the nonlinear nature of acquisition. In contrast, a dense longitudinal design follows the same child (or cohort) continuously across frequent, closely spaced intervals. This enables precise identification of the emergence (onset), productive expansion, temporary overgeneralization (e.g., U-shaped development), and formal stabilization of grammatical categories [Aksu-Koç and Slobin 1985; Boeschoten 2021].

In an agglutinative language such as Uzbek, where lexical roots combine with long strings of unambiguous suffixes expressing diverse grammatical functions, a child's mastery of grammar does not progress via holistic phrase memorization alone, but via the progressive morphological decomposition and productive recombining of affixes. A longitudinal design is the only methodology capable of verifying whether a multi-affix token represents an unanalyzed formulaic chunk or a productive syntactic operation.

A defining strength of the CHILDES framework is its capacity to couple high-density longitudinal data with standardized indexing. In the RusLan-M corpus, two monolingual Russian children were observed longitudinally from age 0;10 to 3;10, generating tens of hours of spontaneous speech linked to time-aligned, anonymized audiovisual records [Lelik et al. 2025]. This design highlights the feasibility of mapping fine-grained developmental trajectories for morphologically complex systems.

Moreover, longitudinal corpora enable concurrent analysis of the child's ambient linguistic environment (input). Newman, Rowe, and Bernstein Ratner empirically established that both the qualitative diversity and quantitative density of maternal input during early infancy directly predict vocabulary size, syntactic sophistication, and cognitive processing speed in later developmental stages [Newman, Rowe, and Bernstein Ratner 2016]. Thus, an Uzbek longitudinal corpus must capture the complete conversational dyad, documenting not only child productions but also child-directed adult speech.

2. Structural Analysis of the Empirical Research Gap in Uzbek Ontolingual Resources

While studies on child language in Uzbek linguistics provide valuable foundations, their scope remains constrained by methodological fragmentation: Existing datasets are predominantly cross-sectional, collected during sporadic, short sessions that lack systematic re-sampling of the same subjects over time; Data collection procedures lack uniformity in audio-visual fidelity, transcription conventions, and contextual documentation; Materials remain isolated in private institutional archives, rendering them inaccessible for independent replication, algorithmic verification, or cross-linguistic pooling.

From an international perspective, although CHILDES contains corpora for dozens of non-Western and typologically diverse languages (such as Inuktitut, Sesotho, Georgian, and Hebrew), the absence of Uzbek is a glaring omission [TalkBank 2026a]. As a major Central Asian Turkic language with over 35 million speakers, Uzbek exhibits unique phonological, morphological, and syntactic characteristics—such as the loss of vocalic harmony in standard dialects, complex auxiliary verb constructions, and diverse converb systems—that provide vital data for testing universal versus typology-specific acquisition models.

Table 1. Comparative matrix of relevant CHILDES corpora and architectural takeaways for the Uzbek corpus

Corpus Resource	Typology & Design	Media Modality	Annotation & Markup Scheme	Core Methodological Takeaway for Uzbek
RusLan-M	Russian (Fusional); Dense Longitudinal; 2 children	Synchronized high-definition video & audio	CHAT/CHILDES; comprehensive media de-identification	Protocol for long-term dense recording; media-text alignment; strict child protection and

				bioethics
Balausa1	Kazakh (Agglutinative); Cross-sectional; 8 children	Audio recordings (~5 hours)	%mor tier: Batchalign2 integrated with UD; expert verification	Feasibility of linking agglutinative Turkic morphology to Universal Dependencies with manual expert validation
Aksu Turkish	Turkish (Agglutinative); Naturalistic interaction	Archival audio-recorded sessions	Standard historical CHILDES transcripts	Comparative reference for verifying developmental milestones across Turkic nominal and verbal inflections
Proposed Uzbek Corpus	Uzbek (Agglutinative); High-density longitudinal	Synchronized digital video and high-fidelity audio	CHAT + hybrid %mor/%gra/UD tiers + deeply verified gold standard	Systematic tracking of inflectional ontogeny and establishment of an open, interoperable national digital resource

Source: Synthesized by the author based on Lelik et al. (2025), TalkBank (2026a, 2026b), and Aksu-Koç & Slobin (1985).

3. CHAT Standard: Transforming Raw Conversational Data into Structured Linguistic Assets

Unprocessed multimedia files do not constitute a corpus until they are systematically segmented, transcribed, time-linked, and enriched with structured metadata. The CHAT format provides the standard markup syntax for all TalkBank repositories. Its architecture separates the transcript into three coordinated planes: Header lines (recording session metadata), the Main transcription tier (*CHI:, *MOT:, etc.), and linked Dependent tiers (%mor:, %gra:, %pho:, %com:) [MacWhinney 2000; TalkBank 2026c].

Transcribing spontaneous child speech requires departing from strict literary orthographic norms. Young children systematically produce sound substitutions, root truncations, affix omissions, idiosyncratic phonological forms, false starts, and overlapping vocalizations. If these spontaneous tokens are forced into conventional literary spelling, critical developmental phenomena are lost.

Consequently, an Uzbek CHAT Project Transcription Manual must be established to provide operational conventions for native dialectal variants, phonetic approximations ([: target]), unintelligible fragments (xxx), micropause demarcations ((.)), and contextual non-verbal actions. Standardizing these practices ensures that transcripts remain fully compatible with CLAN computational algorithms, enabling automated calculation of: Mean Length of Utterance (MLU in words and morphemes via MLU); Lexical diversity indices (Type-Token Ratio, VOCD); Absolute and relative frequency profiles (FREQ); Concordances and Key Word in Context evaluations (KWAL) [MacWhinney and Fromm 2022].

4. Morphosyntactic Annotation Architecture and Universal Dependencies Integration

The core challenge in building an Uzbek child language corpus lies in its agglutinative morphology. Grammatical functions are marked by sequentially chaining distinct suffixes to invariant stems. For example: Nominal chains: kitob-lar-imiz-dan (Book - PLUR - POSS.1PL - ABL; 'from our books'); Verbal chains: yoz-dir-ma-di-ngiz (Write - CAUS - NEG - PAST - 2PL;

'you did not make [someone] write').

Tracking the developmental order in which these morphemes emerge and transition from unanalyzed wholes to productive linguistic modules is a central empirical question. Evaluating child utterances based solely on word counts is inadequate; fine-grained morphological tagging is essential.

The Universal Dependencies (UD) framework provides a cross-linguistically consistent standard for word classes (UPOS), morphological features (Feats), and dependency relations (deprel) [de Marneffe et al. 2021]. In CHILDES, this is parallel to the %mor (lexical/morphological) and %gra (grammatical relations) tiers [MacWhinney 2012]. The construction of the first Uzbek Universal Dependencies treebank by Akhundjanova and Talamo (500 sentences, 5,850 tokens) established baseline parsing rules for written standard Uzbek, providing a valuable foundation for computational parsing [Akhundjanova and Talamo 2025].

However, parsers trained on formal adult written corpora fail when applied directly to spontaneous child language, which features incomplete clauses, omitted roots, and non-canonical word orders. To address this challenge, a two-stage hybrid annotation model is proposed:

1. Macro-Level Broad Corpus: All transcripts are converted to CHAT and processed using semi-automated morphological analyzers (such as adapted UD taggers and CLAN's MOR engine) to assign baseline lemmas and parts of speech.

2. Micro-Level Gold-Standard Subcorpus: A carefully selected, representative subset of transcripts across key developmental ages (e.g., 1;6, 2;0, 2;6, 3;0) is manually corrected and annotated by trained linguists. This tier includes fine-grained morpheme-by-morpheme glossing, error coding (affix omission, overgeneralization), and multimodal tags (gestural scaffolding). This gold-standard subcorpus provides an essential benchmark for evaluating and fine-tuning automated NLP tools on spoken Uzbek child language [Yang et al. 2025].

5. Metadata Architecture, Bioethics, and FAIR Open-Data Governance

A corpus is only as robust as its metadata. Within the CHAT environment, every recording session must be preceded by standardized header definitions [TalkBank 2026c]: @Participants defines standardized actor codes (CHI = target child, MOT = mother, FAT = father, INV = investigator); @ID encodes language, corpus name, participant code, chronological age, sex, socioeconomic profile, and target role; @Age of CHI specifies exact age in the standardized year;month.day format (e.g., 2;04.18), which is critical for tracking developmental milestones; @Situation and @Media detail communicative context (e.g., free play, mealtime, book reading) and multimedia technical specifications. The syntactic integrity of these headers is validated automatically using CLAN's CHECK command to prevent parsing errors.

Given that audiovisual recordings of children involve sensitive personal data, ethical governance must be integrated into the study design from the outset. In accordance with international institutional review standards, informed parental consent must be obtained prior to data collection. Data anonymization requires pseudonyms for all proper names, addresses, and identifying family references within transcripts. Media access should follow TalkBank's tiered access protocols: while text transcripts are made publicly available for scientific inquiry, raw video files containing identifiable features are restricted to password-protected academic repositories, with facial blurring applied where required [MacWhinney and Fromm 2022; Lelik et al. 2025].

This operational structure directly implements the FAIR Guiding Principles for scientific data management [Wilkinson et al. 2016]: Findable (cataloged in TalkBank, assigned persistent DOIs, and indexed with rich metadata); Accessible (openly retrievable under defined academic licenses); Interoperable (formatted in open, non-proprietary standards like CHAT, XML, CoNLL-U); Reusable (supported by rigorous documentation, clear sampling protocols, and explicit annotation manuals).

6. Systematic Blueprint for the Proposed Longitudinal Uzbek Child Speech Corpus

Table 2. Methodological framework of the proposed Uzbek CHILDES/TalkBank longitudinal corpus

Structural Domain	Core Methodological	Operational	Target Scientific

	Requirement	Implementation Mechanism	Outcome
Sampling Design	Dense, continuous observation of identical subjects	Bi-weekly or weekly 45-minute spontaneous recordings across ages 1;0 to 4;0	Uninterrupted individual developmental curves; precise identification of affix emergence
Transcription Standard	Uniform, syntactically verifiable computational format	CHAT standard, automated CLAN CHECK validation, millisecond-level time-alignment	Full interoperability with cross-linguistic corpora; automated extraction of lexical and syntactic indices
Linguistic Annotation	Multi-tier tagging of agglutinative morphemes and syntax	Semi-automated %mor and %gra parsing, UD feature mapping, manual gold-standard verification	Empirical mapping of inflectional ontogeny and syntactic dependency structures
Metadata Structuring	Comprehensive profiling of communicative situations	Standardized CHAT headers (@Age, @Participants, @Situation, linguistic profile)	Quantitative modeling of language acquisition correlated with age, input density, and social factors
Ethical Protocols	Robust participant privacy and bioethical compliance	Informed parental consent, pseudonymization, tiered access models, media de-identification	Adherence to international ethical standards; safe open-science data sharing
Repository Integration	Publication within the global TalkBank repository	Preparation of complete documentation, pilot validation, and formal submission package	Official listing of the Uzbek language within the international CHILDES database

Source: Formulated by the author based on CHILDES/TalkBank research standards.

Establishing this monolingual baseline also provides an empirical foundation for future research on Uzbek-Russian bilingualism. Describing typical monolingual developmental trajectories is an essential prerequisite for accurately assessing cross-linguistic transfer, lexical borrowing, code-switching, and structural convergence in multilingual settings, as underscored by Nematova et al.'s findings on phonological awareness transfer in bilingual children [Nematova et al. 2026].

DISCUSSION

The imperative to establish an Uzbek longitudinal child speech corpus operates across three interrelated dimensions:

1. Typological and Theoretical Imperatives: Mainstream psycholinguistic models of language acquisition have been heavily shaped by isolating and fusional structures. Dense longitudinal data from Uzbek—an agglutinative language with transparent bound morphemes and strictly verb-final syntax—provides a crucial empirical testbed for verifying whether proposed developmental milestones represent language universals or structural artifacts of Indo-European languages. Expanding typological coverage directly refines cross-linguistic theories of early grammar acquisition [Bornstein and Hendricks 2012; Ibbotson et al. 2024].

2. Methodological Standardization: The prevailing limitation of earlier domestic studies is their isolated, non-standardized format. Field recordings collected for individual projects frequently become unusable over time due to missing metadata, non-standard transcription conventions, or a lack of persistent archiving. Adopting the CHILDES/TalkBank ecosystem

secures the entire data lifecycle, ensuring that empirical materials remain permanently accessible and computationally exploitable for the wider scientific community.

3. Computational and Natural Language Processing (NLP) Applications: Spoken child language represents a notoriously challenging domain for speech recognition (ASR), tokenization, and semantic parsing due to its high acoustic variability, structural ellipsis, and non-canonical forms. While existing Uzbek NLP resources are built primarily on formal written registers, a child language corpus tagged with Universal Dependencies provides the training data necessary to make language models robust to spontaneous speech, conversational fragmentation, and colloquial morphosyntactic variants [Akhundjanova and Talamo 2025; Yang et al. 2025].

Importantly, integration into CHILDES cannot be treated as an automatic administrative step. TalkBank maintains strict submission guidelines requiring peer review of sample files, syntax validation, and compliance with data anonymization standards. Research efforts must therefore focus on conducting rigorous pilot sessions, authoring a project-specific transcription manual, and completing automated syntax checks before submitting the final data package for formal archiving and DOI assignment.

CONCLUSION

Designing and constructing a longitudinal corpus of spontaneous Uzbek child language within the CHILDES/TalkBank infrastructure represents a vital milestone for Uzbek developmental linguistics and computational language resources. While prior national studies laid important groundwork, the discipline urgently requires an open, standardized, media-linked, and computationally verifiable longitudinal data repository.

The architectural blueprint proposed in this study synthesizes high-density audiovisual sampling, CHAT transcription standards, multi-tier morphosyntactic tagging adapted to agglutinative structures via Universal Dependencies, and open-science curation governed by the FAIR principles.

The realization of this corpus will:

1. Provide an empirical basis for tracking the acquisition of inflectional morphology, syntactic schemas, and lexical growth in Uzbek;

2. Integrate Uzbek into cross-linguistic child language research and comparative Turkic linguistics;

3. Establish normative developmental baselines to inform future inquiries into Uzbek-Russian bilingualism, diagnostic speech-language pathology, and developmental delays;

4. Enrich the computational infrastructure of Uzbek by providing gold-standard training data for speech recognition and language modeling of spontaneous speech.

Subsequent phases of this initiative will focus on launching pilot recordings with target children, finalizing the Uzbek CHAT transcription manual, and developing the expertly verified gold-standard subcorpus. Establishing this resource will ensure that Uzbek child language is formally represented within the international child language research community.

REFERENCES

1. Abduraxmonova, Sohiba Aktam qizi. 2025. “O‘zbek bolalar nutqining korpusga asoslangan psixolingvistik tadqiqi: CHILDES tizimi asosida yondashuv.” Contemporary Technologies of Computational Linguistics – CTCL.2025, 538–540.
2. Abdumazhitova, Chaman, and Amirkul Amanov, trans. and adapt. 2020. MAIN: Multilingual Assessment Instrument for Narratives – Uzbek Version. ZAS Papers in Linguistics 64.
3. Akhundjanova, Arofat, and Luigi Talamo. 2025. “Universal Dependencies Treebank for Uzbek.” In Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025), 1–6.
4. Aksu-Koç, Ayhan, and Dan I. Slobin. 1985. “The Acquisition of Turkish.” In The Cross-Linguistic Study of Language Acquisition, Vol. 1: The Data, edited by Dan I. Slobin, 839-878. Hillsdale, NJ: Lawrence Erlbaum Associates.
5. Berman, Ruth A., and Dan I. Slobin, eds. 1994. Relating Events in Narrative: A Crosslinguistic Developmental Study. Hillsdale, NJ: Lawrence Erlbaum Associates.
6. Boeschoten, Hendrik. 2021. “Uzbek.” In The Turkic Languages, 2nd ed., edited by Lars Johanson and

Éva Á. Csató, 388-408. London and New York: Routledge.

7. Bornstein, Marc H., and Charlene Hendricks. 2012. “Basic Language Comprehension and Production in >100,000 Young Children from Sixteen Developing Nations.” Journal of Child Language 39 (4): 899-918. https://doi.org/10.1017/S0305000911000407.
8. de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. “Universal Dependencies.” Computational Linguistics 47 (2): 255-308. https://doi.org/10.1162/coli_a_00402.
9. Gagarina, Natalia, Daleen Klop, Sari Kunnari, Koula Tantele, Taina Välimaa, Ute Bohnacker, and Joel Walters. 2019. MAIN: Multilingual Assessment Instrument for Narratives – Revised. ZAS Papers in Linguistics 63. https://doi.org/10.21248/zaspil.63.2019.516.
10. Hamidova, Gulrux Axmat qizi. 2026. “Maktab yoshidagi o‘zbekzabon bolalar og‘zaki nutqini MAIN metodologiyasi asosida o‘rganish.” Tadqiqotlar 88 (6): 110–112.
11. Ibbotson, Paul, et al. 2024. “The Effects of Family, Culture and Sex on Linguistic Development across 20 Languages.” Developmental Science. https://doi.org/10.1111/desc.13547.
12. Lelik, Valeriia, Mariia Diachkova, Svetlana V. Dorofeeva, Anastasiya Lopukhina, Olga Dragoy, and Irina A. Sekerina. 2025. RusLan-M Corpus. CHILDES/TalkBank. https://doi.org/10.21415/1XMG-D508.
13. Lieven, Elena, Dorothé Salomo, and Michael Tomasello. 2009. “Two-Year-Old Children’s Production of Multiword Utterances: A Usage-Based Analysis.” Cognitive Linguistics 20 (3): 481–507.
14. MacWhinney, Brian. 2000. The CHILDES Project: Tools for Analyzing Talk. 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates.
15. MacWhinney, Brian. 2012. “Morphosyntactic Analysis of the CHILDES and TalkBank Corpora.” In Proceedings of LREC 2012.
16. MacWhinney, Brian. 2019. “Understanding Spoken Language through TalkBank.” Behavior Research Methods 51: 1919-1927. https://doi.org/10.3758/s13428-018-1174-9.
17. MacWhinney, Brian, and Davida Fromm. 2022. “Language Sample Analysis With TalkBank: An Update and Review.” Frontiers in Communication 7: 865498. https://doi.org/10.3389/fcomm.2022.865498.
18. Masharipova, Baxtigul, and Barno Qutlimuratova. 2024. “O‘zbek tili bolalar nutqi audio korpusini yaratish.” Computer Linguistics: Problems, Solutions, Prospects 1 (1).
19. Nematova, Shakhlo, Henry Brice, Sarah Akkad, Mary-Claire Ball, and Kaja Jasińska. 2026. “Cross-Linguistic Transfer of Phonological Awareness and Reading Skills in Russian–Uzbek Bilingual Children.” Journal of Research in Reading. https://doi.org/10.1111/1467-9817.70031.
20. Newman, Rochelle S., Meredith L. Rowe, and Nan Bernstein Ratner. 2016. “Input and Uptake at 7 Months Predicts Toddler Vocabulary: The Role of Child-Directed Speech and Infant Processing Skills in Language Development.” Journal of Child Language 43 (5): 1158–1173.
21. Qurbonova, Munavvara A. 2009. O‘zbek bolalar nutqi leksikasining sotsiopsixolingvistik tadqiqi. Filologiya fanlari nomzodi dissertatsiyasi. Toshkent: O‘zMU.
22. TalkBank. 2026a. “CHILDES Corpora.” Accessed August 22, 2026. https://talkbank.org/childes/access/index.html.
23. TalkBank. 2026b. “CHILDES Kazakh Balausa1 Corpus.” Accessed August 22, 2026. https://talkbank.org/childes/access/Other/Kazakh/Balausa1.html.
24. TalkBank. 2026c. “CHAT Transcription Format.” Updated August 5, 2026. https://talkbank.org/0info/manuals/CHAT.html.
25. Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18.
26. Yang, Xiulin, Zhuoxuan Ju, Lanni Bu, Zoey Liu, and Nathan Schneider. 2025. “UD-English-CHILDES: A Collected Resource of Gold and Silver Universal

Dependencies Trees for Child Language Interactions." In Proceedings of the Eighth Workshop on Universal Dependencies (SyntaxFest 2025).