





The Balausa Corpus: A Documentation Study of First Language Acquisition of Kazakh within the Acquisition Sketch Project

by

Ilya Razorenov

A Master's thesis submitted in partial fulfilment of the requirements for the degree of

Master of Arts / Master of Science

At the Joint European Erasmus Mundus Master's Programme in Clinical Linguistics (EMCL++)

University of Groningen, the Netherlands

University of Eastern Finland, Finland

Ghent University, Belgium

Student Number: S5846218 (RUG); 2511307 (UEF)

Address: Karjamäentie 10C, 19c, Joensuu, Finland

E-Mail: i.razorenov@student.rug.nl

First Supervisor: Dr. Michael Rießler, University of Eastern Finland, Joensuu, Finland

Second Supervisor: Dr. Irina Sekerina, City University of New York (CUNY), New York, USA

Date of Submission: June 24, 2026

Abstract

The study of first language acquisition (FLA) has been largely based on a limited number of well-documented languages, leaving many languages, including Kazakh, underrepresented in FLA research. This lack of data restricts our understanding of cross-linguistic variation in early language development and highlights the importance of expanding FLA research to less-studied languages. The present study addresses this gap by introducing the Balausa corpus, a newly compiled dataset of early Kazakh child language, and applying the Acquisition Sketch Project (ASP) framework to these data. The Balausa corpus was compiled from naturalistic caregiver-child interactions and consists of approximately 580 minutes of recorded data of which 284 minutes (7,798 utterances; 17,785 tokens) were transcribed and morphologically annotated in the CHAT format using the BatchAlign2 (BA2) toolkit. The study examined developmental patterns in Kazakh child speech across lexical, morphological, and semantic domains using a combination of quantitative and qualitative approaches. Quantitative analyses included measures such as noun-to-verb ratio (NVR), measure of textual lexical diversity (MTLD), and mean length of utterance in morphemes (MLU-m), with the relationships between age and these measures examined using Spearman correlation analyses. Qualitative analyses focused on lexical frequency patterns, semantic field distributions, and the emergence of nominal and verbal morphological markers across age groups. The findings provide preliminary evidence of developmental changes in Kazakh child speech, including a gradual shift away from strong noun dominance and towards increased verb use, greater grammatical complexity, broader semantic representations, and an increasing ability to refer to past events. Overall, Balausa contributes one of the first systematically prepared corpora of Kazakh child language and demonstrates the applicability of the ASP framework, providing a foundation for future FLA research in Kazakh and cross-linguistic comparison.

Keywords: Acquisition Sketch Project, language acquisition, child language, Kazakh, agglutinative languages, child language corpus

Acknowledgements

I would like to acknowledge my supervisors, Dr. Michael Rießler and Dr. Irina Sekerina, for their continuous support both within and beyond this project. This is my first serious piece of academic work, and I am truly grateful that it was you who guided me throughout its development. I would also like to give special thanks to my family and friends in Kazakhstan for their ongoing support and motivation. In particular, I would like to thank my father, who did not live to see me graduate, yet for whom it would have meant the world. You have been my role model in life and made everything possible from the very beginning.

Table of Contents

Abstract ii
Acknowledgements iii
Table of Contents iv
List of Tables vi
List of Figures vii
CHAPTER 1. INTRODUCTION 8
1.1. First Language Acquisition, Acquisition Sketch Project, and CHILDES 8
1.1.1. First Language Acquisition and Current Research Challenges 8
1.1.2. The Acquisition Sketch Project 9
1.1.3. CHILDES 10
1.2. The Kazakh Language, Its Sociolinguistic Context, and Previous Research on FLA of Kazakh 11
1.2.1. The Kazakh Language 11
1.2.2. Sociolinguistic and Historic Context of Kazakh 12
1.2.3. Existing FLA Research on Kazakh 13
1.3. Research Aims 14
CHAPTER 2. DATA COLLECTION AND PROCESSING 16
2.1. Data Availability and Accessibility 16
2.2. Method 16
2.2.1. Participants 16
2.2.2. Design and Materials 17
2.2.3. Procedure 17
2.2.4. Data Analysis 20
2.3. Automated Transcription and Corpus Preparation 21
2.3.1. Automated Transcription and Processing Pipeline (BatchAlign2) 21
2.3.2. Manual Correction and Common BatchAlign2 Errors 22
2.4. Final Structure and Annotation of the Balausa Corpus (CHAT Format) 26
CHAPTER 3. ACQUISITION SKETCH OF KAZAKH 30
3.1. Lexicon 30
3.1.1. Noun-to-Verb Ratio (NVR) 30
3.1.2. Measure of Textual and Lexical Diversity (MTLD) 32
3.1.3. Lexical Frequency Patterns 34
3.2. Morphology 36
3.2.1. Mean Length of Utterance in Morphemes (MLU-m) 36

3.2.2. Emergence of Nominal and Verbal Morphology 38
3.3. Semantics 43
3.3.1. Semantic Field Distribution of Nominal and Onomatopoeic Forms 43
3.3.2. Temporal Reference in Verbal Production 46
CHAPTER 4. DISCUSSION 49
4.1. Interpretation of Linguistic Findings 49
4.2. Corpus Construction and Methodological Considerations 51
4.3. Limitations and Future Directions 52
CHAPTER 5. CONCLUSION 53
SOCIAL IMPACT OF THE STUDY 55
REFERENCES 56
DECLARATION 64
APPENDICES 65
Appendix A 65
Appendix B 66

List of Tables

Table 1 Demographic Information for Child Participants 17
Table 2 Recording Contexts, Devices, and Durations by Participant 18
Table 3 Most Common Transcription-Level Errors 23
Table 4 Most Common Morphological Annotation Errors 25
Table 5 CHAT Notations Used in the Balausa 27
Table 6 Noun and Verb Token Counts and NVR by Child and Age 31
Table 7 MTLD Values by Child and Age 33
Table 8 Three Most Frequent Nouns and Verbs by Child and Age 35
Table 9 MLU-m Values by Child and Age 37
Table 10 Earliest Attested Nominal Morphological Features by Age 39
Table 11 Earliest Attested Verbal Morphological Features by Age 41
Table 12 Total Number of Nouns and Onomatopoeia per Child and Age 45
Table 13 Distribution of Present/Non-Past and Past-Marked Verbal Forms by Child and Age 47

List of Figures

Figure 1 CHAT Transcript Displayed in the CLAN Interface 27
Figure 2 Relationship between Age and Token-based NVR 31
Figure 3 Relationship between Age and MTLD 33
Figure 4 Relationship Between Age and MLU-m 37
Figure 5 Semantic Field Distribution in Child Speech by Age 44
Figure 6 Relationship between the Proportion of Past-Marked Verbal Forms and Age 47

CHAPTER 1.
INTRODUCTION

1.1. First Language Acquisition, Acquisition Sketch Project, and CHILDES

1.1.1. First Language Acquisition and Current Research Challenges

First language acquisition (FLA) refers to the natural process through which children acquire the language(s) of their environment during early childhood through exposure and interaction (Yule, 2010). Because this process mostly occurs during everyday interactions with caregivers, researchers have long relied on recordings of children's natural speech to study how language emerges (Fernández & Cairns, 2010). Investigating such patterns systematically increasingly requires language resources, including corpora, that capture natural language use (Wilbur & Rießler, 2022). While early FLA research often relied on diary studies and longitudinal observations of individual children to obtain such data (e.g., Darwin, 1877), later work adopted corpus-based methods supported by digital tools. In particular, the development of the CHAT transcription format and CLAN software enabled researchers to store, annotate, and analyze large collections of child language data more systematically (MacWhinney, 2000). These developments facilitated FLA research and supported the emergence of large-scale corpus-based studies of child language.

Child language corpora now play an important role in FLA research, allowing researchers to investigate theoretical questions across many linguistic domains. For example, corpus methods have been used to study verb inflections in Japanese-speaking children (Tatsumi et al., 2021), syntactic categories in early English (Joo et al., 2020), and semantic content categories in Mandarin-speaking preschoolers (Tang et al., 2023). In addition, child language corpora can be used to calculate widely used quantitative measures of lexical and grammatical development. For example, measures such as Mean Length of Utterance (MLU; Brown, 1973) and Type-Token Ratio (TTR; Templin, 1957) are often derived from corpus data to provide complementary information about children's grammatical and lexical development.

At the same time, one of the key challenges of FLA research remains the lack of data from smaller and less documented languages (Bracks & Bardají i Farré, 2023; Kidd & Garcia, 2022). Cross-linguistic reviews have repeatedly shown that most acquisition research focuses on widely studied European languages, while the majority of the world's languages

remain absent from the literature (Defina et al., 2023a; Kidd & Garcia, 2022; Nielsen et al., 2017). Kidd and Garcia (2022), for instance, reviewed 3,123 articles published in four major FLA journals and found that acquisition research had been conducted on only 103 of approximately 7,000 languages worldwide, representing just 1.5% of the world's languages. Among these studies, 84% focused on English and other major Indo-European languages, further illustrating the strong concentration of FLA research on a limited set of well-studied languages. Defina et al. (2023a) similarly note that, despite the importance of linguistic diversity for understanding human language, only around 1-2% of the world's languages are represented in current research. This lack of linguistic diversity in FLA research is problematic because evidence from a wider range of languages could help distinguish universal developmental patterns from those shaped by language-specific structures (Allen, 2024); however, the current linguistic coverage of child language research remains too limited to adequately evaluate such questions across the world's languages (Pye, 2021). In short, these findings highlight the importance of developing child language resources for understudied languages.

1.1.2. The Acquisition Sketch Project

One recent response to the limited typological coverage of FLA research is the Acquisition Sketch Project (ASP), introduced by Hellwig et al. (2023). The project provides a practical framework for documenting child language acquisition and child-directed speech, especially in languages for which little acquisition data currently exist. The framework is organized around two main components: the sketch corpus (Part I) and the acquisition sketch (Part II), which are described in detail in the two Sketch Acquisition Manuals (SAMs; Defina et al., 2023a, 2023b).

The first component of the framework, the sketch corpus (Defina et al., 2023a), focuses on corpus construction. More specifically, ASP provides detailed guidance on corpus design, participant sampling, data collection, and documentation. In brief, it recommends collecting at least 10 hours of recordings, of which approximately 5 hours are selected for detailed analysis and archiving. ASP suggests recruiting children between roughly two and four years of age, often at six-month intervals, using either cross-sectional or longitudinal designs. In addition to the recordings themselves, the ASP framework encourages the collection of metadata on participants, interactional contexts, multilingual language use, and other aspects of the child's communicative environment.

The second component of the framework, the acquisition sketch (Defina et al., 2023b), focuses on describing patterns of child language development based on the sketch corpus. The main focus of this component is the analysis of child speech across different linguistic domains, including phonology, lexicon, morphology, syntax, semantics, pragmatics, and gesture. In addition, a full acquisition sketch may include analyses of child-directed (caregiver) speech, the learning environment, local beliefs and practices surrounding language development, and community-oriented outputs. Such outputs help share the findings with the local community and may take the form of reports, educational materials, videos, presentations, or discussions with parents, educators, and other stakeholders. However, given the scope and length constraints of the present Master's thesis, only a subset of these components could be addressed, with the resulting adaptations discussed in Section 2.2.4.

As a relatively recent initiative, ASP has so far been applied in only a limited number of published studies. Bracks and Bardají i Farré (2023), for example, describe the creation of a sketch corpus for Totoli, a western Austronesian language of Indonesia, showing how methods from adult language documentation can be extended to child language research in an understudied language. Urbanczik (2023), in turn, provides an early evaluation of the sketch format through a study of case marking in German, illustrating the flexibility of the ASP framework in an already well-documented language context. Beyond these initial applications, Lee and Allen (2023) provide the first acquisition sketch to follow both SAMs in full, focusing on Inuktitut, an Inuit language of Canada. These early applications suggest that the ASP framework can be adapted to languages with very different levels of prior documentation and applied to a range of research purposes.

1.1.3. CHILDES

One of the key developments supporting corpus-based FLA research has been the creation of the Child Language Data Exchange System (CHILDES; MacWhinney, 2000), which provides a framework for storing, transcribing, and sharing child language data. Over time, CHILDES has become the largest international repository of child language data. To illustrate its scope, Kempe et al. (2024) report that as of 2024, the database included 436 corpora involving 16,382 children and caregivers. CHILDES was developed in response to a major limitation of earlier FLA research: although many scholars collected naturalistic child speech data, these materials were often kept in private archives and transcribed using project-specific conventions, making them difficult to compare and reuse. By promoting shared transcription and annotation standards alongside open access to corpora, CHILDES has made

child language data more accessible and comparable across languages. However, despite its considerable scope, CHILDES also reveals gaps in linguistic representation. For example, even Turkish, one of the most widely spoken Turkic languages, is currently grouped under the general category “Other” in CHILDES, reflecting the limited representation of Turkic child language data within the database. Nevertheless, despite these gaps in representation, CHILDES remains an important resource for documenting and sharing child language data from a growing number of languages.

1.2. The Kazakh Language, Its Sociolinguistic Context, and Previous Research on FLA of Kazakh

1.2.1. The Kazakh Language

Kazakh is a Turkic language of the Kipchak (Northwestern) branch (Johanson, 2021). It has approximately 17.3 million native speakers worldwide, the majority of whom live in Kazakhstan, with additional large communities in China and Uzbekistan (WorldData.info, 2025). Kazakh holds official status in Kazakhstan, the Russian Republic of Altai, and the Kazakh Autonomous District of Ili in Xinjiang, China (WorldData.info, 2025). Throughout its history, the language has been written in several scripts, including Ancient Turkic, Arabic, Latin, and Cyrillic (Nakispekova, 2026). As a widely spoken language with structural features that differ from well-studied Indo-European languages, Kazakh could represent an important addition to FLA research.

The grammatical overview presented below draws primarily on Dotton and Wagner (2018), one of the few comprehensive descriptions of Kazakh available in English. Morphologically, Kazakh is an agglutinative language in which grammatical meanings are expressed primarily through extensive suffixation. Nouns do not mark grammatical gender, but they inflect for number and seven grammatical cases: nominative, genitive, dative, accusative, locative, ablative, and instrumental. The verbal system is morphologically rich, with suffixes expressing tense, aspect, mood, and person. Kazakh also makes frequent use of auxiliary verb constructions, typically combined with lexical verbs in converbal forms, to express aspectual, modal, and other functional distinctions. Negation is usually marked morphologically through verbal suffixes, although certain constructions use the negative existential form *жоқ* (‘there is no/not’), which functions as an analytic negation marker.

Syntactically, Kazakh is generally described as a Subject-Object-Verb (SOV) language, in which the main verb typically appears at the end of the clause. However, because grammatical relations are overtly marked through case suffixes, word order could be

flexible, and alternative patterns may occur for discourse or pragmatic purposes. Furthermore, Kazakh is a pro-drop language, allowing subjects to be omitted when their reference can be recovered from context or verbal agreement. Within noun phrases, modifiers such as adjectives and numerals usually precede the noun, and the language uses postpositions rather than prepositions. Subordinate clauses commonly precede the main clause and are often formed using participial or converb constructions. Interrogative clauses do not require subject inversion, and yes-no questions are commonly marked by a question particle at the end of the sentence. Overall, these grammatical characteristics make Kazakh a valuable language for examining how developmental patterns observed in well-studied languages extend to typologically different linguistic systems.

1.2.2. Sociolinguistic and Historic Context of Kazakh

Despite having millions of native speakers, the use of Kazakh in modern-day Kazakhstan continues to be shaped by historical and social factors. While ethnic Kazakhs constitute around 70% of Kazakhstan's population, with Russians making up only around 15% (Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan, 2025), this demographic dominance has not translated into linguistic dominance. In fact, only around 80% of Kazakhstan's citizens report proficiency in Kazakh, and roughly half indicate that they use it in daily life (Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan, 2021).

This pattern is closely linked to the historical language hierarchy established during the Soviet period, when Kazakh had a limited role in domains such as economics, science, official documentation, and mass media, and was more strongly associated with cultural expression and rural life (Terlikbayeva & Menlibekova, 2021). Russian, by contrast, functioned as the main language of administration, education, and upward social mobility, a role that continues to shape language practices today (Suleimenova et al., 2021). This hierarchy is also reflected at the institutional level: the 1995 Constitution designates Kazakh as the state language, while Russian retains official use in government institutions and local administration (Jumagaliyeva, 2021). Smagulova (2008) further reports that many members of Kazakhstan's parliament did not speak Kazakh and that much official paperwork was conducted in Russian. Such prestige dynamics continue to influence language use among young people, as some university students report not speaking Kazakh in Russian-speaking contexts to avoid being perceived as less competent by their peers (Kucherbayeva & Smagulova, 2023). Together, these historical and sociolinguistic factors help explain why the

widespread presence of Kazakh speakers has not necessarily translated into equally widespread use of the Kazakh language in the country.

1.2.3. Existing FLA Research on Kazakh

Given the historical and sociolinguistic factors outlined above, it is unsurprising that research on the acquisition of Kazakh remains limited and that much of the existing literature has focused on Russian or Kazakh-Russian bilingualism. The present review identified only one major source focusing directly on the FLA of Kazakh: a report by the National Scientific and Practical Center for Special Education (2019). This report provides a broad description of Kazakh language development in children aged 2-6, based on parent observations collected as part of a monitoring project. Rather than employing systematic methodological procedures, detailed corpus data, or quantitative analyses, the report primarily provides a descriptive overview of the emergence of different language components during the preschool years. Beyond this report, studies focusing specifically on monolingual Kazakh FLA remain limited. Much of the existing literature instead approaches development through related areas, especially Kazakh-Russian bilingualism and language contact. Several studies examine Kazakh-Russian family language practices and children's bilingual speech (Akhmet et al., 2023; Smagulova, 2014). A more specific line of research addresses lexical development in Kazakh-Russian bilingual children, including comparisons with Russian monolingual peers, comprehension of Russian lexical items, and lexical interference in bilingual speech (Aldaberdikyzy, 2013; Zhakupova et al., 2024; Zhakupova et al., 2025). Studies have also examined Kazakh as a foreign and heritage language, including research with migrant families and English-speaking learners of Kazakh (Abdyeva et al., 2015; Boribayeva, 2024; Kuzembayeva et al., 2024; Zhumashova et al., 2023). All in all, while these studies undoubtedly contribute valuable insights into different aspects of Kazakh language development, they do not address the lack of research on early monolingual Kazakh acquisition.

A second major gap concerns the lack of child corpora for studying Kazakh acquisition. As noted by Wilbur and Rießler (2022), empirical research on language use depends on reliable corpora that represent natural speech, which are also important for long-term preservation and future reuse. Although several digital corpora of Kazakh exist, they focus mainly on adult language. For example, the National Corpus of the Kazakh Language includes large collections of written language produced by adults and includes several specialized subcorpora (Zhubanov et al., 2012), while Troiani and Filchenko (2026)

developed a multimedia corpus of modern spoken Kazakh that also consists primarily of adult data. Furthermore, the few studies involving Kazakh-speaking children have not made their recordings publicly available. As a result, there remains a clear lack of accessible corpora of Kazakh child speech. This situation contrasts with other languages that are regionally important or typologically relevant to Kazakh. For example, child language corpora for Russian and Turkish are available through CHILDES, including datasets such as Lelik et al. (2025), Slobin (1982), and Altinkamis (2004). Summing up, the lack of accessible corpora of Kazakh child speech continues to limit opportunities for systematic and reproducible FLA research on Kazakh.

1.3. Research Aims

Against the background outlined above, the present study aims to address two major gaps in research on FLA in Kazakh: the limited availability of empirical research on Kazakh development and the lack of publicly accessible corpora of Kazakh child speech. To address these gaps, the study introduces the Balausa corpus, the first publicly available dataset of Kazakh child speech prepared in a CHILDES-compatible format, and uses this data to develop a preliminary acquisition sketch of Kazakh within the ASP framework. In this way, the project aims to support future research on Kazakh language acquisition while also contributing to broader cross-linguistic comparison. These overall objectives can be divided into four specific aims:

1. *Data Collection Aim (Aim 1)*

Aim 1. To record, transcribe, and annotate naturalistic interactions between Kazakh-speaking children aged 2-5 and their caregivers in everyday settings in order to document early stages of Kazakh child language development.

2. *Research Output Aims (Aims 2-4)*

Aim 2. To compile and organize the recordings collected in Aim 1 into a small corpus prepared in a CHILDES-compatible format to support long-term accessibility and future reuse.

Aim 3. To create a preliminary acquisition sketch of Kazakh following the ASP framework, with particular attention to lexical, morphological, and semantic features of child speech.

Aim 4. To provide structured and accessible data on Kazakh within the ASP framework to support cross-linguistic comparisons of child language development across diverse linguistic typologies.

Beyond its relevance for linguistic research, the creation of Balausa may also have broader value for language technology. In this context, Kazakh is often described as a low-resource language in terms of available digital data (Rakhimova et al., 2025). This lack of resources is particularly evident in the case of child speech data, which is especially useful because automatic speech recognition (ASR) systems often perform less accurately on children's speech than on adult speech, partly due to the limited availability of child training data (Potamianos & Narayanan, 2003). Accordingly, by providing structured recordings of Kazakh child speech, the present study may therefore also support future technological development for the language.

The remainder of the thesis is organized as follows. Chapter 2 describes the methodology, including participants, research design, data collection, transcription, and annotation procedures. Chapter 3 presents the results across the domains of lexicon, morphology, and semantics. Finally, Chapters 4 and 5 discuss the findings and conclude the study.

CHAPTER 2.
DATA COLLECTION AND PROCESSING

2.1. Data Availability and Accessibility

The Balausa corpus, currently released as Balausa1, is publicly available through the CHILDES database (https://talkbank.org/childes/access/Other/Kazakh/Balausa1.html). In addition, the corresponding audio recordings, transcriptions, CMDI metadata files, Python scripts used for data analysis, and the corpus report can be accessed via the project's GitHub repository (https://github.com/langdoc/kaza1248-aquisition). The GitHub repository also includes recordings that were collected as part of the project but were not transcribed or analyzed within the scope of the present thesis, allowing these materials to be reused in future research. The repository infrastructure was established in collaboration with Dr. Michael Rießler (https://github.com/meehkal) and Sven Esser (https://github.com/SvenEss16), and the metadata organization draws on work developed by Esser (2026).

2.2. Method

The Method section follows the general structure and recommendations of the ASP framework. For detailed guidance on sketch corpus construction, see Defina et al. (2023a), and for the acquisition sketch procedures, see Defina et al. (2023b).

2.2.1. Participants

Eight Kazakh-speaking families with children aged 2-5 years (3 male, 5 female) were recruited online from different regions of Kazakhstan, including both urban and rural areas. Children were selected at roughly six-month age intervals, and the inclusion criteria required that children be raised in Kazakh-speaking households and have no reported developmental or language-related disorders. Recorded caregivers were expected to primarily use Kazakh when interacting with the child. However, all participating families reported some degree of Russian use within the household or broader social environment, reflecting the bilingual context of Kazakhstan. According to the ASP recommendations, an ideal sketch corpus includes recordings from at least ten families. To partly compensate for the smaller sample size and to ensure coverage of most age points in Balausa, two children (AE-01, NS-01) were recorded at two different time points, and in future descriptive analyses, recordings from these two children were treated as separate speaker entries at each time point. All demographic information for the child participants is presented in Table 1.

Table 1

Demographic Information for Child Participants

Child ID	Sex	Age[a]	Place of Residence
AE-01	f	1;09	Zerendi
KH-02	f	1;11	Qosshy
AE-01	f	2;00	Zerendi
ZQ-02	f	2;04	Astana
ZQ-01	f	3;00	Kokshetau
MK-01	m	3;05	Almaty
KH-01	f	3;10	Qosshy
NS-01	m	4;01	Astana
NS-01	m	4;05	Astana
BK-01	m	5;01	Astana

[a] Age is presented in the standard year;month format

Because the study involved children under the age of 12, informed consent was obtained from caregivers for themselves and their children prior to participation. As this Master's thesis did not involve medical populations, formal ethical review was not required under the guidelines of the University of Eastern Finland, which remains the owner of the collected data. Nevertheless, the study followed standard ethical research principles, and no sensitive personal information was collected from participants beyond the basic metadata necessary for the corpus.

2.2.2. Design and Materials

The ASP framework allows researchers to use either longitudinal or cross-sectional approaches when documenting child language development (Defina et al., 2023a). Given the time and scope of this Master's project, a cross-sectional design was chosen. No experimental manipulation or elicitation tasks were used. Instead, the recordings consisted of naturalistic caregiver-child interactions that took place in everyday home environments, allowing children to speak freely and comfortably. This methodological choice is further supported by previous research suggesting that naturalistic observations provide a more ecologically valid reflection of children's everyday language use (Lewkowicz, 2001; Tamis-LeMonda et al., 2017).

2.2.3. Procedure

After informed consent was obtained, families were asked to record naturally occurring interactions with their child at home during everyday activities. Caregivers were instructed to place a recording device near the child and allow interactions to unfold

naturally. Participants were additionally encouraged to choose moments when they felt comfortable recording longer interactions, since each participating family was asked to provide approximately one hour of recordings. Caregivers were additionally encouraged to avoid involving too many participants in the same recording to minimize overlapping speech and maintain recording quality. Most recordings therefore consisted of interactions between mothers and children, with occasional contributions from fathers, grandparents, or siblings.

Recordings were made primarily using caregivers' personal mobile phones. Prior to transcription and analysis, all audio files were converted into .wav format, in line with common technical recommendations for corpus creation (Wilbur & Rießler, 2022). In total, 580 minutes of audio recordings were collected. Of this material, 284 minutes were selected for transcription and annotation, consistent with ASP recommendations that approximately half of the collected recordings undergo detailed analysis. The resulting corpus comprised 3,621 child utterances (approximately 7,005 tokens) and 4,177 adult utterances (approximately 10,780 tokens). The selected recordings were then prepared in CHAT format (MacWhinney, 2000) and morphologically annotated using Universal Dependencies resources (UD; de Marneffe et al., 2021). A more detailed description of the transcription and annotation workflow is provided in later sections of this chapter. Detailed information on recording contexts, devices, and durations for each participant is presented in Table 2.

Table 2

Recording Contexts, Devices, and Durations by Participant

Child ID	Balausa1 Audio	Recording Context(s)[a]	Other Participants	Recording Device	Minutes Recorded	Minutes Transcribed
AE-01 (1;09)[b]	AE-01_004; AE-01_005	toyplay	mother, father	iPhone 13	66	31
KH-02	KH-02_001	toyplay	mother	iPhone 15 Pro	61	25
AE-01 (2;00)	AE-01_007; AE-01_008	pictures	mother, sister	iPhone 13	33	33

ZQ-02	ZQ-02_001	toyplay	mother, father	iPhone 15 Pro Max	63	23
ZQ-01	ZQ-01_001; ZQ-01_002	interview	mother, grandmother	Samsung Galaxy S20+	63	30
MK-01	MK-01_002	everyday activities, toyplay	mother, brother	iPhone 15	61	30
KH-01	KH-01_001b	toyplay	mother, sister	iPhone 15 Pro	61	30
NS-01 (4;01)	NS-01_001	interview, narrative	mother	Samsung Galaxy A23	53	26
NS-01 (4;05)	NS-01_003a	toyplay, interview, narrative	mother, grandmother, grandfather	Samsung Galaxy A23	61	30
BK-01	BK-01_001	toyplay, interview	mother	iPhone 15 Pro Max	58	26

[a] Recording contexts were classified using the activity categories established by MacWhinney (2000).
[b] For children recorded at multiple time points, age is provided to distinguish between sessions.

An additional stage of data preparation involved the anonymization of potentially identifying information. Defina et al. (2023a) in their manual suggest that anonymization procedures be discussed together with the archive where the data will be stored. CHILDES guidelines recommend replacing names with placeholders such as “Firstname,” “Lastname,” or “Cityname” (MacWhinney, 2000). However, such replacements are not always ideal for languages in which personal names may carry grammatical marking, as names can function as subjects, objects, or vocatives and therefore contribute to sentence structure (Bick, 2020). For this reason, pseudonymization was adopted in the present study, consistent with approaches used in other corpora, including German and Swedish (Bick, 2020; Megyesi et al., 2018).

In Balausa, pseudonyms were created for personal names, surnames, streets, squares, and village names. Separate pseudonym lists were prepared for male and female children,

parents, and other participants. Lists of popular Kazakh first names (TengriNews, 2025) and common Kazakh surnames (TengriNews, 2022) were used to replace individuals' names. For streets and squares, widely used names found across many locations in Kazakhstan were selected, while fictional but realistic village names were created with the help of native Kazakh speakers. Major city names were not changed, as they were not considered to present a meaningful identification risk. A full list of pseudonyms is provided in Appendix A.

In addition to the recordings themselves, metadata about participants, interlocutors, and recording conditions were collected. Metadata intended for corpus documentation were stored using the Component Metadata Infrastructure (CMDI) format. The ASP framework also recommends collecting more detailed background information about the child and their interactional environment, including daily routines, early language development, regular interlocutors, and the languages typically used in interaction (Defina et al., 2023a). To meet these requirements, additional ASP metadata files were created. CMDI files can be found in the project's GitHub repository, whereas ASP metadata, which contain more sensitive participant information, were stored privately.

2.2.4. Data Analysis

The data analysis combines quantitative and qualitative approaches in line with ASP guidelines. However, not all components of the ASP framework were included in the present study. First, the analysis did not include child-directed speech. Second, the domains of phonology, syntax, and gesture were excluded from the analysis. Third, the community-oriented component of the framework was not implemented. These choices were made to keep the project manageable within the scope of a Master's thesis. As a result, within this scope, child speech was examined across the three selected domains of lexicon, morphology, and semantics. In the lexical domain, analyses included noun-to-verb ratio (NVR), measure of textual lexical diversity (MTLD) for verbs and nouns, and lexical frequency patterns. In the morphological domain, the focus was on MLU-m and the emergence of nominal and verbal morphological features. Finally, in the semantic domain, analyses examined the distribution of semantic fields among nouns and onomatopoeic forms, as well as patterns of temporal reference expressed through verbal morphology.

Across the three domains, quantitative analyses were conducted primarily through frequency searches applied to the morphologically annotated tier in order to extract relevant lexical and grammatical categories. Age-related trends were examined by treating age as a continuous variable measured in months, and Spearman's rank correlation analyses were used

to test relationships between age and selected quantitative measures. Developmental patterns were visualized using scatterplots and heatmaps. Qualitative analyses complemented these findings where closer inspection of the data was required, such as in semantic field classification, identification of earliest attested morphological forms, and interpretation of lexical frequency patterns.

2.3. Automated Transcription and Corpus Preparation

2.3.1. Automated Transcription and Processing Pipeline (BatchAlign2)

This subsection describes the automated transcription and processing pipeline used for the construction of Balausa, focusing on the BatchAlign2 (BA2; Liu & MacWhinney, 2024) toolkit. Manual transcription in corpus building can be extremely time-consuming (Liu et al., 2023). For example, Novotney and Callison-Burch (2010) estimate that one hour of recorded speech may require around eleven hours of transcription work. These challenges are especially pronounced in child language data, which often involves interruptions, incomplete utterances, and overlapping speech, making transcription and annotation more demanding than for adult speech (Bracks & Bardají i Farré, 2023). Consequently, automated approaches to transcription can play an important role in reducing the time required for corpus construction.

To address these challenges, Liu et al. (2023) introduced an automated pipeline called BatchAlign (BA). BA performs several processing steps automatically, including ASR, utterance tokenization, speaker identification, and forced alignment between audio and text. The resulting output is generated directly in CHAT format, making it compatible with the CHILDES infrastructure. BA developers estimate that the system can reduce human transcription time by up to 75%, reporting that the automated processing of a 30-minute recording required approximately seven minutes, followed by around one hour of manual correction. In their evaluation, the reported word error rate (WER) ranged from 2.4% for control speakers to 3.4% for clinical populations. These results suggest that BA can generate reasonably accurate transcriptions while substantially reducing processing time, although human verification still remains necessary at the final stage.

Liu and MacWhinney (2024) later developed BA2, an updated version of the original BA toolkit. This latest version now incorporates the UD resources (de Marneffe et al., 2021) for morphosyntactic tagging, which extracts information such as parts of speech (POS), grammatical features, and grammatical relations. BA2-produced transcriptions also allow researchers to apply various tools to transcriptions, including KIDEVAL, IPSyn, and

GraphViz, which can be used for linguistic profiling or for visualizing grammatical relations. However, it should be noted that many of these tools currently support only a limited set of widely spoken languages, which restricts their direct applicability to Kazakh.

Building on these developments, the present project used BA2 to generate initial transcripts and perform morphosyntactic tagging. Transcription was performed using a local installation of Whisper, ensuring that audio files remained on the researcher's computer throughout processing. Although ASR performance for Kazakh remains limited due to the scarcity of training data, Rakhimova et al. (2025) found that Whisper-based models currently provide the best available performance for the language, reporting an accuracy of approximately 33% and a WER of 0.416. They further note that the performance of such systems could improve substantially if larger publicly available Kazakh speech corpora, particularly child speech corpora, were available. By providing structured recordings of Kazakh child speech, Balausa may therefore contribute to the future development of ASR systems for the language.

Practical use of BA2 in the present study revealed that both hardware configuration and recording quality affected different aspects of the pipeline. First, hardware configuration primarily influenced processing speed. On a CPU-based setup, the ASR stage required approximately one hour to process one minute of audio, whereas GPU processing significantly accelerated this step, reducing processing time to approximately 2.5 minutes per recording, making the process approximately 24 times faster. Morphosyntactic tagging required around eight minutes per hour of recording on CPU compared to approximately one minute on GPU. Second, recording conditions primarily affected output quality. Background noise, frequent interruptions, and other characteristics of naturalistic child speech often resulted in transcription and alignment errors, making manual verification necessary. Therefore, following automatic processing, all transcripts were manually reviewed and corrected with the assistance of native speakers of Kazakh, with the manual correction process and the most common BA2 errors described in the following subsection.

2.3.2. Manual Correction and Common BatchAlign2 Errors

The correction process involved comparing the automatically generated transcripts against the original audio recordings, assigning speakers, and adjusting utterance boundaries and time alignment. In addition, native Kazakh speakers assisted in the interpretation of region-specific vocabulary and colloquial expressions. After transcription correction, the resulting CHAT files were checked using the CLAN's *CHECK* command to identify

formatting inconsistencies. Overall, this manual correction proved to be the most time-consuming stage of the pipeline. On average, transcription and time-alignment correction required approximately eight hours for each 30-minute recording and up to ten hours for recordings involving younger children or particularly challenging acoustic conditions. Additional correction of the morphological tier (*%mor*) required approximately three to four hours per 30-minute recording, even when limited to nouns and verbs. In short, recordings required extensive revision; consequently, BA2 primarily functioned as an aid for initial segmentation and alignment rather than as a source of directly usable transcripts.

A formal evaluation of transcription accuracy was beyond the scope of the present study. As a result, measures such as WER were not systematically calculated, making direct comparison with the results reported by Rakhimova et al. (2025) difficult. Nevertheless, qualitative inspection of the output suggested that transcription accuracy was substantially lower than that reported for controlled benchmark datasets. In this respect, future studies could build on the present corpus to conduct a systematic evaluation of ASR performance on Kazakh child speech by comparing automatically generated output against manually corrected transcripts in Balausa and calculating measures such as WER.

The manual correction provided valuable insight into the types of errors produced by BA2. At the transcription level, the most frequently observed errors in recordings of younger children involved frequent repetitions of the same utterance and hallucinated utterances. Recordings of older children were generally more stable, although spelling inconsistencies and segmentation errors remained. Across all age groups, common issues included variable spelling of the same word, erroneous capitalization, and the insertion of characters that are not part of the standard Kazakh alphabet. These errors often carried over to later annotation stages, making manual correction of this tier necessary before morphological analysis. Examples of the most common transcription-level errors are presented in Table 3.

Table 3

Most Common Transcription-Level Errors

Error Type	Automatically generated output	Correct Annotation	Comments/Notes
Repetition	*PAR0: Қарай . *PAR0: Қалай .	*PAR0: Қара .	The same utterance was generated multiple times in

	*PAR0: Қолай . *PAR0: Қорай .		a row despite being produced only once.
Hallucinated utterance	*PAR0: ҮңғыңңүңҮғүүғ ғҮүҚүҮҚҚғыыҚ ҐҚғыҚ Құрғанда құтап оқу отырсын ба сен ?	*PAR0: Кітап оқып отырсың ба ?	Substantial hallucinated content at the beginning of the utterance
Erroneous capitalization	*PAR0: ҚАбуз	*PAR0: Қабыз	-
Variable spelling of the same word	*PAR0: Машина …. *PAR0: Машинаселінің … *PAR0: Мұшан	*PAR0: Машина	-
Non-Kazakh character insertion	*PAR0: БаҒанда	*PAR0: Барған	-

At the morphological level, a wide range of errors was observed, including occasional errors in POS assignment, lemma identification, and grammatical feature annotation. In some cases, relatively clear items were assigned the general tag *X*, indicating uncertainty in the automatic analysis. Lemma-related errors included truncated stems, incorrect base forms, and distorted lemma shapes. Grammatical feature errors were particularly common for nouns and verbs. Verbal errors involved tense, person, negation, voice, and politeness marking, whereas nominal errors mainly involved case, number, and possessive morphology. Finally, the same word form was occasionally assigned different morphological analyses within the same recording, even when no ambiguity was present. Examples of the main morphological annotation errors identified during manual correction are presented in Table 4.

Table 4

Most Common Morphological Annotation Errors

Error Type	Automatically generated output	Correct Annotation	Translation
POS assignment error	Аю pron|аю-Prs-Nom-S3	Аю noun|аю-Nom	‘a bear’
X-tag assignment	Маған x|маған	Маған pron|мен-Prs-Dat-S1	‘to me’
Incorrect lemma	Бақшамызда noun|бақшамыз-Loc	Бақшамызда noun|бақша-Loc-P1	‘in our garden’
Case error	Жақты noun|жақ-Loc	Жақты noun|жақ-Acc	‘direction’ (accusative form)
Number error	Аяқтармен noun|аяқ-Ins	Аяқтармен noun|аяқ-Plur-Ins	‘with the legs’
Person error	Барамыз verb|бар-Fin-Ind-Pres-S2	Барамыз verb|бар-Fin-Ind-Pres-P1	‘we go’
Negation error	Тиіспейміз verb|ти-Fin-Ind-Rcp-Pres-P1	Тиіспейміз verb|ти-Fin-Ind-Rcp-Neg-Pres-P1	‘we do not touch (something)’
Inconsistent analysis within the same recording	Қырық num|қырық … noun|қырық-Nom	Қырық num|қырық	‘forty’

Overall, observations from both the transcription and morphological annotation stages suggest that automatic processing of Kazakh child speech still requires substantial manual verification and correction. Although this additional effort should be acknowledged, the use of BA2 nevertheless reduced the time required for corpus construction compared to fully manual transcription and annotation procedures.

2.4. Final Structure and Annotation of the Balausa Corpus (CHAT Format)

This section describes the final structure and annotation conventions of the Balausa corpus following automatic processing and subsequent manual correction. The present study relies on the CHAT/CLAN environment (MacWhinney, 2000) for transcription and analysis. CHAT is a standardized transcription format used within CHILDES, which allows researchers to represent speech, speaker roles, metadata, timing information, and additional annotation tiers in a consistent way. CLAN is the accompanying software environment used to search, process, and analyze CHAT transcripts.

At the beginning of each CHAT transcript, metadata headers provide general information about the recording, including the language, corpus name, participant identifiers, speaker roles (e.g., *CHI* for child, *MOT* for mother), and participant age. The main body of the transcript is organized into speaker tiers, where each utterance is transcribed orthographically and can be time-aligned with the audio. In addition to the main orthographic tier, CHAT supports dependent tiers that provide additional layers of annotation. The most relevant tier for the present study is the *%mor* tier, which contains morphological annotation. This tier encodes lexical and grammatical information using the UD resources, typically consisting of POS labels, lemmas, and grammatical features. For example, using UD, the Kazakh noun *балаларға* (‘to the children’) would be glossed as *noun|бала-Plur-Dat,* representing the POS *noun*, the lemma *бала*, the plural suffix *-лар*, and the dative case suffix *-ға*. Notably, some unmarked features, such as nominative case, are also systematically tagged by BA2. The grammatical tier (*%gra*) was not included in the present corpus, as generating accurate grammatical relations would have required substantial additional manual intervention due to the number of errors observed in the orthographic and morphological tiers following automatic processing. Figure 1 presents an example of a CHAT transcript displayed in the CLAN interface, drawn from the Balausa corpus (https://talkbank.org/childes/access/Other/Kazakh/Balausa1.html), which includes speaker tiers, morphological glosses, and English translations.

Figure 1

CHAT Transcript Displayed in the CLAN Interface

```
*CHI:   Үш түсті , үш рет жүресің . •
%mor:   num|үш verb|түс-Fin-Ind-Past-3 cm|cm num|үш noun|рет-Nom
        verb|жүр-Fin-Ind-Pres-S2 .
%eng:   You rolled a three , so move three spaces .
*MOT:   Бір , екі , үш . •
%mor:   num|бір cm|cm num|екі cm|cm num|үш .
%eng:   One , two , three .
*CHI:   Болды . •
%mor:   verb|бол-Fin-Ind-Past-3 .
%eng:   That is enough .
```

In addition to standard orthographic transcription, several CHAT symbols and notations were used in Balausa to represent features of naturalistic spoken interaction (see Table 5).

Table 5

CHAT Notations Used in the Balausa

Notation	Function/description	Example	Notes
●	time alignment marker	*MOT: Қандай тамақтар жедің ? ●	links the utterance to a specific segment of the audio recording
()	truncated speech segment	ғо(й) ‘emphatic particle’	-
[:]	intended or standard form	балдар [: балалар] ‘children’	-
[* p] [* m]	phonological/morphological error	үйқа [: үйге] [* m] ‘to the house’	error code follows corrected form
/	repetition	иә / иә / иә ‘yes, yes, yes’	-
@s:rus	Russian borrowing	бабушка@s:rus ‘grandmother’	-
@s:rus+kaz	Russian borrowing with Kazakh morphology	бабушкаға@s:rus+kaz ‘to the grandmother’	Russian stem ‘бабушка’ with Kazakh suffix ‘-ға’
[- rus]	entire utterance in Russian	[- rus] Я хочу домой ‘I want to go home’	-
xxx	unintelligible segment	мен оған xxx бердім ‘I gave xxx to him/her	-
@l	letter name	к@ дешi ‘say к’	-
@o	onomatopoeic form	бум@o ‘boom’	-

&=	non-verbal vocalization or action	@laughs	-

The most detailed level of morphological glossing in Balausa was given to nouns and verbs. For verbs, annotations included POS, lemma, finiteness (finite vs. non-finite forms such as participles, converbs, and verbal nouns), mood, tense, person, number, politeness, and evidentiality. Mood was restricted to four most commonly described categories in Kazakh (indicative, imperative, conditional, and optative). As future-specific forms were not observed, tense distinctions were limited to present and past. This decision is consistent with accounts of Kazakh that describe the language as having a non-past category, in which the same endings may express both present and future meanings depending on context (McCollum & Chen, 2020). In this study, such forms were glossed as *Pres*. Person and number were annotated in a way that reflects their actual morphological marking. In Kazakh, person endings are often the same for both singular and plural forms, while plurality is expressed through a separate suffix. For this reason, person and number were annotated separately (e.g., *S2* and *Plur*) rather than being combined into a single label (e.g., *P2*). Finally, auxiliary verbs were marked with the *Aux*.

For nouns, annotations included POS, lemma, number (only for plural), and case (including unmarked nominative case). For proper nouns, tagged separately as *propn,* gender was also included if they referred to individuals. Where applicable, possessive and personal endings were annotated in nouns in order to capture two important categories in Kazakh nominal morphology. As UD resources do not provide a satisfactory way of distinguishing between them, the present study proposed a modified annotation convention. First, possessive endings (*мәуелдік жалғау*) were marked using person-number labels such as *S1* or *P1*, since these endings indicate the person and number of the possessor. This is illustrated in Example 1:

(1) бала-м
noun|бала-Nom-S1
‘my child’

Personal endings (*жіктік жалғау*), as found in nominal predicates, were glossed using person labels such as *1*, *2*, or *3* without the *S/P* distinction to avoid confusion with possessive endings. This is illustrated in Example 2:

(2) бала-сың
noun|бала-Nom-2
“you are a child”

Aside from these specific adjustments, CHAT proved suitable for capturing the main morphological features relevant to the analysis of Kazakh child speech in this project.

CHAPTER 3.
ACQUISITION SKETCH OF KAZAKH

Having outlined the procedures used for the construction of Balausa corpus, the study now turns to its second core objective: applying the ASP framework to Kazakh child speech data. Given the relatively limited size of the present dataset, the analyses should be viewed as a preliminary acquisition sketch rather than a comprehensive account of Kazakh language development. Because detailed morphosyntactic verification was carried out only for nouns and verbs, the chapter focuses mainly on patterns that can be examined reliably within these categories. In addition, phenomena explicitly marked during transcription, such as onomatopoeia, are also considered. This chapter is organized into three main sections corresponding to the core domains of the ASP: lexicon, morphology, and semantics.

3.1. Lexicon

3.1.1. Noun-to-Verb Ratio (NVR)

The distribution of nouns and verbs in early child language is an important area of lexical development research, as it provides insight into how children organize their early vocabularies (Stoll et al., 2012). A widely discussed phenomenon in this area is the so-called ‘noun bias’, referring to the tendency for children to acquire and produce nouns earlier and in greater proportions than verbs, possibly because nouns often refer to concrete and perceptually salient entities that are easier to learn (Gentner, 1982). However, cross-linguistic research suggests that the balance between nouns and verbs may vary considerably across languages and may be influenced by language-specific characteristics (Hardegger et al., 2024; Seifart, 2011; Stoll et al., 2012). For example, Seifart (2011) reported that languages with more extensive argument marking on verbs tend to exhibit lower NVRs in children. As Kazakh encodes substantial grammatical information through verbal morphology and also allows omission of overt subjects, it may also be expected to exhibit low NVR values.

To examine these patterns, CLAN’s *freq* command was applied to obtain counts of nouns and verbs across recordings. Proper nouns were included with nouns, and auxiliary forms were included with verbs. To provide a more complete picture, NVR was calculated both on the basis of tokens and types. Token-based NVR reflects the relative frequency with which nouns and verbs were produced, whereas type-based NVR reflects the relative distribution of unique noun and verb lexical items, counting each item only once regardless

of how often it was produced. Table 6 presents the resulting counts and NVR values, while Figure 2 plots token-based NVR values against age in months.

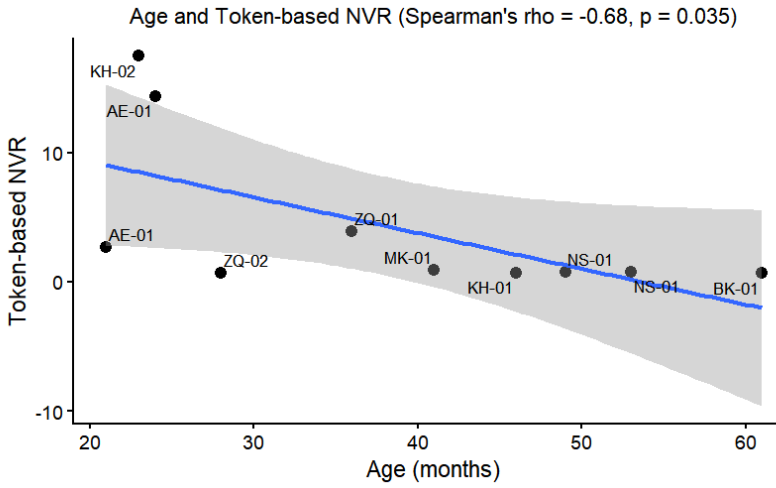
Table 6

Noun and Verb Token Counts and NVR by Child and Age

Child ID	Age	Nouns (tokens)	Verbs (tokens)	NVR (token-based)	Nouns (types)	Verbs (types)	NVR (type-based)
AE-01	1;09	16	6	2.67	6	4	1.50
KH-02	1;11	35	2	17.50	16	2	8.00
AE-01	2;00	144	10	14.40	33	4	8.25
ZQ-02	2;04	90	133	0.68	43	48	0.90
ZQ-01	3;00	105	27	3.89	41	10	4.10
MK-01	3;05	224	252	0.89	57	51	1.12
KH-01	3;10	130	182	0.71	62	42	1.48
NS-01	4;01	224	291	0.77	86	82	1.05
NS-01	4;05	259	325	0.80	80	72	1.11
BK-01	5;01	139	208	0.67	82	57	1.44

Figure 2

Relationship between Age and Token-based NVR



A Spearman's rank correlation revealed a significant negative relationship between age and token-based NVR ($\rho = -.68$, $p = .035$), indicating that the relative frequency of verbs increased with age. Among the youngest children aged 1;09-2;00, NVR values ranged from

2.67 to 17.50, reflecting a particularly strong predominance of nouns over verbs, a pattern consistent with Gentner (1982). From age 2;04 onwards, NVR values clustered around 0.70-0.90 (except the value of 3.89 at 3;00), suggesting a more balanced distribution between nouns and verbs, with verbs occurring slightly more frequently. Type-based NVR values provided a complementary perspective on the token-based results. Although the token-based analysis indicated that verbs were used more frequently than nouns as age increased, type-based NVR values at later ages generally remained above 1.0, suggesting that older children produced more unique noun types than verb types. This pattern indicates that the lower token-based NVR values were not necessarily driven by a larger verb vocabulary, but rather by more frequent use of a smaller set of verbs. At the same time, the youngest children (1;11 and 2;00) exhibited substantially higher NVR values in both the token- and type-based analyses, indicating that the strong noun bias observed at the earliest ages cannot be explained solely by repetition. Overall, NVR findings suggest a developmental shift from strong noun dominance in the youngest children towards increasingly frequent verb use with age.

3.1.2. Measure of Textual and Lexical Diversity (MTLD)

As children grow older, their vocabularies are expected to expand (Owens, 2020). To capture these developmental changes in lexical diversity (LD), many studies have traditionally relied on the use of TTR (Templin, 1957), likely due to its simple interpretation and ease of calculation. However, this measure of LD has been argued by many to be inherently flawed. For instance, Yang et al. (2022) point out that TTR scores are significantly impacted by sample size, where a low number of short utterances would yield higher TTR scores compared to a sample consisting of many long utterances, which would force speakers to repeat themselves (especially grammatical words). Therefore, it became clear that more sophisticated measures of LD were needed.

McCarthy and Jarvis (2010) proposed MTLD, an index that estimates the mean length of sequential word strings maintaining a predetermined TTR threshold (commonly .720). Unlike TTR, MTLD exhibits many advantages. First and foremost, it does not vary as a function of sample length; second, it performs well across all four types of validity when compared to other LD measures; and third, it captures unique aspects of LD that other indices cannot reliably measure (McCarthy & Jarvis, 2010). The result is a single score, where higher values reflect greater LD. Hence, the current project decided to use MTLD to measure LD in Kazakh-speaking children. A Python script for MTLD calculation compatible with CHAT

transcriptions was developed and made available through the project's GitHub repository. Due to the specifics of the transcription refinement process, MTLD was calculated using lemmas only for nouns, proper nouns, verbs, and auxiliary verbs; however, this restricted lexical set was assumed to still provide a meaningful estimate of LD in child speech. Following the recommendations of McCarthy and Jarvis (2010), recordings containing fewer than 100 tokens were excluded from the MTLD analysis, as estimates based on smaller samples may be unreliable. Consequently, the two youngest children, AE-01 (1;09; 22 tokens) and KH-02 (1;11; 37 tokens), were excluded from the MTLD analysis, and the results reported below therefore cover the age range from 2;00 to 5;01. The resulting MTLD values are presented in Table 7, while Figure 3 plots MTLD across age in months.

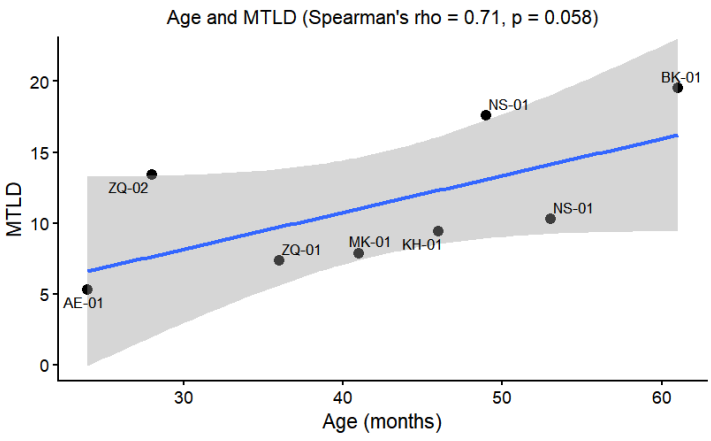
Table 7

MTLD Values by Child and Age

Child ID	Age	MTLD
AE-01	2;00	4.28
ZQ-02	2;04	13.40
ZQ-01	3;00	7.36
MK-01	3;05	7.89
KH-01	3;10	9.41
NS-01	4;01	17.59
NS-01	4;05	10.27
BK-01	5;01	19.53

Figure 3

Relationship between Age and MTLD



Across the dataset, MTLD values ranged from 4.28 at age 2;00 to 19.53 at age 5;01. A Spearman's correlation revealed a positive association between age and MTLD ($\rho = .71$), suggesting that LD tended to increase with age, which should be consistent with Owens's (2020) descriptions of vocabulary growth in child language. However, this relationship did not reach statistical significance ($p = .058$), indicating that the observed trend should be interpreted with caution. Indeed, further inspection of the individual scores indicates considerable variability across recordings. For example, the child aged 2;04 produced an MTLD score of 13.40, exceeding the values observed in all three-year-old children in the sample. Similarly, the recording at age 4;05 yielded a markedly lower score (10.27) than the earlier recording of the same child at age 4;01 (17.59). These fluctuations can suggest that MTLD did not increase uniformly across the present dataset. The lack of statistical significance may also partly reflect the small number of observations and the methodological decision to calculate MTLD using only nouns, proper nouns, verbs, and auxiliary verbs rather than the full lexical repertoire. To sum up, the analysis revealed a positive but non-significant association between age and MTLD, with substantial variability observed across recordings.

3.1.3. Lexical Frequency Patterns

To complement the quantitative MTLD analysis, CLAN's *freq* command was used to qualitatively examine lexical frequency patterns in the corpus. While MTLD provides an overall estimate of LD, it does not reveal which lexical items contribute to developmental changes observed across age groups. Therefore, frequency analysis was conducted to identify the most commonly produced nouns and verbs at different ages by extracting the three most frequent lexical items within each word class for each child, providing an initial qualitative perspective on how children's lexical repertoires may evolve over time. Previous research suggests that children's lexical repertoires develop not only in size but also in composition, with older children gradually incorporating words from a broader range of semantic domains, including mental state vocabulary (Papafragou et al., 2007; Šmit Brleković & Kuvač Kraljević, 2023). Although frequency counts alone cannot directly capture these developmental processes or fully explain patterns of LD, they provide a useful preliminary overview of the lexical items that are most salient in children's speech at different ages. Table 8 presents the three most frequently produced nouns and verbs for each child.

Table 8

Three Most Frequent Nouns and Verbs by Child and Age

Child ID	Age	Top 3 Nouns	Top 3 Verbs
AE-01	1;09	апа (8), папа (3), су (2)	хотеть (3), кел/жат/тұр (1)
KH-02	1;11	папа (7), аю (5), тәте (4)	де/тақ (1)
AE-01	2;00	мама (28), телефон (23), нан, машина (12)	аш (4), хотеть (3), жаз (2)
ZQ-02	2;04	әке (6), гүл/Мирас (5)	бер (12), жаса (11), кес (9)
ZQ-01	3;00	ата/мама (10), автобус (8)	кел (7), бар (5), бер/жүр/біл (3)
MK-01	3;05	билет (54), мама (42), әке (13)	бер (31), бол (18), ал (17)
KH-01	3;10	мама (15), Айкөркем (9), күн (7)	біл (24), айт (21), боя (20)
NS-01	4;01	мама (29), бас (10), кез (9)	бол (26), кет (24), қал (16)
NS-01	4;05	мама (38), Веном (21), бауырсақ (12)	де (57), бол (32), ал (23)
BK-01	5;01	жер (7), рет/уақыт (6)	ал (18), бол (12), кел (10)

Note. Айкөркем and *Мирас* are pseudonyms, and they do not reflect the exact proper nouns produced by the child.

Looking first at nouns, kinship and family-related terms, such as *мама* (‘mother’) and *папа* (‘father’), were frequent in children’s speech across all age groups. Notably, these are Russian-derived forms, and although the Kazakh form *әке* (‘father’) also occurred, alternative Kazakh equivalents for ‘mother’, such as *ана* and *шеше*, were rare in the present data. Younger children more frequently used nouns referring to immediate physical objects and everyday experiences, such as *телефон* (‘phone’), *нан* (‘bread’), and *машина* (‘car’). In contrast, older children more frequently produced abstract or time-related nouns, including *жер* (‘place’), *кез* (‘turn/occasion’), *уақыт* (‘time’), and *күн* (‘day’). Overall, the noun frequency patterns may suggest that while family-related vocabulary is common regardless of age, younger children tend to rely more on words related to their immediate environment, whereas older children use a wider range of abstract nouns. This pattern is broadly consistent with previous findings suggesting that older preschool children increasingly use words referring to more abstract concepts and mental representations (Šmit Brleković & Kuvač Kraljević, 2023).

Turning to verbs, highly frequent items such as *де* (‘say’), *бол* (‘be’), and *бер* (‘give’) were common across all age groups, suggesting that children may rely on a shared set of everyday verbs regardless of age. Younger children more frequently used action verbs referring to immediate activities and interactions, such as *аш* (‘open’) and *кес* (‘cut’). In contrast, older children more frequently produced auxiliary verbs, such as *ал* (‘ability’), *қал* (‘unexpected or resultant state’), and *кет* (‘directional shift’), often in multi-verb constructions. Older children also appeared to use more cognitively oriented verbs, such as *біл* (‘know’), which were absent from younger children’s most frequent verb lists. Notably, the frequent use of this particular verb in older children aligns with Papafragou et al.’s (2007) observation that mental state verbs such as ‘know’ and ‘think’ remain challenging for children until approximately four years of age. Overall, the verbal frequency patterns suggest that while children of all ages may rely on a core set of common verbs, younger children more frequently use verbs related to immediate actions, whereas older children show more frequent use of auxiliary and cognitive verbs.

3.2. Morphology

3.2.1. Mean Length of Utterance in Morphemes (MLU-m)

MLU (Brown, 1973) is one of the most widely used measures of grammatical development in child language research. It is defined as the average length of children’s utterances and is commonly calculated in morphemes (MLU-m), words (MLU-w), or syllables (MLU-s). In agglutinative languages, such as Kazakh, where grammatical information is often expressed through extensive suffixation, MLU-m seems to be the most appropriate measure because it captures internal morphological complexity that may not be reflected in word- or syllable-based counts. Consistent with this view, research on other agglutinative languages, including Georgian, Turkish, and Inuktitut, suggests that MLU-m is more sensitive to grammatical development in such systems than MLU-w or MLU-s (Allen, 2024; Tchintcharauli et al., 2024). Therefore, MLU-m was calculated in this study.

The following procedure was used to calculate MLU-m in the present dataset. To begin with, consistent with recommendations regarding sample size (Matiasovitsová et al., 2024), a sample of 100 utterances was analyzed for each child, with the exception of one recording (1;11), for which only 75 eligible utterances were available. MLU-m was calculated manually because preliminary testing showed that CLAN did not consistently count case markers as separate morphemes in Kazakh, raising concerns about its ability to accurately capture more complex morphological structures. Finally, several adaptations were

made to account for characteristics of spontaneous child speech, following the general principles outlined by Brown (1973). In particular, repetitions within the same utterance were counted only once, fully unintelligible utterances were excluded, and partially intelligible utterances were included based on identifiable linguistic material. Counting sequences and similar list-like productions, which were frequent in some recordings, were also excluded from MLU-m calculations. The resulting MLU-m values are presented in Table 9, while Figure 4 illustrates the relationship between MLU-m and age.

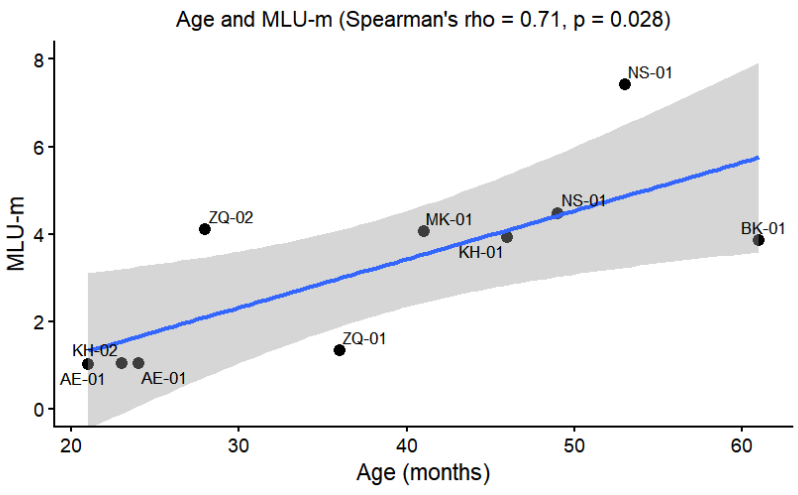
Table 9

MLU-m Values by Child and Age

Child ID	Age	MLU-m
AE-01	1;09	1.03
KH-02	1;11	1.05
AE-01	2;00	1.04
ZQ-02	2;04	4.11
ZQ-01	3;00	1.34
MK-01	3;05	4.05
KH-01	3;10	3.92
NS-01	4;01	4.48
NS-01	4;05	7.42
BK-01	5;01	3.86

Figure 4

Relationship Between Age and MLU-m



Across the dataset, MLU-m values generally increased with age, ranging from very low values in the youngest children (e.g., 1.03 at 1;09) to considerably higher values in older children (e.g., 7.42 at 4;05). A Spearman's correlation analysis confirmed a strong, statistically significant positive relationship between age and MLU-m ($\rho = 0.71$, $p = .0028$), suggesting that morphosyntactic complexity tends to increase with age in the present dataset. At the same time, considerable individual variation was observed. For example, one of the two-year-old participants (2;04) exhibited a relatively high MLU-m (4.11), whereas one of the three-year-old participants (3;00) showed a comparatively low MLU-m (1.34), producing mainly single-word utterances or short two-word combinations similar to those observed in younger children.

The findings also highlight the influence of interactional context on MLU-m values. The highest MLU-m in the dataset (7.42 at 4;05) was associated with highly engaged, narrative-like speech, in which the child discussed preferred topics such as superheroes and computer games. By contrast, the five-year-old participant (5;01) showed a lower MLU-m (3.86), likely reflecting the turn-based guessing game context of the recording, which elicited single-word guesses and brief questions. In conclusion, these findings suggest that MLU-m values generally increased with age in the present dataset, while also appearing to be sensitive to individual differences and interactional context.

3.2.2. Emergence of Nominal and Verbal Morphology

Following an approach inspired by Tchintcharauli et al. (2024), this section examines the emergence of morphosyntactic features in Kazakh child speech by identifying the earliest age at which specific morphological categories are attested in the present dataset. While measures such as MLU-m provide a general indication of increasing morphological complexity, this analysis can offer a more detailed perspective on which nominal and verbal features appear over the course of development. The analysis focused on overtly marked morphology, and only morphological categories explicitly realized in Kazakh forms were considered.

3.2.2.1. Nominal Morphology

For nouns, four major feature domains were analyzed: case, number, possession, and person. Because the analysis focused on overtly marked morphology, unmarked categories such as nominative case and singular number were not included. Thus, case marking included genitive (*Gen*), dative (*Dat*), accusative (*Acc*), locative (*Loc*), ablative (*Abl*), and instrumental

(*Ins*) cases, while number is represented only by plural marking (*Plur*). Possessive morphology (*тәуелдік жалғау*) includes first, second, and third person singular (*S1, S2, S3*), as well as first person plural (*P1*), while plural forms of second and third person were not treated as separate categories, as they are formed by combining the singular possessive suffix with the regular plural marker rather than introducing distinct morphological forms. Finally, person marking (*жіктік жалғау*) was tracked for first and second person (*1, 2*). Table 10 presents the earliest attested examples of these features across age groups. The horizontal axis represents the ages of the children, while the vertical axis lists the feature categories. Each cell contains the first attested example of the corresponding feature, including the original Kazakh form, a simplified morpheme-based gloss (with hyphens marking only the relevant suffix), and an approximate English translation.

Table 10

Earliest Attested Nominal Morphological Features by Age

Feature/age	1;09	1;11	2;00	2;04	3;00	3;05	3;10	4;01	4;05	5;01
Case				**Gen** гүл-**дің** “of a flower” **Acc** ана-**ны** “mother” **Dat** ана-**ға** “to mother” **Loc** үй-**де** “at/in the house” **Abl** жұмыртқа-**дан** “from an egg” **Ins** миы-**мен** “with a brain”						
Number				**Plur** жапырақ-**тар**-ы “leaves (e.g., of a flower)”						
Possession (Тәуелдік жалғау)				**S1** іш-**ім** “my stomach” **S2** үй-**ің**-де “at/in your house” **S3** қол-**ы** “his/her/its hand”		**P1** Комната-**мыз**-ға “to our room”				
Person (Жіктік жалғау)				**2** қызың-**сың** “you are his/her/its girl/daughter”			**1** Қазақ-**пын** “I am Kazakh”			

No overtly marked nominal features were attested in the recordings from children aged 1;09 to 2;00. A clear shift occurs at 2;04, where a wide range of nominal features is observed for the first time, including multiple case categories, plural marking, and singular possessive suffixes (*S1-S3*), along with the first instance of person marking (*2*). At later ages, only a small number of additional features appear: plural possessive marking (*P1*) at 3;05 and the remaining person-related form (*1*) at 3;10. These observations suggest that a broad range of nominal features appears within a relatively short developmental window in the present dataset. It is possible that some of these features emerged earlier but were not captured due to the cross-sectional nature and sampling density of the data. Nevertheless, within the present dataset, age 2;04 appears to represent an important point at which a broad range of nominal morphological categories was observed.

This pattern can be compared with existing descriptions of Kazakh child language development provided by the National Scientific and Practical Center for Special Education (2019). According to their report, early speech up to approximately 1;06 is characterized by single-word utterances, with children relying mainly on isolated lexical items and context-dependent meanings. This aligns with the present findings, where the youngest children (1;09-2;00) produce predominantly uninflected nouns in the nominative case. The same source reports a shift at around 2;06-3;00, when children's speech becomes more structurally complex. A similar pattern is observed here, with a substantial increase in nominal morphology at 2;04, including the emergence of case marking, plural forms, and possessive suffixes. Overall, the findings on the emergence of nominal morphology in the present dataset are broadly consistent with the existing description of the FLA of Kazakh.

3.2.2.2. Verbal Morphology

This subsection examines the emergence of verbal morphology in Kazakh child speech using the same procedure as described for nominal features. The analysis included the following verbal feature domains: finiteness (finite [*Fin*], converb [*Conv*], participle [*Part*], verbal noun [*Vnoun*]); mood (indicative [*Ind*], imperative [*Imp*], optative [*Opt*], conditional [*Cond*]); voice (causative [*Cau*], reciprocal [*Rcp*], reflexive [*Refl*], passive [*Pass*]); tense (present [*Pres*], past [*Past*]); person (*S1, S2, S3, P1, P2*); politeness (*Pol*); negation (*Neg*); evidentiality (*Evid*); and auxiliary constructions (*Aux*). As for auxiliary constructions, only the first attested instance was included, as the extensive range of auxiliary combinations in Kazakh falls beyond the scope of the present analysis. As in the nominal analysis, Table 11

presents the earliest attested examples of these verbal morphological categories across age groups.

Table 11

Earliest Attested Verbal Morphological Features by Age

Feature/age	1;09	1;11	2;00	2;04	3;00	3;05	3;10	4;01	4;05	5;01
Finiteness				**Conv** іш-**е** “while drinking” **Part** ұна-**йтын** “which someone likes” **Vnoun** кес-**у**-ім “my cutting”						
Mood				**Ind** шаштым “I scattered” **Cnd** жасап бол-**са** “once it’s done”				**Opt** үлкей-**гі**-м келеді “I want to become bigger”		
Voice				**Cau** Бат-**ыр**-ып “while making something sink” **Pass** бұз-**ыл**-ып “having been broken” **Refl** жина-**л**-амыз “gather ourselves”				**Rcp** Төбеле-**с**-еді “they fight each other”		
Tense				**Pres** (*Kazakh*) ки-**е**-мін “I wear” **Past** де-**ді**-м “I said”						
Person				**S1** (*Kazakh*) бара-**мын** “I go” **S2** жасап болды-**ң** “you finished doing it” **3** бара-**ды** “he/she/it goes” **P1** бара-**мыз** “we go”		**P2** кетті-**ңдер** “you (plural) left”				
Politeness				**Pol** бер-**ші** “please give”		**Form** кетпе-**ңіз** “don’t go (please)”				

Negation				**Neg** (*Kazakh*) біл-**ме**-ймін “I don’t know”						
Evidentiality						**Evid** екен “it turns out”				
Auxiliary				**Aux** істей **алады** “he/she/it/they can do”						

The table suggests a pattern broadly similar to that observed for nominal morphology. First, no overtly marked verbal features were attested in the recordings from ages 1;09 to 2;00. Then, at age 2;04, a broad range of verbal features was observed for the first time, including both finite and non-finite forms (converbs, participles, and verbal nouns), as well as multiple grammatical categories such as indicative and conditional mood, several voice distinctions (causative, passive, reflexive), present and past tense, and a range of person markers. Negation and auxiliary constructions were also attested at this stage, suggesting the emergence of more complex predicate structures. Subsequent development was characterized by the appearance of more specialized categories. At 3;05, evidential marking (e.g., *екен* ‘it turns out’), formal politeness marking (e.g., *-ңыз*), and second person plural (*P2*) forms were first attested. These categories may reflect children’s increasing ability to encode the source of information and adjust speech according to social context with age. At 4;01, additional features such as optative constructions and reciprocal voice were observed, involving more complex relations between participants and actions. As with nominal morphology, these observations should be interpreted cautiously, as some categories may have emerged earlier but were not captured due to the sampling density of the data.

The observed pattern of verbal morphological emergence can be compared with the developmental stages described by the National Scientific and Practical Center for Special Education (2019). The report suggests that verb use increases substantially during the third year of life, as children begin to coordinate actions in interaction and accompany their activities with speech. This description is broadly consistent with the present dataset. In particular, no overtly marked verbal features were attested in the three youngest children (1;09-2;00), a pattern that also aligns with the relatively low number of verbs produced at these ages (see Section 3.1.1 and Table 6). At 2;04, which falls within the third year of life

mentioned in the report, a wide range of verbal features became attested, including multiple tense, mood, and voice distinctions, as well as auxiliary constructions and non-finite forms. Furthermore, the report describes the period between approximately 3;00 and 4;00 as a stage of active grammatical development, during which children rapidly expand their use of morphological forms. In the present dataset, this stage was reflected in the emergence of more specialized categories, such as evidentiality, politeness marking, and optative constructions. Overall, as was the case for nominal features, these findings suggest that the broad trajectory observed in the present dataset is consistent with existing descriptions of Kazakh language acquisition.

3.3. Semantics

3.3.1. Semantic Field Distribution of Nominal and Onomatopoeic Forms

This section examines the distribution of semantic fields in Kazakh child speech, focusing on nouns and onomatopoeic forms. Following the general approach outlined in the ASP framework (Defina et al., 2023b), these forms were grouped into broader semantic fields to identify patterns in the types of meanings expressed in early language. Previous research suggests that early vocabularies are not random but instead show systematic patterns in the meanings they contain. In particular, studies have shown that children's early lexicons are dominated by nouns referring to concrete entities, such as people, animals, food, and everyday objects, which are linked to the child's immediate environment and are easier to map onto linguistic forms (Gentner, 1982). At the same time, older children gradually acquire vocabulary referring to less concrete meanings, including words related to mental states, time, and other abstract concepts (Šmit Brleković & Kuvač Kraljević, 2023). Examining the distribution of semantic fields across age groups may therefore provide insight into how children's semantic repertoires develop over time.

For the present analysis, semantic fields were adapted from categories proposed in the MacArthur-Bates Communicative Development Inventories (CDI; Fenson et al., 2006). Because the analysis focused exclusively on nouns and onomatopoeic forms, only CDI categories relevant to these lexical classes were retained. The initial set of semantic fields included *onomatopoeia/sounds*, *animals*, *vehicles*, *toys*, *food/drink*, *clothing*, *body parts*, *furniture/rooms*, *household items*, *places/locations*, *people*, and *time*. Kinship terms, which are sometimes treated as a separate category, were included within the broader category of *people*. While the initial set of semantic fields was based on established CDI categories, manual classification of lexical items in the dataset revealed the need for further refinement.

As a result, several additional categories were introduced to better group the observed lexical items, including *nature*, *activities*, and *descriptive/abstract* terms. All lexical items were manually assigned to semantic fields based on their predominant semantic interpretation using this adapted classification scheme. For clarity, Appendix B presents all semantic field categories used in the analysis, together with their definitions and representative examples.

To visualize the distribution of these semantic fields across age groups, a heatmap was constructed (Figure 5), allowing for easier comparison of relative proportions across categories and developmental stages. Table 12 presents the total number of nouns and onomatopoeic forms produced by each child that were used to calculate the percentages shown in the heatmap.

Figure 5

Semantic Field Distribution in Child Speech by Age

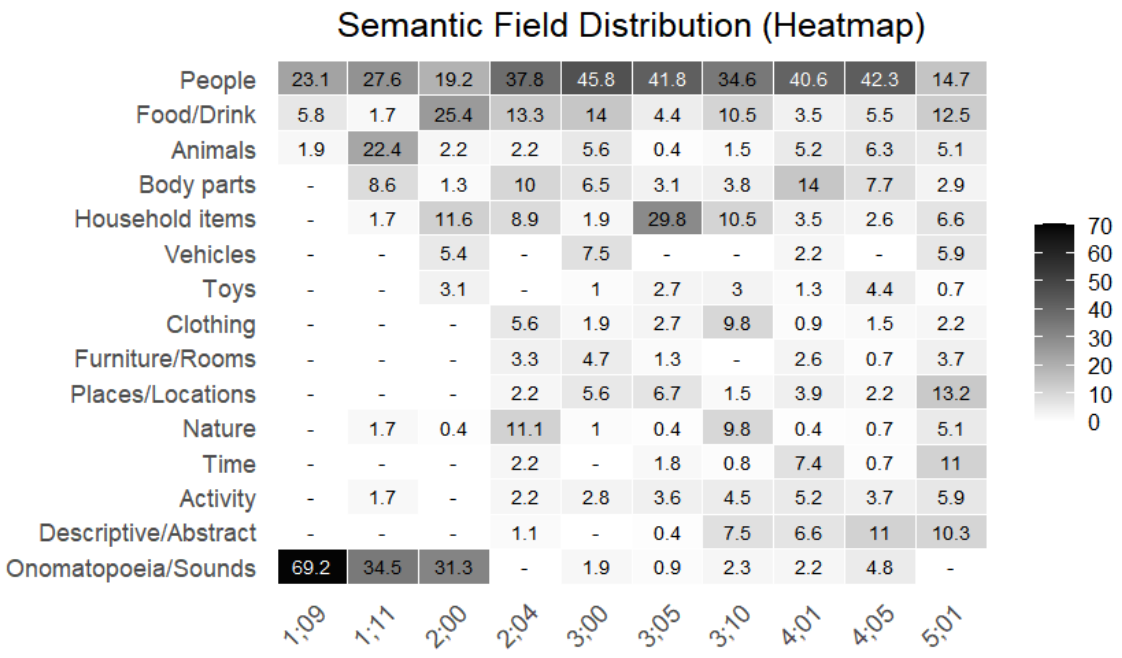


Table 12

Total Number of Nouns and Onomatopoeia per Child and Age

Child ID	Age	Total Number of Tokens
AE-01	1;09	52
KH-02	1;11	58
AE-01	2;00	224
ZQ-02	2;04	90
ZQ-01	3;00	107
MK-01	3;05	225
KH-01	3;10	133
NS-01	4;01	229
NS-01	4;05	272
BK-01	5;01	136

The distribution of semantic fields across age groups suggested several developmental tendencies. Most notably, the youngest children relied heavily on onomatopoeic forms. *Onomatopoeia/sounds* accounted for a substantial proportion of lexical items at the earliest ages (69.2% at 1;09, 34.5% at 1;11, and 31.3% at 2;00), but decreased markedly across development and became virtually absent in older children. Other categories associated with children's immediate environment were also relatively prominent at younger ages, including *animals* (22.4% at 1;11) and *household items*, which remained relatively frequent during the early preschool years and reached a peak of 29.8% at 3;05. The observed patterns can suggest that early semantic development is oriented toward highly salient and concrete aspects of children's everyday experiences, in line with Gentner's (1982) account of early vocabularies.

At the same time, older children appeared to use a broader range of semantic categories. In particular, *descriptive/abstract* terms emerged from approximately age 3;10 onwards and became increasingly represented at later ages (e.g., 11.0% at 4;05 and 10.3% at 5;01). Similarly, vocabularies related to *activities* and *time* were largely absent in younger children but became more consistently represented from around age 4 onwards (e.g., 7.4% at 4;01 and 11.0% at 5;01 for *time*). These findings are broadly consistent with previous descriptions suggesting that children's vocabularies gradually expand toward more abstract

meanings that extend beyond the immediate environment (Šmit Brleković & Kuvač Kraljević, 2023).

Some semantic fields appeared relatively stable across all ages. It included such categories as *people*, *food/drink*, and *body parts*, suggesting that references to caregivers, family members, and the child's own body form a consistent component of children's vocabularies across all age groups. Finally, several other semantic fields remained relatively infrequent throughout the dataset. These included categories such as *vehicles*, *toys*, *nature*, *clothing*, and *furniture/rooms*. Overall, the findings from the semantic fields distribution analysis suggest that semantic development in Kazakh child speech may involve a gradual expansion from early reliance on highly salient meanings related to children's immediate environment toward a broader semantic repertoire that includes abstract and time-related concepts.

3.3.2. Temporal Reference in Verbal Production

This section examines how temporal reference is expressed in Kazakh child speech by focusing on the use of non-past and past verbal forms across age groups. The development of temporal reference is an important aspect of language acquisition because it reflects children's ability to locate events in relation to the moment of speaking and has been linked to broader developments in both language and cognition (Zhang & Hudson, 2018). Previous research suggests that children's early verbal systems are centered on immediate, present-oriented situations, whereas reference to temporally displaced events develops later. For example, Bassano et al. (2004) note that verbal forms referring to present events are generally easier for young children to produce than forms referring to non-immediate temporal contexts. Similarly, Nelson (1989) reports that clearer contrasts between present and past meanings emerge only gradually during the second and third years of life. Based on these findings, a similar tendency may be expected in the present dataset.

To examine whether this pattern can be observed in Kazakh child speech, the distribution of non-past and past forms was analyzed across ages. Since no clearly distinct future forms were identified and Kazakh verbal morphology generally reflects a broader non-past versus past distinction, the analysis focused on these two categories. Tense-marked forms were identified through *freq* searches in CLAN targeting finite indicative verbs and auxiliaries annotated with the features *Pres* and *Past*. As no finite indicative verbal forms were attested at age 1;11, this datapoint was excluded from the visualization and statistical analysis. Table 13 presents the frequencies and proportions of non-past and past forms across

age groups, while Figure 6 illustrates the relationship between age and the proportion of past-marked verbal forms.

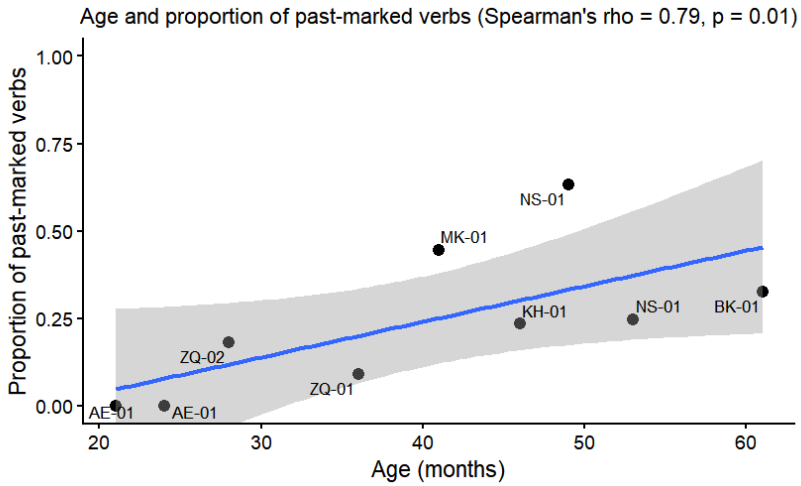
Table 13

Distribution of Present/Non-Past and Past-Marked Verbal Forms by Child and Age

Child ID	Age	Present/Non-Past Verbs	Past Verbs	Total Finite Verbs	Proportion Present	Proportion Past
AE-01	1;09	3	0	3	1.000	0.000
KH-02	1;11	0	0	0	-	-
AE-01	2;00	3	0	3	1.000	0.000
ZQ-02	2;04	59	13	72	0.819	0.181
ZQ-01	3;00	10	1	11	0.909	0.091
MK-01	3;05	72	58	130	0.554	0.446
KH-01	3;10	82	25	107	0.766	0.234
NS-01	4;01	56	96	152	0.368	0.632
NS-01	4;05	115	38	153	0.752	0.248
BK-01	5;01	62	30	92	0.674	0.326

Figure 6

Relationship between the Proportion of Past-Marked Verbal Forms and Age



The results reveal a gradual development in the expression of temporal reference. At ages 1;09 and 2;00, only present forms were attested, and no finite indicative forms occurred

at 1;11. However, the total number of verbs in these recordings is very low, making it difficult to draw firm conclusions about temporal preferences at this stage. From age 2;04 onward, however, a clearer pattern emerges. At this age, past-marked forms appeared for the first time, accounting for .181 of all finite tense-marked forms. Their proportion then substantially increased at later ages, reaching .446 at 3;05 and .632 at 4;01. Although some fluctuation is observed across older age groups, likely reflecting contextual variation in spontaneous speech, the overall trend suggests a growing use of past-marked forms with age. This tendency is supported by a Spearman's rank correlation analysis, which shows a strong, statistically significant positive relationship between age and the proportion of past-marked forms ($\rho = .79$, $p = .01$). Overall, these findings can suggest that the ability to express past temporal reference becomes more prominent over time in Kazakh child speech, broadly aligning with previous research on early verb development (Bassano et al., 2004; Nelson, 1989).

CHAPTER 4.
DISCUSSION

4.1. Interpretation of Linguistic Findings

One of the core objectives of the present study was to produce a preliminary acquisition sketch of Kazakh child language within the ASP framework. Although the findings are based on a relatively limited dataset, they provide an initial overview of developmental patterns in Kazakh child speech.

In the lexical domain, NVR and MTLD were used to examine changes in lexical composition and lexical diversity. First, the observed decrease in token-based NVR with age indicates a gradual shift away from strong noun dominance and towards increased verb use, a developmental pattern reported in previous FLA research (Gentner, 1982; Stoll et al., 2012). At the same time, the type-based analysis showed that older children generally continued to produce more unique noun types than verb types, suggesting that increased verb use was driven primarily by production frequency rather than by a larger verb vocabulary. Second, the positive association between MTLD and age may indicate a tendency toward greater lexical diversity with age, a pattern that would be expected based on Owens's (2020) account of vocabulary development. However, substantial variability was observed across recordings, and the relationship did not reach statistical significance. In addition to NVR and MTLD, a qualitative frequency analysis was also used to explore how vocabulary composition changed across ages. The analysis showed that younger children relied more heavily on nouns and verbs related to immediate objects and everyday actions, whereas older children demonstrated greater use of abstract nouns, auxiliary verbs, and cognitively oriented verbs such as *білл* ('know'), consistent with previous findings on the development of abstract vocabulary and mental state verbs in preschool children (Papafragou et al., 2007; Šmit Brleković & Kuvač Kraljević, 2023).

In the morphological domain, the analysis focused on MLU-m and the emergence of nominal and verbal morphological features. Consistent with previous studies on agglutinative languages, MLU-m was adopted in the present study, as morpheme-based measures are better suited to languages like Kazakh, where grammatical complexity is largely conveyed through suffixation (Allen, 2024; Tchintcharauli et al., 2024). In the present dataset, MLU-m values generally increased with age, suggesting a broader increase in grammatical complexity over development. At the same time, variation across participants indicates that morphosyntactic development is not uniform across children. The findings also suggest that interactional

context may further influence MLU-m values, with narrative-like interactions tending to elicit longer and more morphologically complex utterances than more structured activities.

Regarding the emergence of morphological features, the findings suggest a three-stage pattern in the present dataset. In the youngest children (1;09-2;00), no overtly marked nominal and verbal categories were attested. Around age 2;04, a substantial number of core nominal and verbal features became observable, including case marking, possessive morphology, tense marking, voice distinctions, and auxiliary constructions. At later ages, the emergence of more specialized categories, such as evidentiality, politeness marking, and optative constructions, suggests continued diversification of the morphological system. This general pattern broadly corresponds to the developmental stages described by the National Scientific and Practical Center for Special Education (2019), which characterizes the third year of life as a period of increasing grammatical complexity and expanding verb use. However, given the cross-sectional nature and limited size of the dataset, these observations should be interpreted as preliminary tendencies rather than evidence for the timing of morphological acquisition in Kazakh-speaking children more generally.

In the semantic domain, the analysis first focused on semantic field distribution in nouns and onomatopoeic forms. The results revealed differences in the distribution of semantic categories across age groups, consistent with previous research indicating that children's early vocabularies are dominated by perceptually salient and concrete categories (Gentner, 1982), with more abstract meanings emerging later (Šmit Brleković & Kuvač Kraljević, 2023). In the present dataset, early lexical production was characterized by a high proportion of *onomatopoeia*, alongside categories closely related to children's immediate environment, including *animals* and *household items*. By contrast, older children showed evidence of expanding semantic repertoires, with categories such as *descriptive/abstract* and *time* becoming more prominent from approximately age 3;10 onwards. At the same time, categories such as *people*, *food/drink*, and *body parts* remained consistently high across all age groups, suggesting that some references constitute a stable component of early vocabulary in all children. To sum up, these findings may suggest a developmental shift from highly salient meanings toward a broader range of semantic concepts in the present dataset.

Finally, the analysis examined developmental patterns in temporal reference expressed through verbal morphology. Although the earliest stages were difficult to interpret due to the limited number of finite verbs produced, a clearer pattern emerged from age 2;04 onwards, when past-marked forms became consistently attested in the dataset. Their proportion generally increased across age groups, a tendency further supported by a

significant positive correlation between age and the proportion of past-marked forms. This pattern is broadly consistent with previous research suggesting that children's early verbal systems are primarily centered on present-oriented meanings, whereas reference to temporally displaced events emerges later (Bassano et al., 2004; Nelson, 1989).

4.2. Corpus Construction and Methodological Considerations

Apart from the linguistic findings, the creation of Balausa itself constitutes an important contribution of the present study. The corpus currently contains nearly five hours of transcribed and morphologically annotated caregiver-child interactions, representing over 3,600 child utterances and more than 4,100 adult utterances. Considering that the corpus was developed as part of an individual Master's thesis project, this represents a substantial body of naturalistic speech data. This contribution is particularly important given the limited representation of Kazakh in child language research and the lack of publicly available corpora of Kazakh child speech. Additionally, preparing the data using the internationally recognized CHILDES-compatible CHAT format aligns the Balausa corpus with established international standards in child language research.

Several methodological choices shaped the final design of this corpus. First, the corpus relied on naturalistic caregiver-child interaction recorded in home settings. This offers strong ecological validity, as it captures spontaneous speech in familiar contexts and provides access to everyday patterns of language use that may be less visible in structured tasks. Second, the dataset was primarily cross-sectional, allowing multiple age groups to be examined within a limited timeframe. This made it possible to observe broad developmental tendencies across early childhood and to include a wider age range than would likely have been feasible in a longitudinal design of similar scope. Third, the study relied on a combination of automated tools and manual verification. In particular, the BA2 pipeline helped reduce the amount of manual work required during corpus preparation, while manual correction helped ensure the quality of the final transcripts and annotations. In addition, because the UD framework did not adequately distinguish certain categories relevant to Kazakh morphology, several language-specific analytical decisions were required, including the development of modified conventions for annotating possessive and personal endings. Finally, the linguistic scope of the study was intentionally narrowed, with detailed verification focused mainly on nouns and verbs in order to maintain a focused and manageable analytical scope.

4.3. Limitations and Future Directions

Several limitations should be considered when interpreting the findings. First, the corpus remains relatively small, and the distribution of recordings across age groups is uneven. Second, the sample was not fully gender-balanced, as younger age ranges included mostly female participants, whereas older age ranges included mostly male participants. Third, although participants were recruited from Kazakh-speaking households, the extent of children's exposure to Russian was not systematically quantified. Furthermore, although the naturalistic design offers strong ecological validity, it also introduced variability in recording quality and interactional context. As a result, background noise and spontaneous overlap sometimes affected ASR performance and increased the need for manual correction, and although extensive verification was carried out, some transcription or alignment inaccuracies may remain. Furthermore, detailed verification of the morphological annotation was restricted mainly to nouns and verbs, limiting the range of linguistic features that could be examined in depth. Finally, the grammatical tier (*%gra*) was not included in the corpus, and therefore syntactic development was not systematically assessed.

These limitations point to several directions for future research, including corpus expansion through a larger and more balanced sample, more systematic assessment of children's exposure to languages other than Kazakh, and the inclusion of longitudinal recordings. More controlled recording conditions would also improve transcription quality and reduce manual correction time. In addition, extending detailed annotation to other word classes would allow more comprehensive analyses of children's lexical and grammatical development. The study also highlights the need for stronger computational resources for Kazakh. Future work may therefore improve automatic tagging, annotation guidelines, and ASR systems trained specifically on Kazakh and child speech. Finally, the Balausa corpus may itself serve as a starting point for further linguistic research, including cross-linguistic ASP comparisons and future work on the relationship between caregiver input and child production.

CHAPTER 5.
CONCLUSION

The present study aimed to address a clear empirical gap in research on Kazakh language acquisition by combining corpus construction with the development of an acquisition sketch within the ASP framework. To achieve this, the study developed the Balausa corpus, the first publicly available corpora of Kazakh child speech in a CHILDES-compatible format. The corpus comprises approximately 284 minutes of transcribed and morphologically annotated caregiver-child interactions, including 3,621 child utterances (approximately 7,005 tokens) and 4,177 adult utterances (approximately 10,780 tokens). Using this data, the study provided an initial description of early Kazakh language development across the lexical, morphological, and semantic domains.

The acquisition sketch identified several broad developmental tendencies across the examined domains. In the lexical domain, younger children's speech appeared to be strongly noun-dominated, with very limited verb use, followed by a gradual increase in verb production and a more balanced noun-to-verb distribution over time. In the morphological domain, MLU-m values generally increased with age, suggesting increasing morphosyntactic complexity in the present dataset. In addition, many nominal and verbal categories were first observed around age 2;04. In the semantic domain, children's speech appeared to shift from context-bound and perceptually salient expressions toward a broader range of meanings, including more abstract concepts. Finally, the proportion of past-marked forms increased with age, potentially reflecting children's increasing ability to refer to temporally displaced events. The developmental tendencies observed in the present dataset appear to be broadly consistent with patterns reported in the wider FLA literature and limited existing descriptions of Kazakh language acquisition.

Beyond its findings on language development, the study also contributes a reusable corpus resource. By following standardized transcription and annotation conventions, the Balausa corpus is suitable not only for the present analyses but also for future reuse, expansion, and cross-linguistic comparison. In this way, the project contributes new empirical material for a language that remains underrepresented in acquisition research and provides a foundation for future work on Kazakh child language development. At the same time, the relatively small dataset and focused scope of annotation limit the extent to which the findings can be generalized. Consequently, the patterns identified here are best interpreted as preliminary tendencies rather than definitive developmental norms. Future research can build

on this foundation by expanding the range of recordings, extending annotation coverage, and further investigating developmental patterns in Kazakh-speaking children.

SOCIAL IMPACT OF THE STUDY

Creating social impact was a driving force behind this project from the very beginning. The ways in which this study may benefit society can be viewed from several different angles. First, this study contributes to understanding developmental patterns in Kazakh-speaking children by providing initial measures of lexical, grammatical, and semantic development. This is especially relevant given that no other study on Kazakh has reported these measures before, and most existing developmental knowledge comes from widely studied Indo-European languages. Thus, this study may help challenge the assumption that developmental norms identified in these languages automatically apply to Kazakh-speaking children. Furthermore, establishing typical patterns of Kazakh language development is an important first step toward identifying atypical development, potentially contributing to the creation of evidence-based assessment tools for Kazakh-speaking children. In the longer term, such knowledge may help parents and caregivers better understand the broad range of typical language development in their children.

This study also has a direct impact on language documentation and preservation. As discussed throughout the thesis, despite having millions of native speakers, Kazakh remains underrepresented in child language research. Once published, Balausa will leave a lasting digital record and ensure that Kazakh child speech is documented and represented in international research resources, allowing future generations of researchers to build upon it. The corpus may also contribute to the development of future ASR and natural language processing tools for Kazakh, which could, in turn, facilitate the creation of additional Kazakh corpora and help reduce barriers to the digital documentation and preservation of Kazakh.

REFERENCES

Abdyeva, R. S., Nogaybaeva, J. A., Togzhanova, L. K., & Shynybaeva, A. S. (2015). Kazakh language acquisition and translation from a foreigner's perspective: A case study with English-speaking residents. *Sociological Science, 3*. https://science-sd.com/462-24896

Akhmet, A., Akimbekova, G., & Aliaskar, A. (2023). Otrazhenie kazahsko-russkogo bilingvizma v detskoy rechi v obrazovanii (po dannym sociolingvisticheskogo oprosa) [Reflection of Kazakh-Russian bilingualism in children's speech in education (based on sociolinguistic survey data)]. *Vestnik KazNU. Seriya Filologicheskaya, 190*(2), 4–15. https://doi.org/10.26577/EJPh.2023.v190.i2.ph1

Aldaberdikyzy, A. (2013). Lexical interferences of Kazakh-Russian bilingual children of preschool age: A sociolinguistic survey. *Procedia - Social and Behavioral Sciences, 93*, 1885–1888. https://doi.org/10.1016/j.sbspro.2013.10.134

Allen, S. (2024, June 6). *Is Acquisition any Different in a Polysynthetic Language? The Case of Inuktitut*. Conference of the International Max Planck Research School for Language Sciences, Nijmegen, Netherlands. https://acquisition-sketch.phil-fak.uni-koeln.de/

Altinkamis, F. (2004). *Altinkamis Corpus*. Retrieved from https://childes.talkbank.org doi:10.21415/T5H89W

Bassano, D., Laaha, S., Maillochon, I., & Dressler, W. U. (2004). Early acquisition of verb grammar and lexical development: Evidence from periphrastic constructions in French and Austrian German. *First Language, 24*(1), 33–70. https://doi.org/10.1177/0142723704040704

Bick, E. (2020). An Annotated Social Media Corpus for German. In N. Calzolari, F. Bechet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the 12th International Conference on La3nguage Resources and Evaluation, (LREC2020)* (pp. 6127–6135). European Language Resources Association (ELRA). https://aclanthology.org/2020.lrec-1.752.pdf

Boribayeva, G. (2024). Social emotional learning in early childhood: Language acquisition on Kazakh and English. *Proceedings of the 7th International Scientific Conference*

"Theoretical Hypotheses and Empirical Results", Norway, 33–43. https://doi.org/10.5281/zenodo.12741275

Bracks, C., & Bardají i Farré, M. (2023). Broadening language documentation with child language data: First-hand experience from Tolitoli. *Language Documentation and Conservation Special Publication, 28*, 87–108. https://hdl.handle.net/10125/74721

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan. (2021). *National composition, religion and language proficiency in the Republic of Kazakhstan: Results of the national population census of 2021*. https://stat.gov.kz/upload/medialibrary/cee/3rsfg8ps3xo19orb284esg4rx27ihqf7/%D0%9D%D0%B0%D1%86%D0%B8%D0%BE%D0%BD%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B9%20%D1%81%D0%BE%D1%81%D1%82%D0%B0%D0%B2.pdf

Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan. (2025, March 27). *Chislennost' naseleniya Respubliki Kazakhstan po otdel'nym etnosam i vozrastnym gruppam na nachalo 2025 g*. [Population of the Republic of Kazakhstan by individual ethnic groups and age groups at the beginning of 2025]. https://stat.gov.kz/ru/industries/social-statistics/demography/publications/281562/

Darwin, C. (1877). A biographical sketch of an infant. *Mind, 2*, 287–294. https://psychclassics.yorku.ca/Darwin/infant.htm

Defina, R., Allen, S. E. M., Davidson, L., Hellwig, B., Kelly, B. F., & Kidd, E. (2023a). Sketch acquisition manual (SAM), Part I: The sketch corpus. *Language Documentation and Conservation Special Publication, 28*, 5–38. https://hdl.handle.net/10125/74719

Defina, R., Allen, S. E. M., Davidson, L., Hellwig, B., Kelly, B. F., & Kidd, E. (2023b). Sketch acquisition manual (SAM), Part II: The acquisition sketch. *Language Documentation and Conservation Special Publication, 28*, 39–86. https://hdl.handle.net/10125/74720

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics, 47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

Dotton, Z., & Wagner, J. D. (2018). *A grammar of Kazakh*. Duke University Slavic and East European Language Resource Center.

Esser, S. (2026). *Working with heterogeneous language documentation metadata: A prototype for consolidation, validation, and exploratory visualisation* [Master's thesis, University of Eastern Finland]. University of Eastern Finland eRepository. https://erepo.uef.fi/handle/123456789/38275

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2006). *MacArthur-Bates Communicative Development Inventories, Second Edition (CDIs)* [Database record]. APA PsycTests. https://doi.org/10.1037/t11538-000

Fernández, E. M., & Cairns, H. S. (2011). *Fundamentals of psycholinguistics*. John Wiley & Sons.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczaj (Ed.), *Language development* (Vol. 2, pp. 301–334). Hillsdale, New Jersey: Lawrence Erlbaum.

Hardegger, S., Mazara, J., Küntay, A. C., Hellwig, B., Pfeiler, B., & Stoll, S. (2024). Verbs or nouns? A cross-linguistic study examining the effect of morphological complexity and input on children's early lexical development. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *the Netherlands*, 2354–2361. https://escholarship.org/uc/item/7254q3jb

Hellwig, B., Allen, S. E. M., Davidson, L., Defina, R., Kelly, B. F., & Kidd, E. (2023). Introduction: The acquisition sketch project. *Language Documentation and Conservation Special Publication, 28*, 1–3. https://hdl.handle.net/10125/74718

Johanson, L. (2021). *Turkic* (Cambridge Language Surveys). Cambridge: Cambridge University Press. doi:10.1017/9781139016704

Joo, K. J., Yoo, I. W., & Kim, K. (2020). Syntactic categories in early child English: A corpus study of article production. *Journal of Cognitive Science, 21*(4), 595–618. DOI : 10.17791/jcs.2020.21.4.595

Jumagaliyeva, A. (2021). *Choosing between Kazakh and Russian: A study of young ethnic Kazakhs' language choices through the lens of habitus* [Master's thesis, Nazarbayev University]. Nazarbayev University Graduate School of Education http://nur.nu.edu.kz/handle/123456789/5609

Kempe, V., Brooks, P. J., & Gillis, S. (2024). Four decades of open language science: The CHILDES project. *Language Teaching Research Quarterly, 44*, 15–30. https://doi.org/10.32038/ltrq.2024.44.04

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language, 42*(6), 703–735. https://doi.org/10.1177/01427237211066405

Kucherbayeva, D., & Smagulova, J. (2023). Language revitalization: Challenges for Kazakh in higher education. *Journal of Eurasian Studies, 14*(2), 166–178. https://doi.org/10.1177/18793665231189326

Kuzembayeva, G., Maigeldiyeva, Z., Kussanova, B., & Maydangalieva, Z. (2024). Heritage language experiences and proficiency among ethnic Kazakhs abroad: Exploring challenges, language skill requirements, and acquisition barriers. *Journal of Social Studies Education Research, 15*(3), 348–373. https://jsser.org/index.php/jsser/article/view/5655

Lee, H., & Allen, S. E. M. (2023). An acquisition sketch of Inuktitut. *Language Documentation and Conservation Special Publication, 28*, 135–213. https://hdl.handle.net/10125/74723

Lelik, V., Diachkova, M., Dorofeeva, S. V., Lopukhina, A., Dragoy, O., & Sekerina, I. A. (2025). *RusLan-M Corpus*. Retrieved from https://childes.talkbank.org doi:10.21415/1XMG-D508

Lewkowicz, D. J. (2001). The concept of ecological validity: What are its limitations and is it bad to be invalid? *Infancy, 2*(4), 437–450. https://doi.org/10.1207/S15327078IN0204_03

Liu, H., MacWhinney, B., Fromm, D., & Lanzi, A. (2023). Automation of language sample analysis. *Journal of Speech, Language, and Hearing Research, 66*, 2421–2433. https://doi.org/10.1044/2023_JSLHR-22-00642

Liu, H. & MacWhinney, B. (2024). Morphosyntactic analysis for CHILDES. *Language Development Research, 4*(1), 233–258. http://doi.org/10.34842/j97r-n823

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Matiasovitsová, K., Čechová, P., Sláma, J., Homolková, K., & Smolík, F. (2024). Mean length of utterance in Czech toddlers: Validity estimates and comparison of words, morphemes, and syllables. *Journal of Speech, Language, and Hearing Research, 67*(3), 837–852. DOI: 10.1044/2023_JSLHR-23-00251

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McCollum, A. G., & Chen, S. (2020). Kazakh. *Journal of the International Phonetic Association, 51*(2), 1–23. https://doi.org/10.1017/S0025100319000185

Megyesi, B., Granstedt, L., Johansson, S., Prentice, J., Rosén, D., Schenström, C.-J., Sundberg, G., Wirén, M., & Volodina, E. (2018). Learner corpus anonymization in the age of GDPR: Insights from the creation of a learner corpus of Swedish. *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning, Sweden,* 47–56. https://aclanthology.org/W18-7106/

Nakispekova, A. (2026, April 19). *How Kazakhstan's Writing System Evolved Through History*. The Astana Times. https://astanatimes.com/2026/04/how-kazakhstans-writing-system-evolved-through-history/

National Scientific and Practical Center for Special Education. (2019). *Analiz izucheniya problemy formirovaniya yazykovoy sistemy rodnogo yazyka v khode rechevogo ontogeneza detey rannego i doshkol'nogo vozrasta (analiticheskaya spravka)* [Analysis of the problem of forming the native language system in the course of speech ontogenesis in early and preschool children (analytical report)]. https://special-edu.kz/analitspr/%D0%90%D0%BD%D0%B0%D0%BB.%D1%81%D0%BF%D1%80%D0%B0%D0%B2%D0%BA%D0%B0%20%D0%BA%D0%B0%D0%B7%D0%BB%D0%B0%D0%B1.pdf

Nelson, K. (1989). *Narratives from the crib*. Harvard University Press.

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology, 162*, 31–38. https://doi.org/10.1016/j.jecp.2017.04.017

Novotney, S., & Chris Callison-Burch. (2010). Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, USA*, 207–215. https://aclanthology.org/N10-1024/

Owens, R. E., Jr. (2020). *Language development: An introduction* (10th ed.). Pearson.

Papafragou, A., Cassidy, K., & Gleitman, L. (2007). When we think about thinking: the acquisition of belief verbs. *Cognition 105*, 125–165. DOI:10.1016/j.cognition.2006.09.008

Potamianos, A., & Narayanan, S. (2003). Robust recognition of children's speech. *IEEE Transactions on Speech and Audio Processing, 11*(6), 603–616. https://doi.org/10.1109/TSA.2003.818026

Pye, C. (2021). Documenting the acquisition of indigenous languages. *Journal of Child Language, 48*(3), 454–479. https://doi.org/10.1017/S0305000920000318

Rakhimova, D., Duisenbekkyzy, Z., & Adali, E. (2025). Investigation of ASR models for low-resource Kazakh child speech: Corpus development, model adaptation, and evaluation. *Applied Sciences, 15*(16), 8989. https://doi.org/10.3390/app15168989

Seifart, F. (2011). Cross-linguistic variation in the noun-to-verb ratio: the role of verb morphology and narrative strategies. *Association for Linguistic Typology 9th Biennial Conference*. University of Hong Kong, Japan.

Slobin, D. (1982). Universal and particular in the acquisition of language. In E. Wanner & L. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 128–172). New York: Cambridge University Press.

Smagulova, J. (2008). Language policies of Kazakhization and their influence on language attitudes and use. *International Journal of Bilingual Education and Bilingualism, 11*(3-4), 440–475. https://doi.org/10.1080/13670050802148798

Smagulova, J. (2014). Early language socialization and language shift: Kazakh as Baby Talk. *Journal of Sociolinguistics, 18*, 370–387. https://doi.org/10.1111/josl.12078

Šmit Brleković, L., & Kuvač Kraljević, J. (2023). Parental reports on language development in toddlers and preschoolers based on the Croatian version of Communicative Development Inventories III. *Frontiers in Psychology, 14*, 1188550. https://doi.org/10.3389/fpsyg.2023.1188550

Stoll, S., Bickel, B., Lieven, E., Paudyal, N. P., Banjade, G., Bhatta, T. N., Gaenszle, M., Pettigrew, J., Rai, I. P., Rai, M., & Rai, N. K. (2012). Nouns and verbs in Chintang: Children's usage and surrounding adult speech. *Journal of Child Language, 39*, 284–321. https://doi.org/10.1017/S0305000911000080

Suleimenova, E. D., Akanova, D. Kh., & Aimagambetova, M. M. (2021). Biz birgemiz, or revisiting the diversification of Russian language in Kazakhstan. *RUDN Journal of Language Studies, Semiotics and Semantics, 12*(1), 7–19. DOI: 10.22363/2313-2299-2021-12-1-7-19

Tamis-LeMonda, C. S., Kuchirko, Y., Luo, R., Escobar, K., & Bornstein, M. H. (2017). Power in methods: Language to infants in structured and naturalistic contexts. *Developmental Science, 20*, e12456. https://doi.org/10.1111/desc.12456

Tang, T. P.-Y., Lau, D. K.-Y., & Leung, M.-T. (2023). Corpus of Mandarin child language: A preliminary study on the acquisition of semantic content categories in Mandarin-speaking preschoolers. *Frontiers in Psychology, 14*, 1234525. https://doi.org/10.3389/fpsyg.2023.1234525

Tatsumi, T., Chang, F., & Pine, J. M. (2021). Exploring the acquisition of verb inflections in Japanese: A probabilistic analysis of seven adult-child corpora. *First Language, 41*(1), 41–66. https://doi.org/10.1177/0142723720926320

Tchintcharauli, T., Tsintsadze, N., Damenia, T, Kalkhitashvili, T., Doborjginidze, N. & Uziel-Karl, S. (2024). Establishing guidelines for MLU measurement in an agglutinating language: An illustration of Georgian. *First Language, 44*(4), 345–371. https://doi.org/10.1177/01427237241247930

Templin, M. (1957). *Certain language skills in children*. Minneapolis, MN: University of Minnesota Press.

TengriNews. (2022). *Nazvany samye rasprostranennye familii v Kazakhstane* [The most common surnames in Kazakhstan have been named]. https://tengrinews.kz/kazakhstan_news/nazvanyi-samyie-rasprostranennyie-familii-v-kazahstane-486722/

TengriNews. (2025). *Most popular baby names in Kazakhstan in 2025 revealed.* https://en.tengrinews.kz/kazakhstan_news/most-popular-baby-names-in-kazakhstan-in-2025-revealed-268073/

Terlikbayeva, N., & Menlibekova, G. (2021). The dynamics of language shift in Kazakhstan: Review article. *Journal of English Language Teaching and Applied Linguistics, 3*(2), 12–22. https://doi.org/10.32996/jeltal.2021.3.2.2

Troiani, G., & Filchenko, A. (2026). Introducing the first module of the multimedia corpus of spoken Kazakh language. *Journal of Open Humanities Data, 12*, 67. https://doi.org/10.5334/johd.529

Urbanczik, G. (2023). Exploring case marking in German first language acquisition using the acquisition sketch approach. *Language Documentation and Conservation Special Publication, 28*, 109–134. https://hdl.handle.net/10125/74722

Wilbur, J. & Rießler, M. (2022). Working with language data: Practical, technical and scientific considerations. In F. Brisard, S. D'hondt, P. Gras & M. Vandenbroucke (Eds.), *Handbook of Pragmatics: 25th Annual Installment* (pp. 3–19). John Benjamins Publishing Company. https://doi.org/10.1075/hop.25.wor3

WorldData.info. (2025). *Spread of the Kazakh language*. https://www.worlddata.info/languages/kazakh.php

Yang, J. S., Rosvold, C., & Bernstein Ratner, N. (2022). Measurement of lexical diversity in children's spoken language: Computational and conceptual considerations. *Frontiers in Psychology, 13*, 905789. https://doi.org/10.3389/fpsyg.2022.905789

Yule, G. (2010). *The Study of Language* (4th ed.). Cambridge University Press.

Zhakupova, A., Anishchenko, O., & Temirova, Z. (2024). Strategii osmysleniya motivirovannoy russkoy leksiki det'mi-monolingvami i det'mi-bilingvami Severnogo Kazakhstana [Strategies for understanding motivated Russian vocabulary by monolingual and bilingual children of Northern Kazakhstan]. *Rusistika, 22*(2), 208–224. http://doi.org/10.22363/2618-8163-2024-22-2-208-224

Zhakupova, A. D., Avramenko, M., & Meir, N. (2025). Lexical skills in Russian among Russian-Kazakh bilingual children in Northern Kazakhstan: Evidence from noun and verb naming. *International Journal of Bilingualism, 0*(0), 1–24 https://doi.org/10.1177/13670069251396294

Zhang, M., & Hudson, J. A. (2018). The development of temporal concepts: Linguistic factors and cognitive processes. *Frontiers in Psychology, 9*, 2451. https://doi.org/10.3389/fpsyg.2018.02451

Zhubanov, A. K., Fazylzhan, A. M., & Suleimenova, E. D. (2012). *National Corpus of the Kazakh Language*. https://qazcorpus.kz/indexen.php

Zhumashova, Z., Aksholakova, A., Omirbekova, A., & Kulbekova, B. (2023). Challenges and recommendations for teaching and learning Kazakh as a second language. *International Journal of Society, Culture & Language, 11*(2), 119–131. https://doi.org/10.22034/ijscl.2023.1995923.2967

DECLARATION

I acknowledge the occasional use of ChatGPT (OpenAI) to support language editing, including grammar correction and suggestions for improving clarity and wording. All ideas, analyses, figures/tables, interpretations, and conclusions presented in this thesis are my own.

APPENDICES

Appendix A

Pseudonyms Used in Transcriptions

Category	Pseudonym(s) Used
Target Child (female)	Айлин
Target Child (male)	Айбар
Mother	Томирис
Father	Мирас
Other female individuals	Айкөркем, Айзере, Зере
Other male individuals	Алан, Арлан, Алдияр
Surnames	Ахметов(а), Омаров(а),Оспанов(а), Алиев(а)
Streets/avenues	Абай, Алтынсарин, Жұмабаев
Villages/auls	Ақбелес, Белсай, Күркетау, Суталас

Note. Diminutives, shortened forms, spelling errors, and inconsistent variants in the transcripts were normalized to a single standard pseudonym.

Appendix B

Semantic Field Categories Used in the Analysis

Category Name	Definition	Examples
People	Terms referring to people, family members, and social roles	мама, әже, доктор
Food/Drink	Items referring to food and beverages	нан, шәй, алма
Animals	Names of animals	ит, аю, мысық
Body Parts	Terms denoting parts of the body	қол, көз, аяқ
Household Items	Everyday manipulable domestic objects	сабын, қайшы, щетка
Vehicles	Means of transportation and vehicle-related items	машина, автобус, колесо
Toys	Play objects, games, and fictional play-related referents	доп, кубик, Роблокс
Clothing	Items of clothing and wearable accessories	көйлек, сумка, носок
Furniture/Rooms	Furniture items and parts of domestic interior spaces	үстел, орындық, есік
Places/Locations	Spatial locations, buildings, and other place-denoting nouns	балабақша, дүкен, дала
Nature	Natural objects and environmental elements (excluding animals)	гүл, тас, аспан
Time	Terms related to time and temporal reference	күн, уақыт, минут
Activity	Activities, routines, events, or action-related nouns	той, укол, би
Descriptive/Abstract	Abstract qualities, evaluative terms, and non-concrete concepts	байлық, ақыл, өтірік
Onomatopoeia/Sounds	Sound-symbolic forms and vocal sound imitations	мяу, бум, афаф