

A Kazakh–Russian Corpus of Child-Directed Language for Low-Resource Languages

Albina Mukusheva

IUSS Pavia
Piazza della Vittoria 15
Pavia, Italy

albina.mukusheva@iusspavia.it

Achille Fusco

IUSS Pavia
Piazza della Vittoria 15
Pavia, Italy

achille.fusco@iusspavia.it

Cristiano Chesi

IUSS Pavia
Piazza della Vittoria 15
Pavia, Italy

cristiano.chesi@iusspavia.it

Abstract

Morphologically rich and low-resource languages present challenges for tokenization and other natural language processing (NLP) tasks. Standard tokenization algorithms often fail to segment words into meaningful morphological units. At the same time, datasets for many low-resource languages remain limited. We introduce a new ¹dataset of child-directed language in Kazakh and Russian compiled from multiple sources including child–adult dialogue transcripts, fairy tales, cartoons and children’s literature texts. The dataset is manually collected and contains 1M tokens in Kazakh and 2.3M tokens in Russian. Dialogue data includes meta-data such as child age and speaker identity, enabling research on language acquisition and linguistic development. We describe the process of data collection, corpus organization, and basic statistical properties of the dataset. The corpus provides a new resource for research on low-resource NLP, language acquisition, and morphologically-aware tokenization methods.

1 Introduction

Both agglutinative and morphologically rich languages present persistent challenges for natural language processing because extensive word formation increases surface-form variation and contributes to data sparsity. Kazakh and Russian provide an informative contrast in this respect. Kazakh, an agglutinative Turkic language, forms words by attaching sequences of affixes to stable stems, whereas Russian exhibits complex inflectional morphology across nouns, verbs, and adjectives. These linguistic characteristics make both languages valuable for computational and comparative linguistic research.

At the same time, progress in NLP for low-resource languages is often constrained by the lack of suitable datasets. Although a number of corpora

are available for Russian, resources for Kazakh remain limited, particularly in domains that capture natural child-directed language.

Child-directed texts, such as children’s literature, provide linguistically accessible material while still preserving productive morphological patterns. In addition, transcripts of child–adult interaction offer valuable insight into language acquisition and developmental linguistic patterns that are not typically represented in standard written corpora. In Kazakhstan, many children are exposed to both Kazakh and Russian from an early age, requiring them to acquire and process two typologically distinct morphological systems simultaneously. A bilingual corpus of child-directed input therefore provides not only a resource for NLP but also a foundation for investigating how bilingual children encounter and learn different patterns of word formation across the two languages.

The dataset presented in this work combines narrative texts and dialogue transcripts to support research in low-resource NLP, child language modeling, and tokenization methods. The contributions of this work are:

1. A new Kazakh-Russian corpus of child-directed language.
2. A dataset combining narrative texts and child-adult dialogue transcripts.
3. A statistical and linguistic analysis of the corpus.

2 Related Work

Several corpora have been developed to support research on child language acquisition. One of the most widely used resources is CHILDES (MacWhinney, 2000), which provides transcripts of child–adult interactions across multiple languages. Russian child language data is available

¹<https://huggingface.co/datasets/esimijq/Kazakh-Russian-Child-Directed-Speech-Corpus>

within this resource, though its size remains relatively limited for large-scale NLP experiments.

For low-resource languages, the development of linguistic resources remains a major research challenge. Many NLP studies rely on large datasets such as news corpora or web crawls, which are rarely available for languages with smaller digital presence. Kazakh, despite being widely spoken, still lacks large publicly available corpora suitable for training and evaluating NLP models.

Children’s literature texts have often been used as sources for linguistic datasets because they provide grammatically rich but accessible language. Such texts typically contain simpler syntactic structures while preserving complex morphological patterns. The corpus introduced in this work complements existing resources by combining child speech transcripts and children’s literature texts in Kazakh and Russian.

3 Linguistic Characteristics

Kazakh and Russian differ substantially in how grammatical information is encoded.

Kazakh morphology is characterized by its agglutinative structure, in which affixes are added sequentially to a stable root while preserving clear morpheme boundaries, and each affix typically expresses a single grammatical meaning (Zhanpeyisov, 2002). The language distinguishes between root and affixal morphemes, with affixes functioning either derivationally to form new words or inflectionally to mark grammatical categories (Kaliakbarova, 2013). Important morphological features include the extensive use of possessive, case, and personal endings to indicate grammatical relationships and syntactic roles within the sentence (Momynova et al., 2024). Kazakh morphology is also notable for the absence of grammatical gender and for expressing meanings that many other languages convey through prepositions by means of case endings and postpositions. Kazakh additionally employs multiple writing systems across regions and historical contexts. In Kazakhstan, both Cyrillic and Latin scripts are used, while Kazakh communities in China, primarily use an Arabic-based script (Abish, 2016). The Kazakh corpus used in this study is written in Cyrillic script.

Russian exhibits a fusional synthetic morphological system, in which grammatical meanings are encoded mainly within the word itself through inflection, with a single ending often simultane-

ously expressing categories such as number, gender, case, person, and tense (Bogdanov et al., 2009). An important feature of this system is the use of zero inflections, where the absence of an overt ending still carries grammatical meaning. Alongside these synthetic patterns, Russian also employs analytic constructions, where grammatical meaning is conveyed by function words, as well as combined forms, where meaning is expressed through both a preposition and an inflected word. In some cases, Russian uses suppletion, where different grammatical forms of the same lexeme derive from different roots (Kukatova, 2020). Word forms are created not only through affixation but also by sound alternation, stem modification, and stress shift, and the complete set of these forms constitutes a word’s paradigm, which may be either complete or defective (Kamynina, 1999).

4 Dataset Construction

To reflect the bilingual linguistic environment of many children in Kazakhstan, the corpus includes child-directed materials in both Kazakh and Russian. By representing child-directed input in both languages, the corpus provides a comparative resource for examining how bilingual children process two typologically distinct morphological systems simultaneously. Because the corpus combines dialogue and narrative material, it can support broader natural language processing tasks, including language modeling of early linguistic input, bilingual lexical prediction, morphological analysis, and cross-lingual representation learning.

4.1 Data Sources

The corpus was compiled from multiple sources of child-directed language in Kazakh and Russian.

The Kazakh portion of the dataset contains approximately 1 million tokens collected from:

- Kazakh fairy tales,
- Cartoon subtitles (from OpenSubtitles² and manually collected),
- Texts from children’s journals,
- Stories from children’s books,
- Educational materials.

These texts are primarily written narratives and instructional materials directed toward young children. The table 1 below provides a quantitative

²<https://www.opensubtitles.org/>

summary of the Kazakh corpus, including the number of documents, total tokens, and unique types for each source category.

source	docs	tokens	types
Fairy_tales	176	81609	16375
Cartoon_subtitles	97	143780	28466
Children_journals	1021	219704	41655
Children_books	3	166490	35711
Educational_materials	358	460334	66485
total	1655	1071917	127773

Table 1: Quantitative summary of the Kazakh corpus.

The Russian portion of the dataset consists of approximately 2.3 million tokens collected from three main sources:

- Child–adult dialogue transcripts (from CHILDES (Lelik et al., 2025; Bar-Shalom and Snyder, 1997, 1998) and from *From the First Words to the First Class* (Gvozdev, 1981)),
- Russian folk and fairy tales,
- Fairy tales from other cultures translated into Russian.

The data cover five children recorded across different developmental periods: Varvara (1;6–2;10), Yasha (0;10–3;10), Tosya (0;10–3;10), Tanja (2;5–2;11), and Zhenya (1;0–7;9).

The Russian texts were organized into three groups. The first group contains age-specific fairy tales divided into the following age categories: 0–2, 3, 4, 5, 6, 7, and 8–11 years old. The second group consists of traditional Russian folk tales. The third group includes fairy tales translated into Russian from a wide range of source languages, from Abkhaz to Yakut.

The table 2 below shows a quantitative summary of the Russian corpus, including the number of documents, total tokens, and unique types for each source category.

4.2 Data Format

The dialogue transcripts are stored in spreadsheet format containing three columns: child age, speaker identity, and the utterance text. An example is shown in Table 3; the transliteration and English translation presented in the table were added for illustrative purposes and are not included in the original dataset.

source	docs	tokens	types
Dialogues	1	294608	29623
Tales_by_ages	182	213869	30939
Folk_tales	200	171642	22920
Translated_tales	1605	1589193	84693
total	1988	2269312	116611

Table 2: Quantitative summary of the Russian corpus.

Age	Speakers	Sentences
02;05;14	INTERVIEWER	Что? Chto? “What?”
02;05;14	CHILD	Не может открыть. Ne mozhet otkryt’. “He can’t open it.”
02;05;14	MOTHER	А рот не может открыть. A rot ne mozhet otkryt’. “And he can’t open his mouth.”
02;05;14	MOTHER	А почему? A pochemu? “And why?”
02;05;14	CHILD	Не может. Ne mozhet. “He can’t.”

Table 3: Example of Russian dialogue transcript

The narrative texts are stored in txt format, with each document containing the title followed by the narrative body which separated by a blank line.

5 Corpus Statistics

5.1 Overall corpus statistics

For statistical analysis, each narrative text file in both corpora was treated as an individual document, whereas the Russian dialogue spreadsheet was treated as a single document containing all dialogue utterances. Analyses were conducted on surface word forms in order to capture the full range of lexical and morphological variation present in child-directed input.

Table 4 presents lexical diversity measures for each corpus. To obtain a comprehensive assessment of lexical diversity, several complementary metrics were employed. Type-token ratio (TTR) (Templin, 1957) was included as a traditional and intuitive measure of vocabulary variation, calculated as the proportion of unique word types relative to the total number of tokens. However, because TTR decreases as text length increases, it is

Corpus	TTR	MATTR	MTLD	HLR	Guiraud’s Index	D Index (Vocd-D)
Kazakh	0.1192	0.9134	1.0000	0.5374	123.4123	0.9833
Russian	0.0514	0.8411	1.0000	0.4336	77.4092	0.9299

Table 4: Lexical diversity measures for Kazakh and Russian corpora

primarily suitable for preliminary comparisons between similarly sized texts.

To address this limitation, moving-average type-token ratio (MATTR) (Covington and McFall, 2010) was also calculated. MATTR computes the average TTR across fixed-length moving windows and is substantially less sensitive to document length, making it more reliable for comparing corpora of unequal size. Measure of textual lexical diversity (MTLD) (McCarthy, 2005) was additionally included because it is considered as one of the most robust length-independent lexical diversity measures. The hapax legomena ratio (HLR) (Popescu and Altmann, 2008), representing the proportion of words occurring only once, was used to assess vocabulary novelty and lexical richness. Guiraud’s Index (Guiraud, 1954) was included as a normalized alternative to TTR, reducing text-length effects by scaling the number of word types by the square root of the total number of tokens. Finally, the D Index (Vocd-D) (Malvern et al., 2004) was calculated as a probabilistic estimate of lexical diversity based on Monte Carlo sampling procedures.

The Kazakh corpus demonstrates consistently higher lexical diversity than the Russian corpus across nearly all measures. In particular, Kazakh shows higher values for TTR, MATTR, HLR, Guiraud’s Index, and Vocd-D, indicating a broader and more varied vocabulary. The MTLD values are identical across the two corpora, suggesting comparable levels of length-independent lexical diversity according to this metric.

5.2 Source-level statistics

To provide a more detailed characterization of the dataset, the same statistical analysis was performed on lemmatized texts to reduce inflation of diversity measures caused by morphological variation and to improve comparability across genres.

Lemmatization was performed using the Stanza Python NLP library (Qi et al., 2020) for both the Kazakh and Russian corpora. The choice of Stanza was motivated primarily by the limited availability of lemmatization tools for Kazakh. We do not

Lang	Acc (%)	aLd	Prec	Rec	F1
Kazakh	81.40	0.346	0.835	0.814	0.818
Russian	83.20	0.332	0.857	0.832	0.834

Table 5: Lemmatization accuracy on manually annotated samples.

expect these models to perform perfectly, given the challenges associated with processing under-resourced languages. It was therefore considered important to report performance statistics for Stanza in these languages and to acknowledge the possibility of lemmatization errors.

To assess the reliability of automatic lemmatization, a random sample of 500 tokens from each corpus was manually annotated with gold-standard lemmas and compared with the lemmas predicted by Stanza. Performance was evaluated using exact-match accuracy, average Levenshtein distance, and weighted precision, recall, and F1 score.

Morph. Seg.	Analysis	POS
қожа+мыз+дан master+POSS.1PL+ABL 'from our master'	Gold: қожа Pred.: қожамыз	NOUN
қаш+ай+ық flee+HORT+1PL 'let us flee'	Gold: қаш Pred.: қаша	VERB
ой+ы+мыз+ға thought+POSS.1PL+DAT 'to our thought'	Gold: ой Pred.: ойымыз	NOUN

Table 6: Kazakh lemmatization error examples with glosses and English translations.

Many existing Kazakh morphological tools rely on stemming-based methodologies (Rakhimova and Turganbayeva, 2019; Sultanova et al., 2019; Bogdanchikov et al., 2021), which often fail to preserve lexical consistency and may introduce additional noise into lexical diversity measurements. Although manual review indicated that Stanza’s Kazakh lemmatizer is not entirely error-free, it of-

Corpus	Source category	TTR	MATTR	MTLD	HLR	Guiraud’s Index	D Index (Vocd-D)
KZ	Fairy_tales	0.0882	0.7870	1.0010	0.4720	24.7368	0.9121
KZ	Cartoons_subtitles	0.0980	0.8259	1.0007	0.5347	36.9920	0.9168
KZ	Children_journals	0.0804	0.8280	1.0010	0.4822	37.7399	0.9590
KZ	Children_books	0.0945	0.8864	1.0023	0.4891	38.6663	0.9541
KZ	Educational_materials	0.0630	0.8330	1.0002	0.4964	41.9438	0.9402
RU	Dialogues	0.0688	0.6290	1.0001	0.5412	37.3533	0.8849
RU	Tales_by_ages	0.0730	0.8179	1.0002	0.4357	33.7357	0.9077
RU	Folk_tales	0.0712	0.8035	1.0000	0.4265	29.4929	0.9153
RU	Translated_tales	0.0226	0.8186	1.0001	0.3518	28.4351	0.9005

Table 7: Lexical diversity measures by source category in Kazakh and Russian corpora

fers a unified framework capable of performing full lemmatization for both languages within a single, integrated pipeline. Examples of lemmatization errors in the Kazakh corpus are shown in Table 6.

Table 7 presents the lexical diversity measures calculated for each source category in the Kazakh and Russian corpora. In the Kazakh corpus, *Children_books* exhibit the highest lexical diversity overall, with the highest MATTR, MTLD, and relatively high Guiraud’s Index values, suggesting that this category contains the broadest and most varied vocabulary among the Kazakh sources. *Educational_materials* show the highest Guiraud’s Index, indicating substantial vocabulary growth relative to text length, although their TTR values are comparatively lower. *Cartoons_subtitles* and *Children_journals* display similarly high MATTR scores, reflecting relatively rich local lexical variation. By contrast, *Fairy_tales* show lower MATTR and Guiraud’s Index values, suggesting comparatively less lexical diversity across texts.

In the Russian corpus, lexical diversity measures are more homogeneous across source categories. *Translated_tales* show the highest MATTR value, while *Tales_by_ages* display slightly higher Guiraud’s Index scores than the other Russian written sources. *Folk_tales* and *Tales_by_ages* exhibit broadly similar lexical diversity profiles across most measures. Although *Translated_tales* have a notably lower TTR value, their MATTR and MTLD scores remain comparable to the other Russian categories, indicating that lexical diversity remains relatively stable when text-length effects are controlled. Overall, the Russian written materials demonstrate less variation between source categories than the Kazakh corpus.

6 Future Work

Future work will focus on expanding the corpus with additional child-directed materials in both Kazakh and Russian. In particular, we plan to collect more spoken and written data from age-specific sources such as children’s conversations, educational materials, storybooks, and audiovisual transcripts. Increasing the amount and diversity of child-directed language will improve the linguistic coverage of the corpus and provide a broader representation of developmental language patterns across both languages.

Another direction of future work concerns the development of evaluation benchmarks for morphological generalization. Inspired by the SIGMORPHON 2022 shared tasks (Batsuren et al., 2022), we plan to investigate the construction of benchmark datasets for Kazakh based on child-directed language resources. Such benchmarks would provide a standardized framework for evaluating language models and tokenization methods on morphologically rich and low-resource languages.

In addition, the corpus offers an opportunity to investigate tokenization strategies tailored to distinct morphological systems. Because the dataset contains parallel challenges from agglutinative and fusional languages, it provides a suitable foundation for comparing and developing tokenization approaches that can better capture morphological structure. Recent work in the BabyLM shared task has highlighted the importance of tokenization choices for both intrinsic and extrinsic evaluation (Fusco et al., 2025). Building on these findings, we plan to further investigate how different tokenization methods affect text representations and language model performance in low-resource set-

tings.

Finally, the dataset is intended to support the training of language models for low-resource settings. We plan to investigate how child-directed corpora can improve small-scale multilingual and monolingual language models in Kazakh and Russian, particularly in tasks that require better morphological generalization. This resource may therefore contribute not only to tokenization research but also to broader NLP development for underrepresented languages.

References

- Aynur Abish. 2016. *Modality in Kazakh as Spoken in China*, 1 edition. Harrassowitz Verlag.
- Eva Bar-Shalom and William Snyder. 1997. Optional infinitives in russian and their implications for the pro-drop debate. In Marina Lindseth and Steven Franks, editors, *Formal Approaches to Slavic Linguistics: The Indiana Meeting 1996*, pages 38–47. Michigan Slavic Publications, Ann Arbor.
- Eva Bar-Shalom and William Snyder. 1998. Root infinitives in child russian: A comparison with italian and polish. In Richard Shillcock, Antonella Sorace, and Caroline Heycock, editors, *Language Acquisition: Knowledge Representation and Processing. Proceedings of GALA '97*. The University of Edinburgh, Edinburgh, UK.
- Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. *The SIGMORPHON 2022 shared task on morpheme segmentation*. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Seattle, Washington. Association for Computational Linguistics.
- A. V. Bogdanchikov, O. A. Baimuratov, and D. A. Ayazbayev. 2021. *Stemming of kazakh language. News of the National Academy of Sciences of the Republic of Kazakhstan*.
- S. I. Bogdanov, V. B. Evtukhin, Yu. P. Knyazev, et al. 2009. *Морфология современного русского языка [Morphology of Modern Russian]*. Русский мир: учебники для высшей школы [Russian World: Textbooks for Higher Education]. Faculty of Philology and Arts, Saint Petersburg State University, Saint Petersburg.
- Michael A. Covington and Joe D. McFall. 2010. *Cutting the gordian knot: The moving-average type-token ratio (mattr)*. *Journal of Quantitative Linguistics*, 17(2):94–100.
- Achille Fusco, Maria Letizia Piccini Bianchessi, Tommaso Sgrizzi, Asya Zanollo, and Cristiano Chesi. 2025. *Linguistic units as tokens: Intrinsic and extrinsic evaluation with babyLM*. In *Proceedings of the First BabyLM Workshop*, pages 496–507. Association for Computational Linguistics.
- P. Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.
- Pierre Guiraud. 1954. *Les caractères statistiques du vocabulaire: essai de méthodologie*. Presses universitaires de France, Paris.
- A. N. Gvozdev. 1981. *От первых слов до первого класса: дневник научных наблюдений [From the First Words to the First Class: A Journal of Scientific Observations]*. Saratov State University, Saratov. (In Russian).
- Gulnur Kaliakbarova. 2013. *Қазақ тілі (қысқаша грамматикалық анықтағыш) [Kazakh Language (A Brief Grammatical Reference Guide)]*. Institute for the Development of the State Language, Almaty.
- A. A. Kamynina. 1999. *Современный русский язык. Морфология [Modern Russian Language. Morphology]: Учебное пособие для студентов филологических факультетов государственных университетов [Study Guide for Students of Philological Faculties of State Universities]*. Moscow State University Press, Moscow.
- O. A. Kukatova. 2020. *Современный русский язык. Морфология [Modern Russian Language. Morphology]*. Innovatsion Rivozhlanish Nashriyot-Matbaa Uyi, Tashkent.
- Viktoria Lelik, Maria Diachkova, Svetlana V. Dorofeeva, Anastasia Lopukhina, Olga Dragoy, and Irina A. Sekerina. 2025. *Ruslan-m corpus*. Retrieved from <https://chilides.talkbank.org>.
- Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3 edition. Lawrence Erlbaum Associates, Mahwah, NJ.
- D. D. Malvern, B. J. Richards, N. Chipere, and P. Durán. 2004. *Lexical diversity and language development: quantification and assessment*. Palgrave Macmillan, Basingstoke.
- Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.
- B. Momynova, A. Soltanbekova, U. Anesova, T. Ramazanov, M. Imangazina, and Q. Slyambekov. 2024. *Қазақ тілінің кестелі грамматикасы. Морфология [Tabular Grammar of the Kazakh Language. Morphology]*. Asyl ZhK, Almaty.
- Ion-Iovitz Popescu and Gabriel Altmann. 2008. *Нарх легомена and language typology*. *Journal of Quantitative Linguistics*, 15(4):370–378.

- Prime Minister of the Republic of Kazakhstan. 2021. [Askar mamin chairs meeting of national commission on translation of kazakh alphabet into latin script.](#)
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python natural language processing toolkit for many human languages.](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- Diana Rakhimova and Aliya Turganbayeva. 2019. [Lemmatization of big data in the kazakh language.](#) In *Proceedings of the 5th International Conference on Engineering and MIS, ICEMIS '19*, New York, NY, USA. Association for Computing Machinery.
- Nazerke Sultanova, Kanat Kozhakhmet, Ruslan Jantayev, and Azhar Botbayeva. 2019. [Stemming algorithm for kazakh language using rule-based approach.](#) In *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)*, pages 1–4.
- Mildred C Templin. 1957. Certain language skills in children; their development and interrelationships.
- Erbol Zhanpeyisov. 2002. *Қазақ грамматикасы. Фонетика, сөзжасам, морфология, синтаксис [Kazakh Grammar: Phonetics, Word Formation, Morphology, Syntax]*. Astana.