PERMISSION TO DISTRIBUTE TRANSCRIPT DATA THROUGH TALKBANK
Carnegie Mellon University

Carnegie Mellon University is participating in a data-sharing program known as TalkBank that was established through federal and foundation grants to Brian MacWhinney at Carnegie Mellon. For the program to function properly, contributors will be asked to and will give permission for their work to be made available to other researchers. With a full understanding of the aforementioned, I hereby give permission to TalkBank to make and circulate electronic copies of the language transcripts and media that I describe below. These copies may be distributed to scholars and other responsible parties. I warrant that this use of the data is in accord with Human Subjects review procedures at my institution and that participants have given informed consent to have their data available to researchers. I also warrant that there is no copyright restriction over the transcripts and media being circulated. Any further restrictions that I wish to place on the use of these data are listed under (2) below. I do not hold Brian MacWhinney or Carnegie Mellon responsible for the enforcement of these further restrictions and indemnify and render harmless both Brian MacWhinney and Carnegie Mellon University against any actions at law or in equity or in similar courts of any jurisdictions arising from violations of these restrictions.

1. General description of the data set, and IRB Approval # (if available):

The *Balausa1* corpus comprises a collection of spontaneous caregiver–child interaction samples in Kazakh. The corpus includes 13 recordings from eight typically developing Kazakh-speaking children between the ages of 1;09 and 5;01 interacting with parents and other family members in naturalistic home environments across Kazakhstan. Most children contributed a single audio recording, while several children were recorded on multiple occasions, providing limited developmental coverage across time. Recordings were collected by parents during everyday activities, including play, family conversations, storytelling, mealtime interactions, and routine household activities.

The current release contains approximately five hours of time-aligned audio recordings and transcripts. The corpus includes 3,621 child utterances (approximately 7,005 tokens) and 4,177 adult utterances (approximately 10,780 tokens). Morphological annotation (%mor tier) was generated automatically using Batchalign2 and Universal Dependencies resources and subsequently manually verified for nouns, proper nouns, verbs, and auxiliary verbs with the help of trained native Kazakh speakers.

For each participating child, informed consent for the use and publication of the data was obtained from a parent. Participants are identified only through corpus codes (e.g., AE-01, BK-01), and any potentially identifying information mentioned in the transcripts, such as personal names, addresses, or other private details, was pseudonymized.

2. Restrictions to be placed on the use of the data:

There are no restrictions on the use of the transcripts.

Signed _____ Date: 02.06.2026

Printed Name: Ilya Razorenov

Institutions:
University of Groningen, Groningen, the Netherlands
Ghent University, Ghent, Belgium
University of Eastern Finland, Joensuu, Finland

Please mail this signed form to: Brian MacWhinney, CMU-Psychology, 5000 Forbes Ave. Pittsburgh, PA, 15213, USA or send a scanned signed copy to macw@cmu.edu