Documentation for

Chroma corpus

version 2023.07

Part of
CoCzeFLA | Corpora of Czech as the First Language in Acquisition
coczefla.ff.cuni.cz

Mgr. Anna Chromá /supervision, coordination, morphological annotation/
Faculty of Arts, Charles University, Prague
anna.chroma@ff.cuni.cz

Mgr. Klára Matiasovitsová /supervision, coordination, morphological annotation/
Faculty of Arts, Charles University, Prague
klara.matiasovitsova@ff.cuni.cz

Mgr. Jakub Sláma /morphological annotation/
Czech Language Institute of the Czech Academy of Sciences, Prague
slama@ujc.cas.cz

Mgr. Jolana Treichelová /coordination, morphological annotation/
Faculty of Arts, Charles University, Prague
jolana.treichelova@ff.cuni.cz

Participants: 7
Type of Study: longitudinal, spontaneous
Location: Czech Republic
Media type: transcripts (from audio recordings)

Citation Information

Wenn using data from the *Chroma* corpus in any of its published versions, please cite the following article:

Chromá, A., Sláma, J., Matiasovitsová, K., & Treichelová, J. (in press). A morphologically annotated longitudinal corpus of spoken Czech child-adult interactions. *Language Resources and Evaluation*.

Corpus

Project Description.

Chroma 2023.07 is a corpus of transcribed spontaneous child-adult interactions in Czech. It consists of 99,358 tokens in 41,585 utterances produced by seven children between ca 1.5 to 3.5 years of age, and 238,073 tokens in 60,734 utterances produced by their close caregivers in everyday situations at home. The corpus covers language production of the children from the mean length of 1.04 word per utterance up to 4.84 words per utterance. The length of the recorded period ranges for individual children from 11 to 27 months. The transcripts of both child and adult utterances were lemmatized and tagged using MorphoDiTa, a tool for automatic morphological analysis of Czech. The annotation was transformed into the MOR format.

Details on procedure, participants, and morphological annotation are to find in **Chromá et al. (in press)** (see above) and at the home page of the CoCzeFLA group.

Warnings.

1. During transcription, we did not pay particular attention to the %pho tier, the pseudophonological transcription in *Chroma* corpus is very approximate. Researchers interested in the phonetical/phonological aspects of the data should contact the group CoCzeFLA to get the recordings.
2. There is no translation to English or any other languages. All the transcribed material as well as all the comments are in Czech only.
3. Two transcripts (Julie20221, Klara30424) from the previous version were removed since they did not meet our criteria on dialogical format. All transcripts of recordings made during one day were merged into one file. Thus, version 2023.07 consists of 183 files/transcripts.
4. Some of the coding principles were changed during the longitudinally ongoing project. The first version of *Chroma* corpus (2019.07) was transcribed according to the Manual v1.0; the added seventh child Sara was transcribed according to the Manual v2.0; and the currently published morphologically annotated version (2023.07) was additionally revised according to the Manual v3.0 (overview over the applied codes see in 'CHAT_for_CoCzeFLA' on our homepage). As a result, the main lines of *Chroma 2023.07* differ from the first version 2019.07 consistently in the following:
    i. the code *yyy* was eliminated: mostly, it was replaced by *xxx*; sometimes, the produced form was reconstructed from the %pho tier or the recording;
    ii. a white space was added in front of terminal punctuations and colons;
    iii. the use of underscores was limited to a narrowly defined subgroup of interjections; in other cases, the underscores were simply removed or replaced by parentheses (see also 'CHAT_for_CoCzeFLA' above);
    iv. the codes *@d* for dialectal forms and *@f* for family-specific forms were eliminated; their use was inconsistent and rare;
    v. in error coding, the spike brackets for scope <> were eliminated and the code [*] is now placed consistently at the end of the given utterance (before the terminal punctuation);
    vi. (some) typos and coding errors were corrected.

Some other codes are applied inconsistently in *Chroma 2023.07* because of the changes in the transcription system. It applies to the following:

vii. The use of spike brackets for determining one-word scope of the following code, e.g., [//], was introduced in the transcription of Sara, however, the brackets are not used in the first six children with this function;
viii. The use of <> [!] code for accent was applied in *Chroma 2019.07* but abandoned while transcribing Sara. It was not eliminated but even in the first six children, the code is used inconsistently.
ix. The use of alternative transcript suggestions in case of uncertainty <*uncertain_transcript*> [? *alternative*] was applied in *Chroma 2019.07* but abandoned while transcribing Sara.
x. The use of comments within the main line in form of bracket codes <*transcript*> [=! *comment*] was applied in *Chroma 2019.07* but abandoned while transcribing Sara.

Identity protection. For each participating child, both his/her parents (as well as other participating caregivers) gave an informed consent for the use of the data. For the target children, we consistently use aliases that respect at least some of the morphological and phonological aspects of their original names. The corresponding hypocoristics were created from the aliases as well. With a few exceptions, the first names of other participants are not replaced. Surnames and addresses are replaced by a code *zzz* and commented on the %com tier.

Restrictions. There are no restrictions on the use of the transcripts.

Specific Codes. There are four codes of our own for the usage of interjection with the function of a predicate **@z:ip**, of a nominal phrase **@z:in**, and of a modifier **@z:ia**. There is a code for a foreign expression within otherwise Czech utterance, **@z:c**, but it is very rare.

History and Aknowledgements. Scholarships from Faculty of Arts, Charles University were drawn for transcribing and proofreading students already at the very beginning of CoCzeFLA in the years 2014–2015. In these years, prof. Karel Šebesta supported the project within the AKCES framework. Subsequently, Anna Chromá's project 'Longitudinální korpus raného vývoje řeči' (No. FF_VG_2016_16) received faculty support for the next two years. Between the years 2016–2017, this project provided financial rewards for the coordinator – Anna Chromá – and both transcribing and proofreading students. Thanks to this support, a substantial part of the first version of the *Chroma* corpus (2019.07) was created.

After the end of this project, CoCzeFLA continued to draw scholarships from Faculty of Arts for transcribing and proofreading students within the large infrastructure project LINDAT/CLARIAH-CZ (No. LM2023062, earlier LM2018101) funded by the Ministry of Education, Youth and Sports of the Czech Republic. Since 2021, Anna Chromá has been employed as a data curator at Faculty of Arts within LINDAT.

In the years 2021–2023, the project of Klára Matiasovitsová 'Nominal morphological categories and the mean length of utterance in a longitudinal corpus of early language development' was funded from university sources invested into the program 'Grant Schemes at Charles University' (CZ.02.2.69/0.0/0.0/19_073/0016935). The morphologically annotated version of the *Chroma* corpus (2023.07) was created thanks to this support.