Language Resources and Evaluation (2025) 59:413–436
https://doi.org/10.1007/s10579-023-09710-y

A morphologically annotated longitudinal corpus of spoken Czech child–adult interactions

Anna Chromá[1] · **Jakub Sláma**[1,2] · **Klára Matiasovitsová**[1] · **Jolana Treichelová**[1]

Accepted: 17 November 2023 / Published online: 30 March 2024

Abstract

The paper presents a longitudinal corpus of transcribed spontaneous child–adult interactions in Czech. It consists of 99,388 tokens in 42,103 utterances produced by seven children between ca 1.5 and 3.5 years of age, and 238,211 tokens in 61,252 utterances produced by their close caregivers in everyday situations at home. The corpus covers language production of the children from the mean length of 1.01 word per utterance up to 5.33 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 used within CHILDES, a database dedicated to corpora of first language acquisition. Detailed manual checking was performed on the annotation of all children's utterances. Data from three children were used for a comparison of part-of-speech classification before and after manual checking, data from one child was additionally analyzed for differences in morphological tagging proper. The number of differences was rather low, with (expected) limitations in the areas of part-of-speech classification for uninflected words, annotation of homonymous forms, and annotation of child-specific words. The corpus represents an important contribution to the research of child language with special significance for Slavic languages and other morphologically rich inflecting languages, which are still underrepresented in the study of first language acquisition.

Keywords First language acquisition · Longitudinal corpus · Morphological tagging · Czech language · Slavic languages

Extended author information available on the last page of the article

1 Introduction

Corpora of spontaneous language production play an important role in much of linguistic research. Systematic research of first language acquisition actually began around the 1960s, when spontaneous recordings of several English children and their caregivers were first collected (Brown, 1973). Sampling of child language (in contrast to other linguistic corpora) is rather unique in that it is concentrated in an international database including corpora of many different languages, CHILDES (MacWhinney, 2000). However, the latest CHILDES version (2021.01) includes 75 English-language corpora, which amounts to ca. 28% of all CHILDES corpora, whereas the whole group of Slavic languages is covered only by 11 corpora (4%). This paper presents a Czech corpus of spontaneous child- and child-directed speech, aiming to contribute towards bridging this gap.

Six Slavic languages are currently represented in CHILDES, mostly to a very limited extent. Five of them have at least one longitudinal corpus of the type we are aiming at, i.e., a collection of recordings covering at least one year of a child's life. CHILDES includes such data for eight Serbian children, four Bulgarian children, three Croatian children, one Polish child, and one Russian child. With our corpus, entitled *Chroma* after its initiator, we add another seven Czech children to the Slavic collection.

The *Chroma* corpus is the first published corpus of child Czech. It has been published in two versions so far: the first version with the first six children was published in CHILDES in 2019, version 2022.07 with all the seven children was published in LINDAT (Chromá & Matiasovitsová, 2022), and the latest, revised and morphologically annotated version 2023.07 was recently published in LINDAT (Chromá et al., 2023a) as well as in CHILDES (Chromá et al., 2023b). The corpus was morphologically annotated automatically with a tool which had been developed for Czech, and the annotation of children's utterances was checked manually, ensuring its high accuracy. Such a contribution will facilitate language acquisition research in morphologically complex languages, e.g., making it possible to address questions about the productivity of morphological knowledge, use of word order, and the relations between these domains in the adult input and children's productions.

2 Creation of the corpus

2.1 Participants

The participating families were recruited from among friends and colleagues of the authors. The criteria for recruitment were (1) the adequate age of the child, (2) only one predominant language in their environment (Czech), and (3) typical language and cognitive development so far (parents reported no concerns or observed symptoms of language, speech or sensory impairments in children or themselves).

No developmental disorders or irregularities were identified during the recorded periods. The seven target children (whose aliases, not actual names, are used in the corpus and throughout this paper) were recruited in three phases: the first three children (Aneta, Klára, Viktor) were first recorded in 2014 at the ages of 2;2 to 2;6. The second three (Anna, Jan, Julie) started being recorded in 2016–2017 at the ages of 1;7 to 1;9. The last child (Sára) was first recorded in 2019 at the age of 1;7. The initial and final ages of the seven children, their parent-reported gender (and further details on their recordings) are to be found in Table 1.

Parents and locality The families of Aneta, Anna, Jan and Viktor lived in the (broader) center of Prague during the recorded period. Julie's and Sára's families lived in the Prague metropolitan area, and Klára's family lived in a different university city, Hradec Králové. Both of Aneta's parents come from the northeastern part of Czechia (Czech-Polish borders) and their speech (as well as Aneta's speech) shows some features of this area's dialect. Otherwise, the speech recorded for the corpus mostly represents common Czech. Viktor's and Aneta's fathers have high school education; all the other parents have university education.

Siblings Only Aneta and Klára had older siblings. Aneta's brother is 6 years older than her, and Klára's brother is 2 years older. Jan was an only child throughout the recorded period. During the time of recording, a younger sibling was born to the families of Anna, Julie, Sára and Viktor.

2.2 Recording

The recording was conducted as a general data-collection project without specific research questions in mind. The corpus comprises transcriptions of audio recordings in MP3 format (and, to a very limited extent, in WAV format). The recordings were made by the participating families themselves. The families were provided with an audio recorder and with concise instructions on how to use it (if needed). The specific type of recorder varied from family to family as some of them used their own equipment. Generally, the recorders were of mid-quality with ordinary microphones and the recording was performed with no consistency in mic-to-mouth distance. However, the percentage of fully intelligible utterances in adults as well as in children is high, as indicated in Table 1.

The families were instructed (1) to record as natural situations as possible, (2) to record situations of joint activities of the child and a caregiver, such as eating, playing, and reading, (3) not to elicit rhymes or singing, (4) to avoid media playing and laud surroundings, especially outdoors. The recommended frequency of recording was 20–40 min once per 2 weeks. In the end, 11 to 27 months are covered for each child with the density of 21–54 recorded minutes per month (see Table 1).

Siblings on the recordings The caregivers received no special instructions regarding the presence of the siblings on the recordings. After their birth, younger siblings were present rather often on the recordings, however neither of them started to produce any language in the recorded period. The older brothers of Aneta and Klára were recorded to limited extend (see Table 2).

Table 1 Basic description of the recorded speech in the *Chroma* corpus for individual children and in total

	Aneta	Anna	Jan	Julie	Klára	Sára	Viktor	Total
Gender	Female	Female	Male	Female	Female	Female	Male	–
Most recorded adult(s)	Babysitter; Mother	Father	Mother	Mother	Mother	Mother	Mother	–
Age[a]								
Initial	2;02.08	1;09.30	1;07.05	1;07.05	2;04.22	1;07.06	2;06.23	–
Final	3;03.18	2;07.27	2;09.27	3;09.11	3;04.17	3;09.08	3;09.15	–
MLU[a]								
Initial	1.52	1.47	1.09	1.11	1.94	1.01	3.29	–
Final	2.24	2.94	2.66	2.74	3.19	3.36	5.33	–
Number of recorded months	14	11	15	27	13	27	16	–
Total recorded time	11:25:16	06:29:54	07:11:13	15:09:16	04:32:12	12:26:26	14:22:44	**23:37:01**
Mean number of recorded minutes per month	49	35	29	34	21	28	54	–
Number of utterances (UTT)								
Child's UTT								
All	5286	3127	4483	9345	3627	7750	8485	**42,103**
Intelligible[b]	4747	2459	3886	7951	3079	7632	8124	**37,878**
% of Intelligible	89.8	78.6	86.7	85.1	84.9	98.5	95.7	**90.0**
Adults' UTT								
All	7671	5367	4317	18,769	3197	12,686	9245	**61,252**
Intelligible[b]	7533	5115	4216	18,517	3128	12,674	9195	**60,378**
% of intelligible	98.2	95.3	97.7	98.7	97.8	99.9	99.5	**98.6**
Number of tokens (TOK)								
Child's TOK								
All	11,803	6166	6626	18,436	8937	17,646	29,774	**99,388**
From intelligible[b] UTT	11,331	5567	6279	17,388	7955	17,454	28,924	**94,898**
Adults' TOK								
All	31,812	21,636	15,057	71,059	10,499	48,853	39,295	**238,211**

Table 1 (continued)

	Aneta	Anna	Jan	Julie	Klára	Sára	Viktor	Total
From intelligible[b] UTT	31,699	20,846	14,852	70,429	10,355	48,829	39,209	**236,219**

MLU, number of utterances, and tokens were calculated using the CLAN software (MacWhinney, 2000). Since even the nominative form in Czech often consists of two morphemes and the morphemic analysis of words is generally not straightforward (see Sect. 3.3.1), we do not use MLU in morphemes but in words. For a detailed discussion on the MLU units for Czech, see Matiasovitsová et al. (in press)

[a]Age and MLU at the first and the last recording; age is given in the format Y;MM.DD

[b]Utterances containing any unintelligible speech (coded as *xxx*) were excluded

Table 2 Proportion of fully intelligible utterances (utt) produced by individual caregivers out of all adult's utt; and number of utt produced by siblings in the *Chroma* corpus

	Aneta	Anna	Jan	Julie	Klára	Sára	Viktor
Mother's utt (%)	47.7	1.6	100	96.8	98.4	91.2	92.0
Father's utt (%)	7.3	89.1	–	3.2	1.6	8.7	8.0
Grandparent's utt (%)	2.1	9.3	–	–	–	0.1	–
Babysitter's utt (%)	43.0	–	–	–	–	–	–
N of older sibling's utt	241	–	–	–	144	–	–

In Table 1, the most recorded adults for each child are indicated. In Table 2, all the recorded adults are indicated for each child's subcorpus with the proportion of utterances produced.

2.3 Transcription

We used the standard CHAT format for transcription of spoken language developed within CHILDES. In recent years, regular updates of the CHAT manual have been published online (https://talkbank.org/manuals/CHAT.pdf; MacWhinney, 2000). CHAT transcripts are plain-text files with the special extension .cha dedicated to the CLAN software developed within CHILDES as well. The CLAN enables transcription itself and provides a number of various data analysis programs as well. However, the transcripts are directly processable in any available text editor.

Basic features of CHAT The transcripts in CHAT are structured into separate utterances, each one captured on one main line with dependent tiers. Each main line begins with the specification of the speaker (e.g., **CHI* for the target child or **MOT* for the mother) and contains the orthographic transcription of the given utterance. The dependent tiers specify details about the utterance on the preceding main line. Each dependent tier begins with the specification of its type (e.g., *%pho* for the phonological tier and *%mor* for the morphological tier) and contains the respective information (e.g., phonological transcription, and lemmatization and morphological tags).

Utterances Since the mean length of utterance (MLU) is a key measure of development in child language research (Brown, 1973; Potratz et al., 2022), structuring into utterances is an important feature of CHAT. The end of an utterance is mostly defined prosodically by its falling intonation, but also syntactically (no more than two main clauses joined by the 'and' conjuction may appear in one line).

Optional features of CHAT A comprehensive list of all the other features of CHAT used in our corpus is to be found in Appendix A. See Table 3 for examples of the following selected phenomena:

- dysfluencies (repetitions, false starts, self-corrections, stuttering, trailing off, etc.):

Table 3 Example of transcription with English translation of the relevant lines/tiers. Morphological tiers are omitted for clarity

	@Begin	
	@Languages: ces	
	@Participants: CHI Julie Target_Child, MOT Mother	
	@ID: ces|Chroma2023.04|CHI|2;04.11|female|||Target_Child|||	
	@ID: ces| Chroma2023.04 |MOT|female||||Mother|||	
	@Transcriber: BARBLA	
1	@Situation: dopoledne; dívají se z okna.	@Situation: in the morning; they are looking out of a window .
2	*CHI: letadlo .	*CHI: airplane .
3	%pho: letaďo .	
4	%com: emoční důraz .	%com: affective emphasis .
5	*CHI: &p podíváme 0na letadlo .	*CHI: we will take a look 0at an airplane .
6	%pho: p podivame letaďo .	
7	%com: emoční důraz na 'letadlo' .	%com: with affective emphasis on 'airplane' .
8	*MOT: no, tak se podíváme na letadlo .	*MOT: OK, we will take a look at an airplane .
9	%pho: no tak se podíváme na letadlo .	
10	*MOT: tak si vem židličku .	*MOT: so take a chair .
11	%pho: tak si vem židličku .	
12	*CHI: zišku@c .	
13	%pho: zišku .	
14	%com: zišku = židličku .	%com: zišku = $\text{chair}_{\text{diminutive.accusative}}$.
15	*CHI: židličku .	*CHI: $\text{chair}_{\text{diminutive.accusative}}$.
16	%pho: idišku .	
17	%com: emoční důraz .	%com: affective emphasis .
18	(…)	
19	*MOT: hele@i, už +…	*MOT: look, now +…
20	%pho: hele, už .	
21	*MOT: < už ho > [//] vidíš ho ?	*MOT: < already it > [//] can you see it ?
22	%pho: už ho vidíš ho .	

- o *&* marks stuttering here (line 5 in Table 3); otherwise, it might mark phonological fragments of various kinds as well as hesitation sounds;
- o *+…* marks trailing off (line 19);
- o *< > [///]* marks self-correction (line 21);

- morphological innovations, errors, and idiosyncratic forms:

  - o *0* marks an omitted word; the preposition *na* 'on', omitted by the child, was reconstructed by the transcriber as obligatory (line 5);
  - o *@c* marks an idiosyncratic form; here *zišku@c* is a two-syllable attempt at the target three-syllable form *židličku* '$\text{chair}_{\text{diminutive.accustive}}$' (line 12).

Each recording was transcribed by one trained person and revised by another trained person. The revisions were reviewed and approved by the original transcriber. In the end, all transcriptions were checked for CHAT system conformity, using CLAN procedures, and revised where necessary.Each recording (or all recordings made during one day) was transcribed into one separate file; the collection consists of 13 to 33 files for individual children (183 files in total). Each file was annotated with the child's alias, age-at-recording, and various other metadata (see Table 3).

Anonymization For the target children, we consistently use aliases which respect at least some of the morphological and phonological aspects of their real names. In Czech, it is very common to form various hypocoristics from a given name, so corresponding forms were created from the aliases to replace the original forms. Other people frequently appearing in the recordings were mostly assigned random Czech first names. Last names and addresses were replaced by the code *zzz* and commented on in the multi-purpose %com tier.

3 Lemmatization and morphological tagging

One of the data analysis programs included in the CLAN software (see above) is MOR, a tool for automatic morphological tagging of corpora in the CHAT format. This, however, requires a dedicated MOR grammar for each language, and such grammars are available only for a handful of languages.[1] No MOR grammar is currently available for Czech or any of the other Slavic languages. While it is possible to build a new MOR grammar, we decided to use a readily available tool for lemmatization and morphological tagging of Czech, MorphoDiTa (Spoustová et al., 2009; Straková et al., 2014). The authors of the morphological analysis of a corpus of European Portuguese child- and child-directed speech (Santos et al., 2014) similarly did not create a new MOR grammar for Portuguese but applied a ready-to-use tagger developed previously for other corpora of Portuguese.

MorphoDiTa (Morphological Dictionary and Tagger) is an open-source tool for morphological analysis (tokenization, sentence segmentation, lemmatization, and morphological tagging), which is well-established in the Czech context, having been used for the morphological analysis of natural language texts compiled in corpora of the Czech National Corpus project. MorphoDiTa is available as a web application[2] but, more importantly for natural language processing tasks, also as source code and as a pre-compiled binary package, both available on GitHub.[3] The linguistically challenging subtasks carried out by the tool—lemmatization and morphological tagging—rely on a comprehensive morphological dictionary, a rule-based disambiguation component and a stochastic tagger (for details, see Hnátková et al., 2014).

[1] Information on available MOR grammars is given on the web page of the project TalkBank, which provides various data repositories (including child language banks) as well as CHAT, CLAN, and MOR manuals. Specifically, see https://talkbank.org/morgrams/.

[2] Available at http://lindat.mff.cuni.cz/services/morphodita/.

[3] See https://github.com/ufal/morphodita.

A Python script was developed that strips the main lines of transcripts of all special CHAT-format marking (see Sect. 3.1), uses MorphoDiTa to process this stripped text (see Sect. 3.2), and transforms the output of MorphoDiTa into the morphological tier (%mor), subsequently added to the original CHAT-format transcripts (see Sect. 3.3).

3.1 Text pre-processing

Pre-processing the text of the main line is mostly a straightforward matter of stripping away the speaker ID at the beginning of each main line and omitting the special characters used in the CHAT formatting conventions (for an overview, see Appendix A). For instance, in the token *hol^čič^ka*, the carets (^) mark pauses between the syllables of the word *holčička* ('little girl') and should be ignored for the purpose of morphological analysis.

Sometimes the special characters need to be removed along with some other characters. For instance, in the line *&e není voda* ('there is no water'), the ampersand marks a phonological fragment that is to be ignored in morphological analysis, and thus the script transforms the line into *není voda* by omitting the token beginning with the ampersand. If such a phonological fragment constitutes the whole utterance, the script recognizes that such an utterance is not to be further analyzed and assigned a morphological tier.

Pre-processing the transcript also requires slightly more complex tasks, as illustrated by the following example:

*CHI: vidělas [: viděla jsi] někdy delfínka ?
%pho: vidělas někdy delfínka .
'have you ever seen a dolphin$_{\text{diminutive}}$?'

The square brackets specify that the "contracted" form *vidělas* 'did you see' should be morphologically analyzed as *viděla jsi* (the past participle of the lexical verb *vidět* 'see' and the auxiliary *být* 'be,' which is often reduced to the enclitic *-s* in spoken Czech). The first token *vidělas* is thus ignored in the morphological analysis, while the contents of the square brackets are analyzed (and the square brackets and the colon removed).

Finally, several CHAT markers are removed, but they affect the subsequent tagging: if a token ends in *@i*, *@z:ip*, *@z:ia*, or *@z:in*, marking interjections in the transcript, the script recognizes that these tokens are to be analyzed as interjections in the next step; if a token ends in *@c* (idiosyncratic forms) or *@n* (morphological innovations), the script attaches *-neo* at the end of its morphological tag; finally, for tokens with the marker *@z:c* (foreign expressions), the script attaches *-for* to their tags.

3.2 Morphological analysis with MorphoDiTa

The output of MorphoDiTa provides a Token object for each token of the input (the pre-processed text of the main line). This object contains the original token (i.e., the word form), its assigned lemma, and its assigned morphological tag. For example, the line *viděla jsi někdy delfínka?* 'have you ever seen a dolphin?' is processed as

Table 4 An example of output of MorphoDiTa

Token	Lemma	Tag
viděla 'saw'	*vidět* 'see'	VpQW----R-AAI--
jsi 'are'	*být* 'be'	VB-S---2P-AAI--
někdy 'ever'	*někdy* 'ever'	Db---------
delfínka 'dolphin'	*delfínek* 'dolphin'	NNFS1-----A----
?	?	Z:--------

Table 5 The positions of the morphological tag provided by MorphoDiTa

Position	Value	Found with
1	Part of speech (POS)	All tokens
2	Detailed part of speech	All tokens
3	Gender	Nominal POS and some verb forms
4	Number	All inflected tokens
5	Case	Nominal POS
6	Possessor's gender	Possessive pronouns and adjectives
7	Possessor's number	Possessive pronouns and adjectives
8	Person	Verbs and central (personal + possessive) pronouns
9	Tense	Verbs
10	Degree of comparison	Adjectives and adverbs
11	Negation (by prefix)	Lexical POS
12	Voice	Verbs
13	Aspect	Verbs
14	*Reserved for future use*	–
15	Variant, style, register	Lexical POS and (marginally) interjections

shown in Table 4 (resulting in a list of Token objects, here presented as a table for the sake of convenience).

The tag returned by MorphoDiTa has 15 positions introduced in Table 5, adapted from Hajič (2004). It should be noted that Hajič (2004) describes the thirteenth position of the tag as to be "reserved for future use," since it started to be used for the category of verbal aspect only recently. If a token does not express a particular morphological category, the corresponding position in the tag is filled with a hyphen (this is quite common, since the tag conflates nominal and verbal categories). As a testament to the morphological richness of Czech, note that Straková et al. (2014) report that there are 3922 plausible tags, even though only 1571 unique tags had actually appeared in training data.

3.3 Creating the %mor tier

The information provided automatically by MorphoDiTa was subsequently used by the script to generate the morphological tier (introduced by *%mor*) for each main line, as long as the main line contained some morphologically relevant material

(e.g., lines consisting of hesitation sounds only were not analyzed). The automatic generation of the morphological tier consisted of (1) transforming the morphological tag generated by MorphoDiTa for each token into a representation following the MOR format, and (2) inserting the transformed tags for all tokens on a main line on a separate dependent %mor tier.

3.3.1 The MOR format

The morphological tier is in line with the key features of the MOR format, and is thus readable by CLAN, but had to be adapted for Czech, a highly inflected language, typologically distant from English and the other languages for which MOR was primarily created.

In general, the MOR format mirrors the morphemic structure of tokens on the main line; for instance, the token *switched* is analyzed as *v|switch-PAST*, where *v* specifies the POS, and the morphological information attached to the stem with a hyphen represents the fact that the morpheme following *switch* expresses the past tense. In Czech, there is no one-to-one correspondence between morphemes and their functions, one morpheme typically expresses several morphological categories at once (e.g., the ending *-y* in *kočky* 'cats' functions at once as a marker of the feminine gender, the genitive case, and the singular), and morphonological alternations in the stem are frequent. Nevertheless, in the MOR format, the morphological information expressed through means other than a "well-behaved" suffix, e.g., through a vowel change, can be still specified after the lemma, except that in these cases, it is not separated by a hyphen but by an ampersand (e.g., *ran* is tagged as *v|run&PAST*).

We thus used the hyphen only to separate individual parts of the morphological tag (see the general format detailed in the following section), while most morphological information is included in the core of the tag and the values of individual morphological categories are separated by ampersands, irrespective of whether they are expressed cumulatively in a single morpheme. The main division is therefore not that of stem vs. affixes, as is usual in MOR grammars, but, instead, that of lemma vs. morphological categories, which is the norm in adult Czech language corpora annotated by MorphoDiTa as well and published by Czech National Corpus.[4]

3.3.2 The Czech MOR tag

The Czech MOR tag has the following general format:

POS:SPECIFICATION|LEMMA-AFFIX-INFLECTION-COMMENT

Two parts of the tag are obligatory for all tokens: PART-OF-SPEECH (POS) and LEMMA. The INFLECTION part is obligatory for inflected word classes. Within the INFLECTION part, the values of the categories are separated by ampersands. The SPECIFICATION of POS is used only with some parts of speech (it is practically always used with numerals, and it is obligatory with pronouns and conjunctions). The AFFIX and the COMMENT parts are rather infrequent. The AFFIX part might be used either for marking

[4] See https://www.korpus.cz/.

negated lexical words (value *neg-*), or for marking the comparative/superlative forms of adjectives and adverbs (value *CP-/SP-*). The COMMENT part serves for denoting additional specific information: *-for* marks foreign expressions, *-err* marks an erroneously used form, and *-neo* marks lexical innovations and idiosyncratic forms.

Table 6 provides a summary of POS codes and their possible specifications used in the corpus. The classification conforms to the Czech linguistic tradition (cf. for instance the reference grammars of Czech by Komárek et al. (1986) and by Štícha et al. (2018)) and mostly mirrors POS classifications (and further subclassifications especially of pronouns, verbs, and conjunctions) found in reference grammars of English (e.g., Quirk et al., 1985). What might come as a surprise to a reader unfamiliar with the details of the traditional Czech grammar description is particularly the classification of numerals. It is based on classifications found in grammars of Czech, which has a variety of different types of numerals. However, since some of them are stylistically marked and mostly not expected in spoken language, we applied only a reduced version of the traditional classification.

In special cases, there might be two POS specifications: for instance, the pronoun *svůj* 'my/your/… (own)' is both reflexive and possessive (hence *pro:refl:poss*) and the pronoun *čí* 'whose' is both interrogative and possessive (hence *pro:int:poss*).

Table 7 lists a summary of morphological categories by individual parts of speech mirroring the ordering of these categories in the morphological tag. If the precise value of any of the categories cannot be determined (for instance, due to syncretism and lack of context in the case of the case), the value provided is simply *x*. The table includes all possible categories, and not all of them are always specified; for instance, if a verb expresses the imperative mood, only its person, number, mood, voice, and aspect are given (while its tense and gender are not). For verbs, gender is given only with past and passive participles (cf. *pomohl* 'he helped' vs. *pomohla* 'she helped' vs. *pomohlo* 'it helped'). In the case of the infinitive, only the aspect is given, thus the perfective *pomoci* 'to help' is tagged as *v|pomoci-inf&pf*. (Note that the finiteness listed as a verbal category for the sake of convenience is technically not a morphological category and is not considered as one in the Czech linguistic tradition.)

The following morphological tags represent examples with different levels of complexity. First, the preposition *nad* 'over, above' does not express any morphological categories and its representation thus specifies only the POS and the lemma:

prep|nad

Second, the representation of the conjunction *ale* 'but' also includes a further POS specification:

conj:coord|ale

Third, the word *delfínka* 'dolphin$_{\text{diminutive.accusative}}$' is lemmatized as *delfínek*, classified as a noun (n), and it is identified as a form expressing the accusative case (*4*), the singular number (*SG*), and the masculine animate gender (*MA*):

n|delfínek-4&SG&MA

Table 6 Part-of-speech (POS) tags and their further specification in the *Chroma* corpus

POS	Meaning	Specification	Meaning	Example from the corpus
n	Noun	–	Common	*kočka* 'cat'
		pt	Plurale tantum	*brýle* 'glasses'
		prop	Proper	*Praha* 'Prague'
adj	Adjective	–	Long form	*mladý* 'young'
		short	Short form	*(být) rád* '(to be) glad'
		poss	Possessive	*tátův* 'dad's'
pro	Pronoun	*dem*	Demonstrative	*tento* 'this one'
		pers	Personal	*my* 'we'
		poss	Possessive	*náš* 'our'
		rel	Relative	*který* 'that, which, who'
		int	Interrogative	*který* 'which, who'
		neg	Negative	*nikdo* 'nobody'
		indef	Indefinite	*někdo* 'somebody'
		refl	Reflexive	*se* '-self, -selves'
num	Numeral	*card*	Cardinal	*pět* 'five'
		ord	Ordinal	*pátý* 'fifth'
		mult	Multiplicative	*dvakrát* 'twice'
		indef	Indefinite	*několik* 'a few'
		–	Another type	*kolik* 'how many/much'
v	Verb	–	Lexical	*spát* 'sleep'
		mod	Modal	*muset* 'must'
		aux	Auxiliary	*být* 'be'
		cop	Copular	*být* 'be'
adv	Adverb	–	"Lexical"	*doma* 'at home'
		pro	Deictic	*tady* 'here'
		pro:neg	Negative	*nikde* 'nowhere'
prep	Preposition		–	*nad* 'over, above'
conj	Conjunction	*coord*	Coordinator	*ale* 'but'
		sub	Subordinator	*protože* 'because'
part	Particle		–	*samozřejmě* 'of course'
int	Interjection		–	*mňau* 'meow'
x	Unidentified		–	*boji* '???'

Finally, the attested verb form *ozbílám* is not an existing word in Czech, and yet it is clearly a verb, and it clearly expresses the first person (*1*), the singular number (*SG*), the indicative mood (*ind*), the present tense (*pres*), and, in the respective context, the perfective aspect (*pf*). Since it is an innovative form, the COMMENT *-neo* is given after the morphological tag:

vlozbílám-1&SG ind&pres&pf-neo

Table 7 All possible non-null values of morphological categories in the *Chroma* corpus

POS	Category	Values
n	Case	1 (nominative), 2 (genitive), 3 (dative), 4 (accusative), 5 (vocative), 6 (locative), 7 (instrumental)
	Number	SG (singular), PL (plural)
	Gender	F (feminine), N (neuter), MA (masculine animate), MI (masculine inanimate)
adj/pro/num	Case	1–7
	Number	SG, PL
	Gender	F (feminine), N (neuter), M (masculine, both animate and inanimate)
v	Finiteness	inf (infinitive), *no value* (other forms)
	Person	1–3
	Number	SG, PL
	Mood	ind (indicative), imp (imperative), cond (conditional)
	Tense	pres (present), past (past), futur (future)
	Voice	akt (active), pas (passive)
	Gender	F (feminine), N (neuter), M (masculine, both animate and inanimate)
	Aspect	pf (perfective), impf (imperfective)

In many cases, it was impossible to determine the lemma for the innovative forms; we thus decided to put the attested form in the position for the lemma for all the *-neo* forms, rather than a constructed, potential lemma.

The compatibility of the Czech MOR tag with the general MOR format and the CLAN automatic analysis was approved by CLAN CHECK procedure (MacWhinney, 2000). Also, Matiasovitsová et al. (2023) performed an automatic CLAN analysis of Index of Productive Syntax (Scarborough, 1990) on another Czech data annotated by the same automatic procedure.

3.3.3 Deviations from MorphoDiTa annotation

In some cases, a deviation from the morphological analysis provided by MorphoDiTa was implemented directly in the script. First, a word that had been marked as an interjection manually during the transcription of the recordings was always tagged as an interjection, irrespective of whether MorphoDiTa marked it so or not. Second, a few selected words were always tagged in a way specified in the script; this includes, for instance, the particle *no*, always tagged as a particle by our script (MorphoDiTa often mistakenly tagged it as a noun, as if it were the abbreviation *no.* 'number'). Third, MorphoDiTa analyzes punctuation as well (punctuation marks are assigned a tag starting with *Z*), but in accordance with the CHILDES conventions, we decided not to tag punctuation marks; they thus appear on the morphological line, but are not tagged – see the full stop in the following example:

*CHI:&em nevím .
'&em I don't know .'

%mor:v|vědět-neg-1&SG&ind&pres&akt&impf .

Fourth, the script specified that a limited number of words were always assigned two parts of speech and two lemmas and, in some of these cases, two morphological tags. For instance, the Czech form *abychom* represents an aggregate of the subordinator *aby* 'to, so that' and the conditional form *bychom* of the auxiliary *být* 'to be'; it is thus POS-classified as both a subordinating conjunction and an auxiliary verb (the POS tags are separated by an underscore), and lemmatized as both *aby* and *být* (the lemmas are separated by an underscore as well). With respect to morphological categories, it is nevertheless tagged in the same way as the isolated form *bychom. Aby* is not reflected in the tag, since it is an uninflected conjunction:

conj:sub_v:aux|aby_být-1&PL&cond&akt&impf

The other word forms *abych*, *abys*, *aby*, *abyste*, and the non-standard variant of *abychom*, *abysme*, and the analogical forms starting with another subordinator, *kdyby* (i.e., *kdybych*, *kdybys*, *kdyby*, *kdybychom*, its non-standard variant *kdybysme*, and *kdybyste*) are treated in the same way. Additionally, the words *ses* and *sis* (the reflexive pronoun *se* or *si* + the singular second person verbal marker *-s*, used as an enclitic) and *zač* (*za* 'for' + *co* 'what,' used especially in the common expression *není zač* 'you are welcome,' literally 'isn't for what,' with the alternative form *není za co*) were also assigned to two parts of speech and were assigned two lemmas, and in the case of *ses* and *sis*, since they have (unlike *zač* and the *aby-* and *kdyby-*forms) two inflected components, two morphological tags, again separated by an underscore:

ses: pro:refl_v:aux|se_být-4&SG_2&SG&ind&pres&akt&impf
sis: pro:refl_v:aux|se_být-3&SG_2&SG&ind&pres&akt&impf

Finally, and this represents more of a technical difference from MorphoDiTa's analysis, the part of speech or some parts of the morphological tag were left underspecified in some cases, only to be later identified manually. On data from three children (Anna, Jan and Klára), we analyzed the proportion of POS underspecifications in the output of the automatic annotation. They are given in Table 8.

Naturally, underspecifications and errors in the POS-specified tags were also expected. As a second step, manual checking was thus performed on the automatic output with the following purposes: (1) to check the accuracy of the automatic annotation, (2) to specify the underspecified annotation, (3) to correct errors in automatic annotation, and (4) to correct errors in the transcript on the main lines if necessary.

3.4 Manual checking

All children's lines were revised and checked manually by eight trained research assistants. This included completing the annotation of tokens that had been underspecified in the automatic annotation. The completion concerns particularly cases in which the identification of the further specification of a POS tag is dependent on syntactic criteria rather than purely morphological ones. For instance, MorphoDiTa

Table 8 Proportion of underspecified POS (specification) in the output of automatic annotation for three children (Anna, Jan, Klára)

		N	% of all tokens
Undetermined POS (unrecognized form)	x|	684	6.1
Underdetermined POS specification	v:aux/cop|	1821	16.2
	pro:rel/int|	290	2.6

doesn't make the distinction between auxiliary, copula or lexical use of the verb *být* 'to be', and so the script had automatically tagged the verb as *v:aux/cop*, which was then changed manually to *v:aux*, *v:cop*, or *v* (thus, the number of occurences in Table 8 represents all occurences of the verb *být* in the three children). Similarly, the script had tagged all pronouns functioning as both relatives and interrogatives as *pro:rel/int*, which was later changed manually either to *pro:rel* or to *pro:int*. In problematic cases, the undecided annotation *aux/cop* or *rel/int* might have been preserved.

In other cases of underdetermined or entirely absent values, manual disambiguation or completion was also performed whenever possible. In many cases, for instance, the aspect of the verb had not been marked as either *pf* or *impf* but was underspecified by MorphoDiTa. Especially with children's idiosyncratic and innovative words which had not been recognized (and thus were assigned the technical POS *x* and further unanalyzed), the POS and further morphological information were supplemented manually whenever possible.

Furthermore, given the moot classification of various (especially) uninflected words, several specific rules had been formulated to ensure consistency. For example, words frequently used in one-word utterances such as *ano* 'yes,' *jo* 'yes,' and *ne* 'no' are sometimes considered as particles and sometimes as interjections in the Czech literature (cf. the overview by Vondráček (1998)). We decided to treat these words consistently as interjections whenever they were used as a one-word utterance; this is in accordance with their classification in *Akademický slovník současné češtiny* [The Academic Dictionary of Contemporary Czech], which is currently being written,[5] and whenever the automatic tagging had not been in accordance with this decision, it was manually corrected.

Finally, the main lines were revised by the research assistants as well, eliminating problems such as typing errors and inconsistencies in the CHAT codes.

[5] The dictionary, published by the Czech Language Institute of the Czech Academy of Sciences, is available at https://slovnikcestiny.cz/, and its dictionary-making principles were published in the monograph edited by Kochová and Opavská (2016). Entries starting with *a-* to *f-* have already been published, and entries starting with *h-* to *j-* are at different stages of revisions and editing. One of the co-authors of this paper is a co-author of the dictionary, which granted us access to unpublished entries as well and helped us ensure that the POS-classification in the corpus is in accordance with the most widely accepted POS treatment of moot words in the Czech linguistic tradition.

4 Evaluation of the morphological tagging

While "MorphoDiTa reaches state-of-the-art results for Czech" with success rates upwards of 95% (Straková et al., 2014, pp. 15–16), it was primarily trained for written standard Czech (even though it has also been used successfully for spoken corpora of Czech). For this reason, we were interested in the accuracy of automatic analysis in a corpus of the spontaneous language production of children. In this section, we therefore evaluate the automatic morphological analysis of the corpus, comparing a sample of the automatically annotated data and the same data after manual checking, focusing on the accuracy of the POS classification (Sect. 4.1), including further POS specification (Sect. 4.2), and information regarding morphological categories (4.3). This section illustrates that the automatic analysis turned out to be more reliable than might have been expected, as shown by the relatively low numbers of manual corrections needed. This also shows that the automatic morphological analysis of utterances produced by adults, which have crucially not been checked manually, can still be considered as reliable, albeit with limitations.

4.1 Part of speech

As shown in Table 1, the corpus contains 42,103 utterances produced by children. For the evaluation of the POS tagging, we used the data from three children (Anna, Jan, and Klára), totaling 11,237 utterances, which amounts to roughly one quarter of the corpus (27%). Since they have the smallest collections within the corpus, the data from these children was the only one with completed manual checking at the time of the evaluation analysis.

For the first comparison, we contrasted numbers of tokens assigned to various parts of speech before and after the manual checking; see Table 9. The first row, for instance, shows that there had been 4552 automatically identified nouns; after manual checking, there were 4,786 tokens tagged as nouns in the sample, yielding the difference of 234 tokens.

As the table shows, during the manual checking 357 tokens were removed. These were tokens that were not supposed to be morphologically tagged (e.g., hesitation sounds and remnants of the CHAT format that had erroneously not been removed before tagging, cf. 3.3.1), but the script had processed and tagged them, before it was adjusted in order to not process such cases in further annotation.

Table 9 also shows that, relative to the total number of tokens assigned to the respective parts of speech, the differences between the automatic analysis and the final data are rather small for lexical words (nouns, adjectives, verbs, and adverbs) and for pronouns, numerals, and prepositions.

The difference in the numbers of unidentified tokens was to be expected; roughly 60% of the 684 tokens that had not been POS-classified by MorphoDiTa automatically could be assigned a POS manually. The remaining number of unidentified tokens includes mostly homonymous function words in incomplete utterances and idiosyncratic child words without clear clues, which would enable assigning of POS.

The first two examples below (IDIO1 and IDIO2) represent idiosyncratic words with the code @c which are transparent corruptions of adult words. In both cases,

Table 9 POS counts before and after manual checking in the data from three children from the *Chroma* corpus (Anna, Jan, and Klára)

POS	Automatic	Manually checked	Difference
Noun	4552	4786	+234
Adjective	647	653	+6
Pronoun	4257	4149	− 108
Numeral	238	234	− 4
Verb	5468	5449	− 19
Adverb	2912	2855	− 57
Preposition	683	667	− 16
Conjunction	905	810	− 95
Particle	1142	632	− 510
Interjection	760	1376	+616
Unidentified	684	280	− 404
Total	22,248	21,891	− 357

the corruption consists in omitting one syllable: the middle one in *ja-ho-dy* (IDIO1) and the first one in *vr-tul-ník* (IDIO2). In these cases, POS and some or all of the morphological categories were determined manually. The example IDIO4 represents an idiosyncratic word with identifiable meaning (thanks to the context) but formally very distant from any adult form. In this case, the word remained undetermined. The example IDIO3 represents a form that occurs repeatedly, it seems from the context that it has some meaning for the child, but still we were not able to determine its meaning (nor its formal affiliation to a POS).

IDIO1
*CHI:jadi@c . (Anna 1;09.30)
%mor:n|jadi-1&PL&F-neo .
%pho:jadi .
%com:jadi = jahody ‘strawberries’ .

IDIO2
*CHI:tulník@c . (Jan 1;07.19)
%mor:n|tulník-x&SG&MI-neo.
%com:tulník = vrtulník ‘helicopter’.

IDIO3
*CHI:+ < iše@c . (Anna 1;09.30)
%mor:x|iše-neo .
%pho:iše .
%com:význam nejasný ‘unclear meaning’ .

IDIO4
*CHI:tá@c . (Jan 1;07.19)
%mor:x|tá-neo.
%com:tá = letadlo ‘airplane’.

Finally, the relatively higher differences in the numbers of conjunctions, particles, and interjections were also expected, given the high degree of homonymy among function words in Czech. For instance, some of the most common words in Czech include *a*, most frequently used as a coordinator corresponding to the English ‘and’, but used also as a particle and an interjection; the same applies to other rather frequent words such as *ale* (the coordinator ‘but’ × particle × interjection) and *co* (the pronoun ‘what,’ typically easily differentiated from the other uses: deictic adverb × conjunction × particle × interjection), see e.g. the description of these words in *Akademický slovník současné češtiny* [The Academic Dictionary of Contemporary Czech], mentioned above, and the underlying dictionary-making principles detailed in Kochová and Opavská (2016).

It thus appears that in terms of POS-tagging, MorphoDiTa was successful to a perhaps surprising degree, given the fact that it had been primarily trained for written Czech texts, and the manual changes to its annotation can be mostly accounted for by appealing to linguistic factors, such as the high degree of homonymy that is rather typical of Czech uninflected words.

4.2 POS specification

Apart from changes accounted for at the level of POS (especially the decrease in the number of conjunctions with the specification *coord* discussed in the previous section), the numbers of changes in POS specifications are mostly in single digits, and thus negligible, with a few exceptions, shown in Table 10. The first row points towards the fact that 128 nouns (tagged as *n*) were re-tagged as proper nouns (*n:prop*) during the manual checking. The second row shows that 118 tokens tagged automatically as demonstrative pronouns were later re-tagged manually; a few of these tokens were reassigned to personal pronouns (*pro:pers*), but most of them were reclassified as particles. This concerns especially the central demonstrative in Czech, *ten* 'this, that,' whose inflected forms include *ty*, homonymous with the personal pronoun *ty* 'you (sg.)', and *to*, homonymous with the particle *to*. The penultimate row shows that 54 tokens tagged as lexical verbs (*v*) were retagged as modal verbs (*v:mod*); this follows from the fact that MorphoDiTa does not differentiate between lexical, modal, copular, and auxiliary verbs, and that the script had specified only a very restricted set of verbs that were to be marked as modal (only the core modals *moci* 'can, to be able to,' *muset* 'must, to have to,' and *smět* 'may, to be allowed to').

The other rows of the table are not relevant in this context, since they concern what was mentioned at the end of Sect. 3.3.3., i.e., the fact that MorphoDiTa does not distinguish between lexical (*v*), auxiliary (*v:aux*), and copular (*v:cop*) uses of verbs. The verb *být* 'be' had thus always been automatically tagged as *v:aux/cop* and later classified manually (hence the steep decrease in *v:aux/cop* and the increase in *v:aux* and *v:cop*). The same applies to *pro:rel/int* (a pronoun that can function as a relative or an interrogative), which was manually changed to *pro:int* (interrogative) in the vast majority of the 289 cases.

It thus appears that the automatic analysis had been highly successful, and the specific issues that might be difficult to implement automatically were successfully dealt with manually, including the identification of named entities and the differentiation between different functions of *být* 'to be' or polyfunctional pronouns.

4.3 Morphological categories

For the detailed manual evaluation of the information about changes in morphological categories, we analyzed the data from one child (Anna), totaling 3,127 utterances (7% of the corpus). Anna's subcorpus was selected for this detailed analysis as it is the shortest subcorpus in the whole *Chroma* corpus. In 1,298 utterances (41.5%), a change was made manually. The most frequent types of changes in Anna's subcorpus

Table 10 The most frequent changes in POS specification in the data from three children from the *Chroma* corpus (Anna, Jan, and Klára)

POS	Specification	Automatic	Manually checked	Difference
n	*prop*	303	431	+128
pro	*dem*	2051	1933	− 118
pro	*int*	0	275	+275
pro	*rel/int*	290	1	− 289
v	–	3469	4001	+532
v	*aux*	0	256	+256
v	*aux/cop*	1821	28	− 1793
v	*mod*	178	232	+54
v	*cop*	0	757	+757

are presented in Table 11 (including all types of changes, i.e., even those discussed in the previous two subsections).

The high frequency of the changes in rows 1 to 4 and the number of changes in row 8 follow from what was discussed in the previous subsections. The other, less frequent types of changes (concerning the morphological tagging proper) follow directly from linguistic properties of Czech that have already been mentioned: in the case of nouns (row 5), the main source of changes was the homonymy of nominative and accusative forms found both in the singular and in the plural of neuter and masculine inanimate nouns and of two of the four major types of feminine paradigms, and in the plural of the remaining types of feminine paradigms. In pronouns whose gender was changed manually (row 6), this followed from the fact that selected pronouns (in the data, this concerned most frequently the basic demonstrative, *ten* 'this, that') do not (always) express gender formally, but their gender can often be determined only on the basis of agreement with their head noun, especially in masculine and neuter singular forms of indirect cases (e.g., *s* ***tím*** *psem* 'with that dog_{MA}' and *s* ***tím*** *autem* 'with that car_N' vs. s ***tou*** *kočkou* 'with that cat_F') and in the plural across all genders (e.g., *s* ***těmi*** *psy* 'with those $dogs_{MA}$,' *s* ***těmi*** *auty* 'with those $cars_N$,' and *s* ***těmi*** *kočkami* 'with those $cats_F$' – with the substandard variant *těma* for all genders as well).

In the case of adjectives (row 7), the changes often concerned neuter forms mistagged as masculine; this stems from the fact that while in standard Czech, adjectives of the "hard" declension (such as *nový* 'new') end in *-ý* (in their representative, nominative singular form) when masculine, and in *-é* when neuter, in informal spoken language this distinction is leveled for many speakers, and both neuter and masculine forms take the same ending, *-ý*. Just as with some pronouns, the gender can be determined only on the basis of agreement, and it is not entirely surprising that this might be problematic in the tagging of informal spoken language, which displays characteristics that might further complicate the task, such as the frequent occurrence of ellipsis. Finally, in the case of verbs, most corrections consisted in changing the number of verbs that had been mistagged as singular; this concerned exclusively verbs whose third person

Table 11 The most frequent types of changes in the annotation of the data from one child from the *Chroma* corpus (Anna)

	Type of change	Occurrences
1	Distinguishing *v:aux* and *v:cop* with *být* 'to be'	544
2	Determining the aspect of a verb	183
3	Reclassifying a particle as an interjection	139
4	Distinguishing *pro:rel* and *pro:int*	129
5	Changing the morphological tag of a noun	89
6	Dsambiguating the gender of a pronoun	85
7	Changing the morphological tag of an adjective	66
8	Changing *v* to *v:mod*	34
9	Changing the morphological tag of a verb	23

forms end in *-í* in both the singular and the plural (with verbs of other types of conjugation, the third person singular form and the third person plural form are clearly distinguished). In other words, homonymy of forms within paradigms was responsible for most of the errors in the automatic morphological tagging of nouns, adjectives, pronouns, and verbs.

5 Conclusion

In the present paper, we have introduced a new longitudinal corpus of data on Czech language acquisition from seven children aged between 1.5 and 3.5 years. Thanks to its morphological annotation, the corpus is an important resource for language acquisition research concerning various linguistic areas, especially those of morphology and syntax. The annotated corpus should serve as a source, e.g., for studying growth of morphological complexity in inflectional POS; variability of children's overgeneralization and their recovery from it; or dealing with morphological synonymy. Czech morphology is abundant in various inflectional classes and inflectional variants with the same function, as well as in number of different forms within inflectional paradigms. All of this make our corpus to be a rich source for studying such topics in acquisition of morphology.

The evaluation of the automatic morphological annotation of *Chroma* corpus reached a conclusion similar to that of Santos et al., (2014, p. 1491), who note for their Portuguese corpus that "given a set of well-crafted rules, a statistical model trained and developed for written material can be ported to POS-tag and lemmatize spoken data from children with almost the same performance." In the case of the present Czech corpus, it holds true for identifying the values of morphological categories as well. Based on this, the automatic annotation of the adult utterances in *Chroma* corpus (which has not yet been manually checked) can be considered reliable enough, albeit with limitations, as suggested above. We can conclude that the automatic part of the annotation process (combining MorphoDiTa with the Python script) can be applied as such for processing further child Czech speech samples with good enough results without manual checking. An analysis of the Index of

Productive syntax (IPSyn, Scarborough, 1990) was already performed by Matiasovitsová et al. (2023) on another Czech data with the same automatic procedure. IPSyn might be calculated automatically by the CLAN software, supposing the transcript is provided with morphological annotation in the MOR format. Using transcript-based measures, including IPSyn calculated from an automatically annotated child speech sample, seems to be a promissing tool for the future of developmental language assesment (see e.g. Mooney et al., 2021). We would be very satisfied if such an application of our method would find its way into Czech speech therapist's praxis.

In near future, we intend to apply the automatic annotation procedure on our second child longitudinal corpus in progress (with seven another, a bit younger Czech children, based on videorecordings). Even if we appreciate the high accuracy of the automatic morphological analysis done by MorphoDiTa, we believe that manual revision has added significant value to the *Chroma* corpus by enhancing the accuracy of its annotation, as is desirable for a corpus of this relevance. We thus intend to apply it to our second corpus as well, after further evaluation and optimization of the manual procedure.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10579-023-09710-y.

Acknowledgements The creation of the *Chroma* corpus has been supported by the *Ministry of Education, Youth and Sports of the Czech Republic*, projects No. LM2018101 and LM2023062 (LINDAT/CLARIAH-CZ). The morphological annotation of the corpus has been supported by the funding program *Grant Schemes at Charles University* (CZ.02.2.69/0.0/0.0/19_073/0016935). We are grateful to the following students who participated in the transcription of recordings, the revision of transcripts, and the manual control of the automatic morphological annotation (in alphabetical order): Markéta Baslová, Kateřina Bělehrádková, Tereza Binderová, Barbora Blahnová, Iurii Bochkov, Jan Henyš, Alžběta Macháčková, Anna Marklová, Martin Pavlíček, Jan Pinc, Tereza Šátavová, Denisa Šebestová, Jana Segi Lukavská, Kateřina Šimková, Leona Straková, Tomáš Treichel, Štěpánka Tvrdíková, and Martina Vokáčová. We also thank our collaborator Petra Čechová who participated in the process of morphological annotation, and our mentor, Filip Smolík, whose contribution to the entire project has been invaluable.

Author contributions Conceptualization and supervision [AC, KM]; project administration [AC, KM, JT]; data curation [AC, KM, JT]; methodology [KM, JS]; software [JS]; formal analysis [KM, JS, AC]; writing the original draft [AC, JS]; writing—review and editing [AC, JS, KM, JT].

Funding This work was supported by the *Ministry of Education, Youth and Sports of the Czech Republic* (Grants No. LM2018101 and LM2023062 LINDAT/CLARIAH-CZ) and by the funding program *Grant Schemes at Charles University* (CZ.02.2.69/0.0/0.0/19_073/0016935).

Data availability The *Chroma* corpus is openly available in the CHILDES database (https://childes.talkbank.org/) and in the LINDAT database (https://lindat.cz/) in several versions.

Code availability The MorphoDiTa software used for the morphological tagging is a free software available at https://ufal.mff.cuni.cz/morphodita. The custom Python script written by JS is available at https://github.com/slamajak/The-Chroma-corpus.git.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethical approval The approval of the Research Ethics Committee of Charles University was gained in 2018 under No. 2018UKFF01620.

Informed consent For each participating child, both parents as well as other participating caregivers gave a written informed consent for the use and publication of their anonymized data.

References

Akademický slovník současné češtiny [The Academic Dictionary of Contemporary Czech]. (2017–2023). Czech Language Institute of the Czech Academy of Sciences. https://slovnikcestiny.cz/

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

Chromá, A. & Matiasovitsová, K. (2022). CoCzeFLA Chroma 2022.07. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-4772

Chromá, A., Sláma, J., Matiasovitsová, K., & Treichelová, J. (2023a). CoCzeFLA Chroma 2023.07, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-5183

Chromá, A., Sláma, J., Matiasovitsová, K., & Treichelová, J. (2023b). Chromá Czech Corpus. CHILDES [www.childes.talkbank.org]. https://doi.org/10.21415/3ZNE-HX03.

Hajič, J. (2004). *Disambiguation of Rich Inflection: Computational Morphology of Czech*. Univerzita Karlova v Praze / Nakladatelství Karolinum.

Hnátková, M., Křen, M., Procházka, P., & Skoumalová, H. (2014). The SYN-series corpora of written Czech. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the 9th international conference on language resources and evaluation (LREC 2014)* (pp. 160–164). European Language Resources Association.

Kochová, P., Opavská, Z. (Eds.). (2016). *Kapitoly z koncepce Akademického slovníku současné češtiny*. Ústav pro jazyk český AV ČR, v. v. i.

Komárek, M., Kořenský, J., Petr, J., & Veselková, J. (Eds.) (1986). *Mluvnice češtiny 2: Tvarosloví*. Academia.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. 3rd Edition. Lawrence Erlbaum Associates.

Matiasovitsová, K., Čechová, P., Sláma, J., Homolková, K., & Smolík, F. (in press). Mean length of utterance in Czech toddlers: Validity estimates and comparison of words, morphemes and syllables. *Journal of Speech, Language, and Hearing Research*.

Matiasovitsová, K., Čechová, P., Sláma, J., Treichelová, J., & Smolík, F. (2023). The validity of a transcript-based measure of child language development in Czech. In P. Gappmayr, & J. Kellogg (Eds.), *BUCLD 47: Proceedings of the 47th annual Boston University Conference on Language Development* (pp. 533–547). Cascadilla Press.

Mooney, A., Bean, A., & Sonntag, A. M. (2021). Language sample collection and analysis in people who use augmentative and alternative communication: Overcoming obstacles. *American Journal of Speech-Language Pathology, 30*(1), 47–62.

Potratz, J. R., Gildersleeve-Neumann, C., & Redford, M. A. (2022). Measurement properties of Mean Length of Utterance in school-age children. *Language, Speech, and Hearing Services in Schools, 53*(4), 1088–1100. https://doi.org/10.1044/2022_LSHSS-21-00115

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

Santos, A. L., Généreux, M., Cardoso, A., Agostinho, C., & Abalada, S. (2014). A corpus of European Portuguese child and child-directed speech. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the 9th international conference on language resources and evaluation (LREC 2014)* (pp. 1488–1491). European Language Resources Association. https://repositorio.ul.pt/handle/10451/30661

Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics, 11*(1), 1–22. https://doi.org/10.1017/S0142716400008262

Spoustová, D., Hajič, J., Raab, J., & Spousta, M. (2009). Semi-supervised training for the averaged perceptron POS tagger. In A. Lascarides, C. Gardent, & J. Nivre (Eds.), *Proceedings of the 12th conference of the European chapter of the ACL* (pp. 763–771). Association for Computational Linguistics.

Straková, J., Straka, M., & Hajič, J. (2014). Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In K. Bontcheva, & J. Zhu (Eds.), *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations* (pp. 13–18). Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-5003

Štícha, F. (2018). *Velká akademická gramatika spisovné češtiny: 1. Morfologie: Druhy slov / Tvoření slov*. Část 1. Academia.

Vondráček, M. (1998). Citoslovce a částice – hranice slovního druhu. *Naše Řeč, 81*(1), 29–37.

Authors and Affiliations

Anna Chromá[1] · Jakub Sláma[1,2] · Klára Matiasovitsová[1] · Jolana Treichelová[1]

✉ Anna Chromá
anna.chroma@ff.cuni.cz

Jakub Sláma
slama@ujc.cas.cz

Klára Matiasovitsová
klara.matiasovitsova@ff.cuni.cz

Jolana Treichelová
jolana.treichelova@ff.cuni.cz

[1] Faculty of Arts, Charles University, Prague, Czech Republic

[2] Czech Language Institute of the Czech Academy of Sciences, Prague, Czech Republic