PRETHODNO PRIOPĆENJE
UDK: 811.163.42:811.131.1
81'246.2-053.2
81'322
DOI: https://doi.org/10.22210/strjez/53-1/4
Primljen: 1. 4. 2024.
Prihvaćen: 26. 7. 2024.

# Building MaLi, a Croatian-Italian bilingual child corpus

Mia Batinić Angster
mbatinic@unizd.hr
*Sveučilište u Zadru*

Marco Angster
mangster@unizd.hr
*Sveučilište u Zadru*

The knowledge we have about language and first language acquisition would not have been unveiled in the absence of previous efforts in collecting language data, e.g., recording the spontaneous interactions between children and adults. The CHILDES database (MacWhinney, 2000) gathers child speech in many of the world's languages, including Croatian (documented in Kovačević's 2002 corpus). In this paper, we describe the construction of MaLi, a corpus documenting the language productions of two bilingual children acquiring Croatian and Italian simultaneously. After a short survey of the methods used in collecting child language data with special regard to diary notes and audio recordings, we discuss the background and the details of the data collected in MaLi: we provide an overview of the sociolinguistic context of bilingual first language acquisition of the children observed and a description of the structure of the corpus. We first devote our attention to the data collection, management, and coding of the diary notes. Afterwards, we examine the collection and elaboration of the audio recordings and their ongoing transcription. In our concluding remarks, we offer a short assessment of the advantages and limits of the corpus along with a survey of the future possibilities for the use of this resource.

Keywords: *language documentation, child language, corpus, early bilingualism, simultaneous bilingualism, first language acquisition*

## 1. INTRODUCTION

Much of the knowledge we have about language and linguistic phenomena would not have been discovered without prior efforts in collecting language data, an aspect which is particularly evident in the research of first language acquisition. Documentation of early linguistic development involves the naturalistic observation of the progress of a child's linguistic competence by recording a child's spontaneous speech production with the aim of compiling a data corpus as a resource for researchers interested in child language acquisition.

Working with children is a challenging enterprise. On the one hand, the young age of children asks for a series of ethical requirements to be satisfied. On the other hand, the conditions of data collection require a higher degree of flexibility compared to working with adults, because of the numerous possible organizational issues (e.g., the child's momentary disposition, coping with parents' schedules). Furthermore, the longitudinal nature of conducting such studies presents a demanding and challenging task, both for researchers and for the participants involved, often at least a child and one caregiver, if not the members of entire household. The fact that these imply numerous sessions through several years carries a high risk of decline in the interest of the participants. The recruitment of interested and motivated participants brings about its difficulties, especially if there are no funds available for encouraging the participation of the family whose child is being observed.[1] Thus, it is not surprising that the compilers of corpora, as is the case in most of the available resources and as stated by Kuvač & Palmović (2007, pp. 15–16), are very often researchers who are parents of the observed children – psychologists, linguists, speech therapists, etc. They acknowledge the importance of conducting such studies and at the same time of having the expertise, tools, and assistance necessary to document their children's first language acquisition.

In the Croatian context, a corpus documenting child language exists (Kovačević, 2002), but there is a lack of corpora of early bilingual development for which one of the languages acquired is Croatian. In this paper, the ongoing longitudinal documentation of the linguistic development of two bilingual children who are simultaneously acquiring Croatian and Ital-

[1] To that point it is worth mentioning that, for the creation of the Max Planck dense database (Lieven et al., 2003), as Lieven & Behrens (2012, p. 234) report, a fee – "the equivalent of 50% of a clerical salary" – was offered to families that participated in the study.

ian from birth will be presented. The aim of this collection of naturalistic data is the creation of a corpus, called MaLi, which combines two sampling techniques: (a) diary notes in which the parents – both linguists – record their children's spontaneous language production in the form of short utterances accompanied by comments, and (b) audio-recordings of spontaneous interactions carried out between the children and three adults in both Croatian and Italian as target languages.

## 2. COLLECTING CHILD LANGUAGE DATA

In this section, we will survey the literature concerning the methodology of data collection in first language acquisition. Section 2.1 is focused on reviewing the types of data, with special attention to diary notes and audio-recordings. In section 2.2 we will briefly turn our attention to bilingual first language acquisition and data.

### 2.1. Corpus-based data collection

Keeping a diary has been one of the first ways of monitoring (early) language development, as reported in Demuth (2008, p. 199) and Rowe (2012, p. 193), who mention the observations present in nineteenth-century works such as Darwin (1877), or Deville (1891). As Kuvač & Palmović (2007, pp. 15–16) report, the first diaries about first language acquisition have been compiled by psychologists and linguists taking notes about their own children – e.g., Preyer (1882), Leopold (1939–1949), Gvozdev (1928). Apart from these first accounts devoted to general observations, researchers started to keep diaries devoted to specific topics: e.g., Bowerman (1974) reports on her daughter's generalizations with special regard to causative alternations; Braunwald & Brislin (1979) focus on novel structures produced by two sisters, namely Braunwald's daughters; and Tomasello (1992) analyses his daughter's use of verbs (cf. Behrens, 2008, p. xiii; Demuth, 2008, p. 199).

Diaries, as one of the so-called off-line methods of collecting child language, enable longitudinal observation of child language development (cf. Kuvač & Palmović, 2007, p. 22; Caselli et al., 2015, pp. 59–61). The methodology used in diary notes varies greatly from researcher to researcher. Diaries can be comprehensive or topic-specific, but they can also differ in format and size (cf. Behrens, 2008, pp. xii–xiv), as well as in the frequency of taking notes, the latter aspect being crucially dependent on the parents'/researchers' aim and motivation. The development of technology has enabled the digitalization of some diaries (cf. Behrens, 2008; Rowland et al.,

2008) – i.e., their transferral "into electronic databases", as Behrens (2008, p. xviii) points out, making their electronic versions available to researchers.

Technological progress also enables the application of other methods of monitoring and analysing child language development that include both corpora and experimental methods. As for the latter, thanks to versatile computer programs and applications that are accessible today, it is possible to record targeted, precise and explicit language reactions of a large number of children and even infants, using certain stimuli (cf. Caselli et al., 2015, pp. 57–59). Experiments can be aimed at researching language productions, language processing, and comprehension measuring precisely a series of variables important to researchers, such as reaction times (see, e.g., Swingley, 2012; Fennell, 2012).

The advantages of experimental methods notwithstanding, longitudinal documentation of first language(s) acquisition using diaries, according to Caselli et al. (2015, 59–61), remains a valuable and irreplaceable source of data enabling monitoring of the language development course of an individual child, offering the researcher the possibility of observing a number of variables that can influence the process of first language acquisition. First language acquisition can be documented via audio or video recordings of spontaneous interactions – productions – between children and adults. When these are transcribed, child language corpora are created and can be more easily shared and made publicly available (cf. Caselli et al., 2015). The compilation of child language corpora, their availability and usability in child language research has been possible thanks to the development of computers, audio and video recording equipment, programmes for language processing, and different storage possibilities. At any rate, the aims and the time and funds researchers have at their disposal (cf. Behrens, 2008, p. xv) dictate the techniques used and sampling regime undertaken in recording interactions.

Documenting early language development by corpus creation is not new, and numerous corpora created in this way are publicly available to researchers of child language development in the CHILDES database (*Child Language Data Exchange System*, MacWhinney 2000, http://childes.psy.cmu.edu; also available through Sketch Engine). CHILDES brings together transcripts of audio- and video-recordings of children's productions in different languages and among different clinical populations. There are few existing corpora that monitor the development of the Croatian language. Kovačević (2002) presents a longitudinal corpus of Croatian child language by recording three children up to the age of approximately three in spon-

taneous interaction with adults. This kind of project might appear to have a narrow scope; however, we must take into consideration the following statement by Behrens (2008, p. xviii): "Collecting spoken language data, especially longitudinal data, is a labour-intensive and time-consuming process." Given these difficulties, spoken corpora, even if small – especially if compared to the size of written corpora – can nonetheless be a precious source of insights into different aspects of the language under scrutiny (see Kuvač Kraljević & Hržica, 2016). Along with more spontaneous interactions, guided narrative productions can also be exploited for the study of first language acquisition. Currently, a Croatian narrative child corpus (Hržica & Roch, 2021) is being developed with the aim of researching child storytelling abilities, which is based on short recordings of the speech of Croatian children – some of which are bilingual – between the ages of three and six.

Almost everything we know today about the early acquisition of Croatian was discovered thanks to the Corpus of Croatian Child Language (Kovačević, 2002) based on the observation of three children.[2] However, although there have been studies on bilingual development with Croatian as one of the acquired languages (e.g., Medved Krajnović, 2004, on Croatian-English bilingual acquisition; Čamber, 2020, on Croatian-German bilingual acquisition), there are no available child language corpora attesting early bilingual development where one of the acquired languages is Croatian. On the other hand, some of the bilingual corpora available on CHILDES comprise data collected among children who acquire Italian simultaneously with another language, e.g., van Oosten's (2005) corpus of Italian-Dutch bilingual children.

## 2.2. Bilingual first language acquisition and data

It is known that early bilingual development is in itself characterized by peculiarities compared to early monolingual development, and that the specific language combination, that is, the characteristics of the languages to which children are exposed, is also an important aspect to be taken into account.[3] Although growing up in bilingual and multilingual contexts is a

[2] Even the recently published Croatian frequency dictionary of child language (Hržica et al., 2022, Kuvač Kraljević et al., 2022) is based on the Corpus of Croatian Child Language (Kovačević, 2002).

[3] Among the peculiarities of early bilingual development, we can mention its higher cognitive challenge, which can lead to developmental delays in one or both languages (Hržica et al., 2015: 36). Another issue mentioned in the literature (Caselli et al., 2015: 169) concerns possible disadvantages

common phenomenon today, and there is a large Italian minority in Croatia, which is why Croatian-Italian bilingualism is not uncommon, especially in one part of it – i.e., in Istria and in the city of Rijeka – the characteristics of Croatian-Italian bilingual development have not been defined to date. It is important to discover the differences between early monolingual (Croatian) and early bilingual (Croatian-Italian) development – for example, with regard to the age and pace at which certain language categories are acquired – in order, among other things, to gain knowledge that can feed the practice, e.g., in speech therapy treatment. As pointed out by Hržica et al. (2015, p. 36), there are no developmental scales for bilingual children who acquire Croatian and Italian at the same time, so that such children in Croatia do not have available a speech therapy assessment that would reveal whether there is a kind of early bilingual developmental deviation. Moreover, in the literature it is observed that an equal percentage of both the bilingual and the monolingual population is affected by a developmental language disorder, which corresponds to approximately 7% of children that do not achieve the expected progress for some reason (Detić & Kovačević, 2020, p. 48; Peña et al., 2011). As Detić & Kovačević (2020) point out, bilingual children pose a challenge for speech therapists in Croatia, who report the need for new insights regarding early bilingual development.

Documenting early bilingual language productions can shed light on bilingual L1 patterns, which is a prerequisite for discovering the type and nature of deviations in bilingual (in this case, Croatian-Italian) development (cf. Palmović & Kuvač, 2007, p. 8 for monolingual children). In addition, analysing bilingual language data can also highlight a number of language contact phenomena (such as borrowing, code-switching, and transfer) which are relevant for the study and further development of models of bilingual language competence. In this resides the motivation for starting a data collection such as the one described in section 3.

## 3. BUILDING MaLI

Building a bilingual child language corpus (MaLi) is one of the aims of an ongoing longitudinal bilingual first language acquisition research study approved by the Ethical Committee of the University of Zadar. The research

in the acquisition of the lexicon of the individual languages. In addition, code-switching and transfer are among the peculiar features of the language productions of bilingual speakers (Angster & Batinić Angster, 2022).

study complies with the *Ethical code for conducting research with children* (Ajduković & Keresteš, 2020). Two bilingual siblings have been observed in the study, brother and sister, whose context of acquisition is outlined in section 3.1. The research study is characterized by a combination of sampling techniques, which will be detailed in sections 3.2 to 3.4.

### 3.1. Participants: a sociolinguistic context of bilingual first language acquisition

The observed children have been simultaneously exposed to Italian and Croatian from birth. Both languages belong to the Indo-European linguistic family: Croatian, a South-Slavic language, belongs to the Slavic branch, whereas Italian is a Romance language. The siblings were born to a Croatian mother and an Italian father. Their mother speaks Croatian as her mother tongue and Italian as a foreign language, having C1/C2 competence in the latter. She started studying Italian in school at the age of 10, and she later received a university degree in Italian Language and Literature. The father's mother tongue is Italian (Piedmontese regional variety),[4] and he started learning Croatian at 34 years of age. Both children were born in Croatia, but while the first child in his first year of life lived in Italy for six months (from 0;2 to 0;7 and 0;10 to 0;11), the second child was for the first time in Italy with her brother and family for a month during the summer holidays when she was nine months old.[5] The family is based in Zadar, Croatia, and every year they spend their holidays in Piedmont, Italy, for approximately four to five weeks during summer and two weeks in winter.

Regarding language exposure, the parents decided to use a "one ambient–one language" strategy. Italian has been the dominant language at home, especially until the first child got to the age of six and the second to the age of three, while Croatian is the language they are mostly exposed to outside the familiar environment: in the kindergarten or school, on the playground with friends, or at home when spending time with their nanny. Regarding the varieties of Croatian they are exposed to, it should be mentioned that the varieties in question are the Štokavian ones spoken in Zadar and Split,

[4] It is worth mentioning that, in Italy, regional varieties differ from dialects in that they are not the direct descendants of the Latin spoken in the different parts of Italy, but rather the dialect-influenced versions of the literary language (cf. Marazzini, 1992). Piedmontese regional Italian is close to Standard Italian and different regional varieties are mutually intelligible, but they display lexical, phonological, and syntactic features that can diverge from those of other regional varieties (cf. Cerruti, 2009).

[5] Throughout the article we refer to the chronological age of the children expressed in years and months (e.g. 0;2 '2 months (from birth)').

along with the spoken standard language used in more formal situations. None of their grandparents live in Zadar, but the children are also exposed to Croatian at home when the grandparents from the mother's side visit the family or when the family goes to visit them.

### 3.2. Corpus MaLi: overview

The corpus name MaLi hints at the Croatian word *mali* meaning '(the) small ones'. On the other hand, it includes the first two letters of the Italian words *madre* 'mother' and *lingua* 'language, tongue' hinting at the Italian expressions *madrelingua* 'native speaker' and *lingua madre* 'mother tongue'.

The MaLi corpus comprises a twofold collection of data: diary-collected and audio-recorded data. Consider Table 1, where we summarise the main aspects of the MaLi corpus. The observation of the two children has not been even from their birth. The collection of the second child's language production is characterized by higher systematicity and more representative sampling. The second child's (CH2) data were collected when she was 1;3 years old (diary notes) and 1;7 years old (audio-recordings), and consistency has been maintained ever since. The first child's (CH1) diary data have been collected since the child was 3;2 years old, and at the age of 4;8 his production has been recorded mostly when accompanied by the presence of his sister.[6]

**Table 1**
The data of the MaLi corpus
**Tablica 1.**
Podatci o korpusu MaLi

| **type of data** | **diary data** | | **audio-recorded data** | |
|---|---|---|---|---|
| **child** | **CH1** | **CH2** | **CH1** | **CH2** |
| sampling period (age) | 3;2 – today/ongoing | 1;3 – today/ ongoing | 4;8 – today/ ongoing | 1;7 – today/ ongoing |
| sampling period (dates) | November 2019 – today/ ongoing | January 2021 – today/ ongoing | May 2021 – today/ ongoing | May 2021 – today/ ongoing |

The development of MaLi started rather informally and at the beginning was not guided by specific research questions. Therefore, the sampling re-

[6] In addition to this, the second child's development was followed using the second part of the MacArthur-Bates communication scales (MB-CDI) adapted for Croatian (Kovačević et al., 2007) and Italian (Caselli & Casadio, 1995), which was completed every three to four months starting from the 21st month of the girl's life. The data reported in CDI has not yet been analysed.

gime – its size, frequency, and density – has not been determined *a priori*.[7] However, the parents were advised by an expert to record at least one session per month for each language.[8] In 2021, at the very beginning of the audio data collection, interactions were sometimes recorded more than once a week, but over time the sampling frequency decreased to once a month.

## 3.3. Diary data

Diary data collected in the MaLi corpus cannot be described as a topical diary or a descriptive diary, such as the notes taken by Medved Krajnović (2004) in her study of her daughter's bilingual development. MaLi diary notes are a comprehensive diary of the spontaneous productions by the two observed children.

### 3.3.1. Procedure and structure

Data for the corpus are being collected in an application for taking notes on the mobile phone (Google Keep). The documents are accessed through the parents' accounts and are stored on the cloud. This ensures the privacy of the collection of notes, their cross-platform availability, and protection from data loss. Each diary note in the application is organized in three principal fields reported here in (1):

(1) a. production note,
b. comment,
c. date on which the note has been taken.

The fields are organised so that they can be more easily imported later in the spreadsheet. The production note cannot contain round brackets, but only square brackets, because comments are included in round brackets. After the right round bracket, the date is added.

The notes taken on the mobile phone are then transferred to spreadsheet (Excel) files and preserved in separate files for each child that are stored on a cloud storage service (OneDrive) and shared only by the parents/re-

[7] The informal character of the early stages of the data collection explains why the first child was not recorded et an earlier age.

[8] Regarding the incipient collection of audio-recorded data, the parents got in contact with Gordana Hržica, a researcher from the Department of Speech Therapy, Faculty of Education and Rehabilitation, University of Zagreb, who primarily deals with the acquisition of the Croatian language in her scientific activity and has experience with the methodology of research on children's language development and conducting research based on child language corpora. She suggested the frequency of sampling mentioned above.

searchers on their accounts. In addition, a local copy is saved on the hard-drive of a computer to add a further level of redundancy. Upon their transfer to the spreadsheet file, two columns are added: one for numbering the examples and one for age. In addition, columns are provided for encoding different categories. A non-elaborated excerpt from the MaLi diary for one child is reported in Table 2.

**Table 2**
An excerpt from the diary for one child (CH2)
**Tablica 2.**
Isječak iz dnevnika za jedno dijete (CH2)

| No. of example | Diary note | | | Age |
|---|---|---|---|---|
| | Production note | Comment | Date | |
| 1. | Mamma, gledaj, una ballelina ho našla | () | 15/04/22 | 2;6 |
| 2. | Non ce l'ho fatta | (da un po' di tempo dice non e non più no per la negazione) | 17/04/22 | 2;6 |
| 3. | Mi coperti questo? | (Dandomi lo strofinaccio per metterlo sul tavolino) | 22/04/22 | 2;6 |
| 4. | Ne volim | (non voglio) | 22/04/22 | 2;6 |
| 5. | L'hai metteto! Mamma, l'hai messo! | (subito si autocorregge) | 26/04/22 | 2;6 |
| 6. | Nisam ja kupala, mama | (se kupala) | 15/05/22 | 2;7 |
| 7. | Oni skakaju | () | 16/05/22 | 2;7 |
| 8. | Nemaju više | (per nema više čokolade/ komada čokolade) | 22/05/22 | 2;7 |
| 9. | Ljuljando | (Io dice e canticchia mentre è sdraiata per terra e avvolta nel tappeto di plastica per la vasca e si dondola) | 28/05/22 | 2;7 |
| 10. | L'hai saklato tu | (sakrato, sakrio, nascosto) | 04/06/22 | 2;7 |
| 11. | Livia: To je Paolo stavio tu. Mamma: A što je to? Livia: Pa kruk. Mamma: Što? Livia: Kruk. | (kruh) | 07/06/22 | 2;7 |
| 12. | Nemoj, ja bi spavat | () | 28/06/22 | 2;8 |

A production note corresponds to a child's full utterance or a part of it. The utterance can be a partial – s. ex. 9. –[9] or a full one – s. ex. 5.[10] Sometimes a production note is preceded or followed by a sentence uttered by a parent or a sibling included to provide a wider context – s. ex. 11.[11] A

[9] *Ljuljando* 'swinging', from Croatian *ljuljati se* 'to swing', and the Italian gerundive suffix *-(a)ndo*.

[10] *L'hai metteto! Mamma, l'hai messo!* 'You placed it! Mom, you placed it!', exemplifying a self-repair of the overgeneralized past participle of *mettere* 'to place'.

[11] Livia: *To je Paolo stavio tu*. Mamma: *A što je to?* Livia: *Pa kruk*. Mamma: Što? Livia: *Kruk*. 'Livia: Paolo place it (t)here. Mom: And what's that? Livia: Well, bread. Mom: What? Livia: Bread', the expected outcome of the Croatian word for 'bread' is *kruh*.

production note is often enriched by a comment offering a description of a situation, information about the addressee, the intended or target meaning, or a proposal of the form of the same utterance in the target language as would have been produced by an adult-speaker. In this way, all the relevant information of the speech context of an utterance is kept and preserved from oblivion and available for the researcher to ease interpretation.

### 3.3.2. Coding

So far, one part of the data has been coded for code-switching (on the intersentential, intrasentential, and word-internal level), transfer, overgeneralization or error patterns, and neologisms. Consider the excerpt from the diary coded for code-switching at the word-internal level in Figure 1.

**Figure 1**
An excerpt from the diary for CH2 coded for code-switching at the word-internal level[12]
**Slika 1.**
Isječak iz dnevnika za CH2 označen s obzirom na prebacivanje kodova na razini riječi

| parlato | eta_anni; me | CS | CS_INTE | CS_INTR | CS_WOR | POS_CS_ | INFL_CS_ |
|---|---|---|---|---|---|---|---|
| Dormala je tu. | 1;9 | 1 | 0 | 0 | 1 | V | ptcp.act.f.sg |
| [redacted]a [redacted]: Evo, čitaj ovo! L[redacted]: Io čito! | 2;2 | 1 | 0 | 1 | 1 | V | ind.prs.1sg |
| L[redacted]: Io questo spremo. [redacted]: Lo metti a posto? L[redacted]: Sì. | 2;2 | 1 | 0 | 1 | 1 | V | ind.prs.1sg |
| Io tolgati maglietta | 2;2 | 1 | 0 | 1 | 1 | V | inf |
| Io kačo | 2;2 | 1 | 0 | 1 | 1 | V | ind.prs.1sg |
| Spremo | 2;2 | 1 | 0 | 0 | 1 | V | ind.prs.1sg |
| Miješo | 2;2 | 1 | 0 | 0 | 1 | V | ind.prs.1sg |
| Io segnati questo | 2;2 | 1 | 0 | 1 | 1 | V | inf |
| Guarda, due cavalli trčano | 2;3 | 1 | 0 | 1 | 1 | V | ind.prs.3pl |
| Questo puklato | 2;3 | 1 | 0 | 1 | 1 | V | ptcp.pst.sg |
| L[redacted]: Questa è fatatrice. [redacted]: Hvatatrice? L[redacted]: Aha. | 2;3 | 1 | 0 | 1 | 1 | N | 0 |
| Ja ću beveti | 2;4 | 1 | 0 | 1 | 1 | V | inf |
| Baka [redacted]: Past ćeš /šć/ L[redacted]: No pasco. | 2;4 | 1 | 0 | 0 | 1 | V | ind.prs.1sg |
| Nonna [redacted], io sviro | 2;4 | 1 | 0 | 1 | 1 | V | ind.prs.1sg |
| L: Guada, puklato P: cosa? L: Puklato. | 2;4 | 1 | 0 | 0 | 1 | V | ptcp.pst.sg |
| Spremo | 2;4 | 1 | 0 | 0 | 1 | V | ind.prs.1sg |
| Ja oću pasticu | 2;4 | 1 | 0 | 0 | 1 | N | 0 |
| Questo è puklato | 2;4 | 1 | 0 | 0 | 1 | V | ptcp.pst.sg |
| L: Ja ću carnu. [redacted]: Hoće meso. L: No, carnu. | 2;4 | 1 | 0 | 1 | 1 | N | acc.sg |
| Di je pannolinica | 2;5 | 1 | 0 | 0 | 1 | N | 0 |
| Questa è ballerinica | 2;5 | 1 | 0 | 0 | 1 | N | 0 |
| Mamma, šeto | 2;5 | 1 | 0 | 1 | 1 | V | ind.prs.1sg |
| Miješo | 2;5 | 1 | 0 | 0 | 1 | V | ind.prs.1sg |
| Guarda, pjevano | 2;5 | 1 | 0 | 1 | 1 | V | ind.prs.3pl |

[12] Figure 1 shows a screenshot of the spreadsheet programme used for annotating code-switching at the word-internal level. The annotation has the following fields (columns in the spreadsheet document): "eta_anni;mesi", chronological age expressed in years and months; "CS" presence/absence of code-switching in the collected utterance; "CS_INTER" presence/absence of code-switching at the inter-sentential level; "CS_INTRA" presence/absence of code-switching at the intra-sentential level; "CS_WORD" presence/absence of code-switching at the word-internal level; "POS_CS_WORD" part of speech of the word displaying word-internal code-switching; "INFL_CS_WORD" inflectional features of the word displaying word-internal code-switching. Due to lack of space, some of the labels are not entirely visible in Figure 1. Comments are present in the document, but they are not shown in Figure 1 for the same reason.

Figure 1 shows the CH2's production characterized by code-switching between morphemes in the age span from 1;9 to 2;5. The data show a combination of roots belonging to one language and affixes from the other.

The MaLi diary is not a topical diary. Data reported in it can be described as comprehensive, offering insight on other developmental patterns and various phenomena, mostly on the lexical, syntactic, and morphosyntactic level.

### 3.4. Audio-recorded data

The audio-recordings included in the MaLi corpus document bilingual Croatian-Italian child language through a longitudinal collection of naturalistic data. The whole process of building this component complies with the standard defined in the context of the wider CHILDES database.

#### 3.4.1. Procedure

The collection of data for the corpus started in May 2021. In the study, data is collected from spontaneous interactions of the siblings, their parents, and a woman who is not a family member. The younger child – a girl – has been recorded systematically since she was 19 months old, whereas the older child – a boy – has been recorded in a less systematic way since he was 3 years and 8 months old (see Table 1).

The sessions are recorded in the family home in Zadar, Croatia. Participants are recorded with a Zoom H4n voice recorder in .wav format one to several times a month in sessions that last approximately one hour each.

The recorded interaction takes place in two languages. Interaction between children and parents takes place dominantly, although not exclusively, in Italian, while interaction with the adult who is not a family member takes place in Croatian. The Croatian-speaking adult, a woman close to the family, does not have any knowledge of Italian. She agreed to be recorded, she was previously informed about the purpose of the research, and the method of recording was presented to her, as well as possible risks. In addition, she was asked for her consent every time before the session began and was informed that she could give up and stop the recording at any time.

Family setting determines activities and situations in which recording sessions are organized. Naturally occurring situations are recorded, e.g., playing, dining, reading and telling stories, or doing something else in the household, such as preparing food, doing reparations, etc.

### 3.4.2. Data management, archiving and file naming

Data management is a challenging part of conducting a longitudinal research based on spoken data. Other than planning sessions, taking into consideration time distance from one to another recording for both languages, once recording has been done, a considerable effort needs to be put into the storing and organization of the audio files. The workflow for storing and sharing audio files and related documents involves saving them in the same cloud storage service mentioned in 3.3.1, about the management of diary notes. Also, in this context. audio files are exclusively accessible by the parents/researchers, and an additional level of redundancy is added by saving a local copy of the audio files on a hard drive.

The interactions between children and adults constitute spoken data collected in audio files, which are then transcribed. The organization of the data involves taking a series of decisions in developing and applying different sorts of conventions. In order to keep track of the files, a system of conventions for naming the files was developed. The scheme is laid out in Table 3, and it is followed by descriptions and examples.

**Table 3**
File naming in the MaLi corpus
**Tablica 3.**
Nazivanje datoteka u korpusu MaLi

| **scheme** | MaLiyyyymmdd#tar | e.g. MaLi20210516aita |
|---|---|---|
| **yyyy** | year | e.g. 2021 |
| **mm** | month | e.g. 05 |
| **dd** | day | e.g. 16 |
| **#** | progressive letter (identifying different sessions on the same day) | e.g. a [first session, even if the only one] |
| **tar** | target language | e.g. ita [consisting always of three letters; hrv for Croatian] [13] |

Consider the two examples of audio-file and transcription name below:

(2) MaLi20210516aita
(3) MaLi20210629ahrv

In example (2), the recording took place on 16 May 2021 (20210516) as the first recording of that day (a) in Italian (ita). On the other hand, the file in (3) was recorded on 29 June 2021 (20210629) as the first recording of that day (a) in Croatian (hrv).

[13] Language codes are defined by the international ISO 639-3 standard, as indicated in MacWhinney (2000, p. 30, modified in February 2024).

### 3.4.3. Transcribing and coding

In order to safeguard the participants' anonymity and to protect the children, we decided to name them CH1 'first child' and CH2 'second child', and the adults participating in the study were named after their roles: MOT, FAT for 'mother', and 'father', respectively. Although the most obvious abbreviation for a 'baby-sitter' would be SIT, we opted for TET, considering she is called 'teta Vera' by both children.

In the transcriptions of utterances, taking into consideration the fact that names, just like other nouns in Croatian, belong to a paradigm taking therefore different forms in different case slots, we decided to give them pseudonyms in the form of names that behave similarly to the original names of the participants. These names are: Paolo for CH1, Livia for CH2, teta Vera for TET. These have the same case forms that the real names of the participants do throughout the paradigm. In this way, we acknowledged the relevance of the acquisition of case forms in the development of competence in a highly synthetic language such as Croatian.[14]

The collection of audio recordings and the transcription of the files is in progress, so there still not exist an estimate of the size of the corpus. The files are being transcribed using the CLAN programme, which is why the transcriptions comply with the guidelines of the CHAT format, the notation/coding system developed for the purpose of the development of the CHILDES database.

As MacWhinney (2000, p. 100) states, "transcription is easiest when speakers avoid overlaps." Transcriptions of MaLi data with three to four participants are difficult to do, since they are characterized by more frequent overlaps, or by diverging one-on-one interactions.

Considering the fact that the children are bilingual, one of the peculiarities that distinguishes this corpus from the existent Croatian child language corpus (Kovačević, 2002) is the presence of code-switching in the utterances. Code switching – "a hallmark of bilingual language processing" (Van Hell et al., 2019, p. 459) and a speech style common to fluent bilinguals

[14] For example, although *Marin*, *Darko*, *Stipe*, *Ivo,* and *Dino* are all proper nouns for male referents, they do not follow the same paradigm. *Marin* and *Darko* follow the so-called a-declension paradigm, ending in *-a* in the genitive – *Marina*, *Darka*. On the other hand, *Stipe* and *Ivo* end in in *-e* when in the genitive form – *Stipe* and *Ive*, respectively. At the same time, case forms are sensitive to stress and dialectal varieties, which is why *Dino* can have two genitive forms – *Dina* and *Dine* (see for a discussion Bošnjak Botica & Jelaska, 2008). A consideration of differences in paradigms reflects an informed decision taken regarding the selection of pseudonyms: proper nouns Paolo and Livia for CH1 and CH2, respectively, follow the same paradigm as the real names of the observed children.

(MacSwan, 2005, p. 55) – in the MaLi corpus encompasses a combination of the two languages at the intersentential, intrasentential, and word-internal level. Particular attention is devoted to the coding of code-switched utterances, words, and morphemes. An example is given in (4), which reproduces an excerpt of a transcription of a recording having Italian as target language which documents an interaction when the child was 3 years and 5 months old.

(4)

| | | |
|---|---|---|
| 1 | *MOT: | e quello lì fuori sì è un albero. |
| 2 | *CH2: | [- hrv] <ja ću> [/] <ja ću penjele@s:hrv+ita> [/] ja ću penjele@s:hrv+ita [- ita] a +//. |
| 3 | *CH2: | di là [- hrv] ja ću penjele@s:hrv+ita [- ita] di là |
| 4 | *MOT: | cosa vuoi, spegnere? |
| 5 | *CH2: | no di là penj@s +//. |
| 6 | *CH2: | penjele@s:hrv+ita. |
| 7 | *MOT: | ti vuoi arrampicare sull'albero? |
| 8 | *CH2: | sì. |

When coding code-switching, it is of utmost importance to define the target language of the interaction, which corresponds to the first language the transcriber indicated in the heading of a transcript. In the excerpt in (4), the first – dominant – language is Italian, whereas Croatian is indicated as second in line. In case a word from Croatian is inserted in an otherwise Italian utterance, a code – @s – indicating a foreign word (MacWhinney, 2000, 102–103), is added to it (line 5). In case the utterance is in Croatian, it is introduced by the code [- hrv] (line 2). If the utterance begins with second language and continues with the dominant language, a decision was made to indicate the switching point, making visible the code for Italian code-switching: [- ita] (lines 2, 3). In case code-switching occurs between morphemes, after a word the code @s followed by the colon (:) and both languages hrv+ita in the order that they occur in the word (line 6: "penjele@s:hrv+ita"). In this specific example, the word is a combination of a Croatian root *penj-* 'climb' and the Italian infinitive suffix *-(e)re* pronounced *-(e)le* by the child.[15]

[15] This is a developmental characteristic of child's speech in which the phonetic realisation of the phoneme /r/ overlaps with that of the phoneme /l/.

## 4. CONCLUSION

In this paper we have presented the structure and characteristics of a corpus under construction, MaLi, which will document via naturalistic data the spontaneous linguistic productions of two bilingual children. The twofold corpus – constituted by diary notes and audio-recordings – will develop into a resource complying with the features of the CHILDES database (MacWhinney, 2000).

Spoken naturalistic data are a valuable source for the documentation of language varieties because they provide a sample of spontaneous productions which can spot transitory phenomena and ephemeral features of a language variety. The MaLi corpus focuses on Croatian and Italian, and while it is not uncommon for these languages to be present in an individual's repertoire, especially in the bilingual region of Istria in Croatia, to our knowledge no corpus which collects the spontaneous productions of Croatian-Italian bilingual children in both languages is available in the CHILDES database.

The planned inclusion of MaLi in the CHILDES database will help in achieving the visibility and wider accessibility of data collection. The prospective corpus can be used to study the acquisition of Italian and Croatian in children with a known bilingual background and compare their development to that of monolingual children. Moreover, the corpus will provide examples of borrowing, code-switching, and transfer, phenomena typically present in the speech of bilinguals, the study of which can shed light on bilingual competence.

Following the development of two bilingual children belonging to the same family unit is surely not representative of the population of Croatian-Italian bilingual children: the sample is too restricted; most Croatian-Italian bilingual children are characterized by a different socio-cultural and linguistic context; the children's parents are linguists, unlike most parents of bilingual children. In addition, a single corpus does not obviously allow on its own the definition of developmental scales for Croatian-Italian bilingual children. However, despite these limits of representativeness, the MaLi corpus can be a valuable contribution to the field, on the one hand because it fills a gap in the architecture of the CHILDES database, and on the other hand because it constitutes a first step in the study of Croatian-Italian bilingual first language acquisition.

Concerning the further steps in this incipient enterprise, building MaLi can encourage other researchers to follow the steps described in this paper

and embark in similar small data collections. As for MaLi, our first concern is now to transcribe the bulk of the recordings collected so far, a crucial phase for making MaLi available to the research community.

## REFERENCES


Ajduković, M., & Keresteš, G. (2020). *Etički kodeks istraživanja s djecom (drugo revidirano izdanje)*. Nacionalno etičko povjerenstvo za istraživanje s djecom.

Angster, M., & Batinić Angster, M. (2022). Gli effetti del contatto croato-italiano nelle produzioni di un bambino bilingue. *Studia Romanica et Anglica Zagrabiensia: Revue publiée par les Sections romane, italienne et anglaise de la Faculté des Lettres de l'Université de Zagreb, 67,* pp. 107–122. http://dx.doi.org/10.17234/SRAZ.67.8

Behrens, H. (2008). Corpora in Language Acquisition Research. History, methods, perspectives. In H. Behrens (Ed.), *Corpora in language acquisition research: History, methods, perspectives* (pp. xi–xxx). John Benjamins Publishing Company. https://doi.org/10.1075/tilar.6.03beh

Bošnjak Botica, T., & Jelaska, Z. (2008). Sklonidba muških dvosložnih imena i naglasna dvojnost. *Jezik:* časopis *za kulturu hrvatskoga književnog jezika*, *55(5)*, pp. 170–181.

Bowerman, M. (1974). Learning the Structure of Causative Verbs: A Study in the Relationship of Cognitive, Semantic and Syntactic Development. *Papers and Reports on Child Language Development* 8, pp. 142–178.

Braunwald, S. R., & Brislin, R. W. (1979). The Diary Method Updated. In E. Ochs & B. Schieffelin (Eds.), *Developmental Pragmatics* (pp. 21–41). New York Academic Press.

Caselli, M. C., Bello, A., Rinaldi, P., Stefanin, S., & Pasqualetti, P. (2015). *Il primo vocabolario del bambino: gesti, parole e frasi. Valori di riferimento fra 8 e 36 mesi delle Forme complete e delle Forme brevi del questionario MacArthur-Bates CDI*. Franco Angeli.

Caselli, M. C., & Casadio, P. (1995). *Il primo vocabolario del bambino: guida all'uso del questionario MacArthur per la valutazione della comunicazione e del linguaggio nei primi anni di vita (Vol. 5).* Franco Angeli.

Cerruti, M. (2009). *Strutture dell'italiano regionale. Morfosintassi di una varietà diatopica in prospettiva sociolinguistica*. Peter Lang.

Čamber, M. (2020). *Simultaneous acquisition of Austrian German and Croatian at home and in preschool*. PhD thesis. University of Vienna.

Darwin, C. (1877). A Biographical Sketch of an Infant. *Mind 2*, pp. 285–294.

Demuth, K. (2008). Exploiting corpora for language acquisition research. In: H. Behrens (Ed.), *Corpora in language acquisition research: History, methods, perspectives* (pp. 199–205). John Benjamins Publishing Company. https://doi.org/10.1075/tilar.6.10dem

Deville, G. (1891). Notes sur le développement du langage II. *Revue de linguistique et de philologie comparée 24*: 10–42, 128–43, 242–57, 300–20.

Fennell, C. T. (2012). Habituation Procedures. In E. Hoff (Ed.), *Research methods in child language: A practical guide* (pp. 3–16). Wiley-Blackwell. https://doi.org/10.1002/9781444344035.ch1

Gvozdev, A. N. (1928). Značenje izučenija detskogo jazyka dlja jazykovedenija. *Rodnoj jazyk i leteratura v trudovoj škole. No. 3.4.5.*

Hržica, G., Brdarić, B, Tadić, E., Goleš, A., & Roch, M. (2015). Dominantnost jezika dvojezičnih govornika talijanskog i hrvatskog jezika. *Logopedija, 5(2),* pp. 34–40. https://hrcak.srce.hr/150404

Hržica, G., Kuvač Kraljević, J., & Štefanec, V. (2022). Čestotni *rječnik hrvatskoga dječjeg jezika: morfološki i razvojni oblici*. Naklada Slap.

Hržica, G., & Roch, M. (2021). Lexical diversity in bilingual speakers of Croatian and Italian. In S. Armon-Lotem, & K. Grohmann (Eds.), *LITMUS in Action: Crosscomparison studies across Europe*. John Benjamins Publishing Company. http://dx.doi.org/10.1075/tilar.29.04hrz

Kovačević, M. (2002). *Croatian Child Language Corpus, CHILDES*. https://childes.talkbank.org/access/Slavic/Croatian/Kovacevic.html.

Kovačević, M., Jelaska, Z., Kuvač Kraljević, J., & Cepanec, M. (2007). *Komunikacijske razvojne ljestvice (KORALJE).* Naklada Slap.

Kuvač Kraljević, J., & Hržica, G. (2016). Croatian adult spoken language corpus (HrAL). *FLUMINENSIA: časopis za filološka istraživanja, 28(2),* pp. 87–102.

Kuvač Kraljević, J., Hržica, G., & Štefanec, V. (2022). Čestotni *rječnik hrvatskoga dječjeg jezika: morfološki i razvojni oblici. Natuknice.* Naklada Slap.

Kuvač, J., & Palmović, M. (2007). *Metodologija istraživanja dječjega jezik*a. Naklada Slap.

Leopold, W. F. (1939–1949). *Speech Development of a Bilingual Child: A Linguist's Record* (4 Volumes). Northwestern University Press (Reprint AMS Press 1970).

Lieven, E., & Behrens, H. (2012). Dense sampling. In E. Hoff (Ed.), *Research methods in child language: A practical guide* (pp. 226–239). Wiley-Blackwell. https://doi.org/10.1002/9781444344035.ch15

Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early Syntactic Creativity: A Usage-Based Approach. *Journal of Child Language 30*, pp. 333–370. http://dx.doi.org/10.1017/S0305000903005592

MacSwan, J. (2005). Precìs of a minimalist approach to intrasentential code switching. *Italian Journal of Linguistics*, 17(1). 55.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition*. Lawrence Erlbaum Associates. https://doi.org/10.21415/3mhn-0z89

Marazzini, C. (1992). Piemonte e Valle d'Aosta. In: F. Bruni (ed.), *L'italiano nelle regioni. Lingua nazionale e identità regionali* (pp. 1-44). UTET.

Medved Krajnović, M. (2004). *Razvoj hrvatsko-engleske dvojezičnosti u dječjoj dobi*. PhD thesis. Filozofski fakultet Sveučilišta u Zagrebu.

Preyer, W. (1882). *Die Seele des Kindes*. Grieben.

Rowe, M. L. (2012). Recording, Transcribing, and Coding Interaction. In: E. Hoff (Ed.), *Research Methods in Child: A Practical Guide* (pp. 193–207). Wiley-Blackwell. https://doi.org/10.1002/9781444344035.ch13

Rowland, C. F., Fletcher, S. L. & Freudenthal, D. (2008). How big is big enough? Assessing the reliability of data from naturalistic samples. In H. Behrens (Ed.), *Corpora in language acquisition research: History, methods, perspectives* (pp. 1–24). John Benjamins Publishing Company. https://doi.org/10.1075/tilar.6.04row

Swingley, D. (2012). The Looking-While-Listening Procedure. In E. Hoff (Ed.), *Research methods in child language: A practical guide* (pp. 29–42). Wiley-Blackwell. https://doi.org/10.1002/9781444344035.ch3

Tomasello, M. (1992). *First Verbs: A Case Study of Early Grammatical Development*. Cambridge University Press.

van Oosten, A. (2005). *Lo sviluppo dell'acquisizione del soggetto nei bambini bilingui italo-olandesi*. MA thesis. Utrecht University.

van Hell, Janet G.; Litcofsky, Kaitlyn A.; Ting, Caitlin Y. (2015). Intra-sentential code-switching: Cognitive and neural approaches. In: Schwieter, J. W. (Ed.), *The Cambridge handbook of bilingual processing* (pp. 459–482). Cambridge University Press. http://dx.doi.org/10.1017/CBO9781107447257.020

## Acknowledgments

This work has been realised under the institutional project *Multilingualism: between theory and empirical knowledge* (*ViTE* – Višejezičnost: između teorije i empirije) (IP-01-2023-14), fully financed by the University of Zadar. We would like to thank Gordana Hržica and Vesna Fantina for their help and support.

# Nastajanje hrvatsko-talijanskoga dvojezičnog dječjeg korpusa MaLi

Mia Batinić Angster
Marco Angster

Spoznaje stečene o jeziku i usvajanju prvoga jezika ne bi bile moguće bez prethodno uloženih napora u prikupljanje jezičnih podataka, npr. snimanjem spontanih interakcija između djece i odraslih. Baza podataka CHILDES (MacWhinney, 2000) okuplja dječju jezičnu proizvodnju na mnogim svjetskim jezicima, uključujući i hrvatski (Kovačević, 2002). U ovome radu opisujemo izradu korpusa MaLi kojim se dokumentira jezična proizvodnja dvoje dvojezične djece koja istodobno usvajaju hrvatski i talijanski jezik. Nakon kratkoga pregleda metoda prikupljanja podataka o dječjemu jeziku s posebnim osvrtom na dnevničke zapise i zvučne zapise, raspravlja se o pozadini i pojedinostima podataka prikupljenih za potrebe korpusa MaLi. Nudi se pregled sociolingvističkoga konteksta dvojezičnoga usvajanja prvoga jezika promatrane djece i opis građe. Pozornost se prvo posvećuje prikupljanju podataka, upravljanju podatcima i označavanju dnevničkih zapisa. Nakon toga opisuje se prikupljanje i obrada zvučnih zapisa i njihovo prepisivanje, koje je u tijeku. U završnim napomenama procjenjuju se prednosti i ograničenja korpusa te se daje pregled mogućnosti uporabe ovoga resursa.

Ključne riječi: *jezična dokumentacija, dječji jezik, korpus, rana dvojezičnost, istodobna dvojezičnost, usvajanje prvoga jezika*