Early Child Language Resources and Corpora Developed in Nine African Languages by the SADiLaR Child Language Development Node

Michelle Jennifer White[1,2], Frenette Southwood[2], Sefela Yalala[3] and the South African Communicative Development Inventory team*
University of Plymouth[1], Stellenbosch University[2], Northwestern University[3]
michelle.white@plymouth.ac.uk, fs@sun.ac.za, sefelayalala2027@u.northwestern.edu

Abstract

Prior to the initiation of the project reported on in this paper, there were no instruments available with which to measure the language skills of young speakers of nine official African languages of South Africa. This limited the kind of research that could be conducted, and the rate at which knowledge creation on child language development could progress. Not only does this result in a dearth of knowledge needed to inform child language interventions but it also hinders the development of child language theories that would have good predictive power across languages. This paper reports on (i) the development of a questionnaire that caregivers complete about their infant's communicative gestures and vocabulary or about their toddler's vocabulary and grammar skills, in isiNdebele, isiXhosa, isiZulu, Sesotho, Sesotho sa Leboa, Setswana, Siswati, Tshivenda, and Xitsonga; and (ii) the 24 child language corpora thus far developed with these instruments. The potential research avenues opened by the 18 instruments and 24 corpora are discussed.

Keywords: Communicative development inventory, child language, gesture, vocabulary, grammar

1. Introduction

The dearth of instruments with which to measure early child language development in African languages and of child language corpora in these languages need attention for three main reasons. The first is that life chances are influenced by educational attainment, which requires good literacy, and that the latter is built on adequate language skills (Catts et al., 1999). It is pertinent to identify children who have poor language skills early so that they can receive the intervention necessary for the improvement of said skills (Fricke et al., 2013), and for such identification, one needs reliable measuring instruments and developmental norms. The second reason is related to the first: Child language intervention programmes need to be evidence-based and take typical child language development into account. To gain contextually relevant knowledge on typical child language development, we require instruments with which to measure and track development, and corpora to analyse so that we can answer our child language related research questions. The third reason is that most of what we know about child language development is based on research of English and other European world languages (such as German and French), and that this research (which could present a skewed picture of child language development) is what informs theories of child language development (Kidd and Garcia, 2022). To generate knowledge on child language development in African languages with which to test the generalisability of existing child language theories, we need appropriate child language measuring instruments and sizeable child language corpora in African languages.

In this paper, we report on instruments and corpora developed for isiNdebele, isiXhosa, isiZulu, Sesotho, Sesotho sa Leboa, Setswana, Siswati, Tshivenda, and Xitsonga[1] by a multilingual, multi-site team of linguists, speech-language therapists, and African language specialists. Specifically, one infant and one toddler version of a child language assessment instrument – the MacArthur-Bates Communicative Development Inventory (CDI, Fenson et al., 2007) – was developed for each of these nine official spoken African languages of

* SA-CDI team: Monicca Bhuda (University of Mpumalanga), Nina Brink (North-West University), Heather Brookes (Stellenbosch University), Nomfundo Buthelezi (University of KwaZulu-Natal), Carmen Coetsee (Stellenbosch University), William Jiyana (University of Mpumalanga), Portia Khumalo (Stellenbosch University), Babalwa Ludidi (University of Cape Town), Patricia Makaure (Stellenbosch University), Martin Mössmer (University of Michigan), Muzi Matfunjwa (North-West University), Lufuno Miriri (University of Limpopo), Mikateko Ndhambi (Sefako Makgatho Health Science University), Sibusiso Ndlangamandla (University of South Africa), Helena Oosthuizen (Stellenbosch University), Nomsa Skosana (North-West University), and Katie Alcock (Lancaster University).

[1] The corpora are stored by SADiLaR but have not yet been made available to other researchers. Enquiries about the final versions of the instruments can be directed to the second author.

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 86–93
25 May, 2024.

South Africa[2], and (ii) 24 corpora were built, or are in the process of being built, with these instruments. These corpora comprise the infant and the toddler CDI data as well as transcribed language samples collected from toddlers who speak one of these nine languages. We also discuss the research possibilities these instruments and corpora afford us.

2. The South African Communicative Development Inventories

The MacArthur-Bates CDI was first developed for American English (Fenson et al., 1993) but has since been adapted for more than 100 languages from different language families (see https://mb-cdi.stanford.edu/adaptations.html), under license of, and following the guidelines of, the MacArthur-Bates Board in order to render culturally and linguistically appropriate adaptations rather than mere translations. There are two age versions of the CDI, one for infants (8 to 18 months) and another for toddlers (16 to 30 months), with the 16- to 18-month overlap being intentional[3]. CDIs are caregiver reports: The parents or other primary caregivers check off on a list which language items a child has acquired. Both the infant and toddler CDIs focus on vocabulary. Words from more than 20 semantic domains are listed alphabetically (see Table 2 further below for the domains included in the South African CDIs). Caregivers are asked to indicate which of these words the child knows. On the infant version, a list of approximately 400 words (see Table 2 for precise numbers) can be marked off for either comprehension, or comprehension *and* production. On the toddler version, a list of approximately 700 words can be marked off, for production only. The infant CDI also contains checklists for gestures, play routines, actions, and comprehension of commonly used phrases (see Section 2.2.1), whereas the toddler CDI has grammar checklists for morphology, word combinations, and sentence complexity (see Section 2.2.3).

2.1 Method for Developing the Communicative Development Inventories

2.1.1 General Protocol

Following the MacArthur-Bates Board's guidelines, research teams have utilised a range of methods to adapt CDIs to new languages (see Jarůšková et al., 2023). Many teams make use of Wordbank (Frank et al., 2017), an open access repository of CDI data, to examine which words other CDIs have included. Due to Wordbank only having come into being after the commencement of the current study, this approach was not applied. Another common way to begin the adaptation process is to translate an existing CDI into the target language (Jarůšková et al., 2023), which is subsequently expanded by adding words that are culture-specific and/or language-specific. We began by translating the American English CDI to the target languages. Following, for example, Anđelković et al. (2017) for Serbian and Jackson-Maldonado et al. (1993) for Mexican Spanish, we made use of caregiver interviews to uncover which actions, gestures, and words in the translation might be irrelevant or missing. We also employed focus group discussions and spontaneous language samples to the same effect, as discussed below.

Due to the nature of the differences between the grammars of English and African languages, the grammar sections could not use a translation of the American English CDI as their point of departure. As we will explain below, we consulted the Kiswahili and Kigiryama CDIs (Alcock et al., 2015), the limited literature available on early language development in Bantu languages, caregivers of young children speaking the relevant languages, focus groups, and our recordings of toddlers' spontaneous language samples to create a first version of the grammar section.

The main aims of this pre-pilot phase were to check for completeness and eliminate cultural bias before piloting the CDIs. Below, we discuss the steps that were followed during the adaptation process in more detail.

2.1.2 Testing the First and Second Draft Versions of the CDI

As a first step, the American English CDI was translated by three mother tongue speakers per language for isiXhosa, Sesotho, Setswana and Xitsonga. Initially, funding could only be secured for four languages, and these four were selected because we had an existing network of mother-tongue-speaking researchers available for them. The adaptation process for the remaining five languages (isiZulu, isiNdebele, Sesotho sa Leboa, Siswati and Tshivenda) was initiated two years later, after the CDIs for the first four languages had been piloted twice, and once

[2] We developed similar instruments and resources for the two official Germanic languages of South Africa, i.e., Afrikaans and South African English, but we report on those developed for the Bantu languages only and not on those developed for Afrikaans, which can also be viewed as an African language. Such resource development for South African Sign Language is yet to commence.

[3] There is also a CDI-III for children of 30 to 37 months (see https://mb-cdi.stanford.edu/cdi_iii_form.html). It is a very short questionnaire, and few research teams have developed this CDI age version for their language(s).

further funding had been secured. A main consideration was to harmonise all nine CDI language versions so that they would be comparable and allow for crosslinguistic comparisons and data pooling during future research. Considering that the first four languages' adaptations performed very well during the two pilots, the adaptation of the CDI for the last five languages did not start with a translation of the American English CDI but rather with that of a more closely related language's CDI: For the Nguni languages (isiZulu, isiNdebele and Siswati), the isiXhosa adaptation of the CDI was translated; Sesotho sa Leboa used the Sesotho and Setswana CDIs; and Tshivenda used the Xitsonga CDI. Harmonisation across languages commenced before the first pilot of the first four languages and was further refined after both the first and second pilots.

As indicated above, two rounds of piloting were completed for the first four languages, but only one for the remaining five languages because they were closely based on the four already piloted language versions of the CDI. In the first pilot, each of the preliminary adaptations of the CDIs were completed by 40 caregivers of infants 8 to 18 months old and another 40 caregivers of toddlers 16 to 32 months old. They completed paper copies of the CDIs with the help of fieldworkers who were recruited via Early Childhood Development centers and researcher networks. After the first pilot, some items were removed or replaced based on the caregiver responses.

For the second pilot of the first four languages and the only pilot of the last five languages, online CDIs were used instead of paper-based versions. The online CDIs were built on Qualtrics (Provo, Utah), eliminating possible human error in data capturing. Additionally, the Qualtrics application allowed for the collection of data without the need for internet connectivity, which is a necessity in rural areas and during the frequent electricity blackouts South Africa has been experiencing.

Data from caregivers of more than 100 infants and 100 toddlers per language was collected (see Tables 4 and 5 for exact numbers per language) during Pilot 2 of the first four languages and the only pilot of the remaining five, again with the assistance of fieldworkers, either face-to-face or (when COVID-19 social distancing regulations were in place) telephonically. This was done for respondent comfort, given that many caregivers were not able to complete the CDI themselves due to low literacy levels or technology-related limitations.

2.1.3 Development of the Actions and Gestures Section (Infant CDI)

Actions and gestures that are on the American English CDI were used as a starting point for this section. Those items which were not relevant to our context were excluded or modified after translation, and actions and gestures typically used by speakers of the target languages were added. To make the items relevant to the South African context, some had to be adapted. For example, rather than asking whether the child waved to say hello, as in the American English CDI, we asked whether the child used a gesture to greet such as waving, thumbs up, high five or something culturally similar. This was done to cover the variation that exists in children's first social gestures for greeting across the languages concerned. Imitating adult actions were also changed to be more culturally and/or contextually appropriate. For instance, brooms are used more often for cleaning than vacuum cleaners, therefore sweeping was added to the American English CDI's question about whether the child imitates adults by attempting to mop or vacuum clean.

2.1.4 Development of the Words Section

The translated CDIs were presented to individual language practitioners of each language (e.g., linguists or speech-language therapists) whereafter two focus groups per language were consulted. They consisted of professionals who work with children as well as parents of young children. The feedback from the language practitioners and focus groups led to the removal and addition of some words and/or synonyms. Words which are not relevant to everyday South African life, such as *snow suit*, were removed, whereas words had to be added when, for instance, a single word on the American English CDI could be translated in multiple ways. Consider, for example, porridge, which is a staple food for many South Africans. Various types of porridge (e.g., maize meal porridge or oatmeal porridge) can be referred to with one word, *porridge*, in English but require several words in the African languages concerned, depending on its ingredients and consistency, including *papa, mahleu*, *motoho*, or *mabele* in Sesotho, and *motogo*, *bogobe* or *phaletšhe* in Setswana. All these words for porridge were added to the word lists. When adapting the word lists, dialects or varieties of the specific languages were also considered. For this reason, the focus groups comprised of people speaking various dialects or varieties of the language in question and focus group members were requested to point out those items which were highly dialectal or variety specific.

Subsequently, 30-minute samples of naturally occurring spontaneous language were collected from six toddlers (27 to 32 months) per language. Words that were found to occur in the language samples but were not yet on the word lists were added.

Although the same protocol was followed for all languages in an attempt to facilitate crosslinguistic comparisons, the final number of words varied across languages (see Table 2). This is due, for instance, to some words being polysemous in one language while several related words were required in another.

2.1.5 Development of the Grammar Section (Toddler CDI)

Only a limited number of studies have been conducted on grammar development in children learning Bantu languages, yielding very little available empirical evidence. Such evidence on early acquired grammatical constructs was available for only a few languages (see Demuth, 2003 for a summary), namely Sesotho (e.g., Connelly, 1984; Demuth, 1992), Siswati (Kunene, 1979), isiZulu (Suzman, 1991) and Setswana (Tsonope, 1987; 1993). Thus, the grammar section had to be developed based on this limited existing literature, the language samples referred to above, and the Kiswahili and Kigiriama CDIs (Alcock et al., 2015). These were the only published full CDIs that had been adapted into African languages at the time. The grammar section of the Kiswahili/Kigiriama CDIs appeared to perform well (Alcock et al., 2015) and were thus deemed reliable for use as a starting point. Their structure was followed, yielding grammar sections which each consisted of four subsections, namely small parts of words, word complexity, word combinations, and sentence complexity (see Table 3 for more information). The language professionals and focus groups commented on the preliminary items and were encouraged to suggest examples of constructions that children acquiring the languages are likely to hear or to produce.

Across the languages, there were many similarities but also some distinct grammatical differences. The decision was made to include additional, language-specific items (more than would be needed in the final version of the CDI) for the first pilot, even if the type of construction did not occur in all the languages concerned. This was done to ascertain which items would be most effective because so little data is available on these languages. In Sesotho and Setswana, for instance, there is irregular verb inflection in the past tense, therefore items pertaining to this were included for these two languages only.

A feature common to Bantu languages is that of having several noun classes (which take the form of prefixes), with different numbers in each language. Moreover, some languages and language varieties have pre-prefixes that do not exist in others. Examples of these items that contain structures that would likely be part of a child's early grammar had to be found. The main source of these examples were the language professionals and focus groups.

After the first pilot of the first four languages, the grammar items were improved based on the caregiver responses, and the instructions were clarified to make it easier for caregivers to understand the questions about grammar. Feedback from fieldworkers was especially important to determine what might have been confusing for the caregivers.

The second pilot was conducted with caregivers of 100 toddlers and indicated that the items were suitable; the items correlated significantly with each other and with the child's age and the child's vocabulary size, the latter measured by the word section of the CDI.

The grammar sections of the second group of languages (isiNdebele, isiZulu, Sesotho sa Leboa, Siswati, and Tshivenda), were based on the first four languages', with some adaptation. Their examples came from focus group discussions with caregivers and language professionals and from natural child language recordings. Some items were substituted because they relate to aspects that are irregular in one language but not in another, for instance; or the relevant structure differed across languages. For example, Tshivenda uses a prefix to mark past tense whereas the other languages use a suffix. These grammar sections were piloted once, with 100 caregivers per language.

2.2 Content of the Final Versions of the Communicative Development Inventories

Details of the final versions of the CDIs are summarised in the tables below. Table 1 indicates the five subsections of the actions and gesture section, and the number of items in each subsection. These subsections are (i) first communicative gestures, e.g., deictic gestures such as pointing; (ii) games and routines, e.g., clapping hands, (iii) actions with objects, e.g., drinking from a cup; (iv) pretending to be a parent, which included symbolic gestures and play schemes with a 'baby', e.g., dressing or trying to dress a doll or soft toy; and (v) imitating other adult actions, e.g., writing with a pen/pencil.

Subsection	Examples of questions (English equivalents)	No. of items
First communi-cative gestures	• Requests something by extending arm and opening and closing hand or putting their hands together • Shakes head "no"	12-14
Games and routines	• Plays a hiding game (hiding their face or whole body) • Dances	4-6
Actions with objects	• Combs or brushes own hair • Throws a ball	19

Pretending to be a parent	• Covers [a doll] with a blanket • Pushes [a doll] in a stroller/pram or carries it on his/her back	13-14
Imitating other adult actions	• Cleans with a cloth • Pretends to cook	12-14

Table 1: Number and types of items in the action and gestures sections

Table 2 provides the mean number of words per semantic domain on the Infant forms and the Toddler forms of the CDI. These differ somewhat across languages, as explained in Section 2.1.5.

Subsection	Examples of words (English equivalents)	Mean no. of items	
		Infant forms	Toddler forms
Sounds	*Woof woof, uh oh / yo*	17	17
Animal words (real or toy)	*Bee, cat, donkey*	13	30
Vehicle words	*Car, taxi/combi*	10	11
Words for toys	*Ball, game*	11	14
Food and drink	*Fruit, sourmilk, sweets*	42	69
Words for clothes	*Jersey, shorts*	17	26
Words for body parts	*Arm, eye, tummy*	22	31
Words for small household items	*Bucket, matches, spoon*	36	62
Furniture words	*Bathtub, door, television*	20	27
Outside words	*Garden, mountain, stone*	11	20
Words for places to go	*Creche/school, place, yard*	7	13
Words for people	*Child, mommy, uncle*	12	21
Words for games and routines	*It's hot, high five, please*	28	34
Action words	*Bite, go, sleep*	76	138
Describing words	*Bad, clean, yucky*	16	59
Words about time	*Today, now, morning*	4	6
Words about people and things	*His/hers, me, this*	8	19
Question words	*What, why*	6	7
Words about places	*Behind, here, under*	12	19
Words about amounts	*All, more, some*	6	12
Connecting words	*And, so*	1	5
Total		375	642

Table 2: Number of items per semantic domain of the words section, average across languages

The grammar section of the CDIs is divided into four subsections. The first concerns noun and verb affixes, representing both singular and plural noun classes and past and present tense markers. These are presented in the form of *yes/no* questions. For example, caregivers are asked the equivalent of *"Has your child started adding endings to words to show that an event has already happened?",* with two or three language-appropriate examples provided.

The second subsection asks in more detail about the use of noun class prefixes and verb affixes. The first 10 noun classes are covered as singular and plural pairs, e.g., Class 3 (singular) and Class 4 (the plural of Class 3), but there are only 8 items because, for some of the languages, (i) Classes 8 and 10 have the same prefixes (with nouns in Class 10 occurring more frequently), and (ii) Class 9 has a null prefix and/or occurred less frequently in our language samples and was thus not included. The items are presented as a trio of words with increasing complexity, i.e., a noun stem with no prefix (for instance, in isiXhosa *fazi* '(married) woman'), a noun stem with a 'shadow vowel' or place holder prefix (*mfazi*), and a noun stem with a full, correct prefix (*umfazi*); see Tsonope (1993) and Demuth (1988) for a discussion of these three stages of noun class prefix acquisition. Noun stems that exemplify each item were selected based on word frequency in the language samples. This subsection also asks about the use of verb affixes. These are presented as a pair or trio of words, again in increasing complexity, i.e., a verb stem with no affix, a verb stem with the full affix, and in some cases a middle option of a partial or incorrect affix.

The third subsection asks the caregiver the equivalent of *"Has the child started to combine words to form short sentences?",* with two

language-appropriate examples provided. If the caregiver responds affirmatively, they are asked to provide three examples of the longest sentences they heard the child say that week.

The last grammar subsection asks about the length and complexity of sentences. These are given as an example with two to four options of increasing complexity. The caregiver is asked to select the form that most closely resembles what their child would say – for example, *"I want bread"* or *"I want bread and a drink".*

The number of items per subsection is presented in Table 3, along with an indication of the types of constructions that the questions ask about.

Subsection	Types of constructions	No. of items
Small parts of words	Use of prefixes and suffixes	5
Word complexity	Noun class prefixes Verb suffixes	19-21
Word combinations	Whether the child is combining words, with 3 recent examples of the longest sentences	4
Sentence complexity	*"ball table"* vs *"ball top of table"* vs "*ball is on top of the table*"	13

Table 3: Number and types of items per category of the final grammar sections

3. Child Language Corpora

We developed three types of corpora, namely one corpus for each of the nine languages containing the caregiver responses to the infant CDIs and another containing those to the toddler CDI, as well as a corpus consisting of orthographically transcribed language samples (and their audio recordings) for six of the nine languages[4]. Each of these types of corpora is briefly discussed below.

3.1 Infant CDI Corpora

The infant CDI corpora consist of the answers that the caregivers gave to the CDI items on early communicative gestures and actions and on words that the child either comprehends or comprehends and produces, as well as background information on each infant for which the CDI was completed. The background information was on the infant's birth and medical history, general health, childcare, exposure to languages, household composition, and household resources. The data of 988 infants (approximately 110 per language) are included in the form of one searchable Excel file per language. This file contains instructions for the user and separate tabs for gestures and actions and for the vocabulary items. Table 4 contains the characteristics of the completed CDIs included in the corpus of each language.

Language	Total number	Rural (%)	Female (%)
IsiNdebele	112	62 (55.4%)	55 (49.1%)
IsiXhosa	109	53 (48.6%)	53 (48.6%)
isiZulu	99	52 (52.5%)	45 (45.5%)
Sesotho	111	58 (52.3%)	59 (53.2%)
Sesotho sa Leboa	111	74 (66.7%)	60 (54.1%)
Setswana	97	46 (47.4%)	46 (47.4%)
Siswati	117	55 (47%)	60 (51.3%)
Tshivenda	126	56 (44.4%)	63 (50%)
Xitsonga	105	82 (78.1%)	43 (41%)
Total	987	538 (54.5%)	484 (49%)

Table 4: Characteristics of the infant corpora, per language

3.2 Toddler CDI Corpora

As was the case for the infant CDI corpora, the nine toddler corpora each contain background information on the toddlers. Also included are the responses of the caregivers to the CDI items on words that the child produces (and therefore, by implication, comprehends as well) and the types of grammar constructions that the child can use. Searchable Excel files for each language contain data for 1050 toddlers (approximately 116 per language) in several tabs: As for the infant corpus, one tab contains user instructions; the others contain the background information for each child, as well as the vocabulary and grammar data. The characteristics of the completed CDIs included in the toddler corpora can be seen in Table 5.

Language	Total number	Rural (%)	Female (%)
IsiNdebele	123	63 (51.2%)	61 (49.6%)
IsiXhosa	107	57 (53.3%)	55 (51.4%)
isiZulu	115	53 (46.1%)	55 (47.8%)
Sesotho	112	57 (50.9%)	64 (57.1%)

[4] Development of a language sample corpus for each of the remaining three language (Sesotho sa Leboa, Setswana and Xitsonga) is underway and will also consist of transcribed language samples and their audio recordings.

Sesotho sa Leboa	123	72 (58.5%)	65 (52.8%)
Setswana	119	61 (51.3%)	58 (48.7%)
Siswati	128	61 (47.7%)	64 (50%)
Tshivenda	128	58 (45.3%)	62 (48.4%)
Xitsonga	95	41 (43.2%)	50 (52.6%)
Total	1050	523 (49.8%)	534 (50.9%)

Table 5: Characteristics of the toddler corpora, per language

3.3 Toddler Language Samples

For isiNdebele, isiXhosa, isiZulu, Sesotho, Siswati, and Tshivenda, there are 20 transcribed language samples. These samples were collected from 10 male and 10 female toddlers of 28 to 30 months. The language samples are video recordings of natural interaction between the toddler and (a) familiar adult(s) and/or child(ren). For each toddler, there was collectively 30 to 60 minutes of recordings (see Table 6) which in total contained at least 50 different utterances per child. Some recordings were made by the fieldworker and others by the parents, other caregivers or other adults or children. The toddlers were filmed in and/or around their homes and/or daycares during everyday activities such as indoor/outdoor play or having a meal. Most children had more than one recording, because we wanted to capture conversations in various settings, with recordings ranging in length from 1 to 60 minutes. Recordings were transcribed in CHAT format (see CHILDES; MacWhinney, 2000) in order to render them ready for analysis in CLAN (MacWhinney, 2000). Table 6 indicates the characteristics of the participants who contributed to the language sample corpora and the length of the recordings, for each of the six languages for which this corpus construction has been completed.

Language	Rural / urban	Recording length (in minutes)		
		Range per child	Mean per child	Combined per language
isiNdebele	Semi-urban	50-89	64.6	1292
isiXhosa	Urban	32-69	64.5	1281
isiZulu	Semi-urban	29-69	48.5	970
Sesotho	Rural	50-67	58.7	1174
Siswati	Semi-urban	29-69	51.9	1038
Tshivenda	Rural	60-77	62.8	1255

Table 6: Characteristics of the toddler language sample corpora, per language

4. Possible Research Uses

Although research has been conducted on child language for decades already, there are only a few well-researched languages in terms of child language development, and none of these are African languages (see Kidd and Garcia, 2022). The corpora enable one to answer a range of questions on the nature and size of the vocabulary of young speakers of African languages and on how this changes as the child ages; on the types of morphology that develops first and on how morphological development progresses between the ages of 16 and 30 months; on the mean length of the utterances of toddlers of various languages; on the relationship between child characteristics, household characteristics and child experiences on the one hand and language measures (communicative gestures, vocabulary, and grammar) on the other – for any one of the nine languages or crosslinguistically.

CDIs are used the world over as data collection instruments for research and diagnostic purposes. They allow one to measure and track language development in and of itself, but also as part of studies not pertaining to language development per se, such as studies on the effect of dialogic reading, medical treatment, or creche attendance on a child's development, of which language development forms an important part. Adding nine more language-versions of the CDI to the collection of existing CDIs significantly increases the scope of such research, allowing for the inclusion of child speakers of a wider range of languages. This enables contextually relevant research findings to be generated, which can inform contextually relevant early childhood intervention programmes.

5. Bibliographical References

Alcock, K.J., Rimba, K., Holding, P., Kitsao-Wekulo, P., Abubakar, A., and Newton, C.R.J.C. (2015). Developmental inventories using illiterate parents as informants: Communicative Development Inventory (CDI) adaptation for two Kenyan languages. *Journal of Child Language*, 42:763–785.

Anđelković, D., Ševa, N., Savić, M., and Tutnjević, S. (2014). Izveštaj roditelja kao izvor podataka o ranom razvoju dečijeg govora (Parents' report as a source of information on child language development). XX naučni skup Empirijska istraživanja u psihologij. http://empirijskaistrazivanja.org/wp-content/uploads/2016/06/Knjiga-Rezimea-EIP-2014.pdf

Catts, H.W., Fey, M.E., Zhang, X., and Tomblin, J.B. (1999). Language basis of reading and reading disabilities: Evidence from a longitudinal investigation. *Scientific Studies of Reading*, 3:331–361.

Connelly, M. (1984). *Basotho children's acquisition of noun morphology*. Unpublished Ph.D. dissertation, University of Essex, UK.

Demuth, K. (1992). The acquisition of Sesotho. In D. Slobin (Ed.), *The Crosslinguistic Study of Language Acquisition*, 3. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 557-638.

Demuth, K. (1988). Noun classes and agreement in Sesotho acquisition. In M. Barlow, C.A. Ferguson (Eds.), *Agreement in natural languages: Approaches, theories and descriptions*. CSLI: University of Chicago Press, pp. 305-321.

Demuth, K. (2003). The acquisition of Bantu languages. In D. Nurse, G. Philippson (Eds.), *The Bantu languages*. Surrey, England: Curzon Press, pp. 209-222.

Fenson, L., Bates, E., Dale, P.S., Goodman, J.C., Reznick, J.S., and Thal, D. (1993). *The MacArthur communicative development inventories: User's guide and technical manual*. Baltimore, MD: Paul H. Brookes.

Fenson, L., Marchman, V.A., Thal, D.J., Dale, P.S., Reznick, J.S., and Bates, E. (2007). *MacArthur–Bates Communicative Development Inventories user's guide and technical manual.* Baltimore, MD: Paul H. Brookes.

Frank, M.C., Braginsky, M., Yurovsky, D., and Marchman, V.A. (2017). WordBank: An open repository for developmental vocabulary data. *Journal of Child Language*, 44:677-694. https://doi.org/10.1017/S0305000916000209

Fricke, S., Bowyer-Crane, C., Haley, A.J., Hulme, C., and Snowling, M.J. (2013), Efficacy of language intervention in the early years. *Journal of Child Psychology and Psychiatry*, 54:280-290. https://doi.org/10.1111/jcpp.12010

Jarůšková, L., Smolík, F., Chládková, K., Oceláková, Z., and Paillereau, N. (2023). How to build a Communicative Development Inventory: Insights From 43 Adaptations. *Journal of Speech, Language, and Hearing Research*, 66:2095-2117. https://doi.org/10.1044/2023_JSLHR-22-00591

Jackson-Maldonado, D., Thal, D., Marchman, V., Bates, E., and Gutierrez-Clellen, V. (1993). Early lexical development in Spanish-speaking infants and toddlers. *Journal of Child Language*, 20:523-549. doi:10.1017/S0305000900008461

Kidd, E. and Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, 42:703-735. https://doi.org/10.1177/01427237211066405

Kunene, E. (1979). *The acquisition of Swati as a first language: A morphological study with special reference to noun prefixes, noun classes and some agreement markers.* Unpublished Ph.D. dissertation, University of California at Los Angeles.

Suzman, S. (1991). *The acquisition of Zulu.* Unpublished Ph.D. dissertation, Witwatersrand University, Johannesburg, South Africa

Tsonope, J. (1987). *The acquisition of Tswana noun class and agreement morphology, with special reference to demonstratives and possessives.* Unpublished Ph.D. dissertation, State University of New York, Buffalo.

Tsonope, J. (1993). Children's acquisition of Bantu noun class prefixes. *Botswana Notes & Records*, 25(1):111-117.

6. Language Resource References

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk*. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates. https://childes.talkbank.org/

7. Acknowledgments

This study was financially supported by the South African Center for Digital Language Resources. Additional funding came from The National Research Foundation of South Africa (HSD170602236563). Preliminary work for this research was supported by The British Academy Newton Fund (NG160093) and the National Research Foundation of South Africa/Swedish Foundation for International Cooperation in Research and Higher Education (NRF/STINT160918188417). Any opinion, findings, conclusions or recommendations expressed in this material are those of the authors.