26-05-2025

Data Processing Children and Language Mixing (CALM) from Utrecht University: from LENA recording to quantitative data on language mixing.

Below we will discuss the steps that were taken in the CALM project to process our LENA data. It is explained in detail how we go from a full LENA recording to quantitative data on language mixing behavior. All manuals that our research assistants used for coding, transcribing and analyzing the code-switches can be found on the overarching project page on the Open Science Framework (https://osf.io/c4rm8/). For further questions, feel free to contact me at e.g.e.c.verhoeven@uu.nl.

Sampling the audio

We have 52 recordings, with mean length of 14.1 hours (*sd* = 3.4). As analyzing 14 hours' worth of audio recordings per participant is an unfeasible amount of work, the data is sampled based on number of adult words (AWC), child vocalizations (CVC) and conversational turns (CTC), provided for fragments as small at 5-minutes by the automatic output from the LENA software.

Based on common practice (Marasli & Montag, 2023; Orena et al., 2020; Ramírez-Esparza et al., 2014, 2017), we first remove silent fragments that contain no speech. This is often during naps or tv-time. Then, to portray the language input during an entire day, we sample segments that represented periods of high, medium, and low interaction. Based on personal communication with the LENA foundation, we decided to select 18 5-minute segments (1.5 hours of audio) that contain the highest number of conversational turns, 18 segments that contain the lowest number of conversational turns (but are not silent) and 18 segments that are in between. This way we get 4.5 hours of audio recordings containing different levels of speech quantity in terms of conversational turns. From those 4.5 hours, we analyze every other 30-second segment (Marasli & Montag, 2023; Ramírez-Esparza et al., 2014, 2017), reducing the amount of audio segments that need to be analyzed to 2h 15min per participant (270 30-second segments). This is considered more than sufficient to reliably reflect the full day-long audio recording (Cychosz et al., 2021; Marasli & Montag, 2023)

Coding the segments

Subsequently, we coded the 30-s segments manually for the speaker(s), language(s) spoken, activity, and whether there is speech directed to the target child (CDS). If a segment contained more than one language, it was determined by the coder in a separate column which of these languages occurred most frequently, or whether both languages occurred equally. The coding is exemplified in Table 1. The coding manual is available on the Open Science Framework. Coders were two bilingual Turkish–Dutch and three bilingual Polish–Dutch research assistants. They were trained in their coding abilities and the inter-rater reliability between assistants of the same language was determined based on the average Kappa scores over the columns Speaker, Language, Dominance, and CDS of one participant (270 segments). For both Polish–Dutch (κ = 0.81) and Turkish–Dutch (κ = 0.82), very strong inter-rater reliability was obtained.

Table 1

The coding system of the 30-second segments

Subject	Segment	Speaker	Language	Dominance	CDS	Activity
00001	1	FAT	Dutch	Dutch	CDS-ADULT	Mealtime
00001	2	CHI, SIB, FAT	Dutch, Polish	Polish	ODS	Mealtime
00001	3	CHI, FAT	Dutch	Dutch	CDS-ADULT	Mealtime
00001	4	CHI, FAT, MOT	Dutch, Polish	Both	CDS-ADULT	Playtime

Transcribing the mixed segments

All 30-second segments that contain more than one language, and thus potentially an instance of language mixing, are fully transcribed in CHAT (McWhinney, 2000). The full transcription manual can be found on the Open Science Framework (https://osf.io/c4rm8/). One CHAT file was created per participant, in which all relevant 30-second segments were transcribed. We used codes to mark the instances of language mixing during transcribing (see Table 2). Intra-sentential switches were coded with @s as a suffix to each word. Inter-sentential switches were coded with an int@x in the beginning of the sentence. Transcriptions were done by bilingual speakers of Turkish, Polish and Dutch, most of whom also coded the 30-second segments. All assistants were trained on two English-Dutch examples where they had to correctly transcribe 90% of all instances of language mixing in 10 segments before they were allowed to transcribe any further data.

Table 2

Examples of coding during the transcription of intra- and inter-sentential switches

Type mixing	Example	Transcript
Intra-sentential	1	CHI: mag ik helpen om **mercimek@s çorbası@s** te maken ?
	2	CHI: to jest **op@s mijn@s been@s .**
Inter-sentential (within speaker)	1	*CHI: [- nld] niet op . *CHI: **int@x gdzie jest ?**
	2	*MOT: [- nld] beetje stiekem gaan we ja ? *MOT: [- nld] ja ? *MOT: **int@x to ty idź teraz poczekaj na mnie przy drzwiach .**
Inter-sentential (between speakers)	1	*MOT: no po prostu podnieś . *CHI: **[- nld] int@x die zijn lekker .**
	2	*CHI: [- nld] nee dat betekent dat je welke hebt . *MOT: **int@x ale to musisz sam to zrobić .**

Mixing analysis

As a final step, all instances of language mixing were further analyzed in Excel. A Python script extracted all sentences containing either ‘int@x’ or ‘@s’ and who the speaker was from all transcripts into one document. The same two Turkish-Dutch and two Polish-

Dutch research assistants who had also transcribed the data then coded each instance of language mixing. Every language mix was coded for addressee, language, number of switches within one utterance, direction of the switch, and type of switch. To determine the type of switch, we first code whether the switch was within or between speakers in the column Switch_Type. As language mixing between speakers always takes place between sentences, it is per definition inter-sentential. Within a speaker, different types of switches can occur. We distinguish between inter-sentential switches, insertions, alternations and congruent lexicalization according to the typology of Muysken (1997). An example of the mixing analysis is given in Table 3. The inter-rater reliability was determined based on the average Kappa scores over the columns Switch_Addressee (whether the mix was directed to the target child or not), Language, Switch_Number, Switch_Direction, Switch_Type and Switch_Type_Within of 100 code-switches. For both Polish–Dutch (κ = 0.92) and Turkish–Dutch (κ = XX), almost perfect inter-rater reliability was obtained.

Table 3

Example of language mixing analysis

Switch_ SpelledOut	**Switch_ Speaker**	**Switch_ Addressee**	**Language**	**Switch_ Number**	**Switch_ Direction**	**Switch_ Type**	**Switch_ Type_ Within**
hagelslag@s mı var ?	MOT	SIB	Turkish, Dutch	1	1	Within_ Speaker	Insertion
int@x ja .	SIB	MOT	Dutch	1	1	Between_ Speakers	Inter-sentential
int@x kalk bakalım .	MOT	CHI	Turkish	1	2	Within_ Speaker	Inter-sentential

Note. Direction is 1 = Polish/Turkish → Dutch; 2 = Dutch → Turkish/Polish. 3 = if switch is between two other languages. Abbreviations: MOT = mother, SIB = sibling, CHI = target child

In the present study, we only distinguish between intra- and inter-sentential language mixing (Poplack, 1980). Thus, all instances of insertion, alternation or congruent lexicalization are seen as intra-sentential language mixing. Now we are able to calculate the exact number of intra- and inter-sentential switches that were made by parents when they spoke to the target child.

26-05-2025

Guidelines Mixing Analysis Children and Language Mixing (CALM) project Utrecht University

Variable	**Description**	**Criteria**	**Examples**
Subject	Unique child ID number		
Segment	From which segment the utterance is from.	The number of the segment in which the utterance is from. This can be found in CHAT, and can be used to go back to the original audio file if a larger context is needed.	1 - 270
Row	On which row in CHAT is the utterance?	Number of the row in the CHAT file that belongs to the utterance. This can be used to quickly find the utterance and its context in CHAT.	1 - 100000
Switch_ SpelledOut	The literal utterance from CHAT.	The utterance in which the switch occurs is literally copied from CHAT to here.	kan iemand sól@s aangeven? Ik kan mijn plaszcz@s niet vinden. [- nld] int@x ga maar naar bed.
Switch_ Speaker	Who said the utterance?	The same abbreviation that is used in CHAT to describe who has said the mixed utterance. You can use CHI (Target_child), MOT (Mother), FAT (Father), SIB (Sibling), MAL (Male, for every other male adult), FEM (Female, for every other female adult), OTC (Boy or Girl, for other children who are not siblings of the target child)	CHI (Target_child), MOT (Mother), FAT (Father), SIB (Sibilng), MAL (Male, for every other male adult), FEM (Female, for every other female adult), OTC (Boy or Girl, for other children who are not siblings of the target child)

Switch_ Addressee	Who is the utterance directed to?	Use the same abbreviations as used in CHAT (FAT, MOT, SIB, etc.). If an utterance is directed to multiple people (e.g. all people at the dining table) then we code all those addressees.	MOT: kan iemand sól aangeven? --> CHI, FAT, SIB FAT: Zijn jullie klaar om te gaan? --> CHI, SIB FAT: Trek jij nog even een jas aan? --> SIB SIB: ik kan mijn jas niet vinden! --> FAT
Language	Which languages occur in the sentence?	Note down all the languages that occur in the sentence	MOT: kan iemand sól aangeven? --> Dutch, Polish FAT: Zijn jullie klaar om te gaan? --> Dutch
Switch_ number	What number of switch (from one utterance) is this? Put inter-sentential switches (int@x) and intra-sentential switches (@s) on separate rows.	Any utterance that only has 1 switch will have a value of 1. For utterances with more than 1 switch, copy/paste the row for a separate analysis of inter-sentential switches (int@x) and intra-sentential switches (@s). All intra-sentential switches (@s) can be analyzed on the same row. Each row will get its own number (1, 2, 3…). Bold and italicize the switch in the translation column that is being coded in that row. Complete rest of row for switch of interest.	1, 2, etc…

Switch_Direction	From which language to which language is being switched?	The language that the switch is in is the direction towards the language. For example, if you say "Kan je de sól@s aangeven", then sól is Polish. So the direction of the switch is from Dutch --> Polish. We distinguish three different categories: 1 = from the minority language (Turkish/Polish/English) --> Dutch 2 = from Dutch --> the minority language (Turkish/Polish/English) 3 = if the switch involves any other language	ik kan mijn ayakkabı niet vinden. --> 2 (Dutch --> Turkish) Yağmur yağıyor. Zo ga ik niet naar buiten. --> 1 (Turkish --> Dutch) Ga maar spelen met de voiture@s --> 3 (Dutch --> French)
Switch_Type	What type of switch occurs?	We distinguish three types of switches:Within_Speaker (the same speaker switches between languages)Between_Speaker (a speaker responds in a different language than the previous sentence)Multiple_Conversations (if it seems like there is a switch in the CHAT file, but in reality there are two different conversations happening at the same time. This is often commented on in the CHAT file).	CHI: En toen was er een köpek@s. --> Within_Speaker (within speaker, as the child is switching between both languages)MOT: vandaag eten we spinazie.MOT: jedliśmy wczoraj brokuły. --> Within_Speaker (within speaker, the mother uses both languages)MOT: yatağa gitmelisin.SIS: maar ik ben helemaal niet moe. --> Between_Speaker (the sister responds in a different language than her mother addresses her in).FAT: trek maar snel je coat@s aan . MOT: [- eng] int@x I haven't seen Phil all week . --> Multiple_Conversations (even though the sentences are in a different language, the mother is having a different conversation than the father, and thus the switch is between two different conversations).

26-05-2025

Switch_ Type_Within	If the switch is within speaker, what type of switch is it?	Insertion Alternation Congruent_Lexicalization Intersentential You can find a detailed description and examples for each category below	
		An insertion is one word (group) in the other language within the sentence. If there are multiple words that constituate the same group (e.g. the nice lady are 3 words but 1 NP), it is still considered an insertion.	CHI: En toen was er een köpek@s. --> Insertion (there is 1 word (group) inserted) CHI: En toen köpek@s var@s mıydı@s --> Alternation (there are more word groups inserted, which makes it an alternation and not an insertion) CHI: En toen zag hij bir@s köpek@s --> Insertion (bir köpek is one inserted word group)
		An alternation is the switch from one language to the other within the sentence. Usually, this contains both grammatical and lexical items.	MOT: Hij was aan het wandelen ze@s swoim@s psem@s --> Alternation (ze swoim psem contains multiple word groups (both grammatical and lexical). CHI: Hij eet çok@s güzel@s bir@s dondurma@s --> Insertion (it's only one word group (NP) that is inserted, no grammatical items are inserted).

		Congruent lexicalization is the switching back and fourth between two languages in such a way that you could not distinguish a 'main language'. It looks like multiple cases of alternations and insertions in one utterance.	MOT: Gisteren oynayan iki çocuk gördüm in de speeltuin slaytta --> Congruent Lexicalization (both grammatical and lexical items of both languages are used throughout the whole sentence)FAT: dün iki çocuk gördüm in de speeltuin --> Alternation (there is only one switch from Turkish to Dutch, which would be alternation).
		Inter-sentential is a switch where two adjacent sentences are in a different language.	MOT: vandaag eten we spinazie. MOT: jedliśmy wczoraj brokuły. --> Inter-sentential (the two sentences are adjacent, but in different languages) MOT: vandaag eten we szpinak. MOT: jedliśmy wczoraj brokuły. --> Inter-sentential (even though she switched in the first sentence, the matrix language was still in Dutch, and the following adjacent sentence in Polish). MOT: vandaag eten we spinazie. FAT: nie masz ochoty na brokuły? --> Inter-sentential (the adjacent sentences are in different languages, even though they are uttered by different speakers).
Remarks_ Mixing	If switch cannot be classified into a category	If you had difficulty in coding one of the variables, please elaborate here why the utterance / switch was not suitable to be coded along these guidelines. Other comments about the switch can be put here as well.	If you had any trouble filling out any column, you can place a comment here how/why.