Caregivers Exaggerate Their Lexical Alignment to Young Children Across Several Cultures

Thomas Misiek
Aix-Marseille University
thomasmisiek@gmail.com

Abdellah Fourtassi
Aix-Marseille University
abdellah.fourtassi@gmail.com

Abstract

As soon as they start producing their first words, children engage in dialogues with people around them. Recent work has suggested that caregivers facilitate this early linguistic communication via frequently re-using and building on children's own words. This tendency decreases over development as children become more competent speakers. While this pattern has been observed with data of English-learning children, the question remains as to whether this early child-caregiver dynamics is universal vs. culture-specific. We address this question using large-scale data in six languages belonging to both Eastern and Western cultures. We found that the finding generalizes well cross-linguistically, suggesting that caregivers' early "exaggerating" of lexical alignment is likely a scaffolding strategy used across cultures to facilitate children's early linguistic communication and learning.

1 Introduction

Lexical alignment is a phenomenon whereby interlocutors re-use each other's words in a dialog. For example:

- Speaker 1 :"How do you think this **is going?**"
- Speaker 2 :"Yes, I guess it **is going** well!"

Researchers have hypothesized this mechanism to be associated with dialog coordination, facilitating language processing and production and contributing to the collaborative process of building mutual understanding and, thus, communicative success more generally (Pickering and Garrod, 2004, 2006; Brennan and Clark, 1996).

Interestingly, a similar behavior has been documented in child-adult natural dialog, starting from the early stages of the child's language production (Dale and Spivey, 2006; Fernández and Grimm, 2014; Denby and Yurovsky, 2019; Fusaroli et al., 2021; Misiek et al., 2020; Yurovsky et al., 2016; Foushee et al., 2021).

In particular, two large-scale studies — using data from hundreds of children — by Yurovsky et al. (2016) and Misiek et al. (2020) converged on similar conclusions despite the fact they used different measures and focused on different aspects of alignment. The main finding was that caregivers *exaggerate* their re-use of children's early words/expressions when communicating with them. Another finding was that this exaggerated alignment decreases over time and becomes closer to children's own level of lexical alignment (as well as adult-adult alignment rate) by the end of the preschool period. A similar pattern was also observed in the context of second language (L2) learning between tutors and students (Sinclair and Fernández, 2021).

While lexical alignment is sometimes assumed to be largely automatic and priming-like in spontaneous adult-adult dialog (e.g., Pickering and Garrod, 2004), here the observed patterns of alignment suggest otherwise. In particular, the fact that adults align much more to young children (than the other way around), as well as the fact that there is a negative correlation between the adults' alignment and the children's age — and therefore their language proficiency — provide evidence that caregivers actually align as a *scaffolding strategy* to help the younger — less language proficient — children understand and/or learn (e.g., Vygotsky, 1978; Shafto et al., 2014; Yurovsky, 2018). Such a strategy would be less useful to older children with more developed linguistic skills and who need less communicative scaffolding from the caregiver.

1.1 The current study

The study of child-caregiver early lexical alignment dynamics has focused on data from English-learning children. It is still unknown whether the above-mentioned findings generalize to other languages/cultures, especially in the light of research that has pointed out cross-cultural dissimilarities in

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

the way caregivers interact with children early in development (Bornstein et al., 1992; Saint-Georges et al., 2013; Schick et al., 2022).

Addressing this question is of crucial scientific interest: It allows us to determine if the interactions observed between English-learning children and their caregivers reflect more the specificities of their culture (e.g., in terms of parenting style) or whether they represent universal patterns in human development across cultures. The current study is an effort to fill this gap. We conduct a large-scale study of lexical alignment in child-caregiver dialogues, comparing 6 languages: English, Chinese, Spanish, German, Japanese, and French.

2 Methods

2.1 Data

All the data is derived from CHILDES (MacWhinney, 2000; Sanchez et al., 2019), the largest public repository of child-caregiver dialog corpora. First, as shown in Table 1, we ranked all languages based on the size of their aggregated corpora. We aimed at selecting the subset of languages with the largest sizes, making sure we include at least 2 non-western cultures. Japanese was the second largest non-western language (after Chinese) with around 0.5 million words. We included French, which came next, since it had an approximately similar size as Japanese. We did not include the next language in the list since their size dropped significantly.

We focused on development in the pre-school period, ranging from 2 and 5 years old (data in CHILDES becomes too sparse below and above this range). Table 1 provides some summary statistics of the data we use. We note the heterogeneity in terms of the number of transcripts per child across languages, reflecting heterogeneity in data collection procedures (e.g., cross-sectional vs. longitudinal).

2.2 Measure of lexical alignment

Lexical alignment characterizes the speaker's reuse of words from the interlocutor's previous turns in the dialog. Following previous work (e.g., Fernández and Grimm, 2014; Misiek et al., 2020), we quantified this phenomenon by counting the number of shared unigrams (unique words) and bigrams (sequences of two successive words) across adjacent pairs of turns, normalized by the number of all possible ngrams.

Language	Words	Transc.	Children
English	11,801,282	5894	869
German	2,008,317	1073	54
Chinese	1,023,867	508	329
Spanish	665,789	493	63
Japanese	543,495	652	122
French	538,663	724	192
Slavic	385,839		
Afrikaans	288,927		
Romance	230,101		
Scandinavian	168,629		

Table 1: Top 10 languages with largest (aggregated) corpora in CHILDES. We focused on the top 6 with at least 0.5 million words each. For these languages, we show the number of transcripts (dialog sessions) and unique children aged 2 to 5 years.

We computed both `Child alignment` by comparing the child's turn to the adult's *previous* turn and `Adult alignment` by comparing the child's turn to the adult's *following* turn. In both cases, the pairs of turns have to be adjacent. If the same speaker has multiple consecutive utterances, only the first and the last were taken into account since only the first and last are adjacent to the interlocutor's utterances.

Baselines

In addition to the child's and caregiver's alignment measures, we derived two baselines. The first, which we call the `internal baseline`, computes the alignment of pairs of turns (one belonging to the child and the other to the caregiver) sampled randomly from the same transcript/conversation. The second, which we call `external baseline`, compares pairs of child and caregiver turns sampled randomly from the entire corpus (within a given language).

3 Results

Our first goal is to replicate findings for English data as reported in both Misiek et al. (2020) and Yurovsky et al. (2016). The second goal — and the novel contribution of the current study — is to test how previous findings in English generalize cross-linguistically. The results are shown in Figure 1.

We found the following findings both to replicate in English and to generalize well across languages:

1. Children align consistently to their caregiver, starting from the early stages of language production.

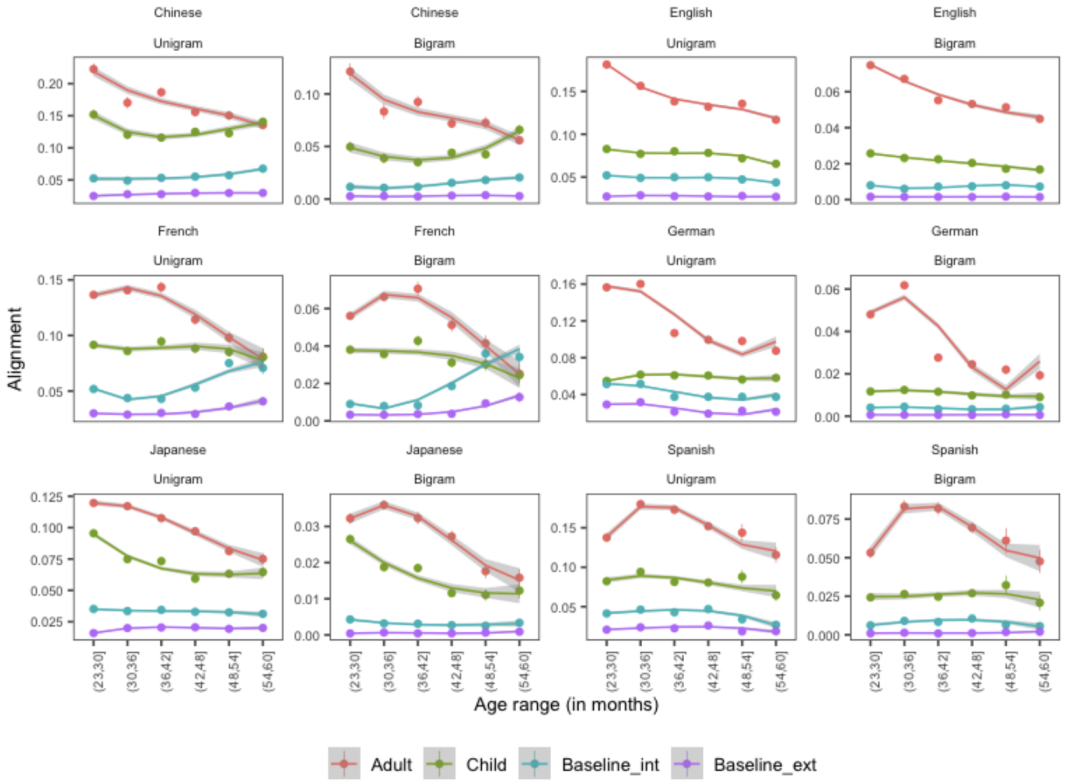

Figure 1: Lexical alignment measured in terms of shared unigrams and bigrams (normalized), as a function of the child's age across languages. The lines are fitted with polynomial splines of degree 3 with 95% confidence intervals.

2. Caregivers align consistently more to children (than the other way around).

3. Caregivers align more when children are younger. Their alignment decreases as children develop.

We corroborate these observations with statistical testing, but first we need to examine the shape of the data and make some simplifications. Figure 2 shows the distribution of (normalized) alignment values. It shows a 0-inflated distribution of a semi-continuous dependent variable. In other words, a substantial chunk of child-caregiver adjacent turns shows no alignment (i.e., the alignment value is exactly 0) and the rest is continuous between 0 and 1.

Standard normality transformations of such data do not solve the zero-inflation issue. One possible solution (to still be able to fit parametric models) is to consider a two-stage approach: a logistic regression predicting the binary 0 vs. non-0 outcome and a linear regression predicting the continuous outcome in the interval]0,1] (e.g., Gelman and Hill, 2006).

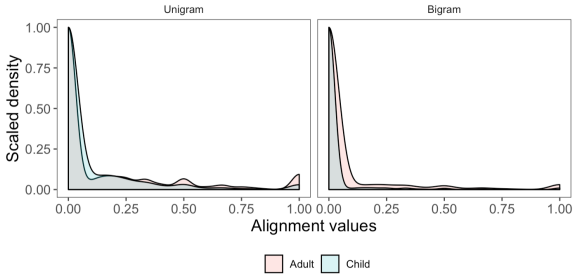

Figure 2: Scaled density plot of normalized alignment values in both unigram- and bigram-based measures, collapsed across all languages.

Here we are interested in comparing data across development, and the linear regression would, however, introduce distortions/biases, e.g., by creating a data size imbalance across ages. The reason is that restricting the data to the]0,1] interval requires selecting, at each developmental stage, only the subset of adjacent turns that include non-zero alignment. This would make it hard to interpret any observed developmental change.

Thus, for simplicity, here we only report results of the logistic regression predicting whether or not adjacent turns have at least one shared lexical un-

igram (for the unigram-based measure) or a least one shared lexical bigram (for the bigram-based measure). The logistic regression (unlike the linear regression on]0,1]) does not require removing data, only reducing its complexity from continuous to binary. This makes the interpretation of developmental *change* much more intuitive.

More precisely, we used mixed-effects logistic regressions, predicting the binary alignment (for both the unigram and bigram measures) as a function of the condition (Child vs. Adult) and age, using the identity of the child and the language as random effects. The results of these two regressions are shown in Table 2.

All predictors were highly significant, confirming the patterns observed in Figure 1: The predictor `Condition` indicates that caregivers align to children to a higher degree (than the other way around). `Age` negatively predicted alignment, showing that alignment decreases with development. The interaction `Condition*Age` shows that caregivers' alignment decreased faster than children's alignment did, confirming the observation that caregivers exaggerate alignment more to younger children than to older ones.

	Alignment	
	Unigram	**Bigram**
`(Intercept)`	-0.658^{***}	-2.207^{***}
	(0.030)	(0.038)
`Condition`	-0.575^{***}	-0.857^{***}
	(0.003)	(0.005)
`Age`	-0.110^{***}	-0.110^{***}
	(0.002)	(0.004)
`Condition*age`	0.125^{***}	0.065^{***}
	(0.003)	(0.005)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 2: Estimates of two mixed-effects logistic regressions models predicting the presence of alignment (Unigram and Bigram) in adjacent child-caregiver turns as a function of `Condition` (who is aligning to whom) and child's `Age` (centered and scaled). The model was specified as `Alignment_or_not` $\sim$ `Condition*Age + (1 | child) + (1 | Language)`.

Cross-linguistic differences

In addition to the consistent cross-linguistic similarities, Figure 1 also shows some (minor) differences. For example, we can observe that the caregivers' decreasing alignment matches that of children by 5 years in some languages (i.e., Chinese, French, and Japanese) but not in others (i.e., English, German, and Spanish). In the latter case, it appears that caregivers are still exaggerating alignment despite children's relatively developed linguistic skills by that age.

Another difference concerns the pattern of children's alignment. While the developmental curve is rather stable in most languages, it tends to decrease in Japanese (although at a slower pace than the caregivers' curve does) and to slightly increase in Chinese.

We can also observe that for some languages, especially Spanish and French, the caregivers' curve tends to show an inverted U-shaped curve whereby the youngest children receive less alignment than the slightly older ones (before the curve starts decreasing again). This observation could be due to the fact that younger children have limited language production skills, providing much fewer opportunities for caregivers to align. This interpretation is supported by the fact that the inverted U-shaped curve is more pronounced in the bigram case, i.e., the case where children's utterance has to contain at least two words to provide the opportunities for the caregivers to align at the bigram level; the youngest children do produce much shorter utterances than older children do.

Finally, we observe that in French, the alignment curves become indistinguishable from the random baseline towards the end of the developmental period under study. However, this is likely due to the fact that in French (unlike all other languages), data of the oldest children had a much smaller sample size in CHILDES than the younger ones (data not shown), leading to noisier data by 5 years old.

4 Discussion

Lexical alignment is an important mechanism for dialog coordination in adults. Recent studies suggest it could play a role in child development as well: Adults tend to re-use children's words more frequently in the earlier stages of language production, perhaps scaffolding children's communicative and linguistic skills.

This paper showed that this finding generalizes

well — beyond English — to five different languages, including in three Western cultures (German, Spanish, and French) and two Eastern ones (Chinese and Japanese). The finding was strikingly similar despite variability in how data was collected across languages, and more important, despite the fact that Eastern and Western cultures are sometimes assumed to differ in terms of parenting style (Foo, 2019). The strong similarity among these languages points toward a rather *universal* pattern that characterizes the evolution of child-caregiver dialog dynamics across the first five years of life.

The developmental literature reports several scaffolding mechanisms that may underlie this finding. For example, caregivers tend to build on words and concepts that children already know in order to introduce new, more sophisticated ones, a strategy sometimes called "anchoring." For example, if the child knows/utters the word "rabbit," the caregiver can build on this knowledge to introduce the more abstract word "animal" that the child may not know yet (e.g., "Yes this is a rabbit, a rabbit is a kind of animal!") (Callanan, 1985) (but see Fourtassi et al., 2020).

Further, when children make mistakes, the caregivers tend to repeat the same utterance while correcting the mistake in it, a strategy known as "reformulation" (Chouinard and Clark, 2003). Caregivers also tend to borrow the children's syntactic structures (e.g., by re-using their verbs and function words), which, in turn, facilitate children's processing of the caregiver's next utterance (Yurovsky et al., 2016).

Future work is needed to examine the relative contribution of these strategies (and others) in explaining the "exaggerated lexical alignment" phenomenon and the potential variability of this relative contribution across cultures. In order to address this question at a large scale (which is crucial for more generalizable results), effort should be devoted to the development of automatic algorithms that characterize the caregivers' scaffolding strategies in naturalistic settings (e.g., Hiller and Fernández, 2016; Jiang et al., 2022; Nikolaus et al., 2021). Such an effort would also have applied implications, especially regarding the design of more effective child-oriented conversational AI for first or second language learning (Huang et al., 2022).

Finally, we return to the issue of cross-linguistic differences in the alignment patterns. While we reported several such differences in the results section, they do not necessarily reflect cultural or linguistic differences. The reason is that the corpora varied widely in terms of their sample size, the number of children involved, whether these children were followed or not in time (longitudinal vs. cross-sectional), as well as the multitude of contexts where the data was collected; these contexts were not necessarily similar across languages, perhaps inducing variability in alignment patterns (Dideriksen et al., 2020).

That said, and if anything, this variability makes our findings about cross-cultural *similarities* (i.e., the main claim of this work) stronger, since these similarities are observed *despite* variability in data sizes, collection procedures, and conversational contexts.

Limitations and future work

We only tested a handful of languages (the ones for which sufficient data was available in CHILDES). However more definitive conclusions would only come from the study of a world representative sample of child-caregiver dialogues, including in non-WEIRD[1] cultures (Henrich et al., 2010; Cristia et al., 2019).

Another limitation is that we focused only on one aspect of alignment (lexical repetition) which provides a partial view of how interlocutors align to each other multimodally in social interaction (Rasenberg et al., 2020). A more comprehensive investigation would require using child-caregiver corpora that facilitate the study of multimodal face-to-face conversations (e.g., Bodur et al., 2021, 2022).

Acknowledgments

The authors of this work have been supported by funding from the Institute of Language Communication and the Brain (ANR-16-CONV-0002) and the MACOMIC project (ANR-21-CE28-0005-01).

References

Kübra Bodur, Mitja Nikolaus, Fatima Kassim, Laurent Prévot, and Abdellah Fourtassi. 2021. Chico: A multimodal corpus for the study of child conversation. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 158–163.

Kübra Bodur, Mitja Nikolaus, Laurent Prévot, and Abdellah Fourtassi. 2022. Backchannel behavior in child-caregiver video calls. In *Proceedings of the*

[1]Western, Educated, Industrialized, Rich, and Democratic.

44th Annual Meeting of the Cognitive Science Society.

Marc H Bornstein, Catherine S Tamis-LeMonda, Joseph Tal, Pamela Ludemann, Sueko Toda, Charles W Rahn, Marie-Germaine Pêcheux, Hiroshi Azuma, and Danya Vardi. 1992. Maternal responsiveness to infants in three societies: The united states, france, and japan. *Child development*, 63(4):808–821.

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482.

Maureen A Callanan. 1985. How parents label objects for young children: The role of input in the acquisition of category hierarchies. *Child Development*, pages 508–523.

Michelle M Chouinard and Eve V Clark. 2003. Adult reformulations of child errors as negative evidence. *Journal of child language*, 30(3):637–669.

Alejandrina Cristia, Emmanuel Dupoux, Michael Gurven, and Jonathan Stieglitz. 2019. Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child development*, 90(3):759–773.

Rick Dale and Michael J Spivey. 2006. Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, 56(3):391–430.

Joseph Denby and Dan Yurovsky. 2019. Parents' linguistic alignment predicts children's language development. In *CogSci*, pages 1627–1632.

Christina Dideriksen, Morten H Christiansen, Kristian Tylén, Mark Dingemanse, and Riccardo Fusaroli. 2020. Quantifying the interplay of conversational devices in building mutual understanding.

Raquel Fernández and Robert Grimm. 2014. Quantifying categorical and conceptual convergence in child-adult dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Koong Hean Foo. 2019. *Intercultural Parenting: How Eastern and Western Parenting Styles Affect Child Development*. Routledge.

Abdellah Fourtassi, Kyra Wilson, and Michael C Frank. 2020. Discovering conceptual hierarchy through explicit and implicit cues in child-directed speech. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Ruthe Foushee, Dan Byrne, Marisa Casillas, and Susan Goldin-Meadow. 2021. Differential impacts of linguistic alignment across caregiver-child dyads and levels of linguistic structure.

Riccardo Fusaroli, Ethan Weed, Deborah Fein, and Letitia Naigles. 2021. Caregiver linguistic alignment to autistic and typically developing children.

Andrew Gelman and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Sarah Hiller and Raquel Fernández. 2016. A data-driven investigation of corrective feedback on subject omission errors in first language acquisition. In *Proceedings of the 20th signll conference on computational natural language learning*, pages 105–114.

Weijiao Huang, Khe Foon Hew, and Luke K Fryer. 2022. Chatbots for language learning—are they really useful? a systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1):237–257.

Hang Jiang, Michael C. Frank, Vivek Kulkarni, and Abdellah Fourtassi. 2022. Exploring patterns of stability and change in caregivers' word usage across early childhood. *Cognitive Science*, 46(7):e13177.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, volume 1. Psychology Press.

Thomas Misiek, Benoit Favre, and Abdellah Fourtassi. 2020. Development of multi-level linguistic alignment in child-adult conversations. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 54–58.

Mitja Nikolaus, Juliette Maes, Jeremy Auguste, Laurent Prevot, and Abdellah Fourtassi. 2021. Large-scale study of speech acts' development using automatic labelling. In *Proceedings of the 43rd annual meeting of the cognitive science society*.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Martin J Pickering and Simon Garrod. 2006. Alignment as the basis for successful communication. *Research on Language and Computation*, 4(2):203–228.

Marlou Rasenberg, Asli Özyürek, and Mark Dingemanse. 2020. Alignment in multimodal interaction: An integrative framework. *Cognitive science*, 44(11):e12911.

Catherine Saint-Georges, Mohamed Chetouani, Raquel Cassel, Fabio Apicella, Ammar Mahdhaoui, Filippo Muratori, Marie-Christine Laznik, and David Cohen. 2013. Motherese in interaction: at the cross-road of emotion and cognition?(a systematic review). *PloS one*, 8(10):e78103.

Alessandro Sanchez, Stephan C Meylan, Mika Braginsky, Kyle E MacDonald, Daniel Yurovsky, and Michael C Frank. 2019. childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior research methods*, 51(4):1928–1941.

Johanna Schick, Caroline Fryns, Franziska Wegdell, Marion Laporte, Klaus Zuberbühler, Carel P van Schaik, Simon W Townsend, and Sabine Stoll. 2022. The function and evolution of child-directed communication. *PLoS Biology*, 20(5):e3001630.

Patrick Shafto, Noah D Goodman, and Thomas L Griffiths. 2014. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, 71:55–89.

Arabella J Sinclair and Raquel Fernández. 2021. Construction coordination in first and second language acquisition. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue*.

Lev Vygotsky. 1978. Interaction between learning and development. *Readings on the development of children*, 23(3):34–41.

Daniel Yurovsky. 2018. A communicative approach to early word learning. *New Ideas in Psychology*, 50:73–79.

Daniel Yurovsky, Gabriel Doyle, and Michael C Frank. 2016. Linguistic input is tuned to children's developmental level. In *CogSci*.