Automatic Coding of Contingency in Child-Caregiver Conversations

Abhishek Agrawal[1], **Mitja Nikolaus**[2], **Benoit Favre**[1], **Abdellah Fourtassi**[1]
[1]Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
[2]CerCo, CNRS, Toulouse, France
{abhishek-amit.agrawal, benoit.favre, abdellah.fourtassi}@univ-amu.fr, mitja.nikolaus@cnrs.fr

Abstract

One of the most important communicative skills children have to learn is to engage in meaningful conversations with people around them. At the heart of this learning lies the mastery of contingency, i.e., the ability to contribute to an ongoing exchange in a relevant fashion (e.g., by staying on topic). Current research on this question relies on the manual annotation of a small sample of children, which limits our ability to draw general conclusions about development. Here, we propose to mitigate the limitations of manual labor by relying on automatic tools for contingency judgment in children's early natural interactions with caregivers. Drawing inspiration from the field of dialogue systems evaluation, we built and compared several automatic classifiers. We found that a Transformer-based pre-trained language model – when fine-tuned on a relatively small set of data we annotated manually (around 3,500 turns) – provided the best predictions. We used this model to automatically annotate, new and large-scale data, almost two orders of magnitude larger than our fine-tuning set. It was able to replicate existing results and generate new data-driven hypotheses. The broad impact of the work is to provide resources that can help the language development community study communicative development at scale, leading to more robust theories.

Keywords: contingency, coherence, child-caregiver conversation, development

1. Introduction

Children's language development involves not only the acquisition of formal structures such as phonology, syntax, and vocabulary but also the learning of how to *use* this formal knowledge to communicate with people around them in day-to-day interactions. Becoming a competent conversational partner requires children to master several skills such as turn-taking (Levinson, 2016; Casillas et al., 2016; Agrawal et al., 2023), active listening (Bavelas et al., 2000; Bodur et al., 2023; Liu et al., 2022), communicative repair (Dingemanse and Enfield, 2024; Clark, 2020; Nikolaus et al., 2022) and interactive alignment (Pickering and Garrod, 2004; Chieng et al., 2024; Fusaroli et al., 2023; Misiek et al., 2020; Misiek and Fourtassi, 2022).

In this paper, we focus on a conversational behavior commonly known in the developmental literature as ***contingency*** (Piaget, 2005; Keenan and Klein, 1975; Bloom et al., 1976; Slomkowski and Dunn, 1996; Hale and Tager-Flusberg, 2005; Melander and Sahlström, 2009; Nadig et al., 2010; Pagmar et al., 2022; Abbot-Smith et al., 2023). It can be defined — broadly speaking — as the collaborative ability to contribute to a dialogue in a relevant fashion, e.g., by connecting with the topic of the ongoing exchange. It is, thus, the glue that makes conversation different from a "succession of disconnected remarks," (Grice, 1975) and "collective monologues" (Piaget, 2005).

Given that contingency is at the heart of the very definition of a conversation; similar concepts have been introduced and studied — beyond the domain of child development — in many scientific fields that deal with dialogue characterization and/or generation such as pragmatics in linguistic theories (e.g., Grice, 1975; Sperber and Wilson, 1986), Conversation Analysis in sociology (e.g., adjacency pairs Schegloff and Sacks, 1973), and dialogue evaluation in human-agent interaction (e.g., Mehri and Eskenazi, 2020).

Cognitive and social impact

Being able to provide contingent conversational turns is believed to be associated with the child's developing cognitive competencies such as Theory of Mind (the ability to infer other people's mental states such as goals, beliefs, and desires) and executive functions such as Inhibitory Control (that is, the ability to inhibit one's impulses vis-à-vis a given stimulus so as to provide a more appropriate response)(see Matthews et al., 2018, for a review). Indeed, learning how to stay on topic requires, amongst other things, the ability to *also* consider the interlocutor's perspective and to inhibit the tendency to *always* talk about one's own interests regardless of what the interlocutor is talking about.

In addition, the mastery of contingency in childhood has important social implications such as the ability to maintain friendships (Hazen and Black, 1989). For instance, peer popularity was found to be negatively correlated with children producing more non-contingent, off-topic comments in conversations with their peers (Place and Becker, 1991). More critically, research such as Garzaniti et al. (2011) and Miczo et al. (2001) suggests that many

LREC-COLING 2024, pages 1856–1870

observed differences between children in terms of conversational skills tend to persist into adulthood, with an impact on their workplace interactions and relationship satisfaction (see Abbot-Smith et al., 2023, for a review).

Towards an automatic annotation of child contingency

Given the connection of conversational contingency with children's broad socio-cognitive development and the persistence of its impact on their later well-being, it is of utmost importance to investigate this phenomenon in its earliest manifestation, i.e., in the context of child-caregiver early *natural* interactions (Pellegrini et al., 2012).

While several corpora of early child-caregiver conversations have been curated (MacWhinney, 2000), a major impediment to the study of contingency is the need for resource-intensive manual annotation. We propose that this impediment can be mitigated through partial or full automation, thanks to recent advances in language and dialogue modeling. Such tools could, in addition, make it possible to study development at a large scale; ideally allowing both an investigation of how current knowledge on the matter – typically based on small-scale studies (e.g., Piaget, 2005; Bloom et al., 1976; Keenan and Klein, 1975) – generalize to a much larger, more diverse sample of children, as well as facilitating the discovery of new insights and hypotheses using bottom-up approaches.

We turn, for inspiration, to the literature on dialogue system evaluation (e.g., evaluating the response relevance of a ChatBot in a free conversation with a human) which has made significant progress, especially since the adoption of pre-trained language models, namely transformer-based models like BERT (Devlin et al., 2019) and GPT2 (Radford et al., 2019). Earlier computational methods tended to be *feature-based*, i.e., extracting several cues and using them as estimators of contingency (as perceived by humans). Such cues included counting repetitions/distribution of certain nouns phrases across turns, the use of speech acts and adjacency pairs, contextual embeddings, and measures of turn similarity (e.g., Barzilay and Lapata, 2008; Cervone et al., 2018; Yi et al., 2019).

More recently, researchers started leveraging pre-trained language models to evaluate the contingency of a turn in the context of the dialogue history. We will call this approach *Language Model-based* (to contrast with the Feature-based approach).

Introducing pre-trained models has allowed researchers to capitalize on rich linguistic knowledge that these models had acquired from data that far exceeds the size of the typical dialogue datasets used to train feature-based methods. This addition resulted in a significant improvement, i.e., a higher similarity with human judgment – compared to previous methods (Sai et al., 2020; Pang et al., 2020; Yeh et al., 2021; Mehri and Eskenazi, 2020; Mehri et al., 2022).

The current study and related work

This work is, to the best of our knowledge, the first attempt to automatize the evaluation of contingency in early child-caregiver natural conversations. This data is different from typical adult conversations used in most above-reviewed work on contingency evaluation (e.g., the SWITCHBOARD corpus: Godfrey et al., 1992). For instance, there is an asymmetry between young children's – rudimentary – language use abilities and the caregiver's mature conversational skills. In addition, the caregiver tends to adapt their language when they talk with children, compared to when they talk with adults. These differences in terms of conversational asymmetry, style, and context call for a dedicated investigation.

In terms of methods, while current research work – with adult data – has largely moved from a Feature-based to a Language Model-based (LM-based) approach, here we study and compare both. Indeed, it is possible that pre-trained language models fail to capture the above-mentioned specifics of child-caregiver interaction, given that these models were pre-trained on data of a very different nature. Conversely, it is possible that child-caregiver dialogues show simpler patterns that can be more adequately captured using a feature-based method. Finally, it is not impossible that neither the feature-based nor the LM-based approach provides a satisfactory account of child-caregiver dialogue contingency if, say, the overall context – which can be crucial for contingency judgment – is not very transparent in the verbal exchange.

For both the Feature-based and LM-based methods, we need a reasonable amount of hand-annotated data from child-caregiver dialogues. This annotated data is necessary for training, fine-tuning, and evaluation. There is – to our knowledge – no publicly available annotation for children's early contingency behavior. Thus, another contribution of this work is to provide such a resource, using a longitudinal corpus of children aged 20 and 32 months old (The New England Corpus, Snow et al., 1996).

The paper is organized as follows. First, we describe how we processed and manually annotated the New England corpus. Next, we describe the various features and models we used to automatically annotate the corpus. Finally, we discuss the results of the automatic annotation and demonstrate the use of these models for a large-scale investigation

of contingency within all of the English-language CHILDES corpora.

Our models, annotations, and code for training the models and running the experiments are all publicly available at https://github.com/abhishek-agrawal94/childes-contingency.

2. Manual Annotation

2.1. Corpus

We annotated contingency behavior in a subset of the New England corpus (Snow et al., 1996). This corpus consists of a longitudinal recording of $N = 52$ children at 14, 20, and then 32 months of age. The context was semi-structured free play between children and their caregivers. The corpus is transcribed and segmented into conversational turns. It is publicly accessible through the CHILDES repository (MacWhinney, 2000) using CHILDES-db R library (Sanchez et al., 2019). We picked this corpus as it covers the age range where children begin developing linguistic and (joint) attention skills that allow them to engage in increasingly extended back-and-forth conversations with the caregiver, thus offering an ideal window to study development from the earliest stages. In addition, the corpus was manually annotated for speech act categories (using the child-adapted INCA-A scheme, Ninio et al., 1994), which we needed for our analyses.

2.2. Data pre-processing

After pilot annotations, it was apparent that verbal data from 14-month-olds was not intelligible enough to enable a precise study of contingency. Thus, our sample included data from children recorded when they were 20 months old and, then, when they were 32 months old.

Starting from the transcripts, we filtered out utterances that weren't intelligible or speech-related, e.g., babbling and other vocalizations. We also filtered out the utterances from the investigator of the study (keeping only utterances from the child or their caregiver). The resulting dataset included a total of 32,343 utterances out of the original size of 81,473 utterances in the New England corpus.

2.3. Procedure

We focused on *turn switches*, i.e., transitions in the conversation when parents or children took a turn following their interlocutor. In other words, if, say, the caregiver made several consecutive utterances, and the child did not intervene (or vice versa), we do not analyze the transition between these consecutive utterances. From a total of 12,981 turn switches across all 85 transcripts that make up the corpus, we annotated – manually – 3,898 turn switches (around 30%), from 28 transcripts that were sampled randomly from the corpus.

The sample can be broken down into 4 equivalent-size conditions as follows: 955 turn-switches of 20-months-olds responding to caregivers, 994 turns of 32-months-olds responding to caregivers, 957 turns of caregivers responding to 20-months-olds, and finally 992 turns of caregivers responding to 32-months-olds.

Two human annotators coded all these turns for contingency on a 3-point scale as **`non-contingent`**, **`contingent`**, and **`ambiguous`**. The annotators made their judgments based on the surrounding verbal context in the dialog. We decided to use the transcripts as our sole source of information for judging contingency as not all the transcripts in CHILDES had accompanying high-quality videos. For turn switches that were not classifiable without information from other modalities, we used the label `ambiguous`. Consider the following example:

> Caregiver: *What's in there?*
> Caregiver: *What do you think they are?*
> Child: *What's this?*
>
> — New England corpus, 32-55.cha

Here, if it can be inferred from the visual modality that the child was pointing/referring to the same thing as the caregiver, then we can consider the child's response to be `contingent`. However, since we are only considering the verbal data, we mark the child's response as `ambiguous`.

Early attempts were used to converge on a common, systematic scheme (see Appendix A.1). Next, both annotators coded all data in batches of approximately 200 turn-switches. After every batch, they adjudicated their disagreements. All original annotations in each batch (i.e., before adjudication) were used to calculate the inter-annotation agreement. The two annotators achieved a weighted Kappa score of $\kappa = 0.728$ (using quadratic weights).

2.4. Results

Figure 1 shows the results of the manual annotation of contingency for children and adults, both grouped (children vs. adults) and broken down by the age of the child (20 and 32 months). When we consider the grouped data (left panel), adults had a higher overall average proportion of contingent utterances (78% of their total turns) compared to children (61% of their total turns). Adults also had lower ambiguous turns (compared to children) and only a very small proportion of non-contingent turns. Children's non-contingent turns represented a minority of their total, but this proportion was still noticeable: 14% of total turns.

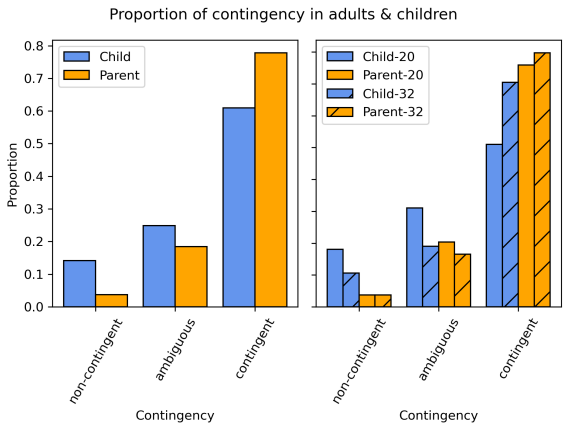

Figure 1: The proportion of `contingent`, `non-contingent`, and `ambiguous` utterances spoken by children and adults in our manually annotated data. The results shown on the right-hand graph are broken down by the age of the child (20 and 32 months).

When we look at the results broken down by the child's age (right panel), we can observe a developmental pattern. First, in terms of children's own responses, the proportion of contingent turns increases from 51% at 20 months to 70% at 32 months. Non-contingency decreased from 18% at 20 months to 11% at 32 months. Second, in terms of caregivers' responses to children, similar findings were observed: Contingency increased from 76% when talking to children at 20 months to 80% when talking to 32-months-old. Ambiguity decreased from 20% at 20 months to 17% at 32 months old (and non-contingent responses remained at floor level).

3. Automatic Annotation

Following recent research on dialogue system evaluation (see Mehri et al., 2022, for an overview), we define the task as labeling the contingency of a turn given a context made of several previous turns in the conversation. We test and compare two different approaches. The first is Feature-based: We extract different verbal features from the dialogue (based on previous research) and evaluate their ability to predict contingency using simple classifiers. The second approach is LM-based: We use pre-trained Language Models and test three levels of fine-tuning on our data (from broad to specific): 1) pre-training only, 2) fine-tuning with self-supervised learning on child-caregiver conversations, and 3) fine-tuning on the supervised task (contingency classification) using manual annotations.

3.1. Feature-based approach

We test the following features:

3.1.1. Speech acts

The speech act categories allow us to infer if, on a high level, the target turn is contingent. For example, we can determine that the category "Yes-no response" is contingent when following a "Yes-no question" and non-contingent when following, say, a "Greeting" (Sacks, 1967; Schegloff and Sacks, 1973; Cervone and Riccardi, 2020; Higashinaka et al., 2014). We use the Inventory of Communicative Acts - Abridged (INCA-A); the most comprehensive coding scheme to date, designed to capture children's emerging speech acts in the context of early interaction with the caregiver (Ninio et al., 1994). INCA-A has 67 different illocutionary categories, which fall into several groups such as directives, declarations, commitments, markings, statements, questions, evaluations, and other vocalizations. The New England corpus, that we use in the current work, was manually annotated for INCA-A by the original authors (Snow et al., 1996).

3.1.2. Noun phrase repetitions

Several previous NLP studies on text coherence or dialogue contingency used repeated named entities across sentences or turns as a feature for contingency prediction (Barzilay and Lapata, 2008; Cervone et al., 2018; Cervone and Riccardi, 2020). The idea is that a turn in which the speaker refers to the same entities as the interlocutor did in a previous turn would be more contingent than one in which the speaker refers to different entities. Given that child-caregiver conversation evolves around simple daily objects or animals instead of the typical entities identified by dedicated NLP tools (e.g., famous people's names and big organizations), we decided to use a broader measure indicating the number of times any noun phrase was repeated across the context and the target turn. To identify the noun phrases, we use the English transformer-based syntactic parser from SpaCy.[1]

3.1.3. Semantic embeddings and similarity

Following Yi et al. (2019), we make use of sentence-level embeddings and cosine similarity as features for contingency prediction. For the embeddings, we used pre-trained Sentence Transformers (Reimers and Gurevych, 2019) to obtain the embedding of the composite {context, turn}. For cosine similarity, we first obtained separate embeddings for context and turn and then computed the cosine similarity

[1] link to model: `https://spacy.io/models/en#en_core_web_trf`

between them. The idea behind using these features is that coherent context-turn pairs would occur closer in the representation space as opposed to non-coherent pairs since they would, for e.g., share similar semantic content.

3.2. Language Model-based approach

Since GPT-2, an auto-regressive transformer language model (Radford et al., 2019), was proven effective in previous research on dialogue evaluation (Pang et al., 2020; Mehri and Eskenazi, 2020), we used it as a starting point to experiment with three levels of fine-tuning on our data. Then, for comparison, we tested another – and more recently introduced – transformer-based model (i.e., DeBERTaV3, He et al., 2023) pre-trained with a different self-supervised objective function (i.e., Replaced Token Detection), compared to GPT-2 (i.e., Next-word prediction based on past context).

3.2.1. GPT-2

GPT-2 is a language model, built of Transformer decoder blocks (no encoder) and pre-trained on WebText: A corpus made of 8 million documents that were linked to in Reddit and received at least three upvotes (to increase the quality of training data) (Radford et al., 2019). We used the version of the model with 124 million parameters[2].

We used this model in three ways, corresponding to the three levels of fine-tuning on our data, ranging from broad to specific, as follows:

a) GPT-2 with pre-training only First, we used the default pre-trained version of the model without any further training on our data. To estimate the contingency, we calculated the perplexity of a turn given the context, quantifying the extent to which this turn naturally follows from the preceding context. This estimation is based on the linguistic knowledge the model has gathered in pre-training.

As GPT-2 is an auto-regressive model (i.e., predicting the next word based on the *past context*), perplexity is well-defined as the exponent of the average of the negative log-likelihood. For a sequence of tokens $X = (x_1, x_2, x_3, \ldots, x_t)$, making up the composite {context, turn}, the perplexity of X is calculated as follows:[3]

$$PPL(X) = \exp\left\{-\frac{1}{t}\sum_{i=0}^{t} log\, p_\theta\left(x_i|x_{<i}\right)\right\}$$

b) GPT-2 with self-supervised fine-tuning The second approach involved fine-tuning a pre-trained GPT-2 model on child-caregiver conversations. We used the same (self-supervised) objective function to fine-tune GPT-2 on all English-language corpora in the CHILDES repository, excluding data from the New England corpus (because it contains our test data). The fine-tuning data consisted of 4,674 transcripts from a total of N=862 children aged 26 months[4] and up to 60 months.

The model was fine-tuned for 3 epochs on the training data. After fine-tuning, we estimated the contingency by computing the perplexity of {context, turn} using the same formula as in the default version of GPT-2 above.

c) GPT-2 with supervised fine-tuning The third approach was to use the pre-trained model and fine-tune it by directly teaching it to classify whether a turn is `contingent`, `non-cotingnent`, or `ambiguous` (given its context) using the manual annotations.

3.2.2. DeBERTaV3

To compare with GPT-2, we use a more recently introduced Transformer called DeBERTaV3 (He et al., 2023); an improved variant of the the DeBERTa model (He et al., 2020), which was, itself, an improved version of BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) transformer models. The most important novelty of DeBERTaV3 is the use of a pre-training objective called Replaced Token Detection (RTD), which proved to be more data-efficient than Mask Language Modeling (MLM) used in DeBERTaV3's predecessors. This model has 304 million parameters and was pre-trained on the English Wikipedia dump, the Book Corpus (Zhu et al., 2015), OPENWEBTEXT which contains reddit content (Gokaslan and Cohen, 2019) and on the STORIES corpus (Trinh and Le, 2019) which is a subset of CommonCrawl.

We fine-tuned DeBERTaV3 on the supervised task of contingency prediction of a turn (given its context) using our manual annotation.

3.3. Task training and Evaluation

All automatic classifications (both feature- and LM-based) were done by training on 80% of our manual annotation and testing on the remaining 20%. The task consists in learning how to associate the pair {context, turn} with one of three labels (`contingent`, `non-contingent`, or `ambiguous`). For each turn – and based on prelimi-

[2]link to model: https://huggingface.co/gpt2

[3]We compute the perplexity for the tokens in the turn only (but conditioned on the entire context).

[4]We did not include younger children to ensure we have a significant proportion of intelligible speech from children.

Classifier	Child		Adult	
	F1 score	MCC score	F1 score	MCC score
Majority classifier	0.46 ± 0.06	0.00 ± 0.00	0.68 ± 0.03	0.00 ± 0.00
Chance classifier	0.38 ± 0.02	0.03 ± 0.04	0.41 ± 0.02	0.00 ± 0.05
Speech acts	0.47 ± 0.06	0.05 ± 0.04	0.68 ± 0.03	-0.01 ± 0.02
Noun phrase reps.	0.51 ± 0.05	0.08 ± 0.04	0.17 ± 0.16	0.00 ± 0.03
Cosine similarity	0.36 ± 0.17	0.05 ± 0.11	0.55 ± 0.03	0.16 ± 0.03
Sentence transformer embedding	0.52 ± 0.07	0.10 ± 0.05	0.68 ± 0.03	0.00 ± 0.04
GPT-2 (no fine-tuning)	0.04 ± 0.02	0.00 ± 0.00	0.28 ± 0.28	0.01 ± 0.01
GPT-2 (self-supervised)	0.53 ± 0.02	0.13 ± 0.06	0.51 ± 0.19	-0.03 ± 0.02
GPT-2 (supervised)	0.62 ± 0.06	0.35 ± 0.06	0.69 ± 0.03	0.22 ± 0.08
DeBERTaV3 (supervised)	0.70 ± 0.03	0.46 ± 0.05	0.76 ± 0.03	0.41 ± 0.04
DeBERTaV3 (supervised)+ optimal context	**0.74 ± 0.01**	**0.53 ± 0.04**	**0.77 ± 0.03**	**0.42 ± 0.06**
Human score	0.82 ± 0.02	0.65 ± 0.03	0.86 ± 0.03	0.58 ± 0.07

Table 1: The mean weighted F1 scores and MCC scores along with the standard deviation across a 5 fold cross-validation for children of all ages and for adults. The results for the feature based models are with a logistic regression classifier.

nary exploration – we fixed the context size for all classifiers to be the five preceding utterances. We evaluate the models with 5-fold cross-validation. Crucially, we decided to split folds using transcripts (entire conversation session) as units instead of turn-switches. The reason is to make sure there were no overlapping passages in training and the test folds regarding the context. Thus, our evaluation method is rather strict and tests the ability of the model to generalize to other conversational sessions.

For the feature-based methods, we used logistic regression classifiers,[5] testing the performance of the features both individually and in combination with each other. As for the LM-based classifiers, we had two cases: Concerning models without fine-tuning or with self-supervised fine-tuning, we used the perplexity value of {context, turn} as a feature in logistic regressions (as we did for feature-based methods). Concerning language models with supervised fine-tuning, we did not need to train further classifiers as these models were trained directly for the classification task.

For each model, we report the F-score and the Matthews Correlation Coefficient (MCC) score. While the F-score remains one of the most popular metrics, it can sometimes show misleadingly inflated results, especially with imbalanced classes as in our case. In contrast, the MCC is more reliable and has been shown to be generally unaffected by the unbalanced data issue (Chicco and Jurman, 2020)

[5]Other classifiers (e.g., random forest) were used but not reported here as their performance did not improve over the simpler logistic regression.

3.4. Results and Analyses

The results of all classifiers are shown in Table 1, together with chance and majority classifiers used as baselines and human inter-annotation agreement as a top-line. The results are broken down for classifiers that were trained/tested either on children's contingency data or on adults' data.[6]

A first inspection of these results confirms that the MCC scores paint a more reliable/interpretable picture than the F-score. For instance, a simple majority classifier for adults' data has a high F1 score of 0.68, but this score only reflects the fact that the overwhelming majority of adults' turns are contingent, and not the accuracy of the classification. In contrast, the MCC score for this same majority classifier for adults is exactly 0, reflecting more faithfully that the classifier has not really learned anything; putting it on par with the performance of the chance classifier. Thus, in the following, we will be analyzing and discussing the results mainly in terms of the MCC scores.

The feature-based classifiers are shown for each feature (tested individually). None of the features managed to surpass a MCC score of 0.16 which is, overall, low. We also trained and tested classifiers with various combinations of features and different classifiers other than the logistic regression (results not shown here, but provided in Appendix A.2), but none of these configurations led to considerable

[6]Training/testing on each age group separately, i.e., 20 and 32 months old led to data-sparsity-related issues, in particular, noisy results with a large variance across folds. These results were not reliable enough to draw clear conclusions. The results are shown in Appendix A.2.

improvement compared to individual scores.

Moving to LM-based classifiers with GPT-2, we can see that the performance increased when GPT-2 was fine-tuned in a self-supervised fashion on CHILDES (compared to GPT-2's original pre-training without any fine-tuning), but this increase was observed only for children's data. The supervised fine-tuning on the manual annotation led to the best results across both children and adults.

When comparing GPT-2 (supervised) to DeBERTaV3 (supervised), we found that DeBERTaV3 improved the results by a fairly large margin. This score was further improved (especially for children) with an optimal context size.[7] Overall, this model learned to classify children's data better than it did for adults' data, echoing a similar difference observed in terms of human agreement scores.

Note, however, that even the best-performing classifier is still lower than the human inter-annotation agreement, suggesting there is still room for improvement.

Effect of training and context size

Using DeBERTaV3 (supervised), we simulated the performance of the classifier when trained on smaller portions of the data. Figure 2 shows that the performance peaks when fed with around 80% of the available training data for both children and adults, indicating that the size of our manual annotation dataset, although relatively small, was sufficient for fine-tuning the language model.

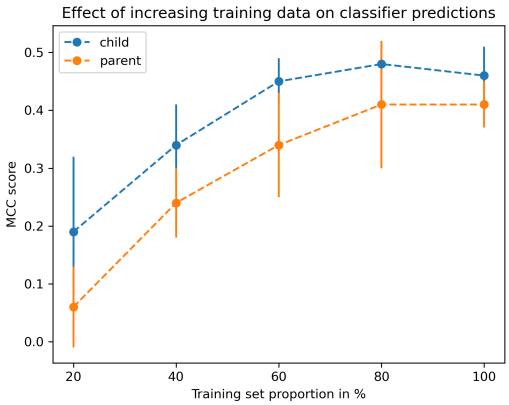

Figure 2: The effect of varying fine-tuning data size (i.e., from the manual annotation data) on the performance of DeBERTaV3 (supervised). The points indicate the mean MCC score across a 5-fold cross-validation and the ranges represent the standard deviation.

[7]The optimal context was 8 preceding turns for children and 2 preceding turns for adults, see Figure 3

Next, we tested how DeBERTaV3 (supervised) performed with different context sizes. The results are shown in Figure 3. We can see that a large improvement occurs by adding only 2 preceding turns as context. For children's data, performance slightly increases; peaking at a context size of around 8 preceding turns. For adults' data, however, adding context beyond 2 preceding turns does not seem to improve performance (if anything, the performance slightly decreases).

Interestingly, performance with no context at all was above zero, suggesting that some turns had intrinsic properties that correlated with their contingency status. Qualitative inspection of a few examples in the 0-context case shows that turns that were successfully classified were often short utterances or backchannels (e.g., 'yeah', 'no', 'mhm', and 'okay').

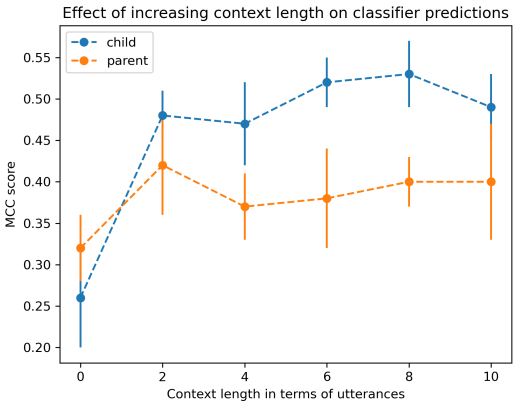

Figure 3: The effect of varying context size on the performance of DeBERTaV3 (supervised). The points indicate the mean MCC score across a 5-fold cross-validation and the ranges represent the standard deviation.

3.5. Toward large-scale investigation

We select the best models from our training and then use them to predict the contingency for turn switches from all English-language CHILDES corpora (excluding the New England corpus) to see how the automatic annotation of the model behaves on new, large-scale data. We test the model's behavior both within and beyond the age range of the training set.

Within-range automatic annotation Since we used manually annotated data from conversations of children aged 20 and 32 months old for our training, we restricted this first exploration to all English-language turn switches in CHILDES corpora belonging to children aged 20 to 32 months

(and their caregivers). Since we did 5-fold cross-validation during DebertaV3's fine-tuning (see Section 3.3), we ended up with 5 different classifiers, one for each fold. We ran all 5 models on this new CHILDES data and did a majority vote to get a final prediction for each {context, turn}. In this manner, we automatically annotated $345,893$ turn-switches for children and $345,133$ turn-switches for caregivers in total, that is, two orders of magnitude larger than the manually annotated training data.

Figure 4 shows the results. First, the automatic annotation captures the broad developmental difference between 20-month-olds (lower contingency) and 32-month-olds (higher contingency). Thus the automatic classifier replicates the same result obtained with manual annotation (shown in Figure 1) using completely different corpora (that have not been seen in fine-tuning), also generalizing it at a large scale. In addition, automatic annotation reveals a new finding: There is a rather *continuous* developmental pattern in children's contingency between 20 and 32 months, although – crucially – no data from children in these intermediate ages were seen in fine-tuning. We can also see a similar (though slower) developmental pattern in parents' contingency, this slight increase appears to be due mostly to a reduction in the number of ambiguous turns, with non-contingent turns remaining largely at floor level (which is also similar to what we obtained with manual annotation).

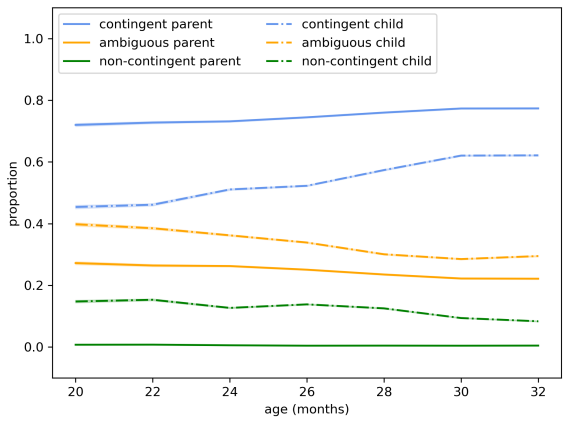

Figure 4: The proportion of contingent, non-contingent, and ambiguous utterances obtained using automatic annotation of new, large-scale data within the age range of the fine-tuning set.

Beyond-range automatic annotation To investigate the extent to which our automatic classifier can be used with data beyond the age range of the fine-tuning set, we now automatically annotate conversations of children aged up to 64 months in all English-language CHILDES, following a similar procedure as above (leading to $911,143$ turn-switches for children and $893,973$ turn-switches for caregivers in total).

Figure 5 reproduces results of Figure 4 in the 20-32 months interval (same data) and shows the automatic annotation beyond this range, up to 64 months. The results show no increase in children's contingent turns beyond 32 to 36 months and no decrease in non-contingent turns either, a finding that is counter-intuitive and most certainly inaccurate. We conclude that the model cannot be used reliably to annotate data beyond the age range seen by the model during fine-tuning.

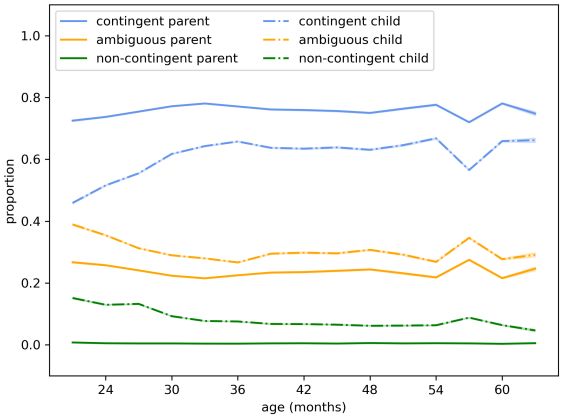

Figure 5: The proportion of contingent, non-contingent, and ambiguous utterances obtained using automatic annotation of new, large-scale data beyond the age range of the fine-tuning set, and up to 64 months.

4. Conclusion

Conversational contingency plays a crucial role in children's communicative and socio-cognitive development. Understanding how this skill develops requires that we study its earliest manifestation in child-caregiver natural interaction, soon after the child becomes able to utter intelligible speech and engage in a verbal back-and-forth with the caregiver. While several studies have investigated contingency behavior around that period and beyond (Piaget, 2005; Keenan and Klein, 1975; Bloom et al., 1976; Abbot-Smith et al., 2023), most are typically based on small-scale samples (with known limitations); due primarily to the fact that the study of this phenomenon in natural interaction requires resource-intensive manual annotations.

Here we explored the possibility of automatizing the process of child-caregiver contingency judgment, with the goal of facilitating more research

into this question, e.g., by testing the generality of our current knowledge at a large scale and by allowing a bottom-up exploration of new hypotheses. We took inspiration from the field of dialogue systems evaluation to build and test various automatic classifiers. The most accurate one was based on a pre-trained language model, which we fine-tuned on a relatively small sample of data that we annotated manually. This classifier was able not only to replicate and generalize findings – obtained with human annotators – on data it had never seen but also to generate a new hypothesis about the shape of the developmental trajectory.

Finally, this work can impact not only research on children's early conversational development but also research on the role of interaction in predicting language learning in the wild. While current methods examine the role of predictors such as children's overall linguistic input (e.g., the quantity of speech heard) or broad interactive measures like the number of turns or temporal contingency (Bergelson et al., 2023; Donnelly and Kidd, 2021; Elmlinger et al., 2023), the current work allows more detailed examination of the verbal content of the interaction and its semantic connectedness, facilitating empirical testing —at scale— of key proposals from interactionist theories and models of language acquisition (Tomasello, 2003; Clark, 2018; Bruner, 1983; Nelson, 2007; Masek et al., 2021; Nikolaus and Fourtassi, 2023, 2021).

5. Limitations

The performance of our best model was still inferior to that of human annotators (although the gap is not huge). How can we improve? The common approach is to annotate more data manually and increase the size of the fine-tuning data. However, as Figure 2 demonstrates, this is unlikely to improve performance as we appear to have already hit a peak. Another – and perhaps more promising way forward – is to use larg*er* language models with a lot more parameters, pre-trained on much more data than say, GPT2 or DeBERTaV3. Up until very recently, such Large Language Models (LLMs) have been closed to researchers with no possibility of fine-tuning their parameters (e.g., GPT4 OpenAI, 2023). This is changing both with the release of more open LLMs (e.g., Llama 2 Touvron et al., 2023, Mistral 7B Jiang et al., 2023) and with improvements in machine learning techniques that allow fine-tuning of LLMs with reasonable computation resources (e.g., Houlsby et al., 2019; Dettmers et al., 2022).

Another limitation of our study is that – for practical reasons – we considered only the transcript of the conversation for our manual annotations and for training/evaluating our models. Nevertheless, the visual context can be very informative for contingency judgment in early childhood, especially in evaluating referring expressions. Adding the visual context would help resolve several instances of what we labeled as "ambiguous." This, however, will depend on the availability of curated multimodal corpora (which are rare, given the concern to protect the anonymity of children) as well as on the ability of models to learn reliably from real-life, naturalistic, and noisy multimodal scenes (which is still an open research question).

6. Acknowledgements

This work, carried out within the Institute of Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government (France 2030), managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX). Furthermore, this study was also supported by the ANR MACOMIC (ANR-21-CE28-0005-01) grant. This work was performed using HPC resources from GENCI–IDRIS (Grant 2022-AD011013886).

7. Bibliographical References

Kirsten Abbot-Smith, Julie Dockrell, Alexandra Sturrock, Danielle Matthews, and Charlotte Wilson. 2023. Topic maintenance in social conversation: What children need to learn and evidence this can be taught. *First Language*.

Abhishek Agrawal, Jing Liu, Kübra Bodur, Benoit Favre, and Abdellah Fourtassi. 2023. Development of multimodal turn coordination in conversations: Evidence for adult-like behavior in middle childhood. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Regina Barzilay and Mirella Lapata. 2008. Modeling Local Coherence: An Entity-Based Approach. *Computational Linguistics*, 34(1):1–34.

Janet B. Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952. Place: US Publisher: American Psychological Association.

Elika Bergelson, Melanie Soderstrom, Iris-Corinna Schwarz, Caroline F. Rowland, Nairán Ramírez-Esparza, Lisa R. Hamrick, Ellen Marklund, Marina Kalashnikova, Ava Guez, Marisa Casillas, Lucia Benetti, Petra van Alphen, and Alejandrina

Cristia. 2023. Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences*, 120(52).

Lois Bloom, Lorraine Rocissano, and Lois Hood. 1976. Adult-child discourse: Developmental interaction between information processing and linguistic knowledge. *Cognitive Psychology*, 8(4):521–552.

Kübra Bodur, Mitja Nikolaus, Laurent Prévot, and Abdellah Fourtassi. 2023. Using video calls to study children's conversational development: The case of backchannel signaling. *Frontiers in Computer Science*, 5.

J.S. Bruner. 1983. *Child's Talk: Learning to Use Language*. W.W. Norton.

Marisa Casillas, Susan C Bobb, and Eve V Clark. 2016. Turn-taking, timing, and planning in early language acquisition. *J. Child Lang.*, 43(6):1310–1337.

Alessandra Cervone and Giuseppe Riccardi. 2020. Is this Dialogue Coherent? Learning from Dialogue Acts and Entities. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 162–174, 1st virtual meeting. Association for Computational Linguistics.

Alessandra Cervone, Evgeny Stepanov, and Giuseppe Riccardi. 2018. Coherence Models for Dialogue. In *Interspeech 2018*, pages 1011–1015. ISCA.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6.

Adriana Chee Jing Chieng, Camille J. Wynn, Tze Peng Wong, Tyson S Barrett, and Stephanie A. Borrie. 2024. Lexical alignment is pervasive across contexts in non-weird adult–child interactions. *Cognitive Science*, 48(3):e13417.

Eve V. Clark. 2018. Conversation and Language Acquisition: A Pragmatic Approach. *Language Learning and Development*, 14(3):170–185.

Eve V. Clark. 2020. Conversational repair and the acquisition of language. *Discourse Processes*, 57(5-6):441–459.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mark Dingemanse and N J Enfield. 2024. Interactive repair and the foundations of language. *Trends Cogn. Sci.*, 28(1):30–42.

Seamus Donnelly and Evan Kidd. 2021. The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. *Child Development*, 92(2):609–625.

Steven L. Elmlinger, Michael H. Goldstein, and Marisa Casillas. 2023. Immature vocalizations simplify the speech of tseltal mayan and u.s. caregivers. *Topics in Cognitive Science*, 15(2):315–328.

Riccardo Fusaroli, Ethan Weed, Roberta Rocca, Deborah Fein, and Letitia Naigles. 2023. Caregiver linguistic alignment to autistic and typically developing children: A natural language processing approach illuminates the interactive components of language development. *Cognition*, 236:105422.

Ivana Garzaniti, Glenn Pearce, and John Stanton. 2011. Building friendships and relationships: The role of conversation in hairdressing service encounters. *Managing Service Quality: An International Journal*, 21(6):667–687. Publisher: Emerald Group Publishing Limited.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Courtney M. Hale and Helen Tager-Flusberg. 2005. Social communication in children with autism: The relationship between theory of mind and discourse development. *Autism*, 9(2):157–178. Publisher: SAGE Publications Ltd.

Nancy L. Hazen and Betty Black. 1989. Preschool peer communication skills: The role of social status and intervention context. *Child Development*, 60(4):867–876. Place: United Kingdom Publisher: Blackwell Publishing.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv*, abs/2006.03654.

Ryuichiro Higashinaka, Toyomi Meguro, Kenji Imamura, Hiroaki Sugiyama, Toshiro Makino, and Yoshihiro Matsuo. 2014. Evaluating coherence in open domain conversational systems. In *Interspeech 2014*, pages 130–134. ISCA.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Elinor Ochs Keenan and Ewan Klein. 1975. Coherency in children's discourse. *Journal of Psycholinguistic Research*, 4(4):365–380.

Stephen C. Levinson. 2016. Turn-taking in Human Communication – Origins and Implications for Language Processing. *Trends in Cognitive Sciences*, 20(1):6–14.

Jing Liu, Mitja Nikolaus, Kübra Bodur, and Abdellah Fourtassi. 2022. Predicting backchannel signaling in child-caregiver multimodal conversations. In *Companion Publication of the 2022 International Conference on Multimodal Interaction*, ICMI '22 Companion, page 196–200.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Lillian R Masek, Brianna TM McMillan, Sarah J Paterson, Catherine S Tamis-LeMonda, Roberta Michnick Golinkoff, and Kathy Hirsh-Pasek. 2021. Where language meets attention: How contingent interactions promote learning. *Developmental Review*, 60:100961.

Danielle Matthews, Hannah Biney, and Kirsten Abbot-Smith. 2018. Individual differences in children's pragmatic ability: A review of associations with formal language, social cognition, and executive functions. *Language Learning and Development*, 14(3):186–223.

Shikib Mehri, Jinho Choi, Luis Fernando D'Haro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, et al. 2022. Report from the nsf future directions workshop on automatic evaluation of dialog: Research directions and challenges. *arXiv preprint arXiv:2203.10012*.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Helen Melander and Fritjof Sahlström. 2009. In tow of the blue whale: Learning as interactional changes in topical orientation. *Journal of Pragmatics*, 41(8):1519–1537.

Nathan Miczo, Chris Segrin, and Lisa E. Allspach. 2001. Relationship between nonverbal sensitivity, encoding, and relational satisfaction. *Communication Reports*, 14(1):39–48. Publisher: Routledge _eprint: https://doi.org/10.1080/08934210109367735.

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, Online. Association for Computational Linguistics.

Thomas Misiek and Abdellah Fourtassi. 2022. Caregivers exaggerate their lexical alignment to young children across several cultures. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*.

Aparna Nadig, Iris Lee, Leher Singh, Kyle Bosshart, and Sally Ozonoff. 2010. How does the topic of conversation affect verbal exchange and eye gaze? A comparison between typical development and high-functioning autism. *Neuropsychologia*, 48(9):2730–2739.

Katherine Nelson. 2007. *Young minds in social worlds: Experience, meaning, and memory*. Harvard University Press.

Mitja Nikolaus and Abdellah Fourtassi. 2021. Modeling the interaction between perception-based and production-based learning in children's early acquisition of semantic knowledge. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, Online. Association for Computational Linguistics.

Mitja Nikolaus and Abdellah Fourtassi. 2023. Communicative feedback in language acquisition. *New Ideas in Psychology*, 68:100985.

Mitja Nikolaus, Eliot Maes, Jeremy Auguste, Laurent Prévot, and Abdellah Fourtassi. 2022. Large-scale study of speech acts' development in early childhood. *Language Development Research*, 2(1):268–304.

Anat Ninio, Catherine E. Snow, Barbara A. Pan, and Pamela R. Rollins. 1994. Classifying communicative acts in children's interactions. *Journal of Communication Disorders*, 27(2):157–187.

OpenAI. 2023. Gpt-4 technical report.

David Pagmar, Kirsten Abbot-Smith, and Danielle Matthews. 2022. Predictors of children's conversational contingency. *Language Development Research*, 2(1). Number: 1 Publisher: Carnegie Mellon University Library Publishing Service.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards Holistic and Automatic Evaluation of Open-Domain Dialogue Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.

Anthony D Pellegrini, Frank Symons, and John Hoch. 2012. *Observing children in their natural worlds: A methodological primer*. Psychology Press.

Jean Piaget. 2005. *Language and Thought of the Child: Selected Works vol 5*. Routledge.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–226. Place: United Kingdom Publisher: Cambridge University Press.

Karen S. Place and Judith A. Becker. 1991. The influence of pragmatic competence on the likeability of grade-school children. *Discourse Processes*, 14(2):227–241. Publisher: Routledge _eprint: https://doi.org/10.1080/01638539109544783.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Harvey Sacks. 1967. Transcribed lectures. *March 9th, University of California, Irvine*.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving Dialog Evaluation with a Multi-reference Adversarial Dataset and Large Scale Pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827. Place: Cambridge, MA Publisher: MIT Press.

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*, 51:1928–1941.

Emanuel A. Schegloff and Harvey Sacks. 1973. Opening up Closings. 8(4):289–327. Publisher: De Gruyter Mouton Section: Semiotica.

Cheryl Slomkowski and Judy Dunn. 1996. Young children's understanding of other people's beliefs and feelings and their connected communication with friends. *Developmental Psychology*, 32(3):442–447. Place: US Publisher: American Psychological Association.

Catherine E. Snow, Barbara Alexander Pan, Alison Imbens-Bailey, and Jane Herman. 1996. Learning How to Say What One Means: A Longitudinal Study of Children's Speech Act Use*. *Social Development*, 5(1):56–84. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9507.1996.tb00072.x.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Citeseer.

Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava,

Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Trieu H. Trinh and Quoc V. Le. 2019. A Simple Method for Commonsense Reasoning. ArXiv:1806.02847 [cs].

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A Comprehensive Assessment of Dialog Evaluation Metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Sanghyun Yi, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2019. Towards Coherent and Engaging Spoken Dialog Response Generation Using Automatic Conversation Evaluators. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 65–75, Tokyo, Japan. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27, Santiago, Chile. IEEE.

8. Language Resource References

MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed*. Lawrence Erlbaum Associates Publishers, The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed. PID https://childes.talkbank.org/. Pages: xi, 366.

Snow, Catherine E. and Pan, Barbara Alexander and Imbens-Bailey, Alison and Herman, Jane. 1996. *Learning How to Say What One Means: A Longitudinal Study of Children's Speech Act Use**. PID https://childes.talkbank.org/access/Eng-NA/NewEngland.html. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9507.1996.tb00072.x.

A. Appendix

A.1. Annotation Scheme

We develop an annotation scheme to annotate contingency for the New England corpus which is one of several corpora of child-caregiver interactions available in the CHILDES databank. We annotate utterances only at the turn switch level between the child and the caregiver i.e., when the role of the speaker in the conversation changes from child to caregiver or vice versa. We don't consider a fixed past context length in terms of the number of utterances that we consider while annotating a target utterance.

We consider topic shifts on a case by case basis. Generally, we consider minor topic shifts to be ambiguous in nature. For instance, while reading the animal picture book, if the caregiver keeps asking "what is this?" we annotate it as ambiguous. Any smooth topic shifts which fall in line with something from the recent past context is annotated as contingent. Consider the below example:

Caregiver: *what is it?*
Caregiver: *a book!*
Child: *yeah.*
Caregiver: *oh you want me to read it?*

In the above example, we consider the turn switch from child to caregiver as contingent since the topic shifts smoothly from the book to reading the book. An example of a non-contingent topic shift is shown below:

Caregiver: *what are you going to do now?*
Child: *going to do.*
Caregiver: *what's that?*
Caregiver: *is that a block?*

In the above example, the turn switch from the child to the caregiver is non-contingent since it is an abrupt change of topic.

If a turn switch can be considered contingent on the assumption that the person is pointing/gesturing to something then we mark it as ambiguous (since we rely only on the transcripts and not visual data for our annotations). If an utterance is a repetition of the previous utterance then we consider it as contingent as it can be a confirmation or acknowledgment of the previous utterance. However, if the interlocutor/s keep repeating an utterance redundantly then we annotate it as ambiguous since we cannot be sure of the intention behind this repetition. Consider the below example:

Caregiver: *this is a no-no.*
Child: *no-no.*
Caregiver: *no-no.*
Child: *no-no.*
Caregiver: *no-no.*

In the above example, we consider the first repetition done by the child to be contingent but all the other repetitions we mark as ambiguous since they are redundant.

We consider all clarification requests to be contingent. If there are two back to back utterances from the same interlocutor where the second utterance can be considered as a continuation of the first utterance and the turn switch is contingent with the second utterance then we mark the turn switch as contingent. Consider the below example:

Caregiver: *what's this?*
Caregiver: *what's in this box?*
Child: *oh.*
Child: *oh this.*

In the above example, we annotate the turn switch as contingent since the second utterance by the child indicates that the child has an idea of what could be in the box.

We annotate any random or off topic responses to questions as non-contingent. If the response to a question is another question then we mark it as non-contingent unless the question in the response is a clarification request. If we are unsure whether the response question is a clarification request then we mark it as ambiguous.

We consider backchannels (short verbal utterances like "mhm", "mm", "uh-huh", "oh", etc.) on a case by case basis. We never treat a backchannel as non-contingent. If there is any doubt concerning the contingent nature of a backchannel response, then we annotate it as ambiguous. Consider the example below:

Caregiver: *is that a cow?*
Child: *mhm.*
Caregiver: *mm.*
Caregiver: *and a baby donkey on the farm.*

In the above example, we consider the child turn switch as contingent since the child is responding to the question. However, consider the following example:

Caregiver: *what's that?*
Caregiver: *wanna sit down and read the book?*
Child: *oh.*
Caregiver: *come here.*

This was marked as ambiguous because one cannot be sure – based on the transcript alone – what the child is trying to express.

Classifier	Child		Adult	
	F1 score	MCC score	F1 score	MCC score
Majority classifier	0.35 ± 0.07	0.00 ± 0.00	0.66 ± 0.06	0.00 ± 0.00
Chance classifier	0.34 ± 0.05	-0.01 ± 0.07	0.40 ± 0.04	-0.01 ± 0.03
Speech acts (SA)	0.35 ± 0.07	0.03 ± 0.04	0.66 ± 0.06	0.00 ± 0.00
Noun phrase reps. (NP)	0.44 ± 0.05	0.12 ± 0.09	0.25 ± 0.25	0.00 ± 0.03
Cosine similarity (CS)	0.28 ± 0.04	0.12 ± 0.05	0.55 ± 0.05	0.17 ± 0.03
NP + CS	0.37 ± 0.07	0.07 ± 0.03	0.43 ± 0.09	0.10 ± 0.05
GPT-2 (no fine-tuning)	0.06 ± 0.04	0.00 ± 0.00	0.31 ± 0.26	-0.01 ± 0.01
GPT-2 (self-supervised, PPL)	0.46 ± 0.08	0.14 ± 0.09	0.55 ± 0.15	-0.03 ± 0.04
PPL + NP	0.49 ± 0.05	0.18 ± 0.09	0.20 ± 0.20	0.00 ± 0.02
PPL + CS	0.51 ± 0.03	0.20 ± 0.05	0.53 ± 0.05	0.14 ± 0.05
PPL + NP + CS	0.52 ± 0.05	0.21 ± 0.07	0.44 ± 0.10	0.09 ± 0.06
DeBERTaV3 (default)	0.19 ± 0.14	0.01 ± 0.01	0.26 ± 0.29	0.03 ± 0.04
DeBERTaV3 (supervised)	**0.64 ± 0.07**	**0.41 ± 0.08**	**0.33 ± 0.05**	**0.73 ± 0.06**

Table 2: The mean weighted F1 scores and MCC scores along with the standard deviation across a 5 fold cross-validation for children aged 20 months and for adults conversing with 20 months old children. The results for the feature based models are with a logistic regression classifier.

Classifier	Child		Adult	
	F1 score	MCC score	F1 score	MCC score
Majority classifier	0.58 ± 0.06	0.00 ± 0.00	0.71 ± 0.02	0.00 ± 0.00
Chance classifier	0.38 ± 0.03	-0.02 ± 0.04	0.41 ± 0.02	0.01 ± 0.02
Speech acts (SA)	0.58 ± 0.06	0.00 ± 0.00	0.71 ± 0.02	-0.01 ± 0.01
Noun phrase reps. (NP)	0.57 ± 0.05	0.05 ± 0.03	0.41 ± 0.05	0.05 ± 0.03
Cosine similarity (CS)	0.51 ± 0.08	0.05 ± 0.09	0.55 ± 0.02	0.14 ± 0.04
NP + CS	0.55 ± 0.08	0.04 ± 0.08	0.50 ± 0.05	0.10 ± 0.03
GPT-2 (no fine-tuning)	0.02 ± 0.01	0.00 ± 0.00	0.28 ± 0.32	0.01 ± 0.01
GPT-2 (self-supervised, PPL)	0.35 ± 0.21	0.08 ± 0.08	0.45 ± 0.19	-0.03 ± 0.04
PPL + NP	0.58 ± 0.05	0.08 ± 0.04	0.42 ± 0.05	0.04 ± 0.04
PPL + CS	0.43 ± 0.15	0.01 ± 0.08	0.55 ± 0.03	0.14 ± 0.05
PPL + NP + CS	0.56 ± 0.07	0.07 ± 0.06	0.51 ± 0.05	0.10 ± 0.04
DeBERTaV3 (default)	0.26 ± 0.25	0.02 ± 0.02	0.31 ± 0.34	0.01 ± 0.02
DeBERTaV3 (supervised)	**0.59 ± 0.14**	**0.24 ± 0.14**	**0.73 ± 0.04**	**0.24 ± 0.11**

Table 3: The mean weighted F1 scores and MCC scores along with the standard deviation across a 5 fold cross-validation for children aged 32 months and for adults conversing with 32 months old children. The results for the feature based models are with a logistic regression classifier.

A.2. Classifier Results Segregated by Age of Child

We also trained separate models for data segregated by the age of the child. Table 2 displays the results for the models trained on data from the 20 months old children. The classifier used in the feature-based methods and for the baselines was the logistic regression classifier. In instances where we compute the perplexity with the GPT-2 model, we then further fit a logistic regression classifier to predict the contingency label from the perplexity values. As you can see in the table, the feature-based models perform quite poorly while the best model is the supervised DeBERTaV3 model for both children and adults.

Table 3 displays the results for the models trained on data from the 32 months old children. Once again, the feature-based models perform quite poorly while the best model is the supervised DeBERTaV3 model for both children and adults.