Automatic Annotation of Grammaticality in Child-Caregiver Conversations

Mitja Nikolaus[1], **Abhishek Agrawal**[2], **Petros Kaklamanis**[3], **Alex Warstadt**[4], **Abdellah Fourtassi**[2]
[1]CerCo, CNRS, Toulouse, France
[2]Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
[3]University of Pennsylvania
[4]ETH Zürich
mitja.nikolaus@cnrs.fr

Abstract

The acquisition of grammar has been a central question to adjudicate between theories of language acquisition. In order to conduct faster, more reproducible, and larger-scale corpus studies on grammaticality in child-caregiver conversations, tools for automatic annotation can offer an effective alternative to tedious manual annotation. We propose a coding scheme for context-dependent grammaticality in child-caregiver conversations and annotate more than 4,000 utterances from a large corpus of transcribed conversations. Based on these annotations, we train and evaluate a range of NLP models. Our results show that fine-tuned Transformer-based models perform best, achieving human inter-annotation agreement levels. As a first application and sanity check of this tool, we use the trained models to annotate a corpus almost two orders of magnitude larger than the manually annotated data and verify that children's grammaticality shows a steady increase with age. This work contributes to the growing literature on applying state-of-the-art NLP methods to help study child language acquisition at scale.

Keywords: language acquisition, grammaticality, acceptability, conversation

1. Introduction

The acquisition of grammar has historically been a central point regarding discussions on the learnability of language from limited input (Chomsky, 1957; Gold, 1967; Harris, 1993; Brown, 1973; Piantadosi, 2023). Traditionally, observational studies on the acquisition of grammar have relied on manual annotations of early child talk. In some cases, notably the question of presence and effectiveness of caregiver corrections following a child's grammatical mistake, research has led to mixed (if not conflicting) results(Brown and Hanlon, 1970; Nelson et al., 1973; Demetras et al., 1986; Marcus, 1993; Morgan et al., 1995; Saxton, 2000; Chouinard and Clark, 2003). The lack of consensus can be attributed, at least partly, to the limited sample size used in each study.

In the current work, we introduce automatic coding as a way forward to address this issue and to help researchers achieve more conclusive results. First, we develop a general coding scheme for the annotation of grammaticality in child-caregiver conversations. Then, we annotate a sample of such conversations to train and evaluate models for automatic annotation, which we use to annotate a large-scale corpus, almost two orders of magnitude larger than the size of the data we coded manually. The developed tools can help researchers perform more cumulative and larger-scale analyses on the development of grammaticality in early childhood and even help adjudicate between general theories of language acquisition (Tomasello, 2003; Clark, 2016).

Our approach differs from typical work on modeling grammaticality using NLP tools, including for research that deals with the linguistic production of adult speakers. While a large portion of this research has dealt with grammaticality (or *acceptability*) of sentences in isolation (Lau et al., 2017; Warstadt et al., 2020, 2019; Huebner et al., 2021), here we study the *grammaticality of utterances in conversations*. This covers a differently distributed set of grammatical phenomena (e.g. high proportion of omission errors), and, more importantly, the utterances are often elliptical, i.e., their interpretation depends on the conversational context.

Contributions of this work This work makes several contributions. First, we propose a new coding scheme for the annotation of grammaticality in child-caregiver conversations, based on which we annotate more than 4,000 utterances from English CHILDES (MacWhinney, Brian, 2000). Additionally, we annotate the specific error category for each ungrammatical utterance.

Based on this data, we train state-of-the-art NLP models to automatically annotate the grammaticality of utterances and find that the performance of the best models is almost on par with human inter-annotation agreement scores.

Finally, we use the trained models to annotate all transcripts from English-language CHILDES of

[1]Work performed while at Aix-Marseille University.

LREC-COLING 2024, pages 1832–1844
20-25 May, 2024.

children aged 2 to 5 years, which allows us to characterize the developmental trajectory based on this large and diverse corpus.

Our models and annotations, as well as the code for all experiments described in the paper are publicly available at https://github.com/mitjanikolaus/childes-grammaticality.

2. Related Work

2.1. Automatic Annotation of Grammaticality

Supervised approaches for the automatic annotation of grammaticality have often relied on data produced by linguists, e.g., example sentences scraped from linguistics publications (Warstadt et al., 2019; Trotta et al., 2021; Mikhailov et al., 2022; Someya et al., 2023), often including textbooks (e.g. Adger, 2003; Kim and Sells, 2008; Sportiche et al., 2013). Using such datasets, early modeling approaches relied on techniques such as n-grams and recurrent neural networks (Wagner et al., 2009; Lawrence et al., 2000). Notably, Lau et al. (2017) additionally controlled for confounding factors such as sentence length and lexical frequency to obtain a better classification performance. More recently, the use of large language models pre-trained on large text corpora has enabled substantial performance improvements as measured by comprehensive evaluation benchmarks (Warstadt et al., 2019, 2020), with the best Transformer-based models achieving scores that are comparable to human inter-annotation agreement (e.g. He et al., 2022).

Here, we examine whether this progress in the study of isolated sentences can be extended to children's talk in a conversational context, requiring the models not only to adapt to children's data but also to take into account the conversational *context* to evaluate the grammaticality of a given utterance.

2.2. Automatic Annotation of Children's Grammaticality in Conversation

Research on automatic annotation of children's productive language in naturalistic conversation has not always focused on grammaticality per se, but instead on other – more readily automatized measures – such as Mean Length of Utterances (MLU; Brown, 1973).

For the specific measurement of grammatical development, Scarborough (1990) proposed the Index of Productive Syntax (IPSyn), in which children are evaluated on how many different syntactic and morphological structures they are correctly producing. Calculating the IPSyn requires the manual scanning of a sample of 100 transcribed utterances for the presence of 56 syntactic and morphological forms. Sagae et al. (2005) proposed a method to speed up the calculation of IPSyn scores using tools for automatic annotation: The output of a statistical dependency parser was used to narrow down the set of sentences where certain structures may be found by manual annotators. While such (semi-)automatic methods can provide us with a general estimate of the linguistic productivity of a child, they do not allow for detailed analyses of grammatical phenomena in a conversational context, or per-utterance analyses.

More recently, Hiller and Fernandez (2016) focused on the specific case of subject omissions using automated annotation. Based on a small set of hand-annotated data, they trained a Support Vector Machine (SVM) to detect subject omissions. They applied it to perform analyses on a substantially larger set of data. In contrast to this previous work, here we developed a more general characterization of the grammaticality of children's utterances in conversation, including subject omissions but also a dozen more error categories. Our models can be used to obtain a general measure of the grammaticality of utterances as well as for calculating the overall grammatical competence of a child. They can also be used as a starting point to investigate various mechanisms of language learning, such as corrective feedback (Brown and Hanlon, 1970) and communicative feedback (Nikolaus and Fourtassi, 2023).

3. Manual Annotation

3.1. Annotation Scheme

3.1.1. Grammaticality of Children's Utterances in Conversation

We develop an annotation scheme adapted for the study of grammaticality of children's utterances in English-language child-caregiver conversations. Based on transcripts of conversations, each child utterance that consists of at least two words is classified as either **`grammatical`**, **`ambiguous`**, or **`ungrammatical`**.[1] Utterances are annotated as **`ungrammatical`** if they contain at least one grammatical error.

The grammaticality of each utterance is judged not only based on the utterance itself, but also on the broader context of the conversation. Many utterances in child-caregiver conversations are non-sentential utterances, with highly context-

[1]We exclude all utterances that are unintelligible or not speech-related, such as babbling and other vocalizations like crying or laughing.

Label	Cases	Examples
ungrammatical	Ellipses with missing verb or determiner	“Cookie Monster.”, “Lunch.”, “No shoes.”
	Ellipses with missing object	”I want.”, “He gave.”
	Ellipses with missing subject, if the context (or verb) clearly points to non-imperative use	“Want to go to the cinema!”
	SVO/SV questions (except if they are used as clarification requests or to express surprise)	“You are coming (to the house)?”
	Ellipses due to the child being interrupted[a]	“I gave.” - “That’s great!”
ambiguous	Onomatopoeia	“Miaow miaow”, “Muuh muh!”, “Vroom vroom”
	Unknown words or vocalizations, baby language, family-internal expression, words spontaneously invented by the child	“Let’s go to the cagriotafer!”, “eh eh.”
	Noun phrases that might be grammatical if accompanied with an appropriate gesture, e.g. a pointing gesture towards an object	”A zebra!”[b], “For the zebra”
	Ellipses with missing subject, for which the context does not clearly discriminate between imperative and declarative use	“Do this.”
	Utterances that are grammatically correct, but not aligned with what the child actually intended to say	“That’s a nice cup.” - “Is it you?”[c], “Hide the table!”[d]
	Reciting, Singing, Counting	“Sweep, sweep, sweep!”, “One, two, three.”
	Utterances that are strictly ungrammatical but very commonly used in spoken Standard American or British English	“Don’t know”, ”You like this?”, ”That all?”
	Transcription errors	He like’s animals.
grammatical	Utterances with missing subject, if they are clearly used as imperatives	Look for it!”, “Take this.”
	Utterances with self-repetitions, disfluencies	“I like I like this.”, “This is uhm a table.”
	Self-corrections/Reformulations (if the final reformulated utterance is grammatical)[e]	“He want she wants a flower!”, “She is she was very happy”
	Exclamations, backchannels	“Uh oh.”, “Mhm hm.”, “All right.”, “Oh no!”
	Self-repetitions over multiple utterances (both utterances should be marked as grammatical)	“I want an apple.” - “An apple.”
	Repetitions from the previous utterance (except if the child is repeating an erroneous part from a previous utterance)[f]	“It is very hot.” - “Very hot.” “This is my funny hat” - “My funny hat.”
	Ellipses that are valid responses to a question[g]	“Who is that?” - “Cookie Monster!”, “What’s this?” - “The pasta that dad made!”, “Are you an artist?” - “I am.”
	Greetings, calls for attention	”Good Morning.”, ”Mummy, mum!”
	Wrong answers, utterances that are logically/semantically wrong or questionable	“Can you say ‘a rat’?” - “A cat!” “The sky is green.”
	Completions of previous utterances	“And then he went” - “To the cinema”
	SVO/SV questions that function as clarification requests or express surprise	“This is big.” - “This is big?”
	Short forms/ contractions commonly used in spoken English	“Cause I went to school”, “You’re sposta go there.”, “That’s a lotta dogs!”, “Gotcha!”
	Utterances with phonological errors (either because of dialect or pronunciation difficulties of the child)	“Sesame Stweet”, “Dis is a dog”, “Let’s go srough this once again.”
	Ellipses that are clearly accompanied by pointing	“Oh this!”, “This one.”, “These cats.”

Table 1: Annotation guidelines with example cases for each label.

[a] As we do not have access to the timing of the utterances, we do not know whether the child was actually interrupted or they just stopped the utterance before completing it. To be consistent, we mark these cases as ungrammatical.

[b] If the determiner is missing (“zebra!”), the utterance should be marked as `ungrammatical`.

[c] In this case, the child’s response is grammatical but they most likely intended to say something like “Is it yours?”.

[d] In this case, the child most probably meant to say “Hide *under* the table”. “Hide the table” is strictly speaking grammatical, but we know that there’s actually a grammatical error (missing preposition) if we can infer from the context what the child actually intended to say.

[e] In the case of reformulations across multiple utterances: “He want” - “She wants a flower” the first utterance should be marked as ungrammatical, the second one as grammatical.

[f] Repetitions that are e.g. missing a determiner should be marked as ungrammatical “I like the book” - “book.”

[g] Usually, questions that ask for a noun (phrase) still require the response to have an appropriate determiner (“What is this?” - “A cat.”). If they are missing the determiner, they should be marked as ungrammatical. However, in case the question directly asks for a concept, a response without a determiner is permitted: “How do we call this?” - “Cat!”.

dependent meanings (Fernández et al., 2007). Consider the following dialog:

> Caregiver: *Here take your coat off.*
> Caregiver: *Where do you wanna put your coat?*
> Child: *On the table.*
>
> — MacWhinney corpus, 030526a.cha

In this example, when judging the grammaticality of the child utterance in isolation, one could be annotating it as `ungrammatical` as it is missing subject and verb. However, within the context of the preceding utterance, it should instead be marked as `grammatical`, as it is a valid response to the preceding question.

In contrast, in the following example, the child's utterance is indeed `ungrammatical` (missing subject and verb), even when taking into account the conversational context:

> Caregiver: *You can play with them on the table.*
> Child: *Lots lots in here.*
>
> — Thomas corpus, 020924.cha

For each utterance, the annotators are instructed to take into account the preceding context of the conversation for judging its grammaticality.[2]

The label `ambiguous` is introduced to cover cases in which the grammaticality depends on context that is impossible to infer from the transcript alone (e.g. information about the visual context) as well as cases in which the concept of grammaticality is not applicable.[3] For example, the utterance *"do this."* could be grammatical as an imperative. It could also be a case of a subject omission error if the child actually intended to say *"I do this"*. In some cases, but not always, it is possible to infer the intended meaning from the context of the conversation (e.g. if the preceding utterance is *"Who does this?"*, it is most likely a case of subject omission).

Utterances that only consist of a noun phrase are annotated as `ungrammatical` (as they are missing a finite verb), except if they function as responses to questions ("What is this?" - "An apple."). Another exception is the case of an isolated noun phrase that can function as a request for attention if accompanied by an appropriate gesture, such as pointing towards an object (e.g. *"A zebra!"*). As we do not have access to the visual context from the transcribed conversations, such utterances are annotated as `ambiguous`. More example cases for each label can be found in Table 1.

3.1.2. Grammatical Error Categories

For analysis purposes, we additionally annotate the fine-grained *types* of errors for each `ungrammatical` utterance. The coding scheme is slightly adapted from Hiller and Fernandez (2016) and Saxton et al. (2005b).[4] Table 2 describes all error categories along with specific examples. An utterance can be assigned multiple error categories, if appropriate.

3.2. Data

Transcribed conversations are taken from English CHILDES (MacWhinney, Brian, 2000) from children between 2 and 5 years of age. Transcripts are randomly selected from the available corpora, in order to increase variability of conversational contexts, parenting styles, and socioeconomic status.[5]

All transcripts are concatenated and then split to create files that each contain exactly 200 children's utterances. In total, 21 files are annotated, resulting in **4200 annotated utterances**.

3.3. Manual Annotation Results

The annotations are performed by 3 annotators. For the first 12 files, the annotations are discussed after each file in order to reach sufficient agreement on the annotation scheme. Each label for which at least 2 annotators disagreed is discussed until a consensus is established. From file 13 on, agreements are not discussed, and final labels are

[2] However, annotators are instructed to *not* take into account the *following* context of the conversation after the end of the current child's turn. This decision was made in order not to influence the grammaticality judgments based, for example, on the presence of clarification requests from the caregivers, which could bias the annotator into considering that the child's utterance is grammatical in retrospect.

[3] Castilla-Earls et al. (2020) introduced a class of `ambiguous` utterances in addition to `grammatical` and `ungrammatical` for similar reasons.

[4] We do not distinguish errors of omission/insertion/substitution (all errors in the `subject`, `verb`, and `object` categories are categorized as errors of omission.) We group regular and irregular past tense errors in the group `tense_aspect` (thereby also including errors with, e.g. participles). Further, regular and irregular plural errors are merged and all kinds of subject-verb agreement errors are included in the group `sv_agreement`).

[5] As our coding scheme was developed for Standard American and British English, we filter the data for diverging dialects. Fine-grained dialect information is not typically available in CHILDES, so we identify cases to be excluded by searching for caregivers whose speech contains a substantial number of indicative bigrams ("she don't", "you was") and exclude the corresponding corpora.

Category (broad)	Category (fine-grained)	Description	Examples	Frequency (Number)
Syntactic	subject	Missing subject	"Is hot?", "Going there."	17.7% (322)
	object	Missing object	"Can we look for.", "I like."	6.4% (116)
	verb	Missing verb (incl. copula)	"This yours.", "Because it."	14.7% (267)
Noun morphology	possessive	Missing or wrong use of possessive	"What's the other boy name?" "Where is Julia house?"	1.4% (26)
	plural	Wrong plural form or use	"No I like mans.", "More truck."	0.9% (17)
Verb morphology	sv_agreement	Subject-verb agreement errors	"He want cake.", "She are happy!"	3.4% (61)
	tense_aspect	Wrong tense or aspect inflection of a verb	"He's forgot me.", "She falled over."	7.9% (143)
Unbound	determiner	Missing or wrong determiner	"Blue wheel.", "A ice cream?"	18.8% (342)
	preposition	Missing or wrong preposition	"I want see it.", "Give it me!"	4.1% (75)
	auxiliary	Missing or wrong auxiliary verb	"We not to put them away.", "Someone been crashed."	11.4% (207)
	present_progressive	Wrong present progressive form	"It coming.", "What's he say?"	4.3% (78)
Other	other	Any other kind of grammatical error	"Many money!" (many/much) "Why it's falling?" (word order) "My want to eat" (wrong case)	8.9% (162)

Table 2: Descriptions of error categories that are used to label ungrammatical utterances. The last column is indicating the frequencies (and number of occurrences) calculated from our manual annotations.

calculated as the majority vote from the 3 annotators. For these files, the **inter-annotation agreement is 0.76** (Krippendorff's Alpha, with ordinal level of measurement (Krippendorff, 2018)).[6]

In total, 1333 (32%) utterances are annotated as `ungrammatical`, 648 (15%) as `ambiguous`, and 2219 (53%) as `grammatical`. For all `ungrammatical` utterances, additional fine-grained error categories (cf. Table 2) are added by one annotator. Their distribution is included in the last column of Table 2.

4. Automatic Annotation

4.1. Models

Based on our survey of the literature on automatic annotation of grammaticality (Section 2.1), we select a range of baseline models and state-of-the-art Transformer-based models for comparison.

We train the models to classify utterances as `grammatical`, `ungrammatical`, or `ambiguous` based on the annotations presented in Section 3.3. We run a majority classifier, SVMs based on n-gram features, and an LSTM (Hochreiter and Schmidhuber, 1997) that we pre-train on English CHILDES using a language modeling objective and fine-tune on the task. Further, we fine-tune on the grammatically task the following pre-trained Transformer models: BERT (bert-base-uncased) (Devlin et al., 2019), GPT2 (Radford et al., 2019), RoBERTa (roberta-large) (Liu et al., 2019), and DeBERTa (deberta-v3-large) (He et al., 2020, 2022). The LSTM as well as the Transformer models are provided with a list of preceding utterances as conversational context in addition to the target utterance (see also Section 4.3.1). We use early stopping by measuring Pearson's Correlation Coefficient (PCC)[7] on a validation set (20% of the training data) to avoid over-fitting during the fine-tuning. Further, we counteract the problem of imbalanced classes (cf. Section 3.3) by applying class weights on the loss. Further implementation details on the models can be found in Appendix A.1.

4.2. Results

We evaluate the models using 5-fold cross-validation, while making sure that there are no transcripts overlapping between training and test sets. As evaluation metrics, we report mean and standard deviation (over the 5 cross-validation folds) of Accuracy as well as PCC.

For models taking into account conversational context, we use a context length of 8 preceding turns. We base this decision on experiments with

[6] Cohen's kappa across the three annotators is on average 0.72 (standard deviation: 0.03).

[7] Related work on grammaticality classification usually relies on Matthews' Correlation Coefficient (Matthews, 1975); we use PCC as it takes into account the fact that we have 3 classes, which are ordinal.

DeBERTa showing that this context length is optimal for that model (cf. Section 4.3.1).

To have an estimate of how the models perform in comparison to inter-annotation agreement, we calculate the same evaluation metrics for human annotators. We report the mean and standard deviation of the pairwise Accuracy and PCC scores across the three annotators.

Table 3 shows the results. Regarding the evaluation metrics, we clearly see the advantage of using the PCC score over Accuracy; the latter tends to – misleadingly – favor classifiers with a majority-class bias. For example, Accuracy shows only a minimal performance difference of a majority class classifier compared to the SVM classifiers, while their PCC scores differ substantially. When comparing PCC scores, we observe that the SVMs show increasing performance with increasing n of their n-gram features, but reaching ceiling starting from 5-grams. The LSTM performs slightly worse than the SVMs according to PCC, and slightly better in Accuracy. The fine-tuned large language models outperform these models by a large margin, with DeBERTa performing best. The PCC score of the best models is very close to human annotators' agreement (0.71 vs. 0.76).

model	PCC	Accuracy
Majority class	0.00 $^{\pm 0.00}$	0.53 $^{\pm 0.11}$
SVM (1-gram)	0.28 $^{\pm 0.09}$	0.55 $^{\pm 0.02}$
SVM (2-gram)	0.29 $^{\pm 0.09}$	0.56 $^{\pm 0.03}$
SVM (3-gram)	0.30 $^{\pm 0.08}$	0.56 $^{\pm 0.03}$
SVM (4-gram)	0.31 $^{\pm 0.08}$	0.56 $^{\pm 0.03}$
SVM (5-gram)	0.32 $^{\pm 0.08}$	0.56 $^{\pm 0.03}$
SVM (6-gram)	0.31 $^{\pm 0.08}$	0.55 $^{\pm 0.03}$
LSTM	0.27 $^{\pm 0.17}$	0.58 $^{\pm 0.07}$
GPT2	0.50 $^{\pm 0.10}$	0.69 $^{\pm 0.04}$
BERT	0.63 $^{\pm 0.07}$	0.73 $^{\pm 0.04}$
RoBERTa	0.70 $^{\pm 0.07}$	0.79 $^{\pm 0.04}$
DeBERTa	0.71 $^{\pm 0.05}$	0.77 $^{\pm 0.03}$
Human annotators	0.76 $^{\pm 0.04}$	0.80 $^{\pm 0.02}$

Table 3: Accuracy and PCC scores on test set. Standard deviation over 5-fold cross-validation with varying model random initializations.

4.3. Analyses

4.3.1. Effect of Context Length

One major contribution of this work is the annotation of grammaticality *in context*, that is, by taking into account the preceding utterances in the conversation. In order to explore to what degree the models benefit from the context, we train the best-performing model (DeBERTa) with varying numbers of preceding utterances as context.

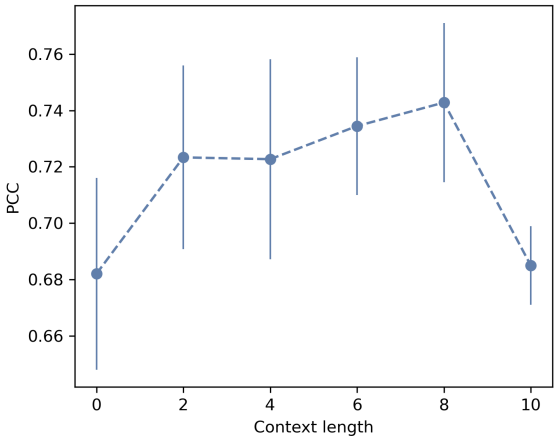

Figure 1: Mean and standard deviation of validation set PCC scores of DeBERTa as a function of the number of preceding utterances in the context.

Figure 1 shows the PCC scores on the validation set for context lengths 0 to 10. We observe a clear increase in performance for models with 2 utterances in the context as compared to no context (i.e. judging the grammaticality only based on the utterance itself). The performance further increases up to a context length of 8, after which it decreases slightly. We conclude that for this version of DeBERTa, a context length of 8 preceding utterances is optimal.

4.3.2. Effect of Training Data Size

Here we explore how the best model (DeBERTa) performs if it is only provided a subset of the training data. Such analyses can provide us insight into the possibilities of further improving model performance by manually annotating additional data. We train models using 20%, 40%, 60%, and 80% of the data. The cross-validation splits and test sets are kept the same.

In Figure 2, we display model performance as a function of training data size. The curve has a logarithmic-like shape. Between a training set size of 1000 and 2000 samples we observe a major improvement in PCC score. Starting from around 2000 training samples the model performance reaches ceiling. We therefore conclude that scaling up our manual annotation efforts is unlikely to lead to substantially improved automatic annotations.

4.3.3. Error Analysis

We perform an analysis of errors of the best-performing model (DeBERTa). Table 4 presents the confusion matrix for the automatic annotations

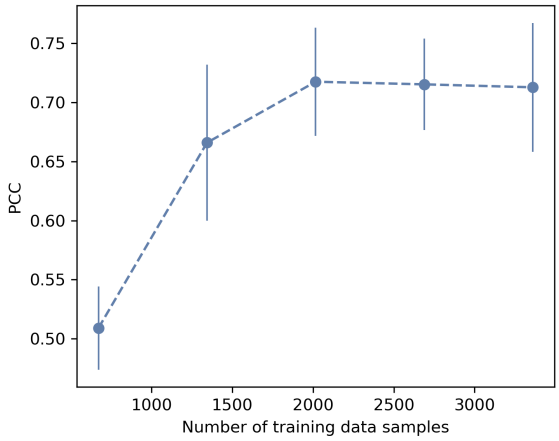

Figure 2: Effect of training data size on test set PCC scores of DeBERTa. The plot displays performance for models trained on 20%, 40%, 60%, 80%, and 100% of the training data.

on the test sets (data aggregated from the 5 cross-validation runs).

	Ungramm.	Ambig.	Gramm.
Ungramm.	0.72	0.13	0.15
Ambig.	0.17	0.56	0.27
Gramm.	0.04	0.09	0.87

Table 4: Confusion matrix for DeBERTa, normalized over the true labels.

We find that the model commits most errors for the `ambiguous` class (only 56% are correctly predicted) and performs best for `grammatical` utterances (87% correct). This pattern reflects the number of training examples available for each class (cf. Section 3.3). Manual inspection of the `ambiguous` utterances reveals that most misclassified examples are cases of missing subject, for which it is unclear whether they are used as imperative or declarative statements as well as noun phrases with missing verbs, for which the visual context was missing to judge whether there was a pointing gesture towards the mentioned object (see also Table 1).

Based on the error category annotations for `ungrammatical` utterances (Table 2) we can additionally analyze the model's performance for different kinds of grammatical errors.

Figure 3 shows the Recall[8] for the `ungrammatical` class split up by the different error categories. The scores do not diverge much from the average Recall, with the exception

[8]We cannot report Precision or F-score as we do not have error category annotations for false positives.

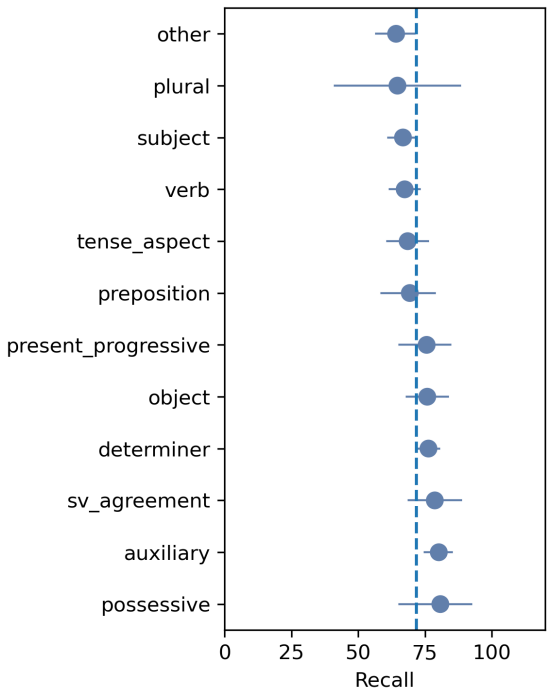

Figure 3: Recall scores for `ungrammatical` utterances with different error types. Error bars indicate 95% confidence intervals estimated using bootstrapping. The dotted line indicates the overall average Recall.

of the `other` and the `plural` class (lowest Recall). One explanation could be that the `other` class includes errors from various sources that are rather scarce (errors with case or word order), and therefore hard to learn for the model. The `plural` class is the least frequent in the training data (only 17 examples). On the other hand, detecting errors of missing auxiliaries and possessives could be easy as there is a large number of training examples for auxiliaries and the error patterns for both classes are rather consistent (a `possessive` error usually involves a missing suffixed "s").

5. Large-scale Annotation of English CHILDES

As a first application and sanity check of the models we introduced in this work, we annotate the grammaticality of all children's utterances in English CHILDES for children aged 2 to 5 years (excluding the manually annotated data). In total, we automatically annotate 276,200 utterances from 321 children and 1900 transcripts. To obtain the labels, we calculate the majority vote of all 5 fine-tuned DeBERTa models (there are 5 models trained on the different cross-validation splits).

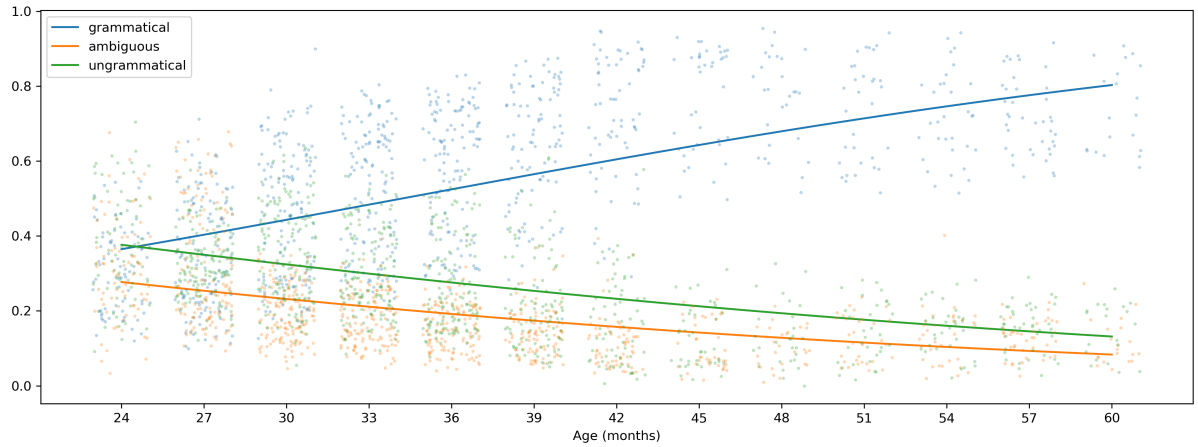

Figure 4: Proportion of `grammatical`, `ambiguous`, and `ungrammatical` utterances for transcripts in English CHILDES of children aged 2 to 5 years. Additionally, we display fitted logistic regression curves.

In Figure 4 we present the proportion of `grammatical`, `ambiguous`, and `ungrammatical` utterances for each annotated transcript.[9] Further, we display fitted curves of a logistic regression for each target label. We observe a clear increase in the proportion of `grammatical` utterances with increasing age. At the same time, the proportion of `ambiguous` and `ungrammatical` utterances decreases. We use mixed effects models to verify these trends. Regarding the proportion of `grammatical` utterances we fit the following model:

$$\texttt{grammatical} \sim \texttt{age} + (1|\texttt{transcript}) \quad (1)$$

We obtain age : $\beta = 0.014, SE = 0.001, p < 0.001$, indicating a significant positive correlation with age. We run equivalent models for the proportion of `ambiguous` and `ungrammatical` utterances and obtain significant negative correlations. For `ambiguous` utterances: age : $\beta = -0.006, SE < 0.001, p < 0.001$ and `ungrammatical` utterances: age : $\beta = -0.008, SE < 0.001, p < 0.001$.

6. Limitations

In order to close the remaining small performance gap between the models and human annotators, one possibility would be to increase the amount of manual annotations. However, our experiments with varying training data sizes show that model performance most probably won't increase substantially with a simple increase in training data size (Section 4.3.2). On the other hand, our error analysis reveals that many failure cases are likely caused by imbalanced classes in the training data (Section 4.3.3). In order to address these issues, future annotation efforts could be targeted to obtain more training data for `ungrammatical` and `ambiguous` utterances.

The current annotations allow for a broad classification of utterances into `grammatical`, `ungrammatical`, and `ambiguous`. While this is a reasonable first step for the study of grammaticality, many patterns are dependent on specific error types. For example, the effects of utterance length on grammaticality differ for errors of omission vs. commission (Castilla-Earls et al., 2022). Further, Saxton et al. (2005a) found that corrective feedback for syntactic errors is more frequent than for morphological errors, and that negative feedback in the form of reformulations is associated with gains in the grammaticality of child speech for 3 out of 13 tested grammatical error categories. More generally, we can gain insight from the study of a specific grammatical phenomenon, such as learning of the English past tense (Saxton, 1997; Rumelhart and McClelland, 1986; Marchman and Bates, 1994; McClelland and Patterson, 2002). To enable more fine-grained analyses of specific error classes, models could be trained to classify the error type in addition to the general grammaticality. As the distribution of error types is highly skewed (cf. Table 2), there is currently not enough manually annotated data to train models for a reliable classification. Again, targeted annotations could be carried out to increase the number of examples of less frequent error types.

Another important limitation of our contribution is that our annotations assume children and caregivers speak Standard American or British English. In some cases, sentences that are labeled ungrammatical (e.g., "I been here.", "You was there.", "She don't like it.") are grammatical in other English dialects, and so our annotated data and classifiers

[9] We excluded transcripts with less than 100 child utterances to reduce clutter.

are not appropriate for the study of other dialects. Even though we make efforts to filter out corpora of diverging dialects (cf. Section 3.2), some instances in the dataset (manually or automatically labeled) may have been missed and, therefore, contain inaccurate labels.

7. Discussion and Conclusion

Research in child language acquisition has recently started to move towards large-scale studies and cross-lab collaborations to overcome issues such as small sample sizes, lack of population diversity, and inconsistent measures (Frank et al., 2017; Byers-Heinlein et al., 2020). The current work contributes to this ongoing effort in the community, providing a tool for the automatic annotation of grammaticality in child-caregiver conversations. This tool will enable researchers to conduct reproducible and cumulative research on a large scale.

We develop a coding scheme for the annotation of the grammaticality of children's utterances in conversation and manually annotate a representative sample. Based on these annotations, we train and evaluate a range of NLP models on this task. We find that the best models are performing on par with human annotators.

Much research in NLP has dealt with the annotation of grammaticality of utterances in isolation (Warstadt et al., 2019, 2020). Here we deal with grammaticality in naturalistic child-caregiver conversations and highlight important differences. Indeed, one of the main contributions of our work is the finding that the grammaticality of an utterance is dependent on the conversational context. Analyzing the dependence of model performance on context length (i.e., how many previous utterances are given to a model in order to best judge the grammaticality of a target utterance) revealed that while it is possible to reach decent performance when annotating the grammaticality of utterances in isolation (without context), the addition of two previous utterances from the conversational context results in a substantial improvement. The best performance is reached with a context length of 8 utterances.

Finally, we show that the developed tool can be used to study the trajectory of grammatical development by applying it to annotate a large-scale corpus, enabling more systematic research into the underlying learning mechanisms.

A promising area of application of the proposed models is the study of grammaticality in language impairment. It has been found that children's productive performance in terms of grammaticality is correlated with specific language impairment, and could probably be used as an early indicator of risk (Rice et al., 2010; Souto et al., 2014; Guo and Schneider, 2016; Eisenberg and Guo, 2013).

Additionally, by allowing for more reproducible large-scale investigations, the models can aid in adjudicating debates about the learning mechanisms, such as the debate about the role of the caregiver's corrective feedback in language acquisition (Brown and Hanlon, 1970; Demetras et al., 1986; Saxton, 2000; Marcus, 1993; Morgan et al., 1995; Nelson et al., 1973) as well as providing a more thorough test to newly proposed mechanisms such as communicative feedback (Warlaumont et al., 2014; Nikolaus and Fourtassi, 2023).

8. Acknowledgements

This work, carried out within the Labex BLRI (ANR-11LABX-0036) and the Institut Convergence ILCB (ANR-16CONV-0002), has benefited from support from the French government, managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

The project leading to this publication has received funding from Excellence Initiative of Aix-Marseille - A*MIDEX (Archimedes Institute AMX-19-IET-009), a French "Investissements d'Avenir" Programme.

Further, this work was supported by the ANR MACOMIC (ANR-21-CE28-0005-01).

This work was performed using HPC resources from GENCI–IDRIS (Grant 2022-D011013886).

9. Bibliographical References

David Adger. 2003. *Core syntax: A minimalist approach*, volume 20. Oxford University Press.

Roger Brown. 1973. *A First Language: The Early Stages*. Harvard University Press.

Roger Brown and Camille Hanlon. 1970. Derivational complexity and order of acquisition in child speech. *Cognition and the development of language*.

Krista Byers-Heinlein, Christina Bergmann, Catherine Davies, Michael C. Frank, J. Kiley Hamlin, Melissa Kline, Jonathan F. Kominsky, Jessica E. Kosie, Casey Lew-Williams, Liquan Liu, Meghan Mastroberardino, Leher Singh, Connor P. G. Waddell, Martin Zettersten, and Melanie Soderstrom. 2020. Building a collaborative psychological science: Lessons learned from ManyBabies 1. *Canadian Psychology / Psychologie canadienne*, 61(4):349–363.

Anny Castilla-Earls, David J. Francis, and Aquiles Iglesias. 2022. The Complex Role of Utterance Length on Grammaticality: Multivariate Multilevel Analysis of English and Spanish Utterances of First-Grade English Learners. *Journal of Speech, Language, and Hearing Research : JSLHR*, 65(1):238–252.

Anny P. Castilla-Earls, Brittany Harvey, Katrina Fulcher-Rood, and Christopher D. Barr. 2020. The Impact of Clinical Review Bias on Child Language Grammaticality. *Communication Disorders Quarterly*, 41(4):214–221.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton.

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

Eve V Clark. 2016. *First Language Acquisition*. Cambridge University Press.

M. J. Demetras, Kathryn Nolan Post, and Catherine E. Snow. 1986. Feedback to first language learners: the role of repetitions and clarification questions*. *Journal of Child Language*, 13(2):275–292.

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.

Sarita L. Eisenberg and Ling-Yu Guo. 2013. Differentiating Children With and Without Language Impairment Based on Grammaticality. *Language, Speech, and Hearing Services in Schools*, 44(1):20–31.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying Non-Sentential Utterances in Dialogue: A Machine Learning Approach. *Computational Linguistics*, 33(3):397–427.

Michael C. Frank, Elika Bergelson, Christina Bergmann, Alejandrina Cristia, Caroline Floccia, Judit Gervain, J. Kiley Hamlin, Erin E. Hannon, Melissa Kline, Claartje Levelt, Casey Lew-Williams, Thierry Nazzi, Robin Panneton, Hugh Rabagliati, Melanie Soderstrom, Jessica Sullivan, Sandra Waxman, and Daniel Yurovsky. 2017. A Collaborative Approach to Infant Research: Promoting Reproducibility, Best Practices, and Theory-Building. *Infancy*, 22(4):421–435.

E Mark Gold. 1967. Language identification in the limit. *Information and Control*, 10(5):447–474.

Ling-Yu Guo and Phyllis Schneider. 2016. Differentiating School-Aged Children With and Without Language Impairment Using Tense and Grammaticality Measures From a Narrative Task. *Journal of Speech, Language, and Hearing Research*, 59(2):317–329.

Randy Allen Harris. 1993. *The linguistics wars*. Oxford University Press.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *Proceedings of the International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *Proceedings of the International Conference on Learning Representations*.

Sarah Hiller and Raquel Fernandez. 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, Berlin, Germany. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Jong-Bok Kim and Peter Sells. 2008. *English syntax: An introduction*. CSLI publications Stanford, CA.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations*, San Diego, CA, USA.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41(5):1202–1241.

S. Lawrence, C.L. Giles, and S. Fong. 2000. Natural language grammatical inference with recurrent neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 12(1):126–140. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

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.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations*.

Virginia A. Marchman and Elizabeth Bates. 1994. Continuity in lexical and morphological development: a test of the critical mass hypothesis. *Journal of Child Language*, 21(2):339–366. Publisher: Cambridge University Press.

Gary F. Marcus. 1993. Negative evidence in language acquisition. *Cognition*, 46(1):53–85.

B. W. Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

James L. McClelland and Karalyn Patterson. 2002. Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in cognitive sciences*, 6(11):465–472.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. RuCoLA: Russian Corpus of Linguistic Acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

James L. Morgan, Katherine M. Bonamo, and Lisa L. Travis. 1995. Negative evidence on negative evidence. *Developmental Psychology*, 31:180–197.

Keith E Nelson, Gaye Carskaddon, and John D Bonvillian. 1973. Syntax Acquisition: Impact of Experimental Variation in Adult Verbal Interaction with the Child. *Child Development*, 44(3):497–504.

Mitja Nikolaus and Abdellah Fourtassi. 2023. Communicative Feedback in Language Acquisition. *New Ideas in Psychology*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12:2825–2830.

Steven Piantadosi. 2023. Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint, lingbuzz*, 7180.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Mabel L. Rice, Filip Smolik, Denise Perpich, Travis Thompson, Nathan Rytting, and Megan Blossom. 2010. Mean Length of Utterance Levels in 6-month Intervals for Children 3 to 9 Years with and without Language Impairments. *Journal of speech, language, and hearing research : JSLHR*, 53(2):333.

David E. Rumelhart and James L. McClelland. 1986. On learning the past tenses of English verbs. In *Psycholinguistics: Critical Concepts in Psychology, Volume 4*, pages 216–271. Routledge.

Kenji Sagae, Alon Lavie, and Brian MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05*, pages 197–204, Ann Arbor, Michigan. Association for Computational Linguistics.

Matthew Saxton. 1997. The Contrast Theory of negative input. *Journal of Child Language*, 24(1):139–161.

Matthew Saxton. 2000. Negative evidence and negative feedback: immediate effects on the grammaticality of child speech. *First Language*, 20(60):221–252.

Matthew Saxton, Phillip Backley, and Clare Gallaway. 2005a. Negative input for grammatical errors: effects after a lag of 12 weeks. *Journal of Child Language*, 32(3):643–672.

Matthew Saxton, Carmel Houston–Price, and Natasha Dawson. 2005b. The prompt hypothesis: Clarification requests as corrective input for grammatical errors. *Applied Psycholinguistics*, 26(3):393–414.

Hollis S. Scarborough. 1990. Index of Productive Syntax. *Applied Psycholinguistics*, 11(1):1–22.

Taiga Someya, Yushi Sugimoto, and Yohei Oseki. 2023. JCoLA: Japanese Corpus of Linguistic Acceptability. ArXiv:2309.12676 [cs].

Sofía M. Souto, Laurence B. Leonard, and Patricia Deevy. 2014. Identifying Risk for Specific Language Impairment with Narrow and Global Measures of Grammar. *Clinical linguistics & phonetics*, 28(10):741–756.

Dominique Sportiche, Hilda Koopman, and Edward Stabler. 2013. *An Introduction to Syntactic Analysis and Theory*. John Wiley & Sons.

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

Daniela Trotta, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. 2021. Monolingual and Cross-Lingual Acceptability Judgments with the Italian CoLA corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2929–2940, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joachim Wagner, Jennifer Foster, and Josef van Genabith. 2009. Judging Grammaticality: Experiments in Sentence Classification. *CALICO Journal*, 26(3):474–490.

Anne S. Warlaumont, Jeffrey A. Richards, Jill Gilkerson, and D. Kimbrough Oller. 2014. A Social Feedback Loop for Speech Development and Its Reduction in Autism. *Psychological Science*, 25(7):1314–1324.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, and Morgan Funtowicz. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

10. Language Resource References

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

A. Appendix

A.1. Model training details

A.1.1. SVM Classifiers

We train Support Vector Machines (SVM) based on n-gram features as simple baseline models for the task.

The data is tokenized using byte-pair encoding (BPE) with a vocabulary size of 10,000. Afterwards, for each n-gram level (1-gram, 2-gram, ...), a vocabulary of the 1000 most commonly occurring n-grams in the training set is constructed. The features for a given utterance consists of a sparse array containing the number of occurrences of each n-gram from the vocabulary. For SVMs with n-gram features of n greater than 1, the features from all smaller n are included (for example, the 2-gram model features are 2000-dimensional, consisting of a concatenation of the 2-gram and the unigram features).

These features are fed into a C-Support Vector Classification model with balanced class weights and the default arguments from the implementation in scikit-learn (Pedregosa et al., 2011) which is based on libsvm (Chang and Lin, 2011).

These baseline models are trained without any conversational context.

A.1.2. LSTM

This model consists of a single-layer LSTM with 512 hidden units and a maximum sequence length of 200 tokens.

Pre-training As a first step, the LSTM is pre-trained with a language modeling objective on the English CHILDES data (cf. Section 5), excluding all the data that is manually annotated (in order not to train on data that will be part of any of the test sets during cross-validation). 10,000 sentences are set aside as a validation set to perform early-stopping based on the validation loss.

The data is tokenized using (BPE) with a vocabulary size of 10,000 and special speaker tokens for child (`[CHI]`) and caregiver (`[CAR]`) that are prepended to each utterance.

The maximum sequence length is set to 200 tokens, the model is trained with a batch size of 100 and Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 10^{-4}.

Fine-tuning After pre-training, the model is equipped with an additional linear classification layer that is fed the output from the last timestep from the LSTM. Then, it is fine-tuned on the grammaticality classification task using a cross-entropy loss with balanced class weights.

The fine-tuning is also performed using an Adam optimizer with initial learning rate of 10^{-4}, batch size of 100, and using early stopping based on the PCC score on a held-out validation set (20% of the training data).

A.1.3. Transformer-based Models

We fine-tune the following Transformer-based models: BERT (bert-base-uncased) (Devlin et al., 2019), GPT2 (Radford et al., 2019), RoBERTa (roberta-large) (Liu et al., 2019), and DeBERTa (deberta-v3-large) (He et al., 2020, 2022).

We leverage pre-trained models from Huggingface (Wolf et al., 2020). We prepend special speaker tokens for child (`[CHI]`) and caregiver (`[CAR]`) to each utterance.

On top of each respective model, a new linear classification layer is fine-tuned on the grammaticality classification task using a cross-entropy loss with balanced class weights.

This fine-tuning is performed using an AdamW optimizer (Loshchilov and Hutter, 2018) with initial learning rate of 10^{-5}, batch size of 100, and using early stopping based on the PCC score on a held-out validation set (20% of the training data).