Large-scale study of speech acts' development in early childhood

Mitja Nikolaus
Aix Marseille Univ, Université de Toulon, CNRS, LIS, LPL, Marseille, France

Eliot Maes
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

Jeremy Auguste
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

Laurent Prévot
Aix-Marseille Univ, CNRS, LPL, Aix-en-Provence, France

Abdellah Fourtassi
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

Abstract: Studies of children's language use in the wild (e.g., in the context of child-caregiver social interaction) have been slowed by the time- and resource- consuming task of hand annotating utterances for communicative intents/speech acts. Existing studies have typically focused on investigating rather small samples of children, raising the question of how their findings generalize both to larger and more representative populations and to a richer set of interaction contexts. Here we propose a simple automatic model for speech act labeling in early childhood based on the INCA-A coding scheme (Ninio et al., 1994). After validating the model against ground truth labels, we automatically annotated the entire English-language data from the CHILDES corpus. The major theoretical result was that earlier findings generalize quite well at a large scale. Further, we introduced two complementary measures for the age of acquisition of speech acts which allows us to rank different speech acts according to their order of emergence in production and comprehension. Our model is shared with the community so that researchers can use it with their data to investigate various question related to language use both in typical and atypical populations of children.

Keywords: language acquisition; conversation; language use; speech acts

Corresponding author: Mitja Nikolaus, Aix-Marseille University, LIS, 163 Avenue de Luminy, 13288, Marseille, France. Email: mitja.nikolaus@univ-amu.fr.

ORCID ID: https://orcid.org/0000-0001-5609-6628

Citation: Nikolaus, M., Maes, E., Auguste, J., Prevot, L., & Fourtassi, A. (2022). Large-scale study of speech acts' development in early childhood. *Language Development Research, 2*(1), 268–305. https://doi.org/10.34842/2022.0532

Introduction

Research on language learning has largely focused on investigating how children acquire language form (e.g., phonology, lexicon, and syntax) and content (e.g., word and sentence meanings). Yet, an important aspect of language learning, which has received less attention, is the mastery of how to use language adequately in natural social interactions (Bloom & Lahey, 1978). This mastery involves, in particular, using linguistic utterances to encoding and decode communicative intents (Grice, 1975) or speech acts that characterize the illocutionary force of an utterance (e.g question, assertion, and request) (Searle, 1976). Children's learning of speech acts is crucial for their ability to engage in coherent conversations. For example, it is important to recognize that an utterance is a "question" requiring an "answer", or that it is a "request" requiring "acceptance" or "refusal", instead.

Several taxonomies have been proposed that purport to capture children's emergent repertoire of speech act categories in the context of early child-caregiver social interactions (for reviews, see Cameron-Faulkner, 2014; Casillas & Hilbrink, 2020), the most comprehensive to date is the Inventory of Communicative Acts and its abridged version INCA-A (Ninio et al., 1994).

Snow et al. (1996) used INCA-A to study the emergence of speech act major classes in a longitudinal corpus of children aged 14 to 32 months old.[1] They documented several important findings that not only informed our understanding of language use development, but also shed light on how children's emerging linguistic skills interface with the development of their social-cognitive competences. By analyzing the development of the number of distinct speech acts as well as the distribution of speech acts used by children, they showed that when children utter their first words, they already express a range of simple communicative intents such as requests and questions. The repertoire of speech acts was observed in this study to increase rapidly within the first years of life, in tandem with development in social-cognitive and linguistic skills: Children become able to express more sophisticated speech acts such as "promise", "prohibit", and "persuade". Using the same coding scheme, Rollins (1999, 2017) has shown that investigating speech act development can also help us study atypical cognitive development such as autism.

While this previous effort has been influential in the study of language use development, it has relied on hand annotation to code the data, which has limited the researchers' ability to explore how their findings generalize to larger population of children and across different interactive contexts. In fact, INCA-A is a rather complex scheme with a

[1] While the terms "speech act" and "communicative intent" have sometimes been used by different researchers to mean slightly different things or to refer to different taxonomies, here — and for simplicity — we use them interchangeably to refer to the categories of communicative intents at the utterance level, as defined in the INCA-A coding scheme.

large number of categories (e.g., 67 different types of illocutionary acts) and its hand-annotation — including the effort of train annotators — is prohibitively expensive to deploy at a large scale.

Current study

The current study aims at addressing this gap using recent advances in automatic speech act labeling. Using Snow et al.'s child-caregiver corpus and its INCA-A annotation, we tested various models on their ability to map utterances to corresponding speech acts and we selected the one that provided the best performance on a testing set made of unseen utterances from the same corpus.

Using this model, we examined how previous findings in speech act development generalized at scale. To this end we proceeded in two steps: First, we validated the chosen model by testing its ability to replicate key findings from Snow et al. (1996). More specifically, we reproduce developmental patterns regarding the number of distinct speech acts as well as the distribution of speech acts used by children from 14 to 32 months of age. Second, and after successful validation, we used the model to automatically label the entire North American English-language section of CHILDES (MacWhinney, 2017) and compared the results of this large-scale analysis to the original findings.

Additionally, we proposed methods for quantifying the age of acquisition of a speech act both in terms of production and comprehension. These measures have allowed us to rank different speech acts according to their order of emergence. We first examined this order of emergence with data in Snow et al. (1996), and second, thanks to our automatic labelling tool, we tested how this developmental trajectory generalized across all English language corpora in CHILDES.

The paper is organized as follows. First, we introduce the dataset and provide an overview of models for automatic annotation of speech acts that we evaluated in our study. Further, we define the measures for speech act emergence in production and comprehension. In the results sections we compare the performance of the selected models and present replications the findings of Snow et al. (1996) using automatically generated labels. Additionally, the results contain predicted ages of acquisition for each speech act using both manually-annotated and automatically-annotated data. Finally, we discuss the results in the context of language development in general and point out limitations of the current approach which offer possibilities for future research.

Datasets and Methods

Datasets

New England Corpus. For model training and validation, we use ground-truth labels from the dataset collected by Snow et al. (1996) which is the largest child-caregiver interaction dataset annotated for speech acts. This dataset was collected for a longitudinal study of 52 children aged 14, 20 and 32 months old. Child-caregiver dyads were invited for three sessions that consisted of semi-structured free play. All conversations were recorded, transcribed, and annotated with INCA-A coding scheme. There were 55,941 labelled utterances in total.

English-Language CHILDES. In order to test how findings from Snow et al. (1996) generalize to a larger dataset of children and across different contexts, we use the entire North American English-language subset of CHILDES made of children in the same age range (i.e., between 14 and 32 month old), resulting in 2078 different transcripts totaling 354 children.[2]

INCA-A Coding Scheme

INCA-A is the most comprehensive coding scheme to date that was designed to capture children's emerging speech acts it the context of spontaneous social interaction with a caregiver (Ninio et al., 1994). The coding scheme has two coding tiers: 1) the interchange level that annotates the topic of the conversation (e.g., "discussing a recent event"), and may span multiple utterances, and 2) the illocutionary force level (e.g., "Ask a yes/no question") which is determined at the utterance level. Here, we focus on the illocutionary force. INCA-A has 67 different speech act types, which are grouped into several high-level categories such as directives, declarations, commitments, markings, statements, questions, evaluations, and other vocalizations.[3]

Automatic Classification of Speech Acts

Speech act classification (also referred to as dialogue act tagging in the field of Natural Language Processing) describes the task of annotating utterances in dialogue with their respective speech act category. Given a transcript of a conversation and a speech act coding scheme, each utterance in the transcript is assigned one of the speech acts in the coding scheme (Stolcke et al., 2000).

Early work used Hidden Markov Models to map utterances to speech acts using a set of lexical, collocational, and prosodic cues (Stolcke et al., 2000). Subsequent work has used

[2]For fair comparison, we excluded very short transcripts where the number of children's utterances was less than the minimum number of children's utterances in transcripts of the New England corpus at the same age.

[3]Refer to the appendix for the full list of speech acts.

Recurrent Neural Networks (RNNs) such as Long short-term memory networks (LSTMs) for encoding transcribed utterances in order to leverage the sequential structure of the data (Khanpour et al., 2016). More recent approaches combine hierarchical deep neural network encoders with Conditional Random Field (CRF) decoders (Kumar et al., 2018). While the encoder is aware of relationships between the different utterances of a transcript and thus models dependencies in the *feature space*, the CRF can model transition probabilities in the *label space*. In this way, it can for example learn common adjacency pairs (Schegloff & Sacks, 1973) in conversation, e.g. that questions are usually followed by answers.

Following this brief review, we considered and compared the following models.

Baselines

As this work is the first to propose automatic speech act annotation using the INCA-A coding scheme on child-caregiver conversations, we run several baselines in order to obtain reference performances on this specific task.

Majority Classifier. As a first simple baseline, we consider the majority classifier, which always predicts the most frequent speech act.

Random Forests. We use the reference implementation of a random forests algorithm from scikit-learn (Pedregosa et al., 2011). As features, we provide the model with the speaker (caregiver or child), bag-of-words, part-of-speech tags (that are present in the corpus[4]), and the number of words in the utterance.

Support Vector Machine. Using the same features as for the random forests model, we train and evaluate a linear support vector machine from scikit-learn.

Conditional Random Field

Next, we consider a CRF as annotation model. We hypothesized this model would outperform the baselines thanks to its ability to track transition probabilities in the label space. We use *pycrfsuite*[5] (Okazaki, 2007) to implement the CRF. We extend the set of features used by the baseline models and add bigrams and repetitions (words that are repeated from the previous utterance, as well as the number of repeated words normalized by the utterances length) to provide the model with some context of the previous utterances.[6] The model uses the whole conversation in a transcript to find the most probable sequences of labels using the Viterbi algorithm.

[4]The POS tags in CHILDES were automatically generated using the Morphological Analysis algorithm (MOR; MacWhinney, 2000) which yields a high accuracy rate on CHILDES adult data (above 99%).

[5]https://github.com/scrapinghub/python-crfsuite

[6]In preliminary experiments we tested adding all the exact words of previous utterances as features to the model but observed, if anything, a small degradation in performance.

Hierarchical LSTM + CRF

We further consider a model that is inspired by state-of-the-art speech act annotation models in other domains. More specifically, we implement a hierarchical LSTM encoder combined with a CRF decoder similar to the implementation of Kumar et al. (2018). The encoder processes the utterances within a transcript on two levels. We add a special token representing the speaker identity to the beginning of each utterance. Afterwards, for each utterance, one-hot encodings of the words are passed through word embeddings, and are then encoded using the word-level LSTM. The last hidden representation of this LSTM forms the latent utterance representation, which is then passed into the utterance-level LSTM. This higher-level LSTM processes the utterances sequentially and generates conversation-context-aware representations. The output of each timestep of the utterances LSTM is then passed as features to a CRF, which predicts the corresponding speech act. The model has access to contextualized utterance representations as well as the history of speech acts for the classification task. A high-level overview of the architecture of this model can be found in the appendix (Figure 9).

BERT

Given recent developments in NLP regarding the success of pre-trained contextualized embeddings (Devlin et al., 2018), we additionally test the performance of a model where utterances are encoded using BERT. The success of these models relies on self-attention mechanisms that allow the model to create contextualized representations with long-range dependencies as well as setups in which the encoder is pre-trained on large-scale data before being fine-tuned on the actual task. Here we replace the word-level LSTM of the Hierarchical LSTM + CRF model with a pre-trained publicly available implementation of DistilBERT (Wolf et al., 2020). The weights of BERT are fine-tuned on the task. Details on the hyperparameters of the neural network models can be found in the Appendix.

Measures of Speech Act Emergence

Here we introduce measures of speech acts' age of emergence, both at the level of children's production and comprehension.

Production

By analogy to work in word learning (Braginsky et al., 2016; Goodman et al., 2008), we define the age of acquisition of a speech act in production as the month by which at least 50% of the observed children produce it.[7] More precisely, for each speech act s,

[7] In line with Snow et al. (1996), we consider that a child acquired a speech act if it is produced at least twice at a certain age.

we proceed as follows:

1. For each age in the dataset (i.e., 14, 20 and 32 months), calculate the proportion of children who are producing S at least twice.

2. Perform a logistic regression over these proportions.

3. Measure the age of first production as the age where the logistic regression curve surpasses the value 0.5.

Comprehension

Studying speech act emergence only from a production point of view may underestimate children's pragmatic competence. Thus, we additionally introduce a measure for children's comprehension, which we define as the ability of children to respond to a target speech act in a contingent fashion (e.g., responding to a "yes/no question" with "yes" or "no"). More precisely, for each speech act S, we proceed as follows:

1. Find all utterances produced by the caregivers labelled as S.

2. Find all cases where these utterances are followed by an utterance of the child.

3. For each occurring follow-up utterance, annotate whether its speech act is contingent as a response to S.[8] We manually annotated the contingency of all combinations of speech act categories that appear in the data. Using this annotation, we could label each child utterance that follows a caregiver utterance as either possibly contingent or non-contingent based on the corresponding speech act category. The contingency annotation can be found in the GitHub repository: https://github.com/mitjanikolaus/childes-speech-acts.

4. For each age (14, 20 and 32 months), calculate the proportion of contingent follow-up utterances.

5. Perform a logistic regression over the proportion.[9]

6. Measure the age of comprehension as the age where the logistic regression curve surpasses the value 0.5.

[8] Annotating contingency was done using a binary scale, indicating whether the speech act was *possibly* contingent (1) or clearly non contingent (0). A speech act was considered contingent (1) if it can form a coherent response with respect to the previous speech act, and non contingent (0) otherwise.

[9] We only regard data points where the proportion was calculated over at least 2 examples, i.e. where there were at least two utterances with follow-ups.

Table 1: Accuracy for all models.

Model	**Accuracy**
Majority Classifier	13.44% ($\pm 2.81\%$)
Random Forests	62.81% ($\pm 6.29\%$)
Support Vector Machine	62.42% ($\pm 6.97\%$)
Conditional Random Field	**72.33%** ($\pm 4.23\%$)
Hierarchical LSTM + CRF	69.77% ($\pm 3.70\%$)
+ BERT	68.50% ($\pm 4.29\%$)
Inter-Annotator Agreement	81% to 89%

Results and Analyses

First, we compare performance across all models presented above on the New England corpus. Second, we choose the best performing model and test the extent to which its predicted labels replicate major findings obtained using gold labels from Snow et al. (1996). Finally, we use the model to automatically label the North American section from CHILDES and explore how original findings from Snow et al. (1996) on the emergence of speech acts generalize to this larger dataset.

Comparing Models of Speech Act Labeling

We evaluate our models on the speech act annotations of utterances in the New England corpus (Snow et al., 1996). We employ 5-fold cross validation so that we evaluate (and later utilize in all analyses) only the predicted labels on the parts of the corpus that were not seen by the model in the training phase. To this end, and to obtain labels for the whole New England corpus, we train models on 5 different training sets, always holding out 20% of the data. Then we use each of the trained models to label their respective test sets which together form a set of predicted speech act labels for the whole New England corpus.

We report the mean and standard deviation (based on the five cross-validation runs) of each model's accuracy in Table 1. The majority classifier had a high score given the relatively large label space. This could be explained by the fact the label distribution is heavily skewed (Figure 1). A small set of speech acts are used very frequently while several others are rarely used. As for other baseline models, i.e., random forests and support vector machine, the scores are relatively high despite the fact that they do not have access to the conversation history or dependencies in the label space. Our more sophisticated models (Hierarchical LSTM with and without BERT) did not improve performance much, which could be explained by the lack of large-scale training data. Further, in the case of the BERT-based model, we hypothesize that we do not see any

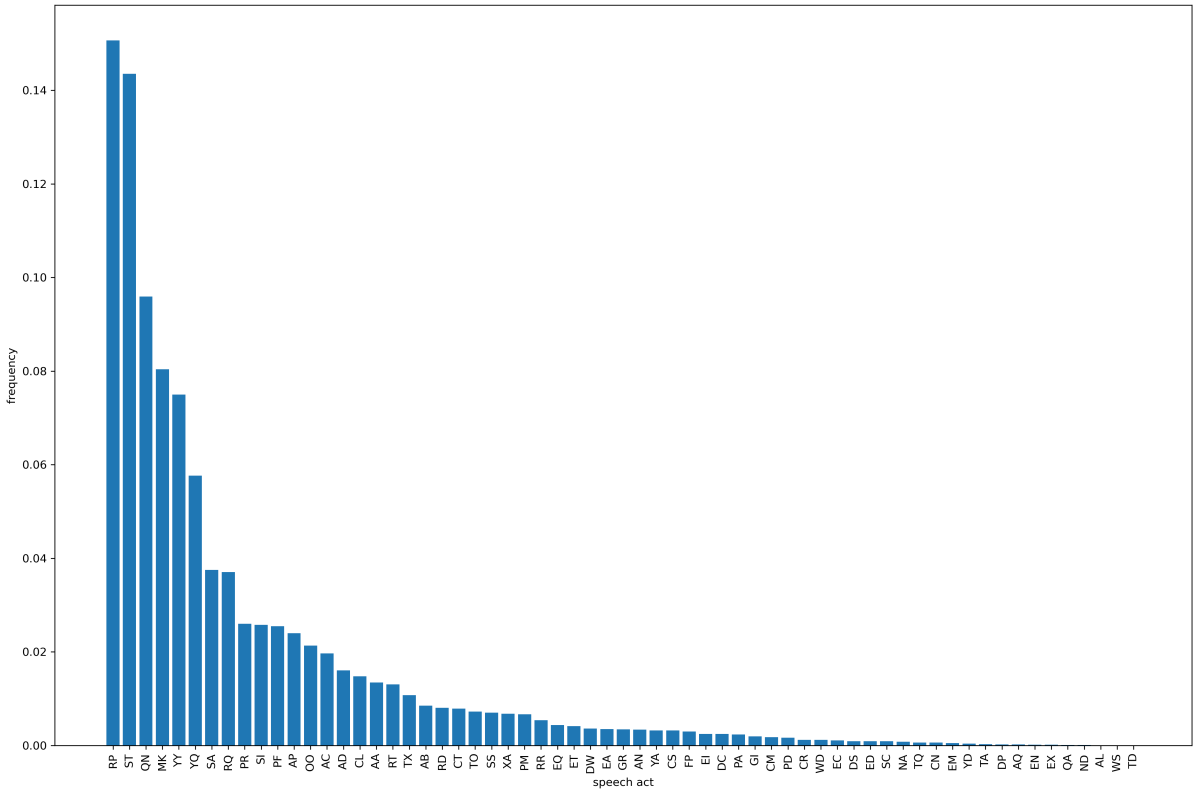

Figure 1. ***Distribution of frequencies of all speech acts in the New England corpus. Labels from the INCA-A tagset are listed in the Appendix.***

performance gains because this model is pre-trained on large text corpora (based on e.g. Wikipedia) that do not have much in common with the dynamics of child-caregiver conversations.

Finally, we find that the CRF model shows the highest accuracy scores, outperforming the baselines as well as the more complex neural network models. Its large performance gains over the baseline are most likely explained by its ability to track transition probabilities in the label space. This property is crucial for the task of speech act annotation; given a speech act sequence, certain speech acts are very likely to follow and others are not. The CRF is the best-performing model, and thus, it is the one we for the rest of analyses in the paper.

Amount of Training Data

We further investigate the effects of the amount of training data on the performance of the CRF model. Figure 2 presents the test accuracy as a function of training set size for this model. The performance indicated in Table 1 was obtained when the model was

trained on 80% of the dataset (around 44,000 utterances). However, from the learning curve in Figure 2 we can see that the model actually achieves decent scores (around 65% accuracy) when trained on only 5,000 annotated utterances, and almost converged when trained on about 20,000 annotated utterances.

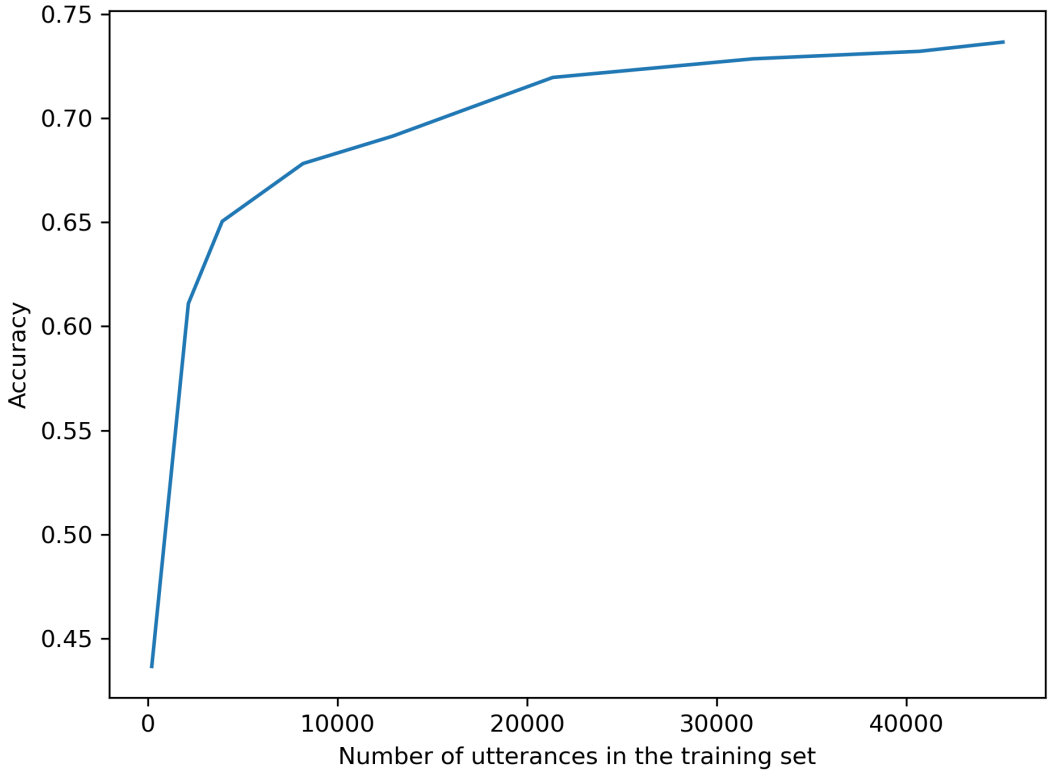

Figure 2. ***CRF: Accuracy as a function of training set size.***

Error Analysis

To gain a better understanding of our best performing model (the CRF), we perform an error analysis. For each speech act category, we calculate precision, recall and f1-score. Results can be found in the Appendix. The variance of the f1-scores for different categories is remarkably high, with values ranging from 0 to 95%. Performance is best for speech acts `QN` ("Ask a product-question") and `EA` ("Elicit onomatopoeic or animal sounds.") and worst for speech acts such as `CR` ("Criticize or point out error in nonverbal act") and `AL` ("Agree to do something for the last time.").

One important factor affecting the per-label performance is the availability of training examples and the distribution of speech acts in the dataset is heavily skewed with a long tail (see Figure 1). For labels with only very few training examples the model struggles to pick up important features. Indeed we find a high correlation between the frequency of

labels and their respective f1-score (Spearman correlation coefficient: 0.59, $p < 1 \cdot 10^{-5}$). The example in Table 2 illustrates this finding. In the conversation, all speech acts have been predicted correctly by our model except for the last utterance ("You're a nut"), which is labelled as ST ("Make a declarative statement") while the ground-truth label is DS ("Disapprove scold protest disruptive behavior"). Indeed, the speech act DS occurs very few times in the training data (only 40 examples, i.e., less than 0.1% of the training data).

Table 2: Excerpt of a conversation from the New England Corpus (Child: Liam, Age: 14 months, Transcript: 99) with manually-annotated speech acts ("Manual") and predicted speech acts ("CRF"). Labels from the INCA-A tagset are listed in the Appendix.

	Speech act	
	Manual	CRF
Mother: We're having a little problem here in the corner. (Mother stands up) (Child unplugs cord from wall again)	ST	ST
Mother: Liam ! (Mother takes hold of Child's hand)	CL	CL
Mother: No! (Mother takes hold of cord and tries to pull it out of Child's hand, Child holds onto cord)	PF	PF
Mother: Let go. (Child lets go of cord, Mother plugs cord back into wall, Child watches what Mother does with cord)	RP	RP
Mother: No. (Mother picks up Child)	PF	PF
Mother: You're a nut.	DS	ST

Another factor that affects the model's performance is what appears to be ambiguities in the definition of some categories in the INCA-A coding scheme. In particular, many pairs of speech acts are either very similar or hierarchically related (see Cameron-Faulkner and Hickey (2011) for a similar observation). More concretely, there are pairs of speech act categories that describe overlapping communicative intents (e.g., "Criticize or point out error in nonverbal act" (CR) can overlap with "Disapprove scold protest disruptive behavior" (DS) and pairs of speech acts where the meaning of one act appears to be covered by the other broader act (e.g., the speech act "Praise for motor acts i.e for nonverbal behavior." (PM) is part of "Approve of appropriate behavior." (AB)). Such overlaps in the definition of some categories do not help the model make clear distinctions between the affected categories and, thus, tend to conflate them.

We provide an example for this phenomenon in Table 3. In this conversation, the mother's utterance "Good girl" is labelled by the CRF as "Approve of appropriate be-

havior." (AB), which is not incorrect, but differs from the human annotation, which categorizes it as "Praise for motor acts i.e for nonverbal behavior." (PM). We hypothesize that collapsing overlapping categories would improve the model performance. Indeed, we experimented with an alternative coding scheme where we collapsed certain categories and the model achieves a higher average performance of 75.35% ($\pm 4.17\%$) accuracy. However, for the remainder of this work, we continue using the original coding scheme to ensure comparability to the work of Snow et al. (1996).

Table 3: Excerpt of a conversation from the New England Corpus (Child: Joanna, Age: 20 months, Transcript: 32) with manually-annotated speech acts ("Manual") and predicted speech acts ("CRF"). Labels from the INCA-A tagset are listed in the Appendix.

	Speech act	
	Manual	CRF
Mother: Take it [= book] out of the box. (*The child struggles with both hands on the open book. Afterwards, the child pulls the book up and out of the box*)	RP	RP
Mother: Good girl.	PM	AB

Replicating Findings from Snow et al. (1996)

Here we validate the CRF model by testing its ability to lead to conclusions similar to the ones obtained in Snow et al. (1996). To this end, and as we mentioned earlier, we proceed in two steps: First, we replicate major findings in Snow et al. (1996) using their hand-annotated labels. Second, we compared them to the corresponding findings obtained using the labels that were predicted using our CRF model. In addition to replicating main analyses from Snow et al. (1996) (i.e., development of the size and distribution of speech acts), we also tested the models with a new, more specific task that consists of predicting the precise normative age of acquisition of speech acts in both production and comprehension.

Development of the Number of Distinct Speech Acts

Figure 3 shows the proportion of children producing a given number of different speech act types for the three age groups studied in Snow et al. (1996) (This is a direct replication of Figure 2 in the original paper). Next to each bar obtained from the hand-annotation (in blue) we plot the corresponding bar from the automatic labeling by CRF on the same dataset (in orange).

We can see that the patterns observed in Snow et al. (1996) are well captured by automatic labeling data: At 14 months, most children produce only a handful of speech act types, such as statements (ST), repetitions (RT) and markings (MK). This number increases on

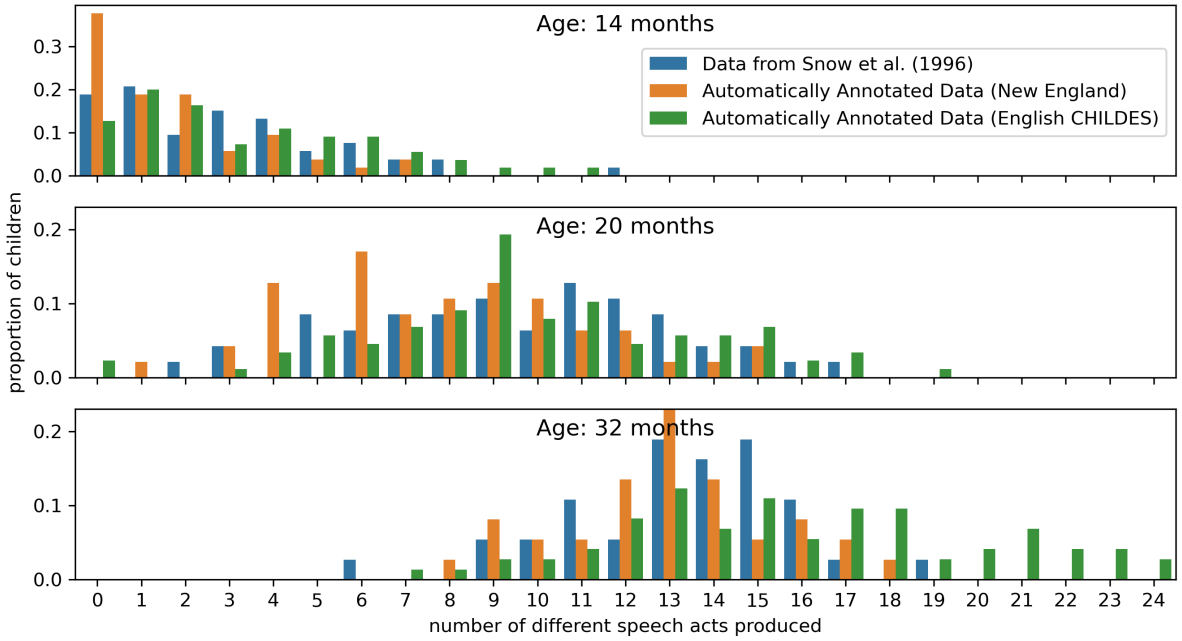

Figure 3. ***Proportion of children producing a given number of distinct speech act types at 14, 20, and 32 months old. Note that the y-axis for the bottom two figures has been shortened for better visibility.***

average for children aged 20 months where now a substantial proportion of children become able to produce around 10 different speech act types (now starting to use for example requests (RP), stating intent (ST) and product questions (QN)). Finally, at 32 months, children typically produce between 10 and 20 different speech act types (starting to use for example polar questions (YQ)). When compared to hand annotated data in the New England corpus, the model was able to capture not only the rough number of speech act types produced at each age range, it was also able to capture quite well the variability between children at each age.

We can quantify the similarity between the hand- and automatic-annotation-based distributions by computing their Jensen-Shannon distances. This measure quantifies the dissimilarity between two probability distributions with values ranging from 0 (maximally similar) to 1 (minimally similar). The similarities of distributions from manually and automatically annotated data were as follows: 0.262 (at 14 months), 0.367 (at 20 months), and 0.186 (at 32 months).

Development of the Distribution of Speech Acts

Figure 4 shows the replication of the analysis on the development of the distribution of speech acts (cf. Table 9 in Snow et al. (1996)). This analysis compares the proportions of

utterances that fall within each speech act category for the three age groups. Similar to the previous graph, next to each bar obtained from the hand-annotation (in blue) we plot the corresponding bar from the automatic labeling by CRF (in orange). We can see that the frequency distributions look remarkably similar in each age group (see Appendix for the legend of what each speech act label refers to). Jensen-Shannon distances of automatically annotated data (New England) compared to data from Snow et al. (1996) were: 0.089 (14 months), 0.103 (20 months), 0.080 (32 months).

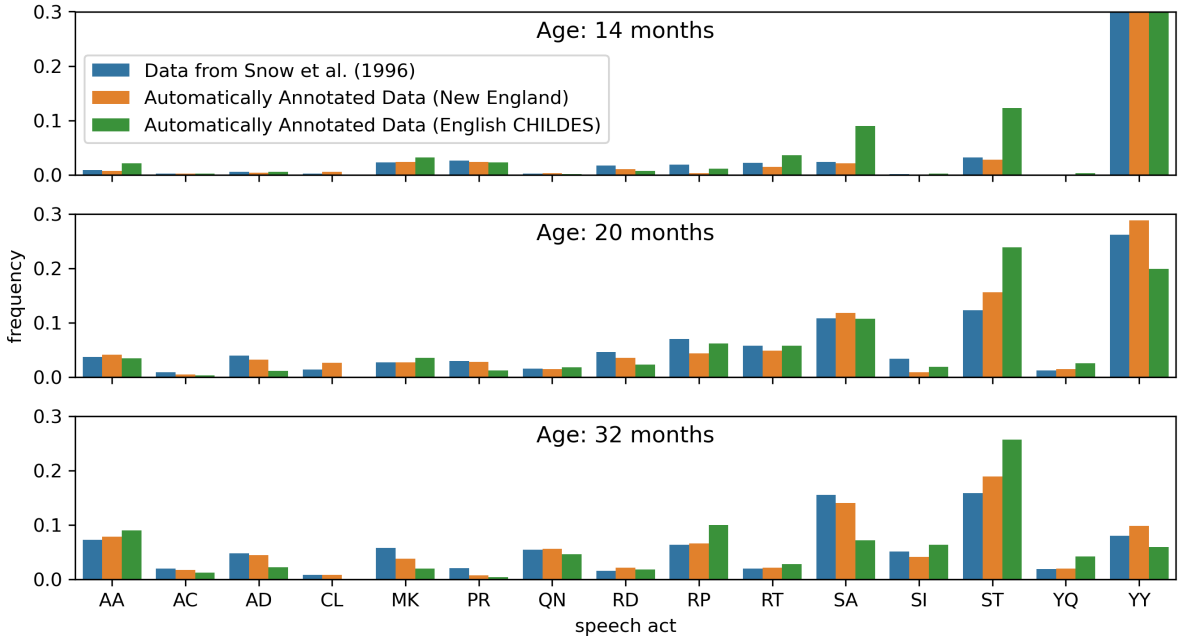

Figure 4. ***Frequency distribution of speech acts for different ages. Note that the y-axes have been trimmed for better visibility (The frequencies for*** `YY` ***at 14 months are around 0.6).***

Generalizing Findings to Data in CHILDES

In the previous subsection, we validated the model by comparing findings from predicted and hand-annotated labels of the same data. Here, we use the trained model to automatically annotate data from English corpora in CHILDES. The goal is to investigate the extent to which findings obtained in Snow et al. (1996) generalize to a larger number of children and to the variety of communicative contexts represented in these new corpora.

More precisely, we trained the CRF on the whole New England corpus (no held-out test set) and used it to annotate speech acts on transcripts of children aged between 14 to 32 months old in the North American English corpora of CHILDES (excluding transcripts

from the New England corpus). Next, we perform the same analyses as in the previous section using the large-scale annotated data.

Development of the Number of Distinct Speech Acts

The green bars in Figure 3 show the number of different speech act types produced by children from CHILDES. Developmental patterns are very similar to the original graphs (in orange), with the exception of the oldest age group (i.e., 32 months) where we found that more children produced a relatively larger number of different speech acts (more than 20). Jensen-Shannon distances of automatically annotated data (English CHILDES) compared to data from Snow et al. (1996) were: 0.209 (at 14 months), 0.222 (at 20 months), and 0.418 (at 32 months).

Development of the Distribution of Speech Acts

We present the frequency distribution of speech acts for children from CHILDES in the green bars of Figure 4. Again, patterns obtained by Snow et al. (1996) generalize very well. Jensen-Shannon distances of automatically annotated data (English CHILDES) compared to data from Snow et al. (1996): 0.204 (14 months), 0.173 (20 months), 0.197 (32 months).

Age of Acquisition of Speech Acts

In this section, we present results for the age of acquisition of speech acts in terms of production and comprehension using the measures defined in the Section "Measures of Speech Act Emergence".

Production

We calculated the age of acquisition for a subset of 25 speech acts[10] using both the manually-annotated labels from Snow et al. (1996) and the automatically generated labels from the CRF on the same dataset. Examples for regression plots and predicted ages of acquisition for all speech acts can be found in the appendix. Then, we calculated the Spearman rank-order correlation[11] to examine whether the *order* of emergence of speech acts is correctly captured by the automatically annotated data.

[10]These were the ones for which we could fit a logistic regression using at least two data points. While the number of acts we keep may seem small compared to the original size (65 possible speech acts excluding categories for unintelligible speech acts, `YY` and `OO`), it is due to the fact that the frequency distribution is highly skewed: Most categories occurred rarely in the corpus (Figure 1) and therefore did not provide enough data to be used in the calculation of age of acquisition.

[11]The rank-order correlation was computed over the subset of 25 speech acts for which an age of acquisition could be calculated, details in the Appendix.

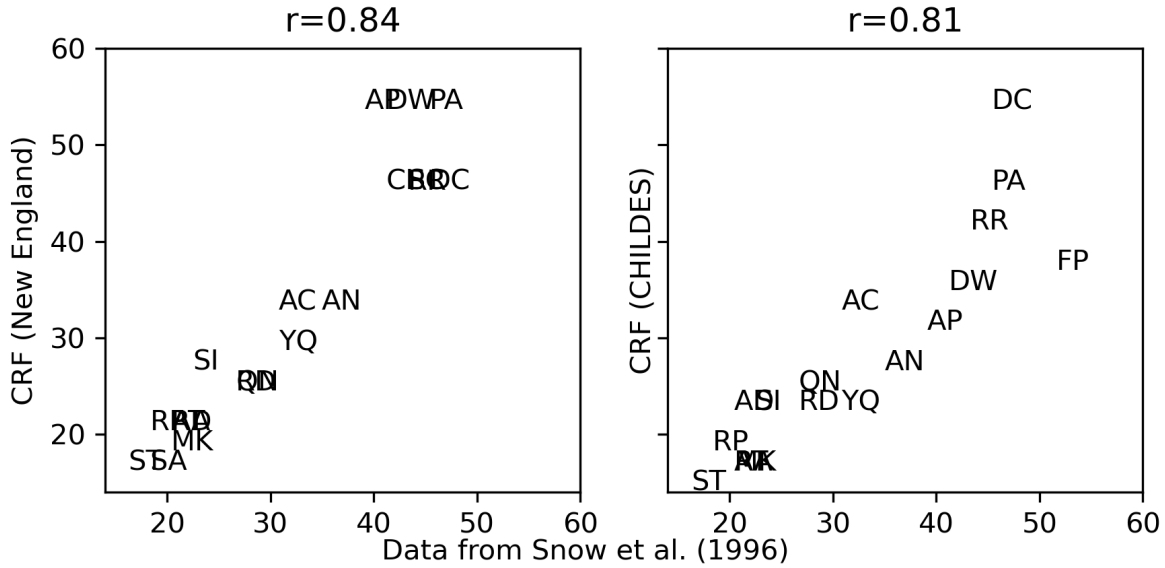

Figure 5. ***Correlation of age of acquisition in terms of production as calculated using data from Snow et al. (1996) and automatically annotated data for the New England corpus and CHILDES. Note that some speech acts are not displayed because the axes limits were set to 60 months for better visibility of early development. However, the correlation was calculated for all values.***

The resulting high correlation (see Figure 5 (left); $r \approx 0.84$, $p < 1 \cdot 10^{-6}$) indicates that the automatically generated labels can provide reasonable estimates for the developmental trajectory of speech acts.

We also calculated ages of acquisition using the predicted labels on CHILDES data. Figure 5 (right) shows the correlation with the ages calculated using New England data. Spearman rank-order correlation was $r \approx 0.81$ ($p < 1 \cdot 10^{-6}$).

Comprehension

To illustrate the emergence of speech acts in terms of comprehension, we first show observed adjacency pairs for adult-child turns for different ages in Figure 6. The youngest children respond with unintelligible utterances or utterances without clear function (YY, OO) in most of the cases displayed. Children at 20 months show some consistent patterns in their response behavior: Polar and product questions (YQ, QN) are answered with adequate responses (AA, SA). Polite requests (RQ) are either accepted (AD) or refused (RD). Requests or suggestions (RP) are also usually accepted or refused, although in some cases children answer with a statement (ST), which is not contingent. Additionally, there is still a large amount of utterances without clear function (YY). Only by the age of 32 months, most of the parents' utterances are addressed with contingent responses (at

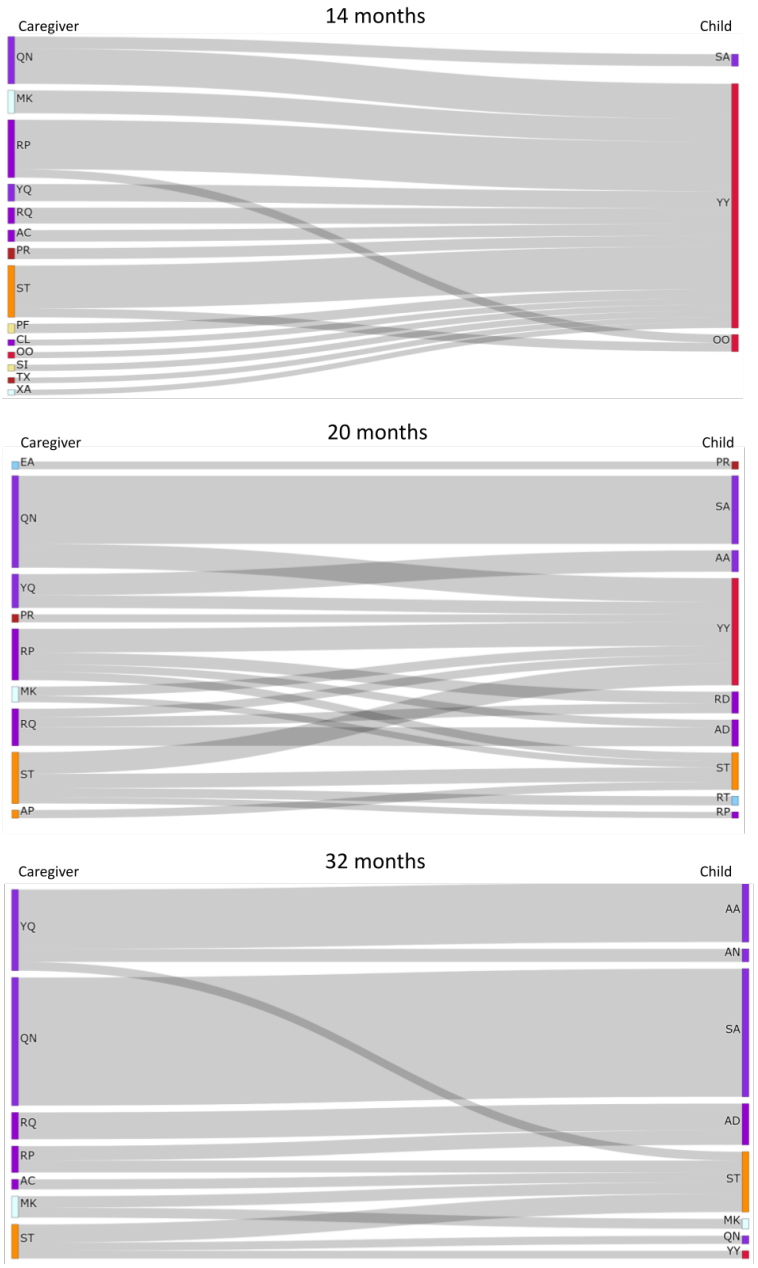

Figure 6. ***Adjacency pairs of speech acts for children of 14, 20, and 32 months. Utterances by the caregiver are on the left, responses by the children on the right. Filtered to display speech acts that occur in at least 0.01% of the data for better visibility. The colors indicate the higher-level interchange type for each speech act (see Snow et al., 1996).***

least as captured at the broad level of speech act categories).

Examples for predicted ages of acquisition for all speech acts can be found in the appendix. We observe that while there are similar trajectories in production and comprehension for some speech acts (e.g. RR), we also observed some striking differences in other cases. For example, "demands for permission" (FP) is produced very late (around 52 months), but they are already understood a lot earlier (around 14 months).

As done for the production measure, we calculated the age of acquisition using both the ground-truth labels from Snow et al. (1996) and the automatically generated labels from the CRF on the same dataset, as well as using generated labels on the English CHILDES data. As in production, the Spearman rank-order correlation coefficient[12] (see Figure 7, left; $r \approx 0.46$, $p < 0.01$) indicates a statistically significant positive correlation (however lower than for the production measure). For the correlation with predicted labels on CHILDES data, the Spearman rank-order correlation was $r \approx 0.63$ ($p < 1 \cdot 10^{-5}$; see Figure 7, right).[13]

Figure 8 shows the full distribution of age of emergence in both production and comprehension. It shows that, overall, comprehension of speech acts precedes their production. Indeed, a paired t-test (using only speech acts for which we could calculate an age of acquisition both in production and in comprehension) shows a mean difference of 2.51 months ($p < 0.05$).[14]

Finally, we ask how the trajectory of emergence in comprehension compares to that of production. For instance, does production follow the same pattern/order of comprehension, only delayed? Pearson's correlation between the two developmental trajectories is $r \approx -0.07$ ($p \approx 0.76$), indicating that speech acts emerge differently in production and comprehension, and suggesting that these two dimensions of development may be explained by different factors.

[12]The rank-order correlation was computed over the subset of 47 speech acts for which an age of acquisition in terms of comprehension could be calculated, i.e. cases in which we could fit a logistic regression using at least two data points, details in the Appendix.

[13]As we said above, we chose to fit the age of acquisition using logistic regressions following the method used for the AoA of words Frank et al. (2021). The main limitation here was the sparsity of available annotated data: The study by Snow et al. (1996) only considers 3 different age groups: Children at 14, 20, and 32 months. While the fitted curves were good for production, this was less obvious for comprehension data based on contingency (see the graphs in the appendix). Note, however, that for our analysis, i.e., correlating AoA from predicted vs. hand-annotated speech acts (Figures 6 and 7), we only needed the ranking of AoA, not necessarily absolute values of ages. So, one simple way to test the robustness of these correlations is the following: Instead of estimating the AoA using logistic regressions, we can estimate the ranking without fitting any model and directly from the data. More specifically, we computed the proportion of children that produced (or understood) a given speech act (averaged over the three-time points) and ranked the speech acts according to these proportions as a proxy for their order of acquisition. The resulting rank-order correlations obtained using this model-free method were very close to the correlations found using the regression method, thus corroborating these findings.

[14]When using the alternative coding scheme with collapsed speech act categories (see Section "Error analysis"), this difference increases to 9.61 months ($p < 0.01$).

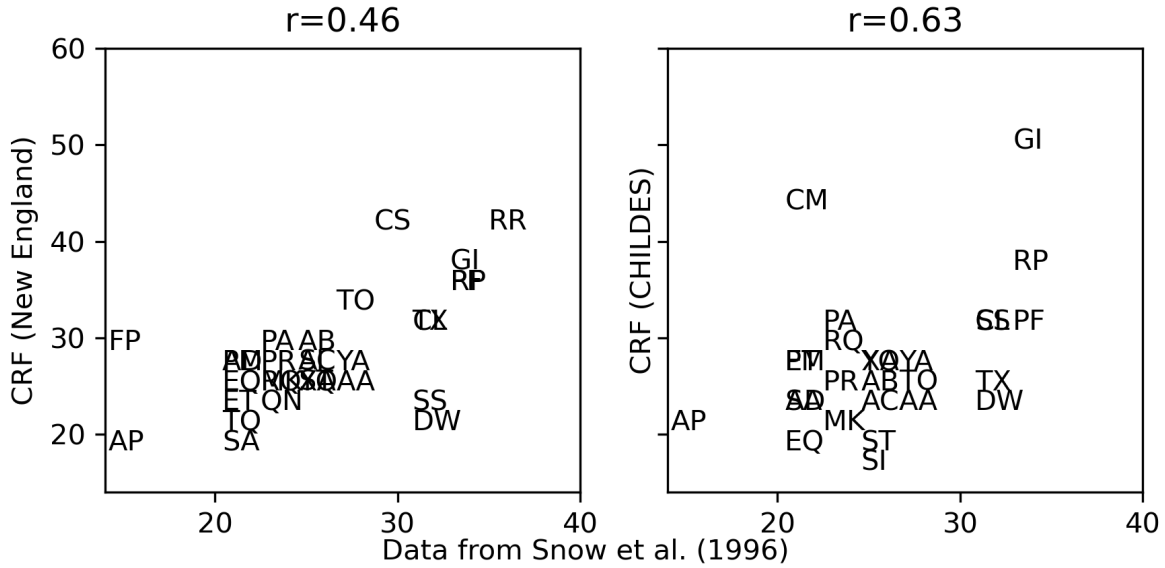

Figure 7. ***Correlation of age of acquisition in terms of comprehension as calculated using data from Snow et al. (1996) and automatically annotated data for the New England corpus and CHILDES. Note that some speech acts are not displayed because the axes limits were set to 60/40 months for better visibility of early development. However, the correlation was calculated for all values.***

Development of Speech Acts Beyond 32 Months

Since CHILDES contains data for children beyond the age range studied in Snow et al. (1996), we could also make predictions about the age of acquisition of some speech acts that could not be calculated using the New England corpus because they were not yet acquired by children by 32 months. To this end, we use all transcripts up to 54 months (data become sparse beyond that age). Using this larger set of annotations, we can for example estimate the age at which children produce speech acts such as prohibitions (`PF`, at 84.9 months), give reason (`GR`, at 87.0 months), polite requests (`RQ`, at 66.2 months), and make promises (`PD`, at 130.7 months)). These predictions are consistent with the developmental literature showing a late acquisition of some of these speech acts (Matthews, 2014). A table of all results can be found in the Appendix.

Discussion

The way children master language use in social interaction is an important frontier in the study of language development (Bloom & Lahey, 1978; Casillas & Hilbrink, 2020; Clark, 2018; Matthews, 2014; Snow et al., 1996). Answering this question has also the potential for impact in clinical applications (e.g., early and automatic detection of communicative difficulties). However, the investigation of this phenomenon in ecological valid settings

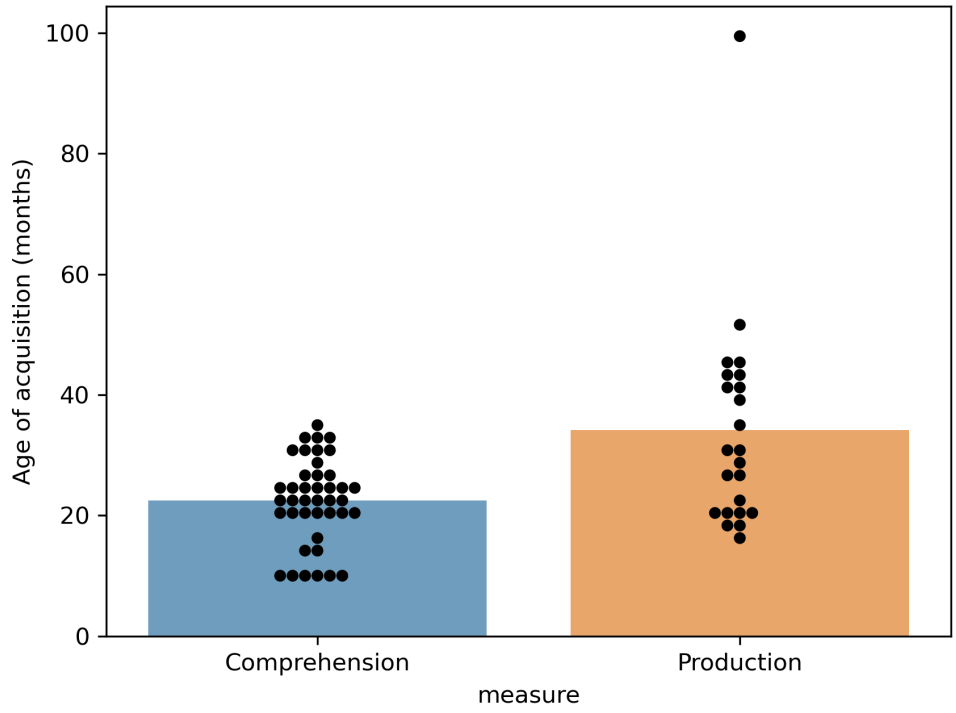

Figure 8. ***The distribution of the speech acts' age of emergence in comprehension and production.***

requires complex, large-scale data annotation which is prohibitively expensive to do by hand only.

In the current work, we introduced a simple model that allows for reliable *automatic* labeling of major speech act categories in the context of child-caregiver social interactions. We trained the model on a dataset that was previously hand-annotated using INCA-A, a comprehensive coding scheme for speech acts in early childhood (Ninio et al., 1994; Snow et al., 1996). When tested on parts of the data it had not seen in the training, the model predicted speech acts that captured quite well the major findings reported in this earlier work such as the average trajectory of speech act development and the patterns of variations between children.

Besides providing a valuable tool that we make available to the community, a major theoretical contribution of the paper was testing how earlier findings — obtained using hand annotation of a small number of children — generalize to a larger and different sample. We tested this generality by automatically labeling the entire American English section of CHILDES for speech acts. We found that, across all major analyses, children

show, overall, patterns that were very similar to the ones reported by (Snow et al., 1996). The main difference was that older children in the larger dataset produced noticeably more speech act types than children of similar age in the original study (Figure 3, bottom). This difference could be due to the fact that the larger dataset contains a richer set of conversational contexts, giving children the opportunity to perform more distinct speech act types.[15]

Anther contribution of this work is the introduction of two measures to quantify the age of emergence of speech acts in children's production and comprehension. We found that these two measures (i.e., comprehension and production) did not correlate, indicating that they provide non-redundant information about development and suggesting that speech acts may develop differently in production and comprehension. In particular, factors that would be relevant for learning in production may not necessarily be the same in comprehension, especially in the rather *asymmetrical* context of child-caregiver interactions.

To illustrate, take the case of "Yes/no requests" (`RQ`) vs. "yes/no questions for information." (`YQ`). In production, we replicated Snow et al. (1996)'s finding that children produce yes/no questions as requests later than yes/no questions for information (very few children produced the first act and only at 32 months). This fact is also in line with the literature on politeness which suggests that children produce polite requests quite late (Axia & Baroni, 1985). Interestingly however, in comprehension we found that on average children responded contingently to the yes/no requests at about the same age as they do to yes/no questions for information.

When using automatically annotated data from our model, we found that their predicted measures of age of acquisition correlated to a high degree with the ages of acquisition predicted from manually labelled data, especially in production. In a direct application, the model allowed us to estimate the age of acquisition of some late emerging speech acts (e.g., "promise" and "give reason") thanks to automatic labeling of new data children that were older in CHILDES than in the original New England corpus.

While the automatic labelling model provides a high average accuracy score, the per-label scores showed high variability. While, as we argued above, some of this variability can be explained by the frequency of occurrence in the training data and by ambiguities in the definition of some categories in the coding scheme, we speculate that other factors could be in play as well, especially the *linguistic variability* with which a speech

[15] Another observation was that the proportion of children producing no speech acts (i.e., 0 in Figure 3) at 14 months is noticeably higher in the automatically annotated data than in the original data. This means that our model classified more utterances as unintelligible or utterance without function than the human annotators. We hypothesize that the highly skewed distribution of speech acts in the dataset for children at this age, with many (but not all) utterances actually being without clear function, leads the model to overfit to this case and miss some actually meaningful utterances.

act can be expressed.[16]

For example, there is a variety of ways one can express the act of "giving reasons" (GR) in linguistic terms, which makes it relatively hard to recognize based only on the linguistic features of its instances (F-score = 0.3). In comparison, the set of linguistic terms typically used to express, say, the act of "requesting repetition" (RR) or "eliciting question" (EQ) is much more constrained, making their recognition easier (F-scores are 0.53 and 0.81, respectively), although all three categories have roughly similar (low) frequency of occurrence in the data. Take also the case of "stating intent" (SI) and "prohibiting" (PF). Both of these speech acts are similarly frequent (around 300 occurrences), but the F-score for PF is much higher than the one for SI (0.76 and 0.43, respectively). This difference could also be due to the fact that "prohibiting" is much more constrained linguistically than "stating intent."

Researchers have made a similar argument about the role that linguistic variability can have on their learnability by children (e.g. Bloom & Lahey, 1978). This analogy is to be taken with a grain of salt though. More generally, it is not warranted to make a direct link between the learnability of speech act categories by our model and their learnability by children: In the first case, the model was aimed at optimizing prediction accuracy and had been trained on labeled data. In the second case, children learn without having access to the true labels of the utterances. Models that aim at "discovering" categories in an unsupervised fashion are more likely to be insightful about the learnability of speech act categories by children (e.g. Bergey et al., 2021).

Limitations and Future Work

Our model learns how to recognize speech acts from their linguistic instances only. While the scores were quite good and allowed us to replicate major findings that were obtained using human annotations, future work should seek to build more comprehensive models that integrate multimodal cues — besides verbal language — that likely play a role in signaling communicative intents including vocal and visual cues (e.g. Fernald, 1989; Senju & Csibra, 2008; Tomasello et al., 1997; Trujillo et al., 2018). This effort will involve collecting multimodal data of spontaneous child-caregiver conversations (e.g. Bodur et al., 2021) as well as the development of machine learning methods for the automatic annotation of speech acts using linguistic, acoustic, and visual features.

Another limitation concerns the measures we used to quantify the age of acquisition. While it is easier to quantify acquisition through production, it is trickier to have a perfect measure of comprehension in a natural, uncontrolled context. Here, we provided a contingency-based measure. Such an operationalization has allowed us to uncover new

[16]Indeed, the higher the variability within a given category, the more examples the model needs to learn it.

interesting phenomena (namely that children understand some speech act before they produce them).

However, measuring contingency is a notoriously difficult task, especially in a naturalistic setting and with verbal data only. First, responses can be contingent in various ways: For example, asking a yes-no question like "Do you want a banana?" can be followed by many speech acts that can all be contingent such as "Yes!", "I just ate one", or "now?". Other speech acts such as declarative statements do not necessarily require a response, so the listener might understand the communicative intent without necessarily giving a response. In this work, we partly avoided these difficulties by using a broad binary annotation that judged whether a response was possibly contingent or totally inappropriate (e.g., a "greeting" after a "yes-no question").

In addition to these theoretical difficulties, there are practical difficulties related to the fact that children (especially the younger ones) may respond contingently but in a non-verbal fashion (a case that is not captured by the current model). Besides, they sometimes respond in an unintelligible fashion (a case which we had to classify as non-contingent). Another case is when they do not respond at all (leading to more data exclusion). However, when children do not respond (e.g., after being asked a question), it does not necessarily mean that they did not understand the speech act. For example, children may lack the appropriate vocabulary to formulate an adequate response or they may just not be interested in following up.

Finally, we did not take into account the timing of responses (as several CHILDES corpora lack timestamps in the transcripts). This is important, because if a child's response only follows a caregiver's utterance after a long temporal delay, it may not be an actually response, but a new initiation. Thus, it would not be appropriate to judge the contingency of this "response" with respect to the caregiver's utterance that preceded it.

All these reasons may contribute to making our contingency measure *under-estimate* children's early age of comprehension. That is, it is very likely that children understand many speech acts at a much earlier age than what we report in this work. That said, some results using this measure, especially the fact that comprehension precedes production in some categories, would still hold. In fact, if anything, a more accurate measure of comprehension would just make such conclusions stronger.

Finally, we found several limitations the INCA-A coding scheme when automatically labeling utterances, including overlapping as well as hierarchically related categories (cf. the error analyses section as well as Cameron-Faulkner (2014) for similar observations). In the future, the coding scheme should be updated in order to make it less ambiguous for automatic annotation.

To conclude, this work has introduced both novel research tools and measures that we hope will pave the way to a more quantitative approach to the study of children's speech act development in the wild.

References

Axia, G., & Baroni, M. R. (1985). Linguistic politeness at different age levels. *Child Development*, 918–927.

Bergey, C., Marshall, Z., DeDeo, S., & Yurovsky, D. (2021). Learning communicative acts in children's conversations: A hidden topic markov model analysis of the childes corpus. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*.

Bloom, L., & Lahey, M. (1978). *Language development and language disorders.*

Bodur, K., Nikolaus, M., Kassim, F., Prévot, L., & Fourtassi, A. (2021). Chico: A multimodal corpus for the study of child conversation. *Proceedings of the 23rd International Workshop on Corpora and Tools for Social Skills Annotation.*

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. (2016). From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. *CogSci*.

Cameron-Faulkner, T. (2014). The development of speech acts. *Pragmatic development in first language acquisition*, 37–52.

Cameron-Faulkner, T., & Hickey, T. (2011). Form and function in irish child directed speech. *Cognitive Linguistics*, *22*(3), 569–594.

Casillas, M., & Hilbrink, E. (2020). 3. communicative act development. *Developmental and Clinical Pragmatics, 13,* 61.

Clark, E. V. (2018). Conversation and language acquisition: A pragmatic approach. *Language Learning and Development, 14*(3), 170–185.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message?. *Child Development, 60*(6), 1497–1510.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The wordbank project*. MIT Press.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? parental input and the acquisition of vocabulary. *Journal of child language, 35*(3), 515–531.

Grice, H. P. (1975). Logic and conversation. *Speech acts* (pp. 41–58). Brill.

Khanpour, H., Guntakandla, N., & Nielsen, R. (2016). Dialogue act classification in domain-independent conversations using a deep recurrent neural network. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2012–2021.

Kumar, H., Agarwal, A., Dasgupta, R., & Joshi, S. (2018). Dialogue act sequence labeling using hierarchical encoder with crf. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1).

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk. transcription format and programs* (Vol. 1). Psychology Press.

MacWhinney, B. (2017). Tools for analyzing talk part 1: The chat transcription format.

Matthews, D. (2014). *Pragmatic development in first language acquisition* (Vol. 10). John Benjamins Publishing Company.

Ninio, A., Snow, C. E., Pan, B. A., & Rollins, P. R. (1994). Classifying communicative acts in children's interactions. *Journal of communication disorders, 27*(2), 157–187.

Okazaki, N. (2007). Crfsuite: A fast implementation of conditional random fields (crfs). http://www.chokkan.org/software/crfsuite/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Rollins, P. R. (1999). Early pragmatic accomplishments and vocabulary development in preschool children with autism. *American Journal of Speech-Language Pathology, 8*(2), 181–190.

Rollins, P. R. (2017). Pathways early intervention program for toddlers with autism. *Journal of Menatl Health and Clinical Psychology, 1*(1).

Schegloff, E. A., & Sacks, H. (1973). Opening up closings. *Semiotica, 8*(4), 289–327.

Searle, J. R. (1976). A classification of illocutionary acts. *Language in society*, 1–23.

Senju, A., & Csibra, G. (2008). Gaze following in human infants depends on communicative signals. *Current biology*, *18*(9), 668–671.

Snow, C. E., Pan, B. A., Imbens-Bailey, A., & Herman, J. (1996). Learning how to say what one means: A longitudinal study of children's speech act use. *Social Development*, *5*(1), 56–84.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V., & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, *26*(3), 339–373.

Tomasello, M., Call, J., & Gluckman, A. (1997). Comprehension of novel communicative signs by apes and human children. *Child Development*, *68*(6), 1067–80.

Trujillo, J. P., Simanova, I., Bekkering, H., & Özyürek, A. (2018). Communicative intent modulates production and comprehension of actions and gestures: A kinect study. *Cognition*, *180*, 38–51.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S., et al. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Data, Code and Materials Availability Statement

Source code of all models and experimentation scripts are publicly available at: https://github.com/mitjanikolaus/childes-speech-acts.

The data is publicly available as part of the CHILDES corpora. Data for the New England corpus has been directly downloaded from the CHILDES database: https://childes.talkbank.org/access/Eng-NA/NewEngland.html. Data for all other corpora has been accessed using childes-db: https://langcog.github.io/childes-db-website/.

Authorship and Contributorship Statement

M.N., E.M., J.A. and A.F. designed research; M.N. and E.M. performed research and analyzed data; and M.N., E.M., J.A., L.P., and A.F. wrote the paper.

Acknowledgements

Research supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI), ANR-21-CE28-0005-01 (MACOMIC), AMX-19-IET-009 (Archimedes Institute) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

Appendix

INCA-A Tagset

Speech acts of the INCA-A coding scheme (Ninio et al., 1994) are listed in Table 4.

Table 4: Speech acts of the INCA-A tagset.

Speech Act	Description
AA	Answer in the affirmative to yes/no question.
AB	Approve of appropriate behavior.
AC	Answer calls/ show attentiveness to communications.
AD	Agree to carry out an act requested or proposed by other.
AL	Agree to do something for the last time.
AN	Answer in the negative to yes/no question
AP	Agree with proposition or proposal expressed by previous speaker
AQ	Aggravated question expression of disapproval by restating a question
CL	Call attention to hearer by name or by substitute exclamations
CM	Commiserate express sympathy for hearer's distress.
CN	Count.
CR	Criticize or point out error in nonverbal act.
CS	Counter-suggestion/ an indirect refusal.
CT	Correct provide correct verbal form in place of erroneous one.
CX	Complete text if so demanded.
DC	Create a new state of affairs by declaration
DP	Declare make-believe reality.
DR	Dare or challenge hearer to perform an action.
DS	Disapprove scold protest disruptive behavior.
DW	Disagree with proposition expressed by previous speaker.
EA	Elicit onomatopoeic or animal sounds.
EC	Elicit completion of word or sentence.
ED	Exclaim in disapproval.
EI	Elicit imitation of word or sentence by modelling or by explicit command
EM	Exclaim in distress pain.
EN	Express positive emotion.
EQ	Eliciting question (e.g. hmm?).
ES	Express surprise.
ET	Express enthusiasm for hearer's performance.
EX	Elicit completion of rote-learned text.
FP	Ask for permission to carry out act.
GI	Give in/ accept other's insistence or refusal.
GR	Give reason/ justify a request for an action refusal or prohibition
MK	Mark occurrence of event (thank greet apologize congratulate etc.).
NA	Intentionally nonsatisfying answer to question
ND	Disagree with a declaration.
OO	Unintelligible vocalization.
PA	Permit hearer to perform act.
PD	Promise.
PF	Prohibit/forbid/protest hearer's performance of an act

PM	Praise for motor acts i.e for nonverbal behavior.
PR	Perform verbal move in game.
QA	Answer a question with a wh-question.
QN	Ask a product-question (wh-question)
RA	Refuse to answer.
RD	Refuse to carry out an act requested or proposed by other.
RP	Request propose or suggest an action for hearer or for hearer and speaker.
RQ	Yes/no question or suggestion about hearer's wishes and intentions
RR	Request to repeat utterance.
RT	Repeat or imitate other's utterance.
SA	Answer a wh-question with a statement.
SC	Complete statement or other utterance in compliance with request.
SI	State intent to carry out act by speaker.
SS	Signal to start performing an act such as running or rolling a ball
ST	Make a declarative statement.
TA	Answer a limited-alternative question.
TD	Threaten to do.
TO	Mark transfer of object to hearer
TQ	Ask a limited-alternative yes/no question.
TX	Read or recite written text aloud.
WD	Warn of danger.
WS	Express a wish.
XA	Exhibit attentiveness to hearer.
YA	Answer a question with a yes/no question.
YD	Agree to a declaration.
YQ	Ask a yes/no question.
YY	Make a word-like utterance without clear function.

Model Details

Hyperparameters

The models were trained until convergence on a held-out dev set (10% of the training data). A small set of hyperparameter configurations based on best practices were evaluated in preliminary experiments. The configuration listed in Table 5 led to the best results.

The learning rate for training the BERT-based model is substantially lower than for the other model as this model is already pre-trained and we are only fine-tuning it on the task.

Table 5: Model hyperparameters

Hierarchical LSTM + CRF	
vocabulary size	1000
word embeddings size	200
word-level LSTM hidden layer size	200
utterance-level LSTM hidden layer size	100
dropout	0.2
optimizer	Adam
initial learning rate	0.0001
+ BERT	
same as above, except for:	
initial learning rate	0.00001

Architecture

A high-level overview of the architecture of the hierarchical LSTM+CRF model can be found in Figure 9.

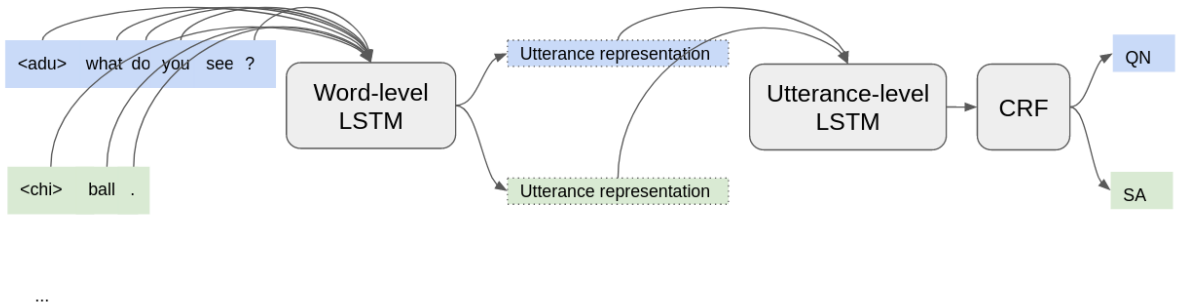

Figure 9. ***Architecture of the Hierarchical LSTM + CRF model.***

Error Analysis

Table 6 contains per-label precision, recall, and F1-scores for a model trained on 80% of the New England corpus and tested on the remaining 20%.

Table 6: Error analysis

	precision	recall	f1-score	support
AA	0.628	0.628	0.628	148
AB	0.690	0.454	0.547	108
AC	0.603	0.527	0.562	245
AD	0.674	0.651	0.662	229
AL	0.000	0.000	0.000	1
AN	0.625	0.571	0.597	35
AP	0.658	0.603	0.629	239
CL	0.800	0.875	0.836	160
CM	0.375	0.231	0.286	13
CN	0.200	0.500	0.286	4
CR	0.000	0.000	0.000	13
CS	0.273	0.086	0.130	35
CT	0.529	0.138	0.220	65
DC	0.750	0.316	0.444	19
DP	0.000	0.000	0.000	8
DS	0.375	0.273	0.316	11
DW	0.633	0.404	0.494	47
EA	0.974	0.884	0.927	43
EC	0.857	0.429	0.571	14
ED	1.000	0.333	0.500	15
EI	0.632	0.800	0.706	15
EM	0.000	0.000	0.000	1
EQ	0.750	0.849	0.796	53
ET	0.739	0.459	0.567	37
EX	0.000	0.000	0.000	1
FP	0.833	0.694	0.758	36
GI	0.375	0.158	0.222	19
GR	0.350	0.226	0.275	31
MK	0.733	0.814	0.772	996
NA	0.000	0.000	0.000	30
ND	0.000	0.000	0.000	1
PA	0.600	0.409	0.486	22
PD	0.800	0.211	0.333	19
PF	0.830	0.702	0.761	272
PM	0.518	0.345	0.414	84
PR	0.769	0.652	0.706	296
QN	0.940	0.958	0.949	1104
RD	0.679	0.494	0.571	77
RP	0.797	0.786	0.791	1689
RQ	0.830	0.848	0.839	506
RR	0.448	0.714	0.550	42

RT	0.467	0.340	0.394	144
SA	0.782	0.662	0.717	417
SC	1.000	0.455	0.625	11
SI	0.551	0.405	0.466	309
SS	0.811	0.664	0.730	116
ST	0.690	0.791	0.737	1620
TA	0.000	0.000	0.000	3
TO	0.333	0.222	0.267	72
TQ	1.000	0.200	0.333	10
TX	0.818	0.863	0.840	73
WD	0.875	0.700	0.778	10
XA	0.671	0.464	0.548	110
YA	0.769	0.408	0.533	49
YD	0.000	0.000	0.000	5
YQ	0.715	0.772	0.742	705
macro avg	0.567	0.446	0.479	10437
weighted avg	0.738	0.725	0.726	10437

Ages of Acquisition

Regression Plots

The regression plots in Figure 10 and 11 illustrate the proportion of children producing a given speech act (in the case of comprehension, the proportion of contingent responses made by children) across time as well as the best logistic fits used to predict the speech acts' precise age of acquisition. We depict only 6 exemplary speech acts for better readability. The data to create these plots was the original annotation data from Snow et al. (1996).

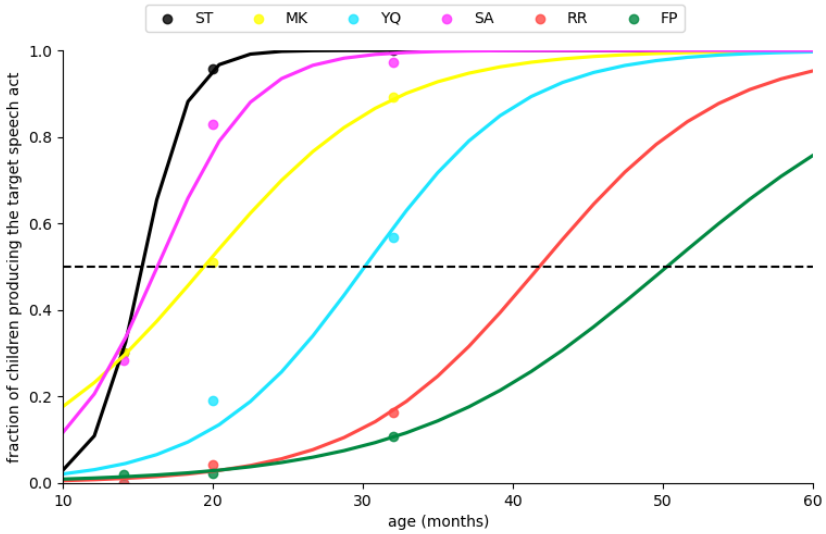

Figure 10. ***Regression plot for production.***

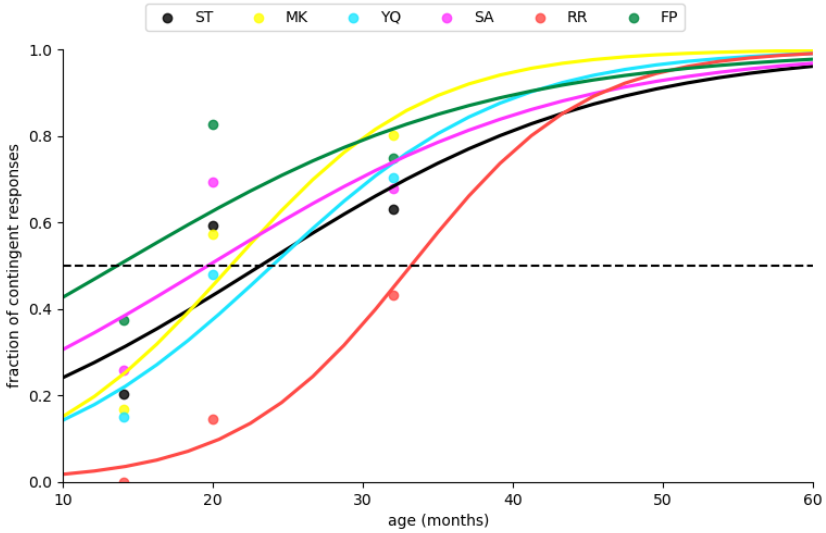

Figure 11. ***Regression plot for comprehension.***

Predicted Ages of Acquisition

The following tables show the age of acquisition (in months) for speech acts calculated using different data sources ("-" indicates that no age of acquisition could be calculated, i.e. at no observed time the proportion of children producing the speech act surpassed 0.5). We calculated the ages of acquisition in terms of production (Table 7) and comprehension (Table 8).

Table 7: Predicted ages of acquisition for production.

Speech act	Snow	CRF	CHILDES
AA	20.4	20.4	16.2
AC	30.8	32.9	32.9
AD	20.4	22.5	22.5
AN	35.0	30.8	26.6
AP	39.1	47.5	30.8
CL	41.2	45.4	70.3
CS	99.5	-	39.1
DC	45.4	45.4	53.7
DW	41.2	64.1	35.0
FP	51.6	-	37.1
MK	20.4	18.3	16.2
PA	45.4	45.4	45.4
PF	-	43.3	35.0
PR	28.7	-	-
QN	26.6	24.6	24.6
RD	26.6	24.6	22.5
RP	18.3	20.4	18.3
RR	43.3	39.1	41.2
RT	20.4	20.4	16.2
SA	18.3	16.2	10.0
SC	43.3	53.7	-
SI	22.5	26.6	22.5
ST	16.2	16.2	14.2
TO	-	35.0	37.1
YQ	30.8	28.7	22.5

Table 8: Predicted ages of acquisition for comprehension.

Speech act	Snow	CRF	CHILDES
AA	26.6	24.6	22.5
AB	24.6	35.0	24.6
AC	24.6	26.6	22.5
AD	20.4	24.6	22.5
AN	-	-	-
AP	14.2	20.4	20.4

AQ	-	-	-
CL	30.8	35.0	30.8
CM	20.4	-	43.3
CN	-	-	-
CR	-	-	-
CS	28.7	30.8	10.0
CT	16.2	16.2	10.0
DC	10.0	-	24.6
DS	-	-	-
DW	30.8	10.0	22.5
EA	10.0	12.1	10.0
EC	-	-	-
EI	10.0	22.5	10.0
EQ	20.4	22.5	18.3
ET	20.4	24.6	26.6
FP	14.2	28.7	10.0
GI	32.9	32.9	49.5
GR	10.0	26.6	20.4
MK	22.5	22.5	20.4
PA	22.5	28.7	30.8
PD	10.0	59.9	10.0
PF	32.9	26.6	30.8
PM	20.4	26.6	26.6
PR	22.5	24.6	24.6
QN	22.5	22.5	10.0
RD	10.0	-	-
RP	32.9	35.0	37.1
RQ	22.5	26.6	28.7
RR	35.0	39.1	99.5
RT	22.5	10.0	10.0
SA	20.4	18.3	22.5
SI	24.6	26.6	16.2
SS	30.8	22.5	30.8
ST	24.6	24.6	18.3
TO	26.6	35.0	24.6
TQ	20.4	12.1	10.0
TX	30.8	28.7	24.6
WD	24.6	-	87.0
XA	24.6	24.6	26.6
YA	26.6	28.7	26.6
YQ	24.6	24.6	26.6

Predicted Ages of Acquisition Including Data of Older Children

Table 9 presents the ages of acquisition in terms of production including data from older children (up to 54 months). We show only speech acts for which the age of acquisition could be calculated, i.e. for which at some age the proportion of children producing the speech act surpassed 0.5 .

Table 9: Predicted ages of acquisition including older children

Speech act	Age of acquisition
AA	18.3
AC	45.4
AD	32.9
AN	41.2
AP	101.6
AQ	155.7
CL	136.9
CN	95.3
CR	141.1
CS	149.4
DP	107.8
DW	76.6
EA	91.2
EI	164.0
EM	180.6
EQ	93.2
FP	78.7
GR	87.0
MK	16.2
PA	139.0
PD	130.7
PF	84.9
QN	35.0
RD	39.1
RP	10.0
RQ	66.2
RR	66.2
RT	10.0
SA	10.0
SI	20.4
ST	10.0
TA	95.3
TQ	62.0
YA	188.9
YQ	26.6

License