Large-scale study of speech acts' development using automatic labelling

Mitja Nikolaus[*,1,2] (mitja.nikolaus@univ-amu.fr)
Juliette Maes[*,2] (juliette.maes@etu.univ-amu.fr)
Jeremy Auguste[1] (jeremy.auguste@lis-lab.fr)
Laurent Prévot[2] (laurent.prevot@univ-amu.fr)
Abdellah Fourtassi[1] (abdellah.fourtassi@gmail.com)

[1]Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
[2]Aix-Marseille Univ, CNRS, LPL, Aix-en-Provence, France
[*]equal contribution

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. Our model will be 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: first language acquisition; speech acts; automatic annotation

Introduction

Research on language learning has mostly focused on investigating how children acquire language structures (e.g., phonology, lexicon, and syntax). 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. 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, etc.) (Searle, 1976)

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. 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. For example, they showed that when children utter their first words, they already express a rich repertoire of communicative intents such as requests and questions. As their social-cognitive and linguistic skills develop (e.g., in terms of Theory of Mind), they become able to express more sophisticated speech acts such as promise, deceive, and persuade. Using the same coding scheme, Rollins (1999, 2017) has shown, in a different work, that the study of speech act development can also help us study atypical cognitive development such as autism.

While this previous effort has been very influential in the study of speech act 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 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.

In order to test how previous findings in speech act development generalize at scale, we proceeded in two steps: First, we validated the chosen model by testing its ability to replicate key findings from Snow et al. (1996). 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.

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 consisited 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. generalize to a larger dataset of children and across different international 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 totalling 354 children.[1]

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, more commonly known as the speech act. 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.[2]

Automatic Classification of Speech Acts

While early work used Hidden Markov Models (Stolcke et al., 2000), later work showed large performance improvements by using Recurrent Neural Networks (RNNs) such as LSTMs (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 very 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 **a) Majority Classifier**. As a first simple baseline, we consider the majority classifier which always predicts the most frequent speech act. **b) 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[3]), and the number of words in the utterance. **c) 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 We use the reference implementation from *pycrfsuite*[4] the CRF. We extend the set of features used by the baseline models and add bigrams and repetitions (number of words that are repeated from the previous utterance) to provide the model with some context of the previous utterances.[5]

Hierarchical LSTM + CRF 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. For each utterance, one-hot encodings of the words (and a prepended speaker token) 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.[6] The model has access to contextualized utterance representations as well as the history of speech acts for the classification task.

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. 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.

Results

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

[1] 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.

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

[3] 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%).

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

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

[6] More details on the model architecture and hyperparameters can be found in the appendix.

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

Table 1: Accuracy for all models.

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 and report mean and standard deviation of the different models' accuracies 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. 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. Finally, we identified the CRF as best-performing. It is the model we use for the rest of the paper.

Though the numbers were obtained when the model was trained on 80% of the dataset (around 44000 utterances), the learning curve in Figure 2 shows that CRF 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.

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, 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. We define this age — by analogy to work on word learning (Braginsky et al., 2019) — as the age when at least 50% of children have acquired the speech act.

We only use predicted labels on parts of the corpus that were not seen by the model in the training phase. To this

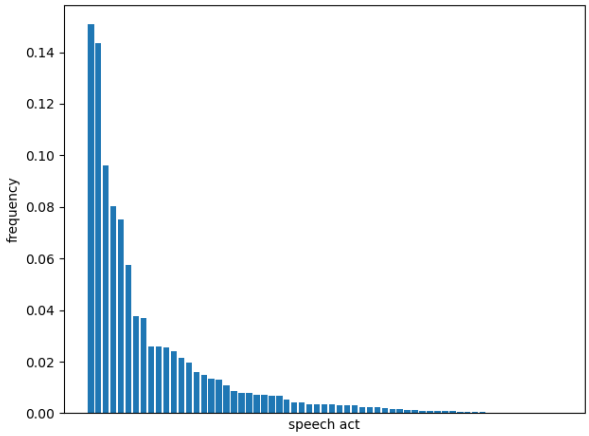

Figure 1: Distribution of frequencies of all speech acts in the New England corpus.

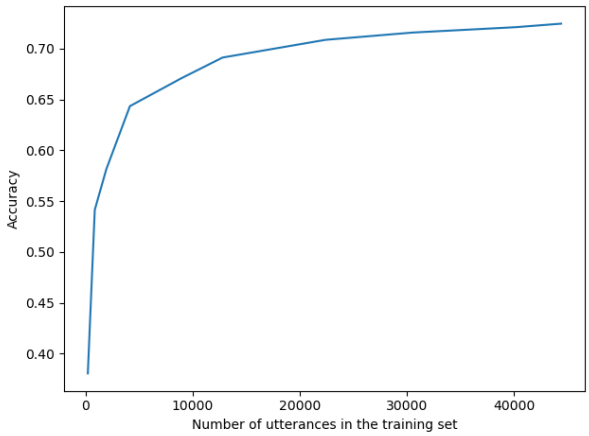

Figure 2: CRF: Accuracy as a function of training set size.

end, and to obtain labels for the whole New England corpus, we used 5-fold cross-validation to 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.

Development of the number of distinct speech acts Figure 3 shows the fraction of children producing a given number of different (and interpretable) speech act types for the three age groups studied in Snow et al. (1996) (This is a direct replication of Figure 2 in that 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 handful of speech act types, such as statements (`ST`), repetitions (`RT`) and markings (`MK`). This number increases on average for children aged 20 months

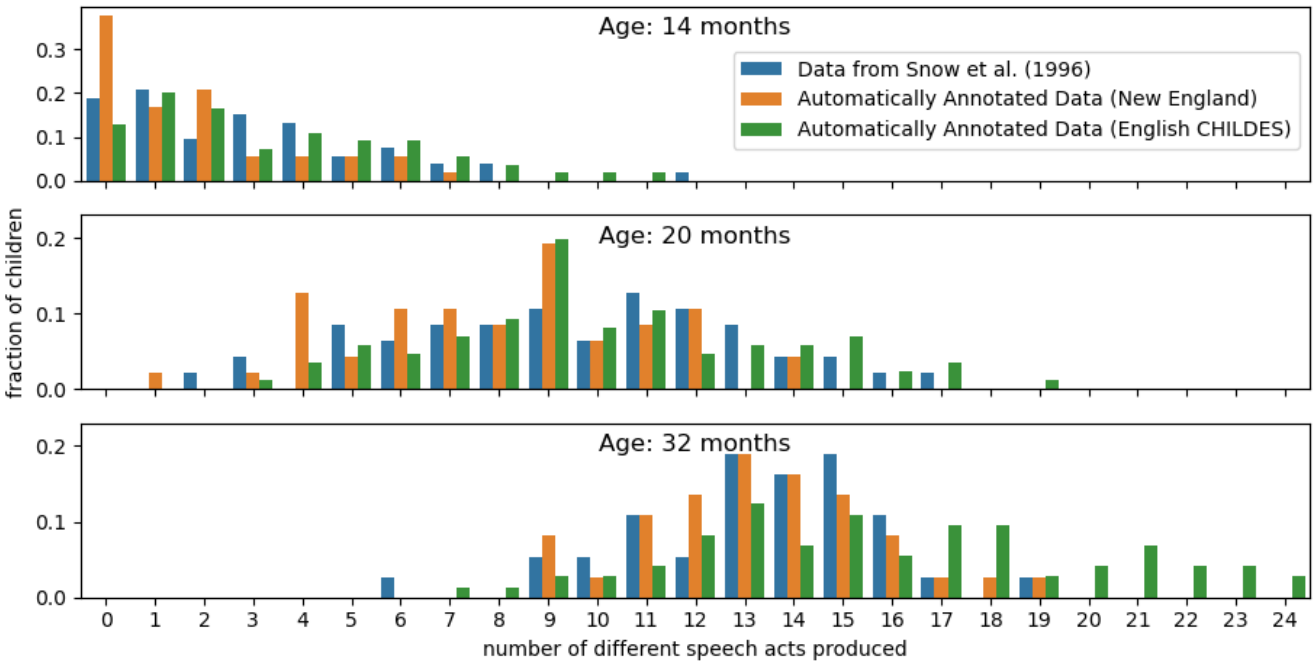

Figure 3: Fraction of children producing a given number of distinct speech act types at 14, 20, and 32 months old. Note that the y-axis for the bottom two figures has been shortened for better visibility. Jensen-Shannon distances of automatically annotated data (New England) compared to data from Snow et al. (1996): 0.262 (14 months), 0.367 (20 months), 0.186 (32 months). Jensen-Shannon distances of automatically annotated data (English CHILDES) compared to data from Snow et al. (1996): 0.209 (14 months), 0.222 (20 months), 0.418 (32 months).

where now a substantial fraction 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). 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.

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)). 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.).

Age of acquisition of speech acts In this section, we do a new analysis that consists in calculating the precise age of emergence of speech acts. By analogy to work in word learning (Braginsky et al., 2019), we say that a speech act is acquired (in terms of production), if at least 50% of the observed children use it[7]. For each speech act S, we proceed as follows:

1. For each age (14, 20 and 32 months), calculate the fraction of children who are producing S at least twice.
2. Perform a logistic regression of the data points.
3. Measure the age of first production as the age where the logistic regression curve surpasses the value 0.5 .

We successfully calculated the age of acquisition for a subset of 25 speech acts[8] using both the ground-truth labels from Snow et al. (1996) and the automatically generated labels from the CRF on the same dataset. Then, we calculated the Spearman rank-order correlation to examine whether the *order* of emergence of speech acts is correctly captured by the automatically annotated data. The resulting high correlation (see Figure 5 (left); $r \approx 0.82$, $p < 1 \cdot 10^{-6}$) indicates that the automatically generated labels can provide reasonable estimates for the developmental trajectory of speech acts.

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

[8] 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, it is due to the fact that the frequency distribution is highly skewed: Most categories occurred rarely in the corpus (Figure 2) and therefore did not provide enough data to be used in the calculation of age of acquisition.

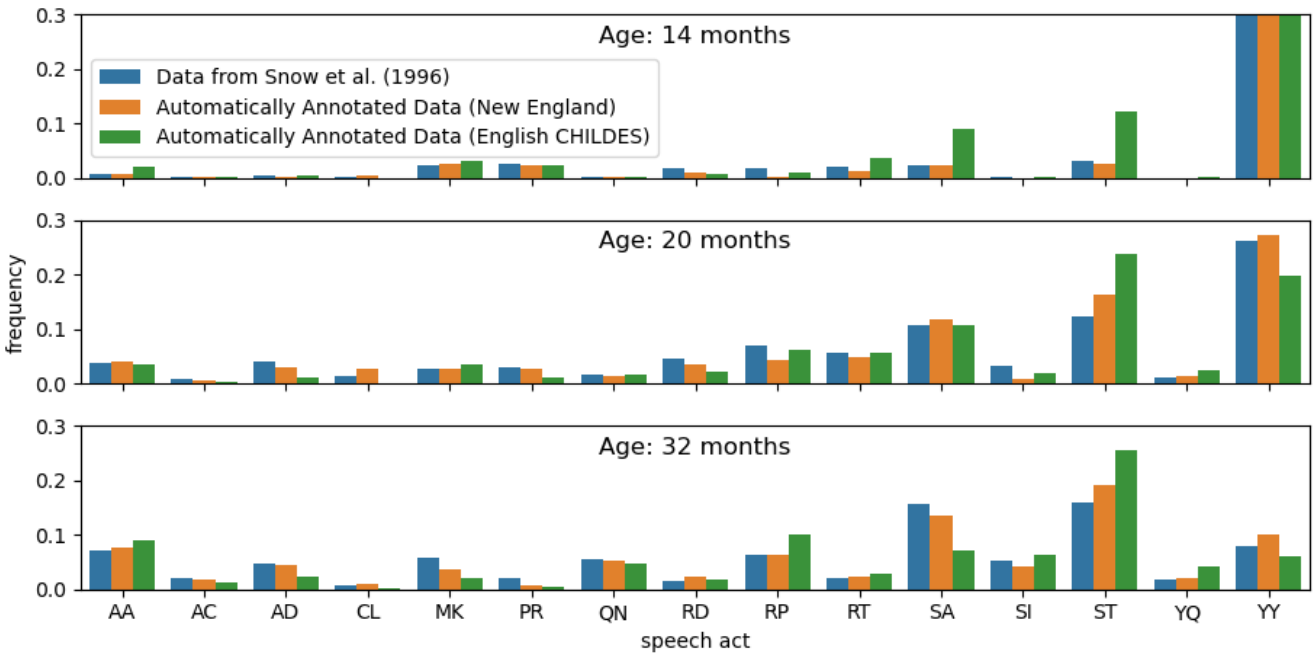

Figure 4: Frequency distribution of speech acts for different ages. Note that the y-axes have been trimmed for better visibility (The frequencies for YY at 14 months are around 0.6). Jensen-Shannon distances of automatically annotated data (New England) compared to data from Snow et al. (1996): 0.089 (14 months), 0.103 (20 months), 0.080 (32 months). 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).

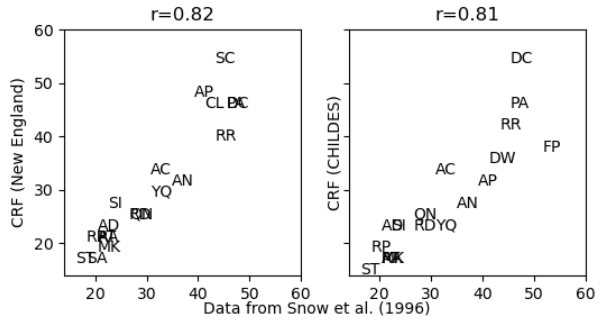

Figure 5: Correlation of age of acquisition as calculated using data from Snow et al. (1996) and automatically annotated data for the New England corpus and CHILDES. Note that some speech acts are not displayed because the axes limits were set to 60 months for better visibility of early development.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

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 investigate the extent to which findings obtained with 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).

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.

Age of acquisition of speech acts We 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 $r \approx 0.81$ ($p < 1 \cdot 10^{-6}$).

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 89.1 months), give reason (GR, at 84.9 months), polite requests (RQ, at 66.2 months), and make promises (PD, at 118.2 months)). These predictions are consistent with the developmental literature showing a late acquisition of some of these speech acts (Matthews, 2014).

Discussion

How children master language use in social interaction is an important theoretical frontier in the study of language development (Sperber & Wilson, 1986) — and human cognition more generally (Tomasello & Rakoczy, 2003) — with the potential for applications in various fields raging from health (e.g., early and automatic detection of communicative issues) to engineering (e.g., design of conversational agents for children). However, the investigation of this phenomenon in ecological valid settings 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, the 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 size. 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 only 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.

Finally, the current model learns how to recognize speech acts from their linguistic properties only. While the scores are quite good and allow us to replicate major findings that were obtained using human annotations, there is still room for improvement. In future work, we 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. Indeed, such cues are picked up on by adults and children and are integrated to optimize language understanding and learning (e.g., Fourtassi & Frank, 2020; Fourtassi et al., 2021). This effort will involve collecting multimodal data of spontaneous child-caregiver conversations as well as the development of machine learning methods for the automatic annotation of speech acts using linguistic, acoustic, and visual features.

Acknowledgements

We thank the anonymous reviewers for their comments and feedback.

This work, carried out within the Labex BLRI (ANR-11-LABX-0036) and the Institut Convergence ILCB (ANR-16-CONV-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)

References

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, *3*, 52–67.

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

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

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*.

Fourtassi, A., & Frank, M. C. (2020). How optimal is word recognition under multimodal uncertainty? *Cognition*, *199*.

Fourtassi, A., Regan, S., & Frank, M. C. (2021). Continuous developmental change explains discontinuities in word learning. *Developmental Science*, *24*(2), e13018.

Grice, H. P. (1975). Logic and conversation. In *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. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 2012–2021).

Kumar, H., Agarwal, A., Dasgupta, R., & Joshi, S. (2018). Dialogue act sequence labeling using hierarchical encoder with crf. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 32).

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

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.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... 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.

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.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Harvard University Press Cambridge, MA.

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

Tomasello, M., & Rakoczy, H. (2003). What makes human cognition unique? from individual to shared to collective intentionality. *Mind & language*, *18*(2), 121–147.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., ... others (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45).

Appendix

The appendix can be downloaded from the following OSF project: `https://osf.io/hvzs2/` .

Source code of all models and experimentation scripts can be found here: `https://github.com/mitjanikolaus/childes-speech-acts`.