Measuring Language Development From Child-centered Recordings

Yaya Sy, William Havard, Marvin Lavechin, Emmanuel Dupoux, Alejandrina Cristia

INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Measuring language development from child-centered recordings

Yaya Sy[1], *William N. Havard*[1,2], *Marvin Lavechin*[1,2,3], *Emmanuel Dupoux*[1,2,3], *Alejandrina Cristia*[1]

[1] Language Acquisition Across Cultures Team, Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'Études Cognitives, ENS, EHESS, CNRS, PSL University, France
[2] Cognitive Machine Learning Team, INRIA, Paris, France
[3] Meta AI Research, France

yayasysco@gmail.com, william.havard@gmail.com, marvin.lavechin@gmail.com, emmanuel.dupoux@gmail.com, alecristia@gmail.com

Abstract

Standard ways to measure child language development from spontaneous corpora rely on detailed linguistic descriptions of a language as well as exhaustive transcriptions of the child's speech, which today can only be done through costly human labor. We tackle both issues by proposing (1) a new language development metric (based on entropy) that does not require linguistic knowledge other than having a corpus of text in the language in question to train a language model, (2) a method to derive this metric directly from speech based on a smaller text-speech parallel corpus. Here, we present descriptive results on an open archive including data from six English-learning children as a proof of concept. We document that our entropy metric documents a gradual convergence of children's speech towards adults' speech as a function of age, and it also correlates moderately with lexical and morphosyntactic measures derived from morphologically-parsed transcriptions.
The source code of the experiments is released at `https://github.com/yaya-sy/EntropyBasedCLDMetrics`
Index Terms: L1 acquisition, child speech, morphosyntax, phonetics, speech technology application

1. Introduction and related work

Children's language production abilities undergo rapid changes in the first three years [1]. Standard ways to measure changes in lexical and morphosyntactic development from spontaneous corpora tend to rely on detailed linguistic knowledge and costly human annotations. For instance, the current best measures of lexicon, morphosyntax, and syntax all depend on morphologically parsed, exhaustively transcribed child speech [2, 3, 4]. To produce such annotations, one requires detailed knowledge of the language, typically including a dictionary, a part-of-speech parser, and a grammar, in addition to an estimated 10x of highly trained annotators' time to produce transcriptions of what the child is saying.

Although undoubtedly useful in high-resource settings, we believe continuing to rely on such measures limits our ability to learn more about child speech production across languages and human populations, and are likely key in explaining why fewer than 1% of languages are represented in mainstream language development journals [5]. Indeed, they are reliant on languages being well-resourced. Moreover, human annotation is particularly impractical for languages which are understudied and in populations that are highly multilingual [6].

Here, in order to derive automatic measures of children language development, we propose to leverage the power of Language Models (LMs) trained on adult utterances to predict the next unit. Indeed, LMs are commonly used to capture contextual dependencies of sequences composed of discrete units, whether these are words, phones or discrete units derived from raw speech. A standard way of evaluating how well a language model is doing is to use the measure of entropy, which quantifies the uncertainty of the language model, i.e., how difficult it is to predict the next unit.

We thought entropy could track language development as follows. Let us suppose we have a language model trained solely on utterances spoken by adults. When children start to speak, their utterances are qualitatively very far from that of adults because they use sounds and combinations of sounds that are rare in the ambient language. If we estimate the entropy of children's utterances using this model when children start to speak and compare it to the entropy computed on utterances spoken by adults, we should observe a large difference, as children's utterances would be more unexpected than adults' under the adult-trained language model. As children grow older, they start producing strings of sounds that are closer and closer to those adults use, not only because they come to use the same sounds and sound combinations, but also the same words, and eventually stringing words together as adults would do. As a result, the entropy of the sentences uttered by children should gradually converge towards that of adults. Thus, we can use adults as a reference point to which children could be compared to, theoretically allowing us to measure development regardless of which language – or set of languages – the child is learning.

The long-term goal we set to ourselves is to borrow from natural language processing to create a language development metric, ideally based on raw audio, all the while being as informative as those that are extracted from costly and linguistically-informed human transcription of children's production. In this paper, we present descriptive results of our efforts to describe development in six English-learning children, whose data are openly available, as a proof of concept, which already allows us to establish some boundary conditions under which our proposed measure behaves as predicted. Through two experiments (Table 1), we make two key contributions: (1) a new language development metric (based on entropy) that does not require linguistic knowledge other than having a corpus of text in the language in question to train a language model, (2) a method to derive this metric directly from speech based on a smaller text-speech parallel corpus. Together, they show that our entropy-based metric documents a gradual convergence of children's speech towards adults' speech as a function of age, and it also correlates weakly to moderately with lexical and morphosyntactic measures derived from morphologically-parsed transcriptions. We close this paper by discussing how our approach may help study individual variation among English-learning children and in more diverse populations.

10.21437/Interspeech.2023-1569

Table 1: *Model, data, and units for the relevant analyses. Training:* LIBRI. *stands for LibriSpeech;* THOM. *stands for* THOMAS. *Test always drew from* PROVIDENCE.

Exp.	Model	Model input units	Train	Test
1A	5-gram language model	phones	LIBRI. text	text
1B		HUBERT-BASE discrete clusters	Libri. audio	speech
1C				synthetic speech
2A	linear regression	speech (+ text entropies at training time)	THOM.	speech
2B			LIBRI.	

2. Experiment 1

2.1. Model and entropy measure

In this paper, we restrict ourselves to a 5-gram language model for the sake of interpretability. However it is only one of many models we could have used and future work might consider other language models. We used the fast and memory-efficient implementation proposed in KenLM [7], which uses a modified Kneser-Ney smoothing to assign probabilities to the unknown ngrams.

Experiment 1A sets a baseline: the 5-gram language model is trained on non-noisy text, an ideal condition. In a more realistic condition, the model of **Experiment 1B** is trained on discrete units derived from raw audio recordings of natural interactions between the child and its mother. To better explain the results of Experiment 1B, **Experiment 1C** reproduces the same experiment with the same corpus but using synthetic audio data. Once the model has been trained, we compute the entropy H of a sequence $s = u_1, u_2, ..., u_T$ composed of T discrete units u as follows:

$$H(s) = -\frac{1}{T}\log\big[p(u_1)p(u_2|u_1)\dots p(u_T|u_1,\dots,u_{T-1})\big]$$

where p is the probability assigned by the 5-gram language model. This quantity tells us how well the model predicts the sequence. In simple terms, we can say that entropy is lower for higher probability, less surprising sequences. For example, grammatical sentences will be assigned a higher probability (i.e., lower entropy) than ungrammatical sentences.

2.2. Data

For training, we used LibriSpeech-960 [8], as it constitutes a large dataset with both text and audio, and may be extended in several languages, which may help future extensions [9]. We tested on the PROVIDENCE data set [10, 11]. We chose it because it is one of the largest and best-established corpora in the archive for child-centred recordings CHILDES, and many previous studies on child language development have employed it. The corpus consists of 364 longitudinal recordings of 6 children, ranging from 11 months to 4 years old, recorded in their homes and thus surrounded by their caregivers and siblings. We focused on the key child (the child wearing the recording device) and the mother, excluding speech by others. In total, the corpus we use for testing contains 178 955 utterance ($\approx$ 161h of speech) for mothers and 112,209 utterances ($\approx$ 95h of speech) for key children. The key child's speech had been transcribed phonetically mostly, and orthographically sometimes; the speech of mothers only orthographically. Utterances have been time-stamped. Sections where speakers overlapped (i.e., time stamps overlapped across speakers) were removed from consideration. The orthographic transcriptions, written in the CHAT format [12], were pre-processed by removing markers specific to this format to leave only what was said (e.g., "ma [: mommy]" becomes "ma", removing the explanation that here "ma" stands for mommy).

2.2.1. Language model inputs

Text. We used PHONEMIZER [13] to transform each utterance into a string of characters, each character representing one phoneme, removing word boundaries (since word boundaries are not explicitly given in the speech experiments, doing so gives us a point of comparison).

Speech. We extracted audio sections corresponding to spoken utterances thanks to available timestamps. We discretized the spoken utterance using a k-means clustering model ($k = 500$) trained on the features of the 9^{th} transformer layer of the second iteration of the HUBERT-BASE [14] which is a self-supervised speech model pre-trained on LibriSpeech-960 [8].

Synthesized speech. We used the orthographic representations to synthesize single-speaker, clean speech with a Coqui TTS model[1] trained on the LJ SPEECH data set [15] that uses a Tacotron2 architecture [16]. For child utterances available only in phonetic transcription, we first generated a likely orthographic transcription (e.g., [əkwʌ ɹɾə] rendered as "aqua otta") using PINCELATE.[2] We then processed the synthesized speech with the exact same HUBERT-BASE as just described above.

2.2.2. Comparison metrics

To benchmark our entropy measure against standard metrics of language development, we applied CLAN's `kideval` command [17] onto the child's transcriptions from each recording session. We focus on three metrics which are resource-dependent because they require morphological transcriptions, and yet they are the current best recommendations for measuring vocabulary, morphology, and syntax, respectively [17]. These measures are vocabulary diversity (VOCD), mean length of utterance (MLU) in morphemes, and the Index of Productive Syntax (IPSyn). VOCD is reputedly the best measure of lexical diversity for several reasons, including its relative independence from transcript length and the fact that lexical diversity is done on lemmas rather than surface forms [18]. MLU measures the average length of sentences, and is thought to reflect both morphological and syntactic development. Finally, IPSyn measures the variability in syntactic constructions found in the child's speech, including categories like noun or verb phrase development or elaboration, question/negation constructions, and sentence phrase structure [4].

[1] `https://www.coqui.ai/`

[2] `https://pincelate.readthedocs.io/en/latest/`

2.3. Results

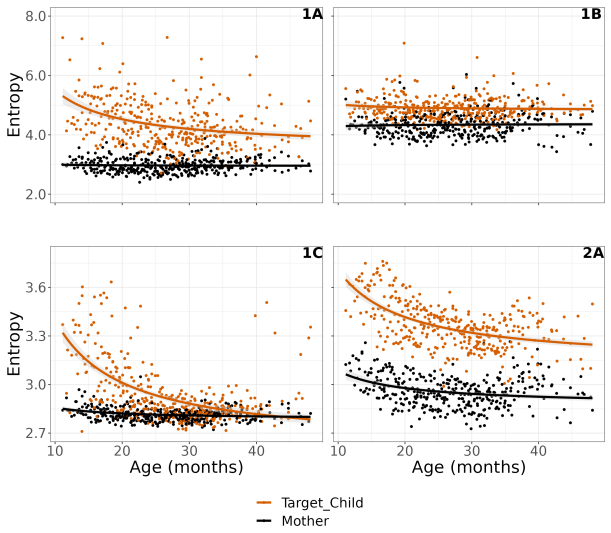

Figure 1: *Entropy for the key child versus the mother, as a function of child age. Each point represents the average entropy for all utterances by the key child (orange) or the mother (black), for a given recording session.* ***1A****=Experiment 1A* on text, ***1B****=Experiment 1B* on speech, ***1C****=Experiment 1C* on synthetic speech, ***2A****=Experiment 2A on* supervised entropy prediction on speech.

Entropy is computed for each utterance by the key child and the mother. Then we derive averages across all utterances for each recording session separately. To analyze how these session-level average entropies relate to age while acknowledging individual variation, we fit a mixed regression model for key child and mother separately, declaring age (subtracting minimum age, so that the intercept corresponds to the minimum age) as fixed and random intercepts for each child, as well as a random slope for age within each child (Table 2).

Figure 1 presents scatterplots for the Experiment 1. More complete information is available in Table 2. In a nutshell, our proposed entropy metric behaved as predicted in Experiment 1A: entropies computed for adult utterances were lower than those by the key child, particularly when the child was young, and this difference narrowed with child age due to entropies decreasing for the child, but remaining stable for the adult. In addition, variation across individual transcriptions across all children was used to see the extent to which our proposed entropy metric correlated with well-established vocabulary, morphoynstax, and syntactic metrics, all of which were calculated from morphologically parsed transcriptions from the same recordings. Fitting all of our predictions, Experiment 1A operates as a non-ideal topline for the other conditions, because although it does not require morphological parses, it still depends on text being transcribed.

Experiment 1B shows unexpected trends, as the child's entropy does not converge to that of the mother. This is probably due to the noises (the background sound of the TV, the child not speaking in front of the microphone or playing with it, etc) and speech variability. By using synthetic speech in Experiment 1C, we remove noises and inter- and intra-speaker speech variability and the results show that, provided these optimal conditions, it is possible to observe the same trend as 1A. The results reported in Table 2 also show that our text-based entropy metric (Experiment 1A) correlates well with standard language development metrics, whereas an unsupervised LM from raw recordings (Experiment 1B) speech does not.

2.4. Discussion

Results of Experiment 1 show how challenging real child-centered data are. When using a noise-free, speaker independent phone-based representation, a 5-gram language model returned entropies that followed precisely the predicted pattern (Experiment 1A). When this is not obtained in Experiment 1B, it can be explained by the many differences between the two conditions, including the fact that discrete units entering the language model are a lot smaller: a phone is about 40ms in duration, so a 5-gram covers about 200m, whereas a 5-gram from speech clusters only likely covers about 50ms. Experiment 1C allows us to discard many of these explanations because it uses the same speech representation model, the same number of clusters, and the same 5-gram covering only 50ms, and yet reproduces the predicted entropy patterns and relation to standard language development metrics. Together, 1B and 1C demonstrate that it is really the use of children's voices and/or the presence of noise in child-centered recordings that poses a hazard for directly extracting entropy from a 5-gram language model applied to discrete speech representations as a potential metric for language development.

3. Experiment 2

A system like that in Experiment 1B would have been ideal, as it could have been applied to any language and it did not rely on any human transcription. However, that failure suggests today's speech representation models may not be equipped for dealing with children's voices and/or the level of background noise found in child-centered recordings [19]. Therefore, in this experiment, we use a layer of supervision to generate a system that predicts entropies from text based on the information in the audio. The hope here is that we can use some text-based supervision, so we are not yet breaking completely free from transcription, but at least we are limiting the amount of human transcription and detailed linguistic knowledge (as required for morphological parsing) that is needed to measure language development.

3.1. Data

The model learns to predict the text entropy of a spoken utterance. So, the training data is composed of paired spoken utterances and the entropies from the text version of these same utterances, as returned by the 5-gram language model from Experiment 1. In Experiment 2A, the training data was 63,276 utterances (30h of speech) randomly sampled from the THOMAS corpus [20], which is another open-source dataset in CHILDES. It contains recordings from a single child learning British English. It was selected because speech for both the key child and his caregivers was transcribed and accurately time-stamped. In Experiment 2B, the training data was LibriSpeech `train-clean-100` (100h of speech). As in the Experiments 1, the models were then tested on PROVIDENCE.

3.2. Model

We use WHISPER [21] as a speech features extractor, because of reports of high performance on ASR and other downstream

Table 2: *Fit of our entropy metric to predictions. From mixed-model regressions,* ***intercepts (Std. Error)*** *indicate how well entropy separates child/mother at the child's youngest age;* ***β age (Std. Error)*** *indicates how entropy changes with age in the child and the mother data.* ***ρ CLD metric*** *shows how entropy correlates with standard metric of language development.* ***** *indicates that the estimate is significantly different from zero. $\downarrow$ (downwards trend) and $\leftrightarrow$ (stable trend) show expected change with age. Negative correlations are consistent with the notion that language development leads to higher VOCD, MLU and IPSyn, and lower (more adult-like) entropies.*

	Intercept (Std. Error)		**β age (Std. Error)**		**ρ CLD metric**		
Exp.	Child	Mother	Child $\downarrow$	Mother $\leftrightarrow$	VOCD	MLU	IPSyn
1A	5.06 (0.23)*	2.97 (0.07)*	-0.31 (0.06)*	0.01 (0.02)	-0.23	-0.56	-0.45
1B	4.91 (0.07)	4.3 (0.11)	0.01 (0.02)	0.01 (0.04)	0.03	-0.02	-0.01
1C	3.13 (0.06)	2.83 (0.01)	-0.1 (0.02)	-0.01 (0.00)	-0.21	-0.56	**-0.54**
2A	3.54 (0.04)*	3.00 (0.03)*	-0.1 (0.01)*	-0.02 (0.01)	**-0.27**	**-0.73**	-0.53
2B	2.67 (0.01)*	2.57 (0.02)*	0.00 (0.00)	0.01 (0.01)	0.14	-0.08	-0.16

tasks, including for child-centered data [22]. The audio is first processed through WHISPER-BASE to return a sequence of vectors representing the audio. We mean-poled these vectors to obtain a single vector $\mathbf{c}$ and a linear model is trained to predict the text entropy: $\hat{e} = \mathbf{w} \times \mathbf{c}$. The parameter $\mathbf{w}$ is estimated using the *Mean Squared Error* loss function. Note that the parameters of the WHISPER model are frozen (i.e, not updated during the training), only $\mathbf{w}$ is estimated.[3]

3.3. Results

To facilitate comparison with Experiment 1, we provide statistical information in Table 2, but due to space limitation, we only give the scatterplot of Experiment 2A in the Figure 1. The results of the Experiment 2A are remarkably close to the results obtained in Experiments 1A and 1C: that suggests that, with appropriate supervision, a system can learn to return entropies that behave in the predicted way as a function of child age, and that are quite similar to those obtained from the text. Except for the VOCD, correlations with the standard language development metrics are strong, reinforcing the idea that this approach based on a small quantity of text-audio pairs could represent children's development as well as more costly full transcriptions. It is also encouraging that the system generalized to a new corpus, collected and annotated by a different researcher, and where there was a single child learning a different dialect. Generalization is not ensured however: Experiment 2B shows that we need to train this system with in-domain data.

3.4. Discussion

Results from Experiment 2A suggest that we can obtain an entropy metric based on the audio signal that relates to language development with 30h of human time-stamped and transcribed child-centered data, which importantly came from a different corpus. The CHILDES archive [23] contains time-stamped and transcribed child-centered data for over 40 languages (although we do not know how many of those have 30h of speech). If Experiment 2B had shown the same pattern, this would have further allowed us to use for training adult-centered data, which is available for many more languages. Unfortunately, the fact that 2B showed a different pattern strongly suggests that training with in-domain data is necessary, imposing important boundary conditions for generalization.

[3] We trained the models with a learning rate of 0.00056 for 5 epochs.

4. Future Directions

Many open challenges lie ahead. To begin with, we do not know how well our metric will work in other languages. As we start studying other languages, the "standard" metrics against which we benchmark ours may come under attack. For instance, IPSyn is currently only available in English on CLAN [17], and MLU in morphemes has been criticized in the context of languages varying in the degree of polysynthesis [24].

Regarding the usefulness of this technique to study individual variation among English learners, we would like to acknowledge that we have simplified the task in several ways, some of which are less problematic than others. For instance, our recordings were less noisy than the increasingly common child-centered long-form audio data [25], but we can easily imagine many applications in which collecting short, noise-free recordings from a given child is feasible. We used human transcriptions for some of the tasks, but Experiment 2A shows a system trained in data from one child learning British English performed well in describing our six American speakers. More problematic is the use of human segmentation to identify spoken sections, which in our experience requires 20-40x the audio length to do accurately.

5. Conclusions

The field of language development has been dominated by the metrics that have been proposed and developed by English-speaking researchers living in high-resource settings, leading to marked inequity in the publication record in terms of language diversity [5]. This paper took a step towards developing a metric that does not rely on morphological parsing, which requires high degrees of training in human annotators and language-specific resources. Another important advantage of our metric, not spelled out above, is that we can employ the data from adults in the child's own recordings to benchmark development, which will prove extremely useful for under-resourced languages and multilingual populations.

6. Acknowledgements

The authors would like to thank the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award; European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ExELang, Grant agreement No. 101001095). This work was performed using HPC resources from GENCI–IDRIS (Grant 2021-AD011013145)

7. References

[1] E. Hoff and M. Shatz, *Blackwell handbook of language development*. John Wiley & Sons, 2009.

[2] M. A. Covington and J. D. McFall, "Cutting the gordian knot: The moving-average type–token ratio (mattr)," *Journal of quantitative linguistics*, vol. 17, no. 2, pp. 94–100, 2010.

[3] M. D. Parker and K. Brorson, "A comparative study between mean length of utterance in morphemes (mlum) and mean length of utterance in words (mluw)," *First language*, vol. 25, no. 3, pp. 365–376, 2005.

[4] J. S. Yang, B. MacWhinney, and N. B. Ratner, "The index of productive syntax: Psychometric properties and suggested modifications," *American Journal of Speech-Language Pathology*, vol. 31, no. 1, pp. 239–256, 2022.

[5] E. Kidd and R. Garcia, "How diverse is child language acquisition research?" *First Language*, 2022.

[6] F. T. Woon, E. C. Yogarrajah, S. Fong, N. S. M. Salleh, S. Sundaray, and S. J. Styles, "Creating a corpus of multilingual parent-child speech remotely: Lessons learned in a large-scale onscreen picturebook sharing task," *Frontiers in Psychology*, vol. 12, p. 734936, 2021.

[7] K. Heafield, "KenLM: Faster and Smaller Language Model Queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, Jul. 2011, pp. 187–197. [Online]. Available: https://aclanthology.org/W11-2123

[8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[9] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.

[10] K. Demuth, J. Culbertson, and J. Alter, "Word-minimality, Epenthesis and Coda Licensing in the Early Acquisition of English," *Language and Speech*, vol. 49, no. 2, pp. 137–173, Jun. 2006. [Online]. Available: https://doi.org/10.1177/00238309060490020201

[11] K. Demuth, "CHILDES English Providence Corpus," 2004. [Online]. Available: https://phon.talkbank.org/access/Eng-NA/Providence.html

[12] B. MacWhinney, *The Childes Project: Tools for Analyzing Talk, Volume I: Transcription format and Programs*, paperback ed. Psychology Press, 5 2014.

[13] M. Bernard and H. Titeux, "Phonemizer: Text to Phones Transcription for Multiple Languages in Python," *Journal of Open Source Software*, vol. 6, no. 68, p. 3958, 2021. [Online]. Available: https://doi.org/10.21105/joss.03958

[14] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[15] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[16] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[17] B. MacWhinney, "Tools for analyzing talk part 2: The clan program," *Talkbank. Org*, no. 2000, 2017.

[18] J. S. Yang, C. Rosvold, and N. B. Ratner, "Measurement of lexical diversity in children's spoken language: Computational and conceptual considerations," *Frontiers in Psychology*, vol. 13, 2022.

[19] M. Lavechin, M. de Seyssel, L. Gautheron, E. Dupoux, and A. Cristia, "Reverse Engineering Language Acquisition with Child-Centered Long-Form Recordings," *Annual Review of Linguistics*, vol. 8, no. 1, pp. 389–407, Jan. 2022. [Online]. Available: https://doi.org/10.1146/annurev-linguistics-031120-122120

[20] E. Lieven, D. Salomo, and M. Tomasello, "Two-year-old children's production of multiword utterances: A usage-based analysis," vol. 20, no. 3, pp. 481–507, 2009. [Online]. Available: https://doi.org/10.1515/COGL.2009.022

[21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[22] M. Lavechin, M. Métais, H. Titeux, A. Boissonnet, J. Copet, M. Rivière, E. Bergelson, A. Cristia, E. Dupoux, and H. Bredin, "Brouhaha: multi-task training for voice activity detection, speech-to-noise ratio, and c50 room acoustics estimation," *arXiv preprint arXiv:2210.13248*, 2022.

[23] B. Macwhinney, "The CHILDES project: tools for analyzing talk," *Child Language Teaching and Therapy*, vol. 8, 01 2000.

[24] S. E. Allen and C. Dench, "Calculating mean length of utterance for eastern canadian inuktitut," *First language*, vol. 35, no. 4-5, pp. 377–406, 2015.

[25] M. Lavechin, M. De Seyssel, M. Métais, F. Metze, A. Mohamed, H. Bredin, E. Dupoux, and A. Cristia, "Statistical learning models of early phonetic acquisition struggle with child-centered audio data," 2023.