INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

BabySLM: language-acquisition-friendly benchmark of self-supervised spoken language models

Marvin Lavechin[1,2], *Yaya Sy*[1], *Hadrien Titeux*[1], *María Andrea Cruz Blandón*[3], *Okko Räsänen*[3], *Hervé Bredin*[4], *Emmanuel Dupoux*[1,2,5], *Alejandrina Cristia*[1]

[1]LSCP, ENS, EHESS, CNRS, PSL University, Paris, France [2]Meta AI Research, France
[3]Unit of Computing Sciences, Tampere University, Finland [4]IRIT, CNRS, Toulouse, France
[5]Cognitive Machine Learning Team, INRIA, France

marvinlavechin@gmail.com

Abstract

Self-supervised techniques for learning speech representations have been shown to develop linguistic competence from exposure to speech without the need for human labels. In order to fully realize the potential of these approaches and further our understanding of how infants learn language, simulations must closely emulate real-life situations by training on developmentally plausible corpora and benchmarking against appropriate test sets. To this end, we propose a language-acquisition-friendly benchmark to probe spoken language models at the lexical and syntactic levels, both of which are compatible with the vocabulary typical of children's language experiences. This paper introduces the benchmark and summarizes a range of experiments showing its usefulness. In addition, we highlight two exciting challenges that need to be addressed for further progress: bridging the gap between text and speech and between clean speech and in-the-wild speech.

Index Terms: spoken language modeling, language acquisition, self-supervised learning, child language

1. Introduction and related work

Machine learning for Natural Language Processing (NLP) has led to models that develop linguistic competence from exposure to written or spoken language. On text, Language Models (LMs) now achieve impressive performance on a wide variety of natural language understanding tasks [1]. More recently, speech-based LMs have also shown impressive linguistic competence on lexical or grammatical acceptability judgment tasks [2, 3], or spoken language generation [4, 5]. Since these models develop linguistic competence without the need for human labels, they promise to advance our understanding of how infants learn language [6, 7, 8]. However, if we want to maximize the impact of the evidence obtained from LMs, it is essential to ensure that our simulations closely emulate real-life situations – as advocated for syntactic acquisition in text-based LMs in [8, 9].

How can we do so? First, we should match the *quantity* of data available to young infants. Although large differences exist across cultures [10] and socioeconomic contexts [11], current estimates of yearly speech input vary between 300 and $1,000$ hours for American English-learning children [6, 12]. This means that by age 3, American English-learning children would have been exposed to approximately 3,000 hours of speech – for those who received the most speech input. Yet, by then, infants know many words and already engage in simple conversations [13]. Second, we should match the *quality* of data available to young infants. Contrary to LMs, infants do not learn language by scraping the entire web or through exposure to a large quantity of audiobooks. Instead, infants' input is speech – not text –, and it contains a relatively small vocabulary arranged in simple and short sentences, sometimes overlapping across speakers and laced with various background noises [7, 14].

Evaluating LMs trained on quantitatively and qualitatively plausible corpora requires the creation of adapted benchmarks, but none exists for speech-based LMs – see [9] or the BabyLM challenge [15] for text-based LMs. Current benchmarks using zero-shot probing tasks, although inspired by human psycholinguistics (e.g., spot-the-word or grammatical acceptability judgment tasks), have been designed for models trained on audiobooks [2]. As a result, these benchmarks use a large vocabulary specific to books (including words like 'rhapsodize', 'zirconium', or 'tercentenary') and probe syntactically complex sentences that are vanishingly rare even in spontaneous adult-adult conversation.

Here, we propose *BabySLM*, the first language-acquisition-friendly benchmark to probe speech-based LMs at the lexical and syntactic levels, both of which are compatible with the vocabulary typical of children's language experiences. Our benchmark relies on zero-shot behavioral probing of LMs [2] and considers a spot-the-word task at the lexical level and a grammatical acceptability judgment task at the syntactic level. To show the utility of our benchmark, we first use it to evaluate text-based and speech-based LMs trained on developmentally plausible training sets. The text-based LM is a long short-term memory (LSTM) trained on phonemes or words. The speech-based LM is the low-budget baseline used in the ZeroSpeech 2021 challenge on unsupervised representation learning of spoken language [2]. Both systems are trained on Providence [16], a dataset of spontaneous parent-child interactions. The comparison between text-based and speech-based LMs shows an important gap that future work should address. Next, *BabySLM* enables us to compare the performance of speech-based LMs when trained on $1,000$ hours of speech extracted from 1) audiobooks, a source of training data commonly used [17, 18]; or 2) child-centered long-form recordings acquired via child-worn microphones as people go about their everyday activities [19]. Our results reveal that speech-based LMs are overly sensitive to the differences between clean speech and in-the-wild speech.

2. Methods

2.1. Metrics

2.1.1. Lexical evaluation: the spot-the-word task

General principle. In the lexical task, the system is presented with minimal pairs of an existing word and a pseudo-word that

We thank HPC resources of GENCI-IDRIS (2022-AD011012554); ANR-19-P3IA-0001; J. S. McDonnell Foundation; ERC (ExELang, 101001095).

10.21437/Interspeech.2023-978

Table 1: ***Lexical task.*** *Minimal pairs of real and pseudo-words. Phonetic (Phon.) transcriptions are given in International Phonetic Alphabet (IPA) standard. Orthographic (Orth.) transcriptions of pseudo-words are proposed for ease of reading.*

Word	Pseudo-words Phon.	Orth.	Word	Pseudo-words Phon.	Orth.
hello h ə l oʊ	l ə l oʊ	lello	thanks θ æ ŋ k s	θ ɛ ŋ k s	thaynks
	p ə l oʊ	pello		θ ɔ ŋ k s	thoanks
	s ə ɹ oʊ	sero		θ ɪ s k s	thisks
	d ə l oʊ	dello		θ æ m p s	thamps
	s ə l oʊ	sello		θ æ n t s	thants
cookie k ʊ k iː	k ʊ t iː	kootie	jump dʒ ʌ m p	dʒ æ m p	jamp
	k ʊ n i	koonie		dʒ ʌ l k	julk
	ɹ ʊ d iː	roodie		dʒ ʌ s k	jusk
	ɹ ʊ t iː	rootie		dʒ ʌ f t	juft
	b ʊ n iː	boonie		dʒ ʌ b s	jubs

is phonologically plausible but does not actually exist [2, 20] (examples in Table 1). The system gets a score of 1 if it returns a higher probability for the former, and 0 otherwise. Contrary to [2], we generate multiple pseudo-words per word. Scores are first averaged across pseudo-words to yield per-word accuracy, which are then averaged across all words to yield a measure of *lexical accuracy*.

Task generation. We first listed all words in the American English CHIld Language Data Exchange System (CHILDES) database [21]. This database contains human-annotated transcripts of various child-centered situations (play sessions, storytelling, etc.), making it a valuable source of vocabulary in real children's input. After excluding items not found in either the Celex [22] or CMU dictionary [23] (e.g., mispronounced, incorrectly annotated or made-up words: 'insectasaurus', 'hiphippopotamus'), we obtained $28,000$ word types. Pseudo-words were produced using the Wuggy pipeline [24], which generates, for a given word, a list of candidate pseudo-words matched for syllabic and phonotactic structure. We applied the same post-processing steps used in [2]. Contrary to [2], to ensure that there is no bias from phone-based unigrams or bigrams, we balanced the count of pseudo-words that had higher (or lower) phonemes unigram and bigram probabilities compared to those computed for the actual word. If a given word had only pseudo-words with higher (or lower) unigram or bigram possibilities, it was discarded from the evaluation set. The resulting $> 90,000$ minimal pairs across $18,000$ words were each synthesized using Google Text-To-Speech (TTS) system using 10 voices (5 males, 5 females).

2.1.2. Syntactic evaluation: grammatical acceptability

General principle. In the syntactic task, the system is presented with minimal pairs of grammatical and ungrammatical sentences across six syntactic phenomena [2, 9] (examples in Table 2), giving the system a score of 1 when it assigns a higher probability to the former, and 0 otherwise. We average scores within each syntactic phenomenon, then across phenomena to obtain our measure of *syntactic accuracy*.

Task generation. We generated templates for each of the six syntactic phenomena explored. For instance, for the noun-verb agreement phenomenon, we used templates such as "The <noun$_1$> <3^{rd} person verb> <noun$_2$>" versus "The <noun$_1$> <1^{st} person verb> <noun$_2$>". Contrary to [2], we restricted this benchmark to simple syntactic phenomena and short sentences which better reflect the type of input children are exposed to. We filled the templates using high-frequency words from CHILDES [21]. For instance, selected animate nouns include words like 'mom', 'girl', or 'cat'; selected adjectives include words like 'good', 'little', or 'big'; and selected verbs include words like 'see', 'know', or 'need'. The resulting $10,800$ minimal pairs were each synthesized using Google TTS system using the same 10 voices (5 males, 5 females).

Table 2: ***Syntactic task.*** *Minimal pairs of grammatical (✓) and ungrammatical (✗) sentences from each of the six syntactic phenomena included in our benchmark. N is the number of* $1,000$ *minimal pairs within each category.*

Phenomenon	N	Sentence example
Adjective-noun order	1.6	✓ *The good mom.* ✗ *The mom good.*
Noun-verb order	1	✓ *The dragon says.* ✗ *The says dragon.*
Anaphor-gender agreement	2	✓ *The dad cuts himself.* ✗ *The dad cuts herself.*
Anaphor-number agreement	1	✓ *The boys told themselves.* ✗ *The boys told himself.*
Determiner-noun agreement	3.6	✓ *Each good sister.* ✗ *Many good sister.*
Noun-verb agreement	1.6	✓ *The prince needs the princess.* ✗ *The prince need the princess.*

2.1.3. Development and test split

For both our lexical and syntactic evaluation sets, we randomly selected one male and one female voice for the development set and the 8 remaining ones for the test. We randomly selected $20\,\%$ of the lexical and syntactic minimal pairs for the development set and the remaining $80\,\%$ for the test.

2.2. Training sets

We built a first training set by extracting human-annotated speech utterances from Providence [16], a publicly available corpus containing transcribed recordings of six American children during spontaneous interactions with their parents. Available utterance-level timestamps were refined with a pretrained voice activity detection (VAD) system [25]. We converted human orthographic transcripts into phonetic transcripts using [26]. This procedure resulted in 128 hours of highly naturalistic infant-parent interactions in audio, orthographic, and phonetic form, allowing us to compare LMs trained on speech, phonemes, or words.

We built a second training set by extracting $1,024$ hours of adult speech utterances – using the same VAD system [25] – from SEEDLingS [19], a corpus of child-centered long-form recordings collected in 61 American English families. This training set enables us to train speech-based LMs in maximally plausible conditions, i.e., directly on what infants hear.

2.3. Models

STELA (speech-based). STELA is a speech-based LM originally proposed in [2, 27]. It comprises an acoustic model that learns discrete representations of the audio and a language

Table 3: ***The BabySLM benchmark.*** *Lexical and syntactic accuracies obtained by different language models trained on developmentally plausible corpora of speech, phonemes, or words. Numbers are computed on the test set, and performances on the development set are reported using small font size. The starred cumulated duration and number of words are estimates based on the* $1.2\,\mathrm{M}$ *of words present in the* 128 *hours of speech from Providence. Data plausibility indicates the extent to which the training set is close to the real sensory signal available to infants.*

System	Input	Training set	Cumulated duration (h)	Number of words (M)	Data plausibility	Lexical acc. (%)	Syntactic acc. (%)
Random baseline	—	—	0	0	—	49.2 52.5	49.3 50.0
STELA [27]	speech	SEEDLingS	1024	9.6*	+++	49.5 45.4	50.3 50.5
STELA [27]	speech	Providence	128	1.2	++	56.8 47.1	50.3 51.1
LSTM	phonemes	Providence	128	1.2	+	75.4 75.2	55.1 55.9
LSTM	words (BPE)	Providence	128	1.2	+	—	65.1 65.3
BabyBERTa [9]	words (BPE)	AO-CHILDES	533*	5	+	—	70.4 70.4

model trained on top of the learned discrete representations. The acoustic model is built from a Contrastive Predictive Coding (CPC) model followed by a K-means clustering algorithm. The language model consists of LSTM layers. We used the same architecture and hyper-parameters as the low-budget baseline proposed in [2]. Contrary to [2] who trained CPC by sampling the positive and negative examples from the same speaker, we applied a second constraint: negative examples were drawn from temporally close speech sequences to reduce mismatch between the positive and negative examples in terms of their local environment as this was found to be helpful when training on long-forms [14].

LSTM (text-based). We include LSTM LMs trained on words – using byte-pair encoding – or on phonemes, using the same architecture and hyper-parameters than [2].

BabyBERTa (text-based). BabyBERTa [9] is a transformer-based LM trained on a $5\,\mathrm{M}$ word corpus of American English child-directed input built from the CHILDES database [21].

3. Results and discussion

3.1. The BabySLM benchmark

Results obtained on our *BabySLM* benchmark are reported in Table 3. Rows are sorted according to the plausibility of the training data. Child-centered long-form recordings (SEEDLingS) have the highest plausibility score as these recordings faithfully capture children's everyday language experiences. In particular, long-forms collect audio data over a whole day – or several – and therefore sample the full range of language experiences across all possible contexts: the child may be in or out of the house, the speech may be directed to the child or others, etc. The audio extracted from in-home recordings of spontaneous infant-parent interactions (Providence) is slightly less plausible as it fails to capture the full range of language experiences: fewer speakers than in a real-life setting, most of the speech is directed to the child, etc. Finally, words and phonemes extracted from AO-CHILDES or Providence have the lowest plausibility score since infants do not learn language from orthographic or phonetic transcriptions but from the continuous signal that is speech.

Results indicate no evidence of lexical and syntactic knowledge for STELA trained on $1,024$ hours of speech from SEEDLingS. This contrasts, in appearance, with what has been found in the ZeroSpeech challenge [2], but this is due to the large variability of speech found in long-forms as we will see in Section 3.3. Results are no different for STELA trained on 128 hours of speech extracted from Providence whose lexical and syntactic accuracies remain close to chance level. However, we hypothesize that the lexical accuracy obtained by STELA might increase with more audio data from semi-controlled recordings of infant-parent interactions as these contain cleaner speech than what is typically found in long-forms. Contrary to speech-based LMs, text-based LMs perform largely above chance level. As expected, the LSTM model trained on words reaches higher syntactic accuracy than the LSTM trained on phonemes. The highest syntactic accuracy is obtained by BabyBERTa, which is a transformer-based LM and has been trained on a larger quantity of data than our LSTM LMs.

Performances on *BabySLM* show a clear gap between text-based and speech-based LMs. Another important finding is that, as of now, spoken language modeling from children's real language experiences seems out of reach, as evidenced by the chance-level lexical and syntactic accuracies obtained by STELA trained on SEEDLingS. We dedicate the remaining sections to illustrating these two challenges: bridging the gap between text and speech and between clean speech and in-the-wild speech.

3.2. Language modeling: from text to speech

Figure 1 shows lexical and syntactic accuracies obtained by text-based (words or phonemes) or speech-based LMs as a function of quantity of data. The LSTM trained on phones requires at least 16 hours of speech, equivalent to $150,000$ words, to start performing above chance level. Once lexical knowledge has emerged, the model follows a logarithmic trend (note the log-scale x-axis), initially improving rapidly and then slowing down. In other words, we need to double the amount of data to obtain the same gain in lexical accuracy. The same patterns hold for the syntactic accuracy obtained by the LSTM model trained on words[1]. For STELA, the lexical accuracy remains close to chance level, although the curve seems to increase between 32 and 128 hours of speech, and there is no evidence for syntactic knowledge.

All in all, the lexical and syntactic accuracy slopes show very different patterns when training from raw speech or phonemes or words. This is despite receiving the same data

[1]Note, however, that the syntactic accuracy obtained by the LSTM model trained on words decreases to $45\,\%$ (below chance level) between 0 and 8 hours ($= 75,000$ words). This effect was found to be driven by co-occurrence statistics in the noun-verb order task. The same pattern was found with a 3-gram model, with a slight decrease between 0 and 8 hours and an increase between 8 and 128 hours.

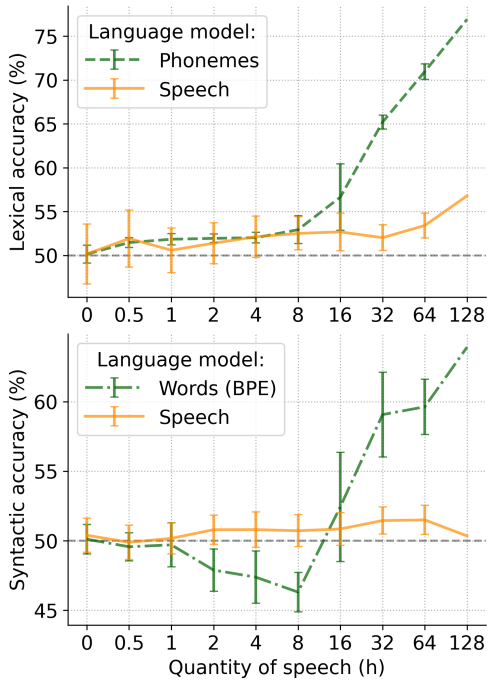

Figure 1: ***Language modeling from text to speech.*** *Top panel shows the lexical accuracy obtained by language models trained on audio (STELA) or phonemes (LSTM). Bottom panel shows the syntactic accuracy obtained by language models trained on audio (STELA) or byte-pair-encoded (BPE) words (LSTM). All models are trained on the Providence corpora in audio, phonetic, or orthographic form. Numbers are computed on the test set. Error bars represent standard errors computed across mutually exclusive training sets.*

in different forms. Admittedly, the speech-based LM faces a more challenging task as it must learn its own discrete units, while text-based LMs must not. Future work might investigate how these slopes change with more data, particularly for the speech-based LM for which 128 hours seems insufficient.

3.3. Language modeling: from clean to in-the-wild speech

So far in the paper, we have little evidence that lexical or syntactic knowledge can emerge in speech-based LMs. To address this concern, we ran one more experiment, this time training STELA under more controlled recording conditions: on up to $1,024$ hours of speech extracted from audiobooks – commonly used to train speech-based LMs [17]. Figure 2 compares this experiment against the performance obtained by STELA when trained on child-centered long-forms (SEEDLingS, Table 3).

Results are unequivocal: we observe a strong improvement on the lexical task for the model trained on audiobooks, while the same model trained on long-forms remains at chance level. On the syntactic task (not shown above), STELA trained on $1,024$ hours of audiobooks obtains an accuracy of $52.8\,\%$ compared to $50.3\,\%$ on long-forms. This is in line with the results in [2] showing that more powerful architectures are necessary to learn at the syntactic level.

Why do we observe chance-level performance when training on long-forms? First, the speech signal found in long-forms is much more challenging than the one found in audiobooks: the speech might be distorted as it is being spoken far from the child; it might overlap with various background noises; and it is

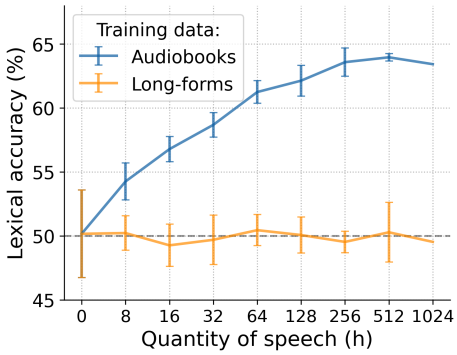

Figure 2: ***Language modeling from clean to in-the-wild speech.*** *Lexical accuracy obtained by STELA trained on audiobooks (Libri-light, in blue) or child-centered long-forms (SEEDLingS, in orange) as a function of speech quantity. Numbers are computed on the test set. Error bars represent standard errors computed across mutually exclusive training sets.*

often produced in short turns that might be under-articulated – see [14] for a comparative analysis. Another essential factor to consider is the domain mismatch between the training and test sets. While the training set contains far-field under-articulated speech as well as close-field storytelling, the test set consists of well-articulated synthesized stimulus to which STELA fails to generalize. However, infants show no difficulties generalizing from uncontrolled real-life conditions to more controlled ones (in-laboratory conditions). We advocate here that generalization is part of the language acquisition problem, and LMs should be evaluated accordingly.

We hypothesize that the discrete units learned by STELA might be too dependent on the various non-linguistic factors found in long-forms, as suggested in [14]. This dependency could prevent the LSTM LM from learning long-term dependencies necessary to solve the lexical or syntactic tasks.

4. Conclusion

Benchmarks are instrumental in allowing cumulative science across research teams. In this paper, we have described how BabySLM has been carefully designed to be adapted to the kinds of words and sentences children hear. We have shown how it can be used to evaluate LMs trained on developmentally plausible text or speech corpus. By doing so, we revealed two outstanding challenges that the community must solve to build more plausible cognitive models of language acquisition. First, we need to reduce the gap between text-based and speech-based LMs, as the latter performed close to chance level on BabySLM. Second, we need to reduce the gap between LMs trained on clean and in-the-wild speech, as evidenced by the striking difference we obtained on the lexical task when training on clean audiobooks versus ecological long-forms.

Future work might consist in evaluating speech-based LMs grounded in the visual modality [28], or linking performances obtained on *BabySLM* with behavioral measures in infants – e.g., age of acquisition as in [29]. A crucial limitation of our benchmark is that it focuses on English, which already accounts for a whopping $54\,\%$ of language acquisition studies [30]. We hope that this paper, together with shared scripts[2], will facilitate the creation of similar benchmarks in other languages.

[2] https://github.com/MarvinLvn/BabySLM

5. References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, A. Baevski, E. Dunbar, and E. Dupoux, "The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling," in *NeuRIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*, 2020.

[3] E. Dunbar, M. Bernard, N. Hamilakis, T. Nguyen, M. de Seyssel, P. Roz'e, M. Rivière, E. Kharitonov, and E. Dupoux, "The zero resource speech challenge 2021: Spoken language modelling," in *Interspeech*, 2021.

[4] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[5] E. Kharitonov, A. Lee, A. Polyak, Y. Adi, J. Copet, K. Lakhotia, T. Nguyen, M. Rivière, A. rahman Mohamed, E. Dupoux, and W.-N. Hsu, "Text-free prosody-aware generative spoken language modeling," in *Annual Meeting of the Association for Computational Linguistics*, 2021.

[6] E. Dupoux, "Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner," *Cognition*, vol. 173, pp. 43–59, 2018.

[7] 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, pp. 389–407, 2022.

[8] A. Warstadt and S. R. Bowman, "What artificial neural networks can tell us about human language acquisition," in *Algebraic Structures in Natural Language*. CRC Press, 2022, pp. 17–60.

[9] P. A. Huebner, E. Sulem, F. Cynthia, and D. Roth, "Babyberta: Learning more grammar with small-scale child-directed language," in *Proceedings of the 25th conference on computational natural language learning*, 2021, pp. 624–646.

[10] A. Cristia, E. Dupoux, M. Gurven, and J. Stieglitz, "Child-directed speech is infrequent in a forager-farmer population: A time allocation study," *Child Development*, vol. 90, no. 3, pp. 759–773, 2019.

[11] S. Dailey and E. Bergelson, "Language input to infants of different socioeconomic statuses: A quantitative meta-analysis," *Developmental science*, vol. 25, no. 3, p. e13192, 2022.

[12] B. Hart and T. R. Risley, *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing, 1995.

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

[14] M. Lavechin, M. de Seyssel, M. Métais, F. Metz, A. Mohame, H. Bredin, E. Dupoux, and A. Cristia, "Statistical learning models of early phonetic acquisition struggle with child-centered audio data," Mar 2022. [Online]. Available: psyarxiv.com/5tmgy

[15] A. Warstadt, L. Choshen, A. Mueller, A. Williams, E. Wilcox, and C. Zhuang, "Call for papers – The BabyLM challenge: Sample-efficient pretraining on a developmentally plausible corpus," 2023. [Online]. Available: https://arxiv.org/abs/2301.11796

[16] B. Börschinger, M. Johnson, and K. Demuth, "A joint model of word segmentation and phonological variation for English word-final/t/-deletion," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013, pp. 1508–1516.

[17] 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 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7669–7673.

[18] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *Journal of Selected Topics in Signal Processing*, 2022.

[19] E. Bergelson, M. Casillas, M. Soderstrom, A. Seidl, A. S. Warlaumont, and A. Amatuni, "What do North American babies hear? A large-scale cross-corpus analysis." *Developmental science*, vol. 22 1, p. e12724, 2019.

[20] G. Le Godais, T. Linzen, and E. Dupoux, "Comparing character-level neural language models using a lexical decision task," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, pp. 125–130.

[21] B. MacWhinney and C. Snow, "The child language data exchange system," *Journal of child language*, vol. 12, no. 2, pp. 271–295, 1985.

[22] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "Celex2," *Linguistic Data Consortium, Philadelphia*, 1996.

[23] R. Weide *et al.*, "The Carnegie Mellon pronouncing dictionary," *release 0.6, www.cs.cmu.edu*, 1998.

[24] E. Keuleers and M. Brysbaert, "Wuggy: A multilingual pseudoword generator," *Behavior research methods*, vol. 42, pp. 627–633, 2010.

[25] M. Lavechin, R. Bousbib, H. Bredin, E. Dupoux, and A. Cristia, "An open-source voice type classifier for child-centered daylong recordings," in *Interspeech*, 2020.

[26] 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.

[27] M. Lavechin, M. de Seyssel, H. Titeux, H. Bredin, G. Wisniewski, A. Cristia, and E. Dupoux, "Statistical learning bootstraps early language acquisition," Dec 2022. [Online]. Available: psyarxiv.com/rx94d

[28] M. Nikolaus, A. Alishahi, and G. Chrupała, "Learning English with Peppa Pig," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 922–936, 2022.

[29] E. Portelance, J. Degen, and M. C. Frank, "Predicting age of acquisition in early word learning using recurrent neural networks." in *CogSci*, 2020.

[30] E. Kidd and R. Garcia, "How diverse is child language acquisition research?" *First Language*, vol. 42, no. 6, pp. 703–735, 2022.