Reverse Engineering Language Acquisition with Child-Centered Long-Form Recordings

Marvin Lavechin,[1,2,3], Maureen de Seyssel[1,2,4], Lucas Gautheron [1], Emmanuel Dupoux[1,2,3], Alejandrina Cristia[1]

[1]Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'Etudes cognitives, ENS, EHESS, CNRS, PSL University, Paris, France
[2]Cognitive Marchine Learning Team, INRIA, Paris, France
[3]Facebook AI Research, Paris, France
[4]Laboratoire de linguistique formelle, Université de Paris, CNRS, Paris, France

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–22

https://doi.org/10.1146/((please add article doi))

Keywords

long-form recordings, LENA, language acquisition, computational studies, ecological validity, reverse engineering

Abstract

Language use in everyday life can be studied using lightweight, wearable recorders that collect long-form recordings - that is, audio (including speech) over whole days. The hardware and software underlying this technique is increasingly accessible and inexpensive, and these data are revolutionizing the language acquisition field. We first place this technique into the broader context of the current ways of studying both the input being received by children and children's own language production, laying out the main advantages and drawbacks of long-form recordings. We then go on to argue that a unique advantage of long-form recordings is that they can fuel realistic models of early language acquisition that use speech to represent children's input and/or to establish production benchmarks. To enable the field to make the most of this unique empirical and conceptual contribution, we outline what this reverse engineering approach from long-form recordings entails, why it is useful, and how to evaluate success.

Contents

1. Introduction

Recent years have seen the rise of data collection through wearable, lightweight, unobtrusive devices that collect audio for tens of hours at a time, allowing a uniquely naturalistic viewpoint of language use as people go about their everyday activities (Wu et al. 2018; Sun et al. 2019; Oller et al. 2010). Although this technique could be used to investigate language use at any age (see Figure 1), it has been extensively used with children; as a result, this body of work has important conceptual, methodological, and ethical contributions that are relevant across fields of linguistics. There exist recent systematic overviews of this prior research (see e.g., Ganek and Eriks-Brophy 2018). Technical aspects of long-form recordings (including step-by-step how-to's and ethical recommendations) have already been largely covered in prior work (see Casillas and Cristia 2019; Cychosz et al. 2020 and Appendix A of our online supplementary materials). Thus, after positioning the methodology in the broader context of language acquisition research methods, we devote most of the article to laying out the promise of the technique in the context of a reverse engineering approach to the study of early language acquisition. Indeed, now that it is possible to capture the full complexity of child language experiences, artificial language learners trained on realistic data can help us build better theories about how humans develop their language perception and production skills.

2. Positioning long-form recordings in a broader methodological landscape

In typical long-form recordings, infants and young children wear a custom-made piece of clothing with a breast pocket, within which a recording device is inserted. This device typically records over many hours. When the full waking day is represented, we may talk of daylong recordings. In this review, however, we will use the term "long-form recordings" to highlight the fact that some researchers do not capture the whole waking day but only 8+ hours; whereas others may also capture nighttime.

Language acquisition can be studied by a variety of means, each of which has unique strengths and weaknesses (See Table 1). It is most reasonable to reflect on the place the use of long-form recordings can have by comparing it with other methods to study (a) children's production and (b) the input afforded to children. By and large, productions and input can be studied jointly using one of four methods: 1. long-form recordings; 2. long observations by third parties (third-party observations, for short); 3. parental reporting; and 4. shorter audio/video recordings. Next, we define each of these techniques and highlight the relative

Figure 1

Examples of wearable recorders. (a) Smartwatch recording audio, heart rate, and movement, adapted from Fig 1 in Liaqat et al. (2018)). (b) Body camera on a South Carolina police officer (Ryan Johnson, CC BY-SA 2.0. (c) A small audio recorder and photo camera worn by a Mayan child in Southern Mexico (by courtesy of Marisa Casillas).

Study method	Context sampling	Ecological validity	Reusability	Complexity
Long-form recordings	High	High	High	High
Third-party observations	Medium-high	Medium-high	Low	Low
Parental reporting	Medium	Low	Low	Low
Short audio-/video-recordings	Variable	Medium	High	Medium

Table 1 Main strengths and weaknesses of the four key methods to study both input afforded to children as well as their production.

strengths and weaknesses by comparing the methods according to the following criteria: **Context sampling** (what proportion of the context of the child experiences is sampled); **Ecological validity** (to what extent the acquired data reflect real characteristics of the situations); **Reusability** (how reusable the acquired data are in light of new hypotheses); and **Complexity** (in analyzing the acquired data).

Long-form recordings rely on a sampling of the full range of a child's experiences in one or several days (and sometimes also during nights) and across all the contexts the child may be in (in or out of the house). Among the four methods we discuss here, long-form recordings are therefore closest to third-party observations, such as those anthropologists employ for time allocation research (Gross 1984) and those that psychologists use for some behavioral observations. For instance, Roopnarine et al. (2005) observed families for 3 hours at a time in 4 separate visits, each time completing a checklist of observed behaviors every 30 seconds. Long-form recordings and third-party observations have the relative advantage over other techniques that the child's carers do not have to do anything special (not even

stay in the same room as the camera). Furthermore, the novelty effect of having an observer should, if anything, decrease over the long observation period, resulting in greater ecological validity for long observations and long-form recordings than the other methods.

Nonetheless, long-form recordings present three advantages compared to third-party observations. First, the observer is less salient, which may result in even lower awareness and fewer perturbations of the natural behavior of participants (high ecological validity). Second, setting aside the potentially longer initial investment to learn the technique and obtain ethical approval, long-form recordings require less effort and time from experts than third-party observations, particularly since the recorder can be mailed. In terms of reusability, recordings, unlike third-party observations, can be consulted, reannotated, and reanalyzed, including to measure behaviors that had not been considered before collecting the data. That said, third-party observations have an important advantage over long-form recordings in that the observer has access to multimodal cues and other information, allowing more nuanced interpretations using the full 360° context. Moreover, such third-party observations often rely on standardized checklists, which are then easy to analyze (low complexity) while long-form recordings require the use of manual and/or automatic annotation tools to extract information of interest (high complexity).

Among parental reporting techniques, the method closest to long-form recording is probably the use of smartphone apps or a similar set-up to collect reports from caregivers over a long time period. Diaries that ask bilingual parents to report on how frequently they use one or another language, at different moments of the day over a whole week, would fall under this category (e.g., Orena et al. 2020). Although apps are not yet prevalent, they could be useful for sampling behavior at different timescales. For instance, we could ask the caregiver to report who is talking to the child right now (through push notifications that pop up at different points in the day); or to report which words the child says or understands at different child ages (as in Wordful[1]). Like long-form recordings and third-party observations, such reports could sample from the full range of experiences afforded to children. That said, there could be reporting biases due to relying on caregiver report, lowering both lower context sampling and ecological validity. As to the former, caregivers may be less able to respond to the app's request about who is talking to the child when they are engaging in hygiene or other "hands-on" routines. As to the latter, the caregiver will be keenly aware of being observed when reporting on their own behavior, and may align their reporting with their beliefs instead of accurately representing what occurs; e.g. a bilingual parent who consistently reports they use each language 50% of the time. Another limitation of parental reports is that, as with third-party observations, they cannot be revisited to code other behaviors not foreseen in the original design (i.e., low reusability of the method). Nonetheless, parental reports have a key relative advantage over long-form recordings and third-party observations in that caregivers can incorporate their background knowledge about the family and the child in their interpretations. Similarly to third-party observations, data acquired through parental reports are also easier to analyze than long-form recordings (low complexity).

Finally, shorter audio/video recordings are probably the most common way in which psycholinguists have described children's input as well as their production. For example, Bergelson et al. (2019) studied input and production in a longitudinal study on infants aged 6-18 months by setting up a video recorder on a tripod and having each infant wear two head-

[1] `http://wordfulapp.com/`

mounted video-recorders. Families were thus recorded at home for one hour, after which the researchers returned to pick up the equipment. Such shorter audio/video recordings share with long-form recordings the relative advantage that they can be revisited as new hypotheses arise (high reusability). One relative disadvantage of short observations over all other methods discussed so far is that investigators must choose whether they want to keep activity heterogeneity low, by asking all families to record during a specific activity (e.g., meal time, play, hygiene), or whether they want to represent the full range of experiences, in which case they still need to choose how to sample from each (e.g., whether to record the same number and length of each, or whether to sample them with the frequency with which they occur in a natural day). Such short observations also probably result in increased consciousness of being observed, which potentially affects participants' behavior and lowers ecological validity. This also leads to a key advantage of short recordings over all the other methods, which is that the investigator can purposefully target an activity or setting that is most relevant to their purposes, for instance by providing a set of toys that leads to an increased use of a relatively rare structure (e.g., eliciting defining and non-defining relative clauses with a purpose-made book or deck of cards). Although shorter audio/video recordings are less complex to analyze than long-form recordings, working on audio and/or videos still requires the use of automatic or manual annotation.

Importantly, no method is perfect, but they are to a certain extent mutually compatible, such that researchers can try to design data collection using one or more in a complementary scheme. For example, Bergelson et al. (2019) collected one full day's audio recording in addition to the hourlong video recording, on separate days, and found some diverging results (notably a larger quantity of speech in the hourlong videos than in the long-form audios), as well as considerable convergence (for instance, in terms of who spoke to the child). Additional mixed methods can be devised to serve the researcher's goals: An investigator could ask families to record over two full days, and send in an observer for part of one of those days; or they could ask the families to play an elicitation game, and write down the time and day they did so. The investigators could then extract the sections of the recording that contain these dual methods, and analyze them further, in order, for instance, to establish the extent to which behaviors are affected by the presence of a third-party observer in the first example, or simply to transcribe and study the speech occurring during the elicitation game in the second example.

Without denying the complementarity of the four methods, we do believe that there are three ways in which long-form recordings are unparalleled. First, long-form recordings are a promising technique to collect naturalistic big data in various populations. Nielsen et al. (2017) documented that developmental journals are heavily skewed towards publishing articles with data from WEIRD (Western, educated, industrialized, rich, and democratic) populations. This sampling bias can lead behavioral scientists to wrongly attribute culturally specific findings as universal traits. Although researchers using systematic behavioral observations and short recordings have certainly made an attempt to broaden the languages and cultures represented in the literature, both of these methods require so much investment and expertise that, in reality, it is mostly outsiders who document language acquisition in such settings. Therefore, despite our best intentions, we may misrepresent the language and culture, and further our research questions and output may not have the optimal impact they can have on the population from which the participants are drawn. Admittedly, most current adopters of long-form recordings are WEIRD Cychosz and Cristia (ming), but we hope that this method will be increasingly used by diverse researchers, including

members of underrepresented and underserved linguistic and cultural communities, so that the mainstream literature can better represent their viewpoints and interests, and so that these populations stand a higher chance of benefiting from the research.

Second, long-form recordings may be ideally suited to address current needs for replicable and reproducible research. To begin with, reproducibility is heightened by the use of audio and video archiving and sharing repositories such as HomeBank (VanDam et al. 2016) and Databrary (Simon et al. 2015), in the wake of the CHILDES tradition (MacWhinney 2000). What is more, by capturing a maximally unbiased sample of the child's language experience, while ensuring maximal ecological validity, the use of long-form recordings should, overall, increase the probability of conceptual replications. Additionally, many of the analyses rely on automated methods that are shared across many laboratories. As a result, it becomes easier to quantify (and possibly fix) biases that may be present in measures extracted by these automatic tools than when relying on human annotation.

Third, such recordings may be ideally suited to foster a new direction of research within the broader field of modeling early language acquisition, namely a reverse engineering approach to the study of infant language development. We will dedicate the rest of this article to laying out this new research direction.

3. Reverse engineering language acquisition

There is a long tradition of modeling in the context of language acquisition (e.g., MacWhinney 2005). A complete review would be beyond the scope of this paper, but to illustrate the field of possibilities we can cite Anderson (1975) for an example of a syntax-learning model or Brent (1996) for a word-discovery model. While computational models of language acquisition traditionally assumed that speech was represented as an error-free string of adult-like phonemes (which is unlikely the case for infants), more recent studies address the problem of language learning from raw speech. This line of research can be illustrated with Nguyen et al. (2020), the last iteration of a challenge[2] organized in the speech processing community that revolves around spoken language modeling without annotation or text.

In view of the substantial progress of these past years in algorithms that can learn from raw speech, there are clear interests in a greater integration of AI and language development studies. Dupoux (2018) set down recommendations for such an enterprise, which we will not repeat here. Like Dupoux, we assume that unsupervised language learning models should be exposed to realistic data; and that they should be evaluated on psycholinguistic benchmarks to compare humans' and machines' language capabilities at various linguistic levels. In this new research direction, child-centered long-form recordings play a crucial role by providing artificial language learners with ecologically valid input data in the form of the speech by adults and other children present in the audio recordings, which is then directly comparable to the input afforded to children. Our proposal builds on and extends Dupoux (2018) by spelling out how a research program centered on long-form recordings could proceed, considering such recordings as an information source on not only children's

[2]Challenges in the machine learning community are events during which participants (usually researchers including students) work on improving the performance of a baseline model on one or multiple tasks (from audio classification Schuller et al. 2019 to unsupervised language modelling Nguyen et al. 2020)

input, but also children's production.

In Section 3.1, we describe the reverse engineering approach to the study of infant language acquisition. In Section 3.2, we explain why language acquisition researchers should consider using machines as a proxy to study infants. We then lay out how to study input effects in Section 3.3. Section 3.4 presents our psycholinguistic-driven framework to measuring language skills of artificial language learners. This benchmarking framework is based on behavioral correlates of language learning observed in humans and allows us to compare the language skills of the artificial language learner with those of the human.

3.1. What does the reverse engineering approach to the study of first language acquisition include?

Since we are discussing reverse engineering in the context of long-form recordings, by and large the experiences under discussion are unimodal, based only on the speech, as the vast majority of long-form recordings being gathered are audio only. That said, recent AI work begins to address the problem of learning language from audiovisual exposure (Chrupała et al. 2017; Harwath et al. 2020; Alishahi et al. 2021), although admittedly these studies do not use realistic data. Using long-form data to train these sighted artificial language learners would require using devices that capture both what children hear and what they see (such a set up has been used in Casillas et al. 2020b,a for instance). Capturing child language experiences across multiple modalities would offer us opportunities to compare audio-only models with audiovisual models and could help us better understand the role of visual experiences during the language acquisition process. Importantly, while audiovisual long-form recordings could be used, touch and smell can not (yet?) be digitized, particularly at the long-form scale, yet these senses can help in the language acquisition process (Abu-Zhaya et al. 2017). This is a current limitation of the technique, and thus we discuss only audio-based and video-based models in this review.

By and large, the models we discuss in this review are passive learners, in the sense that they cannot affect the input data they receive. This aspect of the models makes them somewhat different from human children, who are able to explore and interact with their environment. For example, if a learning human child formulates the hypothesis "Cats are those little hairy animals", and wants to check whether this hypothesis is true, the child could interact with their environment to prove or disprove the hypothesis, such as by pointing at the cat while waiting for a caregiver's reaction. This connects with the importance of embodiment in first language acquisition (Yu 2014) and of the child's role in shaping their environmental input (Tamis-LeMonda et al. 2018). Although the artificial learner may benefit from the child's interactions with their environment, if the two types of learner are not at the same learning stage and/or have not formulated the same hypothesis, ultimately the child will profit more of their own experiences than will an artificial learner who is simply reexperiencing the human child's experiences.

With those considerations in mind, the reverse engineering approach to the study of first language acquisition via long-form recordings can be summarized as follow:

1. We design a computer program to have some learning mechanism(s) we believe are useful to learn language (i.e., we control the mechanisms). Those are discussed in the current section.
2. We provide this program with a controlled and realistic language experience (i.e., we control the input), see section 3.3.

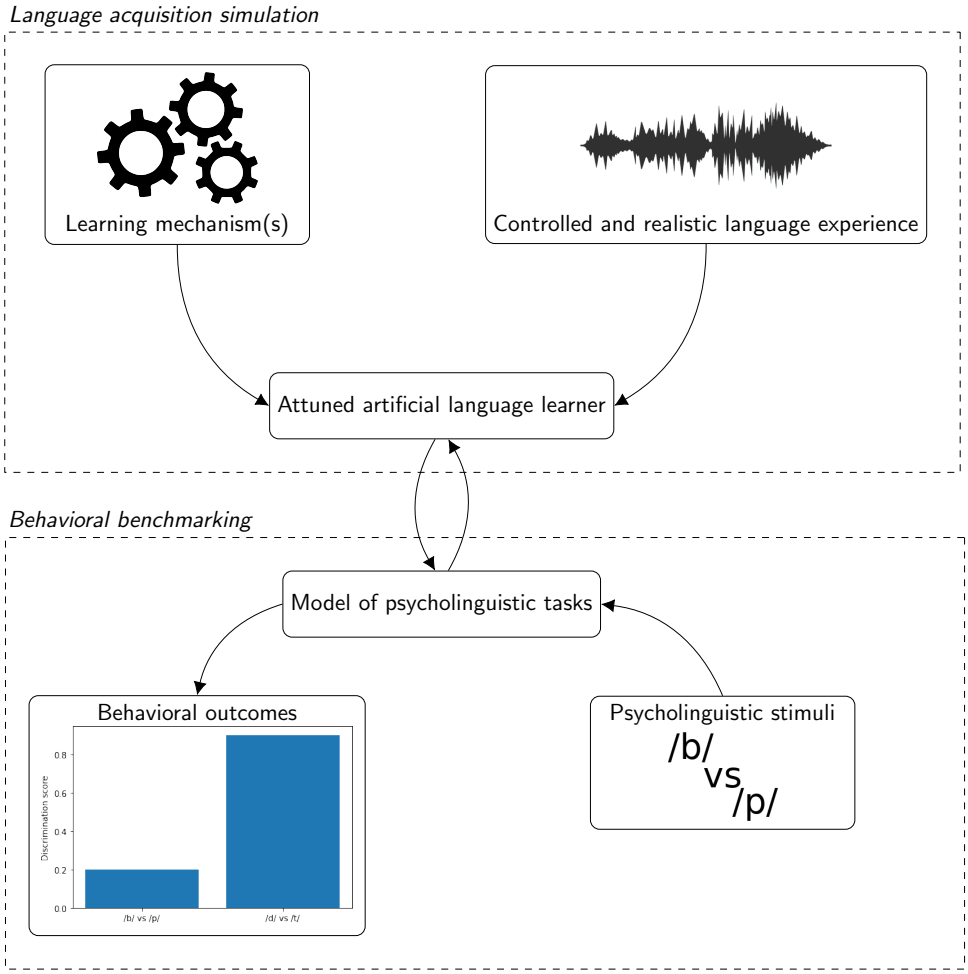

Figure 2

Outline of the reverse engineering approach for language acquisition. First, we perform a language acquisition simulation, by applying learning mechanisms on a controlled and realistic language experience to derive an attuned artificial language learner. These simulations can be used to check for the effect of variation in those language experiences (e.g., multilingual input, variation in household size; see Section 3.3). Second, we use behavioral benchmarking to evaluate learning: the attuned artificial language learner undergoes a battery of psycholinguistic tasks. These tasks are conceptually related to the tasks that are used to study humans, but they are adapted for the machine (see Section 3.4). Note that the psycholinguistic task can also contain a habituation phase that will further modify the state of the attuned artificial language learner (hence the two-sided interaction between the two).

3. We observe what the system learns and how it develops (i.e., we observe the learning outcomes[3]), see section 3.4.

[3]Note that we distinguish output (i.e., production) from outcome (i.e., consequences, that could

Let us take the example in which our goal is to understand *perceptual development*. In such as case, there is often a learning phase and an evaluation phase. The learning phase includes exposing the artificial language learner to a naturalistic and controlled language experience (typically represented by adult speech extracted from child-centered long-form recordings). Then comes the evaluation phase, in which the attuned learner (i.e., the model postexposure) undergoes a battery of psycholinguistic tests. These psycholinguistic tests are conceptually related to experimental protocols used in child studies to assess the language capabilities of infants (looking-while-listening procedure, conditioned head turning, etc.). Behavioral patterns extracted from those tests can then be compared to those observed in humans undergoing the experimental version of the psycholinguistic tests. Figure 2 provides a diagram illustrating this version of the reverse engineering approach.

Having described the approach, we can now discuss the benefits of using in silico modeling to study infant language acquisition.

3.2. Insights that artificial language learners can provide us

Despite thousands of laboratory experiments that isolate learning mechanisms in babies, thousands of hours of observations, and probably millions of hours of study of both bodies of work, still little is known about the inner mechanisms that underlie human language development in the wild, the causal role of the input afforded to the child, and the best explanation for the perception and production outcomes exhibited over the course of acquisition. This is not truly our research community's fault. Language acquisition, like most cognitive processes, is essentially a *black box* and can be studied only as such by language development researchers.

It is important to note that the term "black box" is also used in the AI community and can be defined as the inability to predict an AI model's decisions. Even though the state of the model is fully known at any given time, this state is so complex that its inner processes and workings cannot be fully understood.

It therefore appears that Language Acquisition and AI researchers face a common challenge: studying a black box. So one may wonder: why would one use a black box as a proxy to study another black box?

First, artificial language learners are tireless. We can study their responses over a large quantity of stimuli in ways that would be impossible (or unethical) with infants. Observing behavioral patterns in machines may inspire psycholinguists to run new human studies. Indeed, if a behavior of interest is observed in machines but has never been documented in infants, then it may be worth checking whether infants also exhibit this given behavior. Using thousands or millions of stimuli allows us to generate robust predictions.

The second advantage of the AI black box over the infant's is the adaptability of the former: we can easily tweak an AI model's input data as well as parameters of the learning mechanism(s). Thus, we can observe consequences of the simulated language experience on the language skills of the machine while keeping control over the learning mechanisms, and vice versa.

If there is a mismatch between the learning outcomes of an artificial language learner and the human learner, then we can think of ways to make the machine more human-like. In cases in which we are using input that faithfully represents the input that children are

be visible in production or perception).

exposed to, we can rule out an issue with the input, and thus infer that the mistmatch might be due to other aspects of our experiment: namely, 1. the learning mechanisms in the model; 2. details of the experience presentation (e.g., the quantity of data); or 3. the computational implementation of the psycholinguistic tests used to measure learning outcomes.

In the opposite case, if a computer program yields outcomes similar to those of the infants when provided with the same experience infants had, then we are faced with the tantalizing conclusion that we have found one viable model (probably among many others) of human behavior. Then, based on the results of that viable model, we can suggest new testable predictions regarding how infants' language develops.

3.3. Controlling the input to measure its downstream effects

In this section, we assume that we hold mechanisms constant, and illustrate how the reverse engineering approach can shed light on the causal role of the input, with the goal of measuring learning outcomes through procedures described in Section 3.4.

One way in which we can use this approach is by comparing the input of different infants or groups of infants. Some discussions of, for instance, group variation are based on the idea that some children may be afforded qualitatively (Brookman et al. 2020) or quantitatively (Weisleder and Fernald 2013) different input. Instead of simply describing qualitative and quantitative differences between the children or the groups, a reverse engineering approach would expose the exact same artificial learner to long-form recordings of these different groups, and then assess whether there are substantial differences in learning outcomes.

In fact, one advantage of a reverse engineering approach is that we can go beyond attested environments. Since we can control the quantity of data the artificial language learner is exposed to, we can simulate data deprivation and/or proliferation experiments. We can also plot developmental curves that show the evolution of the model's language skills as a function of the quantity of data it has had access to.

Similarly, some research studies the language input of children exposed to multiple languages using long-form recordings (Orena et al. 2020). Along with checking for outcome differences after exposure to naturally occurring multilingual audio, we can more precisely simulate additional cases by creating bilingual and/or multilingual corpora. This allows us to precisely control for the distribution of these languages during the exposure phase and to use all languages to check for skills in each language during the evaluation phase.

The approach can be generalized to many other aspects of language exposure that are currently hard to control and tease apart, although for some it may be harder to do so well. For instance, we think it is technically possible to vary speaker distribution in terms of household size, number of siblings, and gender distribution, because we can start with recordings done in households with only one adult caregiver and only one sibling and then mix them together to create families with on to ten caregivers and one to ten "siblings", since in the original setting we can trust that adult and child sections will have been correctly attributed to one adult and one other child respectively. Even so, the "siblings" in these simulated large families will be easier to tell apart than siblings found in natural recordings of large families because they will not necessarily sound similar to each other (since they are originally drawn from different families). Other dimensions may be hard to create because the tools they rely on are not necessarily mature yet. For instance, varying vocabulary size would require an accurate automatic speech-to-text model, which

is beyond the state of the art at present; varying the frequency of book readings would require automatically extracting book-reading interactions, but there is no automatic tool to detect such interactions to date. Nonetheless, and given the fast pace of AI research, it is useful to think about this feature of in silico modeling in general terms.

3.4. Evaluating language skills of the artificial language learner

Let us start by assuming that we want to check what was learned. In this case, one may be tempted to ask: do models learn phonemes, nouns, and so forth? This is akin to attempting to establish whether the model has a given linguistic representation. We see two problems with this option. First, it is devilishly difficult to demonstrate that a representation is in place, doing so involves, for instance, additional levels of agreement in terms of how to prove its presence (e.g., for phonemes Jaeger 1980; for a general criticism on inferring mental entities, see Twaddell 1935). Second, representations are precisely the area in which theories of language (acquisition) diverge in the fiercest manner (e.g., Ambridge and Lieven 2015). Although some may disagree with us, we would like to posit that psycholinguistic benchmarking tasks should be theory-agnostic. By not forcing extra assumptions of representations on either the human or the machine learner, the findings have a higher chance of being relevant to a broad range of theories.

A second definition of learning may be to ask about representations in terms of their neural implementation. In the avenue of human-machine comparison, an approach that has interested the AI and neuroscience communities involves comparing activation patterns of neural networks with those of the human brain while being asked to perform a similar task (e.g., Yamins et al. 2014). While this approach promises interesting scientific insights about human brain information processing, it also shows some limitations in the context of early language acquisition. First, data acquisition devices that locate activity precisely in both time and space are very seldom used with infants (but see e.g., Bosseler et al. 2021). Second, these techniques require large sample sizes for reproducible results even among adults (Turner et al. 2018), and given that measurements in infancy are typically even noisier than those gathered in adulthood, we can infer that it is possible that the body of literature on infant neuroimaging will require substantial accumulation of results before we can employ it for our benchmarks.

Therefore, given the current theoretical and methodological landscape, we propose a behavior-oriented approach, in which we study the behavior of infants in parallel to the behavior of machines. [4]

To extract behavioral patterns for the machine, we additionally need to reflect on how numbers returned by the machine can be related to the kinds of behaviors elicited and/or observed among human learners. By numbers, we mean either the output representations[5]

[4]In this review, we are simplifying matters by not going into detail about the fact that, in reality, a perfectly comparable artificial learner would develop not only language but also all of its other cognitive systems, which would allow the influence of attention, memory, and other systems orthogonal to language to be accurately represented in the model. In other words, this would entail not only modeling language development but also modeling how the child approaches the psycholinguistic task. For relevant discussion, see Robinaugh et al. (2021).

[5]Representations learned and returned by the model should not be confused with linguistic representations. Output representations returned by the machine are numbers describing and/or organizing the input that was given to the model. By linguistic representations, we mean hypothesized mental units that represent elements of language (e.g., phonemes, morphemes).

of the input stimuli for *perception models*, or the vocalizations produced by *production models*.

With that in mind, we turn to the following question: What behaviors do infants and adults exhibit through the course of language learning, and which could serve as benchmarks for artificial learners? We discuss these in two subsections, targeting perception (Section 3.4.1) and production (Section 3.4.2).

3.4.1. Measuring perception. Much of the language acquisition literature has attempted to look at perception, mainly through infants' reactions to specific stimuli in clearly defined laboratory tasks. This work suggests that much goes on in the child's mind even before there are obvious changes in production. In this section, we reflect on how long-form recordings paired with computational modeling can help us understand the development of these markers.

As described in Section 3.1, there are two phases in the reverse engineering approach. The first one involves exposing the artificial language learner to a controlled and realistic language experience extracted from long-form recordings, which can be selected to vary certain dimensions parametrically, as explained in Section 3.3. As described in this section on perception, the second phase is one of evaluation of perception skills. Since perception cannot be directly investigated in either human or artificial agents, in this section we rely on behavior that is elicited in controlled conditions. Thus, the AI learner is presented with stimuli like the ones submitted to children in the laboratory, to elicit numbers that can be interpreted as behavioral perceptual patterns in the artificial learner. In other words, we create a computational implementation of infant perceptual benchmarks, to which the artificial language learner is submitted. In Table 2, we draw examples mainly from studies on infants aged from 0 to 12 months, but we trust our reasoning can be generalized to perception tasks beyond 1 year of age. We want to highlight that this is not an exhaustive list, and we would be delighted if other researchers created computational implementations of other infant perceptual benchmarks.

The top section of Table 2, dedicated to sound-only behaviors, represents experiments where the audio is key for eliciting infants' responses, whereas the second half of the Table shows cross-modal behaviors where both audio and visual stimuli are used.

Test tasks may be of two types: distance-based and probability-based. Distance-based benchmarks rely on computing the distance between different stimuli, whereas probability-based benchmarks require the machine to compute the probability of each stimulus. We illustrate these concepts with examples below.

Let us start with a *distance-based* example. Newborn humans can discriminate across rhythmically distinct languages (Nazzi et al. 1998). One can measure a discriminability score in the machine by submitting to it three audio stimuli: A, B and X such that A comes from a first language $\mathcal{L}_1$, B from a second language $\mathcal{L}_2$, and X comes from $\mathcal{L}_1$ but different from A. Under a distance function d, one may expect that $d(A, X) < d(B, X)$ as A and X have been drawn from the same language. While this might not be true for a given stimulus, repeating the procedure across thousands of stimuli allows us to extract robust patterns in the artificial language learner. In other words, we use here the computational implementation of the ABX discrimination task used in psychology. An example of a computational study using this task can be found in de Seyssel and Dupoux (2020), who tested the language discrimination capabilities of an i-vector model to assess the role of monolingual versus bilingual exposure, akin to studies on monolingual and bilingual human infants. The same distance-based

Sound only behaviors	Age (mo)	Task	Data set	Literature
discriminate across rhythmically distinct languages	0	distance-based	bilingual set of stimuli	Gasparini et al. (2021)
discriminate native and non-native consonants	6-8	distance-based	phonetically aligned clean speech	Werker and Tees (1984)
accept novel content words more easily than novel function words	6	few-shot learning + probability-based	jabberwocky sentences	Shi et al. (2006)
prefer high over low phonotactics	9	probability-based	made-up words varying in phonotactics	Jusczyk et al. (1994)
prefer high over low frequency content words	11	probability-based	real words varying in frequency	Jusczyk et al. (1994)
do not discriminate non-native consonants	12	distance-based	phonetically aligned clean speech	Jusczyk et al. (1994)
Cross-modal behaviors	**Age (mo)**	**Task**	**Dataset**	
treat words and monkey vocalizations, but not beeps or coughs, as possible labels	3	few-shot learning + distance-based	images paired with words, monkey vocalizations, beeps or coughs	Ferry et al. (2010)
treat words but not monkey vocalizations as possible labels	6	few-shot learning + distance-based	images paired with words or monkey vocalizations	Ferry et al. (2010)
treat content but not function words as possible labels	6	few-shot learning + distance-based	images paired with function words or content words	Hochmann et al. (2010)
know the meanings of many common nouns	6-9	distance-based	images paired with common nouns	Bergelson and Swingley (2012)
few-shot learning of new word-object pairings	9	few-shot learning + distance-based	images paired with words	Yeung and Werker (2009)
treat words with native but not non-native sounds as possible labels	10	few-shot learning + distance-based	images paired with L1 words and L2 words	May and Werker (2014)

Table 2 A sample of human behavioral correlates of language skills that have been reported in the literature along with their computational implementation. The Task column describes the task that is meant to be submitted to the artificial language learner. Distance-based tasks consist of computing the distance between the output representations of the input stimuli. Probability-based tasks consist of computing the probability of the output representations. Few-shot learning tasks involve a learning phase during which the model is given some examples. The Data Set columns describe the test stimuli that need to be gathered to submit the task of interest. Labels are audio stimuli consistently paired with a visual stimulus, e.g. a monkey vocalization systematically followed by a fish picture. The Literature column suggests entry points in the psycholinguistic literature.

method can be used to benchmark the machine's phoneme discrimination capabilities. In this setup, A, B are X are triphones with A and B differing only in their center phone

(/bet/ vs /bat/), and X the same triphone (but another occurrence) than A. In the same way, one may expect that $d(A, X) < d(B, X)$ as A and X represent the same triphone. Phoneme discrimination capabilities have been evaluated on a Gaussian mixture model in Schatz et al. (2021) who notably showed that a model exposed to Japanese exhibits a lower discrimination score on the [ɹ] vs [l] contrast than does a model exposed to American English. Incidentally, note that Schatz et al. (2021) concluded from their results that the AI learner solves this task without *phonemic representations* per se, exemplifying one way in which not making assumptions about representations may facilitate cross-pollination of findings between developmental science and computational research.

Next, we turn to a probability-based example. At the age of 11 months, infants have been shown to prefer high- over low-frequency content words (Carbajal et al. 2020). Checking that this behavior is present in the machine would consist of submitting to it two stimuli, A and B, with A drawn from high-frequency content words and B drawn from low-frequency content words. One should observe that the probability the model returns for A is higher than the probability returned for B, as the first one is supposed to be more frequent than the second one in the training set.

The same two benchmarking approaches (probability- and distance-based) can be adapted to a cross-modal setting, in which decisions need to be taken by integrating information from the auditory and another (typically visual) modality. As for the visual modality, a task used to test infants' comprehension of words and sentences that seems particularly easy to adapt for the machine is the looking-while-listening task (Fernald et al. 2008). In this task, the child sits in front of a screen that shows two images, only one of which corresponds to the audio the child is concomitantly presented with. During the computational implementation of this test, the machine would receive an audio stimuli A_1 (e.g., "nose") and would be presented with two images, one of them representing the audio stimuli I_1 (e.g., a picture of a nose) and the other one, I_2, representing something else (e.g., a picture of a mouth). The machine would then be asked to output the representations of all three stimuli. To check if the machine was able to map the audio stimuli to the right image, we would consider a distance function d and check that $d(A_1, I_1) < d(A_1, I_2)$ as A_1 and I_1 share the same semantic content. Note that, alternatively, the joint probability distribution between a word and an image might be used instead.

In Table 2, we highlight the fact that such a test phase can be preceded by a learning phase during which the machine is presented with some examples. This phase is called few-shot learning – few-shot means that only a small number of exposure instances are used before the evaluation. For instance, these exposures can take the form of nonsense words in the audio-only setting or image-word pairs in the cross-modal setting.

Finally, let us note that the audio stimuli does not have to be a real word. For instance, a fish picture can be paired with a monkey vocalization, in which case we would evaluate the ability of the machine to learn this new word-object pairing. Indeed, a sizable literature in developmental research investigates the (presumably innate) biases infants bring to word-learning tasks, finding that young infants exposed to words or monkey vocalizations systematically paired with a visual category (e.g., dinosaurs) will generalize the "label" to a new exemplar of the same visual category, whereas the same behavior is not observed when dinosaur pictures are systematically paired with beeps or coughs (an overview of this line of research in Vouloumanos and Waxman 2014). To our knowledge, similar biases have not been investigated in artificial agents.

In sum, the probability- and distance-based evaluation paradigms are extremely pow-

erful. They have already been used in the ZeroSpeech 2021 challenge for computational implementations of human psycholinguistic benchmarks across multiple linguistic levels, including phonetics, lexicon, semantics, and syntax (Nguyen et al. 2020). Nonetheless, it is important to bear in mind that most of the previous computational studies use relatively manicured recordings (such as audiobooks), and to our knowledge, none of them has tried to tackle language learning after exposure to audio from long-form recordings.

3.4.2. Measuring production. In this section, we turn to production, although we start by admitting that this line of research is a great deal less developed than the perception one, and it may take considerable time to make progress in this area. As with perception, the development of production involves learning mechanisms that are still the matter of active debate (Long et al. 2020).

There are two key differences for how perception and production reverse engineering approaches can work. First, models of perception development require the extraction from long-form recordings of only the speech that represents children's input, whereas to model production development it is worthwhile to extract both the input and the child's output. In fact, when considering production development, there is a strong case to be made about children's production being shaped by their own output – and thus their output may, under some accounts, be considered as input too. Second, while perception models are only required to return representations of their input, production models need to integrate those evolving representations of the input; with 1. (potentially changing) biophysical constraints on production (due to the fact that the child's body, including their tongue, is changing with age); 2. mechanisms for learning-related changes in production; and 3. some system for actually generating vocalizations. That said, not all extant models of children's language production consider all of these aspects, and others actually consider additional constraints, such as social constraints (Pagliarini et al. 2020). To take a specific example, Warlaumont et al. (2011)'s model has an articulatory component (which generates the child's output in terms of gestures) as well as a perceptual auditory component (which captures patterns in the input as well as the auditory consequences of the child's production) – but note that the articulatory component does not include biophysical constraints per se. Approaches with realistic models of the developing vocal tract are rare (but see Philippsen 2021).

At least in principle, then, a reverse engineering approach to production development that builds on long-form recordings would proceed as follows: If one believes that input speech can affect production, then the model should use the input to hone perceptual representations (i.e., Section 3.4.1); in all cases, one can use children's production as a benchmark against which to compare the model learner's production. As discussed in Section 3.1, one can control the learning mechanisms and the input, in order to measure outcomes.

In reality, however, we see a considerable gap between how production is typically modeled and long-form recording data, for both the perception and the production aspects of production development. In current work aiming to model production development, input is most typically represented in a simplified manner (e.g., with first and second formant values representing vowels), and output is similarly reduced to such summary representations (although exceptions exist; see e.g., Rasilo and Räsänen 2017). What is more, it is not uncommon to see evaluations akin to an elicited imitation task, where the system is provided with an adult vocalization as input and is asked to imitate them (i.e., produce the articulatory activations corresponding to this auditory input). Such a benchmark does not seem

realistic for children under 2 years of age, given that eliciting repetition is methodologically challenging even at around 20 months (albeit possible, see Hoff et al. 2008). Moreover, it is unclear that studies are actually referring to what human children's performance is in actual imitation tasks. Instead, much of this work appears to operate under the assumption that imitation is prevalent in real-life interaction. However, laboratory observations suggest imitation is vanishingly rare (Athari et al. 2021), even in a setting where parents may be driven to increase their interactions with their child as a consequence of being observed. Convergence, which is a form of imitation, was not found to be systematic in the analysis of long-form recordings (Seidl et al. 2018)

A second issue standing in the way of relating long-form recordings and production models concerns the fact that researchers of production development typically specialize on certain development phases and phenomena. For instance, some study the emergence of syllable structure (Warlaumont and Finnegan 2016), and others study vowel targets (Rasilo and Räsänen 2017). This specialization entails that models of production developed at present do not generate the whole range of vocalizations observed in long-form recordings, but are instead dedicated to features of vocalizations.

When we look at production development more broadly, we see many aspects that should be accounted for, including the following:

- The presence of both speech-like and non-speech-like vocalizations (Long et al. 2020)
- The increase in canonical vocalizations (having at least one adult-like consonant-vowel or vowel-consonant transition) with age (Cychosz et al. 2021)
- The appearance of meaningful utterances, starting with single words (de Boysson-Bardies and Vihman 1991)
- The appearance and increased prevalence of word combinations (Braine and Bowerman 1976)

Not only are these aspects out of reach for any extant model learner, but also the basic description of these phenomena in long-form recordings is rare. In fact, it was only recently found that speech-like vocalizations are prevalent in long-form recordings even among newborns (Long et al. 2020); and that the proportion of vocalizations containing canonical transitions appears to continually increase well beyond the first year of age, according to long-form data (Cychosz et al. 2021). Information from long-form recordings on the other phases (namely, the appearance of meaningful speech and of word combinations) has been documented in only two studies (Casillas et al. 2020b,a), both of which employed human annotation. It would be ideal to develop automated techniques so that they can be applied at scale in multiple languages. Moreover, further development is needed to create computational implementations of these human psycholinguistic benchmarks if we are to succeed in our goal of comparing humans and systems.

In sum, reverse engineering the development of production is farther away on the horizon than the study of perception, awaiting conceptual advances in modeling approaches, which may also necessitate important changes in the way we do descriptive analyses of production data gleaned from long-form recordings. As briefly noted, some work in production also takes a social stance, incorporating caregivers in the developmental loop. Progress in such conceptual settings will require even more work, as they necessitate reverse engineering the caregiver as well.

3.4.3. Limitations of the human/machine behavioral comparison. Before concluding the review, we want to highlight some limitations of our benchmarking approach. All of it relies on the psycholinguistic human data being empirically solid and unbiased – and we believe progress on both of these fronts is necessary to support the backward and forward loop between humans and machines.

To begin with, Table 2 may continue to perpetuate the illusion that infants' skills can be described with simple statements. In reality, conclusions drawn from child studies are rarely as clear as "children do X" or "they don't do Y". Results may vary depending on the sample size, the methodology used, and the age of the participants, such as reported in a study of 12 meta-analyses in Bergmann et al. (2018). Ultimately, we should probably instead look at the distribution of effects sizes emerging from meta- or mega-analyses, rather than an arbitrary yes/no binomial assessment. This point applies both to perception and to production benchmarks.

Another issue with evaluating AI learners against extant human infant literature comes from the fact that this literature is biased towards specific populations and languages. A study of three leading developmental journals (where perception experiments are often published) showed that over three-quarters of their papers bore on North American and/or European infants (Nielsen et al. 2017), and a study in the Journal of Child Language (which often publishes articles on children's naturalistic production) showed a shocking 69% of the papers bore on English learners, with a mere 15% bearing on non-Indo-European languages (Slobin 2014). And although there is less evidence about this, it is likely that the samples from, for instance, North American infants are not representative of the greater populations. Thus, a characteristic observed in an American infant growing up in a high-socioeconomic-status setting may or may not be observed in infants growing up in other communities. In the absence of systematic observations across cultures, the AI learner seems doomed to reproduce bias found in the language acquisition literature. Several researchers are working hard to collect long-form recordings from more diverse populations (see e.g., Cychosz et al. 2021) and we hold out hope that, at least in terms of long-form audio, the bias may be weakened in years to come. However, for our perceptual benchmarks we require something like laboratory experiments, and there are currently very few researchers collecting perception data from more diverse communities (but look out for Marisa Casillas' output in coming years; Casillas et al. 2020b; and the ManyBabies efforts).

Setting aside these two issues of robustness and bias in the data, we foresee that further work is needed to reflect on which tasks we want to incorporate in our benchmark. Being able to solve a given task does not tell us if solving this task is *required* for language learning. For instance, divergent discrimination responses to rhythmically similar versus different languages may be neither necessary nor sufficient for language learning. Alternatives include that this difference in behavior is an acquired response bias; or a side effect of auditory development as affected by ambient sounds (similar to how infants prefer their mother's voice but not their father's voice at birth, Lee and Kisilevsky 2014), in other words, behaviors that emerge but that are neither necessary nor sufficient for language learning. If such a task is used to evaluate the AI learner, observations in humans and machines may be divergent for uninteresting reasons.

That said, we think this problem is less worrisome than the other two. In fact, deciding which behaviors necessarily occur as a function of language development could be a problem with which computational models can help. The intuition is that if a behavior is necessary and sufficient for language development, then it should be systematically ob-

served for any and all artificial agents that do acquire language. Large-scale cross-linguistic studies assessing languages skills of AI learners across different learning mechanisms and language experiences may indeed help us gain insight into which behaviors are merely side effects, and which behaviors are necessary for language learning. In the opposite case, a single computational model solving task T without exhibiting behavior B would be enough to conclude that B is not necessary for T – as in the case of Schatz et al. (2021)'s model, which shows perceptual attunement to the exposure language without phonemic representations. To take another example not yet attested, we can imagine a model that could learn some level of semantics, while being unable to detect word boundaries. This would constitute a proof of principle of the computational tractability of semantic learning without word boundaries.

4. Conclusion

Long-form recordings offer an ecological view of language use in everyday life. Aside from capturing child language experiences in an ecologically valid way, they offer new and exciting research opportunities in reverse engineering infant language development. Building upon the work of Dupoux (2018), we have defined two key aspects of the reverse engineering approach: 1) the language acquisition simulation, or how to use controlled and realistic data to create simulated language experiences and 2) the behavioral benchmarking, or how to assess language skills of the artificial language learner with psycholinguistic tests. This two-sided approach has the potential to increase our understanding of how language is acquired and how it develops through exposure, both in humans and in machines. The more we understand language acquisition in humans, the more human-like artificial language learners we can create. Similarly, the closer artificial language learners are to humans, the more we understand how language outcomes are shaped by exposure.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We are grateful to DARCLE, LAAC, and CoML members for helpful discussion. All errors remain our own. AC gratefully acknowledges financial and institutional support from Agence Nationale de la Recherche (ANR-17-CE28-0007 LangAge, ANR-16-DATA-0004 ACLEW, ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017); and the J. S. McDonnell Foundation (Understanding Human Cognition Scholar Award). This work was also partly funded by l'Agence de l'Innovation de Défense. Early study of unsupervised phonetic learning from child-centered long-form recordings was performed using HPC resources from GENCI-IDRIS (Grant 2020-AD011011829). Their computational support largely contributed in instantiating our methodological approach.

LITERATURE CITED

Abu-Zhaya, R., Seidl, A., Tincoff, R., and Cristia, A. (2017). Building a multimodal lexicon: Lessons from infants' learning of body part words. In *Proc. GLU 2017 International Workshop on Grounding Language Understanding*, pages 18–21.

Alishahi, A., Chrupała, G., Cristia, A., Dupoux, E., Higy, B., Lavechin, M., Räsänen, O., and Yu, C. (2021). Zr-2021vg: Zero-resource speech challenge, visually-grounded language modelling track, 2021 edition.

Ambridge, B. and Lieven, E. (2015). A constructivist account of child language acquisition. *The handbook of language emergence*, 87:478.

Anderson, J. R. (1975). Computer simulation of a language acquisition system: A first report. *Carnegie Mellon University*.

Athari, P., Dey, R., and Rvachew, S. (2021). Vocal imitation between mothers and infants. *Infant Behavior and Development*, 63:101531.

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., and Tor, S. (2019). Day by day, hour by hour: Naturalistic language input to infants. *Developmental Science*, 22(1).

Bergelson, E. and Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258.

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., and Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development*, 89(6):1996–2009.

Bosseler, A. N., Clarke, M., Tavabi, K., Larson, E. D., Hippe, D. S., Taulu, S., and Kuhl, P. K. (2021). Using magnetoencephalography to examine word recognition, lateralization, and future language skills in 14-month-old infants. *Developmental Cognitive Neuroscience*, 47:100901.

Braine, M. D. and Bowerman, M. (1976). Children's first word combinations. *Monographs of the Society for Research in Child Development*, pages 1–104.

Brent, M. R. (1996). Advances in the computational study of language acquisition. *Cognition*, 61(1-2):1–38.

Brookman, R., Kalashnikova, M., Conti, J., Xu Rattanasone, N., Grant, K.-A., Demuth, K., and Burnham, D. (2020). Depression and anxiety in the postnatal period: An examination of infants' home language environment, vocalizations, and expressive language abilities. *Child Development*, 91(6):e1211–e1230.

Carbajal, M. J., Peperkamp, S., and Tsuji, S. (2020). A meta-analysis of infants' word-form recognition. *Infancy*.

Casillas, M., Brown, P., and Levinson, S. C. (2020a). Early language experience in a Papuan community. *Journal of Child Language*, pages 1–23.

Casillas, M., Brown, P., and Levinson, S. C. (2020b). Early language experience in a Tseltal Mayan village. *Child Development*, 91(5):1819–1835.

Casillas, M. and Cristia, A. (2019). A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. *Collabra: Psychology*, 5(1). 24.

Chrupała, G., Gelderloos, L., and Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 613–622.

Cychosz, M. and Cristia, A. (forthcoming). Using big data from long-form recordings to study development and optimize societal impact. In Lockman, J. J. and Gilmore, R., editors, *Advances in Child Development and Behavior*, volume 62. Elsevier.

Cychosz, M., Cristia, A., Bergelson, E., Casillas, M., Baudet, G., Warlaumont, A. S., Scaff, C., Yankowitz, L., and Seidl, A. (2021). Vocal development in a large-scale crosslinguistic corpus. *Developmental Science*.

Cychosz, M., Romeo, R., Soderstrom, M., Scaff, C., Ganek, H., Cristia, A., Casillas, M., De Barbaro, K., Bang, J. Y., and Weisleder, A. (2020). Longform recordings of everyday life: Ethics for best practices. *Behavior research methods*, pages 1–19.

de Boysson-Bardies, B. and Vihman, M. M. (1991). Adaptation to language: Evidence from babbling and first words in four languages. *Language*, 67(2):297–319.

de Seyssel, M. and Dupoux, E. (2020). Does bilingual input hurt? a simulation of language discrimination and clustering using i-vectors. In *CogSci - 42nd Annual Virtual Meeting of the Cognitive Science Society*.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language learner. *Cognition*, 173:43–59.

Fernald, A., Zangl, R., Portillo, A. L., and Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language. *Developmental psycholinguistics: On-line methods in children's language processing*, 44:97.

Ferry, A., Hespos, S., and Waxman, S. (2010). Categorization in 3- and 4-month-old infants: An advantage of words over tones. *Child Development*, 81:472 – 479.

Ganek, H. and Eriks-Brophy, A. (2018). Language ENvironment analysis (LENA) system investigation of day long recordings in children: A literature review. *Journal of Communication Disorders*, 72:77–85.

Gasparini, L., Langus, A., Tsuji, S., and Boll-Avetisyan, N. (2021). Quantifying the role of rhythm in infants' language discrimination abilities: A meta-analysis. *Cognition*, 213:104757. Special Issue in Honour of Jacques Mehler, Cognition's founding editor.

Gross, D. R. (1984). Time allocation: A tool for the study of cultural behavior. *Annual Review of Anthropology*, 13(1):519–558.

Harwath, D., Hsu, W.-N., and Glass, J. (2020). Learning hierarchical discrete linguistic units from visually-grounded speech. In *International Conference on Learning Representations*.

Hochmann, J.-R., Endress, A., and Mehler, J. (2010). Word frequency as a cue to identify function words in infancy. *Cognition*, 115:444–57.

Hoff, E., Core, C., and Bridges, K. (2008). Non-word repetition assesses phonological memory and is related to vocabulary development in 20-to 24-month-olds. *Journal of Child Language*, 35(4):903.

Jaeger, J. J. (1980). Testing the psychological reality of phonemes. *Language and Speech*, 23(3):233–253.

Jusczyk, P. W., Luce, P. A., and Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33(5):630–645.

Lee, G. Y. and Kisilevsky, B. S. (2014). Fetuses respond to father's voice but prefer mother's voice after birth. *Developmental Psychobiology*, 56(1):1–11.

Liaqat, D., Wu, R., Gershon, A., Alshaer, H., Rudzicz, F., and de Lara, E. (2018). Challenges with real-world smartwatch based audio monitoring. In *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*, WearSys '18, page 54–59. Association for Computing Machinery.

Long, H. L., Bowman, D. D., Yoo, H., Burkhardt-Reed, M. M., Bene, E. R., and Oller, D. K. (2020). Social and endogenous infant vocalizations. *PloS one*, 15(8):e0224956.

MacWhinney, B. (2000). *The CHILDES project: The database*, volume 2. Psychology Press.

MacWhinney, B. (2005). A unified model of language acquisition. *Handbook of bilingualism: Psycholinguistic approaches*, 4967.

May, L. and Werker, J. (2014). Can a click be a word?: Infants' learning of non-native words. *Infancy*, 19.

Nazzi, T., Bertoncini, J., and Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human perception and performance*, 24(3):756.

Nguyen, T. A., de Seyssel, M., Rozé, P., Rivière, M., Kharitonov, E., Baevski, A., Dunbar, E., and Dupoux, E. (2020). The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. *NeuRIPS*.

Nielsen, M., Haun, D., Kärtner, J., and Legare, C. H. (2017). The persistent sampling bias in

developmental psychology: A call to action. *Journal of Experimental Child Psychology*, 162:31–38.

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., Yapanel, U., and Warren, S. F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences*, 107(30):13354–13359.

Orena, A. J., Byers-Heinlein, K., and Polka, L. (2020). What do bilingual infants actually hear? evaluating measures of language input to bilingual-learning 10-month-olds. *Developmental Science*, 23(2):e12901.

Pagliarini, S., Leblois, A., and Hinaut, X. (2020). Vocal imitation in sensorimotor learning models: a comparative review. *IEEE Transactions on Cognitive and Developmental Systems*.

Philippsen, A. (2021). Goal-directed exploration for learning vowels and syllables: a computational model of speech acquisition. *KI-Künstliche Intelligenz*, pages 1–18.

Rasilo, H. and Räsänen, O. (2017). An online model for vowel imitation learning. *Speech Communication*, 86:1–23.

Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., and Waldorp, L. J. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, 16(4):725–743.

Roopnarine, J. L., Fouts, H. N., Lamb, M. E., and Lewis-Elligan, T. Y. (2005). Mothers' and fathers' behaviors toward their 3-to 4-month-old infants in lower, middle, and upper socioeconomic African American families. *Developmental psychology*, 41(5):723.

Schatz, T., Feldman, N., Goldwater, S., Cao, X. N., and Dupoux, E. (2021). Early phonetic learning without phonetic categories–insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences of the United States of America*.

Schuller, B., Batliner, A., Bergler, C., Pokorny, F. B., Krajewski, J., Cychosz, M., Vollmann, R., Roelen, S.-D., Schnieder, S., Bergelson, E., et al. (2019). The interspeech 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity. In *Interspeech*.

Seidl, A., Cristia, A., Soderstrom, M., Ko, E.-S., Abel, E. A., Kellerman, A., and Schwichtenberg, A. (2018). Infant–mother acoustic–prosodic alignment and developmental risk. *Journal of Speech, Language, and Hearing Research*, 61(6):1369–1380.

Shi, R., Werker, J. F., and Cutler, A. (2006). Recognition and representation of function words in english-learning infants. *Infancy*, 10(2):187–198.

Simon, D. A., Gordon, A. S., Steiger, L., and Gilmore, R. O. (2015). Databrary: Enabling sharing and reuse of research video. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 279–280.

Slobin, D. I. (2014). Before the beginning: The development of tools of the trade. *Journal of Child Language*, 41(S1):1.

Sun, J., Harris, K., and Vazire, S. (2019). Is well-being associated with the quantity and quality of social interactions? *Journal of Personality and Social Psychology*.

Tamis-LeMonda, C. S., Kuchirko, Y., and Suh, D. D. (2018). Taking center stage: infants' active role in language learning. In *Active learning from infancy to childhood*, pages 39–53. Springer.

Turner, B. O., Paul, E. J., Miller, M. B., and Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, 1(1):62.

Twaddell, W. F. (1935). On defining the phoneme. *Language*, 11(1):5–62.

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., and MacWhinney, B. (2016). Homebank: An online repository of daylong child-centered audio recordings. In *Seminars in Speech and Language*, volume 37, page 128. NIH Public Access.

Vouloumanos, A. and Waxman, S. R. (2014). Listen up! speech is for thinking during infancy. *Trends in cognitive sciences*, 18(12):642–646.

Warlaumont, A., Westermann, G., and Oller, D. K. (2011). Self-production facilitates and adult

input interferes in a neural network model of infant vowel imitation. *Society for the Study of Artificial Intelligence and the Simulation of Behaviour*.

Warlaumont, A. S. and Finnegan, M. K. (2016). Learning to produce syllabic speech sounds via reward-modulated neural plasticity. *PloS one*, 11(1):e0145096.

Weisleder, A. and Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological science*, 24(11):2143–2152.

Werker, J. F. and Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7(1):49–63.

Wu, R., Liaqat, D., de Lara, E., Son, T., Rudzicz, F., Alshaer, H., Abed-Esfahani, P., and Gershon, A. S. (2018). Feasibility of using a smartwatch to intensively monitor patients with chronic obstructive pulmonary disease: Prospective cohort study. *JMIR mHealth and uHealth*, 6(6):e10046.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.

Yeung, H. H. and Werker, J. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition*, 113:234–243.

Yu, C. (2014). An embodiment perspective. *The Routledge Handbook of Embodied Cognition*, page 139.