SCALa: A blueprint for computational models of language acquisition in social context

Sho Tsuji, Alejandrina Cristia, Emmanuel Dupoux

[Preprint]

The final version of this article has been accepted as:
Tsuji, S., Cristia, A., & Dupoux, E. (2021). SCALa: A blueprint for computational models of language acquisition in social context. *Cognition, 213,* 104779. doi: 10.1016/j.cognition.2021.104779

SCALa: A blueprint for computational models of language acquisition in social context

Sho Tsuji[1], Alejandrina Cristia[2], & Emmanuel Dupoux[2,3,4]

[1] International Research Center for Neurointelligence, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan

[2] Laboratoire de Sciences Cognitives et Psycholinguistique, Departement d'Etudes Cognitives, ENS, EHESS, CNRS, PSL University, 29 Rue d'Ulm, 75005 Paris, France

[3]Cognitive Machine Learning Team, INRIA, 2 rue Simone Iff, Paris 75012, France

[4]Facebook Artificial Intelligence Research, Paris, France

Author Note

Declarations of interest: none

Correspondence concerning this article should be addressed to Sho Tsuji, International Research Center for Neurointelligence, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033 Japan, Contact: tsujish@gmail.com

Abstract

Theories and data on language acquisition suggest a range of cues are used, ranging from information on structure found in the linguistic signal itself, to information gleaned from the environmental context or through social interaction. We propose a blueprint for computational models of the early language learner (SCALa, for Socio-Computational Architecture of Language Acquisition) that makes explicit the connection between the kinds of information available to the social learner and the computational mechanisms required to extract language-relevant information and learn from it. SCALa integrates a range of views on language acquisition, further allowing us to make precise recommendations for future large-scale empirical research.

Keywords

statistical learning; word-to-world; language socialization; computational modeling

SCALa: A blueprint for computational models of language acquisition in social context

1. Introduction

Infants' ability to learn their native language with astonishing speed and efficiency has fascinated and puzzled researchers for many years. Human language is a richly structured and complex communication system with multiple levels, each with their own sets of rules, ranging from speech sounds over words to grammar. This system's complexity and expressive power is unrivaled in the animal world (Hauser, Chomsky, & Fitch, 2002), and its acquisition is all the more surprising when considering diversity in the learning targets (linguistic diversity) and learning environments (cultural diversity). Infants might face languages with consonant inventories ranging from 6 to more than 100 consonants (Maddieson, 2013), or with gender markings ranging from 0 to 5 or more genders (Corbett, 2013), to give just two examples. As for learning environments, they differ qualitatively but also quantitatively: Children growing up in industrial societies might hear 2-10 times more input compared to children growing up in preindustrial societies (Vogt, Mastin, & Schots, 2015), and children growing up in households with higher socioeconomic status hear more varied and complex utterances compared to children growing up in households with lower socioeconomic status (Hoff, 2003).

The multiple sources of information an infant can potentially consult to acquire language are considered key to infants' amazing ability to learn their native language, at the same time as they are considered instrumental to explain the diversity in learning outcomes. For instance, listening to the speech stream an infant can learn about the order of phonemes, syllables, and words; seeing a caregiver point at something while uttering a word, an infant can learn the name of an object; being provided with corrective feedback can shape infants'

language output to the community norms. Often, the role of these different sources of information has been studied in isolation, preventing an evaluation of their relative role to explain commonalities and specificities of developmental trajectories across age range, languages and cultures. We propose to bridge this gap in an integrative computational approach.

Traditionally, two kinds of approaches have been leveraged to understand the underpinnings of early language learning: experimental laboratory studies and corpus studies. Since the seminal work of Eimas et al. (1971), Mehler (1981), and many others, the field of experimental infant psychology has produced a wealth of measures of developmental landmarks (for instance, infants know a few words by 6 months, Tincoff & Jusczyk, 2012), of potential mechanisms underlying the learning of language structure (for instance, infants pick up on aspects of their language's prosodic and morphosyntactic structure by 7 months, Gervain & Werker, 2013), or of the social-communicative cues that are involved in learning at different ages (for instance, infants' increasing reliance on gaze cues over the second year of life, Hollich, Hirsh-Pasek, & Golinkoff, 2000). As regards corpora studies (e.g., see MacWhinney, 2000, chapter 2, for a brief history), the recording, annotation and distribution of child-caretaker interactions in open archives has provided ways to investigate how infants learn language from their interactions with others, involving realistic input data and opening the door to a quantitative perspective. Further, portable lightweight audio and video devices (e.g., Language ENvironment Analysis LENA, Xu et al., 2009) have made possible the collection of large-scale datasets capturing infants' multimodal interactions in their natural environments (Bergelson, Amatuni, Dailey, Koorathota, & Tor, 2019) from diverse human cultures (Casillas, Brown, & Levinson, 2019). These observational approaches promise novel insights into the role of linguistic input and the social-communicative environment on early language acquisition.

Despite the remarkable progress achieved by these two experimental and observational approaches, we struggle in answering even simple questions about the relative role of different kinds of information for language learning: does the acquisition of syntax depend primarily on observing the distribution of linguistic patterns in the input, on the correlation of input and the world, or on corrective feedback? Is it the same for phonetics and semantics? What happens if one of these sources is reduced or degraded? Experiments are limited in their scope, on the one hand because their design often comes at the cost of felicitousness to any real life learning situation an infant might face, both by controlling for confounding factors and by any lab study being restricted in terms of participant numbers, populations, and age ranges. On the other hand, they are limited in their intervention on language learning by practical and ethical considerations, and natural experiments (e.g., language learning in the visually impaired, see Landau, Gleitman, & Landau, 2009, or in the hearing impaired, see Goldin-Meadow & Mylander, 1998), despite being illuminating, are scarce and do not cover the full scale of questions regarding the role of input types on language learning. Corpus studies may capture more naturalistic situations, but the data can be hard to analyze and establishing causality is challenging.

Here, we propose to complement experimental and observational studies with a third approach, aiming for algorithmic and computational specificity to facilitate integration of diverse perspectives and generation of predictions (see e.g., van Rooij & Baggio, 2020). Specifically, we follow a *reverse engineering approach* (Dupoux, 2018), which states that such causal and quantitative theories have to fully specify three components: the *learner* (taking in input, learning latent representations and producing outputs), the *environment* (providing input that is independent of the learner, as well as input that is reactive to the learner), and the *outcome measure* (linking the learned representations to traditional outcome measures in the child

language literature). As an illustration of this approach in the domain of phonetic learning, Schatz et al. (2020) instantiate the *learner* as a state-of-the-art representation learning algorithm which is fed spectral representations of the speech input and learns its probabilistic distribution, the *environment* as several hours of audio recordings of read or spontaneous speech, and the *outcome measure* as a phoneme discrimination task run on the posterior probabilities of input stimuli. In such an example, the environment is a static recording, and the learner passively accumulates statistics over the recording. Could such an approach be extended to address questions about how language is learned in more ecological situations, where the learner not only listens, but also sees, feels and moves in a physical environment, as well as interacts with and receives feedback from a caretaker?

The contribution of this paper is precisely to explore what needs to be done to apply the reverse engineering approach to this more realistic setting. We will not offer a fully fledged implemented model like Schatz et al. (2020). Rather, we will offer a *blueprint*, which can be viewed as the set of specifications of what an implemented model should look like. In addition, we will focus the paper on the *learner*, only referring to the *environment* and *outcome measure* in so far as it is needed to explain our description of the learner. We take the viewpoint of an *ideal learner*, i.e., a learner whose ultimate goal is to learn language, and who is optimal at it. While real infants may of course depart from optimal learners for a variety of reasons (e.g, their still developing attention and memory capacities), having an optimal model is at least a good starting point in developing causal theories of the infant learner, and can answer questions as to the functional role of particular types of inputs for instance.

We defer the specification of a blueprint of the environment to further work, since moving beyond static environments while staying realistic is complicated even in the case of simple physical interactions, leading studies on simulated caretakers to focus on toy problems (see

Hermann et al., 2017; Lazaridou and Baroni, 2020). As for outcome measures, these have started to be addressed elsewhere (see Lavechin, de Seyssel, Gautheron, Dupoux, & Cristia, 2021, for some propositions).

In Section 2, we start by classifying information for language development into three broad types, each of which is indispensable for understanding the learning process, but each of which on its own is not sufficient to characterize the information available in the learning environment as a whole. In Section 3, we argue that these information types require different types of learning algorithms to be fully exploited by an ideal learner. In addition, we introduce the notion of filters, whose role is to route the sensory information available to the learner to the correct learning algorithm. In Section 4, we motivate the need to annotate datasets of infant-caretaker interactions in terms of this routing problem, in order to quantify (and therefore to be able to model), the prevalence of each of the relevant types of information in the infant's environment. In Section 5, we finish by presenting a roadmap of the work that needs to be done to fully integrate this approach into the more traditional experimental and corpus approaches.

2. Three Types of Information Relevant for Language Acquisition

In the following, we present a classification of information relevant to language acquisition into three broad types. The purpose of this overview is not to give an exhaustive account of types of information that could ever play a role in early language, nor to argue for this precise classification as being the most important one. Our main goal is to establish a distinction of different types of information for the language learner which will map onto the learning problem from a computational point of view. Summarizing the extensive empirical evidence showing that language acquisition involves extracting and learning each type of information is beyond the

scope of this paper, which is mainly theoretical. In its stead, we justify each type by making reference to different theories that focus on it.

2.1 Language as Structure

Language has a multi-level and complex internal structure, and no matter which language background an infant is born into, it is indispensable to learn about her native language's structural aspects, ranging from phonetic[1] categories to syntax. Many scholars agree on the importance of the distribution of linguistic patterns as a source of information, even when they diverge widely in terms of other theory-relevant characteristics, such as the extent to which they believe infants' ability to acquire those patterns or structures is innate or learned.

At one end of the spectrum, generativists assume strong innate predispositions that guide the acquisition of language structures. This assumption is motivated by the observation that any type and amount of language input that infants receive is compatible with multiple grammars, and thus their input alone would not suffice to explain why a given infant prefers one generalization over another. For instance, it has been proposed that infants have an innate capacity to acquire language (the Language Acquisition Device, LAD; Chomsky, 1964), and that structural aspects of languages, such as their phonology or syntax, can be described by a common set of principles and parameters, some of which are universal, and others are turned on or off based on experience with a particular language. Later approaches (e.g., Optimality Theory; Prince & Smolensky, 2004) introduced constraints instead of parameters, allowing for a more flexible and general framework to accommodate the structures of many languages, as well as probabilistic instead of deterministic grammar variants (see Pater, 2019, for an overview). Common to these views is a strong focus on innate knowledge of structural aspects of language.

[1] Note that, while we use words like "phonetic", "talked about", "speech", or “verbal”, our framework is also intended to cover sign (including tactile) languages.

On the other end of the spectrum, the learnability of language structure based on the input is stressed (e.g., PRIMIR, Werker & Curtin, 2005). Lab studies on input-based learning highlight infants' capability to deduce language structure such as phonetic categories, word forms, word-object mappings from tracking statistical regularities in the input. These studies crucially show that infants are able to learn from patterns present in their language input outside of any referential or communicative context. Whether such capacity could extend to naturalistic input is not yet known, as meta-analyses over different stimulus sets and experiments suggest that some of these studies might not be robust across different conditions (Black & Bergmann, 2017; Cristia, 2018).

While the long-standing debate about the poverty or richness of the stimulus remains unresolved, both sides agree on the importance of linguistic patterns, both as a source of input and a target for learning. We believe this strongly suggests that any comprehensive framework of language acquisition will thus need to accommodate a way for a language learner to acquire this source of information, as well as to test different assumptions about the role of innate predispositions and environmental input in this process.

2.2 Language as a Description of the World

In order to become competent native language users, infants need to learn not only about the structure of language, but also about how the content of language connects to the outside world, involving both visible and invisible states as well as concrete and abstract concepts. This link is rendered evident when language is used to describe aspects of the outside world. Such instances are inherent in the social interactions taking place in an infant's environment (see Section 2.3), and could actually help or even guide the learning of structure (see Section 2.1). The accounts described in this Section 2.2 have in common that they describe ways in which

language links to the world. As above, we have selected representative theories to show the divergence in views outside of this common focus. For instance, these theories vary in whether they assume domain-general or language-specific mechanisms, and to what extent learning of language structure and the prioritization of social cues are innate or develop through exposure.

To begin with, one group of proposals calls attention to how language structure links with the world as a general property of language, via semantic bootstrapping (Pinker, 1982; 1984) and/or syntactic bootstrapping (Gleitman, 1990). Both types of bootstrapping highlight the regular correspondences between classes of syntactic objects (nouns, verbs, etc.), and types of entities in the world (objects, events, etc.; see Brown, 1957). Both also argue that such correspondences are possibly innate and drive learning, but they diverge on which direction the correspondence is most potent.

A second example is the Natural Pedagogy theory (Gergely & Csibra, 2009), which assumes a specifically human learning system based on ostensive-referential communication, positing that under certain social triggers infants might be innately biased to link language and the world at the level of abstract and generic knowledge, rather than at the level of specific instances. This nativist view thus draws attention to how social cues can mediate learning in the context of language-world associations.

Diverging from the two groups of theories just mentioned on several counts, associative proposals argue that infants can link language to the world by domain-general associative mechanisms, such as co-occurrence statistics between auditory and visual objects (Smith & Yu, 2008). Social interaction can enhance this link by guiding attention and leading to ideal learning opportunities, for instance by increasing the salience of co-occurrence, such as happens when waving an object in an infant's central visual focus while labeling it. In other words, this particular

group of theories stresses that social interaction can improve the quality of the language-world link.

A final example we want to mention here is the Emergentist Coalition Model (Hollich et al., 2000). According to this model, when infants learn to link form and meaning, they are sensitive to attentional cues (such as perceptual salience), social cues (such as eye gaze), and linguistic cues (such as prosody). However, the precise weighting of these cues changes over time, broadly speaking from a stronger reliance on attentional cues early on to a stronger reliance on social and linguistic cues later on, and this change in weighting comes about by infants' discovery that some cues are more reliable indicators than others.Thus, this account is potentially compatible with both nativist and empiricist perspectives, and their crucial contribution is that, in the context of making language-world correspondences, different mechanisms' importance is proposed to change over time via learning.

Despite their diversity of focus and scope, these views all agree on the core assumption that learning language involves learning a correspondence between language structure and the world, thus lending credence to the view that this second type of information is crucial when aiming to explain early language acquisition.

2.3 Language as a Social Construct

Section 2.1 and 2.2 readily map onto external information that the infant can pick up on, such as language structures and links between language and the world, which can be conceived of as completely external to the infant. In the current section, we want to operate a mind shift, and bring to the fore the fact that the infant is not a passive observer of a reality that plays like a movie, independent of the infant's actions. Thus, the third class of information we discuss is not "external" to the infant, but instead is information found in the interaction between the infant and

the environment. To do so, we build on the social constructivist and language socialization traditions, positing that an individual's language learning takes place as an emerging property from his or her interactions in a group (e.g., Vygotsky, 1978; Snow, 1977; Lieven, 1978; Ochs & Schieffelin, 2011). There was some mention of social interaction in some of the views discussed in Section 2.2; however, to clarify the distinction with theories discussed in this Section 2.3, those theories could be based on an asymmetric view of the infant and others around her, whereby the interactor organizes the infant's experience and the infant could be consuming it passively.

Social constructivism and language socialization approaches focus on the interaction itself as a necessary driver of language learning, so that children are not only patients but also actors in the construction of their knowledge. To begin with, social constructivism proposes that the caregiver *scaffolds* the input to adapt it to the child's needs. Thus, the child can affect the caregiver, and ultimately the input she receives. A remarkable example of this comes from Ratner and Bruner (1978), where structure and content of ritualized play routines (like peekaboo) observed in two infants between the ages of 5 and 14 months of age documented, on the one hand, their mothers' apt highlighting of game structure and junctures (providing more detail when the infant is younger and at the beginning of the game, less as the game progressed and the infant aged), and on the other, the infants' growing mastery of key aspects of language (and the game), including "interchangeability of roles", "appropriacy conditions", and "phonetically consistent form[s]" (p. 397).

In language socialization approaches, the infant learns language at the same time as she learns her place in social interactions (Ochs & Schieffelin, 2011). For instance, Ochs (1982) explains that "how a Samoan child speaks - both form and content - is strongly influenced by social norms for using language in Samoan households and by certain attitudes and beliefs

concerning individuality, knowledge, and human competence" (p. 78). Mothers *enact* relative status: In a situation where there is a mother, an older sibling, and the young child (whose language acquisition we focus on here), requests by the young child will not directly be responded to by the mother, and instead the mother will indicate to the older sibling to reply to the child. The child can thus draw linguistic information (from the overheard command), and social status information: Since the mother is highest ranking in this situation, she signals her status by minimizing activity, awareness, and involvement.

Once again, we have quite divergent perspectives represented across theories discussed in this section, divergences that we believe are important and meaningful. Nonetheless, despite the differences in the way interactors are conceived in social constructivism and language socialization theories, they agree among them (and differ from the other theories) in terms of their central focus on children in a truly social situation. The interplay between caregivers and children plays a key shaping role by which input is embedded in a mutual feedback loop.

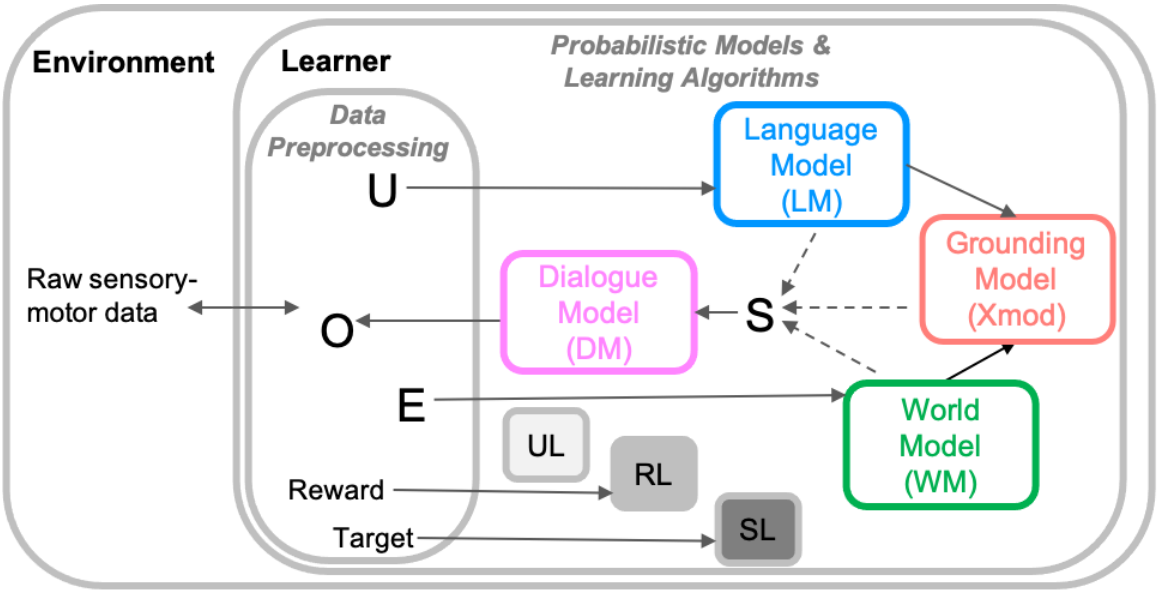
2.4 Summary: How to Integrate these Three Types of Information?

Our conceptual distinction above illustrates the multifaceted nature of the language learning process: Infants simultaneously learn about the structure of language, its connection to the outside world, and how to use language in social interactions. At the same time, this highlights how wide an array of input the infant needs to process during the acquisition process, ranging from verbal forms over sensory input to multimodal social interactions. We propose that all three aspects of language learning are important, and that researchers would benefit from a way to examine the relative contributions of different types of information to the learning process in a common framework.

3. A Socio-Computational Architecture for the Language Learner

In this section, we provide a description of a Socio-Computational Architecture of Language Acquisition (SCALa) which provides a blueprint of a model of the socially embedded learner. This fits well with the recent push towards formal and/or computational models and away from weak verbal theories (e.g., Guest & Martin, 2020; for a range of views in this topic, see Fried, 2020, and replies in the same issue). However, as outlined above, we do not present a fully implemented computational model, but rather a blueprint that can be used to generate many such models. To this end, we incorporate references to actual algorithms that are used in speech and natural language processing, which we group into generic classes rather than focusing on implementation details. This list is not exhaustive and given fast changes in the machine learning field, additional algorithms may emerge in the future, which could also be informative. Additionally, since we hope that SCALa will become a framework within which individual theories and models can be instantiated, we do not believe the full list is essential, given that some theories may focus on spelling out specific subareas. That said, we hope that by embedding their narrower theories and models within SCALa, researchers will be cognizant of the aspects of the learning process that they are remaining silent on.

As outlined above, we assume an *environment* in which the infant (hereafter: *learner*) is embedded. This environment includes the physical world that the infant can sense and interact with as well as one or multiple interaction partners (hereafter: *interactor*) that provide social-communicative input to the infant.



Probabilistic Models

- **Language Model**. Estimates P(U), the probability distribution of message U.
- **World Model**. Estimates P(E), the probability of event E.
- **Grounding Model**. Estimates probabilities of association between verbal form and event (P(U,E)). Assumes that the intended meaning is accessible here-and-now.
- **Dialogue Model**. Computes the probability of communicative output O given message and current state of world S (P(O|S)). S is computed from a representation of past events and utterances.

Learning Algorithms

- **Unsupervised Learning (UL)**. Tries to optimize the likelihood of observing a given input (U or E). Language Models (LM) and World Models (WM) can be learned in this fashion.
- **Reinforcement Learning (RL)**. Tries to optimize the expected reward (Reward), Dialogue Models (DM) can be learned this way.
- **Supervised Learning (SL)**. Tries to minimize the discrepancy between an expected response (Target) provided by the environment and actual response O. DMs can be learned in this way.

Data Preprocessing

- **Filtering:** what sensory data counts as a language input (U), a world input (E), a Reward, a Target ?
- **Segmenting:** what are the units of the language stream (U), what is an event (E) ?
- **Routing:** is there an intended/corrective target (Target), and if so, what output O is it supposed to correct? If there is a referential act, which parts of U map to which part of E for cross modal learning?

Figure 1. The learner's internal probabilistic models and learning algorithms. Input from the outside world first undergoes data preprocessing. The preprocessed data can take the form of utterances (U), objects and events (E), intended/corrective Targets, and positive or negative Rewards, and the learner can produce outputs (O).

As illustrated in Figure 1, in SCALa, the learner is imbued with two sets of components : (1) *probabilistic models* and associated *learning algorithms, (2) data preprocessing* components. As to the first set of components, probabilistic models represent their inputs as parametrized probability distributions and update these parameters through the application of learning algorithms. Each algorithm has characteristic *inductive biases*, i.e., a particular way to generalize to novel unseen input given a finite set of training examples. Such inductive biases are also sometimes called '*priors*' in statistics or in cognitive psychology, or innate knowledge in generative linguistics, and can be conceptualized as assumptions that the algorithm makes about the nature of the entity to be learned (such as the "Whole object assumption"; Woodward, Markman, & Fitzsimmons, 1994).

The second set of components enables the *intake* of information (generically grouped under the machine learning term '*Data Preprocessing*'), which we take to be a subset of the available input, since the infant may ignore some aspects and focus on others. Such components have been often presented in psychology as *'filters'* that determine *which* inputs are processed by which learning algorithms. Such filters are rarely described in current theories of language acquisition, and similarly, in standard machine learning applications, data preprocessing is often not considered as part of the learning problem. Filters apply at different levels of processing, and for some filters, we can assume their explanation lies outside of language because they stem from phylogenetically old abilities. For example, speech is filtered

into distinct frequency bands thanks to a brain network found in humans and other mammals (Saenz & Langers, 2014). We suspect many filters necessary for language acquisition do not benefit from a robust cross-species basis, and theories assuming them needs to explain how they came about.[2] In the context of a social learner embedded in the real world, the data preprocessing problem is less specified and more complex. In addition to the sensory, attentional, and informational filters signaling language-relevant input to the learner, the learner requires advanced operations like referential selection and communicative event classification to process input adequately, (see Section 3.3).

In the following sections, we first give a more detailed description of probabilistic models (Section 3.1) and associated learning algorithms relevant to language learning (Section 3.2). In Section 3.3, we then describe how an ideal learner should filter the input as a function of the socio-communicative context; in other words, what the output of filters for referential selection and communicative event classification would ideally look like to be processed as input by the learning algorithms and probabilistic models. We finally describe how an actual algorithm might approach this task (Section 3.4).

3.1 The Learner's Probabilistic Models

[2] While we highlight the importance of optimal filters for an ideal learner, it is possible that a less than ideal learner using a rough proxy for a filter might still work in practice. One example is a computational model of cross modal learning (e.g., Harwath & Glass, 2017). This model is able to directly map whole sentences to whole images. The model, applied without any filter, localises which part of the image maps best onto which part of the sound sequence. This is thus a case where a filter can be learned directly from the unfiltered data. A later model (Hsu, Harwath, Song and Glass, 2020) is able to produce a spoken caption for an image by incorporating a similar approximate filter (words and objects are just rectangular regions of space or time) which is learned at the same time as the captioning system. Thus, what seems difficult or conversely easy from a verbal description can turn out very different when software is implemented and tested. Our contribution here is simply to state that some notion of filters need to be addressed in the final software, not how it will be implemented.

Some of the recent successes of machine learning rest on powerful **generative probabilistic models**. Briefly, a probabilistic model specifies a probability distribution $P_\theta(X)$ over the set of all possible observations X, where θ is the set of parameters of the model.[3] It is called generative when it can generate new observations by sampling from the distribution. A probabilistic model can also be associated with a *learning algorithm* which updates the parameters θ given a finite sample of observations (called the training set: $X_1, \ldots X_T$). Learning can be construed as an optimization problem which consists in searching for the parameters θ^* that maximize an objective function (also sometimes called the 'Loss function' *L*, in which case it has to be minimized), which is a quantity that depends on the parameter θ and the training examples ($X_1, \ldots X_T$) ($\theta^*=\arg\min_\theta L_\theta(X_1, \ldots X_T)$).[4] Once learned, the model with its parameters can be used to estimate the probability of novel unseen samples (generalization) or generate new samples from the model. In order to make SCALa address all of the relevant aspects of the language learning problem as outlined in Section 2 (Language as Structure, Language as a Description of the World, Language as a Social Construct), we assume a learner with several such models: A *Language Model*, a *World Model,* a *Grounding Model* and a *Dialogue Model* (Figure 1). Let us review these models one by one.

The **Language Model** captures the probability distribution $P_{\theta_L}(U)$ of verbal forms, let's say, an utterance U, in the linguistic environment of the infant. This probability distribution represents the *tacit knowledge* that the learner has about his or her language at a given time t. Such a model can be built using lists, dictionaries, episodic memories, probabilistic grammars, neural networks, or other computational structures. The free variables in these structures

[3] A very simple example probability distribution is the Gaussian distribution, which specifies a density probability over real numbers; its parameters θ are the mean and standard deviation.

[4] A simple loss function is the negative log probability of the training examples $L=-\log p_\theta(X_1, \ldots X_T)$. Minimizing this loss amounts to finding the parameters that make the training examples maximally probable.

constitute the model parameters θ^L. In systems used in language technology, the utterance U is typically represented as a sequence of discrete symbols: for text it may be characters; for spoken languages, phones or phonetic features. Recent work suggests that it is possible to learn discrete units from raw audio, and use them as if they were phonemes or letters (see Nguyen et al. 2020; Lakhotia et al, 2021 for the feasibility of this approach).[5] The learning of the model parameters θ^L is typically done by a class of algorithms that is called in machine learning, somewhat confusingly, *Language Modeling* (LM)[6]. The most powerful of them have millions of parameters trained on large sections of the world web. They can give a good approximation of adult human performances on a variety of metrics (e.g., grammaticality judgments, see Warstadt et al., 2019) and can be turned into surprisingly creative language generators (see Brown et al., 2020). Cognitively, postulating a Language Model in the infant would account for their capacity of displaying preferences for patterns (over phonemes, words, sentences) that are frequent versus infrequent in their language, and their ability to produce such patterns.

The **World Model** captures the probability of occurrences of an event E in the world and represents the tacit knowledge of the learner about objects, agents and events. Like the

[5] Note that in general, the probabilistic models described here are formulated over *sensory representations* (audio patterns, visual patterns, etc). This may seem overly behavioristic; cognitive models would rather include reference to abstract entities like phonemes, morphemes, words, syntactic trees, objects, substances, forces, agents, concepts, beliefs, intentions, etc. However such abstract entities are ultimately grounded in sensory data; in probabilistic models such as graphical models (e.g., Pearl, 1998; Tenenbaum & Griffiths, 2001) they are explicitly represented as *latent* representations (latent because not part of the observations, and their values are calculated during inference); in neural networks, they emerge in the hidden layers and connectivity pattern of the system.

[6] Simple LM algorithms based on n-grams consist in counting the occurrence of each n-gram in the training set. More sophisticated LM algorithms are recurrent or convolutional neural networks trained using prediction objectives. For instance, if an utterance U is composed of a sequence of word tokens (or some other units like phonemes or speech frames) $u^1 ... u^N$, the system is trained to predict the token u^i based on the past ones $u^1 ... u^{i-1}$ (increasing i from 1 to N). This objective makes it easy to compute the probability of the entire sequence U as the product of the conditional probabilities of each successive unit: $P(U)=\mathbf{\Pi}_{i=1..N}\ P(u^i|u^1..u^{i-1})$. This algorithm thus tracks the temporal structure of language in order to predict the likelihood of a given verbal form occurring given the preceding verbal forms. Newer algorithms like BERT do not respect the linear order and attempt to reconstruct a certain proportion of masked or corrupted tokens (Devlin, Chang, Lee, & Toutanova, 2018).

Language Model, it can be constructed in various ways (as an episodic memory or in a parameterized probabilistic model). The associated algorithm, *World Modeling* (WM), refers to updating the parameters of the probability model of a particular percept ($P_{\theta w}(E)$)[7]. Current models can generate realistic short videos (see Clark, Donahue, & Simonyan, 2019) or be used to plan the action of robots. Applied to the modeling of an infant learner, such an algorithm would store in its parameters θ^{W} world knowledge computed over sensory primitives that the infant has experienced related to occurrences of world events, which can be external (visual, auditory, tactile, ...), internal to the learner (emotions, pain, ...), or even internal to another person, as long as there are external cues or perceptible referents available that can evoke a concept or percept in the infant. Such algorithms succeed in modeling infant behavior such as distinguishing possible from impossible physical events (Riochet, Sivic, Laptev, & Dupoux, 2020), and would thus account for infants' capability to display 'surprise' when presented with unlikely events (Kellman & Spelke, 1983) by assuming they assign probabilities to newly encountered events (see also Battaglia et al., 2013; 2016). Even though the World Model does not contain any language information, it is useful for a range of perception and action routines that learners are likely developing alongside their language, and, crucially for SCALa, for understanding speech or sign in context.

Two classes of algorithms are important to ground the learning of language into the larger context. A **Grounding Model** expresses relationships between whole utterances or subcomponents of utterances (words or phrases, denoted here with *U*) and events or subcomponents of world events (objects or actions, denoted here with *E*) as a joint probability

[7] Some of the most successful models in this area are generative adversarial models; a component of the model is tasked with generating an image or a short video based on a random number as input, while another component is tasked with discriminating the produced output against real inputs. The two parts are trained jointly. Other models generate videos sequentially, like in an LM, allowing to compute the probability of entire videos (Riochet et al., 2020).

(P(U,E)). Bayes rule can be used to express this knowledge in a directional fashion: P(U|E): distribution over the possible linguistic descriptions of event E; or P(E|U): distribution over the possible referential meanings of expression U. The associated learning mechanism is often termed *Cross-Modal Learning* (Xmod) or, in psychology, cross-situational learning (e.g., Smith & Yu, 2008), and consists in probabilities of the association between verbal forms and their referred meaning. Such algorithms have been used in machine learning to generate captions based on images or videos (P(U|E); e.g., Sun, Myers, Vondrick, Murphy, & Schmid, 2019), or vice versa to generate images based on verbal description (P(E|U); e.g., Reed et al., 2016), although the performance of these kinds of models when done at the utterance level still suffers from severe limitations (Sun et al., 2019). As applied to infants, the Grounding Model can be evidenced in paradigms like looking-while-listening, where infants show a preference at looking to the correct referent of an utterance when shown multiple options (Fernald, Zangl, Portillo, & Marchman, 2008). Cross-Modal Learning can be applied every time a verbal form and the associated meaning are accessible simultaneously in the caregivers' input. It has therefore been proposed as a mechanism accounting for how infants link language to the world (Smith & Yu, 2008).

The **Dialogue Model** addresses the same issue in a more general way: it tries to capture the dynamic of language in context without necessarily assuming that each form corresponds to a particular meaning observable here and now. As defined here, *Dialogue Modelin*g (DM) establishes the probability that the learner produces a given *output* O (P(O|S)). It is conditioned on a representation of the *state of the world* S, which is itself computed on the basis of past interactions with the world (utterances and events). Dialogue Modeling is more general than Cross-Modal Learning, because it includes the possibility that, depending on context, the learner might (or might not) name or designate an object. It is similar to Language

Modeling because it outputs a probability over utterances, but there are two differences: (1), the utterances are produced by the learner, not perceived, and (2) the probabilities are conditioned by past interactions, including events in the environment. This is only one of multiple ways of modeling dialogues, and also the area of machine learning where there is most progress to be made. Evidence of infants' learning about conversational structure is available both from perception (Casillas & Frank, 2017) and production (Hillbrink, Gattis, & Levinson, 2015).

3.2 Types of Learning Algorithms: The Role of Feedback and Supervision

How are the parameters of any model (Language Model, World Model, Grounding Model, Dialogue Model) learned? There are basically three classes of algorithms: Unsupervised Learning, Reinforcement Learning, and Supervised Learning.

Unsupervised Learning (UL, sometimes Self-Supervised Learning, see a review in Jing & Tian, 2020) can be performed by a passive observer. It consists in updating the parameters of the probability models based only on a set of sensory inputs. The objective function to maximize is the probability of the input. Such algorithms have the advantage that they do not require any overt action from the learner and can work on any sensory data available. Their downside is that they may require a lot of data to correctly model the probability distribution. This is because Unsupervised Learning only has access to the data that belongs to the distribution, and not to the data that does not belong to it (i.e., it is provided with no negative evidence). Unsupervised Learning is also quite sensitive to the presence of noise in the input data: unless equipped with a robust input filter, it has no way to separate signal from noise -- and this problem is still considered a hard problem for naturalistic inputs. World Models (Kim, Sano, De Freitas, Haber, & Yamins, 2020), Language Models (e.g., for representation learning, Baevski, Schneider, & Auli, 2019), and most recently Dialogue Models (at least in terms of learning sequences of

utterances, Lan et al., 2019) have all been demonstrated to be learnable from Unsupervised Learning alone. One important limitation of Unsupervised Learning is that while it teaches an organism what to expect, it does not tell it what to do.

Reinforcement Learning (RL) requires the learner to perform actions (not just to observe passively). It is a class of algorithms which assumes that the environment can return *rewards,* positive or negative scalar numbers, in response to the past overt actions of the learner (Sutton & Barto, 1998). The learning algorithm attempts to maximize not the probabilities of the sensory inputs, but the *expected reward*. This class of algorithm exploits the lessons hidden in even long sequences of interactions, for instance in a game like chess or go, by interacting with other players (or playing against itself), whereas the mere Unsupervised Learning of stored examples alone does not yield as good of a performance (Schrittwieser et al., 2019). As applied to Dialogue Modeling, such an algorithm would basically allow infants to test whether their output is in the required distribution or not (by encoding a smile or a frown as a positive or negative reward), thereby providing them with positive or negative evidence, which was absent in the Unsupervised Learning setup. Note that “positive” or “negative” should not be interpreted in absolute terms, but needs to be interpreted within conversational context and culture - for instance, in some cultural contexts negative feedback might be accompanied by a smile. In cases where World Modeling is extended to include motor outputs, Reinforcement Learning could be used to test hypotheses about the physical properties of objects, and learn faster than by just observation (e.g., dropping a light versus heavy item on one's foot).

Supervised Learning (SL) is an even stronger algorithm where the environment provides a desired output directly to the learner. Here, the objective function to optimize is the similarity between the predicted outputs and the desired output. A large section of the field of deep

learning is concerned with this type of algorithm (Goodfellow et al, 2016)[8]. There are two relevant cases for language acquisition. The first situation is when caregivers provide corrective feedback in response to an incorrect output by the learner. Here the intention is not to teach a particular regularity that when one says ‘goed’ the interactor should say ‘went’ (regularity captured in the Dialogue Model via Unsupervised Learning); here, the output of the caregiver is to be inserted as the desired output of the infant. A more general class of this latter situation is sometimes called *imitation learning*, where the caregiver performs something that the infant should perform (Piaget, 1951; Rosenblith 1959). This mode of learning is used in robotics to speed up the learning of behaviors that would otherwise require too many reinforcement learning episodes (see a tutorial here: https://sites.google.com/view/icml2018-imitation-learning/). The second situation is when caregivers directly provide pairs of words plus meaning, using pointing and other attention-grabbing devices. In such situations, caregivers essentially do the ‘filtering’ (data preprocessing) for the infant, and provide the right input for a Cross-Modal Learning algorithm to work well.[9] It is to be noted that the intention of the caregiver is not to teach that one should always point to objects while saying their names (what the infant should learn if she were only applying Unsupervised Learning), but rather that the pointing is a way to directly insert the pair {object,word} in the relevant learning algorithm. To the extent that it is used in sufficient amounts by humans across cultures, and can be detected as such by infants, such cases of supervised

[8] There is a weaker form of Supervised Learning (Weakly Supervised Learning), where the systems attempt to learn patterns of associations between two types of inputs. The model is provided with two inputs (e.g., a word and an image), and the desired output is 1 if the two inputs are related and 0 if not. This type of learning enables to learn many-to-many mappings as in the case of word-picture associations.

[9] Note that this is still not a compelling explanation of a filter (see introductory part of Section 3), because how does the caregiver know how to do the filtering? What kind of (innate or acquired) mechanism leads them to provide infants with the right information? We stress that any theory making such an assumption needs to explain this as part of the "language teaching device" in the caregiver.

learning could speed up the learning of language, dialogue and meaning in comparison to Unsupervised Learning.

3.3 Selecting and Classifying Social-Communicative Input from an Ideal Learner Perspective

In this section, we argue that a particular piece of input is differentially informative for the models and algorithms described in Sections 3.1 (Language Model, World Model, Grounding Model, Dialogue Model) and 3.2 (Unsupervised Learning, Reinforcement Learning, Supervised Learning) as a function of the referential and communicative context in which it occurs.

Before diving into the type of complex data preprocessing a social learner embedded in the real world needs to undergo, we want to point out that each learning algorithm expects a particular type and format of data as input. Fed with the wrong data, it will produce garbage. For instance, an Unsupervised Language Model will require input in the shape of a *sequence* of samples, and it will only learn language abstractions if fed with proper language (as opposed to natural sounds or a mix of language and sounds). Similarly, a World Model based on images will require retinotopically organized pixels and will learn language-relevant concepts only if fed with images of nameable objects and actions (as opposed to, say, patterns of smoke or photographs of fluid dynamics). This requires organizing the sensory data in a format compatible with the model, segmenting and filtering the sensory input according to types of data (utterances, turns, human-relevant macroscopic objects and actions, etc.).[10] Reinforcement learning models require an identification of what counts as positive or negative reinforcement. While emotional cues have a universal basis, there is considerable culture-dependence as well (Elfenbein &

[10] Note that these filters may seem more plausible than some of the others we have discussed, but are still under-defined, since we still need to show how infants create and/or detect categories such as "utterance","turn","human-relevant objects and actions", or any other type of data format posited by a given model.

Ambady, 2002). Finally, Supervised Learning models require a means of conveying that a given stimulus is a supposed output instead of an observation to be passively predicted. In the context of a social learner, the data preprocessing problem is therefore quite complex, and requires advanced operations like referential selection and communicative event classification to process feedback adequately.[11]

For instance, when a caregiver says: "A dog!", how to use this information will depend on what was said and done before this utterance, as follows. It could be used as a referential signal to trigger cross modal learning if the infant was pointing or looking at a dog, it could be used as a supervised target (corrective feedback) if the infant had wrongly said "a cat" or "a gog" while pointing or looking at a dog. It could be used as a reinforcement if the infant had said “a dog” in the correct context. It could also be useless in case the caregiver comments on something that the infant cannot see and will never discover.

In Figure 2, we try to systematically enumerate the different cases that could arise depending on the referential and communicative aspects of a simple interaction event of this sort, and link each case to the learning algorithms that would be most appropriate in this particular situation. We view this figure as representing an ideal learner, i.e., a learner whose ultimate goal is to learn language, and who is optimal at it. Thus, we operate under the best case scenario whereby the learner can pre-process the input adequately and can decide which case (cell) in Figure 2 it should be classified into (see Section 4.1 for further discussion).

[11] An advanced form of filtering is embodied in a class of learning algorithms called “active learning”. Here, instead of waiting for the relevant data to present itself to the learner, the learner acts on the environment to get the data she needs. For instance, in the case of supervised learning, the child would point to an object in the hope of eliciting a label from the caregiver, or vice versa would say a word in the hope of eliciting a pointing from the caregiver. In our framework, this would require two communicative turns, the first being initiated by the learner. Active learning can be modelled as an intrinsic reinforcement signal to promote exploration within a reinforcement learning algorithm.

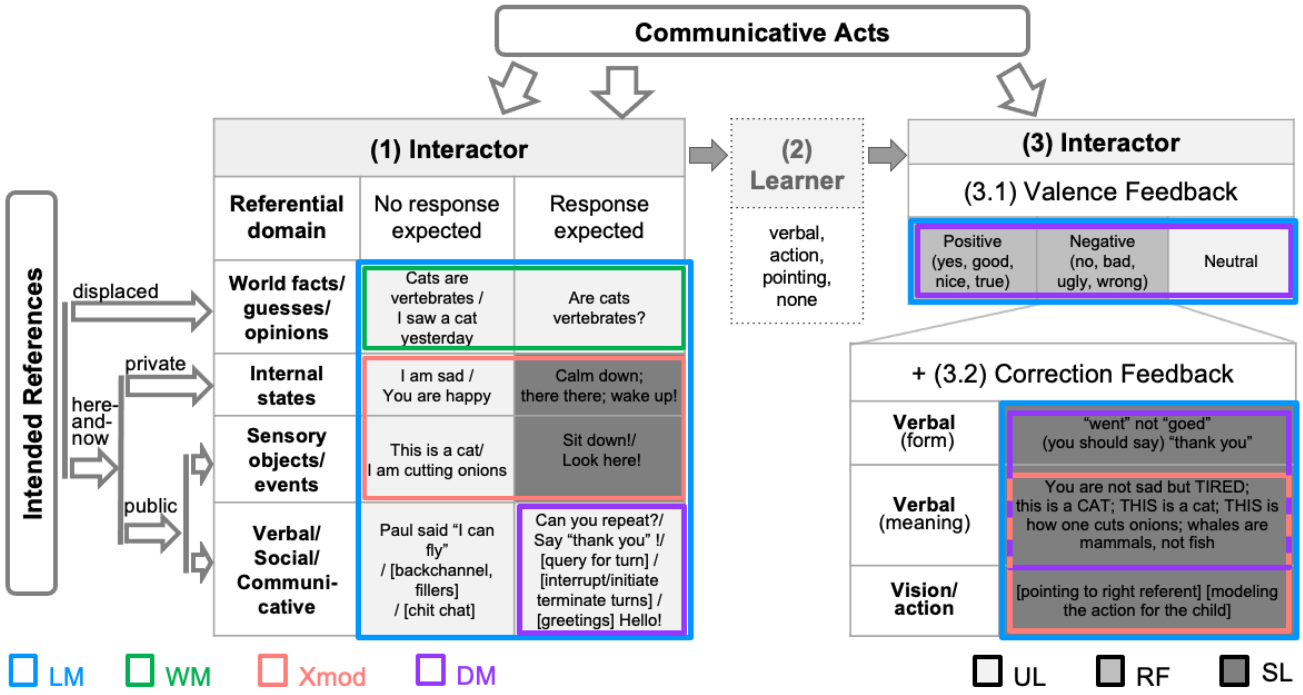


Figure 2. Possible events based on reference and communicative intentions and how they would relate to various probabilistic models and learning algorithms in an ideal learner. The models and algorithms are explained in more detail in Figure 1. We assume an interaction with a maximum of three turns, starting with the interactor. As explained in more detail in the text, the

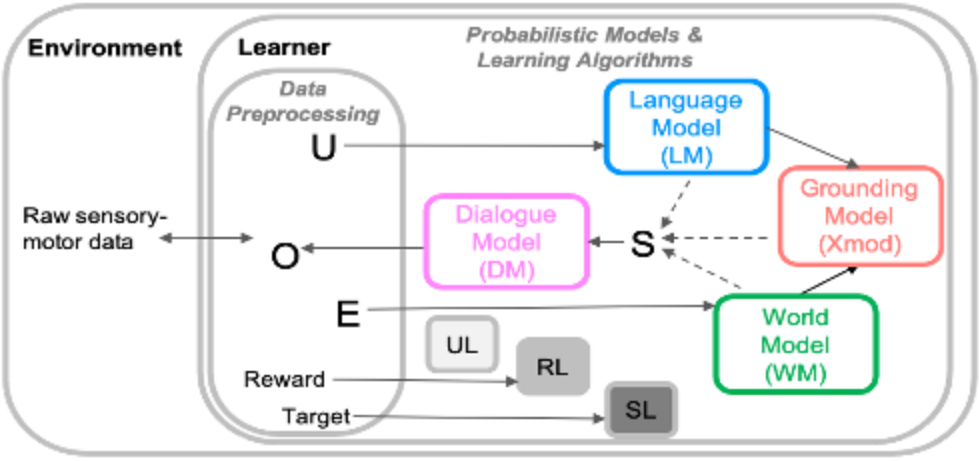


Figure 1. The learner's internal probabilistic models and learning algorithms. U: utterances, E: events, O: outputs

learner chooses the referential domain in terms of Intended References, and classifies Communicative Acts. The learner's turn is depicted, but is not part of the classification process. Colors indicate probabilistic models, and grey shading indicates learning algorithms.

We distinguish two dimensions necessary for filtering or data preprocessing of a social learner embedded in the real world, dimensions that can be used to classify the possible space of social-communicative input in an informative way with regard to the possible learning mechanisms involved (Figure 2). The first dimension regards **communicative acts,** and it determines how the communication as a whole is to be interpreted. The second dimension concerns **intended reference**, which establishes what in the environment or common ground is being talked about. We will now go into details for these two dimensions, asking the reader to bear with us, since the relevance of the classification process might only become obvious later on, when we link these classifications to the learning algorithms.

Communicative acts are used to parse a sequence of communicative actions *sequentially and causally* (see Section 4.1 for discussion of desirable expansions of this section). Much like in Speech Act Theory (Austin, 1962; Searle, 1969), its function is to describe and classify the actions that a communicator accomplishes by producing an utterance. This classifier allows the learner to decide (1) that the interactor is talking/signing; (2) whether the interactor requires a response or not; and (3) if the interactor's action actually occurs after an action by the learner, whether the interactor's action actually constitutes feedback, and whether this is positive or negative. For instance, the statement “Cats are vertebrates” would be classified into the column “No response expected”, while the question “Are cats vertebrates?” would go into the column “Response expected”. These two examples could (but do not have to) occur as a first conversational turn. The other two types of possible classifications in our scheme necessarily occur as a response to the learner (which does not need to be verbal, but could also be an action like pointing). Valence feedback constitutes a feedback with positive (e.g. “Well done”), negative (e.g. “That’s wrong”) or neutral (e.g. “I see”) valence. Only in case valence feedback is negative, it can be accompanied by corrective feedback (e.g., “THIS one is a cat, not that one”).

Intended references are primarily residing in the head of the speaker embedded in a particular context; they can be expressed by the listener via verbal and non verbal material plus the listener’s own perception of the context. Here, we consider the intended reference and separate different domains using a flowchart of binary questions. First, the interactor may be referring to something that is present here and now, or something that is abstract or displaced. For instance, the phrase “This is a cat” likely refers to something in the here and now, while “I saw a cat yesterday” does not have a referent in the present. If the interlocutor is referring to something occurring here and now, this reference may be to a "public" state. These public

states cover a wide range of occurrences, ranging from comments on objects and actions that are sensorily available to the learner (as in the previous cat example via vision, but also i.e., via audition, touch, smell, taste), to purely social comments including backchannels. Else, the reference could be “private” and refer to the interlocutor's (“I am sad”) or learner’s (“Calm down”) internal states. The former is inaccessible to the learner, but might be accompanied by sensorily accessible information (e.g., the caregiver has a sad face).

Communicative acts and intended references cross-cut each other resulting in 16 cells as described in Figure 2. In each cell one or several learning algorithms should be applied in the ideal learner case. We review these associations in the next section.

3.4 Linking Input Type and Algorithm

How can we link the types of preprocessed inputs described in Section 3.3 to the models and algorithms described in Sections 3.1 (Language Model, World Model, Grounding Model, Dialogue Model) and 3.2 (Unsupervised Learning, Reinforcement Learning, Supervised Learning)?

Language Modeling can be applied to all of the (clean) segmented utterances, since it only requires verbal language input. Moreover, Language Modeling can often be achieved without taking into account feedback, thus by Unsupervised Learning alone. This input can not only be used to update the structural probability of verbal forms, but also to learn the appropriate scripts, responses, and actions in a dialog in the interactor and learner's culture. This is why Language Modeling occurs in all of the cells of Figure 2, and Dialogue Modeling, which shares the property of learning about the sequences of events, in that case of a dialogue, in many of them.

In contrast, other learning algorithms are much more demanding in their inputs, as becomes clear when contemplating the problem of reference selection in learning of the meaning of words or utterances. For instance, the learner must establish whether the content of what is said refers to events in the here-and-now or to displaced or abstract contents. While both types of contents may be useful to build or update the World Model (i.e, are informative with respect to the world), the latter cannot or should not be interrogated to try to do Cross-Modal Learning (i.e., to learn sensory referents of verbal forms), since no sensory referent is present. Moreover, Cross-Modal Learning can be attempted for both private and the subset of public references that are accessible to the infant - but only when no response is expected from the infant. In the case of imperatives or questions, the learner cannot be certain that sensory references are available, and thus Cross-Modal Learning may be risky. To take a specific example, a request to "sit down" may be followed by the learner pointing to a chair, which the interactor may consider as appropriate or not depending on the situation. If considered appropriate, then the learner would have to further infer that the interactor was being lax in the definition, because a chair is not a good referent for the verb phrase "to sit down".

Next we turn to a more complex scenario, in which the learner is not just parsing each individual statement by the interactor, but rather considering a whole sequence in which the learner said or did something, and the interactor acts next. To begin with, the learner must determine that the interactor's action is causally connected to their own action. Assuming that this was determined to be the case, the learner still needs to decide the valence of this intervention, namely whether it is positive or negative (in which case it fully deserves the name of feedback) or neutral instead. All three cases are useful for Language and Dialogue Modeling, but only the former two, in which the learner needs to resort to Supervised Learning or Reinforcement Learning, can be used to reinforce a behavior (either positively or negatively). In

fact, inspecting negative feedback more closely, it becomes obvious that the learner must decide what level the negative feedback pertains to: The interactor may be correcting the verbal form the learner used, in which case the learner should try to update their Language or Dialog Model. Alternatively, the interactor may be correcting the meaning of what the infant said, or they may be correcting the sensory objects/events the infant was referring to. The latter two may be useful for Cross-Modal Learning. As in all other scenarios, we again observe that Language Modeling can be done in all of these cases.

In linking the inputs that infants can occur with learning algorithms, SCALa highlights the existence of a *feedback assignment problem*, which is the problem of determining when social feedback occurs and what to do with it in terms of learning algorithms.[12] This problem is more general than the traditional referential ambiguity problem noted in the philosophy of language (Quine, 1960). This problem is related to but also more general than the credit assignment problem (Minsky, 1961) in Reinforcement Learning. This problem runs as follows: The reinforcement often comes after a long series of actions; which action should be suppressed in case of failure, or enhanced in case of success? Moreover, what *counts* as positive versus negative feedback is not necessarily obvious (silence could be approving or disapproving, depending on the culture and situation). Also, it may not be obvious whether the feedback concerns the form or the content of the message. Referential cues like pointing are intrinsically ambiguous and need to be parsed and interpreted before being taken as face value. Corrective feedback needs to be identified as such and passed on to the correct algorithm. In the next sections, we explain how, despite this unresolved issue, SCALa helps reconcile previously expressed opposing views and may allow us to gain new insights into the language acquisition process.

[12] Note that theories formulated within the framework that do not rely on feedback need not worry about this problem.

4. Integrating Different Types of Language Inputs within SCALa

We have previously discussed three different sources of information that are pertinent to language acquisition: Language as Structure, Language as a Description of the World, and Language as a Social Event. In this section, we want to show that while each of these sources are relevant to a particular component or mode of operation of SCALa, they can be integrated as complementary information within this framework.

The Language as Structure view seems best captured as the Unsupervised Learning mode of operation for Language/Dialogue Models, which enable the learner to acquire structural aspects of language by tracking regularities in the input (positive evidence only). Yet, as we have seen, a given Language Model/Dialogue Model can be improved through other signals, especially through the presence of (interpretable) feedback, thus expressed in the Reinforcement Learning or Supervised Learning modes of the Language and Dialogue models (as proposed by Language as Description and Language as a Social Construct).

Language as a Description of the World can be linked to the Cross-Modal Learning algorithm. Crucially, this algorithm relies not only on a selection of pairs of utterances and events in the world, but may require the prior learning of structured representations for language and for the world. This is shown in the fact that while it is possible to learn correspondences between objects and words, it is more difficult to do so at the level of whole utterances and scenes. Since words do not come very often in isolation nor do objects, being able to segment utterances into words, and scenes into objects would probably help the algorithm (something that could be done in part through Unsupervised Language Modeling and World Modeling). In other words, Language as a Description of the World is not incompatible

with Language as Structure. On the contrary, it would seem that both types of learning should benefit from one another, something that can be explicitly tested within an integrated architecture like SCALa.

Finally, Language as a Social Construct corresponds to cases of Supervised Learning and Reinforcement Learning applied within and across linguistic and non-linguistic domains. Both of these learning components reflect the necessity of extracting the right feedback signal, which may itself require some culture-dependent tuning, which could be learned with Unsupervised Learning. Here, we only considered simple cases of social interactions (two turns in an interaction, feedback from caregiver to infant). But there could be more sophisticated feedback loops, including ones where the caregiver also learns.

Highlighting one view (e.g., Language as Structure versus Language as Social Construct) may lead to the impression that the views are incompatible. However, machine learning offers a very simple way to integrate these views, once they have been cast within a computation system: **Multi-Task Learning.** It boils down to optimizing the sum of several loss functions instead of a single one. When the losses are of different kinds, like when trying to maximize both input probability and expected reward, a scaling parameter is simply added to specify which of the two tasks is more important to the learner. An implementation of SCALa therefore would not only allow to integrate the different sources of information within a single system, but also to study their relative contribution quantitatively through the manipulation of this scaling factor. Multi-Task Learning also helps us understand the synergy between the different algorithms. Research into Reinforcement Learning has shown that while it is a powerful learning algorithm, it can require many iterations before a randomly initialized agent can learn anything. Combining it with Unsupervised Learning over observed agents can help. Similarly here, an integrated model like SCALa could help evaluate in a quantitative fashion the relative

contribution of the three sources of information at different points in development, and across linguistic and cultural contexts.

Cutting across these three different sources of information in language development, researchers have been debating for decades about the relative role of nature versus nurture, and about the role of domain specific versus domain-general mechanisms without reaching a clear conclusion. In a way, SCALa could be viewed as neutral with respect to these debates, as it is compatible with a wide range of theories. However, we argue that it is more than that: SCALa provides a roadmap, and perhaps the only one available, to solve these long-standing debates. Indeed, while most researchers would agree that learning is done through the interaction of a learner (who has inductive biases) and data (which has structure embedded into it), the difficult question is how much and what kinds of information is contained in the inductive biases as opposed to the data. It is very difficult to answer this question by sticking to verbally expressed theories that are not specific enough to make predictions on the basis of real input data. In contrast, an algorithmic implementation of SCALa, when tested on realistic data, will be able to answer in a quantitative way whether the inductive biases of the algorithms are sufficient or not to account for the emergence of linguistic knowledge, given this or that experience.

4.1 Limitations of SCALa

In this Section, we have so far highlighted some advantages of using SCALa, so before moving on we feel it is also important to highlight some limitations of our current paper, which we hope future researchers will improve upon. To begin with, SCALa is at present only a blueprint, and thus it does not provide fully elaborated computational or formal models allowing direct predictions that can be tested against data (for a discussion on how to relate models of the data

against theoretically-derived predictions in a post-verbal-theory age, see Robinaugh et al., 2020).

Moreover, this blueprint assumes an ideal learner, who would be able to route any input received in the right format into the relevant learning algorithm. Real infants may depart from optimal learners for a variety of reasons (e.g., their still developing attention and memory capacities). However, having an optimal model is a good starting point in developing causal theories of the infant learner, and can answer questions as to the functional role of particular types of inputs, as we have done in this paper. We return to violations of the ideal assumption in Section 5.2.

Additionally, we think that the current blueprint with its possible implementations in current machine learning covers quite well phenomena related to infant language perception, which also means at least two limits on the current scope of SCALa: (1), it does not contemplate as such cognitive skills that we view as less essential to language, including memory, attention, and executive functions; and (2), it may not appropriately cover many (perhaps most) aspects of pragmatic development (e.g., implicatures).

In fact, we are aware of limitations particularly in the socio-pragmatic components. To begin with, modeled conversations are described lasting for up to three conversational turns -- but this is not the only way conversation is structured. We trust readers with more expertise in conversation analysis and pragmatics will be better able to elaborate our framework further to establish which strategies the learner takes to classify acts and the sources of information feasible available to do so. For instance, syntactic properties of the phrases encountered can help classify them into statements and questions, but this is insufficient since both a statement and a question can require a response (e.g, "it's hot in here" and "can you let some air in?" both invite a behavioral response of opening the window; Searle, 1969). More likely, ideal listeners

take into account how a given turn in an interaction needs to be interpreted given the preceding turns (see Clark, 1996, p. 29-124, for an introduction; Casillas & Hilbrink, 2019, for an overview). Once these strategies are fully specified, we should be able to determine the learning algorithm that needs to be applied to the input in order for the learner to have the necessary information: the syntactic properties of the input could be parsed by Language Modeling, while keeping track of conversational turns would call for Dialogue Modeling.

5. Roadmap

While constructing SCALa, our blueprint of the ideal learner, we established a framework to classify the social-communicative content of a learner's input into intended reference and acts, which in turn can be processed by dedicated learning algorithms. In this section, we want to point out research avenues suggested by SCALa, as necessary to further our understanding of the role of social-communicative environmental input on learning (see Table 1 for a summary).

Table 1. Overview of proposed differential contributions by corpus analysts, computer modelers, and experimentalists to different research avenues.

	Algorithms	Input Data	Outcome measures	Integration
Corpus Analysis		Estimate	Measures of	

		prevalence of the various referential and event types	language output maturity	Explanations of outcome/input relationships in infants across cultures Predictions of outcomes of interventions
Computer Modeling	Implementation of probabilistic models, learning and preprocessing algorithms	Estimate of outcomes as a function of prevalence of referential/event types in the input for each combination of algorithm and preprocessing		
Experimental Studies	Proof-of-concept of preprocessing and learning algorithms		Measure of tacit knowledge (probabilistic models of infants)	

5.1 Priorities for Corpus Analysts

SCALa allows us to gain a systematic view of the different social-communicative types in an infant's daily speech input so as to assess their *prevalence*. Attempts at characterizing the prevalence of socio-communicative events are not new. For instance, early work suggests that mothers' speech to infants is mostly focused on the here-and-now, describing what is happening right now, what just happened or what is about to happen (Phillips, 1971). Other work has measured how often instances that allow Cross-Modal Learning, for instance when an object is named, are clear and unambiguous. A study measuring infants' attention to a target object named by their caregiver during naturalistic play found that instances in which infants attend to this target object in the time-window around the naming event account for only around 30% of instances, while infants do not attend to the target at all for 35%, and for some of the time for another 35% of instances (Yu, 2020). While studies as the one just described allow a glimpse into aspects of the information content of social-communicative interaction, to date there is no systematic assessment of the prevalence of different types of social-communicative input in a

standardized way. Our proposed scheme (Figure 2) offers a way to quantify the prevalence of different types of communicative input based on such data, and, as a consequence, provides insight into how often our ideal learner would need to apply the different learning algorithms. Of course, this table should be supplemented with a standardized method to enable the systematic and replicable annotation of corpora. Coming up with an implementation of an annotation scheme would be in and of itself a significant contribution. This is not a straightforward task. Before we can actually apply our classification scheme to natural communication settings, we need to devise a methodology to decide which utterance goes into which cell. This includes defining the appropriate units of reference, which could, for instance, be utterances or turns. The next step in the roadmap for putting SCALa to use would be to develop guidelines for annotation, ideally honing them by testing the annotation in a range of cultural and linguistic settings. In parallel, we should develop plausible and unsupervised versions of these classifiers integrated with the machine learning algorithms to produce quantitative predictions. The ACLEW Project (https://sites.google.com/view/aclewdid/home), and in particular Casillas' datasets (Casillas et al., 2019; 2020), may provide the ideal starting point for this, since it is cross-cultural and contains images that may allow the identification of nonverbal elements of the social interactions. ACLEW project members have illustrated the importance of coordinated data annotation for developing initial annotations (Casillas et al., 2017), as well as the usefulness of collaborating with experts of speech technology and machine learning to develop tools that speed up annotation and generalize analyses from the hand-annotated fraction to the day-long scale (Al Futaisi et al., 2019; Le Franc et al., 2018; Räsänen et al., 2020).

We would then be able to quantitatively assess whether there are systematic individual and/or cultural differences in the prevalence of the different types. For instance, it is possible that an infant encounters many examples of input that could be parsed using Unsupervised

Learning, but only few examples that could contribute to the Language Model using Supervised Learning. Or perhaps the two trade off, as in the case of the "elema" particle in Schieffelin's reports of the Kaluli: Kaluli infants are said to be talked to very little, but by around age two years, mothers and others actively model sentences and dialogues for the child, asking the child to repeat verbatim with the *elema* particle (Schieffelin, 1990). Thus, although American infants may benefit from a higher quantity of directed input, they do not enjoy these overtly supervised Language Modeling cases. This general ethnographic description can be improved by systematically annotating both types in an American and a Kaluli database.

Let us take a different example. Rabain-Jamin (e.g., Rabain-Jamin & Sabeau-Jouannet, 1997) reports that Wolof mothers tend to speak little of the neighboring physical environment and organize conversations in a multiadic fashion (involving others), whereas French mothers make frequent reference to the physical environment in essentially dyadic interaction, based on relatively short observations of 4-5 mothers of 4-month-olds. What is the extent of the differences in the structure and topic of conversations across a wider range of ages and cultures when a daylong coverage is obtained? These findings will have crucial consequences for considering the extent to which Dialogue Modeling versus Cross-Modal Learning are called for across such settings.

A final example is based on ethnographic evidence: Kulick (1997) describes a village in which parents systematically give wrong information to their infants, for instance saying "Look at the pig" while pointing to nothing in particular (p. 121). Kulick describes this as such a common feature of child-directed speech that children's understanding of this routine reflects errors in comprehension, whereby children point to a tree while saying "the pig", errors which are overtly corrected by parents. However, it is unclear from this description how frequently caregivers engage in misleading pointing, and whether this only occurs in the context of one particular

routine. A more systematic approach to these kinds of cultural reports would provide language acquisition experts with crucial food for thought regarding the extent of variation of infants' experiences.

5.2 Priorities for Computational Modelers

SCALa invites us to run computational experiments on existing or synthetic datasets to test the effect of variation of the prevalence of different types of input for an optimal learner. Although it will be crucial to use the product of investigations as those described in Section 5.1, we do not need to await them. Instead, we can (and should) generate synthetic corpora with different raw quantities and proportions of these types of information, and program model learners with access to a variety types of algorithms, to study whether there can be trade-offs in their effects on learning, and if so, how.

To take a precise example, recall the case of Wolof versus French mothers, who differ on their object-orientedness. We could constitute artificial corpora with these properties and test a range of model learners. We may find that, regardless of how we set up our model learning, the learning of certain word categories (e.g., nouns for objects) depends crucially on Cross-Modal learning and is affected by differences in the prevalence of object-orientedness, whereas other word categories (e.g., social terms) are not. We would then be able to predict precise differences in vocabulary development corresponding to differences in prevalence of these kinds of inputs in actual human learners (see Section 5.4).

Going further, to the extent that it is possible to evaluate the computational complexity of these different algorithms, would a learner profit more from highly prevalent and easy-to-process, but low-information data, or from rare, hard-to-process, but high-information data? The information content of instances that can be parsed by Unsupervised Language

Modeling is lower compared to those that can be parsed by Supervised Language Modeling. Would the profit model learners draw change as a function of performance modules, representing processing skills and working memory? These are fascinating questions, which will be extremely hard to address even with today's advanced methods to study infant language development (see Section 5.4). By adopting a computational approach like SCALa, we can attempt to study them *in vitro*, assessing the potential effects of variation in processing skills and working memory in our model learner without wasting infants' and caregivers' precious resources, and even including in our experiments cases of deprivation that we would be unable to find in the real world. For instance, if we find a range of experiences whereby advances in, say, Language Modeling and Dialogue Modeling can trade off with each other to nonetheless result in apparently normal language acquisition, then all it may remain to do is check whether all human cultures (as described by the corpora studied as proposed in Section 5.1) do fall within this range. If any culture does not, then we need only test our predictions in that one culture.

Finally, computational models would also enable us to test what would happen in the case of a non-ideal learner. What happens if Cross-Modal Learning is applied indiscriminately to everything instead of to only highly informative pairs of words and objects? What happens if corrective feedback is undetected or ignored? This would allow us to make more precise predictions, in particular if it turns out that the automatic filtering by input types is a difficult problem (which we suspect it is). Generating all of these predictions is already feasible given some of the computational tools and datasets currently available, and it is only ethically and technically plausible when dealing with computational learners who can be tested 24/7 and in extreme conditions.

5.3 Priorities for Experimentalists

We believe one of the key contributions experimentalists stand to make is by checking for the presence of learning mechanisms via proof-of-concept studies. We believe there is ample work documenting the basic mechanisms underlying Language Modeling, for instance in statistical learning and artificial language learning work (Black & Bergmann, 2017; Gomez & Gerken, 2000). Cross-Modal Learning also has a rich literature backing up the presence of these mechanisms in the young child (e.g., Smith & Yu, 2008). Work on naïve physics contributes to documenting basic mechanisms contributing to World Modeling (Schilling & Clifton, 1998). In contrast, we believe fewer of our colleagues have tried to assess the presence of mechanisms underlying children's Dialogue Modeling. There is certainly a rich descriptive literature based on corpus studies (Guijarro & Sanz, 2008, to cite just one example of many), and some work focusing on infants' use of e.g. lexical, syntactic, and phonological cues when parsing a third-party interaction (e.g., Casillas & Frank, 2017). However, we invite experimentalists to consider creating artificial languages and situations based on dyadic and multiadic interactions, in order to more carefully isolate properties, and provide proof-of-concept evidence for young children's use of these properties when learning about the structure and content of conversations.

5.4 Weaving Everything Together

Corpora can provide us with data to characterize the input; computational modeling with constraints regarding what kind of learning mechanisms a learner would need to profit from such input; and experimental work with proof-of-concept demonstration of the presence of such mechanisms. But we are missing one crucial step: The demonstration that actual children, in real life, do use the input in that predicted way.

The final step is to use SCALa help to interpret inter-individual or cross-cultural differences in language outcomes. Armed with the knowledge gained in research projects described in Sections 5.1-5.3, we can sample human cultures based on descriptions of the input to check for the predicted patterns of faster versus slower development of specific kinds of words or structures, of words in specific types of sentences, and of interaction patterns. We can then employ any of the tools in our kit. We can use perception or production experiments, including studies where we check whether a mechanism is active regardless of how much relevant data parents in that culture provide their children with (Liszkowski, Brown, Callaghan, Takada, & De Vos, 2012). We can measure child outputs in targeted or week-long observations. And we can check for vocabulary counts and compositions using parental questionnaires, or any other similar strategy). The key point is that, by developing these predictions via studies in Sections 5.1-5.3, we will truly be in a hypothesis-testing setting, and we can even modify our experiments to avoid confounds (e.g., we can test our stimuli on model learners to make sure that only model learners having algorithm X can succeed, but not if they have algorithms P-Q).

6. Conclusion

Influence factors on infant language acquisition are multifaceted and show individual and cultural variation. SCALa allows us to understand and classify different socio-communicative information, by linking them to described learning algorithms. This opens the doors to resolving outstanding issues on the relative contributions of innate versus acquired biases, as well as to integrate multiple theoretical views on language acquisition in an unprecedented way.

Acknowledgment

This paper is written in honour of Jacques Mehler (JM), who has shaped the field of early language acquisition through groundbreaking experimental investigations of the human newborn, and more generally by pushing a theory-driven view of cognitive development nourished by work in theoretical linguistics and philosophy of mind. JM has also been a very influential PhD advisor and mentor, including to the last author of this paper (ED). It is ED's opinion that JM might have been ambivalent regarding this paper. In a way, it strays too far from the approach he always professed, which can be summarised by the idea that the only way to study development is to study (1) the stable state (in the adult) (2) the initial state (in the newborn), while (3) avoiding anything in the middle that might be too descriptive, behavioristic, or otherwise uninteresting. After all, the idea was that (2) was going to reveal such a rich initial state that learning (3) would become rather trivial. Here, this paper is doing exactly the opposite, i.e. focusing exclusively on (3) by investigating learning algorithms and describing the input based on corpus of interactions. To make things worse, this paper does NOT take a strong nativist view and is quoting a great deal of connectionist work. Is this a case of high treason? We think it may be more useful to view this in light of the evolution of the entire field. With the replication crisis, experimentation in infants may no longer be the primary path to uncover infants' skills; at the same time, big data approaches have renewed the field of corpus studies allowing them to be more than just descriptive. However, the standard Mehlerian way of framing language acquisition in terms of an unfolding of an initial-state into a stable state is not abandoned, but is here taken as a starting point that needs elaboration. His insistence on adopting a theoretical stance on development is not relinquished but strengthened with a computational approach. More generally, Jacques Mehler’s taste for strong interdisciplinarity

and methodological opportunism applied to the strengthening of cognitive science would probably have found some redeeming qualities to this paper.

ST has been supported by the Japanese Society for the Promotion of Science KAKENHI [grant no. 19K23361], AC by the Agence Nationale pour 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. ED's contributions at EHESS have been supported by the Agence Nationale pour la Recherche [ANR-19-P3IA-0001 PRAIRIE 3IA Institute] and a CIFAR grant [Learning in Machines and Brain]. We thank Marisa Casillas for helpful discussion.

References

Akhtar, N., Carpenter, M., & Tomasello, M. (1996). The role of discourse novelty in early word learning. *Child Development, 67*(2), 635-645. https://doi.org/10.1111/j.1467-8624.1996.tb01756.x

Al Futaisi, N., Zhang, Z., Cristia, A., Warlaumont, A., & Schuller, B. (2019). VCMNet: Weakly supervised learning for automatic infant vocalisation maturity analysis. *Proceedings of the 2019 International Conference on Multimodal Interaction* (pp. 205-209). https://doi.org/10.1145/3340555.3353751

Austin, J. L. (1962.) *How to do things with words.* Oxford: Oxford University Press.

Baevski, A., Schneider, S., & Auli, M. (2019). *vq-wav2vec: Self-supervised learning of discrete speech representations.* ArXiv. https://arxiv.org/pdf/1910.05453.pdf

Baevski, A., & Mohamed, A. (2020). Effectiveness of self-supervised pre-training for asr. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7694-7698). IEEE.

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., & Tor, S. (2019). Day by day, hour by hour: Naturalistic language input to infants. *Developmental Science, 22*(1), e12715. https://doi.org/10.1111/desc.12715

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2018). What do North American babies hear? A large-scale cross-corpus analysis. *Developmental Science, 22*(1), 1-12. https://doi.org/10.1111/desc.12724

Black, A., & Bergmann, C. (2017). Quantifying infants' statistical word segmentation: A meta-analysis. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 124-129). Cognitive Science Society.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. ArXiv. https://arxiv.org/pdf/2005.14165.pdf

Casillas, M., Brown, P., & Levinson, S. C. (2019). Early language experience in a Tseltal Mayan village. *Child Development, Early View.* https://doi.org/10.1111/cdev.13349

Casillas, M., Brown, P., & Levinson, S. C. (2020). Early language experience in a Papuan community. Journal of Child Language.

Casillas, M., & Frank, M. C. (2017). The development of children's ability to track and predict turn structure in conversation. *Journal of Memory and Language, 92*, 234-253. https://doi.org/10.1016/j.jml.2016.06.013

Casillas, M., Bergelson, E., Warlaumont, A. S., Cristia, A., Soderstrom, M., VanDam, M., & Sloetjes, H. (2017). A new workflow for semi-automatized annotations: Tests with long-form naturalistic recordings of children's language environments. *Proceedings of Interspeech 2017* (pp. 2098-2102).

Casillas, M., & Hilbrink, E. (2019). Communicative act development. In K. P. Schneider, & E. Ifantidou (Eds.), *Developmental and Clinical Pragmatics*. De Gruyter Mouton.

Clark, H. H. 1996. *Using language.* Cambridge: Cambridge University Press.

Clark, A., Donahue, J., & Simonyan, K. (2019). Efficient video generation on complex datasets. ArXiv. https://arxiv.org/abs/1907.06571

Cristia, A. (2018). Can infants learn phonology in the lab? A meta-analytic answer. *Cognition, 170*, 312-327. https://doi.org/10.1016/j.cognition.2017.09.016

Chomsky, N. (1964). *Current Issues in Linguistic Theory.* The Hague: Mouton.

Corbett, G.G. (2013). Number of Genders. In: Dryer, M. S. & Haspelmath, M. (eds.) *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/30

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences, 13*(4), 148-153. https://doi.org/10.1016/j.tics.2009.01.005

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. ArXiv. https://arxiv.org/abs/1810.04805

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition, 173,* 43-59. https://doi.org/10.1016/j.cognition.2017.11.008

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171,* 303-306. https://doi.org/10.1126/science.171.3968.303

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological Bulletin, 128*(2), 203-235. https://doi.org/10.1037/0033-2909.128.2.203

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language. In: Sekerina, I.A., Fernández, E.M., & Clahsen, H. (Eds.).*Developmental Psycholinguistics: On-line Methods in Children's Language Processing* (pp. 97-135), John Benjamins: Amsterdam

Fried, E. I. (2020) Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry, 31*(4), 271-288, https://doi.org/10.1080/1047840X.2020.1853461

Ganek, H. V., & Eriks-Brophy, A. (2018). A concise protocol for the validation of Language ENvironment Analysis (LENA) conversational turn counts in Vietnamese. *Communication Disorders Quarterly, 39*(2), 371-380. https://doi.org/10.1177/1525740117705094

Gervain, J., & Werker, J. F. (2013). Prosody cues word order in 7-month-old bilingual infants. *Nature Communications, 4*(1), 1-6.

Goldin-Meadow, S., & Mylander, C. (1998). Spontaneous sign systems created by deaf children in two cultures. *Nature, 391*(6664), 279-281.

Gómez, R. L., & Gerken, L. A. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4*(5), 178-186. https://doi.org/10.1016/S1364-6613(00)01467-4

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep Learning* (Vol. 1). Cambridge: MIT Press.

Guest, O., & Martin, A. E. (2020). *How computational modeling can force theory building in psychological science.* PsyArXiv. https://doi.org/10.31234/osf.io/rybh9

Guijarro, J. M., & Sanz, M. J. P. (2008). Compositional, interpersonal and representational meanings in children's narrative: A multimodal discourse analysis. *Journal of Pragmatics, 40*(9), 1601-1619. https://doi.org/10.1016/j.pragma.2008.04.019

Harwath, D., & Glass, J. R. (2017). *Learning word-like units from joint audio-visual analysis.* ArXiv preprint. arXiv:1701.07481.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science, 298*(5598), 1569-1579. DOI: 10.1126/science.298.5598.1569

Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., ... & Wainwright, M. (2017). Grounded language learning in a simulated 3d world. ArXiv. https://arxiv.org/abs/1706.06551

Hilbrink, E. E., Gattis, M., & Levinson, S. C. (2015). Early developmental changes in the timing of turn-taking: a longitudinal study of mother–infant interaction. *Frontiers in Psychology, 6*, 1492. https://doi.org/10.3389/fpsyg.2015.01492

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 74*(5), 1368-1378. https://doi.org/10.1111/1467-8624.00612

Hollich, G., Hirsh‐Pasek, K., & Golinkoff, R. M. (2000). II. The emergentist coalition model. *Monographs of the Society for Research in Child Development, 65*(3), 17-29.

Jing, L., & Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* https://doi.org/*10.1109/TPAMI.2020.2992393*

Kellman, P. J. and Spelke, E. S. (1983). Perception of partly occluded objects in infancy. *Cognitive Psychology, 15*(4), 483–524. https://doi.org/10.1016/0010-0285(83)90017-8

Kim, K., Sano, M., De Freitas, J., Haber, N., & Yamins, D. (2020). *Active World Model Learning with Progress Curiosity.* ArXiv. https://arxiv.org/pdf/2007.07853.pdf

Kuhl, P. K. (2007). Is speech learning ‘gated’ by the social brain? *Developmental Science, 10*(1), 110-120. doi: 10.1111/j.1467-7687.2007.00572.x

Kulick, D. (1997). *Language shift and cultural reproduction: Socialization, self and syncretism in a Papua New Guinean village, 14.* New York: Cambridge University Press.

Landau, B., Gleitman, L. R., & Landau, B. (2009). *Language and Experience: Evidence from the Blind Child* (Vol. 8). Cambridge: Harvard University Press.

Lavechin, M., de Seyssel, M., Gautheron, L., Dupoux, E., & Cristia, A. (2021). Reverse-engineering language acquisition with child-centered long-form recordings. PsyArxiv. https://doi.org/10.31234/osf.io/pt9xq

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). *Albert: A lite bert for self-supervised learning of language representations.* ArXiv. https://arxiv.org/pdf/1909.11942

Lazaridou, A., & Baroni, M. (2020). *Emergent Multi-Agent Communication in the Deep Learning Era.* arXiv preprint arXiv:2006.02419.

Le Franc, A., Riebling, E., Karadayi, J., Wang, Y., Scaff, C., Metze, F., & Cristia, A. (2018). The ACLEW DiViMe: An Easy-to-use Diarization Tool. *Proceedings of Interspeech* (pp. 1383-1387).

Lieven, E. (1978). Conversations between mothers and young children: Individual differences and their possible implications for the study of language learning. In N. Waterson & C. Snow (Eds.), *The Development of Communication*. Chichester: John Wiley.

Liszkowski, U., Brown, P., Callaghan, T., Takada, A., & De Vos, C. (2012). A prelinguistic gestural universal of human communication. *Cognitive Science, 36*(4), 698-713. https://doi.org/10.1111/j.1551-6709.2011.012

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk.* 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates. https://doi.org/10.21415/3mhn-0z8928.x

Maddieson, I. (2013). Consonant Inventories. In: Dryer, M. S. & Haspelmath, M. (eds.). *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/1

Mehler, J. (1981). The role of syllables in speech processing: Infant and adult data. *Philosophical Transactions of the Royal Society of London B, Biological Sciences, 295*(1077), 333-352. https://doi.org/10.1098/rstb.1981.0144

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE, 49*(1), pp. 8-30.

Ochs, E., & Schieffelin, B. B. (2011). The theory of language socialization. In A. Duranti,

E. Ochs, & B. Schieffelin (Eds.), *The Handbook of Language Socialization* (pp. 1–21). Malden, MA: Wiley-Blackwell.

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language, 95*(1), e41-e74.

Pearl, J. (1998). Graphical models for probabilistic and causal reasoning. In *Quantified representation of uncertainty and imprecision* (pp. 367-389). Springer, Dordrecht.

Pereira, A. F., Smith, L. B., & Yu, C. (2014). A bottom-up view of toddler word learning. *Psychonomic Bulletin & Review*, *21*(1), 178-185. https://doi.org/10.3758/s13423-013-0466-4

Phillips, J. R. (1971). Formal characteristics of speech which mothers address to their young children. Dissertation Abstracts International, 31(7-B), 4369–4370.

Piaget, J. (1951). *Play, Dreams, and Imitation in Childhood.* New York: Norton.

Prince, A., & Smolensky, P. (2004). *Optimality Theory: Constraint Interaction in Generative Grammar.* Oxford: Blackwell.

Quine, W. V. O., 1960. *Word and Object.* Cambridge: MIT Press,

Rabain-Jamin, J., & Sabeau-Jouannet, E. (1997). Maternal speech to 4-month-old infants in two cultures: Wolof and French. *International Journal of Behavioral Development, 20*(3), 425-451.

Räsänen, O., Seshadri, S., Lavechin, M., Cristia, A., & Casillas, M. (2020). ALICE: An open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings. Accepted for publication in *Behavior Research Methods*.

Ratner, N., & Bruner, J. (1978). Games, social exchange and the acquisition of language. *Journal of Child Language, 5*(3), 391-401.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. ArXiv. https://arxiv.org/abs/1605.05396

Robinaugh, D., Haslbeck, J., Ryan, O., Fried, E. I., & Waldorp, L. (2020). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. https://psyarxiv.com/ugz7y

Riochet, R., Sivic, J., Laptev, I., & Dupoux, E. (2020). Occlusion resistant learning of intuitive physics from videos. ArXiv. https://arxiv.org/abs/2005.00069

Rosenblith, J. F. (1959). Learning by imitation in kindergarten children. *Child Development,* 69-80.

Saenz, M., & Langers, D. R. (2014). Tonotopic mapping of human auditory cortex. *Hearing Research, 307*, 42-52. https://doi.org/10.1016/j.heares.2013.07.016

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926-1928. https://doi.org/10.1126/science.274.5294.1926

Schieffelin, B. B. (1990). *The give and take of everyday life: Language, socialization of Kaluli children*, 9. Cambridge: CUP Archive.

Schilling, T. H., & Clifton, R. K. (1998). Nine-month-old infants learn about a physical event in a single session: Implications for infants' understanding of physical phenomena. *Cognitive Development, 13*(2), 165-184. https://doi.org/10.1016/S0885-2014(98)90037-5

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., ... & Lillicrap, T. (2019). *Mastering atari, go, chess and shogi by planning with a learned model.* ArXiv. https://arxiv.org/abs/1911.08265

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language.* Cambridge: Cambridge University Press.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3), 1558-1568. https://doi.org/10.1016/j.cognition.2007.06.010

Snow, C. E. (1977). Mothers' speech research: from input to interaction. In Catherine E. Snow & Charles A. Ferguson (eds.). *Talking to Children, (*pp. 31--49). Cambridge University Press.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review, 27*(4), 501-532. https://doi.org/10.1016/j.dr.2007.06.002

Soderstrom, M., Grauer, E., Dufault, B., & McDivitt, K. (2018). Influences of number of adults and adult: child ratios on the quantity of adult language input across childcare settings. *First Language, 38*(6), 563-581. https://doi.org/10.1177/0142723718785013

Suarez-Rivera, C., Smith, L. B., & Yu, C. (2019). Multimodal parent behaviors within joint attention support sustained attention in infants. *Developmental Psychology, 55*(1), 96. https://doi.org/10.1037/dev0000628

Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). Videobert: A joint model for video and language representation learning. *Proceedings of the IEEE International Conference on Computer Vision* (pp. 7464-7473).

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 135). Cambridge: MIT Press.

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1536), 3617-3632. https://doi.org/10.1098/rstb.2009.0107

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In *Advances in neural information processing systems* (pp. 59-65).

Tincoff, R., & Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy, 17*(4), 432-444. https://doi.org/10.1111/j.1532-7078.2011.00084.x

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition.* Harvard University Press.

van Rooij, I., & Baggio, G. (2020). Theory development requires an epistemological sea change. *Psychological Inquiry, 31*(4), 321-325.

Vogt, P., Mastin, J. D., & Schots, D. M. (2015). Communicative intentions of child-directed speech in three different learning environments: Observations from the Netherlands, and rural and urban Mozambique. *First Language, 35*(4-5), 341-358. https://doi.org/10.1177/0142723715596647

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge: Harvard University Press.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S. F., & Bowman, S. R. (2020). BLiMP: *The benchmark of linguistic minimal pairs for English. Transactions of the Association for Computational Linguistics*, 8, 377-392.

Weisleder, A., & Fernald, A. (2013). Talking to children matters early language experience strengthens processing and builds vocabulary. *Psychological Science, 24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development, 1*(2), 197-234. https://doi.org/10.1080/15475441.2005.9684216

Werker, J. F., Yeung, H. H., & Yoshida, K. A. (2012). How do infants become experts at native-speech perception? *Current Directions in Psychological Science, 21*(4), 221-226. https://doi.org/10.1177/0963721412449459

Woodward, A. L., Markman, E. M., & Fitzsimmons, C. M. (1994). Rapid word learning in 13-and 18-month-olds. *Developmental psychology, 30*(4), 553.

Xu, D., Yapanel, U., & Gray, S. (2009). *LENA TR-05: Reliability of the LENA Language ENvironment Analysis System in young children's natural home environment.* Boulder, CO: LENA Foundation.

Yu, C. (2020). *Multiple sensorimotor pathways to parent-infant coordinated attention in naturalistic toy play* [Oral presentation]. International Conference on Infant Studies 2020, Online Conference.