

Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner*

Emmanuel Dupoux

EHESS, ENS, PSL Research University, CNRS, INRIA, Paris, France
emmanuel.dupoux@gmail.com, www.syntheticlearner.net

Abstract

Spectacular progress in the information processing sciences (machine learning, wearable sensors) promises to revolutionize the study of cognitive development. Here, we analyse the conditions under which 'reverse engineering' language development, i.e., building an effective system that mimics infant's achievements, can contribute to our scientific understanding of early language development. We argue that, on the computational side, it is important to move from toy problems to the full complexity of the learning situation, and take as input as faithful reconstructions of the sensory signals available to infants as possible. On the data side, accessible but privacy-preserving repositories of home data have to be setup. On the psycholinguistic side, specific tests have to be constructed to benchmark humans and machines at different linguistic levels. We discuss the feasibility of this approach and present an overview of current results.

Keywords

Artificial intelligence, speech, psycholinguistics, computational modeling, corpus analysis, early language acquisition, infant development, language bootstrapping, machine learning.

Highlights

- Key theoretical puzzles of infant language development are still unsolved.
- A roadmap for reverse engineering infant language learning using AI is proposed.
- AI algorithms should use realistic input (little or no supervision, raw data).
- Realistic input should be obtained by large scale recording in ecological environments.

- Machine/humans comparison should be run on a benchmark of psycholinguistic tests.

1 Introduction

In recent years, artificial intelligence (AI) has been hitting the headlines with impressive achievements at matching or even beating humans in complex cognitive tasks (playing go or video games: ?, ?, ?; processing speech and natural language: ?, ?, ?; recognizing objects and faces: ?, ?, ?) and promising a revolution in manufacturing processes and human society at large. These successes show that with statistical learning techniques, powerful computers and large amounts of data, it is possible to mimic important components of human cognition. Shockingly, some of these achievements have been reached by throwing out some of the classical theories in linguistics and psychology, and by training relatively unstructured neural network systems on large amounts of data. What does it tell us about the underlying psychological and/or neural processes that are used by humans to solve these tasks? Can AI provide us with scientific insights about human learning and processing?

Here, we argue that developmental psychology and in particular, the study of language acquisition is one area where, indeed, AI and machine learning advances can be transformational, provided that the involved fields make significant adjustments in their practices in order to adopt what we call the *reverse engineering approach*. Specifically:

The reverse engineering approach to the study of infant language acquisition consists in constructing *scalable* computational systems that can, when fed with *realistic* input data, *mimic* language acquisition as it is observed in infants.

The three italicised terms will be discussed at length in subsequent sections of the paper. For now, only an intuitive understanding of these terms will suffice. The idea of using machine learning or AI techniques as a means to study child's language learning is actually not new (to name a few: ?, ?, ?, ?, ?) although relatively few studies have concentrated on the early phases of language learning (see ?, ?, for a pioneering collection of essays). What is new, however, is that whereas previous AI approaches were limited to proofs

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

of principle on toy or miniature languages, modern AI techniques have scaled up so much that end-to-end language processing systems working with real inputs are now deployed commercially. This paper examines whether and how such unprecedented change in scale could be put to use to address lingering scientific questions in the field of language development.

The structure of the paper is as follows: In Section ??, we present two deep scientific puzzles that large scale modeling approaches could in principle address: solving the bootstrapping problem, accounting for developmental trajectories. In Section ??, we review past theoretical and modeling work, showing that these puzzles have not, so far, received an adequate answer. In Section ??, we argue that to answer them with reverse engineering, three requirements have to be addressed: (1) modeling should be computationally scalable, (2) it should be done on realistic data, (3) model performance should be compared with that of humans. In Section ??, recent progress in AI is reviewed in light of these three requirements. In Section ??, we assess the feasibility of the reverse engineering approach and lay out the road map that has to be followed to reach its objectives

, and we conclude in Section ??.

2 Two deep puzzles of early language development

Language development is a theoretically important subfield within the study of human cognitive development for the following three reasons:

First, the linguistic system is uniquely *complex*: mastering a language implies mastering a combinatorial sound system (phonetics and phonology), an open ended morphologically structured lexicon, and a compositional syntax and semantics (e.g., ?, ?). No other animal communication system uses such a complex multilayered organization (?, ?). On this basis, it has been claimed that humans have evolved (or acquired through a mutation) an innately specified computational architecture to process language (see ?, ?, ?).

Second, the overt manifestations of this system are extremely *variable* across languages and cultures. Language can be expressed through the oral or manual modality. In the oral modality, some languages use only 3 vowels, other more than 20. Consonants inventories vary from 6 to more than 100. Words can be mostly composed of a single syllable (as in Chinese) or long strings of stems and affixes (as in Turkish). Semantic roles can be identified through fixed positions within constituents, or be identified through functional morphemes, etc. (see ?, ?, for a typology of language variation). Evidently, infants acquire the relevant variant through learning, not genetic transmission.

Third, the human language capacity can be viewed as a finite computational system with the ability to generate a (virtual) infinity of utterances. This turns into a *learnability problem* for infants: on the basis of finite evidence, they

have to induce the (virtual) infinity corresponding to their language. As has been discussed since Aristotle, such induction problems do not have a generally valid solution. Therefore, language is simultaneously a human-specific biological trait, a highly variable cultural production, and an apparently intractable learning problem.

Despite these complexities, most infants spontaneously learn their native(s) language(s) in a matter of a few years of immersion in a linguistic environment. The more we know about this simple fact, the more puzzling it appears. Specifically, we outline two deep scientific puzzles that a reverse engineering approach could, in principle help to solve: solving the bootstrapping problem and accounting for developmental trajectories. The first puzzle relates to the ultimate outcome of language learning: the so-called *stable state*, defined here as the stabilized language competence in the adult. The second puzzle relates to what we know of the intermediate steps in the acquisition process, and their variations as a function of language input.¹

2.1 Solving the bootstrapping problem

The stable state is the operational knowledge which enables adults to process a virtual infinity of utterances in their native language. The most articulated description of this stable state has been offered by theoretical linguistics; it is viewed as a grammar comprising several components: phonetics, phonology, morphology, syntax, semantics, pragmatics.

The *bootstrapping problem* arises from the fact these different components appear *interdependent* from a learning point of view. For instance, the phoneme inventory of a language is defined through pairs of words that differ minimally in sounds (e.g., "light" vs "right"). This would suggest that to learn phonemes, infants need to first learn words. However, from a processing viewpoint, words are recognized through their phonological constituents (e.g., ?, ?), suggesting that infants should learn phonemes before words. Similar paradoxical co-dependency issues have been noted between other linguistic levels (for instance, syntax and semantics: ?, ?, prosody and syntax: ?, ?). In other words, in order to learn any one component of the language competence, many others need to be learned first, creating apparent circularities.

The bootstrapping problem is further compounded by the fact that infants do not have to be taught formal linguistics or language courses to learn their native language(s). As in other cases of animal communication, infants *spontaneously* acquire the language(s) of their community by merely being

¹The two puzzles are not independent as they are two facets of the same phenomenon. In practice, proposals for solving the bootstrapping problem may offer insights about the observed trajectories. Vice-versa, data on developmental trajectories may provide more manageable subgoals for the difficult task of solving the bootstrapping problem.

immersed in that community (?, ?). Experimental and observational studies have revealed that infants start acquiring elements of their language (phonetics, phonology, lexicon, syntax and semantics) even before they can talk (?, ?, ?, ?), and therefore before parents can give them much feedback about their progress into language learning. This suggests that language learning (at least the initial bootstrapping steps) occurs largely *without supervisory feedback*.²

The reverse engineering approach has the potential of solving this puzzle by providing a computational system that can demonstrably bootstrap into language when fed with similar, supervisory poor, inputs³.

2.2 Accounting for developmental trajectories

In the last forty years, a large body of empirical work has been collected regarding infant's language achievements during their first years of life. This work has only added more puzzlement.

First, given the multi-layered structure of language, one could expect a stage-like developmental tableau where acquisition would proceed as a discrete succession of learning phases organized logically or hierarchically (e.g., building linguistic structure from the low level to the high levels). This is not what is observed (see Figure ??). For instance, infants start differentiating native from foreign consonants and vowels at 6 months, but continue to fine tune their phonetic categories well after the first year of life (e.g., ?, ?). However, they start learning about the sequential structure of phonemes (phonotactics, see ?, ?) way *before* they are done acquiring the phoneme inventory (?, ?). Even before that, they start acquiring the meaning of a small set of common words (e.g. ?, ?). In other words, instead of a stage-like developmental tableau, the evidence shows that acquisition takes places at all levels more or less simultaneously, in a *gradual* and largely *overlapping* fashion.

Second, observational studies have revealed considerable *variations* in the *amount of language input* to infants across cultures (?, ?) and across socio-economic strata (?, ?), some of which can exceed an order of magnitude (?, ?, p. 2146; ?, ?; see also Supplementary Section S1). These variations do impact language achievement as measured by vocabulary size and syntactic complexity (?, ?, ?, ?, ?, among others), but at least for some markers of language achievement, the differences in outcome are much less extreme than the variations in input. For canonical babbling, for instance, an order of magnitude would mean that some children start to babble at 6 months, and others at 5 years! The observed range is between 6 and 10 months, less than a 1 to 2 ratio. Similarly, reduced range of variations are found for the onset of word production and the onset of word combinations. This suggests a surprising level of *resilience* in language learning, i.e., some minimal amount of input is sufficient to trigger certain landmarks.

The reverse engineering approach has the potential of accounting for this otherwise perplexing developmental tableau, and provide quantitative predictions both across linguistic levels (gradual overlapping pattern), and cultural or individual variations in input (resilience).

3 Standard approaches to language development

It is impossible in limited space to do justice to the rich and diverse sets of viewpoints that have been proposed to account for language development. Instead, the next sections will present a non exhaustive selection of four research strands which draw their source of inspiration from a mixture of psycholinguistics, formal linguistics and computer science, and which share some of the explanatory goals of the reverse engineering approach. The argument will be that even though these strands provide important insights into the acquisition process, they still fall short of accounting for the two puzzles presented in Section ??.

3.1 Psycholinguistics: Conceptual frameworks

Within developmental psycholinguistics, *conceptual frameworks* have been proposed to account for key aspects of the bootstrapping problem and developmental trajectories (see Table ?? for a non exhaustive sample).

Specifically addressing the bootstrapping problem, some frameworks build on systematic correlations between linguistic levels, e.g., between syntactic and semantic categories (syntactic bootstrapping: ?, ?; semantic bootstrapping: ?, ?, ?), or between prosodic boundaries and syntactic ones (prosodic bootstrapping: ?, ?, ?). Others endorse Chomsky's (?) hypothesis that infants are equipped with an innate Language Acquisition Device which constrains the hypothesis space of the learner, enabling acquisition in the presence of sparse or ambiguous input (?, ?, ?).

Other conceptual frameworks focus on key aspects of developmental trajectories (patterns across ages, across languages, across individuals), offering overarching architectures or scenarios that integrate many empirical results. Among others, the competition model: ?, ?, ? ; WRAPSA: ?, ?; the emergentist coalition model: ?, ?; PRIMIR: ?, ?; usage-based theory: ?, ?. Each of these frameworks propose a collection of mechanisms linked to the linguistic input and/or the social environment of the infant to account for developmental trajectories.

²Even in later acquisitions, the nature, universality and effectiveness of corrective feedback of children's outputs has been debated (see ?, ?, ?, ?, ?, ?, ?).

³A successful system may not necessarily have the same architecture of components as described by theoretical linguists. It just needs to behave as humans do, i.e., pass the same behavioral tests. More on this in section ??.

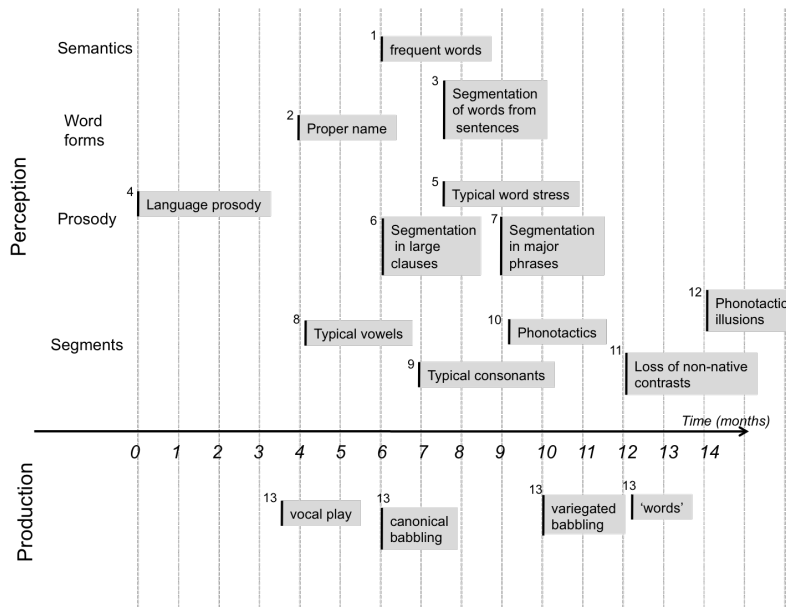


Figure 1. Sample studies illustrating the *time line* of infant's language development. The left edge of each box is aligned to the earliest age at which the result has been documented. ¹ Tincoff & Jusczyk, (1999); Bergelson & Swingle, (2012); ² Mandel et al. (1995); ³ Jusczyk & Aslin (1995) ⁴ Mehler et al. (1988) ⁵ Jusczyk et al. (1999) ⁶ Hirsh-Pasek et al. (1987) ⁷ Jusczyk et al (1992) ⁸ Kuhl et al. (1992) ⁹ Eilers et al. (1979) ¹⁰ Jusczyk et al. (1993) ¹¹ Werker & Tees (1984) ¹² Mazuka et al. (2011) ¹³ Stark (1980).

While these conceptual framework are very useful in summarizing and organizing a vast amount of empirical results, and offer penetrating insights, they are not specific enough to address our two scientific puzzles. They tend to refer to mechanisms using verbal descriptions (*statistical learning*, *rule learning*, *abstraction*, *grammaticalization*, *analogy*) or boxes and arrows diagrams. This type of presentation may be intuitive, but also vague. The same description may correspond to many different computational mechanisms which would yield different predictions. These frameworks are therefore difficult to distinguish from one another empirically, or for the most descriptive ones, impossible to disprove. In addition, because they are not formal, one cannot demonstrate that these models can effectively solve the language bootstrapping problem. Nor do they provide quantitative predictions about the observed resilience in developmental trajectories or their variations as a function of language input at the individual, linguistic or cultural level.

3.2 Psycholinguistics: Artificial language learning

Psycholinguists sometimes supplement conceptual frameworks with propositions for specific *learning mechanisms* which are tested using an artificial language paradigm. As an example, a mechanism based on the tracking of statistical modes in phonetic space has been proposed to underpin phonetic category learning in infancy. It was tested in infants through the presentation of a simplified language (a contin-

uum of syllables between /da/ and /ta/) where the statistical distribution of acoustic tokens was controlled (?, ?). It was also modeled computationally using unsupervised clustering algorithms and tested using simplified corpora or synthetic data (?, ?, ?). A similar double-pronged approach (experimental and modeling evidence) has been conducted for other mechanisms: word segmentation based on transition probability (?, ?, ?), word meaning learning based on cross situational statistics (?, ?, ?, ?), semantic role learning based on syntactic cues (?, ?), etc.

Although studies with artificial languages are useful to discover candidate learning algorithms which could be incorporated in a global architecture, the algorithms proposed have only been tested on toy or artificial languages; there is therefore no guarantee that they would actually work when faced with realistic corpora that are both very large and very noisy (see ??).

3.3 Formal linguistics: learnability studies

Even though much of current theoretical linguistics is devoted to the study of the language competence in the stable state, very interesting work has also been conducted in the area of formal models of *grammar induction*. These models propose algorithms that are provably powerful enough to learn a fragment of grammar given certain assumptions about the input. For instance, ? (?) proposed an algorithm that provided pairs of surface and underlying word forms can

Table 1

Non-exhaustive sample of conceptual frameworks accounting for aspects of early language acquisition. (BP: Bootstrapping Problem; DT: Developmental Trajectories)

Conceptual framework	Reference	Puzzle addressed (age range)	Proposed mechanism
Semantic Bootstrapping	?, ?	BP: syntax (production, 18 mo-4 y)	inductive biases based on syntax/semantic correlations
Syntactic Bootstrapping	?, ?	BP: verb semantics (perception, 18 mo-4 y)	inductive biases on syntax/semantic correlations
Prosodic Bootstrapping	?, ?	BP: word segmentation, part of speech, syntax (perception, 9-18 mo)	inductive biases on prosodic/lexicon/syntax correlations
Knowledge-driven LAD	?, ?	BP: syntax (perception & production, 18 mo -8 y)	perceptual intake; Universal Grammar; inference engine
WRAPSA	?, ?	DT: phonetic categories, word segmentation (perception, 0-12 mo)	auditory processing; syllable segmentation; attentional weighting; pattern extraction; exemplar theory
PRIMIR	?, ?	DT: phonetic, speaker, phonological and semantic categories (perception, 0-2 y)	exemplar theory, statistical clustering, associative learning, attentional dynamic filters
Competition Model	?, ?, ?	DT: syntax (production, 18 mo - 4 y)	competitive learning (avoidance of synonymy)
Usage-Based Theory	?, ?	DT: semantics, syntax (perception & production, 9 mo - 6 y)	construction grammar; intention reading; analogy; competitive learning; distributional analysis

learn the phonological grammar (see also ?, ?). Similar learnability assumptions and results have been obtained for stress systems (?, ?, ?). For learnability results of syntax, see ? (?).

These models establish important learnability results, and in particular, demonstrate that under certain hypotheses, a particular class of grammar is learnable. What they do not demonstrate however is that these hypotheses are met for infants. In particular, most grammar induction studies assume that infants have an error-free, adult-like symbolic representation of linguistic entities (e.g., phonemes, phonological features, grammatical categories, etc). Yet, perception is certainly not error-free, and it is not clear that infants have adult-like symbols, and if they do, how they acquired them.

In other words, even though these models are more advanced than psycholinguistic models in formally addressing the effectiveness of the proposed learning algorithms, it is not clear that they are solving the same bootstrapping problem than the one faced by infants. In addition, they typically lack a connection with empirical data on developmental trajectories.⁴

3.4 Developmental artificial intelligence

The idea of using computational models to shed light on language acquisition is as old as the field of cognitive science itself, and a complete review would be beyond the scope of this paper. We mention some of the landmarks in this field which we refer to as *developmental AI*, separating three learning subproblems: syntax, lexicon, and speech.

Computational models of syntax learning in infants can

be roughly classified into two strands, one that learns from strings of words alone, and one that additionally uses a conceptual representation of the utterance meaning. The first strand is illustrated by ? (?). It views grammar induction as a problem of representing the input corpus with a grammar in the most compact fashion, using both a priori constraints on the shape and complexity of the grammars and a measure of fitness of the grammar to the data (see ?, ? for a probabilistic view). The first systems used artificial input (generated by a context free grammar) and part-of-speech tags (nouns, verbs, etc.) were provided as side-information. Since then, manual tagging has been replaced by automatic tagging using a variety of approaches (see ?, ? for a review), and artificial datasets have been replaced by naturalistic ones (see ?, ?, for a review). The second strand can be traced back to ? (?), and makes the radically different hypothesis that language learning is essentially a translation problem: children are provided with a parallel corpus of speech in an unknown language, and a conceptual representation of the corresponding meaning. The Language Acquisition System (LAS) of ? (?) is a good illustration of this approach. It learns context-free parsers when provided with pairs of representations of meaning (viewed as logical form trees) and sentences (viewed as a string of words, whose meaning are

⁴A particular difficulty of formal models which lack a processing component is to account for the observed discrepancies between the developmental trajectories in perception (e.g. early phonotactic learning in 8-month-olds) and production (slow phonotactic learning in one to 3-year-olds).

known). Since then, algorithms have been proposed to learn directly the meaning of words (e.g., cross-situational learning, see ?, ?), context-free grammars have been replaced by more powerful ones (e.g. probabilistic Combinatorial Categorical Grammar), and sentence meaning has been replaced by sets of candidate meanings with noise (although still generated from linguistic annotations) (e.g., ?, ?). Note that both types of models take textual input, and therefore make the (incorrect) assumption that infants are able to represent their input in terms of an error-free segmented string of words.

Computational models of word discovery tackle the problem of segmenting a continuous stream of phonemes into word-like units. One idea is to use distributional properties that distinguish within word and between word phoneme sequences (?, ?, ?, ?). A second idea is to simultaneously build a lexicon and segment sentences into words (?, ?, ?, ?). These ideas are now frequently combined (?, ?, ?). In addition, segmentation models have been augmented by jointly learning the lexicon and morphological decomposition (?, ?, ?), or tackling phonological variation through the use of a noisy channel model (?, ?). Note that all of these studies assume that speech is represented as an error-free string of adult-like phonemes, an assumption which cannot apply to early language learners.

Finally, a few computational model have started to address language learning from raw speech. These have either concerned the discovery of phoneme-sized units, the discovery of words, or both. Several ideas have been proposed to discover phonemes from the speech signal (self organizing maps: ?, ?; clustering: ?, ?; auto-encoders: ?, ?; HMMs: ?, ?; etc.). Regarding words, ? (?) proposed a model that learn both to segment continuous speech into words and map them to visual categories (through cross situational learning). This was one of the first models to work from a real speech corpus (parents interacting with their infants in a semi-directed fashion), although the model used the output of a supervised phoneme recognizer. The ACORNS project (?, ?) used raw speech as input to discover candidate words (?, ?, see also ?, ?, ?, etc.), or to learn word-meaning associations (see a review in ?, ?, and a comprehensive model in ?, ?), although the speech was collected in the laboratory, not in real life situations.

In sum, developmental AI represents the clearest attempt so far of addressing the full bootstrapping problem. Yet, although one can see a clear progression, from simple models and toy examples, towards more integrative algorithms and more realistic datasets, there is still a large gap between models that learn from speech, which are limited to the discovery or phonemes and word forms, and models that learn syntax and semantics, which only work from textual input. Until this gap is closed, it is not clear how the bootstrapping problem as faced by infants can be solved. The research itself is unfortunately scattered in disjoint segments of the literature,

with little sharing in algorithms, evaluation methods and corpora, making it difficult to compare the merits of the different ideas and register progress. Finally, even though most of these studies mention infants as a source of inspiration of the models, they seldom attempt to account for developmental trajectories.

3.5 Summing up

Psycholinguistic conceptual frameworks capture important insights about language development but are not specified enough to demonstrably solve the bootstrapping problem nor can they make quantitative predictions. Artificial language experiments yield interesting learning mechanisms aimed at explaining experimental data but not necessarily to scale up to larger or more noisy data. These limitations call for the need to develop *effective computational models* that work at scale. Both linguistic models and developmental AI attempt to effectively address the bootstrapping problem, but make unrealistic assumptions with respect to the input data (linguistic models take only symbolic input data, and most developmental AI models take either symbolic data or simplified inputs). As a result, these models may address a different bootstrapping problem than the one faced by infants. This would call for the need to use *realistic data* as input for models. Both linguistic models and developmental AI models take as their gold standard description of the stable state in adults. This may be fine when the objective is to explain ultimate attainment (the bootstrapping problem), but does not enable to connect with learning trajectory data. This would call for a direct *human-machine comparison*, at all ages.

Obviously, the four reviewed research traditions have limits but also address part of the language development puzzles (Table ??). Before examining how the reverse engineering approach could combine the best of these traditions, we examine next with more scrutiny the requirements they have to meet in order to fully address these puzzles.

4 The three requirements of the reverse engineering approach

Here, we argue that to be of scientific import, models of development should (1) go beyond conceptual and box-and-arrow frameworks and be turned into effective, scalable computational systems, (2) go beyond toy data and be fed with realistic input, and (3) be evaluated through human/machine comparisons.

4.1 Why scalable computational models?

Scalable computational systems can provide a proof of principle that the bootstrapping problem can be solved, and generate quantitative predictions. But there is an even more compelling reason to strive for them: verbal reasoning and toy

Table 2
Summary of different theoretical approaches to the study of early language acquisition.

	Effective Model	Realistic Data	Human/Model Comparison
Conceptual Frameworks	No (verbal)	Yes	No (verbal)
Artificial Language Learning	Yes (but not scalable)	No	Yes
Formal Linguistics	Existence proof	Idealized	In the limit
Developmental AI	Yes	Simplified	Qualitative / In the limit
Reverse Engineering	Yes	Yes	Yes

models tend to badly misjudge how a combination of contradictory tendencies will play out in practice, resulting in sometimes spectacularly incorrect predictions. We illustrate this with three examples.

'Easy' problems proving difficult.

How do infant learn phonemes? A popular hypothesis ('*distributional learning*') states that they track the statistical modes of speech sounds to construct phonetic categories (?, ?). How do we turn such a verbal description into a scalable algorithm?

? (?) and ? (?), among others, have proposed that it can be done with *unsupervised clustering* algorithms. As it turns out, these algorithms were only validated only on toy data (points in formant space generated from a Gaussian distribution) or on manually obtained measurements. This is a problem because many if not most clustering algorithms are sensitive to data size, variability and dimensionality (?, ?). When tested on continuous audio representations which are large, variable and of high dimension, very different result ensue. For instance, ? (?) have shown that a clustering algorithm based on Hidden Markov Models and Gaussian mixtures does not converge on phonetic segments, but rather, on much shorter (30 ms), highly context-sensitive acoustic clusters (see also ?, ?). This is not surprising given that phonemes are not realized as discrete acoustic events but as complicated overlapping gestures. For instance, a stop consonant surfaces as a burst, a closure, and formant transitions into the next segment.

This shows that contrary to the distributional learning hypothesis, finding phonetic units is not only a problem of *clustering*, it is also includes *continuous speech segmentation* and *contextual modeling*. These problems are not independent and have therefore to be addressed jointly by the learning algorithms. Despite the optimistic conclusions of ? (?) and ? (?), the unsupervised discovery of phonetic categories is still an unsolved problem in speech technology (see ?, ?, ?).

'Impossible' approaches turning out feasible. The second example relates to the popular hypothesis that acquiring the meaning of words is essentially a problem of associating word form to referents in the outside world (or to conceptual representations of these referents; see ?, ? for possible learning mechanisms).

Under such a view, it would seem impossible to learn any word meaning from language input only. However, research in natural language processing has shown that it is in fact possible to derive an approximate representation of the word meanings using only cooccurrence patterns within the verbal material itself. These distributional techniques (?, ?, ?) construct vector representation of word meanings which correlate surprisingly well with human semantic similarity judgments (?, ?, ?)⁵. ? (?) found that it is possible to derive such vectors even without any properly segmented lexicon, and even without adult-like phonetic categories. It turns out that the approximate meaning representation so derived can provide top-down feedback helping clustering phonetic information into phonemes. Thus, computational systems can suggest a priori implausible, but potentially effective, mechanisms. The empirical validity of such mechanisms in infants remains to be tested.

Statistically significant effects ending up unimportant.

A third example relates to the so-called '*hyperspeech hypothesis*'. It has been proposed that parents adapt their pattern of speech to infants in order to facilitate perception (?, ?). ? (?) observed that parents tend to increase the separation between point vowels in child directed speech, possibly making them easier to learn. Yet, ? (?) ran a word discovery algorithm borrowed from developmental AI on raw speech and failed to find any difference in word learning between child and adult directed speech; if anything, the former was slightly more difficult. This paradoxical result can be explained by the fact that in child directed speech, parents increase phonetic variability even more than they increase the separation between point vowels, the two effects not only cancel each other out, but even result in a small net *degradation* in category discriminability (?, ?; see also ?, ?, ?). The lesson is that it is only through a completely explicit model that the quantitative effect of linguistic and phonetic variables on learning can be assessed.

⁵Interestingly, text-based distributional semantic tend to predict human semantic similarity judgments better than image-based representations (?, ?, ?, ?).

4.2 Why using realistic data?

We turn here to the most controversial of the three requirements: the idea that one should address language learning in its full complexity by running computational models on inputs that are as close as infants' sensory signals as possible.

This may seem an exaggeration. Simplification is the hallmark of the scientific method, which usually proceeds by breaking down complicated problems into smaller, more manageable ones. Here, we claim that an exception has to be made for language learnability. Why? In a nutshell: learning is a process whose outcome is exquisitely sensitive to details of the input signal. If one makes even slightly incorrect assumptions about the input of the learning process, one ends up studying a different learning problem altogether. We illustrate this with three cases where simplifications is a learnability game changer. We conclude that since the learnability-relevant properties of infant's input are currently unknown, the only possibility left is to go with the real thing.

Data selection matters. The entire set of sensory stimulations available to the child is called the input. The subset of this input which is used to learn about the target language(s) is called the intake. The difference between input and intake defines a *data selection problem* which, we claim, is an important part of the learning problem itself. Unfortunately, many computational models of language acquisition short-circuit the selection problem and use human experts to prepare pre-selected and pre-cleaned data. We illustrate this with three data selection problems.

The first problem relates to defining what counts as linguistic versus non-linguistic information. There is no language-universal answer to this question. For instance, gestures are typically para- or extra-linguistic in communities using oral communication (?, ?, ?), but they are the main vehicle for language in sign language (?, ?) which is learned by children in deaf or mixed hearing/deaf communities (?, ?). Within the auditory modality, some vocal sounds like clicks are considered as non-linguistic in many languages, but in others they are used phonologically (?, ?); similarly for phonatory characteristics of vowels like breathiness and creakiness (?, ?, ?).

The second problem is that even if linguistic and non-linguistic signals are defined for a language, the actual mixing of these signals may be difficult. For instance, infants hear a superposition of many audio sources, only some of which contain linguistic signals. Auditory source separation is a computationally difficult problem (untractable in general). In human adults, it is influenced by top-down word recognition (e.g. ?, ?). In pre-verbal infants such sources of top-down information have themselves to be learned.

The third problem is that even if non-linguistic signals are separated from linguistic ones, what to do with non-linguistic signals? In most instances, they should be considered as noise and discarded. In other cases, however, they can be

useful for language learning. For instance, non-linguistic contextually relevant information in the form of visually perceived objects or scenes may help lexical learning (?, ?) or bootstrap syntactic learning (the semantic bootstrapping hypothesis, see ?, ?). Social signals (eye gaze, touch, etc), have also been taken as crucial for language learning (?, ?, ?, among others). Here again, the proper channeling of these non-linguistic cues is part of the learning problem.

In brief, data selection is a critical component of a learning problem. It should *not* be performed by the modeler, who has inside information about the target language and culture, but by the model, whose task is precisely to discover it.

Variability and ambiguity matter. Assuming data selection is solved, ambiguity and variability are prevalent properties of language, at all level of structure, from phonetics up to semantics and pragmatics. Yet, many modeling approaches simplify this complexity by replacing real input with synthetic or idealized data. Although doing so is a useful practice to debug algorithms or prove mathematical results, generalizing from the simplified to real input is risky business.

We already discussed how clustering algorithms that discover phonetic categories when run on synthetic or simplified phonetic data yield much totally different results when run on speech signals. One level up, word segmentation algorithms that recover word boundaries when fed with (errorless) phoneme transcriptions (?, ?) utterly fail when run on speech signals (?, ?, ?). The problem is pervasive. Learning algorithms work because they incorporate models of the shape of the data to be learned. Mismatches between the models and the data will likely result in a learning failure.

Vice versa, however, oversimplifying the input can make the learning problem harder than it is in reality. As an example, syntax learning models often operate from abstract transcriptions, and as a result ignore prosodic information which could prove useful for the purpose of syntactic analysis, or lexical acquisition (e.g. ?, ?, ?, ?).

'Presentation' matters. The notion of 'presentation' comes from formal learning theory (?, ?). It corresponds to the particular way or order in which a parent selects his or her language inputs to the child. There are well known examples where presentation has extreme consequences on what can be learned or not. For instance, if there are no constraints on the order in which environment presents grammatical sentences, then even simple classes of grammars (e.g., finite state or context free grammars, ?, ?) are unlearnable. In contrast, if the environment presents sentences according to a computable process (an apparently innocuous requirement), then even the most complex classes of grammars (recursive grammars) become learnable.⁶ This result extends to a probabilis-

⁶The problem of unrestricted presentations is that, for each learner, there always exists an adversarial environment that will trick the learner into converging on the wrong grammar. Vice versa,

tic scenario where the input sentences are sampled according to a statistical distribution (see ?, ?).

The importance of presentation boils down to the question of whether parents are being 'pedagogical' or not, i.e., whether they present language according to a *curriculum* which facilitates learning⁷. Importantly, such curriculum may also include phonetic aspects (e.g. articulation parameters: ?, ?), para-linguistic aspects (e.g., communicative gestures or touch: ?, ?, ?), as well as the communication context (e.g., availability of a perceptible reference: ?, ?, ?).

To the extent that presentation matters, it is of crucial importance neither to oversimplify by assuming that parents are always pedagogical, nor to overcomplexify by assuming that there is no difference with adult-directed observations.

How realistic does it need to be? We discussed three ways in which the specifics of the input available to the learner matter greatly as to which models will succeed or fail. If one is interested in modeling infant language learning, one should therefore use inputs that are close to what infants get. How to proceed in practice?

One possible strategy would be to start simple, i.e., to work with idealized inputs generated by simple formal grammars or probabilistic models and to incrementally make them more complex and closer to real data. While this approach, pursued by formal learning theory has its merits, it faces the challenge that there is currently no known model of the variability of linguistic inputs, especially at the level of phonetics. Similarly, there is no agreed upon way of characterizing what constitutes a linguistic signal (as opposed to a non-linguistic one), nor what constitutes noise versus useful information. The particular presentation of the target language and associated contextual information that result from caretaker's communicative and pedagogic intentions has not been formally characterized. Even at the level of the syntax, the range of possible languages is not completely known, although this is perhaps the area where there are current propositions (e.g., ?, ?). This approach therefore runs the risk of locking researchers in a bubble universe where problems are mathematically tractable but are unrelated to that faced by infants in the real world.

A second strategy is more radical: use actual raw data to reconstruct infant's sensory experience. This data-driven solution is what we advocate in the reverse engineering approach: it forces to confront squarely the problem of data selection and removes the problems associated with the idealization of variability, ambiguity and mode of presentation. Importantly, the input data should not be limited to a single dataset: what we want to reverse engineer is infant's ability to learn from any mode of presentation, in any possible human language, in any modality. One practical way to address this would be to *sample* from the finite although ever evolving set of attested languages, and split them into development set (to construct the algorithm) and test set (to validate it). It may be

interesting to sample typologies and sociolinguistic groups in a stratified fashion to avoid overfitting the learning model to the prevalent types.

How far should sensory reconstruction go? Obviously, it would make no sense to reconstruct stimuli outside of the sensory range of infants, or with a precision superior to their discrimination abilities. Hence, input simplifications can be done according to known properties of the sensory and attentional capacities of infants. If other idealizing assumptions have to be made, at the very least, they should be explicit, and their impact on the potential oversimplification or overcomplexification of the learning problem should be discussed (as an example, see Sections ?? and ??).

4.3 Why human-machine comparisons?

We now turn to the apparently least controversial requirement: everybody agrees that for a modeling enterprise of any sort, a success criterion should be specified. However, there is little agreement on which criterion to use.

Which success criterion? To quote a few proposals within cognitive psychology, ? (?) proposed nine criteria, ? (?) nine other criteria, ? (?) six criteria, ? (?) three criteria, ? (?) two criteria. These can be sorted into conditions about effective modeling (being able to generate a prediction), about the input (being as realistic as possible), about the end product of learning (being adult-like), about the learning trajectories, and about the plausibility of the computational mechanisms proposed. For formal learning theorists, success is usually defined in terms of *learnability in the limit* (?, ?): a learner is said to learn a target grammar in the limit, if after a finite amount of time, his own grammar becomes equivalent to the target grammar. This definition may be difficult to apply because it does not specify an upper bound in amount of time or quantity of input required for learning (it could take a million years, see ?, ?), nor does it specify an operational procedure for deciding when and how two grammars are equivalent⁸. More pragmatically, researchers in the AI/machine learning area define success in terms of the performance of their system as measured against a gold standard obtained from human adults. This may be an interesting procedure

as computable processes can be enumerated, and hence a stupid learner can test increasingly many grammars and presentations and converge.

⁷Parents may not be conscious of what they are doing: they could adjust their speech according to what they think infants hear or understand, imitate their speech, etc. By pedagogical we refer to the result, not the intent.

⁸Two grammars are said to be (weakly) equivalent if they generate the same utterances. In the case of context free grammars, this is an undecidable problem. More generally, for many learning algorithms (e.g., neural networks), it is not clear what grammar has been learned, and therefore the success criterion cannot be applied.

for testing the end-state of learning but is of little use for measuring learning trajectories.

We propose to replace all these criteria by a single operational principle, *cognitive indistinguishability* defined in terms of cognitive tests:

A human and a machine are cognitively indistinguishable with respect to a given set of cognitive tests when they yield numerically overlapping results when ran on these tests.

Now, this definition is not sufficient in itself: it shifts the problem of selecting a good success criterion to the problem of selecting the tests to be included in the cognitive benchmark. At least, it enables to get rid of arbitrary or aesthetic criteria (I like this model because it seems plausible, or, it uses neurons) and forces one to define operational tests to compare models. Yet, it leaves open a number of questions: should the tests measure behavioral choices, reaction times, physiological responses, brain responses? Should they include meta- or paralinguistic tests (like the ability to detect accent, emotions, etc.)? In addition, given the range of theoretical options that have been formulated on language development (e.g., ?, ?, ?), and disagreements on the essential properties of language (e.g., ?, ?, ?), one would think our proposed cognitive benchmark will be difficult to come about.

How to construct a cognitive benchmark? The benchmark that we propose to construct within the reverse engineering approach has a very specific purpose. Its aim is not to tease apart competing views of language acquisition, but to target the two developmental puzzles presented in Section 2: how do infant bootstrap onto an adult language system? how are gradual, overlapping and resilient patterns of development possible?

Answering these puzzles requires only to measure the state of linguistic knowledge present in the learner at any given point in development, and across the different linguistic structures (phonetic all the way to semantics and pragmatics).

This objective can be expressed in terms of the top level of Marr's hierarchy: the computational/informational level. It abstracts away from considerations about processing or neural implementation. This means that under such benchmark, will be considered 'cognitively indistinguishable', models of the child that have little similarity to infants psychological or brain processes (e.g. Bayesian ideal learners, artificial neural networks), so long as they have acquired the same language-specific information. Of course, one could enrich the benchmark by adding more tests that address lower levels of Marr's hierarchy (see Supplementary Section ?? for a discussion of biological plausibility).

In addition, we propose to guide the construction of the benchmark by selecting tests that

satisfy three conditions: they should be *valid* (measure the construct under study as opposed to something else), *reliable* (with a good signal to noise ratio), and *administrable* (to adults, children and computers alike).

The first two conditions are standard best practices in psychometrics and psychophysics (e.g., ?, ?). Test validity refers to whether a test, both theoretically and empirically, is sensitive to the psychological construct (state or process) it is supposed to measure. As a counterexample, the famous imitation game ? (?) tests whether machines can 'think' by measuring how well they can appear to be humans in an on-line text-based interaction.

This test has dubious theoretical validity, as 'thinking' is not a well defined cognitive construct, but rather an underspecified folk psychology concept, and dubious empirical validity, as it is easy to fool human observers using simplistic text manipulation rules (see ELIZA, ?, ?). Section ?? presents Turing test replacements.

Test reliability refers to the signal to noise ratio of the measure. It can be estimated by computing the between-human or test-retest agreement, or by sampling over initial parameters for the machines.

Test administrability does not belong to standard psychometrics, but very important for comparing the performance of different systems or organisms.

To test a human adult with most tasks, one simply provides instructions in his or her native language.

This is not directly to human infants nor to machines. In infants, a testing apparatus has to be constructed, i.e., a controlled artificial environment whereby responses to test stimuli are measured using spontaneous tendencies of the participants (preference methods, habituation methods, etc; see ?, ?, for a review).⁹ As for machines, the learning algorithms are not constructed to run linguistic tests, but to optimize a particular function which may have nothing to do with the test. Therefore, they need to be supplemented with particular *task interfaces* for each of the proposed tests in order to extract a response that would be equivalent to the response generated by humans.¹⁰ In all cases, administering the task should not compromise the test's validity. Biases or knowledge of the desired response has to be removed from the instructions (adults), testing apparatus (infants) and interface (machines).

In brief, we motivated the importance of a human-machine benchmark and presented principles to construct it. The construction of the benchmark should be viewed as part of the research program itself. It should seek a common

⁹In animals, before tests can be run, an extensive period of training using reinforcement learning is often necessary, in order for the animal to comply with the protocol. Such procedures are not possible in human infants.

¹⁰A task interface can be viewed as a function which takes as input the internal states of the algorithm generated by the stimuli and delivers a binary or real valued response.

ground between competing views of language acquisition, and be periodically revised as understanding of the language competence progresses, and as new experimental protocols for language competence are established.

5 Deep learning to the rescue?

The combination of the first two requirements discussed above, namely, scalable computation and realistic input would have been, up to a recent period a major stumbling block for achieving a reverse engineering approach. Indeed, for many years, computers were struggling with language processing. It was customary in psycholinguistic courses to mock the dismal performance of automatic dictation or translation systems. All of this started to change with a paper by Hinton and colleagues on speech recognition (1, 2): after years in the making, neural networks were starting to perform better than the dominating technology based on probabilistic models (Gaussian Mixtures, Hidden Markov Model). A few years later, the entire speech processing pipeline has been replaced by neural networks trained end-to-end, with performance claimed to achieve human parity on a dictation task (3, 4, but see 5, 6). In the following, we very briefly review how such systems are constructed before turning on whether they could be used to inform infants language acquisition studies.

5.1 The new AI spring

One important characteristics of the new systems is they get rid of the specialized design features of their predecessors, and replace them with generic neural network architectures trained in large annotated corpora. Continuing with the example of speech, specialized audio features are replaced by spectrograms (some systems even work from raw audio input) and phonetic transcriptions and pronunciation lexicons are eliminated: systems are trained to directly map speech to orthographic transcriptions, in an end-to-end fashion.

We do not need a phoneme dictionary, nor even the concept of a 'phoneme.' (7, 8).

As it turn out, the basic architectures and many core ideas are not very different from those proposed in the early days of connectionism. For instance, Figure ?? shows the architecture of Deep Speech 2 (9, 10) a state-of-the-art speech recognition system composed of rather classical elements popularized in the late 80's the (the multi-layer perceptron, back-propagation training, convolutional networks, recurrent networks: Rumelhart & McClelland, 1986; 11, 12).

What has changed, though, is the scale of the networks and the volume of data on which they are trained, enabled by tremendous progress in computer hardware and in mathematical optimization techniques (13, 14, for an advanced introduction). As a result, neural networks have grown at a

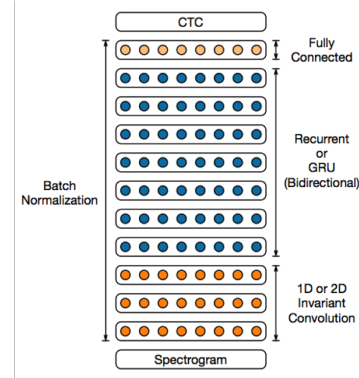


Figure 2. Network architecture for Deep Speech 2 (reprinted from Amodei et al. 2016). The input is a spectrogram, the output is a sequence of characters. Each layer incorporate particular patterns of connectivity. Convolutional layers are organized in terms of local patches sharing their connections along the time and/or frequency dimensions. Recurrent layers accumulate activations through time. Fully connected layers do not have any particular topology. Batch normalization is a process by which the activations at each layers are rescaled to be of means 0 and variance 1 over a small set of examples during training.

pace slightly faster than Moore's law: the speech processing network in 1 (10) had 8000 parameters; 28 years later, Deep Speech 2 is twelve thousand times larger.

Speech is not the only area where deep learning have shaken the AI landscape: object recognition (15, 16, 17), language translation (18, 19, 20), and speech synthesis (21, 22), are all areas where neural networks have displaced by a large margin the previous state-of-the-art, while approaching human performance. This explosion of research is facilitated by the large distribution of programming frameworks (tensorflow, pytorch, dynet, mxnet, etc.), the open sourcing of datasets and state-of-the-art systems which can be downloaded pre-trained and tested on new inputs.

These successes are generating interest for taking machine learning systems trained on large corpora as quantitative models of cognitive functions. Indeed, despite their a-priori lack of neural or biological plausibility¹¹, the performance of these systems show surprising convergences with biological organisms. For instance, a deep neural network trained to recognize artefacts and natural kind categories from images turn out to be good predictors of multi-unit responses of neurons in the Inferior Temporal cortex of primates (e.g., 23, 24, 25). There are also surprising divergences such as the strange way in which neural networks can be fooled by adversarial examples (26, 27), and limits such as their inability to perform

¹¹In fact, from their inception, neural networks have been heavily influenced by research in neuroscience and psychology, see the review by 28 (?).

causal reasoning or display systematic behavior (?, ?). This gives rise to an exciting area of research applying cognitive psychology or cognitive neuroscience methods to machine learning systems (?, ?, ?, ?, ?, among others).

The crucial question that we raise here is whether any of these algorithms could be good candidates for modeling language acquisition?

5.2 Does machine learning model human learning?

Statistical learning mechanisms have been claimed to be at the core of language acquisition (?, ?), so a-priori, there are reasons to be optimistic (?, ?). However, there is a fundamental gap between what this term means in cognitive studies and how it is used in machine learning. The big difference is that in machine learning, statistical techniques are used as a convenient way to construct systems, not as models of human acquisition processes.

Interpreted cognitively, machine learning procedures would correspond to a caricature of 19th century schooling: the learner, initially, a kind of tabula rasa, is relentlessly fed with inputs paired with desired responses, which are annotations of the input provided by a human supervisor. The drill is repeated until the learner gets it right.

This setup is called *supervised learning*, because for a given input there is only one correct answer. As an example, in speech recognition, the system is trained to associate a speech utterance with its written transcription. In natural language processing tasks, the system is presented with sequences of words (in orthographic format) as input, and trained to associate each word to a part-of-speech, a semantic role tag, or a co-reference in the text, and so on. This differs in how infants learn language in two important ways.

First, children do not learn their first language by being asked to associate sensory inputs with linguistic tags. Long before they are even exposed to linguistic tags by going to school and learn to read and write, they have acquired what amounts to a fully functional speech recognition and language processing system. They have done so on the basis of sensory input alone, and if there are supervisory signals from the adults, these are neither unambiguous nor systematic. This moves the problem of language learning in the area of *unsupervised* or *weakly supervised* machine learning: to an input, there is no unique desired output, but rather a probabilistic distribution of outcomes (with relatively unfrequent rewards or punishments)¹².

The second difference is in the sheer amount of data required by artificial systems compared to infants. For instance, the Deep Speech 2 system described above is trained with over 10000 hours of transcribed speech (plus a few billion words worth of text to provide top-down language statistics). In comparison, a four-year-old child, who admittedly has functional speech recognition abilities, is being spoken to for a total amount varying between 700h and 4000h (cor-

responding to 8 and 44M words, respectively), depending on the language community (for estimates, see Supplementary Section ??). This means that Deep Speech 2 requires around 14 times more speech, and 240 times more words than what a four-year old Mayan child get. A recent time allocation study in the Tsimane community (?, ?) shows that the amount of child directed input may even be lower than the Maya yet by a factor of 3 (less than one minute of speech per waking hour). This shows that the human infant is equipped with a learning algorithm which enables him or her to learn language with very scarce data.

5.3 Summing up.

Machine learning has made progress to the point that 'cognitive services' (speech recognition, automatic translation, object and face recognition, etc.) are incorporated in everyday life applications. This means that one of the major road block for the reverse engineering approach, i.e. the feasibility of building language processing systems that can deal with realistic input at scale is now lifted. Instead of being locked with simplified data or toy problems, for the first time, it becomes possible to address the bootstrapping problem in its full complexity, and derive quantitative developmental predictions along the way.

Still, there are challenges ahead; current machine learning systems fail to provide models of infant acquisition, not because they discard or simplify the input, but because they use too much of it, both in sheer quantity and in adding extra inputs that the infant could not possibly get (linguistic labels). What needs to be done, therefore is to adapt some of the existing algorithms or construct new ones, so that they can learn with as few data as infants do. How far are we?

6 The road ahead

We now turn to the feasibility of the reverse engineering approach as applied to early language development. To do so, we limit ourselves to the following simplifying assumption:

The total input available to a particular child provides enough information to acquire the grammar of the language present in the environment.

This may seem reasonable, but it essentially puts us in the open loop situation described in Figure ??), where the environment delivers a fixed curriculum of inputs (utterances and their sensory contexts) and the learner recovers the grammar that generated the utterances. In this situation, the output

¹²This raises the issue about what is the internal reward for the infant which pushes him or her to acquire language. A drive for learning statistical patterns? A drive to interact with others in his or her group?

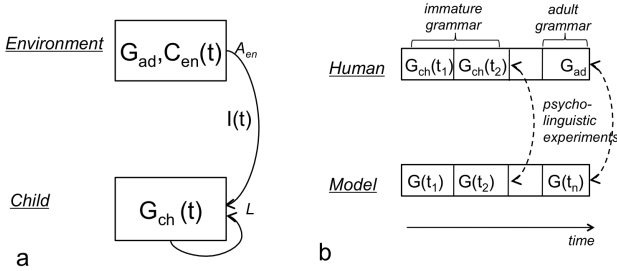


Figure 3. a. The (simplified) learning scenario: The Child’s internal state is a grammar $G_{ch}(t)$ that can be updated through the learning function L based on input $I(t)$. The environment’s internal state is a constant adult grammar G_{ad} and a variable context C_{en} , which produces the input to the child. b. Method to test the empirical adequacy of the model by comparing the outcome of psycholinguistic experiments with that of children and adults.

of the child is not modeled, and the environment does not modify its inputs according to her behavior or inferred internal states. This input-driven idealization may overestimate the difficulty of the task compared to a more realistic close-loop scenario. We think however, that it is useful to study the input-driven scenario in its own sake, as it gives an estimate of what can be learned in the worse case scenario where parents have other priorities than optimizing their children’s language learning.

We examine how this simplifying assumption can be relaxed in Supplementary Section ??

Within this scenario, we claim that recent advances in AI and big data now make the reverse engineering roadmap actionnable. We discuss current avenues of research and the challenges that need to be met. Following our three requirements, we review, in turn, the feasibility of constructing systems that can learn without expert labels, the collection of large realistic dataset, and the establishment of human-machine benchmarks, and illustrate it with a selection of recent work.

6.1 Unsupervised / weakly supervised algorithms

Bringing machine learning to bear to language development requires to construct systems that discover linguistic structure with little or no expert supervision. This is obviously more difficult than learning to associate inputs to linguistic labels. Here, the learner has to discover its own labels given the input. This class of machine learning problems is unfortunately less well studied and understood than supervised learning, but is an expanding field of research in machine learning. Two main, non exclusive, ideas are being explored to address this challenge.

Inductive biases. The first idea is to build into the learner prior knowledge about the underlying nature of the

data, so that generalization can be made with few or noisy datapoints. With strong prior knowledge, some logically impossible learning problems become easily solvable.¹³ Some models of the acquisition of syntax mentioned in Section 3.1 favor very strong priors, where the only thing to learn (besides the meaning of words) is a small number of syntactic binary parameters. The learning problem becomes so constrained that a single sentence (called a trigger) may be sufficient to decide a parameter’s value ($?$, $?$, $?$). The notion of inductive biases can be formulated elegantly using Bayesian graphical models ($?$, $?$, $?$). In these models, prior knowledge is specified as probability distributions over the model’s parameters, which are updated for each new input (see $?$, $?$ for a general presentation).

For the purpose of illustration, let us revisit the discovery phonetic categories from continuous speech. We have mentioned previously that generic clustering algorithms fail to learn phonemes, because of a mismatch between what clustering algorithms expect (relatively well delimited clusters) and what the data consists in (a complicated gesture unfolding in time). $?$ ($?$) proposed a Bayesian graphical model, where phonemes are defined as sequences of three acoustic states (schematically, a state for the beginning, the central part and the end of the phoneme). Each state is modeled as a mixture of 8 Gaussians in the space of acoustic parameters (MFCCs, a representation derived from spectrograms). Phoneme durations are also controlled through a binary boundary variable (modelled with a poisson distribution), and the number of phonemes is specified by a Dirichlet prior, which expects the distribution of phonemes to follow a power law (a few phonemes are used often, many phonemes are used rarely). Far from being a general purpose clustering algorithm, the algorithm of Lee & Glass uses language-universal information about the phonemes, (their shape, their duration, their frequency) to specify a model that will be inductively biased to discover this kind of structure in the data. Bayesian probabilistic models are also used in natural language processing to infer syntactic structures from raw data without supervision ($?$, $?$, $?$). Some of these models have been recently used on child directed input (CHILDES transcripts) to account for developmental results ($?$, $?$).

The challenge with these types of models is that the optimization of the parameters is very computationally intensive, which becomes prohibitive for large models and/or large datasets. For instance, the Lee & Glass model has only been applied to a relatively small corpus of read speech (TIMIT), and the $?$ ($?$) model on textual input. Current research is devoted to develop efficient approximations of these algorithms

¹³One good illustration is the following: can you tell the colors of 1000 balls in an urn by just selecting one ball? The task is impossible without any prior knowledge about the distribution of colors in the urn, but very easy if you know that all the balls have the same color.

to deploy them in more naturalistic datasets (see for instance ?, ? for a scalable reimplementation of Lee & Glass).

Synergies. Here, the idea is that the different components of language being interdependant, it may help to jointly learn these components rather than to learn them separately. This is actually turning the bootstrapping problem on its head: instead of being a liability, the codependancies between linguistic components become an asset. Of course, it is an empirical issue as to whether joint learning between any two language components is always more successful than separate learning. The existence of synergies has been documented using Bayesian models between phonemes and words inventories (?, ?), syllables and words segmentation (?, ?), referential intentions and word meanings (?, ?).

The existence of synergies can be leveraged in models other than Bayesian ones, including deep learning or more algorithmic speech engineering systems. For instance, returning to the issue of phonetic learning, several lines of research indicate that words could help the discovery of subword units (?, ?, ?), and that even an imperfect, automatically discovered proto-lexicon can help (?, ?, ?). The model described in Figure ?? implements this idea. It consists in a word discovery system which extracts similar segments of speech across a large corpus. The discovered segments constitute a proto-lexicon of acoustic word forms (?, ?), which are then used to train a neural network in a discriminative fashion. The resulting output of the network is a representation of speech sound which is much more invariant to a change in talker than the original spectral representation on which the system started with (?, ?). In a similar spirit, ? (?) and ? (?) showed that by training a neural network to associate an image with a speech input corresponding to a short description of this image, the network develops phone-like and word-like intermediate representations for speech.

In brief, even though unsupervised/weakly supervised learning is difficult, there is a growing interest within machine learning for the study of such algorithms, as shown by special sessions on this topic in machine learning conferences, and the organization of challenges involving laboratories in cognitive science and speech technology communities (e.g. the zero resource speech challenge, ?, ?, ?).

6.2 Large scale data collection in the wild

A large number of datasets across languages have been collected and organized into repositories that have proved immensely useful to the research community. One prominent example of this is the CHILDES repository (?, ?), which has enabled more than 5000 research papers (according to a google scholar search as of 2016). These datasets, however, contain only relatively sparse datapoints (a few hours per infants). Perhaps the most ambitious large scale and dense data collection effort to date is the Speechome project (?, ?), where video and audio equipment was installed in each room


Pronunciation	b	a	n	a	n	a					
	[b]	[ax]	[n]	[ae]	[n]	[ax]					
											
Frame index (t)	1	2	3	4	5	6	7	8	9	10	11
Speech feature (x_t^i)	x_1^i	x_2^i	x_3^i	x_4^i	x_5^i	x_6^i	x_7^i	x_8^i	x_9^i	x_{10}^i	x_{11}^i
Boundary variable (b_t^i)	1	0	0	1	0	1	0	1	1	0	1
Boundary index (g_t^i)	g_0^i	g_1^i		g_2^i	g_3^i		g_4^i	g_5^i		g_6^i	
Segment ($p_{j,k}^i$)	$p_{1,1}^i$		$p_{2,4}^i$		$p_{5,6}^i$		$p_{7,8}^i$		$p_{9,9}^i$		$p_{10,11}^i$
Duration ($d_{j,k}^i$)	1	3	2	2	1	2					
Cluster label ($c_{j,k}^i$)	$c_{1,1}^i$		$c_{2,4}^i$		$c_{5,6}^i$		$c_{7,8}^i$		$c_{9,9}^i$		$c_{10,11}^i$
HMM (θ_i)	θ_1		θ_2		θ_3		θ_4		θ_5		θ_6
Hidden state (s_t^i)	1	1	2	3	1	3	1	1	1	3	3
Mixture ID	1	1	6	8	3	7	5	2	8	2	8

Figure 4. Outline of a clustering algorithm with a hierarchical generative architecture for learning phonemes from raw speech (reprinted from Lee, & Glass, 2012). The model is provided with speech described by speech features (3rd line), and infers all of the other parameters (4th line downwards: boundary variables, segment identify and duration, states of the hidden markov model and parameters of the Gaussian mixtures) from the data according to the hierarchical model by sampling in the space of possible values for these parameters.

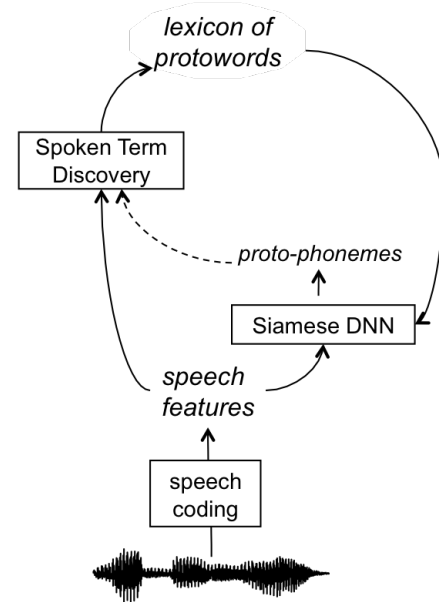


Figure 5. Architecture illustrating a top-down synergy between learning phonemes and words. Auditory spectrograms (speech features) are computed from the raw speech signal. Then, protowords are extracted using Spoken Term Discovery; these words are then used to learn a more invariant speech representation using discriminative learning in a siamese Deep Neural Network architecture (from Thiolliere et al., 2015).

of an apartment, recording 3 years' worth of data around one infant. This pioneering work illustrates several key technological, analysis and ethical issues that arise in 'ecological' data collection.

Regarding technological issues, the falling costs in digital sensors and data storage make it feasible to duplicate Speechome-like projects across many languages. More challenging is the fact that to be usable for modeling, the captured should enable the reconstruction of infant's sensory experience from a first person point of view. Already, relatively inexpensive out-of-the box wearable technology can go some way in that direction. Miniaturized recorders (see for instance the LENA system, ?, ?) enable recording the infant's sound environment for a full day at a time, even outside home, and will become more and more usable as microphone array and advanced signal processing enable source reconstruction even in noisy environment. Proximity and accelerometer sensors can be used to categorize activities (?, ?); 'life logging' wearable devices capture images every few seconds and help to reconstruct the context of speech interactions (?, ?). Head-mounted cameras can help to reconstruct infant's field of view (?, ?). Upcoming progress in the miniaturization of 3D sensors would enable to go further in the reconstruction of infant's visual experience.

Regarding analysis issues, the challenge is to supplement raw data with reliable linguistic/high level annotations. Manual annotations are too costly to scale up to large and dense datasets. In the Speechome corpus, more than 3000 hours of speech have been transcribed, which represents only a fraction of the total 140000 hours of audio recordings (?, ?). The recent breakthroughs in machine learning discussed in Section ?? (speech recognition: ?, ?; object recognition: ?, ?; action recognition: ?, ?; emotion recognition: ?, ?) will enable the semi-automatic annotations of large amounts of data.

As for ethical issues, the main challenge is to find a point of equilibrium between the requirement of sharability and open scientific data, and the need of protecting the privacy of the families' personal data. Up to now, the response of the scientific community has been dichotomous: either make everything public (as in the open access repositories like CHILDES, ?, ?), or completely close off the corpora to anybody outside the institution that has recorded the data (as in the Riken corpus, ?, ?, or the Speechome corpus ?, ?). Neither solutions are acceptable.

Alternative strategies are being considered by the research community. The Homebank repository contains raw and transcribed audio, with a restricted case by case access to researchers (?, ?, <http://homebank.talkbank.org>). Databrary has a similarly organized system for the secure storage of large sets of video recordings of developmental data (?, ?, <https://nyu.databrary.org>). Progress in cryptographic techniques would make it possible to envision preserving privacy while enabling more open exploitation of

the data. For instance, the raw data could be locked on secure servers, thereby remaining accessible and revokable by the infants' families. Researchers' access would be restricted to anonymized meta-data or aggregate results extracted by automatic annotation algorithms. The specifics of such a new type of linguistic data repository would have to be worked out before dense speech and video home recordings can become a mainstream tool for infant research.

In brief, large scale data collection of infant data is within reach and is under in a number of research projects (see www.darcle.org), although its exploitation in an open source format requires specific developments in privacy-preserving storage and computing infrastructures.

6.3 Cognitive benchmarking of language acquisition

Our final requirement, the construction of a cognitive benchmark for language processing, can draw from work in linguistics and psycholinguistics.

One can indeed find relatively easy-to-administer, valid and reliable tests of the main components of linguistic competence in perception/comprehension (see Table ??). These tests are easy to administer because they are conceptually simple and can be administered to naive participants; most of them are of two kinds: goodness judgments (say whether a sequence of sound, a sentence, or a piece of discourse, is 'acceptable', or 'weird') and matching judgments (say whether two words mean the same thing or whether an utterance is true of a given situation, which can be described in language, picture or other means). As for validity, (psycho)linguistic tests often use a *minimal set design* where one linguistic construct is manipulated while every other variable is kept constant (for instance: 'the dog eats the cat' and 'the cat eats the dog' contain the same words, but one sequence is syntactically correct, the other not). Regarding test reliability, as it turns out, many linguistic tests are quite reliable, as 97% of the results of a grammaticality judgment from textbooks are replicable using on-line experiments (?, ?)¹⁴.

Given the simplicity of these tasks, it is relatively straightforward to apply them to machines. Indeed, matching judgments between stimulus A and stimulus B can be derived by extracting from the machine the representations triggered by stimulus A and B, and compute a *similarity score* between these two representations. Goodness judgments are perhaps more tricky; they can easily be done by generative algorithms that assign a *probability score*, a *reconstruction error*, or a *prediction error* to individual stimuli. As seen in Table ??, some of these tests are already being used quite standardly

¹⁴Of course, even in simple psychophysical tasks, humans can be affected by many other factors like attention, fatigue, learning or habituation to stimuli or regularities in stimulus presentations, etc. Methods try to minimize but never totally succeed in neutralizing these effects.

Table 3

Example of tasks that could be used for a Cognitive Benchmark.

Task description in human adults	Linguistic level	Equivalent task in children	Equivalent task in machines
Well-formedness judgement <i>does utterance S sound good?</i>	phonetic, prosody, phonology, morphology, syntax	preferential looking (9-month-olds: ?, ?), acceptability judgment (2-year-olds: de Villiers and de Villiers, 1972; Gleitman, Gleitman, and Shipley, 1972)	reconstruction error (?, ?), probability (?, ?), mean or min log probability (?, ?)
Same-Different judgment <i>is X the same sound / word / meaning as Y?</i>	phonetic, phonology, semantics	habituation / deshabituation (newborns, 4-month-olds: ?, ?, ?), oddball (3-month-olds: ?, ?)	AX/ABX discrimination (?, ?, ?), cosine similarity (?, ?)
Part-Whole judgment <i>is word X part of sentence S?</i>	phonology, morphology	Word spotting (8-month-olds: ?, ?)	spoken web search (?, ?)
Reference judgment <i>does word X (in sent S) refer to meaning M?</i>	semantics, pragmatics	intermodal preferential looking (16-month-olds: ?, ?), picture-word matching (11-month-olds: ?, ?)	picture/video captioning (e.g., ?, ?), Winograd's schemas (?, ?)
Truth/Entailment judgment <i>is sent S true (in context C)?</i>	semantics	Truth Judgment Task (3-year-olds: ?, ?, ?)	visual question answering (?, ?)
Felicity judgement <i>would people say S to mean M (in context C)?</i>	pragmatics	Ternary reward task (5-year-olds: ?, ?), Felicity judgment task (5 years olds: ?, ?).	?

in the evaluation of unsupervised learning systems, in particular, in the evaluation of phonetic and semantic levels while for others they are less widespread.¹⁵

The challenge comes from the applicability of these tests to infants and children. As seen in Table ??, there are considerable variations on the age at which different linguistic levels have been tested in children, and generally, the younger the child, the more difficult it is to construct reliable tests. Addressing this challenge would require improving substantially the signal-to-noise of some of these techniques. There is also the possibility to increase the number of participants through community-augmented meta-analyses (?, ?; see also <http://metalab.stanford.edu/>), collaborative testing (?, ?), or remotely run experiments (Shultz, 2014, <https://lookit.mit.edu/>; Izard, 2016, <https://www.mybabylab.fr>).

Before the full reverse engineering roadmap has been put into place, it is already possible to test specific predictions using existing techniques. One can use the patterns of errors made by computational models when run on infant input data to generate new predictions. The reasoning is that these errors should not be viewed as 'bugs', but rather signatures of intrinsic computational difficulties that may also be faced by infants. For instance, even very good word discovery algorithms make systematic segmentation errors: under-segmentations for frequent pairs of words (like "readit" instead of "read"+"it") or over-segmentations ("butter"+"fly"

instead of "butterfly") (see ?, ?).

? (?) showed that it is possible to use the preferential listening paradigm in eleven month infants to probe for signature mis-segmentations. Deriving predictions from a very simple model of word discovery (an ngram model) run on a CHILDES corpus, she constructed a set of otherwise matched frequent versus unfrequent mis-segmentations. Eleven month olds preferred to listen the frequent mis-segmentations, and did not distinguish them from real words of the same frequency. ? (?) found that it was possible to compare the outcome of different segmentation algorithms in measuring their ability to predict vocabulary acquisition as measured by parental report.

In brief, while a cognitive benchmark can be established, and it is already possible to test in infants some predictions of computational models, large scale model comparison will require progress in developmental experimental methods.

7 Conclusions

During their first years of life, infants learn a vast array of cognitive competences at an amazing speed; studying this development is a major scientific challenge for cognitive science in that it requires the cooperation of a wide variety of

¹⁵Regarding the evaluation of word discovery systems, see the proposition by ? (?) but see ? (?) for a counter proposal and a discussion in ? (?).

approaches and methods. Here, we proposed to add to the existing arsenal of experimental and theoretical methods the reverse engineering approach, which consists in building an effective system that mimics infant's achievements. The idea of constructing an effective system that mimics an object in order to gain more knowledge about that object is of course a very general one, which can be applied beyond language (for instance, in the modeling of the acquisition of naive physics or naive psychology) and even beyond development.

We have defined three methodological requirements for this combined approach to work: constructing a computational system at scale (which implies 'de-supervising' machine learning systems to turn them into models of infant learning), using realistic data as input (which implies setting up sharable and privately safe repositories of dense reconstructions of the sensory experience of many infants), and assessing success by running tests derived from linguistics on both humans and machines (which implies setting up benchmarks of cognitive and linguistic tests). We've showed that even before these challenges are all met, such an approach can help challenging verbal theories, help characterize the learning consequences of different kinds of inputs available to infant across cultures, and suggesting new empirical tests.

Before closing, let us note that the reverse engineering approach we propose does *not* endorse a particular model, theory or view of language acquisition. For instance, it does *not* take a position on the rationalist versus empiricist debate (e.g., Chomsky, 1965, vs. Harman, 1967). Our proposal is more of a methodological one: it specifies what needs to be done such that the machine learning tools can be used to address scientific questions that are relevant for such a debate. It strives at constructing at least one effective model that can learn language. Any such model will both have an initial architecture (nature), and feed on real data (nurture). It is only through the comparison of several such models that it will be possible to assess the *minimal* amount of information that the initial architecture has to have, in order to perform well. Such a comparison would give a quantitative estimate of the number of bits required in the genome to construct this architecture, and therefore the relative weight of these two sources of information. In other words, our roadmap does not start off with a given position on the rationalist/empiricist debate, rather, a position in this debate will be an outcome of this enterprise.

Acknowledgments

This paper would not have come to light without numerous inspiring discussions with Paul Smolensky and Alex Cristia. It also benefitted from insightful comments by Paul Bloom, Emmanuel Chemla, Ewan Dunbar, Michael Frank, Giorgio Magri, Steven Pinker, Thomas Schatz, Gabriel Synnaeve, the members of the Cognitive Machine Learning team of the Laboratoire de Sciences Cognitives et Psy-

chologique, and three anonymous Cognition reviewers. This work was supported by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the Fondation de France, the Ecole de Neurosciences de Paris, and the Region Ile de France (DIM cerveau et pensée).

Appendix: Supplementary Materials

S1. Estimate of input to child

In this section, we describe the way in which we estimated the amount and variability of speech input to infants. We are mainly interested in the number of hours and number of words, since these are two common metrics used in automatic speech recognition and natural language processing. We therefore use these metrics when they were available in the original data, and estimate them otherwise.

Table ?? lists the sample of four studies that we have included in our survey, which incorporate large variations in languages and cultures. ? (? , H&R) studied English speaking infants splitted into three groups according to the Socio-Economic Status (SES) of the family. In our analysis, we only include the two extreme groups (N=13 and 6, respectively). ? (? , S&G) studied two groups, one rural Mayan speaking community (N=6), one English speaking urban community in the USA (N=6). ? (? , W&F) studied one group of low SES Spanish speaking family in the USA (N=29). Finally, ? (? , VdW) extensively measured one Dutch speaking child in the Netherlands.

One methodological problem is that the four studies reported different kinds of metrics (H&R: number of words and utterances, S&G: number of utterances, W&F: number of words, and VdW: number of hours, words and utterances). In order to compare them, one has therefore to estimate how to convert one metric into another, which requires possibly incorred assumptions about the conversion parameters. They should therefore be taken with a large grain of salt, and are subject to revision when more precise data comes along.

Table ?? lists the results and indicate the value of the conversion factor that we used. To compute the total number of hours per year, we used a waking time estimate of 9h for all of the studies except VdW which directly estimated speaking time per day. To convert number for words into hours, we used an estimate of word duration of 400ms. This is compatible with the numbers reported by VdW. To convert between number of utterances and number of words, we used an SES-dependant estimate of Mean Utterance Length of 4.43 for high SES and 3.47 for low SES (from H&R). Finally, to estimate the total amount of speech heard by infants, we used a proportion of Child Directed Input of 64% for high SES (for S&G) and of 62% for low SES (from W&F). To see an updated version of this analysis including a new population

Table S1

Four studies used to estimate infant's speech input

study	reference	mode of acquisition;age	population
H&R	? (?)	observer, 1h every month; 12-36 months	urban high, mid & low SES, English
S&G	? (?)	observer, 1h every month; 12-36 months	urban high SES, English & ru- ral low SES, Maya
W&F	? (?)	daylong recording; 19 months	low SES, Spanish
VdW	? (?)	daylong recording; 6-9 months	high SES, Dutch

Table S2

Estimates of yearly input, in total, and restricted to Child Directed Speech (CDS) , in number of hours and words (millions) per year in four studies (see the references in Table ??) as a function of sociolinguistic group (SES: Socio Economic Status). The numbers between brackets provide the range [min, max] of these numbers across families. ^t uses a wake time estimate of 9 hours per day. ^w uses a word duration estimate of 400ms. ^c uses S&G's estimate of %CDS for high SES. ^d uses W&F's estimate of %CDS for low SES. ^m uses H&R's MLU's estimates (according to SES).

	Yearly total				Yearly CDS			
	Hours		Words (M)		Hours		Words (M)	
Urban, high SES								
H&R (N=13) ^t	1221 ^{w,c}	[578,1987]	11.0 ^c	[5.20, 17.9]	786 ^w	[372, 1279]	7.07	[3.35, 11.5]
S&G (N=6) ^t	2023 ^{w,m}	[1243, 2858]	18.2 ^m	[11.2, 25.7]	1223 ^{w,m}	[853, 1574]	11.0 ^m	[7.7, 14.2]
VdW (N=1)	931		9.28		140		1.39	
Urban, low SES								
H&R (N=6) ^t	363 ^{w,d}	[136, 558]	3.26 ^d	[1.22, 5.02]	225 ^w	[84, 346]	2.02	[0.76., 3.11]
W&F (N=29) ^t	363 ^w	[52, 1049]	3.27	[0.46., 9.44]	225 ^w	[32, 650]	2.03	[0.29, 5.85]
Rural, low SES								
S&G (N=6) ^t	503 ^{w,m}	[365, 640]	4.53 ^m	[3.28, 5.76]	234 ^{w,m}	[132, 322]	2.10 ^m	[1.19, 2.90]

of forager-farmers, see (?, ?).

S2. A biological plausibility requirement?

In this section, we briefly discuss one issue which often comes up when computational systems are used as models of human processing: the issue of *biological plausibility*. By this, we mean that the hypothetical algorithm be compatible with what we know about the biological systems that underlie these computations in human infants/adults.

While this constraint is perfectly reasonable, we argue that it is difficult to apply to the modeling of early language acquisition for the following reasons: First, the computational power of a human brain is currently unknown. Current supercomputers can simulate at a synapse level only a fraction of a brain and several orders of magnitude slower than real time (?, ?). If this is so, all computational models run in 2016 are still massively underpowered compared to a child's brain. Second, a particular algorithm may appear to be too complex for the brain, but a different version performing the same function will not. For instance, some word segmentation algorithms require a procedure called Gibbs sampling,

which, in theory, require an infinite number of time steps to converge. This would seem to discredit the algorithm altogether. Yet, it turns out that a truncated version of this algorithm running in finite time works reasonably well. Similarly, algorithms that require a lot of time steps can be rewritten into algorithms that require less steps and more memory. This makes a priori claims of biological plausibility difficult to make.

Still, biological plausibility can place some theoretical bounds on *system complexity at the initial state*. Indeed, the initial state is constructed on the basis of the human genome plus prenatal interactions with the environment. This allows to rule out, for instance, a 100% nativist acquisition model that would pre-compile a state-of-the-art language understanding systems for all of the existing 6000 or more languages on the planet, plus a mechanism for selecting the most probable one given the input.¹⁶

¹⁶The reason such system would not be biologically realizable is that the parameters of a state-of-the-art phoneme recognition system for a single of these languages already require 10 times more memory storage than what is available in the fraction of the genome that

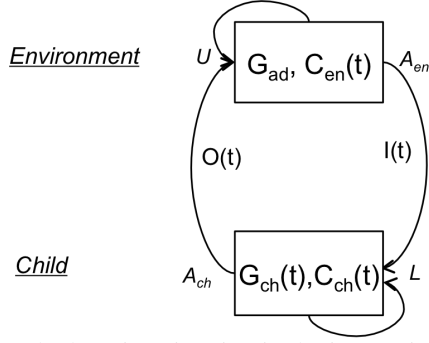


Figure S1. The learning situation in the interactive scenario, viewed as two coupled dynamic systems: the Child and the Environment.

Apart from this rather extreme case, biological plausibility may not affect much of the reverse engineering approach until more is known about the computational capacity of the brain. Yet, it is compatible with our approach, since as soon as diagnostic tests of language computation in the brain are available, they could be added to the cognitive benchmark, as defined in Section ?? (see also Frank, 2014¹⁷).

S3. Can reverse engineering address the fully interactive learning scenario?

In this section, we revisit the simplifying assumptions of the input-driven scenario endorsed in Section ?? and displayed in Figure ??a. This scenario does not take into consideration the child's output, nor the possible feedback loops from the parents based on this output. Many researchers would see this as a major, if not fatal, limitation of the approach. In real learning situations, infants are also agents, and the environment reacts to their outputs creating feedback loops (?, ?, ?, ?, ?, ?).

The most general description of the learning situation is therefore as in Figure ??. Here, the child is able to generate observable actions (some linguistic, some not) that will modify the internal state of the environment (through the monitoring function). The environment is able to generate the input to the child as a function of his internal state. In this most general form, the learning situation consists therefore in two *coupled dynamic systems*.¹⁸

Could such a complex situation be addressed within the reverse engineering approach? We would like to answer with a cautious yes, to the extent that it is possible to adhere to the same three requirements, i.e., realistic data (as opposed to simplified ones), explicit criteria of success (based on cognitive indistinguishability), and scalable modeling (as opposed to verbal theories or toy models). While none of these requirements seem out of reach, we would like to pinpoint some of the difficulties, which are the source of our caution.

Regarding the data, the interactive scenario would require

accessing the full (linguistic and non linguistic) output of the infant, not only her input. While this is not intrinsically harder to collect than the input, and is already been done in many corpora for older children, the issue of what to categorize as linguistic and non linguistic output and how to annotate it is not completely trivial.

Regarding computational modeling, instead of focusing on only one component (the learner) of one agent (the child), in the full interactive framework, one has to model two agents (the child and the adult) for a total of four components (the learner, the infant generator, the caregiver monitor, and the caregiver generator). Furthermore, the internal states of each agent has to be split into linguistic states (grammars) and non-linguistic (cognitive) states to represent the communicative aspects of the interaction (e.g., communicative intent, emotional/reinforcement signals). This, in turn, causes the split of each processing component into linguistic and cognitive subcomponents.

Although this is clearly a difficult endeavor, many of the individual ingredients needed for constructing such a system are already available in the following research areas. First, within speech technology, there are available components to build a language generator, as well as the perception and comprehension components in the adult caregiver. Second, within linguistics, psycholinguistics and neuroscience, there are interesting theoretical models of the learning of speech production and articulation in young children (?, ?, ?, ?). Third, within machine learning, great progress has been made recently on reinforcement learning, a powerful class of learning algorithms which assume that besides raw sensory data, the environment only provides sporadic positive or negative feedback (?, ?). This could be adapted to model the effect of the feedback loops on the learning components of the caregiver and the infant. Fourth, developmental robotics studies have developed the notion of intrinsic motivation, where the agent actively seek new information by being reinforced by its own learning rate (?, ?). This notion could be used to model the dynamics of learning in the child, and the adaptive effects of the caregiver-child feedback loops.

The most difficult part of this enterprise would perhaps concern the evaluation of the models. Indeed, each of these new components and subcomponents would have to be evaluated on their own in the same spirit as before, i.e., by running them on scalable data and testing them using human-validated tasks. For instance, the child language generator

differentiate humans from apes. A DNN-based phone recognizer has typically more than 200M parameters, which barring ways to compress the information, takes 400Mbytes. The human-specific genome is 5% of 3.2Gbase, which boils down to only 40Mbytes.

¹⁷<http://babieslearninglanguage.blogspot.com/2014/02/psychological-plausibility-considered.html>

¹⁸We thank Thomas Schatz, personal communication, for proposing this general formulation.

should be tested by comparing its output to age appropriate children's outputs, which requires the development of appropriate metrics (sentence length, complexity, etc) or human judgments. The cognitive subcomponents would have to be tested against experiments studying children and adults in experimentally controlled interactive loops (e.g., ?, ?, ?). In addition, because a complex system is more than the sum of its parts, individual component validation would not be sufficient, and the entire system would have to be evaluated.¹⁹

Fully specifying the methodological requirements for the reverse engineering of the interactive scenario would be a project of its own. It is not clear at present how much of the complications introduced by this scenario are necessary, at least to understand the first steps of language bootstrapping. To the extent that there are cultures where the direct input to the child is severely limited and/or the interactive character of that input circumscribed, it would seem that a fair amount of bootstrap can take place outside of interactive feedback loops. This is of course entirely an empirical issue, one that the reverse engineering approach should help to clarify.

References

- Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., & Steedman, M. (2017). Bootstrapping language acquisition. *Cognition*, 164, 116–143.
- Abrams, K., Chiarello, C., Cress, K., Green, S., & Ellett, N. (1978). Recent advances in the psychology of language. In R. Campbell & P. Smith (Eds.), (Vol. 4a, chap. The relation between mother-to-child speech and word-order comprehension strategies in children). New York: Plenum Press.
- Allen, J., & Seidenberg, M. S. (1999). The emergence of grammaticality in connectionist networks. In B. MacWhinney (Ed.), *Emergentist approaches to language: proceedings of the 28th carnegie symposium on cognition* (p. 115–151). Lawrence Erlbaum Associates.
- Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... Zhu, Z. (2016, 20–22 Jun). Deep speech 2: End-to-end speech recognition in english and mandarin. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd international conference on machine learning* (Vol. 48, pp. 173–182). New York, New York, USA: PMLR.
- Anderson, J. R. (1975). Computer simulation of a language acquisition system: A first report. In R. Solso (Ed.), *Information processing and cognition*. Hillsdale, N.J.: Lawrence Erlbaum.
- Angluin, D. (1988). *Identifying Languages from Stochastic Examples* [Technical Report 614. New Haven, CT: Yale University].
- Antetomaso, S., Miyazawa, K., Feldman, N., Elsner, M., Hitczenko, K., & Mazuka, R. (2017). Modeling Phonetic Category Learning from Natural Acoustic Data. In M. LaMendola & J. Scott (Eds.), *Proceedings of the 41th Annual Boston University Conference on Language Development* (pp. 32–45). Somerville, MA: Cascadilla Press. (OCLC: 992973124)
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 2425–2433).
- Badino, L., Canevari, C., Fadiga, L., & Metta, G. (2014). An Auto-encoder based approach to unsupervised learning of subword units. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Acl (1)* (pp. 238–247).
- Bates, E., & MacWhinney, B. (1987). Competition, Variation and Language learning. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition* (pp. 157–193). Hillsdale, N.J.: Lawrence Erlbaum.
- Bergelson, E., & Swingle, D. (2012, February). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9), 3253–3258.
- Bertoncini, J., Bijeljac-Babic, R., Blumstein, S. E., & Mehler, J. (1987). Discrimination in neonates of very short cvs. *The Journal of the Acoustical Society of America*, 82(1), 31–37.
- Berwick, R. (1985). *The acquisition of syntactic knowledge*. MIT Press.
- Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human perception and performance*, 14(3), 345.
- Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.
- Bornstein, M. H., & Tamis-LeMonda, C. S. (2010). The wiley-blackwell handbook of infant development. In J. G. Bremner & T. D. Wachs (Eds.), (pp. 458–482). Wiley-Blackwell.
- Botha, J. A., & Blunsom, P. (2013). Adaptor grammars for learning non-concatenative morphology. In *Emnlp* (pp. 345–356).
- Boves, L., Ten Bosch, L., & Moore, R. K. (2007). ACORNS- Towards computational modeling of communication and recognition skills. In *6th IEEE International Conference on In Cognitive Informatics* (pp. 349–356). IEEE.
- Brent, M. R. (1996a). Advances in the computational study of language acquisition. *Cognition*, 61(1), 1–38.
- Brent, M. R. (1996b). *Computational approaches to language acquisition*. MIT Press.
- Brown, R. (1973). *A first language; the early stages*. Cambridge, Mass: Harvard University Press.
- Bruner, J. S. (1975, April). The ontogenesis of speech acts. *Journal of Child Language*, 2(01).
- Bruner, J. S. (1983). *Child's Talk: Learning to Use Language*. New York, N.Y.: Norton.
- Bruni, E., Boleda, G., Baroni, M., & Tran, N. K. (2012). Distributional semantics in technicolor. In *Proceedings of the ACL* (pp. 136–145).
- Cadiu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., ... DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *arXiv preprint arXiv:1406.3284*.

¹⁹For instance, a combined learner/caregiver system should be able to converge on a similar grammar as a learner ran on realistic data. In addition, their interactions should not differ in 'naturalness' compared to what can be recorded in natural situations, see (?, ?).

- Carlin, M. A., Thomas, S., Jansen, A., & Hermansky, H. (2011). Rapid evaluation of speech representations for spoken term discovery. In *Proceedings of Interspeech*.
- Casillas, M. (2016). Age and turn type in mayan children's predictions about conversational turn-taking. to be presented at. In *Boston university child language development*. Boston, USA.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT press.
- Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of child language*, 30(03), 637–669.
- Christiansen, M. H., Conway, C. M., & Curtin, S. (2005). Multiple-cue integration in language acquisition: A connectionist model of speech segmentation and rule-like behavior. *Language acquisition, change and emergence: Essay in evolutionary linguistics*, 205–249.
- Christodoulopoulos, C., Goldwater, S., & Steedman, M. (2010). Two Decades of Unsupervised POS induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 575–584). Association for Computational Linguistics.
- Christophe, A., Millotte, S., Bernal, S., & Lidz, J. (2008). Bootstrapping lexical and syntactic acquisition. *Language and Speech*, 51(1-2), 61–75.
- Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition. *arXiv preprint arXiv:1601.02970*.
- Clark, A., Giorgolo, G., & Lappin, S. (2013). Statistical representation of grammaticality judgements: the limits of n-gram models. In *Proceedings of the fourth annual workshop on cognitive modeling and computational linguistics (cmcl)* (pp. 28–36).
- Clark, A., & Lappin, S. (2011). *Linguistic Nativism and the poverty of the stimulus*. Wiley and sons.
- Connor, M., Fisher, C., & Roth, D. (2013). Starting from scratch in semantic role labeling: Early indirect supervision. In T. Poibeau, A. Villavicencio, A. Korhonen, & A. Alishahi (Eds.), *Cognitive aspects of computational language acquisition* (pp. 257–296). Springer.
- Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, 14, 597–612.
- Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2017). Child-directed speech is infrequent in a forager-farmer population: a time allocation study. *Child Development*. doi: 10.1111/cdev.12974
- Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in cognitive sciences*, 13(4), 148–153.
- Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. Mit Press.
- Daland, R., & Pierrehumbert, J. B. (2011). Learning Diphone-Based Segmentation. *Cognitive Science*, 35(1), 119–155.
- Dehaene-Lambertz, G., Dehaene, S., et al. (1994). Speed and cerebral correlates of syllable discrimination in infants. *Nature*, 370(6487), 292–295.
- de Marcken, C. G. (1996). *Unsupervised Language Acquisition* (Unpublished doctoral dissertation). MIT.
- de Villiers, P. A., & de Villiers, J. G. (1972). Early judgments of semantic and syntactic acceptability by children. *Journal of Psycholinguistic Research*, 1(4), 299–310.
- Devlin, J., Gupta, S., Girshick, R., Mitchell, M., & Zitnick, C. L. (2015). Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*.
- Dresher, B. E., & Kaye, J. D. (1990). A computational learning model for metrical phonology. *Cognition*, 34(2), 137–195.
- D'Ulizia, A., Ferri, F., & Grifoni, P. (2011). A survey of grammatical inference methods for natural language learning. *Artificial Intelligence Review*, 36(1), 1–27. doi: 10.1007/s10462-010-9199-1
- Dunbar, E., Xuan-Nga, C., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., ... Dupoux, E. (2017). The zero resource speech challenge 2017. In *Proceedings of ASRU*.
- Dupoux, E. (2016). *Evaluating models of language acquisition: are utility metrics useful?* Retrieved from <http://bootphon.blogspot.fr/2015/05/models-of-language-acquisition-machine.html>
- Eilers, R. E., Gavin, W., & Wilson, W. R. (1979). Linguistic experience and phonemic perception in infancy: A crosslinguistic study. *Child Development*, 14–18.
- Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, 171(3968), 303–306.
- Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2), 179–211.
- Elman, J. L., & Zipser, D. (1988). Learning the hidden structure of speech. *The Journal of the Acoustical Society of America*, 83(4), 1615–1626.
- Elsner, M., Goldwater, S., & Eisenstein, J. (2012). Bootstrapping a Unified Model of Lexical and Phonetic Acquisition. In *Proceedings of the 50th Annual Meeting of the Association of Computational Linguistics* (pp. 184–193).
- Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, 32(5), 429–448.
- Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., ... Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, 2(3), 267–279.
- Feldman, N., Myers, E., White, K., Griffiths, T., & Morgan, J. (2011). Learners use word-level statistics in phonetic category acquisition. In *Proceedings of the 35th Annual Boston University Conference on Language Development* (pp. 197–209).
- Fernald, A. (2000). Speech to infants as hyperspeech: Knowledge-driven processes in early word recognition. *Phonetica*, 57(2-4), 241–254.
- Ferrucci, D. A. (2012). Introduction to “this is watson”. *IBM Journal of Research and Development*, 56(3.4), 1–1.
- Fiscus, J. G., Ajot, J., Garofolo, J. S., & Doddington, G. (2007). Results of the 2006 spoken term detection evaluation. In *Proc. sigir* (Vol. 7, pp. 51–57).
- Foppolo, F., Guasti, M. T., & Chierchia, G. (2012). Scalar implicatures in child language: Give children a chance. *Language learning and development*, 8(4), 365–394.
- Fourtassi, A., & Dupoux, E. (2014). A rudimentary lexicon and semantics help bootstrap phoneme acquisition. In *Proceedings of the 18th conference on computational natural language learning (CoNLL)*.
- Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: cross-modal contributions to speech perception. *Journal of Ex-*

- perimental Psychology: Human Perception and Performance*, 17(3), 816.
- Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... Lew-Williams, C. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*.
- Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010, November). Modeling human performance in statistical word segmentation. *Cognition*, 117(2), 107–125.
- Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5), 578–585.
- Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245), 273–278.
- Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, 25(3), 407–454.
- Gilmore, R. O., & Adolph, K. E. (2017). Video can make behavioural science more reproducible. *Nature Human Behaviour*, 1, s41562–017.
- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2016). Region-based convolutional networks for accurate object detection and segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 38(1), 142–158.
- Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition*, 1(1), 3–55.
- Gleitman, L. R., Gleitman, H., & Shipley, E. F. (1972). The emergence of the child as grammarian. *Cognition*, 1(2), 137–164.
- Gold, E. M. (1967). Language identification in the limit. *Information and control*, 10(5), 447–474.
- Goldin-Meadow, S. (2005). *Hearing Gesture: How Our Hands Help Us Think*. Belknap Press of Harvard University Press.
- Goldstein, M. H. (2008). Social Feedback to Babbling Facilitates Vocal Learning Michael H. Goldstein and Jennifer A. Schwade. *Psychological Science*.
- Goldwater, S. J. (2007). *Nonparametric Bayesian models of lexical acquisition* (Unpublished doctoral dissertation). Brown.
- Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., & Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of child language*, 14(01), 23–45.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. (<http://www.deeplearningbook.org>)
- Gregory, R. J. (2004). *Psychological testing: History, principles, and applications*. Allyn & Bacon.
- Grimshaw, J. (1981). Form, function and the language acquisition device. In C. L. Baker & J. J. McCarty (Eds.), *The logical problem of language acquisition* (pp. 165–182). The MIT Press.
- Guenther, F. H., & Vladusich, T. (2012, September). A neural theory of speech acquisition and production. *Journal of Neurolinguistics*, 25(5), 408–422. doi: 10.1016/j.jneuroling.2009.08.006
- Guevara-Rukoz, A., Mazuka, R., Thiollière, R., Martin, A., Schatz, T., Cristia, A., & Dupoux, E. (2017). Are words in infant directed speech easier to learn? a corpus study of acoustic clarity and phonological density. *arXiv preprint, arXiv:1712.08793*.
- Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... Ng, A. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.
- Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3), 146–162.
- Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young american children*. Paul H Brookes Publishing.
- Harwath, D., & Glass, J. R. (2017). Learning word-like units from joint audio-visual analysis. *arXiv preprint arXiv:1701.07481*.
- Harwath, D., Torralba, A., & Glass, J. (2016). Unsupervised learning of spoken language with visual context. In *Advances in neural information processing systems* (pp. 1858–1866).
- Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, 95(2), 245–258.
- 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.
- Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, 39(3), 379–440.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026–1034).
- Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., ... Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6), 82–97.
- Hirsh-Pasek, K., Nelson, D. G. K., Jusczyk, P. W., Cassidy, K. W., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition*, 26(3), 269–286.
- Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child development*, 74(5), 1368–1378.
- Hoff, E. (Ed.). (2012). *Research methods in child language: a practical guide*. Malden, MA: Wiley-Blackwell.
- Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., ... Bloom, L. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the society for research in child development*, i–135.
- Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive psychology*, 61(4), 343–365.
- Jackendoff, R. (1997). *The architecture of the language faculty* (No. 28). MIT Press.
- Jäger, G., & Rogers, J. (2012). Formal language theory: refining the chomsky hierarchy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1598), 1956–1970.
- Jansen, A., Dupoux, E., Goldwater, S., Johnson, M., Khudanpur, S., Church, K., ... Thomas, S. (2013). A summary of the 2012 JH CLSP workshop on zero resource speech technologies and models of early language acquisition. In *Icassp-2013 (IEEE international conference on acoustics speech and signal processing)* (p. 8111–8115). Vancouver, BC, Canada. doi: 10.1109/icassp.2013.6639245
- Johnson, K. (2004). Gold's theorem and cognitive science. *Philosophy of Science*, 71(4), 571–592.

- Johnson, M. (2008). Using Adaptor Grammars to Identify Synergies in the Unsupervised Acquisition of Linguistic Structure. In *ACL* (pp. 398–406).
- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... Dean, J. (2016). Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.
- Johnson, W., & Reimers, P. (2010). *Patterns in child phonology*. Edinburgh University Press.
- Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, Mass.: MIT Press.
- Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive psychology*, 29(1), 1–23.
- Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of memory and language*, 32(3), 402–420.
- Jusczyk, P. W., Hirsh-Pasek, K., Nelson, D. G. K., Kennedy, L. J., Woodward, A., & Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. *Cognitive psychology*, 24(2), 252–293.
- Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive psychology*, 39(3), 159–207.
- Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., ... Bengio, Y. (2015). Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 1–13.
- Katsos, N., & Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, 120(1), 67–81.
- Kelley, K. (1967). *Early syntactic acquisition* (Tech. Rep. No. P-3719). Santa Monica, California: Rand Corp.
- Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific reports*, 6, 32672.
- Kiela, D., & Bottou, L. (2014). Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. In *EMNLP* (pp. 36–45).
- Kohonen, T. (1988). The 'neural' phonetic typewriter. *Computer*, 21(3), 11–22.
- Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. Cambridge, MA: MIT Press.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).
- Kuhl, P., Williams, K., Lacerda, F., Stevens, K., & Lindblom, B. (1992, January). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044), 606–608. doi: 10.1126/science.1736364
- Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences*, 97(22), 11850–11857.
- Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., ... Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, 277(5326), 684–686.
- Kunkel, S., Schmidt, M., Eppler, J. M., Plesser, H. E., Masumoto, G., Igarashi, J., ... others (2014). Spiking network simulation code for petascale computers. *Frontiers in neuroinformatics*, 8(78).
- Kwiatkowski, T., Goldwater, S., Zettlemoyer, L., & Steedman, M. (2012). A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. *EACL 2012*, 234.
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building machines that learn and think like people. *arXiv preprint arXiv:1604.00289*.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), 211.
- Langley, P., & Carbonell, J. G. (1987). Language acquisition and machine learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 115–155). Hillsdale, N.J.: Lawrence Erlbaum.
- Larsen, E., Cristia, A., & Dupoux, E. (2017). Relating unsupervised word segmentation to reported vocabulary acquisition. In *Proceedings of Interspeech* (p. 2198–2202).
- Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cognitive science*, 41(5), 1202–1241.
- Lee, C.-y., & Glass, J. (2012). A nonparametric Bayesian approach to acoustic model discovery. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1* (pp. 40–49).
- Levesque, H. J., Davis, E., & Morgenstern, L. (2011). The winograd schema challenge. In *Aaai spring symposium: Logical formalizations of commonsense reasoning* (p. 552–561).
- Liang, P., Jordan, M. I., & Klein, D. (2011). Learning dependency-based compositional semantics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 590–599). Association for Computational Linguistics.
- Lidz, J., & Gagliardi, A. (2015, January). How Nature Meets Nurture: Universal Grammar and Statistical Learning. *Annual Review of Linguistics*, 1(1), 333–353. doi: 10.1146/annurev-linguist-030514-125236
- Lidz, J., & Musolino, J. (2002). Children's command of quantification. *Cognition*, 84(2), 113–154.
- Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4, 521–535.
- Lu, C., & Tang, X. (2014). Surpassing human-level face verification performance on lfw with gaussianface. *arXiv preprint arXiv:1404.3840*.
- Ludusan, B., Seidl, A., Dupoux, E., & Cristia, A. (2015). Motif discovery in infant-and adult-directed speech. In *Conference on empirical methods in natural language processing (emnlp)* (p. 93–102).
- Ludusan, B., Versteegh, M., Jansen, A., Gravier, G., Cao, X.-N., Johnson, M., & Dupoux, E. (2014). Bridging the gap between speech technology and natural language processing: an evalu-

- ation toolbox for term discovery systems. In *Proceedings of LREC* (p. 560-567).
- MacWhinney, B. (1978). Conditions on acquisitional models. In *Proceedings of the ACM annual conference* (pp. 421-427). ACM.
- MacWhinney, B. (1987). The Competition model. In B. MacWhinney (Ed.), (pp. 249-308). Hillsdale, N.J.: Lawrence Erlbaum.
- MacWhinney, B. (2000). The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database. *Computational Linguistics*, 26(4), 657-657.
- Magri, G. (2015). Noise robustness and stochastic tolerance of OT error-driven ranking algorithms. *Journal of Logic and Computation*.
- Mandel, D. R., Jusczyk, P. W., & Pisoni, D. B. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science*, 6(5), 314.
- Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, 46(1), 53-85.
- Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cognitive Science*, 37, 103-124.
- Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., & Cristia, A. (2015). Mothers speak less clearly to infants: A comprehensive test of the hyperarticulation hypothesis. *Psychological Science*, 26(3), 341-347.
- Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3), B101-B111.
- Mazuka, R., Cao, Y., Dupoux, E., & Christophe, A. (2011). The development of a phonological illusion: a cross-linguistic study with Japanese and French infants. *Developmental science*, 14(4), 693-699.
- Mazuka, R., Igarashi, Y., & Nishikawa, K. (2006). *Input for learning Japanese: Riken Japanese mother-infant conversation corpus* (Vol. 106(165); Tech. Rep. No. TL 2006-16).
- McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, 12(3), 369-378.
- McMurray, B., Kovack-Lesh, K. A., Goodwin, D., & McEchron, W. (2013, November). Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? *Cognition*, 129(2), 362-378.
- Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, 29(2), 143-178.
- Meltzoff, A. N., Kuhl, P. K., Movellan, J., & Sejnowski, T. J. (2009). Foundations for a new science of learning. *science*, 325(5938), 284-288.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of workshop at iclr*.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529-533.
- Morgan, J., & Demuth, K. (1996). *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. L. Erlbaum Associates.
- Muscariello, A., Gravier, G., & Bimbot, F. (2009). Audio keyword extraction by unsupervised word discovery. In *INTERSPEECH 2009: 10th Annual Conference of the International Speech Communication Association* (p. 2843-2846).
- Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (Non)words, (non)words, (non)words: evidence for a protollexicon during the first year of life. *Developmental Science*, 16(1), 24-34.
- Nguyen, A., Yosinski, J., & Clune, J. (2014). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *arXiv preprint arXiv:1412.1897*.
- Olivier, D. C. (1968). *Stochastic grammars and language acquisition mechanisms* (Unpublished doctoral dissertation). Harvard University Doctoral dissertation.
- Ondel, L., Burget, L., & Černocký, J. (2016). Variational Inference for Acoustic Unit Discovery. *Procedia Computer Science - Proceedings of SLTU*, 81, 80-86. doi: 10.1016/j.procs.2016.04.033
- Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.
- Oudeyer, P.-Y., Kaplan, F., & Hafner, V. V. (2007, April). Intrinsic Motivation Systems for Autonomous Mental Development. *IEEE Transactions on Evolutionary Computation*, 11(2), 265-286.
- Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families. *Child development*, 76(4), 763-782.
- Park, A. S., & Glass, J. R. (2008, January). Unsupervised Pattern Discovery in Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1), 186-197.
- Pearl, J. (1997). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Francisco, Calif.: Morgan Kaufmann Publishers.
- Pearl, L., & Phillips, L. (2016). Language, cognition, and computational models. In A. Villavicencio & T. Poibeau (Eds.), (chap. Evaluating language acquisition models: A utility-based look at Bayesian segmentation). Cambridge Univ Press.
- Peters, A. M. (1983). *The units of language acquisition* (Vol. 1). Cambridge University Press Archive.
- Pinker, S. (1984). *Language learnability and language development*. Cambridge, Mass: Harvard University Press.
- Pinker, S. (1987). The bootstrapping problem in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 399-441). Lawrence Erlbaum.
- Pinker, S. (1989). *Learnability and cognition: the acquisition of argument structure*. MIT Press.
- Pinker, S. (1994). *The language instinct*. Harper.
- Podesva, R. J. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of sociolinguistics*, 11(4), 478-504.
- Poizner, H., Klima, E., & Bellugi, U. (1987). *What the hand reveals about the brain*. MIT Press Cambridge, MA.
- Pons, C. G., Anguera, X., & Binefa, X. (2013). Two-Level Clustering towards Unsupervised Discovery of Acoustic Classes. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on* (Vol. 2, pp. 299-302). IEEE.

- Rahmani, H., Mian, A., & Shah, M. (2016). Learning a deep model for human action recognition from novel viewpoints. *arXiv preprint arXiv:1602.00828*.
- Räsänen, O. (2012). Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions. *Speech Communication*, 54(9), 975–997. doi: 10.1016/j.specom.2012.05.001
- Räsänen, O., & Rasilo, H. (2015). A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological review*, 122(4), 792.
- Rowe, M. L., & Goldin-Meadow, S. (2009). Differences in early gesture explain sex disparities in child vocabulary size at school entry. *Science*, 323(5916), 951–953.
- Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, 112(41), 12663–12668.
- Roy, D. (2009). New horizons in the study of child language acquisition. In *Proceedings of interspeech* (p. 13–20). Brighton, England.
- Roy, D. K., & Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1), 113–146.
- Rumelhart, D. E., & McClelland, J. L. (1987). Mechanisms of language acquisition. In B. MacWhinney (Ed.), (pp. 195–248). Erlbaum Hillsdale, NJ.
- Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent-child discourse. *Children's language*, 4, 1–28.
- Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current directions in psychological science*, 12(4), 110–114.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926–1928.
- Sakas, W. G., & Fodor, J. D. (2012). Disambiguating syntactic triggers. *Language Acquisition*, 19(2), 83–143.
- Sangwan, A., Hansen, J., Irvin, D., Crutchfield, S., & Greenwood, C. (2015). Studying the relationship between physical and language environments of children: Who's speaking to whom and where? In *Signal processing and signal processing education workshop (sp/spe)*, 2015 IEEE (pp. 49–54).
- Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., ... Hall, P. (2017). English conversational telephone speech recognition by humans and machines. In *arXiv preprint arXiv:1703.02136*.
- Saxton, M. (1997). The contrast theory of negative input. *Journal of child language*, 24(01), 139–161.
- Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline. In *INTERSPEECH-2013* (p. 1781–1785). Lyon, France.
- Seidl, A., Tincoff, R., Baker, C., & Cristia, A. (2015). Why the body comes first: effects of experimenter touch on infants' word finding. *Developmental science*, 18(1), 155–164.
- Shneidman, L. A., & Goldin-Meadow, S. (2012, September). Language input and acquisition in a Mayan village: how important is directed speech?: Mayan village. *Developmental Science*, 15(5), 659–673.
- Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-month-old infants. *Proceedings of the National Academy of Sciences*, 108(15), 6038–6043. doi: 10.1073/pnas.1017617108
- Siklossy, L. (1968). *Natural language learning by computer* (Tech. Rep.). DTIC Document.
- Silberer, C., Ferrari, V., & Lapata, M. (2016). Visually Grounded Meaning Representations. *IEEE transactions on pattern analysis and machine intelligence*.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587), 484–489.
- Silverman, D., Blankenship, B., Kirk, P., & Ladefoged, P. (1995). Phonetic structures in jalapa mazatec. *Anthropological Linguistics*, 37(1), 70–88. Retrieved from <http://www.jstor.org/stable/30028043>
- Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1), 39–91.
- Siu, M.-h., Gish, H., Chan, A., Belfield, W., & Lowe, S. (2013). Unsupervised training of an HMM-based self-organizing recognizer with applications to topic classification and keyword discovery. *Computer Speech & Language*.
- Smith, K., Smith, A. D., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, 35(3), 480–498.
- Smith, L. B., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, 16(3), 407–419.
- Smith, N. A., & Trainor, L. J. (2008). Infant-directed speech is modulated by infant feedback. *Infancy*, 13(4), 410–420.
- Snow, C. E. (1972, June). Mothers' Speech to Children Learning Language. *Child Development*, 43(2), 549.
- Song, J. J. (2010). *The oxford handbook of linguistic typology*. Oxford Univ. Press.
- Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, 134, 219–248.
- Stark, R. (1980). Child phonology. Vol. 1: Production. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), (chap. Stages of development in the first year of life). New York: Acad. Press.
- Steedman, M. (2014). Evolutionary basis for human language: Comment on "Toward a computational framework for cognitive biology: Unifying approaches from cognitive neuroscience and comparative cognition" by tecumseh fitch. *Physics of life reviews*, 11(3), 382–388.
- Sundara, M., Polka, L., & Genesee, F. (2006). Language-experience facilitates discrimination of /d-/in monolingual and bilingual acquisition of english. *Cognition*, 100(2), 369–388.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.
- Swingle, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1536), 3617–3632.

- Tamis-LeMonda, C. S., & Rodriguez, E. T. (2008). Parents' role in fostering young children's learning and language development. In (pp. 1–11).
- Ten Bosch, L., & Cranen, B. (2007). A computational model for unsupervised word discovery. In *INTERSPEECH* (pp. 1481–1484).
- Tesar, B., & Smolensky, P. (1998). Learnability in optimality theory. *Linguistic Inquiry*, 29(2), 229–268.
- Tesar, B., & Smolensky, P. (2000). *Learnability in optimality theory*. MIT Press.
- Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, 56(1), 16–34. doi: 10.1016/j.jml.2006.07.002
- Thiollière, R., Dunbar, E., Synnaeve, G., Versteegh, M., & Dupoux, E. (2015). A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling. In *INTERSPEECH-2015* (p. 3179–3183).
- Thomas, D. G., Campos, J. J., Shucard, D. W., Ramsay, D. S., & Shucard, J. (1981). Semantic comprehension in infancy: A signal detection analysis. *Child development*, 798–803.
- Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, 10(2), 172–175.
- Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition*. Cambridge, Mass: Harvard University Press.
- Trueswell, J. C., Lin, Y., Armstrong, B., Cartmill, E. A., Goldin-Meadow, S., & Gleitman, L. R. (2016). Perceiving referential intent: Dynamics of reference in natural parent-child interactions. *Cognition*, 148, 117–135.
- Tsivdis, P. A., Pouncy, T., Xu, J. L., Tenenbaum, J. B., & Gershman, S. J. (2017). Human learning in atari." In *The aaai 2017 spring symposium on science of intelligence: Computational principles of natural and artificial intelligence*.
- Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses toward cumulative data assessment. *Perspectives on Psychological Science*, 9(6), 661–665.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460.
- Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37, 141–188.
- Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, 104(33), 13273–13278.
- Van Cleve, J. V. (2004). *Genetics, disability, and deafness*. Gallaudet University Press.
- VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., & MacWhinney, B. (2016). Homebank: An online repository of daylong child-centered audio recordings. In *Seminars in speech and language* (Vol. 37, pp. 128–142).
- van de Weijer, J. (2002). How much does an infant hear in a day. In *GALA 2001 Conference on Language Acquisition, Lisboa*.
- Varadarajan, B., Khudanpur, S., & Dupoux, E. (2008). Unsupervised learning of acoustic subword units. In *Proceedings of acl-08: Hlt* (p. 165–168).
- Versteegh, M., Anguera, X., Jansen, A., & Dupoux, E. (2016). The zero resource speech challenge 2015: Proposed approaches and results. In *SLTU-2016 - procedia computer science* (Vol. 81 (2016), pp. 67 – 72).
- Versteegh, M., Thiollière, R., Schatz, T., Cao, X.-N., Anguera, X., Jansen, A., & Dupoux, E. (2015). The zero resource speech challenge 2015. In *INTERSPEECH-2015* (p. 3169–3173).
- Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, 167(3917), 392–393.
- Weisleder, A., & Fernald, A. (2013, November). Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary. *Psychological Science*, 24(11), 2143–2152.
- Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45.
- Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language learning and development*, 1(2), 197–234.
- Werker, J. F., & Tees, R. C. (1984). Cross-language Speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7, 49–63.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., ... Zweig, G. (2016). Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*.
- Xu, D., Yapanel, U. H., Gray, S. S., Gilkerson, J., Richards, J. A., & Hansen, J. H. (2008). Signal processing for young child speech language development. In *Wocci* (p. 20).
- Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619–8624.
- Yang, C. D. (2002). *Knowledge and learning in natural language*. Oxford University Press.
- Yu, C., & Smith, A. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5), 414–420.