Cognition 170 (2018) 312–327

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Can infants learn phonology in the lab? A meta-analytic answer

Alejandrina Cristia

LSCP, Département d'études cognitives, ENS, EHESS, CNRS, PSL Research University, France

A R T I C L E I N F O

Keywords:
Infant language learning
Artificial grammars
Experimental psychology
Replication

A B S T R A C T

Two of the key tasks facing the language-learning infant lie at the level of phonology: establishing which sounds are contrastive in the native inventory, and determining what their possible syllabic positions and permissible combinations (phonotactics) are. In 2002–2003, two theoretical proposals, one bearing on how infants can learn sounds (Maye, Werker, & Gerken, 2002) and the other on phonotactics (Chambers, Onishi, & Fisher, 2003), were put forward on the pages of *Cognition*, each supported by two laboratory experiments, wherein a group of infants was briefly exposed to a set of pseudo-words, and plausible phonological generalizations were tested subsequently. These two papers have received considerable attention from the general scientific community, and inspired a flurry of follow-up work. In the context of questions regarding the replicability of psychological science, the present work uses a meta-analytic approach to appraise extant empirical evidence for infant phonological learning in the laboratory. It is found that neither seminal finding (on learning sounds and learning phonotactics) holds up when close methodological replications are integrated, although less close methodological replications do provide some evidence in favor of the sound learning strand of work. Implications for authors and readers of this literature are drawn out. It would be desirable that additional mechanisms for phonological learning be explored, and that future infant laboratory work employ paradigms that rely on constrained and unambiguous links between experimental exposure and measured infant behavior.

1. Introduction

One of the fundamental tasks facing human infants pertains to learning the ambient language's phonology, including both determining the sound inventory and the constraints regarding the position and sequencing of those sounds. An overwhelming body of evidence suggests that attunement to the ambient language's phonology begins in infancy (Werker & Hensch, 2015), raising the question of what mechanisms may begin to operate at such an early age. Given that infants' lexicon is highly constrained, a mainstream assumption posits that infants may begin to learn their language's sound system by applying simple statistical mechanisms directly onto the spoken signal they hear.

For learning native sound categories specifically, Maye, Werker, and Gerken (2002) proposed that infants track the distribution of acoustic cues in the input, as modes or peaks in the distribution of one acoustic cue could reflect the presence of a sound category implemented in that location of acoustic space. The underlying intuition runs as follows: if a language has two sounds that contrast along an acoustic dimension, then there will be two peaks in the distribution of tokens along that dimension (one corresponding to each sound), whereas if there is only one category in those regions of acoustic space, there will be a single peak in the frequency distribution. A sensitivity to modes in the frequency distribution would then help infants learn the sound system of their ambient language. For learning phonotactics (the regularities concerning the position and sequencing of sounds), Chambers, Onishi, and Fisher (2003) proposed that infants keep track of the frequency with which sounds occur in a given syllabic position and/or in a specific order.

These interesting proposals can be evaluated in a number of ways. One crucial way of assessing the potential explanatory value of such proposals is to carry out proofs of principle demonstrating that, at least in ideal situations, infants extract the kind of information one predicts they would. For those two theoretical proposals, this has been done using what may termed artificial grammars, languages, or phonologies: Researchers devised a simplified language to represent some feature of human language, exposed infants to some exemplars generated from that grammar, and later tested infant perception with new items that could be distinguished if the artificial phonology had been learned.[1]

---

[1] It may be relevant to point out that the artificial phonologies constructed for these studies are not only much simpler than those of natural languages, but may not have the same properties. For example, a key notion of phonology is that of minimal pair, two words that are identical except for one sound, yet mean different things; none of the artificial phonologies built for these studies contain any reference to meaning. Furthermore, there is no evidence that infants learn them using the exact same mechanisms they employ to learn natural languages' phonologies. Nonetheless, we have no evidence that they do employ the same cognitive mechanisms available for learning other types of categories and regularities outside the domain of phonology, in the auditory or any other modality. Instead, the choice of the term 'artificial phonology' reflects what the experimenter put in, rather than what the infant actually does. The actual learning may turn out to be as general as 'auditory learning' (applying equally to speech and non-speech) or as specific as 'adaptation to a syllable repeated more than 20 times' (and not applicable to any other situation).

http://dx.doi.org/10.1016/j.cognition.2017.09.016
Received 25 February 2016; Received in revised form 29 June 2017; Accepted 27 September 2017

Indeed, the original theoretical proposals in Maye et al. (2002) and Chambers et al. (2003) were supported each by two such experiments. Additionally, the same authors and others continued to explore the proposed basic mechanisms in additional laboratory-based experiments that could be considered conceptual replications of the original studies (e.g., Liu & Kager, 2014; Maye, Weiss, & Aslin, 2008; Wanrooij, Boersma, & Zuijen, 2014; Yoshida, Pons, Maye, & Werker, 2010, all followed up on Maye et al., 2002's sound category learning proposal).[2] Undoubtedly, such experiments are only the first step in laying out the explanatory power of such proposals. For instance, additional steps may include checking whether the spoken input available to infants contains the characteristics attributed to it (i.e., whether the number of modes in acoustic cue distributions corresponds to the number of sound categories in the language; Bion, Miyazawa, Kikuchi, & Mazuka, 2013); and designing computational models that mimic the procedure attributed to the learner, and assessing whether such a learner can succeed in detecting the native phonological inventory when presented with idealized or realistic input (e.g., Vallabha, McClelland, Pons, Werker, & Amano, 2007; Versteegh et al., 2015). Those two proposals have inspired a great deal of research, leading to numerous experimental extensions, and inviting both corpora studies and computer models. In view of the important theoretical implications of Maye et al. (2002) and Chambers et al. (2003), as well as their widespread effects in the study of early language acquisition, the present paper seeks to assemble extant empirical evidence on such proofs of principle, and assess how convincing this evidence is through meta-analytic tools.

1.1. *Why carry out a meta-analysis?*

Some readers may wonder: How can a meta-analytic approach be integrated into research on cognition? A single experiment sometimes appears more compelling than the accumulation of varied evidence. Indeed, well-designed, robust experiments can be extremely useful in shedding light on cognitive processes whose effects are fundamental, yet so slight or transitory that they are not obvious in simple observations of behavior in the 'wild'. What a meta-analytic approach can add is an estimation of how *robust and replicable* an experiment is. When we carry out a laboratory experiment, the assumption is that we are creating replicable conditions for observing the effects of a hypothesized cognitive construct. If those conditions are re-created at another point in time, or by a different group of researchers, provided that the cognitive construct is available to that second group of infant participants, we as experimentalists would predict that we will observe similar effects. Naturally, since any single experiment is a noisy observation of underlying reality, we can expect that effect sizes will vary across experiments, and can then use a cumulative approach to estimate the underlying effect size despite variations due to noise.

A meta-analytic approach can further help us assess whether variance in results across conceptual replications is systematic. For instance, it can help us measure the effects of changes in the design across different implementations of the same general conceptual goal, or test the reliable effect of a factor that has been invoked on theoretical or empirical grounds. For example, infant age is often discussed when attempting to explain changes in performance, as older infants may be more entrenched in their native language and more resistant to the exposure (Yoshida et al., 2010), and younger infants, although more flexible, may be cognitively limited and learn more slowly (Seidl, Cristia, Onishi, & Bernard, 2009). These differences often occur within the same paper (Seidl et al., 2009), but sometimes those factors are invoked to explain differences across data sets published in different papers (Cristia, Seidl, & Gerken, 2011). If age is a variable that structures performance, then it should explain some variance in results even when age does not vary by design, and should thus be evident when comparing the size of the effects found in different experiments, including those that are unrelated to the scientists making the initial claim.

Finally, meta-analyses can uncover a further source of structure in public data, namely that emerging from conscious or unconscious biases affecting the producers of those data. The psychological sciences are seeing today a revival of concern in the robustness and replicability of our results. The problem likely emerges from the fact that the current reward scheme pushes researchers, reviewers, and editors to value more results where there is a p-value below the .05 threshold than results that are not significant (see e.g., Ioannidis, 2005; Nosek, Spies, & Motyl, 2012 for the general argument, and Open Science Collaboration, 2015 for a recent set of results in psychology). As a consequence, there is an over-representation of significant results in the literature, and quite likely an increase in the number of false positives that are present in published studies (Sterling, Rosenbaum, & Weinkam, 1995). Beyond the why's and how's, what is certain is that the community should be careful when interpreting published literature, as it may be the joint result of actual findings and biases. A meta-analytic approach can help us determine whether there is evidence of biases in reporting, through the study of the systematic patterns found in the literature.

1.2. *The present work*

My main goal is to inform readers about the overall empirical value of artificial phonology studies for our understanding of putative language learning mechanisms in infancy, covering distributional learning for establishing the phonological inventory on the one hand, and phonotactic learning on the other. A secondary goal, which will be evident in the discussion, is to explore implications of the results found for the empirical literature on infant laboratory learning at large.

To address the primary goal, I apply meta-analytic tools to the body of infant artificial phonology studies in order to describe it in three ways. First, I assess the overall robustness of effects found, estimated through the weighted mean effect size. Second, I explore whether certain factors are moderators of overall effect sizes, meaning that there is variance that may be systematically attributed to them, for example design characteristics that regularly lead to greater or smaller effect sizes. Finally, I investigate the possibility that there is selective reporting in this literature via inspection of funnel plots and reported p-values, techniques that will be introduced in further detail below. Although for conceptual reasons one may prefer to present the sound category learning literature before the phonotactic learning one, the methods are more complex in the former than the latter. Therefore, I will present them in the opposite order to facilitate readers' comprehension.

2. Phonotactic learning

I have followed the 'Preferred Reporting Items for Systematic Reviews and Meta-Analyses' (PRISMA) statement (Moher, Liberati, Tetzlaff, & Altman, 2009). A file with inclusion/exclusion decisions, a flow-chart summarizing this process, the PRISMA checklist, a spreadsheet containing the meta-analytic data, and all analysis scripts are available for download from https://osf.io/9zd2a/.

---

[2] A note may be needed regarding the word "replication". Some hold that a replication is only an experiment that keeps all factors constant; in the extreme, the only way to replicate would be to go back in time and run the experiment again, as it is possible that changes in e.g., weather, society, prevalence of mild otitis, etc. could affect results via their effects on the infants tested. More commonly, an experiment constitutes a *strict* replication when the original experiment is repeated on a new sample of the same population, in the same or another lab. Finally, the term *conceptual* replication is used for cases in which only the conceptually key factors are kept constant, but methods are allowed to vary. Studies included in the present meta-analyses fall in the latter category, as stimuli, infant age, and sometimes preference method can change between the original study and the follow-ups, but they all share the same conceptually key events: exposure followed by some test in which preference could only be explained if infants had picked up on an underlying contrast or regularity.