The way we learn

Monica Barbir

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Ecole Normale Supérieure

The Way We Learn

Soutenue par
Monica BARBIR
Le 4 Octobre 2019

Ecole doctorale n°158
Cerveau, Cognition, Comportement

Spécialité
Sciences cognitives

Composition du jury :

Judit, GERVAIN
Directrice de recherche, Université Paris Descartes
Présidente/ Examinatrice

Nuria, SEBASTIAN-GALLES
Professeure, Universitat Pompeu Fabra
Rapportrice

Chen, YU
Professeur, Indiana University Bloomington
Rapporteur

Luigi, RIZZI
Professeur, Università di Siena
Examinateur

Isabelle, DAUTRICHE
Chargée de recherche, Université Aix-Marseille
Examinatrice

Anne, CHRISTOPHE
Directrice de recherche, Ecole Normale Supérieure
Directrice de thèse

ENS | PSL

The Way We Learn

For Mother Nature

En présence de la création fleurie, embaumée, aimante et charmante, le ciel splendide inondait d'aurore la Tourgue[1] et la guillotine, et semblait dire aux hommes : Regardez ce que je fais et ce que vous faites.

Victor Hugo, Quatrevingt-treize[2]

[1] La Tourgue est une forteresse où eut lieu une bataille décisive (et fictive) entre républicains et royalistes.
[2] Hugo, V. (1979). *Quatrevingt-treize* (Ed. Folio). Paris: Gallimard. (p. 517).

Thank yous

Je n'ai pas réussi à redresser le monde, à vaincre la bêtise et la méchanceté, à rendre la dignité et la justice aux hommes, mais j'ai tout de même gagné le tournoi de ping-pong à Nice, en 1932.
Romain Gary, La promesse de l'aube[3]

And I. I wrote this thesis. And lived this brief eternity that permeated deep into the nooks and crannies of my existence.

Written in that fashion, you must be imagining something like a keener actress prepping for a one-woman show, who gets so far into character that the line between character and self blurs: she repeats her lines while waiting for the bus, stirring in spaghetti sauce, or riding her bike. Then, one day she finds herself ordering sushi in the same intonation as her character, gesturing like her, feeling her emotions. And yet, a thesis is not a one-person show. Sure. There's a main character, and likely a keener at that. But. Without the other roles in the play, the main character's performance would have been an incoherent monologue that drags on and on.

Each of the roles in the thesis play requires actresses and actors who are exceptional in their own right. Some are on set from beginning to end, leaving just briefly. Others flash across the stage, leaving just an afterimage of their presence. Without even one of them, would our impression of the main character's performance have been the same? Would the play have touched the audience in the way it had? It is unlikely. It is the totality of the whole cast and, indeed, crew, from the performers to the stage technicians to the make-up artists, that makes a thesis what it is. At the end of the performance, the team joins hands on stage and bows. The applause is for everyone.

***

Mise en scène (Producer)
Anne Christophe

To my amazing PhD supervisor. Anne guided me through the PhD labyrinth, masterfully balancing my curious exploration with her years (decades?) of experience. She was at my side through thick and thin, pushing me to be a better researcher at every step. Beyond research, Anne was a role-model in many ways, and inspired me to use my strengths and do what I love. French people always think I exaggerate when I label things 'awesome', but none would contend that Anne is an awesome PhD supervisor.

How to thank a researcher? With ideas? I hope this thesis and any ideas I may have brought or sparked in regard to your research can serve as thank yous, Anne.

May we be *bamoules* who *nuve* together forever.

[3] Gary, R. (1978). *La promesse de l'aube* (Ed. Folio). Paris: Gallimard. (p. 60).

Avec (Starring)

Isabelle Dautriche, Judit Gervain, Luigi Rizzi, Núria Sebastián Gallés, & Chen Yu

To my wonderful thesis committee. Writing this, I am still in awe that such a superb group of researchers will be making up the committee. I am grateful for all the future comments and discussion, and I hope we will continue to work together into the future.

I would like to especially thank Núria and Judit, who were also part of my PhD committee, and followed my research from the very first year. Their comments were invaluable, and it was an enormously enriching experience to discuss with them.

Décor (Set)

Anne-Caroline Fiévet & Isabelle Brunet

To the best lab managers on this earth. Anne-Caroline and Isabelle were magnificently accommodating and hard-working, from managing recruitment for complicated experimental designs (sending videos home to adults on a Sunday must have been the worst of all) to finding participants in a pinch so an abstract could be completed. They were always there to make this PhD the smoothest ride possible.

Dramaturgie (Dramaturges)

Mireille Babineau, Naomi Havron, Samuel Recht, & Kishore Sivakumar

To my splendid co-authors.
The two post-docs with whom I spent three, wait, my whole PhD, with, Naomi and Mireille, you taught me so much. Spending three years with you was worth more than all my cog sci class added together. Praat pro Mireille showed me all the secrets of Praat and how to do fine-graining acoustic analyses and stats pro Naomi taught me the beauty of statistics and how to make powerful analyses. Most of all, both of them anchored me into Open Science and all the positive aspects it implies. You have been my role models and I can't imagine doing science without you. I sincerely hope we continue to collaborate throughout our careers.
Kishore was my intern in my second year of PhD, when I had had over a year of null results. He showed me the pure strength of a positive outlook. For someone who stuck through slews of testing, debugging of codes, and creating wonderful Praat scripts, I would not know where to begin to thank you.
And Samuel. The meta-cognition specialist, who inspired and fueled my studies, always pushing me to think of the bigger implications of my work, reciting infinite follow-up studies, and sending me all and any potentially interesting articles. Python pro, who created elegant work-arounds for limitations in the Python source codes, and professional infant who spent most of his PhD walking around the department with a target sticker on his forehead. To this day, when our conversation wanes, his gaze drifts off to the distance, and he utters, in perfect child-directed speech, "Oh, regarde ko bamoule".

Régie générale (Stage management)

Radhia Achheb, Chase Crook, Michel Dutat, Clémence Lebon, Camille Straboni, Vireack Ul, & Zhiguo Wang

To the top-notch tech and admin teams at the LSCP, the DEC, and Eyelink. A big thank you to you all for your help and patience. A special shout out to Michel and Vireack for intervening

each time the Babylab computers ran awry, and Radhia for her streamlined admin, even at the very last minute.

Lumières (Lighting)
Sylvain Charron, Vincent de Gardelle, Jérôme Sackur, & Brent Strickland

To the brilliant researchers at the LSCP and beyond. A heartfelt thank you for the hours spent making my experiments the best they could be.

Surtitrages (Surtitles)
Clémence Baly, Héloïse Hennebelle, Anna Jondot, Filippo Milani, Mariana Erin McCune, & Emilie Rolland

To my marvelous interns. Thank you for your helping hands. It was wonderful having your input and energy shine into my PhD. Special props to Emilie, my very firstest intern and future scientist.

Scenographie (Scenography)
Julia Carbajal, Alex de Carvalho, Adriana Guevara Rukoz, Cathal O'Madagain, & Sho Tsuji

To my senpais. Thank you for all your advice and guidance, and general awesomeness.

Composition et musique (Composition and music)
Si Berebi, Mélissa Berthet, Guillaume Dezecache, Ava Guez, Ghislaine Labouret, Alice Latimier, Georgia Loukatou, Camila Scaff, Leticia Schiavon Kolberg, & Camille Williams

To the greatest PhD (and post-doc) (and master) room ever. Thank you for all the laughs and good times (and all the science of course).

Collaboration artistique (Artistic collaboration)
Fams and friends

To the greatest family in the universe and just as awesome friends. I'd never have arrived to this magnificent place called a PhD had it not been for you. Each smile. Each shared laugh. Each moment. Thank you for encouraging me to continually grow in both academic and personal ways, and for loving my quirks. Because of you, this life has the best journey a person could ask for.

Mécénat (Sponsorship)
Ecole Normale Supérieure

To my grad studies fellowship granting institution. We've all heard that saying "The best things in life are free". Sure, maybe. But basic physical subsistence, like food and lodging, is not. Thank you to the ENS for making these years possible.

***

Durée du spectacle (Length of the performance)
3 ans sans entracte (3 years without intermission)

TABLE OF CONTENTS

1. GENERAL INTRODUCTION

> Jane stepped off the asphalt and onto a dusty earth that was caught in limbo between sand and clay. She let her gaze drift softly across landscape. The quiet. The stillness. The… In the distance, she saw a tiny speck that stood out, ever so slightly, from the coffee-coloured ground. She focused her binoculars on the speck. It was a bamoule. She could not believe she was seeing a bamoule; it was the first time she'd seen one that wasn't on a picture or in a video. Suddenly, the bamoule pirdaled. It had pirdaled quickly, but to Jane everything had slowed down, each second its own eternity. She scanned a radius around the bamoule. Bamoules were one of the rare creatures to pirdale when faced with a predator. She couldn't believe her luck: a bamoule sighting and to top it all off, one that had pirdaled.

The very first time you read '*bamoule'* or '*pirdale'*, they likely stood out. You may have registered them as unknown, and you may have wondered what they meant. Yet, as you read on, the unknown began to dissipate and their meanings slowly took form. You could almost see the *bamoule* and how it *pirdaled*. There was no accompanying visual scene; there was no illustrative description of *bamoules* or *pirdaling*. Where then did these proto-meanings come from?

Intuitively, words appear to keep their meanings to themselves: words are the locus of each meaning, and combined, they convey an aggregate meaning. This corresponds quite well to the impression words give, but this impression may be just that: an impression. To discern whether meanings fit neatly within a word's boundaries, we can quickly look at three words in order.

Cat

That word, alone, may very well be sufficient for a reader to access a sort of dictionary meaning (e.g., a carnivorous mammal of the feline family, domesticated since antiquity).

Fortuitous

Dictionary aficionados apart, this word, in isolation, may not be sufficient for a reader to cite a meaning. The reader may instead feel a certain frustration, as if the meaning were at the tip of her tongue, as if she should be able to define it. She knows the word; she has encountered it before; but here and now, it seems to only be accompanied by a vague valence: it feels like it refers to something positive. One may however argue that this is a rare word, one few would be comfortable using themselves.

Engine

This word, despite being both common and frequent, may not be sufficient for a reader to muster anything more than a sketch of a meaning. As the reader scours her mind in a vain attempt to pinpoint a meaning, the familiarity of the word begins to dissolve into the blankness that

surrounds it; the word's meaning, a fading memory. Nonetheless, the same reader will have no trouble interpreting the following phrases:

The engine roared.
The engine was broken.
It is a well-oiled engine.

Of course, the reader may say, "But what is a meaning? What were the isolated words supposed to evoke? Ought I have a dictionary entry for each word? Could it be an image? A feeling?"

"This, precisely, is the question," the author would reply.

There is a clear distinction in the reader's mind between words a reader knows (e.g., cat, fortuitous, engine) and words she does not know (e.g., *bamoule*, *pirdale*). Yet, there is an inconspicuous underlying convergence between known and unknown words. When known words are presented in isolation, their precise meanings can escape us; and when unknown words are presented in sentences, they can appear meaningful. Word boundaries may actually be porous: a word's meaning may seep into co-occurring words and co-occurring words may infuse meaning into a word.

Thinking of meaning outside the frame of a word has allowed us to grasp how individuals may begin to accomplish the ostensibly insurmountable task of acquiring word meanings. Any language learner will have to acquire the links between words and their meanings to become a competent language user. Moreover, the learner will have to begin establishing links without an answer sheet. These links will have to be formed and verified on the set of evidence available in the learner's environment. However, upon scrutiny, it can seem that the links created are over-and-above what the evidence attests. The insufficiency of an evidence set to fix meanings, as we picture them, was famously exposed by Willard van Orman Quine as the '*gavagai* problem' (Quine, 1960). When a native speaker utters '*gavagai'* in the presence of a rabbit hopping across a field, a learner could derive a number of hypothesized meanings from the observable evidence: a rabbit hopping, a rabbit in a field, a quick rabbit, a whole rabbit, a set of undetached rabbit parts, and so on. The learner could go on to reduce the set of hypothesized meanings after multiple encounters with the utterance in slightly different visual contexts: *gavagai* uttered in the presence of a rabbit in a forest and then in the presence of a rabbit hiding in the bushes. However, it is highly unlikely that the learner could reduce the set of possible meanings down to just one: *gavagai* could always mean rabbit or set of undetached rabbit parts. In this example, the evidence set is composed of an utterance (*gavagai*) and co-occurring visual scenes (including the native speaker's reactions or emotions, Quine, 1960). If we take word meanings to extend beyond word boundaries, then we could argue that this evidence set is insufficient because it is just a subset of the available evidence.

A language learner will be highly unlikely to encounter just one utterance multiple times, and much more likely to encounter many different utterances, some of which may repeat. For instance, the learner may hear '*gavagai'* in a scene with a rabbit, '*gabibo'* in a scene with a chicken, and '*gadine'* in a scene with a horse. The learner may hypothesize that '*ga'* refers to something common to the three scenes (e.g., the presence of an animal, the presence of something cute, the act of eating) and the rest of the utterance to something unique about each scene (e.g., the specific animal). In other words, the learner will be able to reduce the possible meanings using the external visual context, and the linguistic context, adding a whole slew of extra evidence with which to work. If we consider word meanings to extend beyond word

boundaries, then these linguistic contexts can be more than just *extra* evidence; they can be particularly rich sources of information about a word's meaning (Gleitman, 1990). Better understanding meaning can thus allow us to better understand how word meanings can and may be acquired.

A devil's advocate would be quick to point out that the learner must first acquire linguistic contexts, without any linguistic context with which to scaffold acquisition. Specifically, learning must begin somewhere, with a constrained set of evidence.

Yet, the insufficiency of evidence may be a general problem for cognition, and not one specific to learners of language. A mind with finite access to evidence will have to simulate whether a set of evidence is representative of how the world is objectively. However, without omniscience, without access to all possible evidence, the mind cannot be one hundred percent certain; the simulation will thus need to be based on the likelihood that the world is so. Anything above a predefined threshold of likelihood will be considered representative of how the world is; otherwise, the mind would remain in a perpetual state of uncertainty. Similarly, evidence beyond a predetermined threshold or 'enough evidence' for a word's meaning (e.g., hearing '*gavagai'* in twenty scenes with rabbits; so '*gavagai'* likely means rabbit-like-thing) will be considered as representative of a word's meaning (e.g., '*gavagai'* means rabbit-like-thing).

In this dissertation, we argue that the evidence a learner has at hand may be insufficient to provide her with a complete, self-standing meaning for a word, the kind of meaning we have the impression we have; however, the very same evidence may be sufficient to provide the learner with the word-meaning link necessary for cognition. In other words, the defused, spread-out meaning words may actually have may indeed be acquirable from the kind of, seemingly insufficient, evidence that a learner has. More broadly, an evidence set is or is not sufficient for acquisition in respect to that which needs to be acquired. Whereas the brunt of research has focused on revealing overlooked sources of evidence and expanding the evidence set (e.g., linguistic context, Gleitman, 1990; communicative cues, Ferguson & Waxman, 2016; social cues, Tsuji, Mazuka, & Swingley, 2019), here we switch gears and tackle the kind of knowledge the evidence set supports. The gap between what we think we know about meanings and how meanings may actually be processed by the mind could reveal that we may aim to have sufficient evidence for something that is not a cognitive reality. Here, thus, we explore broadly what it is 'to know', from the point of view of the learner.

We begin the dissertation with an introductory section, in which we present evidence about how meanings are processed and acquired by the mind, and then frame the gap between what we know and what we think we know in the broader context of how minds make sense of an ever changing external environment.[4] Next, in the first chapter, we respond directly to the devil's advocate, by experimentally testing whether a constrained set of evidence (i.e., a handful of known words and a frequent novel grammatical context: '*Ga* rabbit plays. *Ga* rabbit is so cute.') can give rise to productive knowledge of linguistic contexts, or generalization (i.e., we test whether the learner can use this new grammatical context to infer the meanings of unknown words: '*Ga bamoule*'). To ensure ecological validity with how infants actually approach language learning, we design an experimental protocol that mimics grammatical context learning 'in the wild'. A set of pilot experiments for this first study revealed an unanticipated pattern of generalization across the lifespan, and motivated the series of experiments presented in the second chapter. In this chapter, we investigate experimentally whether the same evidence

[4] The introductory section serves to situate the questions addressed in the dissertation in the context of the literature. An in-depth literature review is spread over the body of the dissertation, in the introductions of each paper.

set (maximized as much as possible) gives rise to productive knowledge of linguistic contexts, and we interpret our results within the extant literature, pointing to the live possibility that what counts as knowledge depends on where an individual places her 'knowledge threshold'. Finally, we experimentally examine whether evidence that reflects a knowledge-state (e.g., explicitly hearing a direct translation: '*Bamoule'* means 'cat') is inherently more informative, or merely more appealing, for a learner. We advance the conclusion that pre-packaged knowledge-state evidence uniformly boosts confidence, but not performance. Across three chapters, we put forth a set of fundamental features about what it is to know: (1) a little evidence can go a long way, (2) how much evidence is enough may depend on a modifiable threshold, and (3) the mind may crave certainty. Broadly, we suggest that knowledge may depend on where one is in the learning process, rather than an objective ideal. Therefore, we propose that to better understand how we learn, we need to probe what it is to know from the point of view of someone who has yet to know (i.e., the learner) and not the point of view of someone who knows already (e.g., the competent language user).

1.1 Quantum meanings

How the mind processes and acquires meaning

At the beginning of the dissertation, the reader formed proto-meanings for the newly encountered words, *bamoule* and *pirdale*: *bamoule* is an animal that *pirdales*, and *pirdale* is an act of defense. Lacking a concordant visual scene and a detailed description, the reader had to rely on information from the linguistic context. Perhaps surprisingly, just one accompanying very general word can narrow down the possible meaning of a hitherto unknown word. If one hears 'a *bamoule'*, one knows that '*bamoule'* likely refers to an artefact, animal or abstract concept; it is a noun. If one hears 'they *pirdale'*, one knows that '*pirdale'* likely refers to an action; it is a verb. In this manuscript, the words *bamoule* and *pirdale* appeared for the first time in simple sentential contexts, 'It was a *bamoule'* and 'The *bamoule pirdaled'*, which allowed the reader, nonetheless, to extract fundamental information about their meanings. This information (i.e., 'bamoule' is a noun, 'pirdale' is a verb) then likely served as a foundation for the elaboration of a more precise meaning (e.g., 'bamoule' refers to an animal). The capacity to glean a word's proto-meaning from its linguistic context suggests that at least part of a word's meaning may lie beyond its boundaries and that, as such, these contextual clues can be an invaluable tool for acquiring word meanings.

Studies have found evidence for both of these aspects of the linguistic context, carriers of meaning and fuel for learning, on a cognitive level. When a native speaker of a language listens to a sentence, she can use the linguistic context to narrow down the set of possible upcoming words. For instance, if a native French speaker sees two images, one of a banana and one of a boat, and hears the sentence '*Oh, regarde la banane*' (Oh, look at the-*feminine* banana), she will be able to orient her gaze toward the banana even before she hears the word *banane* (Dahan, Swingley, Tanenhaus, & Magnuson, 2000; Van Heugten, Dahan, Johnson, & Christophe, 2012; Lew-Williams & Fernald, 2007). In French, nouns belong to one of two grammatical genders, feminine or masculine: banana is feminine, while boat is masculine. The determiner that precedes the noun is marked for grammatical gender (e.g., *la* is the feminine determiner, while *le* is the masculine determiner), and matched to the grammatical gender of the noun (e.g., *la banane* versus *le bateau*). Banana and boat, having different grammatical genders, are preceded by distinct determiners. Accordingly, when the French speaker hears '*la'*, she can be sure that the word that follows will be 'banana' (i.e., *la banane*) and not 'boat' (i.e.,

le bateau).[5] Native speakers, from toddlerhood to adulthood, have been shown to use contextual cues during online processing (Dahan et al, 2000; Van Heugten et al, 2012; Lew-Williams & Fernald, 2007). More globally, individuals interpret a word in a sentential context (e.g., This is a well-oiled engine) faster than an isolated word (e.g., engine; Lieberman, 1963). This is so even when the sentential context is seemingly neutral (e.g., Look at the engine, Fernald & Hurtado, 2006). The context thus appears to bear pertinent information for interpreting a word's meaning. Furthermore, the relation between a word and co-occurring words is not unidirectional, but rather reciprocal. Isolated words seem to contain information beyond the meaning itself: they also activate grammatical features, such as grammatical gender (e.g., in French, *banane* will activate the feature 'feminine'; Wang, Chen, & Schiller, 2019). The research seems to indicate that the point at which one word ends and another begins may not correspond to the point at which one meaning ends and another begins.

The very same indices that guide comprehension of a known word's meaning, can also foster a rudimentary understanding of a hitherto unknown word's meaning. In other words, determiners tend to precede nouns, be they known nouns or novel nouns, while pronouns tend to precede verbs, be they known or novel. If a learner has acquired the association between determiners and nouns, and pronouns and verbs, she can use this knowledge to fuel acquisition of novel words. Instead of learning each word's meaning from scratch, a learner can scaffold acquisition of novel words with elements she already knows (Gleitman, 1990). Notably, very frequent contexts, such as grammatical words (e.g., determiners: the, a; pronouns: she, they, it), provide reliable information about an upcoming word's meaning (Gleitman, 1990). Using the grammatical context to narrow down meanings of unknown words, also called 'syntactic bootstrapping', allows young language learners to extract a wealth of information about a word's meaning: whether a word is a noun or verb (He & Lidz, 2017); whether a verb is transitive or intransitive (Gertner & Fisher, 2012); fine-grained differences between semantically similar words such as 'look' and 'see' (Gleitman, 1990). Furthermore, and perhaps most importantly, this capacity is present early in development. Infants, as young as 18 months old, use syntactic bootstrapping to infer words' meanings (He & Lidz, 2017). For instance, when an 18-month-old infant hears 'a *bamoule'*, she can infer that '*bamoule'* refers to an object; and when she hears 'they *pirdale'*, she can deduce that '*pirdale'* refers to an action. This appears quite similar to how the reader may too have formed proto-meanings for '*bamoule'* and '*pirdale'*. Nevertheless, young children may have more coarse-grained interpretations of grammatical contexts than adults. Some studies have shown that young children begin with broad, simple interpretations, that are later fine-tuned (Gertner & Fisher, 2012). For instance, young children seem to initially consider the first noun in a sentence to be the subject and the second noun the object (Gertner & Fisher, 2012). They do not thus distinguish between two sentences such as 'The cat and the mouse *pirdale'* and 'The cat *pirdales* the mouse'. In other words, young children may either feel they have sufficient evidence to correctly interpret the grammatical contexts (but be wrong about it) or they may feel that they likely do not necessarily have sufficient information but likely may have enough information to make an educated guess. The literature provides substantial evidence that young children use syntactic bootstrapping to guide meaning acquisition, but less is known about how young children go about stitching together their syntactic bootstraps. Importantly, the necessary information needs to be available in and derivable from a constrained evidence set (both in content and in quantity). The first chapter of the dissertation thus investigates experimentally whether infants can make use of little evidence to make viable and productive hypotheses about grammatical contexts.

[5] Note that all gender-marked contexts do not seem to contribute equally to rapid identification of a target word. For instance, a listener will not orient their gaze to the banana based on the preceding context, if she hears '*des **grandes** bananes*' (the **big**-*feminine plural* bananas, Dahan et al, 2000).

1.2 A taste of omniscience

How a finite mind makes sense of an external word

A learner who scaffolds acquisition of word meanings using a grammatical context must first have identified the kind of information said context provides. The learner, not having access to all the existent evidence, will have to extract information from the evidence at hand and then weigh the probability that her interpretation is correct (see Goodman, 1955). In other words, a learner could potentially extract the structure '*ga* + animal' from three exemplars (e.g., '*gavagai'* in a scene with a rabbit, '*gabibo'* in a scene with a chicken, '*gadine'* in a scene with a horse), but at this point the posited structure may just be a hypothesis. The learner may need more examples or particularly informative examples to feel certain that her hypothesis reflects how the language actually is. A learner who generalizes early, risks guiding acquisition with incorrect information. For instance, if '*ga'* in reality means 'one', the learner may interpret a novel word '*gabamou'* as referring to the ducks in the scene rather than the solitary chair. On the other hand, a learner who waits too long to generalize may miss opportunities to fuel further learning. The learner must thus have a threshold at which she generalizes, at which she begins to use extracted information to guide learning. An accurately calibrated generalization threshold is thus essential for efficacious and efficient learning.

Generalization is, nonetheless, a capacity that extends beyond learning; it allows an individual to act in an ever-changing world. Subjectively, we may feel that few of our interactions with our environment are truly novel. The moments of our lives appear to be composed of ostensibly similar situations. In western industrialized settings these may include morning commutes, coffee breaks, answering emails, conference calls, restaurants, pigeons. Objectively, however, it is highly unlikely that an individual will encounter the exact same situation twice. A mouse may be in the garden, another in the metro, one may be tiny, another swift. Yet, somehow, all of them will be interpreted, with little second thought, as mice. This capacity to generalize knowledge from one situation to another, to interpret different situations as belonging to the same overarching kind of situation, is primordial for any interaction between an individual and her external environment (Shepard, 1987). The capacity to group situations via an internal metric of similarity (e.g., this is like a mouse, it is a mouse; that is not like a mouse, it is a hedgehog), to generalize, was considered so fundamental and pervasive that it was even proposed in 1987 as a universal law of psychology (Shepard, 1987). It is generalization that allows an individual to interpret a hitherto unexperienced situation as if it were already known. When an individual sees Tom and Jerry on the screen for the first time, she will immediately interpret the animated characters as a cat and a mouse. Visually and conceptually however they are quite different from real cats and mice. Tom sings, but cats do not; Jerry runs on his two hind legs, mice do not. At no point however does the individual think to herself, these two creatures are just somewhat like cats and mice.[6] Via the capacity to generalize, an individual can side-step processing all the evidence to know how the world *is*: the individual just needs a metric of similarity for a set of evidence at hand to predict how the world *may be*. Generalization may thus guide mind-world interaction broadly, whether it be in highly similar situations or novel learning situations (for a great discussion, see Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018).

[6] Note that this does not entail that the individual lacks the capacity to distinguish the difference in similarity between a field mouse and wood mouse (small), and a field mouse and Jerry (medium), and a field mouse and Mickey Mouse (large). Rather, generalization allows these differences to appear smaller (or less important for 'mouse-ness') than across situation kinds (e.g., a field mouse and a hedgehog, Shepard, 1987).

Intuitively, we would expect an individual to begin to generalize when she has enough evidence. In accordance with our intuitions, individuals have been shown to generalize more readily when the quantity and diversity of evidence is greater (e.g., Quinn & Bhatt, 2005; Gerken, Balcomb, & Minton, 2011). Simple, Bayesian ideal observer computational models, which extract abstract structures and weight likelihoods, can account for a number of generalization results in the experimental literature, but they run short of accounting for them all (e.g., Frank and Tenenbaum (2011)'s model for Gervain & Endress (2017)'s experimental data). Other than the evidence set itself, one of the most common factors that seems to affect rate of generalization is age. A number of studies reveal that younger learners will generalize novel patterns that peers just a couple months older do not (e.g., 7.5-month-olds versus 9-month-olds, Gerken & Bollt, 2008; see also Gerken et al, 2011). In contrast, a number of other studies reveal that adults generalize more readily than children (e.g., 7-11-year-olds versus adults, Schulz, Wu, Ruggeri, & Meder, 2019; 4-6-year-olds versus adults, Deng & Sloutsky, 2015). The source behind these developmental differences is often attributed to a bias stemming from prior knowledge or cognitive maturity (e.g., Gerken & Bollt, 2008; Deng & Sloutsky, 2015; modeled by Xu & Tenenbaum, 2007). However, these biases, though justifiable, can appear to be attributed on an *ad hoc* basis: when the younger generalize more than the older, it is because they have less prior knowledge, and when the older generalize more than the younger, it is because they have a mature cognitive system (for a detailed discussion about *ad hoc* parameters, see Endress, 2013). Most developmental studies looking at generalization focus narrowly on the difference in generalization between children of different ages or narrowly on the difference in generalization between children (as a category ranging from infancy to adolescence) and adults: few studies have investigated generalization of novel patterns on a lifespan scale, starting in infancy and going through to adulthood.[7] To investigate modulations in generalization rate over development, in the second chapter, we present infants, pre-school children and adults with the same set of evidence, in a paradigm intuitive[8] for all ages. We then interpret our results within the broader scope of the generalization literature to pinpoint the possible source of the observed differences in generalization rate: the incapacity to extract an abstract structure, a lower internal likelihood that the extracted structure represents the world, or a higher generalization threshold.

1.3 Learning in a world of knowledge
The gap between intuition and cognition

An individual may be able to identify a mouse without having introspective access to the metric of similarity that determines 'mouse-ness'. In other words, the individual may be able to generalize correctly without being able to conceptualize the essential feature or features necessary for generalization. This is quite similar to the reader's experience with known words such as 'cat', 'fortuitous', and 'engine': the reader has no problem using the words even though she cannot muster a precise meaning. Broadly, individuals may lack full introspective access to the nuts and bolts of their knowledge. This may explain why an individual feels she knows the precise meaning for each word, but cannot say what it is out loud; however, it does not explain why word meanings appear condensed between word boundaries.

[7] Adulthood here refers to early adulthood (18-45). For some generalization differences in early and late adulthood, see: Mutter & Plumlee, 2014; Worthy, Gorlick, Pacheco, Schnyer, & Maddox, 2011.
[8] As intuitive as possible.

At first glance, word meanings may appear to fit neatly between word boundaries simply because words need to be, and are, manipulated, inserted into any plethora of possible sentences, and from time to time uttered in isolation (e.g., 'Look!'). Yet, it has been shown that competent speakers, who can use and understand words correctly, do not automatically consider words as isolable, self-standing entities (Havron & Arnon, 2017a; Havron & Arnon, 2017b; for a review, see: Morais & Kolinsky, 2002). This capacity appears to develop when a child begins to read (Havron & Arnon, 2017a) and remains largely undeveloped in illiterate adults (Havron & Arnon, 2017b). A pre-literate child, for example, can understand and use the word 'up' correctly in a number of different contexts (e.g., 'The balloon is up in the sky', 'She went up the stairs'), but she will have trouble extracting the word from frequent contexts (e.g. 'Look up', Havron & Arnon, 2017b). In other words, meanings may appear concentrated within word boundaries because the capacity to read (and write) requires explicitly extracting and isolating words. This impression may thus not be an accurate reflection of how languages, and meaning, are intrinsically.

A skewed perception of word meanings can have an impact on second language teaching methods, insofar as teachers and policy makers are often literate adults. Teaching word meanings as confined to word boundaries may equip learners with incomplete or incorrect knowledge, or it may equip the learner with ungraspable knowledge. A telltale example is direct translation of words from one language to another (e.g., in French, '*souris'* means mouse). Translation presupposes that word meanings in the two languages have the same scope, and moreover offers little indication of how the word ought to be used.[9] A native English speaker learning French may know that '*ronronner'* is 'to purr', but she will not know that in French the sound of purring is not exclusive to cats (e.g., engines can purr). However, the same native English speaker learning French, who instead hears a set of sentences such as '*Le chat ronronne*' ('The cat purrs') and '*Le moteur ronronne*' ('The engine hums'), will begin to associate '*ronronne'* with '*chat'* and '*moteur'* (perhaps even before knowing what those two words mean). In other words, she will acquire the very contexts that may carry some of a word's meaning. Therefore, a learner who acquires word meanings by way of direct translation may acquire partial (or, in extreme cases, incorrect) meanings. If that same native English speaker is a pre-literate child, she may have difficulty extracting '*ronronne'* from associated contexts and associating it with a unique meaning. She may thus be at a loss as to what 'to purr' might signify. Accordingly, a pre-literate learner who acquires meanings via translation may not be able to benefit from the information provided. There are some studies that show that children (literate, 6-12 years old) tend to learn word meanings better when they are presented with corresponding pictures (e.g., '*souris'* and a picture of a mouse) than when they are presented with translations (e.g., '*souris'* means mouse; Emirmustafaoğlu & Gökmen, 2015), and that adults may learn best with a combination of the two (e.g., '*souris'* means mouse and a picture of a mouse; Plass, Chun, Mayer, & Leutner, 1998; Kost, Foss, & Lenzini Jr, 1999; but *cf*, Lotto & De Groot, 1998; Morimoto & Loewen, 2007). However, the effectiveness of translation is highly dependent on what is measured (e.g., the capacity to provide translations, the capacity to pick out pictures, the capacity to use the word in a sentence; Prince, 1996). Most of the studies, however, compare two or more supervised learning methods (e.g., direct translation or picture naming). Each method provides the learner with what we call 'knowledge-state' information (e.g., 'cat' means this), and thus with the kinds of meaning we have the impression of possessing. In the third chapter, we thus probe why we choose to learn the way we do, why we opt for supervised learning methods that provide us with potentially imprecise knowledge-states (such as translation). To determine what is appealing about knowledge-state information,

[9] 'Word use' is canonically taught via grammatical rules (e.g., '*la'* is the article used before feminine nouns).

we directly compare the effect of supervised and unsupervised[10] meaning acquisition on performance and preference, in three kinds of learners, adults, literate first-grade children, and pre-literate first-grade children.

***

Over three chapters, we examine what it is to know, from the standpoint of the learner. We begin by proposing that sufficient evidence for learning need not be sufficient for knowing in all certainty. Next, we suggest that knowledge thresholds may depend on where one is in the learning process. Finally, we propose that finite minds may, inherently, seek knowledge-states. We advance the conclusion that 'knowledge', far from an omniscient ideal, may be a tractable standard, which allows a mind to make sense of the 'blooming buzzing confusion' (James, 1890) that is the external world.

References

Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. S. (2000). Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language*, *42*(4), 465-480.

Deng, W. S., & Sloutsky, V. M. (2015). The development of categorization: Effects of classification and inference training on category representation. *Developmental Psychology*, *51*(3), 392.

Emirmustafaoğlu, A., & Gökmen, D. U. (2015). The effects of picture vs. translation mediated instruction on L2 vocabulary learning. *Procedia-Social and Behavioral Sciences*, *199*, 357-362.

Endress, A. D. (2013). Bayesian learning and the psychology of rule induction. *Cognition*, *127*(2), 159-176.

Ferguson, B., & Waxman, S. R. (2016). What the [beep]? Six-month-olds link novel communicative signals to meaning. *Cognition*, *146*, 185-189.

Fernald, A., & Hurtado, N. (2006). Names in frames: Infants interpret words in sentence frames faster than words in isolation. *Developmental Science*, *9*(3), F33-F40.

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*(3), 360-371.

Gerken, L., Balcomb, F. K., & Minton, J. L. (2011). Infants avoid 'labouring in vain' by attending more to learnable than unlearnable linguistic patterns. *Developmental Science*, *14*(5), 972-979.

Gerken, L., & Bollt, A. (2008). Three exemplars allow at least some linguistic generalizations: Implications for generalization mechanisms and constraints. *Language Learning and Development*, *4*(3), 228-248.

[10] Such as, hearing sentences and seeing multiple possible referents.

Gervain, J., & Endress, A. D. (2017). Learning multiple rules simultaneously: Affixes are more salient than reduplications. *Memory & Cognition*, *45*(3), 508-527.

Gertner, Y., & Fisher, C. (2012). Predicted errors in children's early sentence comprehension. *Cognition*, *124*(1), 85-94.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3-55.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Cambridge, Massachusetts: Harvard University Press.

Havron, N., & Arnon, I. (2017a). Minding the gaps: literacy enhances lexical segmentation in children learning to read. *Journal of Child Language*, *44*(6), 1516-1538.

Havron, N., & Arnon, I. (2017b). Reading between the words: The effect of literacy on second language lexical segmentation. *Applied Psycholinguistics*, *38*(1), 127-153.

He, A. X., & Lidz, J. (2017). Verb learning in 14-and 18-month-old English-learning infants. *Language Learning and Development*, *13*(3), 335-356.

James, W. (1890). *The Principles of Psychology* (Vol. 1). Cambridge, MA: Harvard University Press.

Kost, C. R., Foss, P., & Lenzini Jr, J. J. (1999). Textual and pictorial glosses: Effectiveness on incidental vocabulary growth when reading in a foreign language. *Foreign Language Annals*, *32*(1), 89-97.

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, *18*(3), 193-198.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, *6*(3), 172-187.

Lotto, L., & De Groot, A. M. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning*, *48*(1), 31-69.

Morais, J., & Kolinsky, R. (2002). Literacy effects on language and cognition. In L. Bäckman & C. von Hofsten (Eds.), *Psychology at the turn of the millennium, Vol. 1. Cognitive, biological, and health perspectives* (pp. 507-530). Hove, England: Psychology Press/Taylor & Francis (UK).

Morimoto, S., & Loewen, S. (2007). A comparison of the effects of image-schema-based instruction and translation-based instruction on the acquisition of L2 polysemous words. *Language Teaching Research*, *11*(3), 347-372.

Mutter, S. A., & Plumlee, L. F. (2014). The effects of age on associative and rule-based causal learning and generalization. *Psychology and Aging*, *29*(2), 173.

Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (1998). Supporting visual and verbal learning preferences in a second-language multimedia learning environment. *Journal of Educational Psychology*, *90*(1), 25.

Prince, P. (1996). Second language vocabulary learning: The role of context versus translations as a function of proficiency. *The Modern Language Journal*, *80*(4), 478-493.

Quine, W. V. O. (1960). *Word and Object*. Cambridge, Massachusetts: MIT press.

Quinn, P. C., & Bhatt, R. S. (2005). Learning perceptual organization in infancy. *Psychological Science*, *16*(7), 511-515.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323.

Schulz, E., Wu, C. M., Ruggeri, A., & Meder, B. (2019). Searching for rewards like a child means less generalization and more directed exploration. *BioRxiv*, 327593.

Tsuji, S.,Mazuka, R., & Swingley, D. (2019). Temporal contingency augments attention to a referent in a word learning task. To appear in the *Proceedings of the 43th Boston University Conference on Language Development*, Boston, MA, USA.

Van Heugten, M., Dahan, D., Johnson, E. K., & Christophe, A. (2012). Accommodating syntactic violations during online speech perception. In *Poster presented at the 25th Annual CUNY Conference on Human Sentence Processing, New York, NY*.

Wang, M., Chen, Y., & Schiller, N. O. (2019). Lexico-syntactic features are activated but not selected in bare noun production: Electrophysiological evidence from overt picture naming. *Cortex*, *116*, 294-307.

Worthy, D. A., Gorlick, M. A., Pacheco, J. L., Schnyer, D. M., & Maddox, W. T. (2011). With age comes wisdom: Decision making in younger and older adults. *Psychological Science*, *22*(11), 1375-1380.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245.

2. EXPONENTIAL LEARNING

Infants have to begin making sense of the external world with a small set of evidence. This set of evidence serves as a source of information about the world. If an infant's evidence set contains the utterances –'The dog', 'The fish', 'The rabbit'–, she may surmise that 'the' precedes animals. However, the infant will not have access to all the evidence, and thus will not know for sure if her hypothesis is correct (in this example, it is not). She will thus have to weigh the likelihood that her hypothesis represents the world, based on the same small evidence set that gave rise to her hypothesis. When the infant has decided that her hypothesis likely represents the world, she frees up cognitive resources for further learning, and moreover can scaffold learning with this new knowledge. An infant who has deduced that 'the' precedes nouns, will be able to use this information to fuel acquisition of yet to be learned words: she will be able to narrow down the meanings of words preceded by 'the' (e.g., 'the *bamoule'*) to objects. Therefore, for an infant, any knowledge will be invaluable, in so far as it is a potent catalyst for learning: rather than relying uniquely on the evidence set at hand, the infant can bolster learning with derived knowledge. In a pre-registered study, we present infants with a novel grammatical distinction ('*ko'* precedes animal, and '*ka'* objects) in their native language, French. We investigate whether infants can deduce the uses of these grammatical elements and propel learning with them, from a small set of evidence (just 6 animals and 6 objects). Our results reveal that infants accelerate native language acquisition using knowledge derived from scanty evidence. Learning may thus be supported by a half-baked knowledge, rather than a formal step-by-step progression from indubitable knowledge about one element to indubitable knowledge of another.

2.1 Infants quickly use newly-learned grammar to guide acquisition of novel words

Barbir, M., Babineau, M., Fiévet, A.-C., & Christophe, A.

One-sentence summary.
Knowledge of just a handful of vocabulary items spurs a virtuous circle of concordant grammar and vocabulary acquisition.

Abstract.
Infants use grammar to narrow down the set of potential meanings for a novel word, but how do they learn that grammatical elements co-occur with certain words but not others? We investigated whether a novel grammatical element, inserted alongside some common words infants already know, could be later used to guide interpretation of novel words. After watching a short video in which a novel grammatical element was introduced alongside previously known words (e.g., *ko* rabbit, *ka* book), infants correctly categorised novel words using the newly-acquired grammatical element (*ko* <u>*bamoule*</u>, therefore <u>*bamoule*</u> = animate). These results demonstrate infants' incredible capacity to fuel acquisition, by using what little they know to lay the structural foundations for future language learning.

***

Grammar reliably delimits the possible meanings of neighbouring words: for instance, determiners are most often followed by nouns (which tend to refer to objects), while pronouns by verbs (which refer to actions). For a young learner faced with a slew of unknown words and a visual scene teeming with possible referents, grammar can be a valuable tool with which to zoom-in on the potential meaning of a word. If a learner hears 'the *bamoule'*, and knows that 'the' precedes nouns, then she can subset the possible meanings of '*bamoule'* to only objects; on the other hand, if the learner hears 'I *bamoule'*, and knows that 'I' precedes actions, she can subset the possible meanings to actions.

The theory that infants use grammar to guide acquisition of word meanings is called 'syntactic bootstrapping' (*1*). Recent research shows that infants as young as 18 months old use syntactic bootstrapping during word learning (*2, 3*). For example, infants can base themselves on the presence of determiners (e.g., 'the', 'a') or pronouns (e.g., 'she', 'he') to infer that what follows is a noun or a verb, respectively (*4*). That grammar is both useful and used by infants to learn vocabulary is widely accepted, but the way in which infants go about harnessing the sheer potency of grammatical structures remains a matter of debate.

In order to actively exploit the clues grammar provides about neighbouring words, infants must first roughly categorise grammatical contexts (e.g., 'the' + object, 'she' + action). An ostensibly straightforward way to achieve this task would be to use known words. After hearing, '<u>the</u> bottle', '<u>the</u> baby', '<u>the</u> teddy', but never '<u>the</u> eat*' or '<u>the</u> sleep*', an infant could deduce that 'the' is paired with objects; likewise, after hearing, '<u>you</u> sleep', '<u>you</u> eat', but never '<u>you</u> bottle*' or '<u>you</u> teddy*', and infant could deduce that 'you' is paired with actions. In sum, infants would scaffold grammar acquisition with vocabulary and vocabulary acquisition with grammar. To overcome the blatant circularity of this proposition, either vocabulary or grammar would have to be learnable, at least to a certain extent, without the other.

Concrete nouns have been shown to be a subset of vocabulary that is a particularly favourable candidate for acquisition without the aid of grammar (*5*). Moreover, these words are most likely to be used when the referred-to object is present in the child's environment (*6*), potentially facilitating word-referent associations. Indeed, at as early as 6 months old, infants can identify a handful of concrete nouns, such as, 'baby', 'bottle', 'foot' (*7, 8*). Nevertheless, some more abstract words are also acquired early in infancy, though a bit later on: at 10 months old, infants know a handful of verbs, such as, 'eat', 'sleep' (*6, 9*). Though it is unclear how akin infant knowledge of these words is to full-fledged adult knowledge, there does appear to be a certain level of referential precision: young infants can correctly identify the referents even when the foils are from the same overarching category (e.g. milk vs. juice, *8*). Thus, infants appear to be capable of acquiring some vocabulary without much grammar knowledge.

Likewise, rudimentary grammatical knowledge emerges well before the age of one, before infants have access to a substantial vocabulary. Early in life, infants can differentiate function words, words that serve a structural, grammatical purpose (e.g., determiners and pronouns) and lexical words, with a meaning-conveying purpose (e.g., nouns and verbs) (*10*). Refinement of grammatical knowledge makes great strides during the second year of life. At 14 months, infants expect nouns to be preceded by determiners (e.g., 'the bottle', 'your bottle'), manifesting surprise if they are preceded by pronouns (e.g., 'I bottle'*, 'you bottle'*), but they do not yet have clear expectations for verbs (*11*). Then, some months later, at 18 months, infants know that words preceded by pronouns are likely to be action-referring words (i.e., verbs) and words preceded by determiners are likely to be objects (i.e., nouns, *12*). Therefore, infants seem to be able to glean some information about grammar without a big vocabulary.

These findings suggest that a fledgling vocabulary and grammar emerge presumably without much aid from the other. It is unclear, however, whether, and more importantly, how, such a constrained quantity of knowledge can be sufficient to establish bootstrapping tools. Recently, computational modeling studies have aimed at determining how infants could categorize grammatical contexts with limited linguistic knowledge. Algorithms were found to most easily categorise those contexts when they were provided with a *semantic seed* –that is, knowledge of a few common words (e.g., bottle, teddy, etc.)– and when they could memorize the *frequent contexts* associated with these words (e.g., the bottle, the teddy, the baby, therefore 'the X', X=object) (*13, 14, 15*). Thus, if they are anything like algorithms, infants would simply need to have rudimentary knowledge of a small set of words and of the contexts in which those words appear. They would thus not need a sophisticated vocabulary or complex structural knowledge to be able to derive the uses of grammatical contexts: a seedling of each may be a good enough scaffold. Here, we, therefore, asked whether infants can guide acquisition of grammatical contexts by relying on a small semantic seed, and in particular, whether this acquisition is robust enough to allow for these novel grammatical contexts to soon be used to learn new words.

We chose a categorical distinction that is marked grammatically in the world's languages but not in the native language of the infants we tested, French, and that is salient for infants: animate versus inanimate (for review, *16*). Very early in life, infants demonstrate the capacity to distinguish animates from inanimates, and just a little later on, the ability to use this categorization during language acquisition. From the age of 9 months, babies can distinguish different species of animals, from objects, such as cooking utensils, furniture, electronic appliances (e.g., *17, 18*). During their first year of life, from the age of 19-months, infants can deduce the animacy of a noun, based on the semantic properties, or roughly the meaning, of a co-occurring verb. When an infant hears "The dax is crying", knowing that only animate nouns

can co-occur with certain verbs like 'cry', she will associate the novel word 'dax' with an animate (*19*). Infants can also learn novel category-level words for animals versus vehicles in the lab, as early as 22 months old (*20, 21*). Animacy is thus a generally salient category for infants and also more specifically, it is actively used in language learning tasks.

In our study, we presented infants with two novel determiners, '*ko*' for animates and '*ka*' for inanimates. To be certain that infants would interpret our novel determiners as grammatical elements and not lexical, meaning-bearing words, we inserted them into full French sentences, replacing existent French determiners ('*le'*, '*la'*, '*un'*, '*une'*: 'the' and 'a'). The novel determiners were paired with a semantic seed of 12 common nouns –6 animate and 6 inanimate– that infants were likely to know (e.g., 'book', 'bottle', 'rabbit', 'dog'; Fig. 1A).[11] The novel determiners functioned on a structural level just like French determiners; for instance, when an adjective modified the noun, it was inserted in the usual place in French, between the determiner and the noun (e.g., ***le*** *grand chien* = ***ko*** *grand chien*, **the** big dog). These determiner-noun pairings were presented to the infants in an engaging and ecologically valid situation: infants watched a short video in which a woman acts out stories with toys (Fig. 1A; for the script, see S1, Annex S1).

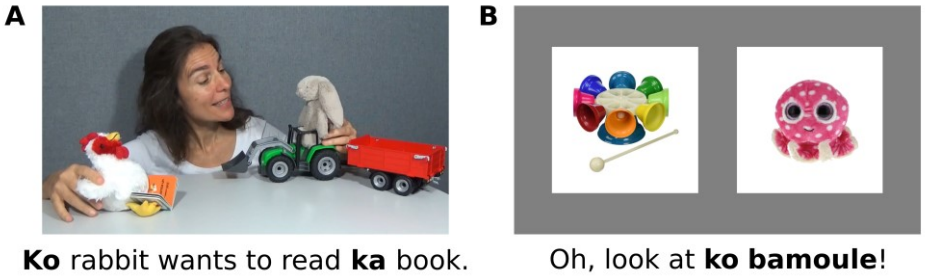


Fig. 1. Schematic of study design. (**A**) Infants first watched a training video with novel determiners (*ko* and *ka*), three times at home and once in the lab. (**B**) Then, they completed a test phase, during which they saw two novel images (one animate and one inanimate) and heard a prompting sentence to look at one of the two images The prompting sentence contained a novel word infants did not know (e.g., *bamoule*). The only way to solve the task was to use knowledge about the associated novel determiners.

While it is pragmatically sound to teach a novel word in just one laboratory visit, as there may be a new object, action or concept that one of any age may just not have encountered before, it is not clearly so when teaching a novel grammatical element. Grammatical elements, especially determiners, are amongst the most frequent items across languages (e.g., English, *22*; Japanese and Italian, *23*), and any new grammatical elements ought to embody this aspect too. Infants thus watched the training video at home for three consecutive days before the laboratory visit, and once more in the lab. This procedure allowed for increased total time exposed to the novel determiners (~40minutes) and an increased period of time during which these determiners had been used (4 days), allowing for further processing of, or simply familiarization with, the use of these linguistic elements. Additionally, infants slept between each viewing of the video, and this may have been beneficial for encoding the grammatical distinction: it has been shown that sleep boosts grammatical rule learning in both infants and adults (e.g., *24, 25*).

At the lab, infants watched the training video and then proceeded to the test phase. During the test phase, infants were presented with novel items and pseudo-words. They saw

[11] Based on previous MacArthur-Bates Communicative Development Inventory data gathered in our lab for French learning infants.

two images on the screen, a novel animate and a novel inanimate, and heard a prompting sentence to look at one of the two images (e.g., Oh look at *ko bamoule*! Fig. 1B). The novel items were original toys that infants were unfamiliar with, and thus did not have a French label for. The novel words were paired with a novel determiner. If infants had categorised the novel determiners during training (*ko*+animate and *ka*+inanimate), they could use this categorisation to narrow down the potential meaning of the novel words (e.g., *ko*+*bamoule*, *bamoule* = animate).

Infants' gaze was recorded with an eye-tracker, and served as an index of infants' interpretation of the novel nouns. Gaze was measured in two ways: the fine-grained 'looking-while-listening' method which encodes the evolution of gaze patterns from moment to moment during the trial, and the more broad 'preferential looking' method which sums overall gaze time to a target during the trial (*26*). These two methods allow for a complementary and comprehensive analysis, with both specific information as to when precisely infants orient their gaze to the correct image, as well as a way to capture general preference in light of individual variation (e.g., some infants might look right away to the target image, others may take more time and only do so toward the end of the trial).

Specifically, we analyzed the proportion of gaze to the animate image, when infants had heard a prompting sentence with *ko* versus one with *ka* (Fig. 1B). If infants had categorised and were able to use the novel determiners to narrow down the meanings of words they do not know, they ought to show distinct gaze patterns for novel nouns paired with *ko* and for those paired with *ka*. A cluster-based permutation analysis (*27*) revealed that infants looked more to the animate image when they heard *ko* than when they had heard *ka* during the time-window from 300 to 2440ms after hearing the novel determiners ($p = 0.01$, Fig. 2A). A mixed-effects regression analysis confirmed that infants looked longer overall, throughout the whole trial, to the animate image when they had heard *ko* than when they had heard *ka*: $\beta = -0.09$, $SE = 0.04$, $t = -2.075$, $p = 0.049$, Cohen's $d = 0.53$ (means are displayed in Fig. 2B). These results suggest that after little exposure infants are able to categorize novel grammatical elements when paired with a handful of known nouns and just as quickly use these novel grammatical elements to constrain the set of potential meanings of unknown lexical words.

The capacity to quickly stitch together bootstraps indicates that syntactic bootstrapping is likely an even more powerful tool than previously demonstrated. Research on syntactic bootstrapping has unveiled infants' and toddlers' grammatical knowledge to be surprisingly sophisticated and fine-grained (*28, 29, 30*), though not always reaching adult level (*31*). Even though infants can make use of coarse-grained knowledge to do syntactic bootstrapping, for instance, interpreting agents and patients as the nouns that appear first and second in a sentence respectively, how quickly they could stitch together provisional syntactic bootstraps was largely unknown. Our results point to the live possibility that even very rudimentary grammatical knowledge, based on on-the-fly categorizations, can spark syntactic bootstrapping. As such, syntactic bootstrapping can be a highly potent tool, particularly early in life, when knowledge resources are few. Further research is needed to determine how much evidence and resources are enough. In our experiment, infants watched the training videos three times at home, but just one time at home or just once in the lab may actually have sufficed. There may even be something along the lines of *one-shot* grammar learning, where incredibly little exposure leads to syntactic bootstrapping. Such evidence would add to findings from studies investigating generalization of grammatical rules in artificial languages, which have found young infants to generalize from as little as 2-minutes of exposure and from as few as just one exemplar (*32, 33*).

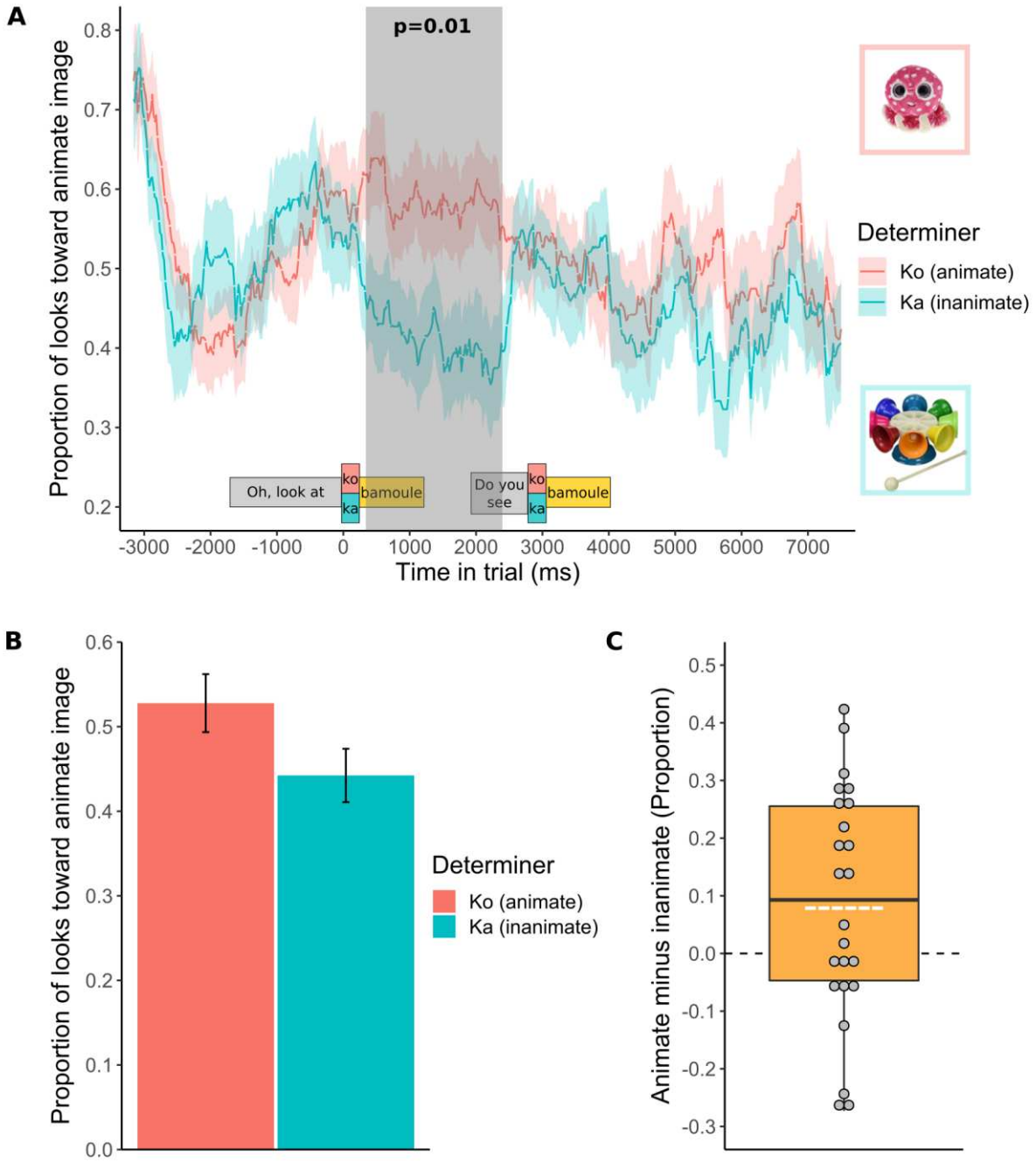


Fig. 2. Results. (A) Time-course. Proportion of looks toward the animate image at each point in time during the trial, when the infants heard the animate determiner *ko* paired with a novel noun (e.g. *ko bamoule*) in pink and when they heard the inanimate determiner *ka* (e.g., *ka pirdale*) in blue. Dark lines represent mean across participants, and light shading the SEM (95% confidence intervals of the mean). Grey shading indicates the time-window during which the two conditions, animate determiner sentences versus inanimate determiner sentences, diverge (300-2440ms, p = 0.01, cluster-based permutation analysis). **(B) Overall looking preference.** Proportion of looks toward animate image averaged over the whole trial, when the infants heard the animate determiner *ko* paired with a novel noun in pink and when then heard the inanimate determiner *ka* in blue. Error bars represent SEM (95% confidence intervals of the mean). Infants look significantly longer to the animate image when they heard the determiner *ko* than *ka* (p = 0.049). (**C**) **Difference per participant.** The difference between the proportion of looks toward the animate image when the animate or inanimate determiner was heard, per participant. Dots indicate participants. Dashed white line indicates mean. Upper and lower regions of the box indicate the first and third quartiles (25^{th} to 75^{th} percentiles). The upper whisker represents the third quartile up to the 1.5 interquartile smallest value, while the lower whisker the 1.5 interquartile smallest value to the first quartile. Dashed line black indicates no difference between proportion of looks the animate image, when infants heard the animate or inanimate determiner.

Beyond the clear advantage of harnessing incomplete grammatical knowledge to further vocabulary learning, partial grammatical knowledge may too be catalyst for further grammatical learning. Artificial language learning studies have demonstrated that experience with one kind of grammar rule, well before the rule is acquired, speeds up learning of analogous more complicated grammatical rules (*34*). A clever way to investigate this idea would be to compare acquisition of our novel determiners by infants learning a language that has just one overarching determiner, like English (e.g., English = 'the' versus French = '*le'* or '*la'* depending on whether the word is masculine or feminine). Having a determiner system in which there is a distinction may propel learning of an analogous system, with French-learning infants acquiring the distinction faster than English-learning infants.

Yet, the kind of categorisation may also play a role in ease of learning. The categorical distinction between animates and inanimates may have been particularly salient, and thus speedily remarked. The animate and inanimate categories are even considered to be part of 'core knowledge', the basic foundational cognitive categories (*35*). Grammatical distinctions based on non-core knowledge categories, or just less salient categories, may take longer to acquire (e.g., based on size: small or big, *36*). Additionally, age of the learner may interact with grammatical category acquisition. Previous studies on grammatical rule-learning in artificial languages have found that some rules are learned more readily early in life (*37*). 20 months may have been an age at which the grammatical distinction animate-inanimate is easily learnable. Preliminary work suggests that 3-year-olds may be less adept at this task than their younger peers (see SI: Supplemental Experiments 1-3). Further work will be required to determine the specificities of this impressive capacity.

In our experiment, the rule governing the use of the novel determiners could only be discovered alongside known words, and just 6 exemplars of each category at that. In stark contrast to algorithms that had been trained with a semantic seed, infants were learning in an entirely unsupervised manner: they were not told 'rabbit = animal' and 'car = object' (*14, 15*). They had to figure out the dimension on which these items differed that corresponded to the novel determiners. The algorithm, however, had been explicitly told that each of its semantic seeds was either a noun or a verb, and did not have to figure out for itself how to categorise the seed (*14, 15*). Whether this, or another, algorithm would have been as successful if it had been presented with sets of possible classifications of each seed (cat = noun, animal, furry, small) is difficult to tell. There is however evidence that algorithms can categorise words based on the kinds of sentences in which they appear (*38*). Thus, infants may form sensorial classifications hand-in-hand with linguistic classifications. It is nonetheless astounding that despite the numerous possible classifications, infants were capable of zooming into the right distinguishing feature quickly.

These results go a step beyond investigations of how grammar can bootstrap vocabulary acquisition or how vocabulary can bootstrap grammar acquisition, and demonstrate the sheer capacity infants have to use any scraps of information as a bootstrap. They push us to rethink too how infant language learning experiments are done: perhaps testing just the effect of one tiny element may be so unnatural that an effect may be missed and knowledge not attributed until a much later age. Broader more ecologically valid testing methods may reveal knowledge much earlier than expected. Such testing methods may too expose a highly intertwined network of bootstraps (see *39* for how words can, independently, help acquisition of phonology, vocabulary and grammar). The next step in infant language acquisition research will be to understand these networks, and discover how the elements infants use interact to fuel language learning.

References

1. L. Gleitman. The structural sources of verb meanings. *Lang. acquisition* ***1***, 3-55 (1990).
2. S. Yuan, C. Fisher. "Really? She blicked the baby?" Two-year-olds learn combinatorial facts about verbs by listening. *Psych. Sci.* ***20***, 619-626 (2009).
3. A. X. He, J. Lidz. Verb learning in 14-and 18-month-old English-learning infants. *Lang. Learning and Dev.* ***13***, 335-356 (2017).
4. S. Bernal, J. Lidz, S. Millotte, A. Christophe. Syntax constrains the acquisition of verb meaning. *Lang. Learning and Dev.* ***3***, 325-341 (2007).
5. J. Gillette, H. Gleitman, L. Gleitman, A. Lederer. Human simulations of vocabulary learning. *Cognition*, ***73***, 135-176 (1999).
6. E. Bergelson, D. Swingley. The acquisition of abstract words by young infants. *Cognition* ***127***, 391-397 (2013).
7. R. Tincoff, P. W. Jusczyk. Six-month-olds comprehend words that refer to parts of the body. *Infancy* ***17***, 432-444 (2012).
8. E. Bergelson, D. Swingley. At 6–9 months, human infants know the meanings of many common nouns. *Proc. of the Natl. Acad. of Sci.* ***109***, 3253-3258 (2012).
9. S. M. Pruden, K. Hirsh-Pasek, M. Maguire, M. Meyer. Foundations of verb learning: Infants categorize path and manner in motion events. *Proceedings of the 28th annual Boston University conference on language development* ***2004***, 461-472.
10. R. Shi, M. Lepage. The effect of functional morphemes on word segmentation in preverbal infants. *Dev. Sci.* ***11***, 407-413 (2008).
11. R. Shi, A. Melançon. Syntactic categorization in French-learning infants. *Infancy* ***15***, 517-533 (2010).
12. E. Cauvet, R. Limissuri, S. Millotte, K. Skoruppa, D. Cabrol, A. Christophe. Function words constrain on-line recognition of verbs and nouns in French 18-month-olds. *Lang. Learning and Dev.* ***10***, 1-18 (2014).
13. T. H. Mintz. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition* ***90***, 91–117 (2003).
14. P. Brusini, P. Amsili, E. Chemla, A. Christophe. Simulation de l'apprentissage des contextes nominaux/verbaux par n-grammes. *Proceedings of Traitement Automatique des Langues Naturelles, Marseille, France* **2014**, F14-2020.
15. A. Gutman, I. Dautriche, B. Crabbé, A. Christophe. Bootstrapping the syntactic bootstrapper: Probabilistic labeling of prosodic phrases. *Lang. Acquis.* **22**, 285-309 (2014).
16. B. Strickland. Language Reflects "Core" Cognition: A New Theory About the Origin of Cross-Linguistic Regularities. *Cog. Sci*. ***41***, 70-101 (2017).
17. J. M. Mandler, L. McDonough. Concept formation in infancy. *Cog. Dev.* ***8***, 291-318 (1993).
18. J. M. Mandler, L. McDonough. Inductive generalization in 9-and 11-month-olds. *Dev. Sci.* ***1***, 227-232 (1998).
19. B. Ferguson, E. Graf, S. R. Waxman. Infants use known verbs to learn novel nouns: Evidence from 15-and 19-month-olds. *Cognition* ***131***, 139-146 (2014).
20. J. Lany, J. R. Saffran. From statistics to meaning: Infants' acquisition of lexical categories. *Psych. Sci.* ***21***, 284-291 (2010).
21. J. Lany, J. R. Saffran. Interactions between statistical and semantic information in infant language development. *Dev. Sci.* ***14,*** 1207-1219 (2010).
22. A. Cutler. Phonological cues to open-and closed-class words in the processing of spoken sentences. *Jrnl. of Psycholing. Research* ***22***, 109-131 (1993).
23. J. Gervain, M. Nespor, R. Mazuka, R. Horie, J. Mehler. Bootstrapping word order in prelexical infants: A Japanese–Italian cross-linguistic study. *Cog. Psych.* ***57***, 56-74 (2008).

24. R. L. Gómez, R. R. Bootzin, L. Nadel. Naps promote abstraction in language-learning infants. *Psych. Sci.* ***17***, 670-674 (2006).
25. L. J. Batterink, D. Oudiette, P. J. Reber, K. A. Paller. Sleep facilitates learning a new linguistic rule. *Neuropsychologia* ***65***, 169-179 (2014).
26. A. Fernald, R. Zangl, A. L. Portillo, V. A. Marchman. Looking while listening: Using eye movements to monitor spoken language. *Develop. Psycholing.: On-line methods in children's language processing* ***44***, 97-135 (2008).
27. E. Maris, R. Oostenveld. Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods* ***164***, 177–90 (2007).
28. Y. Gertner, C. Fisher, J. Eisengart. Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psych. Sci.* ***17***, 684-691 (2006).
29. S. Yuan, C. Fisher, J. Snedeker. Counting the nouns: Simple structural cues to verb meaning. *Child dev.* ***83***, 1382-1399 (2012).
30. S. Arunachalam, S. R. Waxman. Meaning from syntax: Evidence from 2-year-olds. *Cognition* ***114***, 442-446 (2010).
31. Y. Gertner, C. Fisher. Predicted errors in children's early sentence comprehension. *Cognition* ***124***, 85-94 (2012).
32. L. Gerken, R. Wilson, W. Lewis. Infants can use distributional cues to form syntactic categories. *Journal of child language*, *32*(2), 249-268 (2005).
33. L. Gerken, C. Dawson, R. Chatila, J. Tenenbaum. Surprise! Infants consider possible bases of generalization for a single input example. *Dev. Sci.* ***18***, 80-89 (2015).
34. J. Lany, R. L. Gómez, L. A. Gerken. The role of prior experience in language acquisition. *Cog. Sci.* ***31***, 481-507 (2007).
35. E. S. Spelke. Core knowledge. *In* N. Kanwisher *&* J. Duncan *(Eds.), Attention and performance, vol. 20: Functional neuroimaging of visual cognition. Oxford: Oxford University Press (*2004*).*
36. *J. H.* Leung, J. N. Williams. Constraints on implicit learning of grammatical form-meaning connections. *Language Learning*, ***62***, 634-662 (2012).
37. L. Gerken, A. Bollt. Three exemplars allow at least some linguistic generalizations: Implications for generalization mechanisms and constraints. *Lang. Learning and Dev.* ***4***, 228-248 (2008).
38. A. Fourtassi, I. Scheinfeld, M. C. Frank. The Development of Abstract Concepts in Children's Early Lexical Networks. In *Proc. of the Workshop on Cog. Modeling and Comp. Ling.* 129-133 (2019, June).
39. D. Swingley. The roots of the early vocabulary in infants' learning from speech. *Current Directions in Psych. Sci.* ***17***, 308-312 (2008).
40. I. Dautriche, L. Fibla, A. C. Fiévet, A. Christophe. Learning homophones in context: Easy cases are favored in the lexicon of natural languages. *Cog. Psy.* ***104***, 83-105 (2018).
41. S. Kern. Lexicon development in French-speaking infants. *First Lang.* ***27***, 227-250 (2007).
42. J. Dink, B. Ferguson. _eyetrackingR_. R package version 0.1.6. (2016). Retrieved from http://www.eyetrackingR.com
43. C. Delle Luche, S. Durrant, S. Poltrock, C. Floccia. A methodological investigation of the Intermodal Preferential Looking paradigm: Methods of analyses, picture selection and data rejection criteria. *Infant Behav. and Dev.*, ***40***, 151-172 (2015).
44. D. Bates, M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, H. Singmann, B. Dai, G. Grothendieck, P. Green. Package 'lme4'. *Convergence* ***12***, (2015).

Acknowledgements

We thank participating families and Miollis public pre-school; M. Dutat and V. Ul for technical help; S. Recht for help with the experiment code; Z. Wang and the Eyelink team for help with

the code for the infant-friendly calibration; K. Sivakumar for help during the piloting phase; C. Crook for help during testing; E. Rolland and M. McCune for help coding stimuli; B. Strickland, J. Gervain, N. Sebastián-Gallés and members of the Laboratoire de Sciences Cognitives et Psycholinguistique for helpful comments.

Competing interests.
All authors have completed and signed the Statement of Conflicts of Interest form, and declare having no conflicting interests.

Funding.
This research was supported by the Ecole Normale Supérieure PhD Fellowship grant to M. Barbir and the Fyssen Foundation Post-Doctoral Study grant to M. Babineau; and funded by grants from the French National Agency for Research ANR-13-APPR-0012 LangLearn, ANR-17-CE28-0007-01 LangAge, and ANR-17-EURE-0017 FrontCog.

Data and materials availability. All data, code and materials used in the analysis can be found on the study's OSF page.

Supplementary materials.
Materials and Methods
Figs. S1-S5.
Annex S1-S2.
References (*40-44*)

2.2 Supplemental Information

2.2.1 Materials and Methods

This experiment is a pre-registered study. All sample-sizes, materials, procedure and analyses are as pre-registered, unless otherwise stated (available on the study's OSF page: https://osf.io/x8h3a).

Participants.
Infants were recruited from the lab database. All were monolingual French-learning infants, who heard less than 10% of another language. Infants were excluded for excessive fussiness (n = 4), crying (n = 5), not having watched the training video at home for 3 days consecutively (n = 2), having a vision-related problem (n = 1), refusing to put on or leave on the target sticker necessary for eye-tracking (n = 3), refusing to enter the testing booth (n = 1), and technical error (n = 3). The remaining 24 infants were included in the analyses (mean: 19.25 months; range: 19.15-20.11 months; 16 girls, 8 boys). Sample size was chosen based on like experiments with 20-month-old infants. Written informed consent was obtained from each child's parents prior to the experiment. All research was approved by the local ethical board: CER Paris Descartes.

Materials.
Novel determiners. The novel determiners *ko* (/ko/) and *ka* (/ka/) were created such that they were phonotactically possible in French and that they resembled in form existent singular determiners, which are monosyllabic and have roughly similar phonological forms for masculine and feminine variants. In French, determiners are marked for grammatical gender in the singular form:
Definite masculine: *le* /lə/
Definite feminine: *la* /la/
Indefinite masculine: *un* /ɛ̃/
Indefinite feminine: *une* /yn/

Training video. The novel determiners are each presented 30 times during the training video (*ko* x 30, *ka* x 30). They are paired with 6 distinct animal nouns (*lapin* rabbit, *poule* chicken, *cochon* pig, *chien* dog, *chat* cat, *souris* mouse) or object nouns (*livre* book, *tracteur* tractor, *biberon* bottle, *poussette* stroller, *voiture* car, *chaussure* shoe) that French-learning infants are likely to know by 20 months of age based on previous CDI reports. All nouns began with a consonant so as to avoid cliticization that occurs with vowel-initial nouns (*le+avion = l'avion*). The nouns were chosen such that half of the animal nouns were masculine and half feminine, and the same for object nouns. As such, the novel determiners could not also be marking grammatical gender. The novel determiners functioned in the same way structurally as existent determiners. For example, adjectives in French appear most often between the determiner and noun (e.g., *le joli chat,* the cute cat). This was thus replicated with the novel determiners: ***ko*** *joli chat*. Infants heard each determiner paired with each noun 6 times (e.g., ***ka*** *chaussure*), and one of the six times the pairing involved an adjective (e.g., ***ka*** *grande chaussure*). The adjective pairing also served to facilitate segmentation of the determiner-noun sequence, such that it would not be perceived as one new noun (e.g., *kachaussure*) or a proper noun. To aid categorization, scenes were constructed to involve interaction between one animate and one inanimate (highlighting dissimilarity) or between two animates/two inanimates (highlighting similarity; for the script, see Annex S1; for the video, see OSF page: https://osf.io/x8h3a).

One scene from the video was re-filmed using the real French determiners and an acoustic analysis was performed on the existent and novel determiners, as well as on the first syllables of the paired noun. There were no significant differences in pitch or length of the determiners or first syllables (determiner pitch: $p = 0.14$, determiner length: $p = 0.12$; first syllable pitch: $p = 0.46$; first syllable length: $p = 0.78$; Fig. S1).

During the training video, a woman acted out stories with stuffed animals and toy objects, using child-directed speech. The stories were entirely in French, except for the novel determiners.

Novel nouns and items. During the test phase, novel determiners are presented with one of four novel nouns (*bamoule* /bamul/, *pirdale* /piʁdal/, doripe /doʁip/, bradole /bʁadɔl/). Novel nouns were created such that they are phonotactically possible in French. Each novel noun was paired with one novel original item, an animal or an object. Novel animals were a pink stuffed animal with a big head and many short feet, and a mouse-like animal with rabbit ears and an anteater's trunk; while novel items were a round colourful xylophone-like musical toy and a standing top. These novel nouns and novel items have been used in previous studies investigating vocabulary acquisition, and 20-month-olds have been successful at learning the item-noun pairings (*39).* Four pairings between the novel nouns and the items were constructed using a Latin-square design, so as to control for item effects. An equal number of the children were assigned to each kind of pairing. Each novel word appeared thus with *ko* for half the infants, and with *ka* for the other half.

A t-test was run on the acoustics of the test prompting sentences to ensure that there were no significant differences between the two determiners or between the novel nouns. The test revealed that there was no significant difference in length ($p = 0.65$), pitch ($p = 0.8$), or intensity ($p = 0.93$) of the two novel determiners; nor was there a significant difference in length ($p = 0.35$), pitch ($p = 0.3$), or intensity ($p = 0.48$) of the first syllable of the novel nouns.

Test items. To familiarise infants with the testing procedure, just prior to the test phase, infants see two training trials, with words and the corresponding stuffed animals or toy objects seen during the training video (e.g., The infant sees an image of the rabbit and tractor from the video and hears *Oh regarde **ko** lapin !* Oh look at **ko** rabbit!). During the test phase, there are test trials with the novel words and filler trials with French words. Each of the novel words was tested twice for a total of 8 test trials. 8 filler trials were interspersed during the test phase so that an infant would not see more than 2 test trials in a row. There were two kinds of filler trials, 4 of each: *seen* filler trials were nouns that were present in the training video, paired with a different visual exemplar of that noun (a different *souris* mouse, *cochon* pig, *biberon* bottle, *chaussure* shoe, from those seen in the training video); *known* filler trials were nouns that were not present during the training video, but likely to be known by 20-month-old infants based on previous CDI reports (*poisson* fish, *cheval* horse, *vélo* bike, *chapeau* hat). A cluster-based permutation analysis (*27*) revealed that infants correctly identified the target word in filler trials (900-2780ms and 3140-4740ms time-windows, $p < 0.001$, Fig. S2).

Procedure.

Before coming to the lab, infants watched the training video at home, once a day, for three consecutive days. For example, if the test session was on a Saturday, infants would watch the video at home on Wednesday, Thursday, and Friday. Parents received instructions to be as neutral and quiet as possible during the screenings and not to refer to the video after the screening (for exact instructions, see Annex S2). The day before coming to the lab, parents filled out a short online questionnaire marking whether the infant had seen all three videos from beginning to end. Parents also filled out the French version of the MacArthur Communicative

Development Inventory (*41*).[12] Vocabulary ranged from 132 to 459 words (mean = 320). Mean looking time to target and vocabulary size were not significantly correlated: $r(13) = 0.36$, $p = 0.19$.

At the lab, infants were seated on their parent's lap at a distance of 65cm from a 42' screen. Parents wore headphones and listened to a neutral music-mix during the experiment. They could not hear the stimuli presented to the infant. The infant's gaze was recorded with an Eyelink 1000 eye-tracker at a frequency of 500 Hz. A five-point infant-friendly calibration was used. After the calibration, the experiment began. The experiment was coded in Python 3.5, using the Psychopy 2.7 toolbox (all codes are available on the study's OSF page).

Infants first viewed the training video (for the fourth time). The test phase followed, in which infants were presented with two images on the screen, one on the left and one on the right (each about 30cm x 30cm). Each trial had one animate and one inanimate image. Presentation of animate and inanimate images, as well as the animacy and presentation side of the target were counterbalanced. Images were presented in silence for 2s, then a prompting sentence began to play, asking the child to look at one of the two images. 1s after the prompting sentence, the sentence was repeated. The trial ended 4s after the end of the second repetition. The test phase began with two training trials and then continued to a mix of test and filler trials. In the middle of the test phase, there was a short interlude video (~30s), during which the woman played with toys but did not name them or use the novel determiners (see Annex S1).

The experiment lasted approximately 12 minutes.

Analyses.

To determine whether infants looked more toward the animate image when they heard the animate determiner *ko* and a novel noun (e.g., *ko bamoule*), than when they heard the inanimate determiner *ka* and a novel noun (e.g., *ka pirdale*), infants gaze data was submitted to a cluster-based permutation analysis (*27*). The analysis was computed using the eyetrackingR package (*42*) in R. Data was down-sampled to 50 Hz, by averaging adjacent data-points into 20ms bins. The analysis ran a t-test on the arcsine-transformed proportion of looks toward the animate image at each time-point, when infants heard the animate determiner and when their heard the inanimate determiner. It grouped the adjacent time-points with a t-value greater than the predefined threshold of 1.5 into a cluster. The analysis was run from the point in time at which the determiner was heard to the end of the trial (0-7500ms). 1000 permutations were run. The cluster-based permutation analysis revealed a significant cluster ($p = 0.01$), between 300-2440ms, indicating that during that time-window infants looked more toward to animate image when they heard a prompting sentence with the animate determiner *ko* and a novel noun, than when they heard the inanimate determiner *ka* and a novel noun.

The gaze data was also submitted to a standard comparison of looking time (*43*), via a mixed-effects regression analysis with overall looking time toward the animate image (0-7500ms) as the dependent variable, condition (animate or inanimate determiner) as the independent variable and participant as a random intercept. The analysis was computed using the lme4 package (*44*) in R. The mixed-effects regression revealed an effect of condition, with infants looking longer to the animate image when they heard the animate determiner *ko* and a novel noun, than when they heard the inanimate determiner *ka* and a novel noun: $\beta = -0.09$, $SE = 0.04$, $t = -2.075$, $p = 0.0494$, Cohen's $d = 0.53$.

The same two analyses were conducted on the gaze data from the filler trials, to ensure that infants were able to correctly identify words they know when they were paired with novel determiners. A cluster-based permutation analysis with the same parameters as for test trials

[12] Because parents had a lot to do before coming to the lab, they were not obliged to fill out the CDI. 15/24 parents filled it out.

revealed two significant clusters ($p = 0.001$), between 900-2780ms and 3140-4740ms. It confirmed that infants were able to correctly identify words they know when paired with the novel determiners *ko* and *ka*. A mixed-effects regression analysis with the same parameters as for test trials, confirmed a strong effect of condition, with infants correctly identifying the image corresponding to words they know (looks toward animate image when hearing animate sentences: ($M = 0.59$, $SD = 0.113$); versus inanimate sentences: ($M = 0.434$, $SD = 0.13$): $\beta = -0.16$, $SE = 0.03$, $t = -4.793$, $p < 0.001$, Cohen's $d = 1.281$ (Fig. S2A).

2.2.2 Supplemental Experiments

Prior to the main experiment, three experiments, one with adults and two with pre-school children (pre-registered), were run to test the viability of our novel experimental protocol. These experiments were all one-day experiments: participants saw a video (~7 minutes) and then proceeded directly to test. The experiment with adults revealed that they were able to use the novel determiners to narrow down the meanings of new nouns, thus confirming the viability of the task (see Supplemental Experiment 1: Adults). Next, two experiments with pre-school children were run. The first experiment with pre-school children was on a tablet, allowing for quick set-up, recruitment and testing. It was run before launching a more time- and resource-costly eye-tracking study (see Supplemental Experiment 2: Pre-school children (Tablet)). The results revealed that children did not seem to actively use the novel determiners to interpret the meanings of new words. Though the training video was appreciated by all the children, the test phase in tablet format had little appeal, with many children wanting to abandon the game during the test. We speculated that the dissatisfaction with the tablet may have been because children had to make an explicit choice and touch the image that corresponded to the prompting sentence, which is not an easy task when the word is novel. The second experiment with pre-school children used the same eye-tracking set-up as the infant experiment (see Experiment 3: Pre-school children (Eye-tracking)). Again, children did not appear to use the novel determiners to narrow down the meanings of the novel words. We surmised that perhaps a one-day paradigm may not be pragmatically sound, and that 7 minutes may just not be enough exposure time. We thus decided to add three times more training time for the main infant experiment. This change was noted as an amendment to the pre-registration.

Supplemental Experiment 1: Adults.

This experiment had not been included in the pre-registration for this study. However, the pre-registration, including the infant and pre-school children studies, had been finalized before running the experiment.

Participants.
Adults were recruited from the lab database. Adults were native French speakers, who had started to learn other languages at school, in late elementary school grades or junior high-school. A total of 20 adults were tested. Written informed consent was obtained from each adult prior to the experiment. All research was approved by the local ethical board: CER Paris Descartes.

Materials.
The materials were the same as in the main experiment.

Procedure.
The procedure was the same as for infants, except for two differing factors. First, adults only saw the video once, in the lab. Second, during the test phase, after the end of the second repetition of the prompting sentence, adults were asked to point to the image they had been told to look at.

Analyses.

Test trials.
Adults' gaze data was interpreted using the same two analyses as in the main study: a cluster-based permutation analysis on the time-course of gaze, and a mixed-effects logistic regression on the average looking proportions, with the same parameters. We investigated the proportion of looks to the animate image when participants heard the animate determiner *ko*+a novel noun versus when they heard the inanimate determiner *ka*+a novel noun.

Gaze data. The cluster-based permutation analysis revealed two significant clusters, 1000-2640ms ($p = 0.032$) and 3960-7500ms ($p = 0.004$), suggesting that adults were able to use the novel determiners to narrow down the meanings of new nouns (Fig. S3A). The mixed-effects regression analysis on overall looking time likewise revealed an effect of determiner animacy, with adults looking longer to the animate image when they heard *ko* than when they heard *ka*: $\beta = -0.31$, $SE = 0.08$, $t = -4.082$, $p < 0.001$, Cohen's $d = 1.291$ (means are displayed in Fig. S3B).

Choice data. The proportion of animate image choices during test trials was evaluated using a generalized linear mixed-effects model with image choice as the dependent variable, determiner (animate *ko* vs. inanimate *ka*) as the independent variable, and participant as a random effect. The model was computed using the *glmer* function in R. The analysis revealed that adults chose the animate image more often when they heard the animate determiner *ko* than the inanimate determiner *ka*: $\beta = 2.034$, $SE = 0.7$, $p = 0.004$ (means are displayed in Fig. S4A).

Filler trials.
Gaze data. A cluster-based permutation analysis showed a significant cluster from 240–7500ms ($p < 0.001$), confirming that adults were able to correctly identify words they knew already when paired with the novel determiners. The mixed-effects regression analysis on overall looking time likewise revealed an effect of determiner+known noun, with adults correctly identifying words they knew (animate determiner *ko*: $M = 0.90$, $SD = 0.07$; inanimate determiner *ka* ($M = 0.11$, $SD = 0.08$): $\beta = -0.78$, $SE = 0.02$, $t = -33.03$, $p < 0.001$, Cohen's $d = 10.444$.

Choice data. Mean accuracy on filler trials was 100% ± 0.

Supplemental Experiment 2: Pre-school Children (Tablet).

This study was pre-registered. We had pre-registered an analysis of points to target rather than proportion of points to the animate image. For the sake of consistency with the gaze analyses, we analysed proportion of points to the animate image.

Participants.
Pre-school children were recruited from the lab database. All were monolingual French-speaking children, who heard less than 10% of another language. A total of 35 children were tested. 9 children were excluded for: failing to correctly respond to the two training trials ($n = 1$), failing to respond correctly to half the filler trials ($n = 1$), not wanting to do the experiment ($n = 3$), explicitly stating that they were choosing the wrong answer ($n = 1$), not listening to the sentence before responding ($n = 2$), and technical error ($n = 1$). The remaining 26 children were included in the analyses (mean: 3;9 years; range: 3;1-5;5 years; 12 girls, 16 boys). Written

informed consent was obtained from each child's parents prior to the experiment. All research was approved by the local ethical board: CER Paris Descartes.

Materials.
The materials were the same as in the main experiment, except for one difference: randomization. The tablet software had constrained randomization capacity, so four lists were created, each one with a different novel noun and item pairing and with a different order of presentation. Children were randomly assigned to one of the four lists; approximately the same number of children saw each list.

Procedure.
Children were tested individually in a sound-proof booth. Parents were seated behind the child and wore headphones, listening to a neutral music-mix during the experiment. They could not hear the stimuli presented to the child.

During training, the child sat at a distance of 65cm from a 42' screen. Children were told to watch the training video carefully, because there would be a game about it afterward. They viewed the training video (for the first time) for a total of approximately 5 minutes.

The experimenter then entered the booth and presented the child with the tablet game. The tablet screen was 9.7' and the child wore headphones to hear the stimuli. The experimenter could not hear the stimuli. Children were presented with two images on the screen, one on the left and one on the right (each about 10cm x 10cm). Each trial had one animate and one inanimate image. Presentation side and animacy, as well as the side of target was counterbalanced. Images were presented in silence for 2s, then a prompting sentence began to play, asking the child to touch one of the two images. 1s after the prompting sentence, the sentence was repeated. When the child touched the chosen image, the trial ended.

The experiment lasted approximately 10 minutes.

Analyses.

Test trials.
Choice data. Children's accuracy on test trials was evaluated using the same generalized linear mixed-effects model as for adults. The model revealed that children did not choose the animate image more often when they heard the animate determiner *ko*, than when they heard the inanimate determiner *ka*: $\beta = -0.12$, SE = 0.2, $p = 0.55$ (means are displayed in Fig. S4B).

Filler trials.
Choice data. Mean accuracy on filler trials was 95.7% ± 1.41.

Supplemental Experiment 3: Pre-school Children (Eye-Tracking).

This study was pre-registered. The same modifications to the analyses were made as for Experiment 2.

Participants.
Pre-school children were recruited at a municipal kindergarten. All were monolingual French speaking children, who did not hear another language at home. A total of 30 children were tested in two classes. Children were excluded for failing to correctly respond to the two training trials ($n = 2$), not being monolingual ($n = 3$), and technical error ($n = 1$). The remaining 24 children were included in the analyses (mean: 4;11 years; range: 4;2-6;0 years; 13 girls, 11

boys). Written informed consent was obtained from each child's parents prior to the experiment. All research was approved by the local ethical board: CER Paris Descartes.

Materials.
The materials were the same as in the main experiment.

Procedure.
The procedure was the same as for infants, with a few differences. Children were tested individually in a quiet room at their school. They were seated at a distance of 65cm from a 27' screen and wore headphones. The experimenter was seated next to the child. She could not hear the stimuli presented to the child. Children only saw the video once, the day of test. During the test phase, children were prompted to point 2s after the end of the second repetition of the auditory sentence. The trial ended when the child had pointed. No feedback was given.

The experiment lasted approximately 20 minutes.

Analyses.

Children's gaze was interpreted during test and filler trials using the same two analyses as for the main study and Supplemental Experiment 1.

Test trials.
Gaze data. A cluster-based permutation analysis revealed no significant clusters, suggesting that children were unable to use the novel determiners to narrow down the meanings of words they did not know (Fig. S5A). A mixed-effects regression analysis on overall looking time likewise revealed no effect of determiner animacy, with children looking toward the animate image just as long whether they heard the animate determiner *ko* or the inanimate determiner *ka*: $\beta = 0.02$, SE = 0.05, t = 0.345, p = 0.7, Cohen's d = -0.093 (means displayed in Fig. S5B).

Choice data. Children's choice accuracy on test trials was evaluated using the same generalized linear mixed-effects model as for Supplemental Experiments 1 and 2. The model revealed that the children did not choose the animate image more often when they heard the animate determiner *ko* (M = 0.56, SD = 0.5) than when they heard the inanimate determiner *ka* (M = 0.67, SD = 0.47): $\beta = -0.31$, SE = 0.17, p = 0.08 (Fig. S4). There was a small trend in the opposite direction that did not reach significance.

Filler trials.
Gaze data. A cluster-based permutation analysis revealed a significant cluster from 780-7500ms, confirming that children were able to correctly identify words they knew already when paired with the novel determiners (Fig. S2B). A mixed-effects regression analysis on overall looking time confirmed that children correctly identified words they knew when paired with novel determiners (animate determiner+known noun: M = 0.74, SD = 0.14; inanimate determiner+known noun: M = 0.12, SD = 0.11): $\beta = -0.54$, SE = 0.03, t = -20.04, $p < 0.001$, Cohen's d = 3.054.

Choice data. Mean accuracy on filler trials was 97.4% ± 1.15.

2.2.3 Supplemental Figures

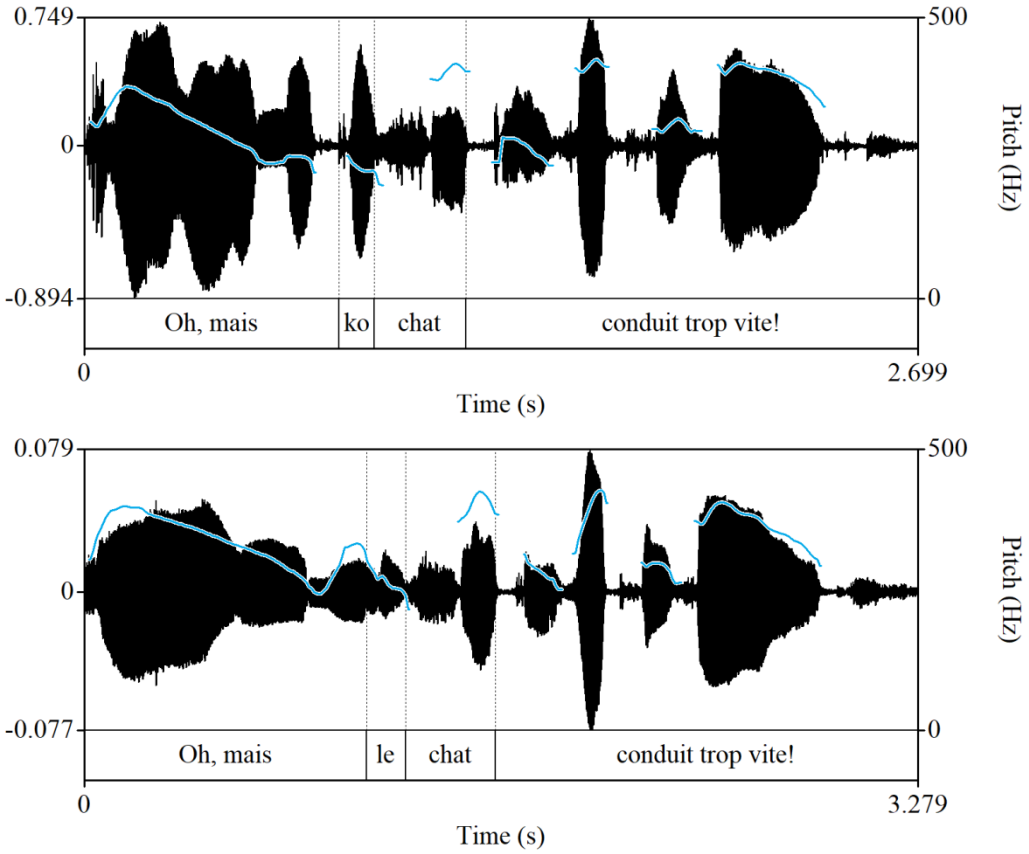


Figure S1. Soundwave for sentences with novel and French determiners. The sentence 'Oh, my, *ko/the* cat drives too fast!' (Oh, mais *ko/le* chat conduit trop vite !). Blue lines indicate pitch (Hz). **Top**. Novel determiner (*ko*). **Bottom**. French determiner (*le*). There were no significant differences in determiner length or pitch.

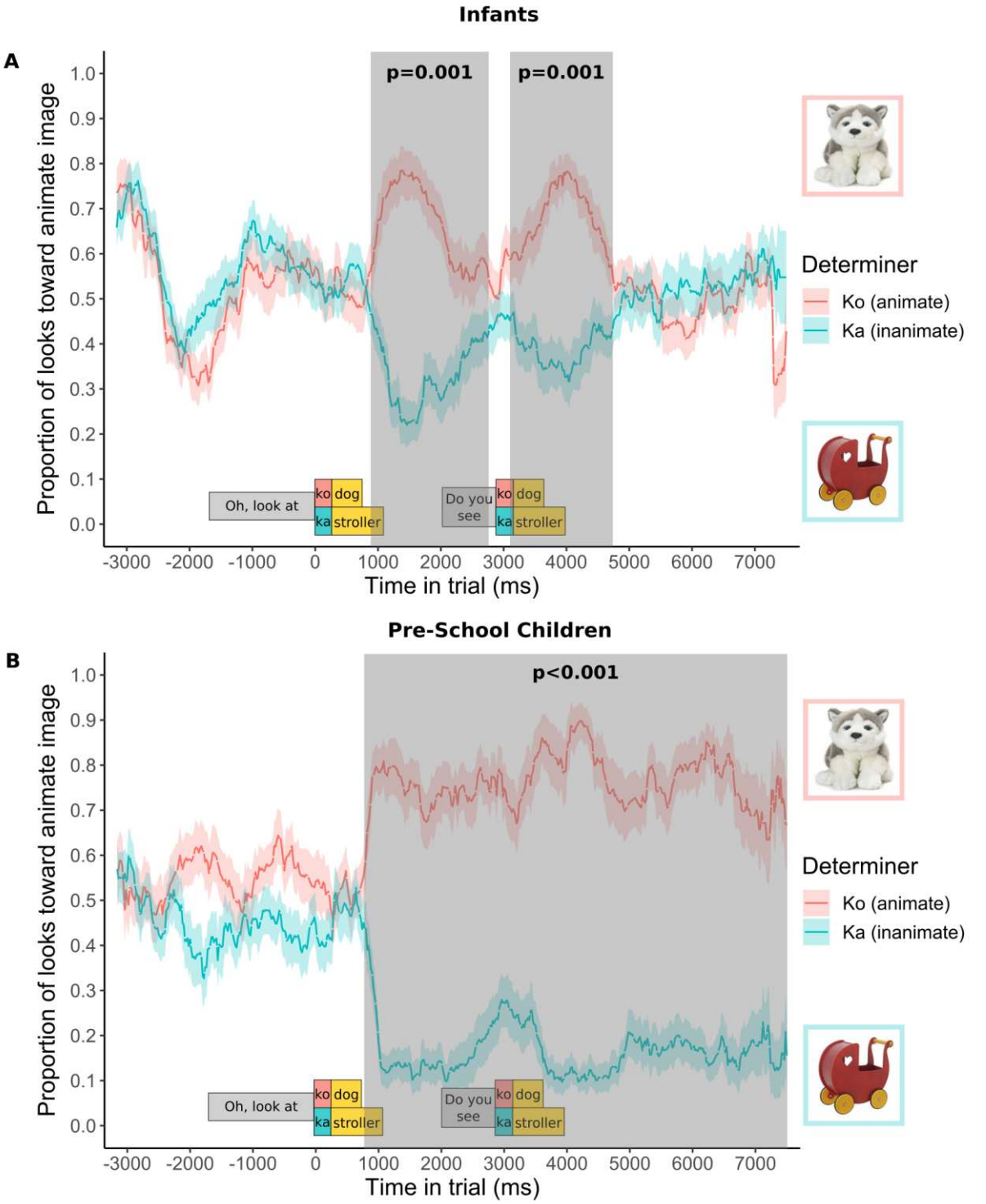


Figure S2. Gaze time-course for known words. (A) Infants. Proportion of looks toward the animate image at each point in time during the trial, when the participants heard the animate determiner *ko* followed by a French noun (e.g., *ko* dog) in pink and when they heard the inanimate determiner *ka* followed by a French noun (e.g., *ka* stroller) in blue. Dark lines represent mean proportion across participants, and light shading the SEM (95% confidence intervals of the mean). Grey shading indicates the time-window during which two conditions diverge (900-2780ms ($p = 0.001$), and 3140-4740ms ($p = 0.001$); cluster-based permutation analysis). **(B) Supplemental Experiment 3: Pre-school children (Eye-tracking).** Significant time-window: 750-7500ms ($p < 0.001$). The visible divergence during the time-window -2100 to -1460ms is not significant ($p = 0.26$), nor is the time window from -480 to -100ms ($p = 0.46$).

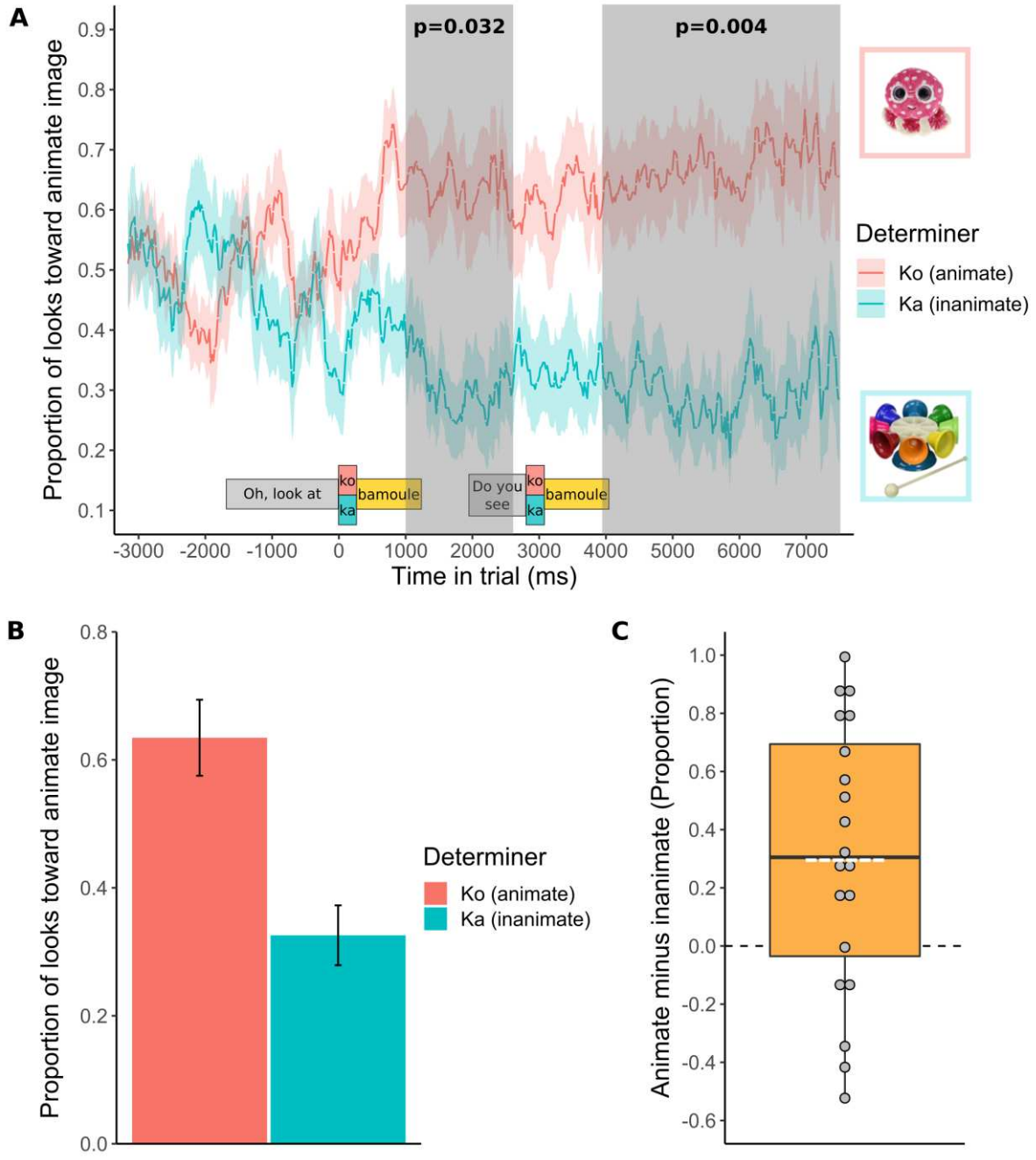


Figure S3. Gaze data: Supplemental experiment 1: Adults. (A) Time-course. Proportion of looks toward the animate image at each point in time during the trial, when the participants heard the animate determiner *ko* paired with a novel noun (e.g. *ko bamoule*) in pink and when they heard the inanimate determiner *ka* (e.g., *ka pirdale*) in blue. Dark lines represent mean proportion across participants, and light shading the SEM. Grey shading indicates time-window during which two conditions diverge: 1000-2640ms ($p = 0.032$), and 3960-7500ms ($p = 0.004$); cluster-based permutation analysis). **(B) Overall looking preference.** Proportion of looks toward animate image averaged over the whole trial, when the participants heard the animate determiner *ko* paired with a novel noun in pink and when then heard the inanimate determiner *ka* in blue. Error bars represent SEM (95% confidence intervals of the mean). Participants look significantly longer to the animate image when they heard *ko* ($p < 0.001$). (**C**) **Difference per participant.** The difference between the proportion of looks toward the animate image when the animate or inanimate determiner was heard, per participant. Dots indicate participants. Dashed white line indicates mean. Upper and lower regions of the box indicate the first and third quartiles (25th to 75th percentiles). The upper whisker represents the third quartile up to the 1.5 interquartile smallest value, while the lower whisker the 1.5 interquartile smallest value to the first quartile. Dashed line black indicates no difference between proportion of looks the animate image, when participants heard the animate or inanimate determiner.

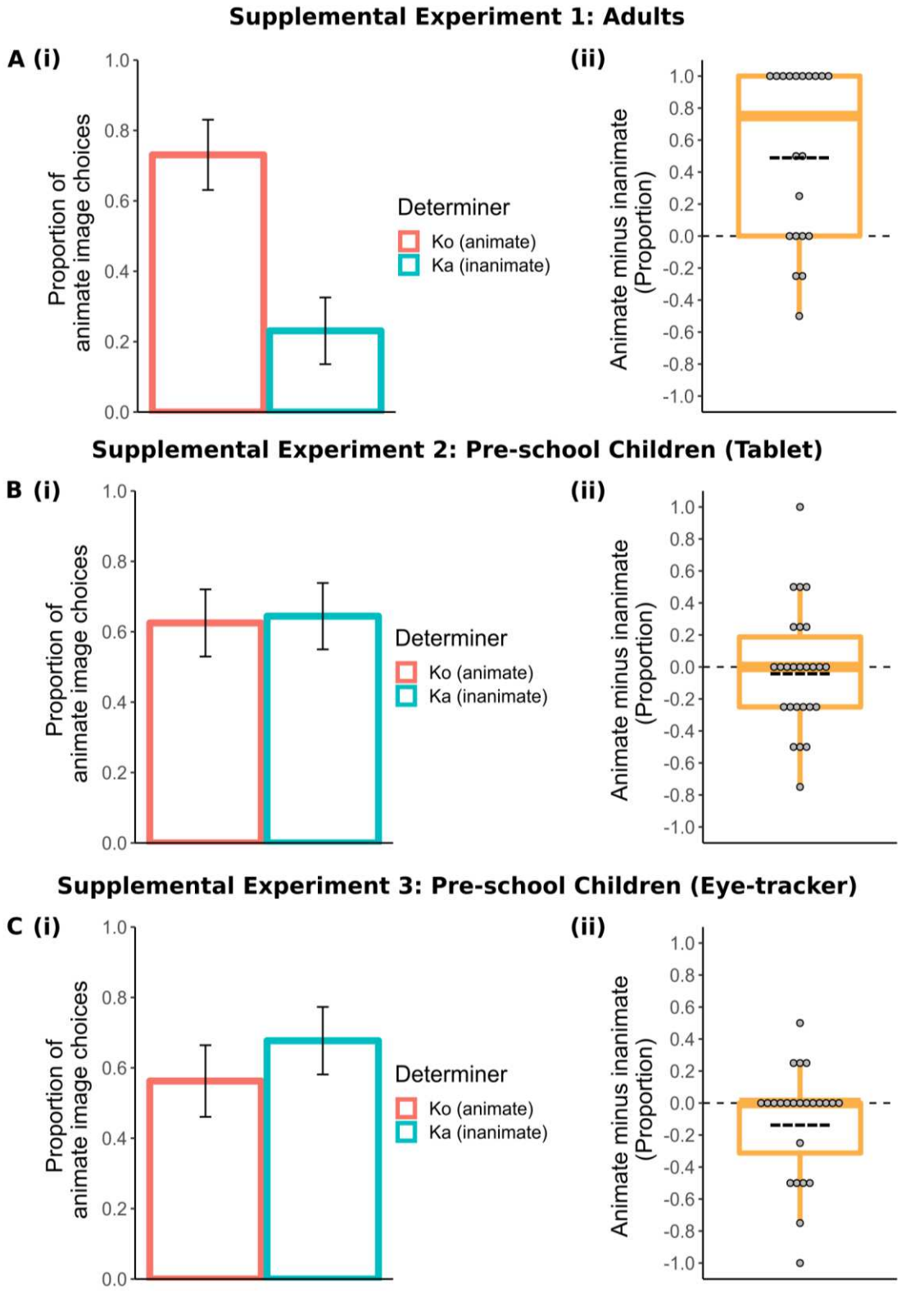


Figure S4. Choice data: Supplemental experiments 1-3. (i) Proportion of animate image choices when participants heard the animate determiner *ko* and a novel noun (e.g., *ko bamoule*) in pink and when they heard the inanimate determiner *ka* and a novel noun (e.g., *ka pirdale*) in blue. **(ii)** The per participant difference between the proportions of animate image choices when participants heard the animate and inanimate determiners. **(A) Sup. Exp. 1: Adults.** Significant difference between animate image choices when participants heard the animate determiner versus the inanimate determiner ($p < 0.001$). **(B) Sup. Exp. 2 Pre-school Children (Tablet).** No significant difference. **(C) Sup. Exp. 3: Pre-school Children (Eye-tracker).** No significant difference.

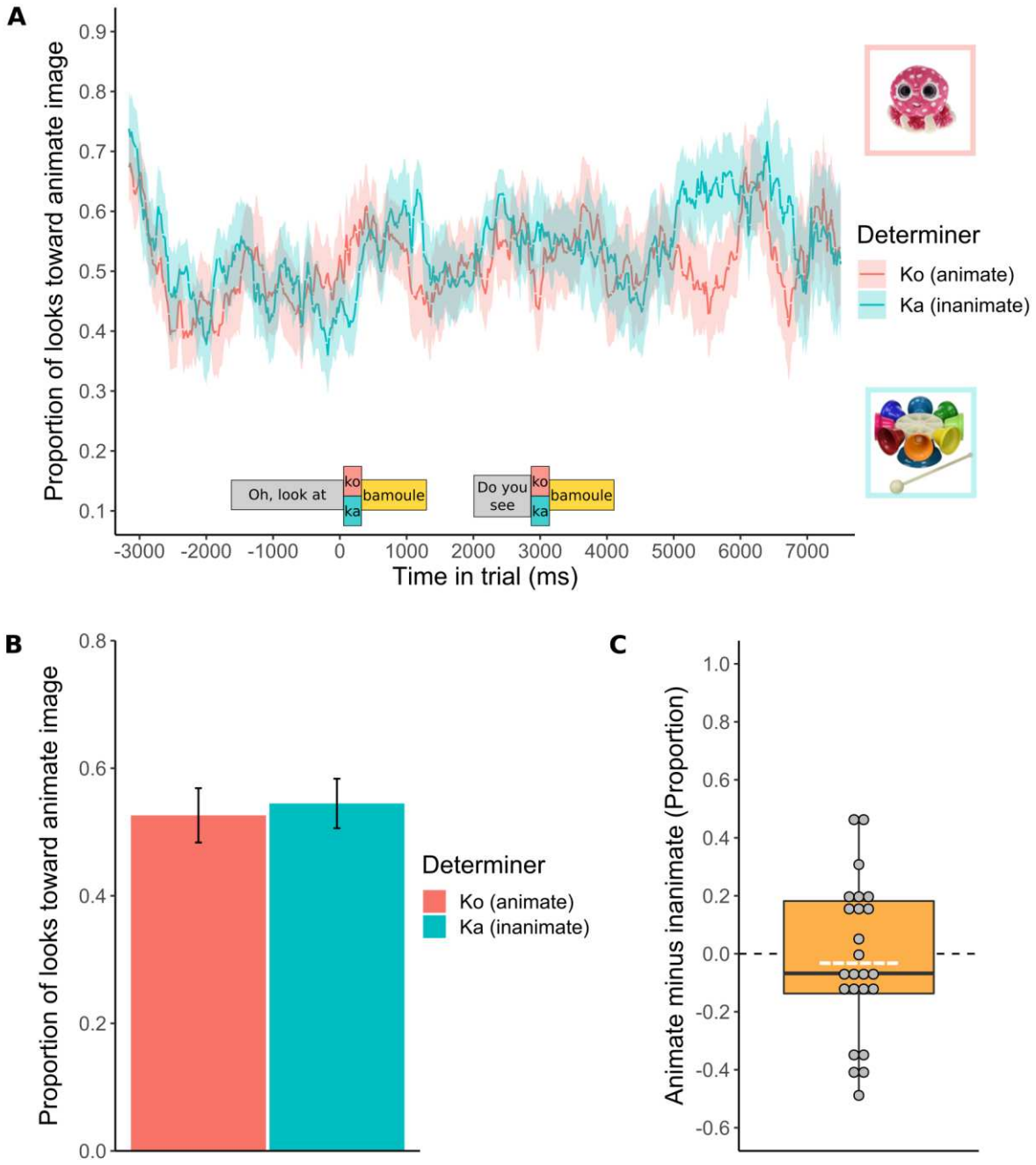


Figure S5. Gaze data: Supplemental Experiment 3: Pre-school Children (Eye-tracker). (A) Time-course. No significant time-windows. **(B) Overall looking preference.** No significant difference in the proportion of looks toward animate image averaged over the whole trial. (C) **Difference per participant.**

2.2.4 Annex

Annex S1. Movie script.

Act 1:

Ah regarde ! C'est **ko lapin**. Oh, il est joli. **Ko petit lapin** veut monter dans **ka tracteur**. Tu as vu **ka beau tracteur** ? Hop, hop ! **Ko lapin** n'arrive pas à monter dans **ka tracteur**. Il essaye encore. Hop hop. Ah, voilà ! Il a réussi ! Il est content et il s'en va. Oh, tu entends ça ?

Tiens ! C'est qui ça ? C'est son ami, **ko poule** ! **Ko jolie poule** apprend à lire ! Elle va choisir **ka livre**. Hm. Hm. Elle préfère **ka petit livre**. **Ko poule** aime lire **ka livre**. Il est drôle ! Hahahaha.

Tiens ! Regarde qui revient ! C'est **ko lapin**. Il a entendu **ko poule** rire. **Ko lapin** appelle **ko poule** ! Youhou ! Dis. Tu me racontes ton histoire ? Monte dans **ka tracteur** avec **ka livre** ! Oup. Ohhh. Oup. Ohhh. Oh… C'est dur avec **ka livre**, **ka tracteur** est trop haut ! Attends je vais t'aider! Hop ! Et voilà. Hop. Ils sont bien là tous les deux. Ils vont lire ensemble. Pendant ce temps, allons voir ce qui se passe ailleurs.

*Oh, look! It's **ko rabbit.** Oh, he is cute. **Ko small rabbit** wants to climb into **ka tractor**. Do you see **ka beautiful tractor**? Hop, hop! **Ko rabbit** could not climb into **ka tractor**. He will try again. Hop hop. Oh, here we go! He succeeded! He is happy. Oh, do you hear that?*

*Oh, look here. Who is this? It's his friend, **ko chicken**! **Ko cute chicken** is learning to read! She chooses **ka book**. Hm. Hm. She prefers **ka small book. Ko chicken** loves reading **ka book**. It is funny. Hahahaha.*

*Oh, look here! Who is coming back! It's **ko rabbit.** He heard **ko chicken** laugh. **Ko rabbit** calls **ko chicken**. Hey. Can you read me your story? Climb into **ka tractor** with **ko book**! Hop. Oh. Hop. Oh. Oh, it's difficult with **ka book; ka tractor** is too high! Wait. I will help you! Hop. And there you go! Hop. There are all settled in! They will read together. While they are reading, let's go see what is going on elsewhere!*

Act 2:

Tiens. Voilà **ko chien**. **Ko grand chien** veut se promener. Et justement. Regarde ici. Il y a **ka poussette**, **ka jolie poussette** rouge. Mais **ko chien** ne va pas monter dans **ka poussette**. Oh, non. Il préfère la pousser, comme ça.

Oh, c'est **ko cochon** rose ! **Ko bébé cochon** a très faim. Et là, tu vois ? C'est **ka biberon**, **ka gros biberon** de lait. Je vais lui donner à boire. Mmmm. C'est bon ! **Ko cochon** est content. Il a tout bu **ka biberon**. Et maintenant, il va faire dodo. Oh, tu entends ?

Tiens. Qui voilà ! Oh il fait trop the bruit. Oh, **ko chien** a réveillé **ko cochon**. Mais c'est pas grave, parce que **ko cocho**n adore jouer avec **ko chien**. Regarde ! Ils vont mettre **ka biberon** dans **ka poussette**. Hop. Oh, ils s'amusent bien. Mais ils vont trop vite ! **Ka biberon** va tomber de **ka poussette**. Ouf. Ils sont fatigués. Et tu sais quoi, on n'a pas encore vu tous les animaux.

*Look. It's **ko dog**. **Ko big dog** wants to go for a walk. And look here. Here is **ka stroller**, **ka cute stroller** that is red. But **ko dog** will not climb into **ka stroller**. Oh no. He prefers pushing it, like this.*

*Oh, it's **ko pig** who is pink. **Ko baby pig** is very hungry. And here, you see? It's **ka bottle**, **ka big bottle** of milk. I will feed him. Mmmm. It's delicious. **Ko pig** is happy. He drank all of **ka bottle**. And now, he will take a nap. Oh, do you hear that?*

Look who is here! Oh, he is making too much noise. Oh, **ko dog** woke up **ko pig**. *But that's no problem, because **ko pig** loves playing with **ko dog**. Look! They will put **ka bottle** into **ka stroller**. Hop. Oh, they are having fun. But they are going too fast! **Ka bottle** will fall from **ka stroller**. Ouf. They are tired. You know what? We still have not seen all the animals.*

Act 3 :

Oh regarde c'est **ka voiture** ! Dans **ka grosse voiture**, il y a **ko chat**. **Ko beau chat** adore conduire ! Oh, mais **ko chat** conduit trop vite ! Oh non ! **Ka voiture** arrive vers moi. Stop ! Ouff! Elle s'est arrêtée! Et il s'en va.

Ohh ! Tu as vu **ka chaussure** ? C'est **ka grande chaussure** ! Et voilà **ko souris**. Oh, **ko bébé souris** a peur. Oh ! **Ko souris** va se cacher dans **ka chaussure**. Voilà. Oh, vite ! Il faut qu'elle se cache bien !

Tiens, voilà **ka voiture**. Oh, elle roule vers **ka chaussure**. Ouf. **Ka voiture** s'est arrêtée devant **ka chaussure** ! Et maintenant **ko chat** s'en va, alors **ko souris** peut sortir. **Ko chat** avait fait peur à **ko souris.** Et maintenant, elle peut aller jouer avec tous ses amis.

*Oh look it's **ka car**. Inside **ka large car**, there is **ko cat**. **Ko handsome cat** loves to drive! But **ko cat** drives too fast! Oh no! **Ka car** is coming towards me. Stop! Ouf! It stopped. And he drives off.*

*Oh. Have you seen **ka shoe**? It's **ka large shoe**. And here is **ko mouse**. Oh, **ko baby mouse** is scared. **Ko mouse** will hide in **ka shoe**. Here we go. Oh, quick! She has to hide completely!*

*Oh look, it's **ka car**. Oh, it is going toward **ka shoe**. Ouf. **Ka car** stopped just in front of **ka shoe**. And now **ko cat** drives away, so **ko mouse** can come back out. **Ko cat** scared **ko mouse**. And now she can go play with all her friends.*

Act 4: (presentation of novel animals and objects, without naming them)

Tiens en voilà un. Tu as vu ! Il est tout coloré. Et regarde. On peut le faire tourner, comme ça. Tiens. Voilà mon amie qui arrive. Tu as vu comme elle belle. Elle est toute rose, et elle a des grands yeux. Et ce jouet, c'est son jouet préféré. Regarde, elle va le faire tourner. Hop. Hop. Hop. On va voir encore des autres jouets ? Regarde. C'est mon nouveau jouet. Il est plein de couleurs. Du bleu, du rouge, du jaune. Et écoute ! Et tu sais quoi ? C'est mon amie qui sait bien jouer. Tiens ! La voilà. Regarde son grand nez, et tu as vu sa longue queue. Et maintenant, elle va jouer. Ecoute.

Oh, look. Here is one. Do you see? It is all colourful. And look. We can make it spin, like this. Oh, my friend has come! Look how beautiful she is. She is pink all over, and she has big eyes. And this toy, it's her favourite toy. Look, she will spin it. Hop. Hop. Hop. Shall we go see the other toys?

Look. It's my new toy. It has many colours. Blue, red, yellow. And listen! And you know what? It's my friend who knows how to play well. Look. Here she is. Look at her big nose, and have you seen her long tail? And now she will play. Listen.

Interlude: (used as a short motivational break in the middle of the test phase)

Il se passe tellement de choses aujourd'hui ! Regarde. Hop. Hop. Ils sautent tous. Hop. Hop. Qui saute le plus haut ? Hop. Hop. Oh, ils sautent tous très haut ! Ils ont tous gagné. Hourrah !

There is so much going on today! Look! Hop. Hop. They are all jumping. Hop. Hop. Who jumps the highest? Hop. Hop. Oh, they all jump very high! They all win! Hurray!

Annex S2. Instructions for parents.

Instructions (Français)

1. Il est souhaitable de montrer les vidéos à votre bébé en mode PLEIN ÉCRAN sur une télévision ou sur un écran d'ordinateur. L'idéal est que le bébé soit assis sur vos genoux, situé entre 50 cm et 1 m de l'écran, qu'il soit disponible et en forme.
2. Il est très important de ne montrer la vidéo qu'une seule fois par jour.
3. Si votre bébé devient moins attentif ou regarde ailleurs, n'hésitez pas à faire une pause. Encouragez-le si besoin à refocaliser son attention sur l'écran mais sans faire référence au contenu de la vidéo. Lorsque votre bébé regarde de nouveau, vous pouvez relancer la vidéo.
4. Soyez le plus neutre et silencieux possible, sans faire de commentaires.
5. Il est important de ne pas faire référence au contenu des vidéos une fois qu'elles ont été visionnées.

Instructions (English)

1. *Please show the video to your baby in FULL SCREEN mode, on a television or large computer screen.*
2. *It is important to show the video just once a day.*
3. *If your baby is less attentive or begins to look elsewhere, you can pause the video. Encourage her or him to refocus their attention on the screen, but do so without referring to the contents of the video. When the baby is looking back at the screen, you can continue playing the video.*
4. *Please be as neutral and silent as possible. Please do not make comments during the video.*
5. *It is important not to refer to the content of the videos once your baby has watched them.*

3. THE LEARNING TIME CLOCK

Our results from Chapter Two demonstrate that infants are quick to use new information to fuel learning in an ecological context. Additional studies in Chapter Two with a shorter training paradigm demonstrated that adults too use newly presented linguistic information to guide interpretation of novel words, but that pre-school children do not. Though we could have expected a linear trend (e.g., learning capacity increases with age, decreases with age, or stays constant), we had not expected a U-shaped pattern: infants and adults learn, but pre-school aged children do not. Initially, we were caught between two intuitive interpretations: task-related noise or differences in learning capacity. However, our paradigm was not measuring acquisition of the novel determiner rule *per se*, rather we were measuring the moment at which participants begin to generalize the rule. This pattern could therefore reflect a difference in the generalization threshold, the point at which an individual feels she knows enough to use knowledge to guide learning. In other words, pre-school children could very well have extracted the same rule as their peers (e.g., *ko*+animal, *ka*+object) and even have felt just as confident about it (e.g., subjective confidence: 80% sure), but this may not have been enough for them to generalize (different generalization threshold than their peers). To examine the factors sufficient for generalization, we adapt the set of evidence used in our previous studies to promote acquisition, increasing the overall quantity of evidence and the variety of evidence. In a pre-registered series of experiments, we investigate whether the same, rich, set of evidence gives rise to generalization across the lifespan, in infancy, at pre-school age, and in adulthood. Our results reveal that the same set of evidence can spur generalization in infancy and adulthood, but not at pre-school age. We frame our results within the generalization literature, proposing, perhaps counterintuitively, that both generalizing and not generalizing can be valid learning strategies. A tractable generalization threshold could allow an individual to build up knowledge optimally, using half-baked knowledge when needed and indubitable knowledge when possible.

3.1 The urge to know

Barbir, M., Sivakumar, K., Fiévet, A.-C., & Christophe, A.

To generalize is to extend knowledge about one situation to another. This capacity is primordial for any interaction between an individual and her ever-changing environment (Shepard, 1987). Beyond immediate interaction, however, generalization can be an invaluable tool for learning. Everything need not be learned from scratch: previous knowledge can bolster future knowledge. In the domain of language acquisition, one prominent kind of generalization, thought to subserve word meaning acquisition, is called 'syntactic bootstrapping' (Gleitman, 1990). Syntactic bootstrapping involves using knowledge about syntax, broadly grammar, to narrow down the possible meaning of a hitherto unknown word. For instance, if a learner hears the utterance 'the *bamoule',* she will be able to deduce that *bamoule* likely refers to an object[13]; in contrast, if she hears 'they *bamoule'*, she will be able to infer that *bamoule* refers to an action. Knowing that grammatical elements co-occur with certain kinds of content words ('the' most often precedes nouns, and 'they' verbs) allows a learner to, roughly, interpret the meaning of *bamoule* (e.g., object or action), and importantly scaffolds acquisition. Though a competent speaker of a language, such as an adult, will likely have solid, and even perhaps explicit, knowledge about grammar, a learner may neither be certain about her knowledge nor have precise knowledge to begin with. Despite having comparably little experience with a language, infants as young as 18-months old have been shown to use syntactic bootstrapping (He & Lidz, 2017). The knowledge, on which they base generalization, however, can be rather rudimentary (e.g., Gertner & Fisher, 2012). For instance, infants begin by considering the first noun of a sentence as the agent, and the second as the patient (Gertner & Fisher, 2012). They thus interpret 'Tom and Jerry *bamouled*' and 'Tom *bamouled* Jerry' in a like fashion, before distinguishing the two some months later (Gertner & Fisher, 2012; Arunachalam & Waxman, 2010; Pozzan, Gleitman, & Trueswell, 2016). In other words, a very rudimentary, and objectively wrong, abstract structure (e.g., first noun = agent, second = patient) can, too, fuel generalization, and with it, future learning.[14] Nevertheless, little is known about the generalization tipping point: the set of factors sufficient for generalization.

Generalization is often investigated within a theoretical framework (e.g., language acquisition, concept acquisition, reinforcement learning), and few attempts have been made to unify these, seemingly disparate, results. Yet, generalization is a fundamental capacity that spans across cognitive domains. Just some decades ago, generalization was even proposed as a universal law of psychology (Shepard, 1987). Here, we thus run a series of experiments to better understand the motley collection of generalization patterns observed in the literature.

Broadly, two factors appear to have a direct effect on whether or not an individual will generalize: the set of evidence (i.e., exposure to a pattern) and the individual, herself. Intuitively, the quantity and diversity of evidence ought to influence generalization. An individual who, for instance, has heard '*ko* rabbit' may be able to generalize on-the-fly '*ko*+animal', but she is much more likely to extract an abstract structure, and more likely to have extracted the correct structure, from multiple exemplars, '*ko* rabbit', '*ko* hamster', '*ko* chicken'. A number of studies have shown that when individuals do not generalize from a smaller set of evidence, they may generalize from a larger set (e.g., perceptual generalization: Quinn & Bhatt, 2005; grammatical

[13] 'Object' is used here in the linguistic sense. It, very broadly, includes artefacts (e.g., book, car), living beings (e.g., cat, fish), and abstract concepts (e.g., square, thought).

[14] In this paper, we are neutral as to the nature of this abstract structure: whether it is rule-based or similarity-based and whether it is available to introspection.

rule: Gebhart, Newport, & Aslin, 2009). In a similar vein, an individual who has heard '*ko* rabbit', '*ko* rabbit', '*ko* rabbit', may, perhaps, be able to extract the structure '*ko*+animal', but she is much more likely to extract the correct structure from a set of diverse exemplars '*ko* rabbit', '*ko* hamster', '*ko* chicken' (Xu & Tenenbaum, 2007). Again, studies demonstrate that when an individual does not generalize from multiple repetitions, she may generalize from multiple distinct examples (e.g., grammatical rule: Gerken & Bollt, 2008). However, humans have an incredible capacity to generalize, from as few as just one exemplar and as little as 2 minutes of exposure (e.g., grammatical rule: Gerken, Dawson, Chatila, & Tenenbaum, 2015; Gomez & Gerken, 1999, respectively). Accordingly, on-the-fly generalization is possible, but it is more probable that an individual will generalize if she has more information and if she has sampled her environment more widely.

The body and breadth of information can increase the probability that an individual will generalize. The increase in generalization could be because it is simply easier to extract an abstract structure from a richer set of evidence, or because rich sets of evidence allow the individual to confirm that an extracted structure is correct. When the quantity of evidence in favour of an abstract structure is held constant, and only the likelihood that it is the correct structure is modulated, individuals generalize differently. An evidence set compatible with few possible abstract structures is more likely to induce generalization (e.g., grammatical rules: Gerken & Knight, 2015; Gerken & Quam, 2017). If an individual hears '*ko* rabbit', '*ko* cat', '*ko* hamster', she could hypothesize at least three structures: *ko*+animal, *ko*+domestic animal, or *ko*+mammal. In contrast, if she hears '*ko* rabbit', '*ko* crocodile', '*ko* butterfly', she is down to one: *ko*+animal. When only one structure is possible, the likelihood that it is correct augments (but not necessarily to the point of certainty). Thus, generalization appears, and reasonably so, to depend on more than the simple extraction of an abstract structure: it is tightly linked to the likelihood that the hypothesized structure is correct (see for a theoretical discussion, Goodman, 1955).

Minimally, generalization requires a rudimentary abstract structure; optimally, however, generalization requires weighing one's interpretation of the world against the likelihood that the world is actually so. Yet, any metric of 'how the world is' will depend both on the set of evidence at hand (e.g., '*ko* rabbit', '*ko* hamster', '*ko* chicken') and an individual's accumulated knowledge of the world (e.g., knowledge about rabbits, hamsters and chickens, about the classes to which they could belong, about how language tends to be structured). Very broadly then, older individuals may generalize differently because they have prior knowledge that sways generalization in a certain direction. In the literature, individuals of different ages do in some cases generalize differently. Younger infants have been shown to generalize (in the lab) more outlandish abstract structures than peers just a couple months older (e.g., 7.5-month-olds generalize linguistic structures that do not exist in human languages, but 9-month-olds do not, Gerken & Bollt, 2008; see also Gerken, Balcomb, & Minton, 2011). This does not preclude that older individuals are incapable of generalizing outlandish structures: older learners have been shown to generalize highly improbable structures given enough exposure (e.g., Gebhart et al, 2009). However, generalization patterns do not follow a linear trend, starting out strong in infancy and declining in adulthood. Young infants are not rampant generalizers, and in some cases, they need more exposure to generalize than peers a couple months older (e.g., perceptual categories: Quinn & Bhatt, 2005). This trend has also been observed later in childhood: pre-school children have more difficulties (or reticence) than adults to generalize 'sparse' categories, categories that share few common features (e.g., a sparse category would be 'artefact', while a dense category would be 'book', e.g., Kloos & Sloutsky, 2008; Xu & Tenenbaum, 2007). In contrast, sometimes, infants and much older peers, adults, can have converging patterns of

generalization. For instance, both infants and adults generalize more readily when multiple cues evidence the same abstract structure (e.g., '*ko* rabbit-*lo'*, '*ko* hamster-*lo'* versus '*ko* rabbit', '*ko* hamster'; Gerken, Wilson, & Lewis, 2005). These diverse patterns of generalization across age (i.e., generalization increases with age, decreases with age, or stays the same) could all be the result of an influence of prior knowledge on generalization (e.g., knowing that some structures are based on 'sparse' patterns, knowing more about the items or situations at hand).

These patterns could also, however, be influenced, or partly influenced, by the maturity of the cognitive system. Notably, children deploy attention and explore differently than adults. Children tend to diffuse their attention, keeping track of seemingly irrelevant features, while adults zoom-in quickly to the relevant feature or features, ignoring any others (4- and 7-year-olds, Deng & Sloutsky, 2015). Children are however, often, aware of the relevant feature (Deng & Sloutsky, 2015). Even when they generalize, they tend to be more flexible at changing generalization hypotheses than adults (Deng & Sloutsky, 2015; Best, Yim & Sloutsky, 2013). Infants have been shown to accurately generalize, like adults, from a set of evidence (e.g., '*ko* rabbit', '*ko* hamster', '*ko* chicken', therefore *ko*+animal), but quickly switch generalization hypotheses in the face of novel evidence, unlike adults who take more time (e.g., now add '*ko* ball', '*ko* house', therefore actually *ko*+noun, 6-8-month-olds versus adults; Best et al, 2013). Furthermore, when faced with novel situations, adults will quickly generalize any gleaned knowledge, but children will guide learning via directed exploration (ages 7-11 years old; Schulz, Wu, Ruggeri & Meder, 2019). Directed exploration, in contrast to random exploration, involves sampling selectively the uncertain hypotheses: after hearing '*ko* rabbit', '*ko* hamster', '*ko* chicken', an adult will thus expect 'ko dog' (*ko*+animal), but a child may choose to probe '*ko* book' (*ko*+noun). Thus both adults and children guide structure exploration based on hypothesized patterns, but children may place comparably more weight on uncertain options or more generally place less importance on certainty. Differences in attention and exploration could very well arise from the development of the cognitive apparatus (e.g., children have a more diffuse attentional system), but they could just as well arise from divergences in prior knowledge (e.g., adults know that categories boil down to one or few essential distinguishing features, adults know that a certain number of examples reliably predicts how the world is, adults have a better grasp of what relevant categories are likely to be).

A complementary way of probing how the set of evidence and age influence generalization is to model experimental results. Very basic, Bayesian ideal observer models, which posit hypotheses and then weigh the likelihood that they are correct, capture a vast array of lab results (e.g., Frank & Tenenbaum, 2011, but not all, e.g., Gervain & Endress, 2017). These models can account for observed disparities in generalization from different sets of evidence (e.g., a set of evidence that is compatible with one possible hypothesis versus multiple possible hypotheses, Frank & Tenenbaum, 2011). Though these models can mirror generalization results from different age groups, ranging from infants to adults, when two age groups with the same set of evidence generalize differently, additional parameters are often needed. For example, modeling divergences in generalization patterns of subordinate, object-level and superordinate words (e.g., Dalmatian, dog, and animal, respectively) in childhood and adulthood requires a supplementary 'object-level bias' parameter for adults (i.e., the assumption that a word refers to object-level things: dogs, spoons, or books; Xu & Tenenbaum, 2007). Though supplementary parameters can be intuitively and conceptually plausible (e.g., prior knowledge induces bias), they can appear rather ad hoc (Endress, 2013). This is especially the case when 'development' is synonymous with 'child versus adult'. There are indeed observed differences in generalization between children and adults, but there also divergences between children of different ages (e.g., Schulz et al, 2019; Gerken & Bollt, 2008, respectively). Few

studies however examine generalization through development, from infancy, into childhood and all the way to adulthood, on the same set of evidence. An infant may generalize subordinate, object-level and superordinate words in the same way as a 4-year-old, or she may not. It is thus difficult to determine whether the same factors drive divergences in generalization observed between children of different ages and between children and adults, and it is, accordingly, difficult to discern broad, non-*ad-hoc*, 'development' parameters for models. Our study thus investigated generalization of a grammatical rule in infancy, childhood and adulthood.

We adapted a paradigm, initially designed for infants, to be intuitive and pragmatically sound for all three age groups (Barbir, Babineau, Fiévet, & Christophe, in prep). The paradigm consists of a video training phase followed by a generalization test phase (Fig. 1A). Training videos are composed of short stories in the participants' native language (here, French), with the exception of a novel grammatical element (e.g., '*ko'* in '*Ko* rabbit is going to play'). In the videos, two novel determiners *ko* and *ka*, marked for animacy, replace existent French determiners (*le, la, un, une*; the, a) which are marked for grammatical gender. *Ko* preceded animates and *ka* inanimates (e.g., '*ko* rabbit', '*ko* chicken' and '*ka* shoe', '*ka* book'). The participants' task is to glean the patterns of use of *ko* and *ka*. One reliable source of information is the noun with which the determiner is paired. Participants would have to notice that nouns that appear with *ko* are animates, while those that appear with *ka* are inanimates. Then, in the test phase, participants' capacity to generalize this rule is measured. Participants see two images on the screen, a novel animate and a novel inanimate, and hear a prompting sentence with a novel content word (e.g., 'Look at *ko bamoule*', Fig. 1B). The only way to determine the referent of the novel content word is to use one's knowledge about the determiners *ko* and *ka*. If participants have learned the associations between *ko* and animates and *ka* with inanimates, then they could deduce whether the novel word is an animate or inanimate.

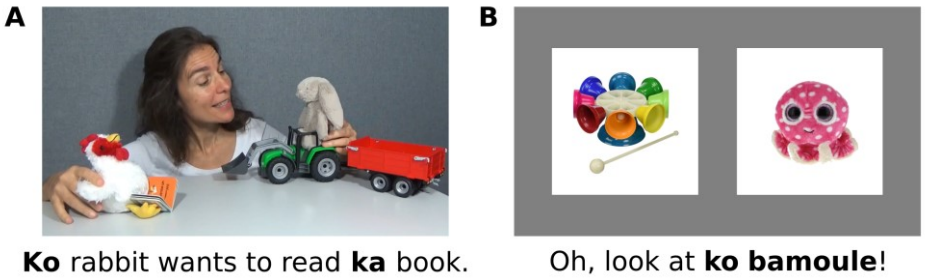


Fig. 1. Schematic of study design. (**A**) Participants first watch a training video with novel determiners (*ko* and *ka*), three times at home and once in the lab. (**B**) Then, they complete a test phase, during which they see two novel images (one animate and one inanimate) and hear a prompting sentence to look at one of the two images. The prompting sentence contains a novel noun participants did not know (e.g., *bamoule*). The only way to solve the task is to use knowledge about the associated novel determiners (e.g., *ko*+animate, therefore '*bamoule'* refers to the animate).

While it is pragmatically sound to teach a novel word in just one laboratory visit, as there may be a new object, action or concept that one of any age may just not have encountered before, it is not clearly so when teaching a novel grammatical element. Grammatical elements, especially determiners (the, a), are amongst the most frequent items across languages (e.g., English: Cutler, 1993; Japanese and Italian: Gervain, Nespor, Mazuka, Horie, & Mehler, 2008), and any new grammatical elements ought to embody this aspect too. Participants, thus, watch the training videos at home for three consecutive days prior to the lab visit, and then once more in the lab (total of 4 times; e.g., Wednesday, Thursday and Friday at home, and Saturday in the lab). The plausibility of encountering a novel grammatical element in one's native language, however, also declines with age. We, therefore, adapted the stories to promote learning of the

grammatical element at any age (Fig. 2). This was done as a precautionary measure; the stories from Barbir et al (Fig. 2A) have been shown elicit generalization in all three age groups (infants and adults: Barbir et al; pre-school children: Babineau, de Carvalho, Trueswell, & Christophe, submitted). First, we increased the quantity and variety of exposure. Increasing exposure has been shown to augment the probability that participants will generalize a novel, even outlandish, structure (e.g., Gebhart et al, 2009). Second, we paired the new grammatical elements with a handful of novel content words (e.g., *Ko nuve* loves to play, Fig. 2B). As such, older participants, though likely acquainted with all of the French words used in the stories, would encounter a 'word learning situation', a situation in which they would have to rely on, and as such be attentive to, the linguistic context (as well as the visual context) to determine the meaning of the novel content words. Third, we presented the test items, to which participants would have to generalize, alongside corresponding novel content words in a referentially ambiguous context (e.g., "Look at them all! There's *ko pirdale, ko doripe, ka bradole,* and *ka bamoule*", Fig. 2D). As such, older participants would not be tempted to name a novel test item with a real French word (e.g., that thing looks a bit like a pig, so it must be called 'pig'). Lastly, we reduced extraneous linguistic and visual information to a minimum in simple referential scenes where participants saw a static image with a neutral sentence containing a novel grammatical element (e.g., Look at *ka* ball, Fig. 2C). Children, who have a more diffuse attention, may get snarled up in a plethora of irrelevant information during stories, but may be more likely to focus on the most critical elements in simple referential scenes (e.g., children's attention can be successfully drawn to critical elements, Deng & Sloutsky, 2015). The adaptations were designed to reduce, as much as possible, 'task perception' related noise between age groups.

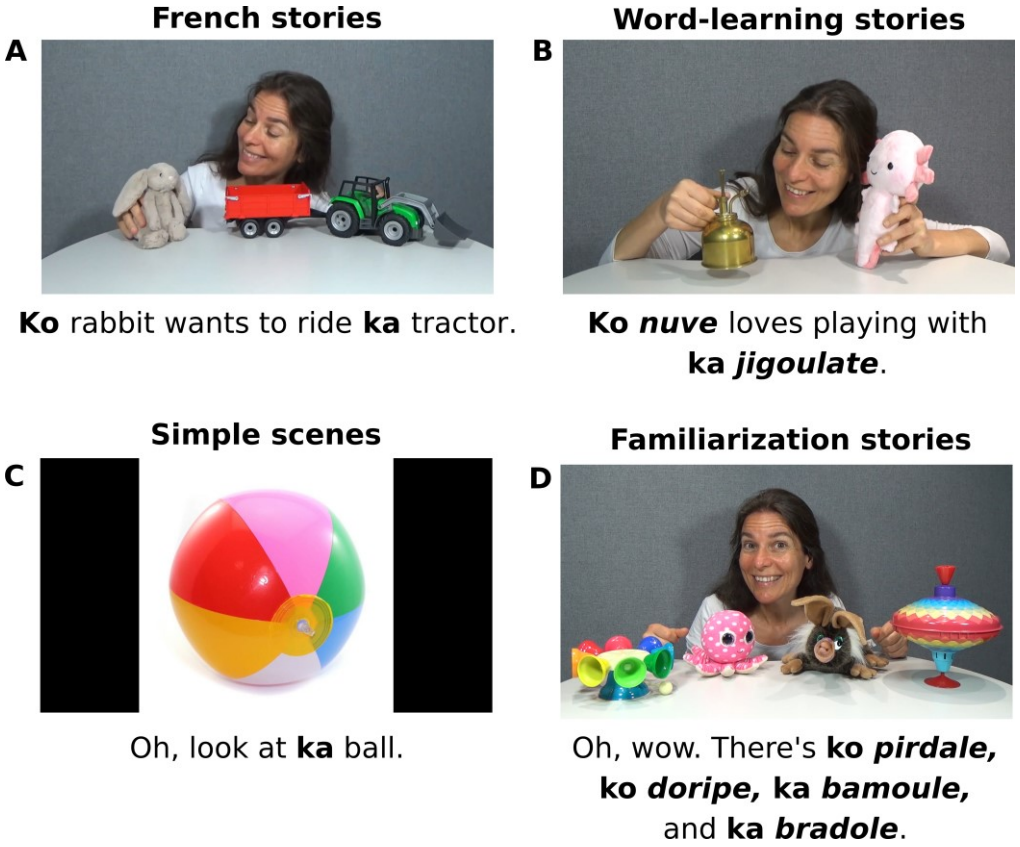


Fig. 2. Training video stories. (**A**) French stories (same as in Barbir et al). Stories are entirely in French, except for the novel determiners. (**B**) Word-learning stories. Novel determiners are presented alongside novel nouns. **(C)** Simple scenes. Static images are displayed while simple sentences are played. **(D)** Familiarization stories. Novel toys and novel nouns, used during test, are presented in referentially ambiguous scenes.

To promote learnability, we chose a grammatical distinction that exists in human languages but not in the participants' native language, French, and that is salient from a very young age: animate versus inanimate (for review, see: Strickland, 2017). Very early in life, infants demonstrate the capacity to distinguish animates from inanimates (e.g., Mandler & McDonough, 1993; McDonough & Mandler, 1998). Notably, however, 20-month-old infants are capable of learning, and generalizing, a grammatical distinction based on animacy (Barbir et al). Adults too can learn to generalize an animacy-related grammatical distinction (Leung & Williams, 2012; Barbir et al). Given that infants and adults can acquire a grammatical distinction based on animacy in the lab, it seems highly likely that older children, of pre-school age, too could. Fewer studies have directly tested an animate versus inanimate grammatical distinction with pre-school aged children (Cf. supplemental studies of Barbir et al). Nevertheless, preschool children have been shown to be able to learn and extend a category label for animals (e.g., cats, dogs and fish are *bamoules,* what else is a *bamoule*?; 3 and 4-year-old children, Xu & Tenenbaum, 2007) and to learn grammatical distinctions between other categories, such as aliens and planets (Culbertson, Smith, Jarvinen, & Haggarty, 2018). In choosing a highly salient distinction and one that does not require a large quantity of world knowledge, such as animate-inanimate, our goal was to give learners of all ages as equal as possible a chance of generalizing.

We tested 20-month-old infants (n = 27), 4-year-old children (n = 25) and adults (n = 26) in a pre-registered study. There were two main differences between the three age groups: knowledge about their native language (and more broadly the world) and cognitive maturity (e.g., focused versus diffused attention). Intuitively, knowledge about one's native language increases with age. 20-month-old infants have rudimentary knowledge about their native language: they already understand over a 100 words (Fernald, Marchman, & Weisleder, 2013) and even have knowledge of basic grammatical categories (e.g., nouns and verbs, He & Lidz, 2017). 4-year-old children, though still actively learning new content words and complex grammatical structures (e.g., complex passives, Lempert, 1978), are quite competent in their native language: they understand the meanings of very abstract words like know and think (Dudley, Orita, Hacquard, & Lidz, 2015) and comprehend less frequent grammatical structures, such as simple passive sentences (Messenger & Fisher, 2018). Adults, in stark contrast, are not actively learning their native language; they are proficient speakers. Similarly, cognitive functions mature with age. Though infants and children have diffused attention, adults' attention is focused (Best et al, 2013; Deng & Sloutsky, 2015). Intuitively, we could expect two basic patterns: generalization decreases with age (more knowledge reduces the hypothesis space; Gerken & Bollt, 2008) or generalization increases with age (more knowledge or cognitive maturity aids in the generalization of sparse categories such as animates and inanimates[15]; Deng & Sloutsky, 2015). Yet, results from prior studies suggest that adults can generalize a novel animate-inanimate grammatical distinction (Leung & Williams, 2012, and supplemental study from Barbir et al) and that infants too can generalize the animate-inanimate distinction (Barbir et al, via the same paradigm but a slightly different evidence set). We thus expected both adults and infants to generalize the animate-inanimate distinction in our study. However, if adults generalize and so do infants, pre-school children ought not encounter difficulties generalizing because of prior knowledge (adults who have a lot of knowledge about their native language appear to generalize) nor because of diffuse attention (infants who also have diffuse attention generalize the sparse animate-inanimate distinction). We hypothesized that if all three age groups generalize, then there may be cases in which knowledge does not impede generalization (perhaps with enough exposure, e.g., Gebhart et al, 2009) and in which

[15] Animates are likely a more dense category than inanimates. However, the category 'animate' is nonetheless comparatively sparse.

sparse categories can be learned in spite of diffuse attention (perhaps when they contain highly salient features like eyes, e.g., Jones, Smith, & Landau, 1991). Alternatively, if infants and adults generalize but not pre-school children, there may be cases in which generalization depends on some additional factor, or there may be an entirely different, hitherto unreported, factor underlying generalization patterns *in general*.

We used two behavioral measures to determine whether participants generalize: gaze and explicit choice. We investigated gaze focus using two metrics: the fine-grained 'looking-while-listening' method which encodes the evolution of gaze patterns from moment-to-moment during the trial, and the broader 'preferential looking' method which sums overall gaze time to a target during the trial (Fernald, Zangl, Portillo, & Marchman, 2008). These two metrics allow for a complementary and comprehensive analysis, with both specific information as to when precisely participants orient their gaze to the correct image, as well as a way to capture general preference in light of individual variation (e.g., some participants might look right away to the target image, others may take more time and only do so toward the end of the trial). We expected participants to look more to the animate image when they heard '***ko*** *bamoule'* (a novel noun with the animate determiner) than when they heard '***ka*** *bamoule'* (a novel noun with the inanimate determiner), if they were generalizing the grammatical rule. In addition, we examined explicit choices, as indexed by pointing, for pre-school children and adults. In a similar vein as with gaze, we expected participants to choose the animate image more often when they heard '***ko*** *bamoule'* than when they heard '***ka*** *bamoule'*, if they were generalizing the rule. Multiple measures allowed us to capture a potentially weak effect in noisy data. In addition, we ran the same analyses on the aggregated data from the current study and a like study, with the same grammatical distinction (animate-inanimate), a similar paradigm (videos and test), and an overlapping set of evidence, to increase power (from Barbir et al, infants n = 51, pre-school children n = 100, adults n = 46, see Materials and Methods for more detail). Our results reveal that the same set of evidence spurs generalization in infants and adults, but not pre-school children.

3.1.1 Results

A. Experiment 1: Infants

Participants were 20-month-old French-learning infants, who heard less than 10% of another language (n = 27). Infants saw a total of 8 test trials, with novel items and nouns, half animate and half inanimate (e.g., Look at ***ko*** *bamoule*). There were an additional 8 filler trials with French nouns (e.g., Look at ***ko*** dog; analyses in SI). An eye-tracker recorded infants gaze during the trial. Gaze data was analysed from the onset of the determiner to the moment in the trial at which older participants were prompted to point (i.e., 2s after the second repetition of the auditory sentence, 0-5850ms, Fig. 3).

Looking-while-listening.
To determine whether participants were generalizing the grammatical rule, we investigated the proportion of looks to the animate image when participants heard a sentence with the animate determiner *ko* compared with when they heard one with the inanimate determiner *ka* (e.g., 'Oh look at *ko bamoule'* versus 'Oh look at *ka pirdale*') at each time-point in the trial (0-5850ms). A cluster-based permutation analysis (Maris & Oostenveld, 2007) revealed that infants looked more to the animate image when their heard *ko* during the time-window from 120 to 1100ms after hearing the novel determiner (p = 0.028, Fig. 3A; in this paper, we report all clusters that have a p-value equal to or less than 0.1). Infants, thus, oriented their gaze as a function of the

determiner's animacy, indicating that they were able to generalize the rule governing determiner use to novel, unknown, nouns.

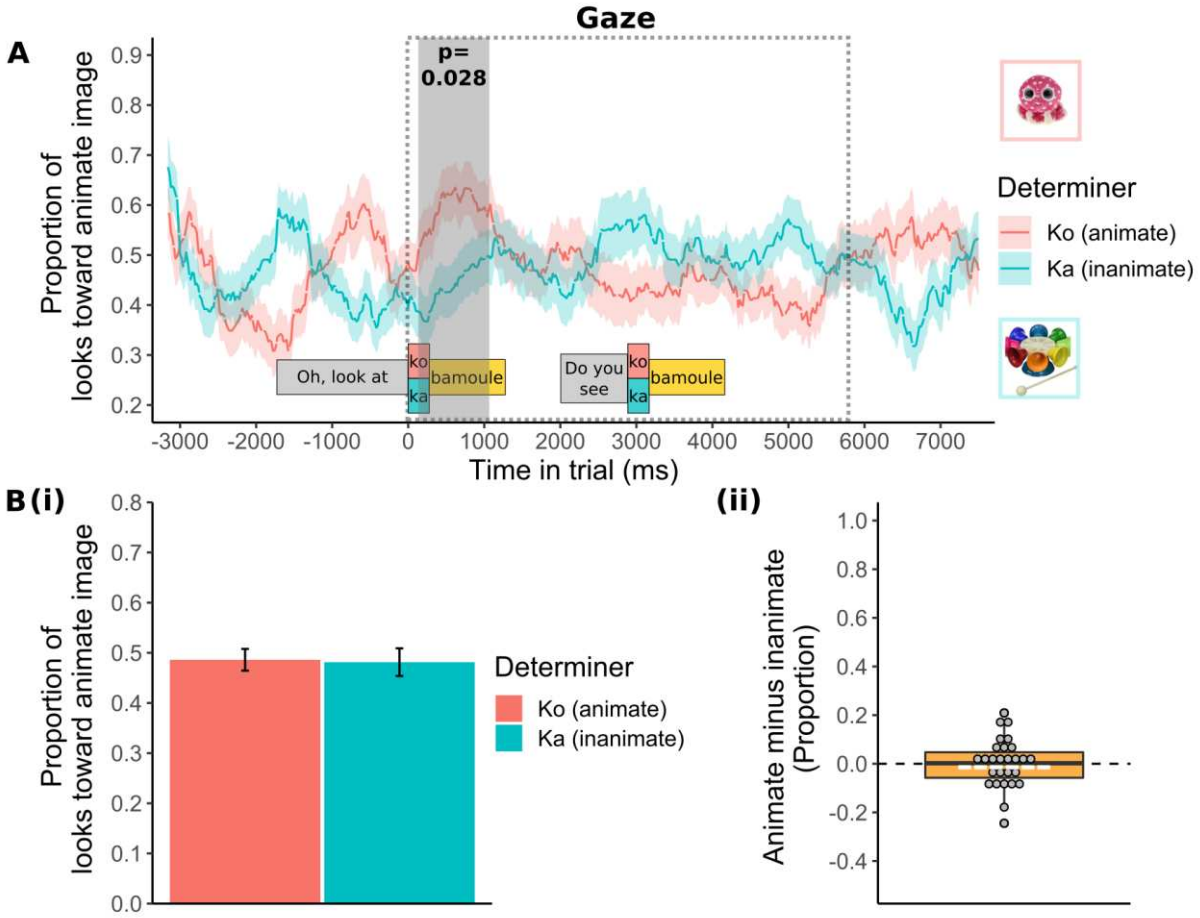


Fig. 3. Experiment 1 Results (Infants). (A) Time-course. Proportion of looks toward the animate image at each point in time during the trial, when the infants heard the animate determiner *ko* paired with a novel noun (e.g. *ko bamoule*) in pink and when they heard the inanimate determiner *ka* paired with a novel noun (e.g., *ka pirdale*) in blue. Dark lines represent mean across participants, and light shading the SEM (95% confidence intervals of the mean). The grey dotted square represents the time from the onset of the determiner to the moment older age groups were prompted to point (0-5850). Grey shading indicates the time-window during which the two conditions, 'animate determiner sentences' versus 'inanimate determiner sentences', diverge (120-1100ms, p = 0.028, cluster-based permutation analysis). Other visible divergences were not significantly different than what could be found by chance (all p > 0.1; -1940- -1340ms: p = 0.32; -940- -280ms: p = 0.165; 2600-3360ms: p = 0.284; 4800-5380ms: p = 0.356; 6280-6840ms: p = 0.365). **(B) Overall looking preference.** Proportion of looks toward animate image averaged over the whole trial, when the infants heard the animate determiner *ko* paired with a novel noun in pink and when then heard the inanimate determiner *ka* paired with a novel noun in blue. Error bars represent SEM (95% confidence intervals of the mean). Infants did not look significantly longer to the animate image when they heard the determiner *ko* than the determiner *ka*. (**C**) **Difference per participant.** The difference between the proportion of looks toward the animate image when the animate determiner was heard and when the inanimate determiner was heard, per participant. Dots indicate participants. Dashed white line indicates mean. Upper and lower regions of the box indicate the first and third quartiles (25^{th} to 75^{th} percentiles). The upper whisker represents the third quartile up to the 1.5 interquartile smallest value, while the lower whisker the 1.5 interquartile smallest value to the first quartile. Dashed line black indicates no difference between proportion of looks the animate image, when infants heard the animate determiner minus when they heard the inanimate determiner.

Preferential looking.
Next, we ran an analysis on the proportion of overall looking time (0-5850ms) to the animate image when participants heard the animate determiner *ko* versus when they heard the inanimate determiner *ka*, as a complementary measure of generalization. Means are displayed in Fig. 3B. A mixed-effects regression analysis showed that infants did not look significantly longer for the whole the duration of the trial to the animate image when they had heard *ko* than when they had heard *ka*: $\beta = -0.009$, SE = 0.04, $t = -0.23$, $p = 0.82$, Cohen's $d = 0.03$ (Fig. 3B). Though infants used the determiner's animacy to direct their gaze right after hearing the determiner, the determiner did not appear to drive a general gaze preference.

Gaze data indicates that infants are able to and do generalize the features associated with the novel determiners (namely, animate-inanimate) to unknown nouns, and that this effect is localized in time, right after hearing the determiner.

Aggregated data.
The current experiment replicated results from Barbir et al. at a like time-window (early onset effect: 300-2440ms). The protocol was the same in the two experiments, but the evidence set differed in the quantity and variety of exposure. To pinpoint across-experiment trends and to increase overall statistical power, we aggregated the data from the two experiments (20-month-old infants $n = 51$, trials $n = 336$), and ran the same analyses. Analyses on aggregated data were not pre-registered, and accordingly, to be as conservative as possible, we analysed the whole time-window from the onset of the determiner to the end of the trial, approximately 2s after pre-school children and adults were prompted to point (0-7500ms).

A cluster-based permutation analysis confirmed that infants looked significantly more to the animate image when they heard a sentence with *ko* than when they had heard one with *ka* during an early time-window: from 140 to 1340ms after hearing the novel determiners ($p = 0.01$, Fig. 4A). A mixed-effects regression analysis revealed that infants did not look significantly longer, on average during the trial, to the animate image when they heard a sentence with *ko* than when they one with *ka*: $\beta = 0.04$, SE = 0.03, $t = 1.39$, $p = 0.17$, Cohen's $d = -0.14$ (Fig. 4B(i)).

Thus, 20-month-old infants appear to exploit the animacy feature of the determiner when interpreting novel nouns, generalizing to novel contexts right after hearing the determiner. However, infants' generalization is localised in time, and is not manifest in a significant overall looking preference to one of the images.

B. Experiment 2: Pre-school children

Participants were 4-year-old French-speaking children, who did not hear more than 10% of another language ($n = 25$). Pre-school children did the same experiment as infants. They were explicitly told to pay careful attention to the stories because they would hear important clues for the game that followed (test phase). At the end of each trial, they were asked to point to the image they had been told to look at. An eye-tracker recorded the direction of their gaze.

Looking-while-listening.
A cluster-based permutation analysis revealed that there were no time-windows during which kids looked significantly more to the animate image when they heard the animate determiner *ko* than when they heard the inanimate determiner *ka* (Fig. 5A). Children, thus, did not appear

to be directing their gaze in accordance with the animacy feature of the determiners at specific points in time during the trial.

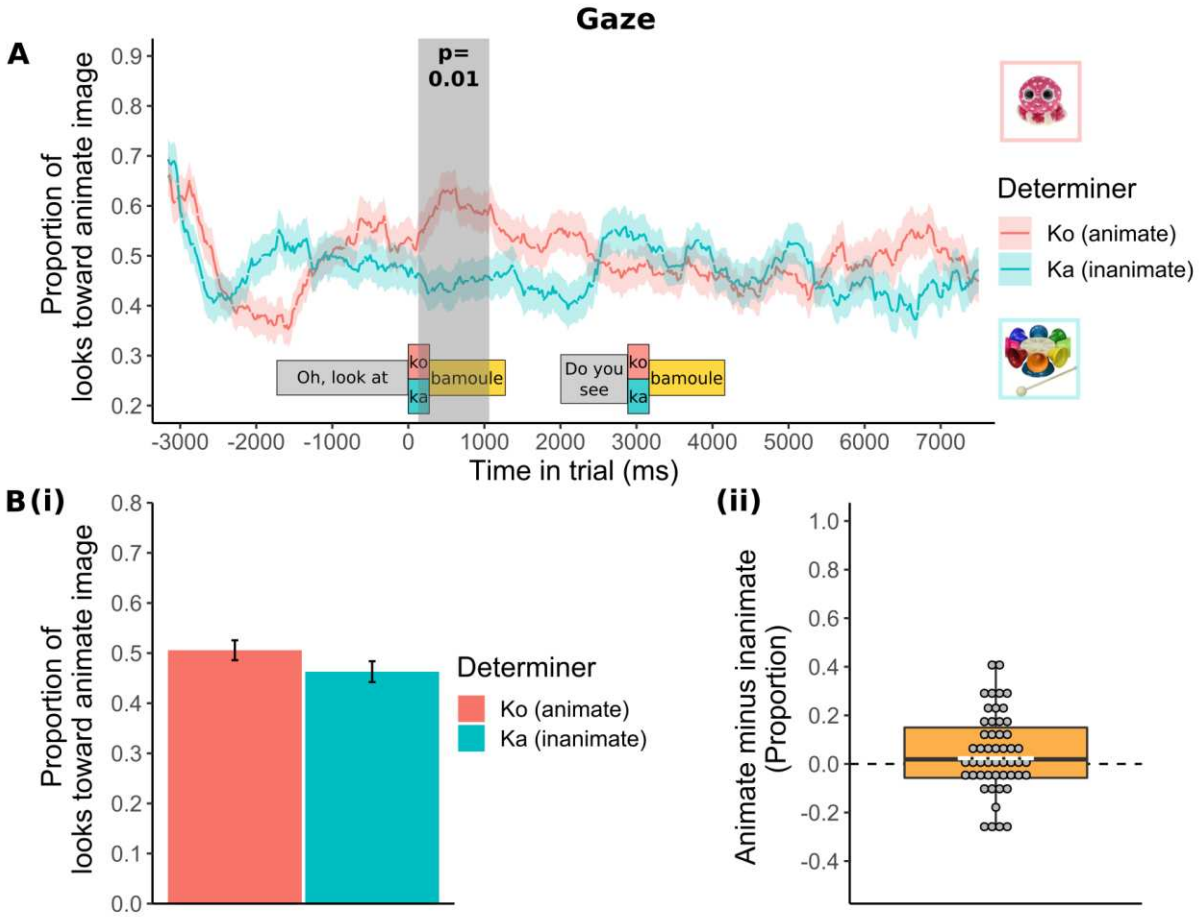


Fig. 4. Aggregated Results: Infants. Data from current study and Barbir et al: infants *n* = 51. **(A) Time-course.** Infants looked longer to the animate image when they heard the animate determiner than when they heard the inanimate determiner from (140-1340ms, p = 0.01). Other visible divergences were not significantly different than what could be found by chance (all p > 0.1). **(B) Overall looking preference.** Infants did not look significantly longer to the animate image when they heard the determiner *ko* than when they heard *ka*. (C) **Difference per participant.**

Preferential looking.

A mixed-effects regression analysis showed that kids did not look longer overall, throughout the whole trial, to the animate image when they had heard the animate determiner *ko* than when they had heard the inanimate determiner *ka*: β = 0.02, SE = 0.03, t = 0.72, p = 0.48, Cohen's d = 0.08 (means displayed in Fig. 5B(i)). As such, children did not seem to guide gaze focus using determiner features.

Explicit choice.

Next, we investigated the proportion of animate image choices when children heard the animate determiner *ko* and when they heard the inanimate determiner *ka*, to determine whether children generalized when they had to choose an image explicitly. A generalized linear mixed-effects model showed that children did not choose the animate image significantly more often when they heard the determiner *ko* than when they heard the determiner *ka*: β = 0.55, SE = 0.41, model comparison: $\chi^2(1) = 1.73$, p = 0.19, Cohen's d = 0.24 (means in Fig. 5C(i)). Explicit

choice data and gaze data provide convergent evidence that children do not seem to generalize, at least significantly, the grammatical rule to novel nouns.

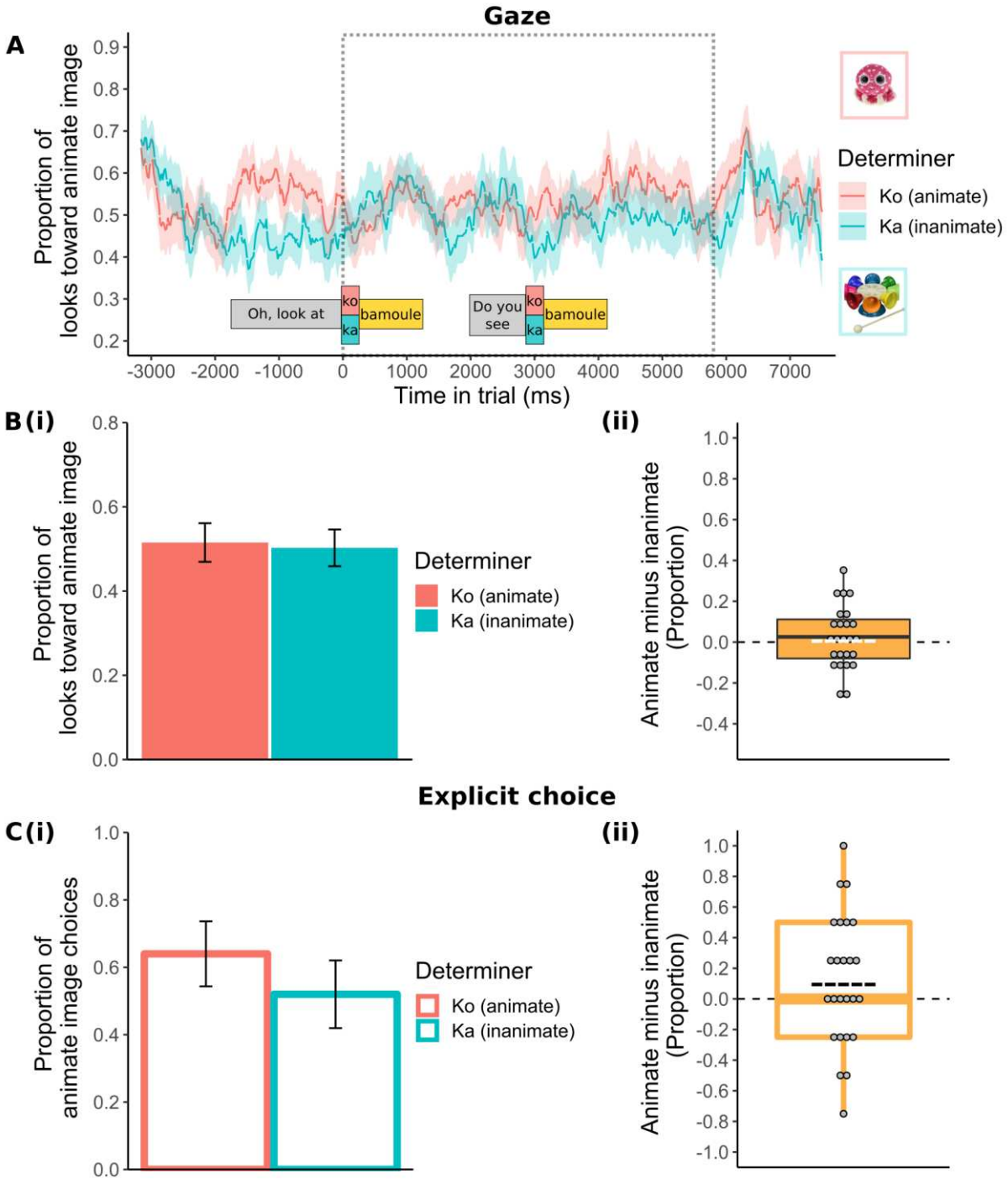


Fig. 5. Experiment 2 Results (Pre-school children). (A) Time-course. Pre-school children did not look more to the animate image when they heard the animate determiner *ko* and a novel noun than the inanimate determiner *ka* and a novel noun at any time-point during the trial. **(B(i)) Overall looking preference.** Pre-school children did not look significantly longer to the animate image when they heard the animate determiner. (**B(ii)**) **Overall looking preference per participant. (C(i)) Explicit choice.** Proportion of animate image choices (pointing) when the pre-school children heard the animate determiner *ko* in pink and when they heard the inanimate determiner *ka* in blue. Pre-school children did not choose the animate image more often when they heard the animate determiner. **(C(ii)) Explicit choice per participant.** The per participant difference between the proportion of animate image choices when participants heard the animate and inanimate determiners.

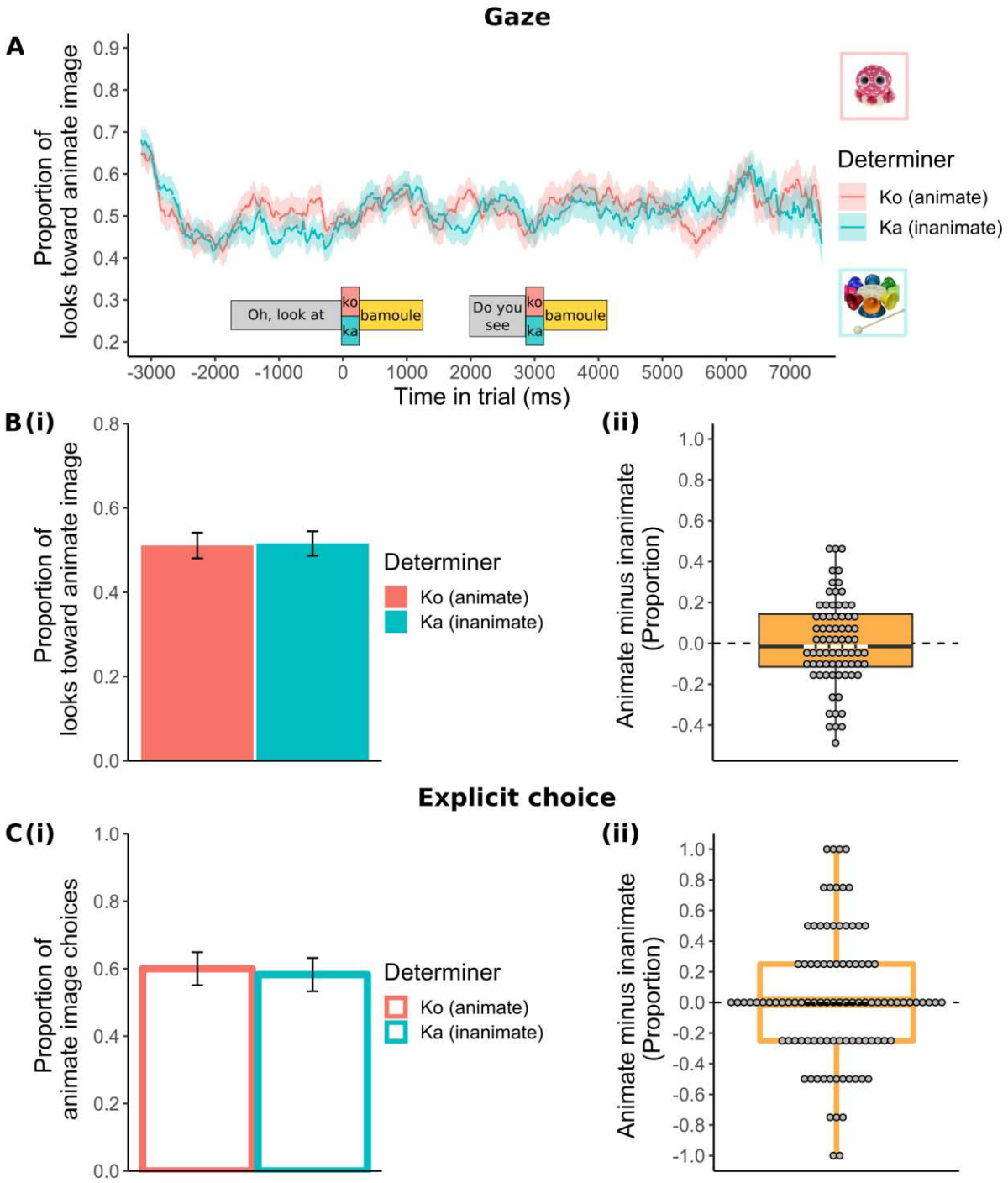


Fig. 6. Aggregated Results: Pre-school Children. Data from current study, supplemental study and two supplemental studies from Barbir et al: children n = 74 (gaze data) and children n = 100 (explicit choice data). **(A) Time-course.** Pre-school children did not look more to the animate image when they heard the animate determiner than when they heard the inanimate determiner at any point during the trial. **(B (i)) Overall looking preference.** Pre-school children did not look significantly longer to the animate image when they heard the animate determiner than the inanimate determiner. **(B(ii)) Overall looking preference per participant. (C(i)) Explicit choice.** Pre-school children did not choose the animate image more often when they heard the animate determiner than the inanimate determiner: they chose the animate image more often than chance in both cases ($p < 0.001$). **(C(ii)) Explicit choice per participant.**

These results cannot be reduced to general confusion about the task. Children did not have trouble interpreting French nouns when paired with novel determiners: mean accuracy on filler trials was 96.5% ± 1.3 (see SI for an analysis of gaze data for filler trials).

Aggregated data.
Children's data, however, may have been very noisy, occulting an effect of generalization. We thus aggregated data from pre-school children (ages 3-6) in four experiments: the main experiment, the supplemental experiment (see SI), and two supplemental experiments from Barbir et al. All four experiments share a common protocol (videos then test phase) and the same determiner distinction (animacy), but differ in the quantity and diversity of the evidence set. Notably, only the main experiment involves a four-day training phase; in the three other experiments children saw the training video just once and were tested immediately afterward, on the same day (one-day experiments). In total, there was gaze data from three experiments (children $n = 74$; total of 535 trials, mean age: 4;9 years; range: 4;2-6;0 years; 36 girls, 38 boys) and explicit choice data from four experiments (children $n = 100$; mean age: 4;6 years; range: 3;1-6;0 years; 47 girls, 55 boys).

A cluster-based permutation analysis confirmed that children did not look significantly more to the animate image when they heard the animate determiner *ko* than when they heard the inanimate determiner *ka*, during any time-window (Fig. 5A). A generalized linear mixed-effects model too showed that children did not look significantly more, overall, at the animate image when they heard the animate determiner *ko* than when they heard the inanimate determiner *ka*: $\beta = -0.002$, SE = 0.02, t = -0.07, p = 0.94, Cohen's d = -0.02 (means shown in Fig. 5B(i)). A generalized linear mixed-effects analysis affirmed that children did not choose the animate image more often when they had heard the animate determiner *ko* than when they had heard the inanimate determiner *ka*: $\beta = 0.02$, SE = 0.21, model comparison: $\chi^2(1) = 0.004$, p = 0.95, Cohen's d = 0.04 (means in Fig. 5C(i)). However, the model also revealed a strong animacy bias: children chose the animate image more often than chance: $\beta = 0.45$, SE = 0.11, $p < 0.001$.

The aggregated data, like data from the main experiment, indicate that children do not appear to be generalizing determiner features to novel nouns.

C. Experiment 3: Adults

Participants were native French-speaking adults, who had heard French from birth ($n = 26$). Adults did the same experiment as infants and pre-school children. They were told that they were participating in an experiment for children (see SI for full instructions). Like pre-school children, they were prompted to point to the image they had been told to look at. An eye-tracker recorded the direction of their gaze. At the end of the experiment, a post-experiment questionnaire was administered to determine whether adults were aware of the rule governing determiner use and whether generalization was driven by explicit knowledge.

Looking-while-listening.
A cluster-based permutation analysis revealed that adults looked more to the animate image when they heard *ko* than when they had heard *ka* during the time-window from 740 to 5850ms (the end of the analysed time-window) after the onset of the novel determiners ($p < 0.001$, Fig. 7A). Adults, thus, appeared to generalize the determiner features to novel nouns.

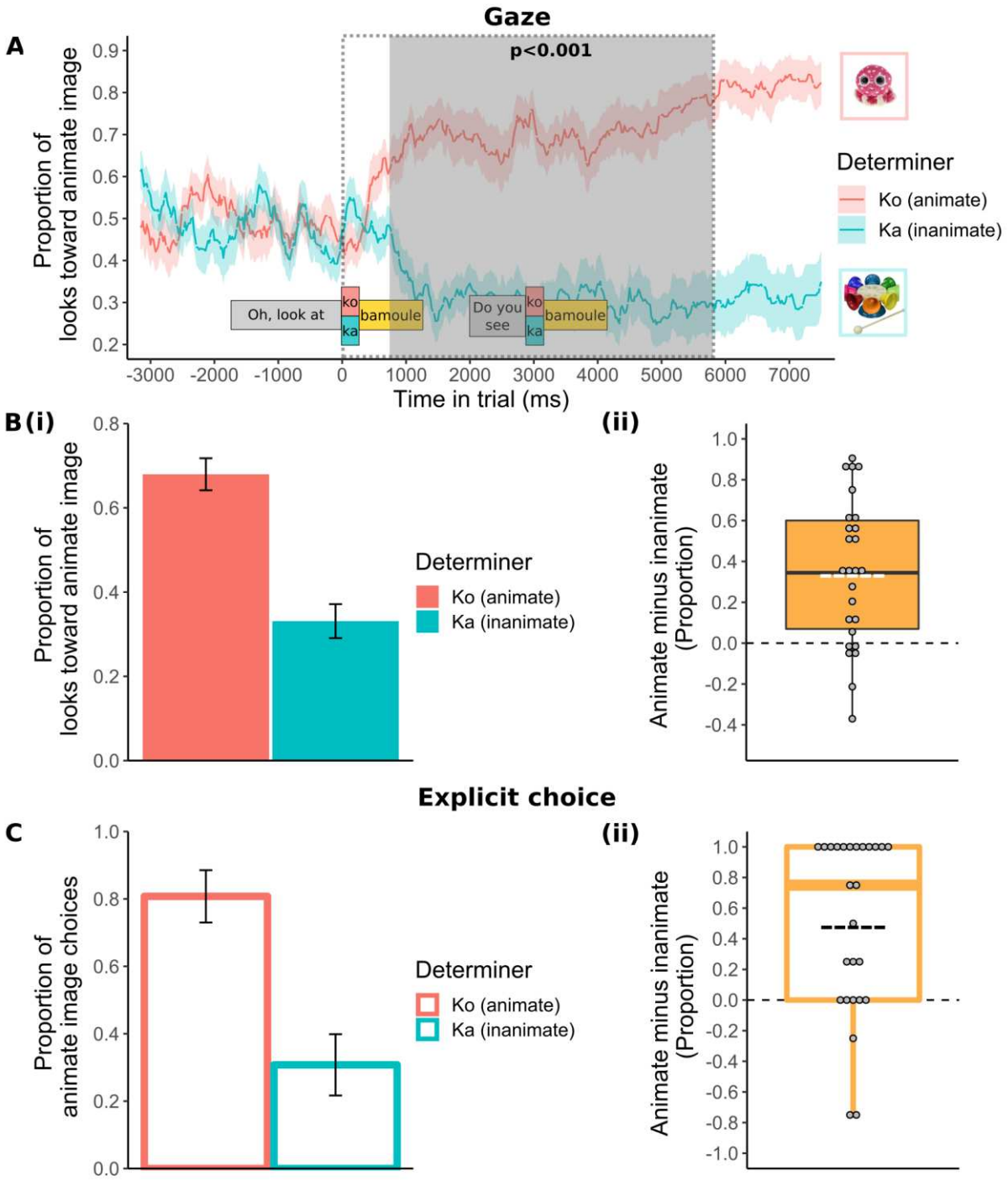


Fig. 7. Experiment 3 Results (Adults). (A) Time-course. Adults looked more to the animate image when they heard the animate determiner *ko* than the inanimate determiner *ka* during the time-window: 740-5850ms (the moment participants were prompted to point). **(B(i)) Overall looking preference.** Adults looked significantly longer to the animate image when they heard the animate determiner than the inanimate determiner ($p < 0.001$). (**B(ii)) Overall looking preference per participant. (C(i)) Explicit choice.** Adults choose the animate image more often when they heard the animate determiner than the inanimate determiner ($p < 0.001$). **(C(ii)) Explicit choice per participant.**

Preferential looking.

A mixed-effects regression analysis confirmed that adults looked longer overall, on a whole trial basis, to the animate image when they had heard the animate determiner *ko* than when they

had heard the inanimate determiner *ka*: $\beta = 0.3$, SE = 0.07, $t = 5.04$, $p < 0.001$, Cohen's $d = 1.31$ (means shown in Fig. 7B(i)). Adults therefore seemed to focus their gaze on the image that corresponded to the features of the determiner (i.e., animate-inanimate), indicating that they were indeed generalizing.

Explicit choice.
A generalized linear mixed-effects model showed that adults chose the animate image significantly more often when they heard the animate determiner *ko* than when they heard the inanimate determiner *ka*: $\beta = 5.03$, SE = 1.61, model comparison: $\chi^2(1) = 15.15$, $p < 0.001$, Cohen's $d = 1.16$ (means in Fig. 7B(i)). Explicit choice data and gaze data thus provide convergent evidence that adults were generalizing the rule governing determiner use to infer the potential meaning of novel nouns.

Kind of knowledge.
To probe whether adults were generalizing via explicit rule application (e.g., *ko*+animate, *ko*+*bamoule*, therefore *bamoule* is an animate), we grouped together adults who could explicitly state the rule governing determiner use, during the post-experiment questionnaire (Explicit Knowledge group, $n = 14$) and those who could not (No Explicit Knowledge group, $n = 12$). We measured between-group generalization differences as well as group-specific levels of generalization. To probe group-specific performance, an analysis was run on each group separately, and the alpha was Bonferroni-corrected for multiple comparisons, $\alpha = 0.05/2$.

Two separate cluster-based permutation analysis revealed that the Explicit Knowledge group looked more to the animate image when they heard the animate determiner *ko* than when they had heard the inanimate determiner *ka* during the time-window from 480 to 5850ms ($p < 0.001$, Fig. 8A); but the No Explicit Knowledge group did not look significantly more to the animate image when they heard *ko* than when heard *ka* at any time-point during the trial (there was a trend in that direction, during the time-window from 4060-5300ms, but it did not reach significance, $p = 0.06$, Fig. 8B). Thus, adults appear to be orienting their gaze to the image that matches determiner animacy, for the most part, via explicit rule application.

A mixed-effects regression analysis showed that the Explicit Knowledge group looked more overall, throughout the trial, to the animate image when they heard the animate determiner *ko* than when they heard the inanimate determiner *ka*, when compared with the No Explicit Knowledge group [interaction between Knowledge and Determiner: $\beta = 0.34$, SE = 0.12, $t = 2.7$, $p = 0.013$, Fig. 9A(i)]. Two separate mixed-effects regressions showed that the Explicit Knowledge group looked longer overall to the animate image when they had heard the animate determiner *ko* than when they had heard the inanimate determiner *ka* [$\beta = 0.51$, SE = 0.1, $t = 5.06$, $p < 0.001$, Cohen's $d = 2.18$, means displayed in Fig. 9A(i)]; and though there was a trend for the No Explicit Knowledge group to look longer to the animate image when they heard the animate determiner *ko* than the inanimate determiner *ka*, it did not reach significance [$\beta = 0.17$, SE = 0.07, $t = 2.55$, $p = 0.027$, Cohen's $d = 0.62$, means in Fig. 9A(i)]. Accordingly, adults appear to generalize more when they have explicit knowledge, but we cannot conclude that generalization was uniquely driven by explicit knowledge.

A generalized linear mixed-effects analysis revealed that the Explicit Knowledge group chose more often the animate image when they heard the animate determiner *ko* than the inanimate determiner *ka*, when compared with the No Explicit Knowledge group [interaction between Knowledge and Determiner: $\beta = 5.93$, SE = 3.02, model comparison: $\chi^2(1) = 4.97$, $p = 0.026$, Fig. 9B(i)]. Two distinct generalized linear mixed-effects models showed that the

Explicit Knowledge group chose the animate image significantly more often when they had heard the animate determiner *ko* than when they had heard the inanimate determiner *ka* [β = 12.57, SE = 5.35, model comparison: $\chi^2(1) = 15.53$, $p < 0.001$, Cohen's d = 1.75, means shown in Fig. 9B(i)]; but that this was not the case for the No Explicit Knowledge group [β = 2.08, SE = 1.32, model comparison: $\chi^2(1) = 2.25$, $p = 0.12$, Cohen's d = -0.67, means displayed in Fig. 9B(i)].

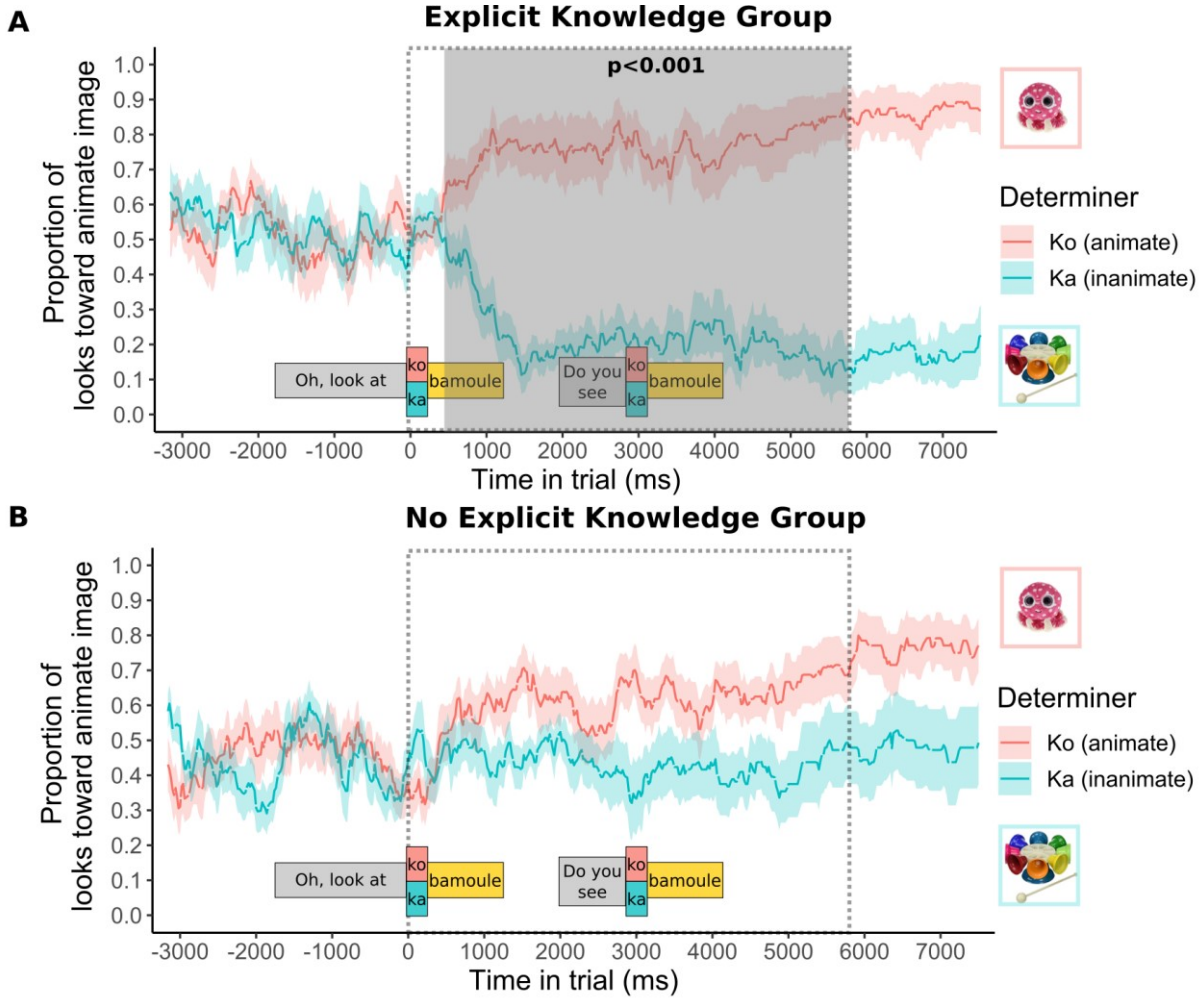


Fig. 8. Experiment 3 Knowledge Kind (Adults): Gaze Time-Course. Explicit Knowledge Group (*n* = 14), No Explicit Knowledge Group (*n* = 12). **(A) Explicit Knowledge Group.** Adults who had explicit knowledge of the rule governing determiner use looked more to the animate image when they heard the animate determiner *ko* than the inanimate determiner *ka* during the time-window: 480-5850ms (the moment participants were prompted to point). **(B) No Explicit Knowledge Group.** Adults who did not have explicit knowledge of the rule governing determiner use did not look significantly more to the animate image when they heard the animate determiner *ko* than the inanimate determiner *ka* at any time-point during the trial.

Accordingly, both gaze data and explicit choice data provide convergent evidence that explicit knowledge appears to boost generalization. Nevertheless, these results do not unequivocally indicate that adult generalization is solely the result of explicit knowledge. Trends may have been diluted in noise because these analyses were based on a small sample size (*n* = 12-14).

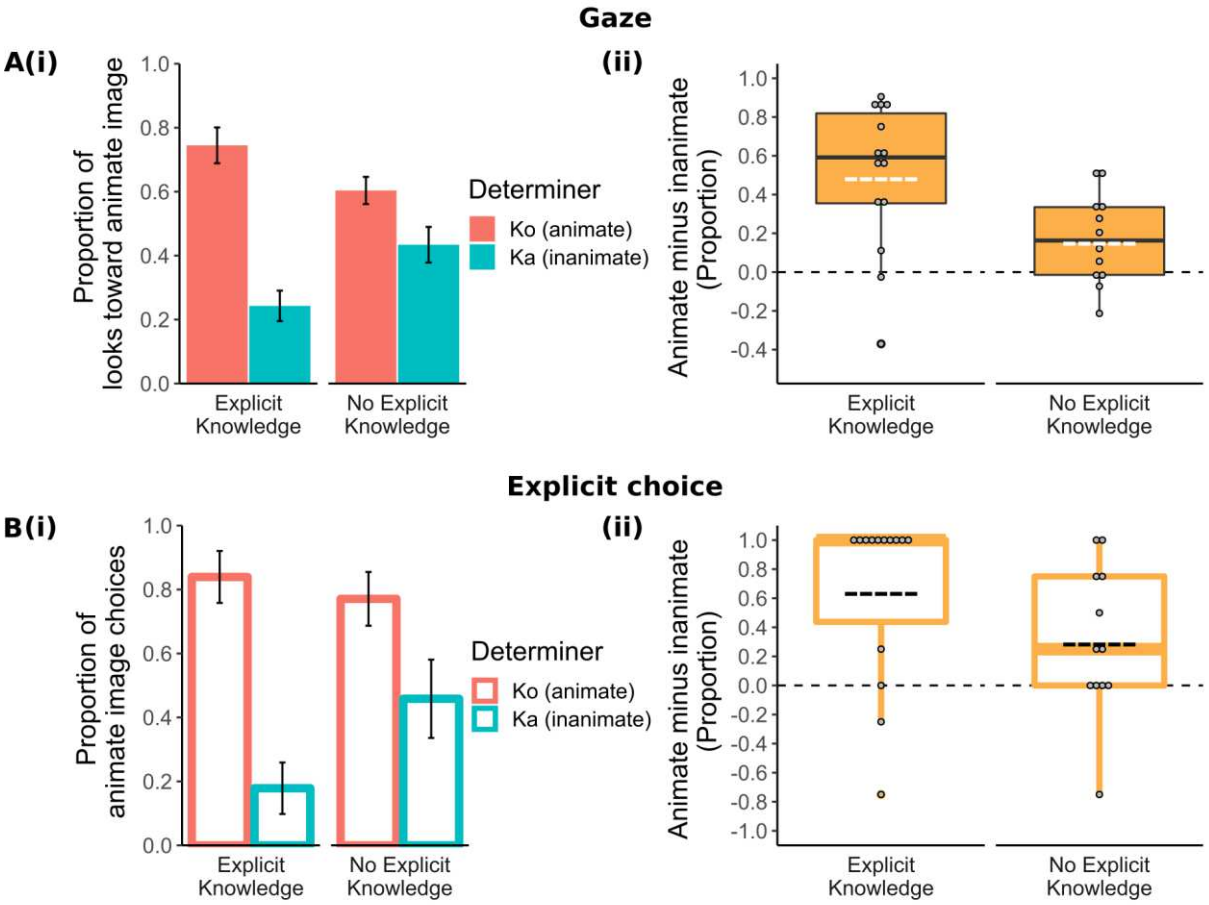


Fig. 9. Experiment 3 Knowledge Kind (Adults): Gaze & Explicit Choice Results. (A(i)) Overall looking preference per knowledge group. The Explicit Knowledge group was significantly better at distinguishing the two determiners than the No Explicit Knowledge group ($p = 0.018$). The Explicit Knowledge group looked significantly longer to the animate image when they heard the animate determiner ($p < 0.001$), but the No Explicit Knowledge group did not ($p = 0.027$, not significant after Bonferroni correction). (**A(ii)**) **Overall looking preference per participant and per knowledge group. (B(i)) Explicit choice.** The Explicit Knowledge group was significantly better at distinguishing the two determiners than the No Explicit Knowledge group ($p = 0.026$). The Explicit Knowledge group chose the animate image significantly more often when they heard the animate determiner ($p < 0.001$), but the No Explicit Knowledge group did not ($p = 0.12$). (**B(ii)**) **Explicit choice per participant and per knowledge group.**

Aggregated data.

To investigate first whether more statistical power would reveal effects weakened by noise in each individual experiment, we aggregated the adult data from a supplemental study in Barbir et al (adults $n = 20$) and the current study (adults $n = 26$, total $n = 46$; there were no significant differences in performance between the two experiments, analyses reported in the SI). The Barbir et al study had a similar protocol (videos and then test) and the same determiner distinction (animacy), but differed in the quantity and variety of exposure. Notably, the Barbir et al study was a one-day study: participants saw a 7-minute video and then performed the test. The aggregate data had a total of 24 participants who could explicit state the rule governing determiner use (One Day: $n = 10$; Four Day: $n = 14$) and 22 who could not (One Day: $n = 10$; Four Day: $n = 12$). The alpha for subgroup analyses (Explicit Knowledge group/No Explicit Knowledge group) was corrected to ($\alpha = 0.05/2$).

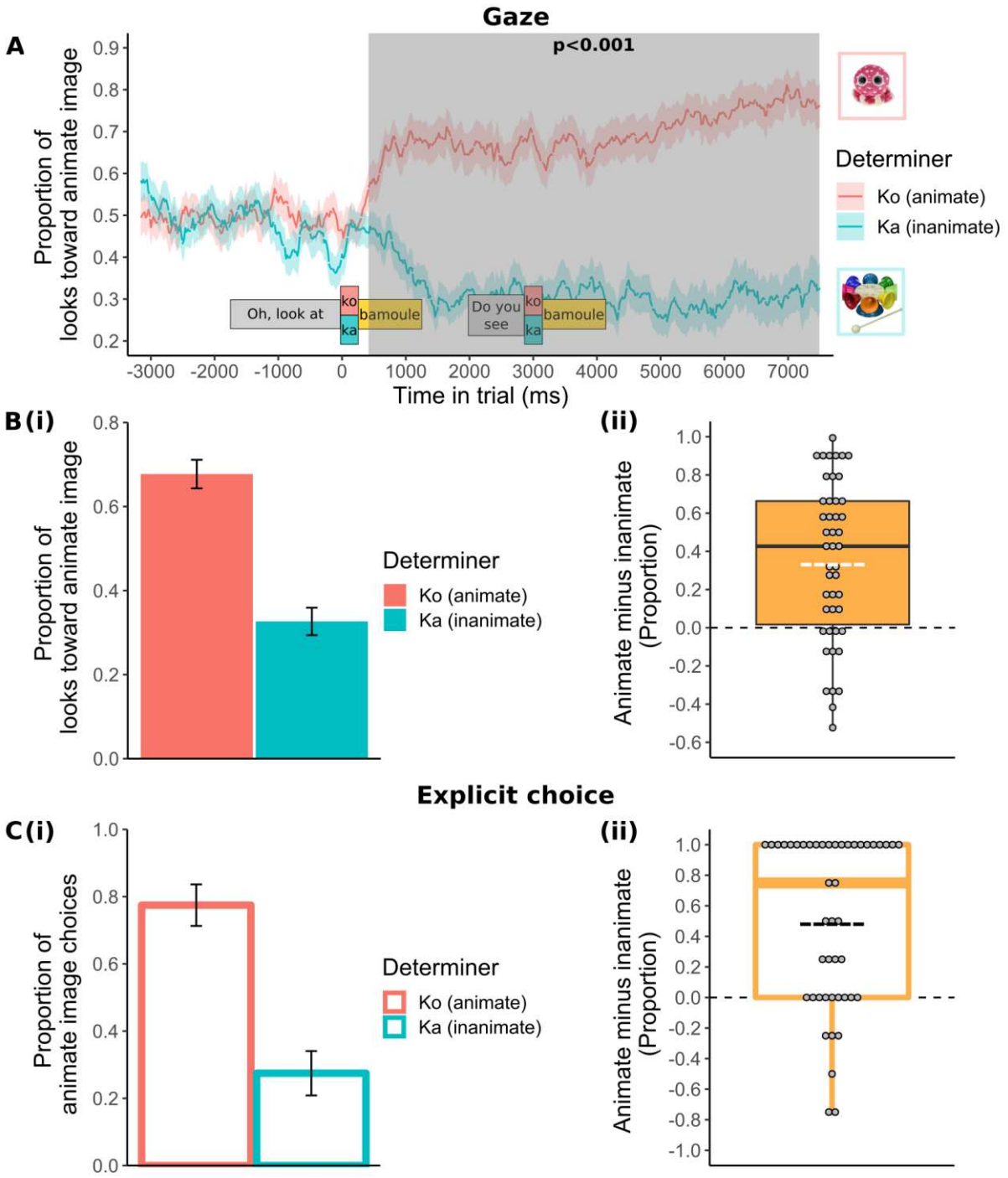


Fig. 10. Aggregated Results: Adults. Data from current study and supplemental studies from Barbir et al: adults $n = 46$. **(A) Time-course.** Adults looked more to the animate image when they heard the animate determiner *ko* than the inanimate determiner *ka* during the time-window: 540-5850ms (the moment participants were prompted to point). **(B(i)) Overall looking preference.** Adults looked significantly longer to the animate image when they heard the animate determiner than the inanimate determiner ($p < 0.001$). **(B(ii)) Overall looking preference per participant. (C(i)) Explicit choice.** Adults choose the animate image more often when they heard the animate determiner than the inanimate determiner ($p < 0.001$). **(C(ii)) Explicit choice per participant.**

A cluster-based permutation analysis on the looking-while-listening data revealed that adults looked more to the animate image when they heard the animate determiner *ko* versus when they heard the inanimate determiner *ka* during the time-window from 540ms to the end of the trial (7500ms, Fig. 10A). A mixed-effects regression on the gaze focus data confirmed

that adults looked more overall to the animate image when they heard the animate determiner *ko* than when they had heard the inanimate determiner *ka* [$\beta = 0.33$, SE = 0.05, t = 6.27, $p < 0.001$, Cohen's d = 1.13, means in Fig. 10B(i)]. A generalized linear mixed-effects analysis on the explicit choice data also showed that adults chose the animate image more often when they heard the animate determiner *ko* than when they heard the inanimate determiner *ka* [$\beta = 4.52$, SE = 1.02, model comparison: $\chi^2(1) = 30.74$, $p < 0.001$, Cohen's d = 1.15, means displayed in Fig. 10C(i)]. The aggregated data, thus, provide convergent evidence that adults generalize the determiner rule to a novel context.

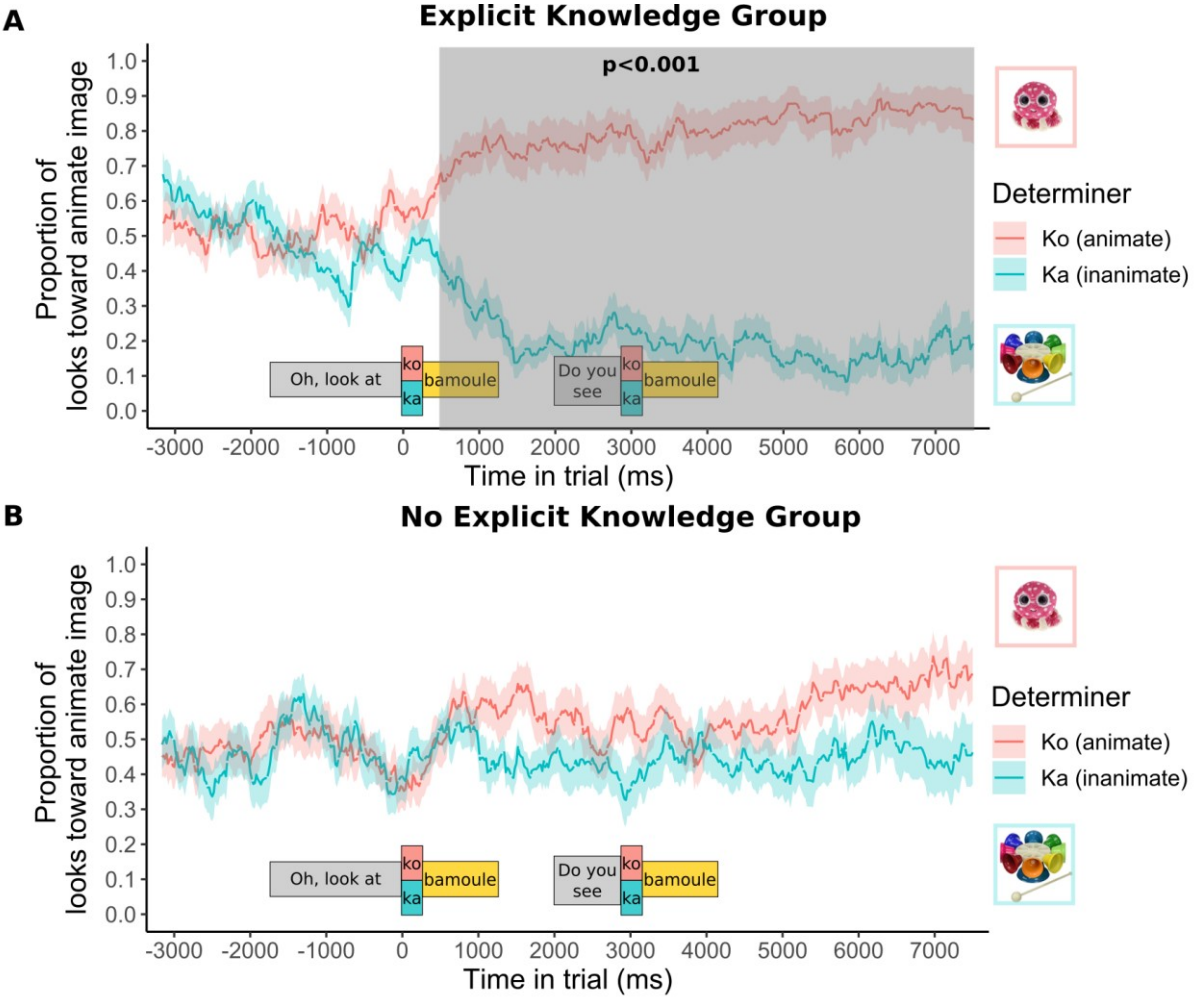


Fig. 10. Aggregated Data by Knowledge Kind (Adults): Time-Course. Explicit Knowledge Group (n = 24), No Explicit Knowledge Group (n = 22). **(A) Explicit Knowledge Group.** The Explicit Knowledge group looked more to the animate image when they heard the animate determiner *ko* than the inanimate determiner *ka* during the time-window: 360-5850ms. **(B) No Explicit Knowledge Group.** The No Explicit Knowledge group did not look significantly more to the animate image when they heard the animate determiner *ko* than the inanimate determiner *ka* at any time-point during the trial.

Two additional cluster-based permutation analyses on the looking-while-listening data showed that the Explicit Knowledge group looked longer to the animate image when they heard the animate determiner *ko* during the time-window from 360ms to the end of the trial (7500ms, Fig. 11A), but that the No Explicit Knowledge group did not look significantly longer to the animate image when they heard the animate determiner during any time-window (Fig. 11B). A mixed-effects regression showed that the Explicit Knowledge group looked more overall,

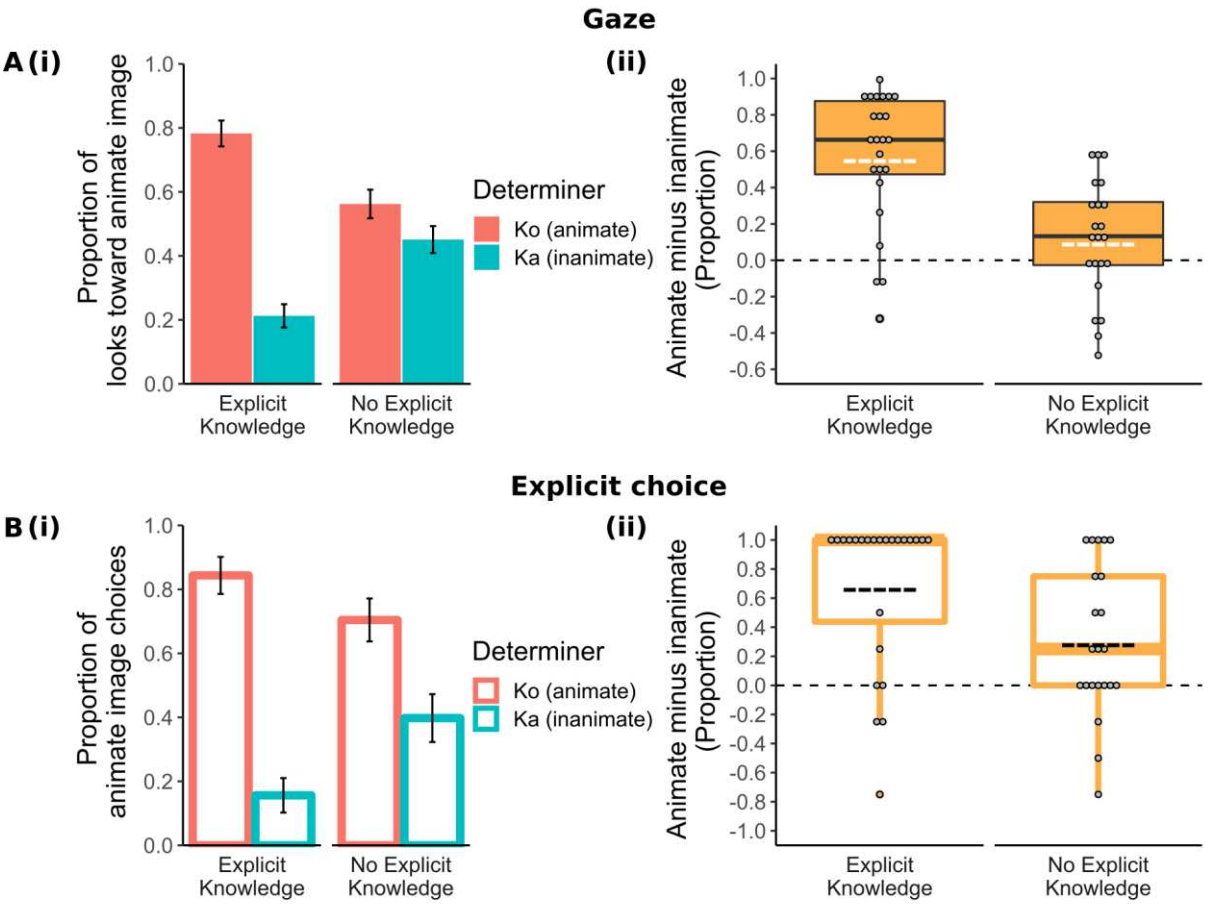


Fig. 11. Aggregated Data by Knowledge Kind (Adults): Gaze and Explicit Choice. (A(i)) Overall looking preference per knowledge group. The Explicit Knowledge group was significantly better at distinguishing the two determiners than the No Explicit Knowledge group ($p < 0.001$). The Explicit Knowledge group looked significantly longer to the animate image when they heard the animate determiner ($p < 0.001$), but the No Explicit Knowledge group did not ($p = 0.14$). (**A(ii)**) **Overall looking preference per participant and per knowledge group. (B(i)) Explicit choice.** The Explicit Knowledge group was significantly better at distinguishing the two determiners than the No Explicit Knowledge group ($p = 0.002$). The Explicit Knowledge group chose the animate image significantly more often when they heard the animate determiner ($p < 0.001$), and so did the No Explicit Knowledge group ($p < 0.001$). (**B(ii)**) **Explicit choice per participant and per knowledge group.**

throughout the trial, to the animate image when they heard the animate determiner *ko* than when they heard the inanimate determiner *ka*, in comparison to the No Explicit Knowledge group [interaction between Knowledge and Determiner: $\beta = 0.46$, SE = 0.1, $t = 4.48$, $p < 0.001$, Fig. 12A(i)]. Two separate mixed-effects regression analyses revealed that the Explicit Knowledge group looked significantly longer overall to the animate image when they had heard the animate determiner *ko* than when they had heard the inanimate determiner *ka* [$\beta = 0.57$, SE = 0.07, $t = 7.64$, $p < 0.001$, Cohen's $d = -2.3$, means in Fig. 10C(i)]; but that this was not the case for the No Explicit Knowledge group [$\beta = 0.11$, SE = 0.07, $t = 1.53$, $p = 0.14$, Cohen's $d = 0.34$, means in Fig. 12B(i)]. A generalized linear mixed-effects analysis on the explicit choice data revealed that the Explicit Knowledge group chose the animate image more often when they heard the animate determiner *ko* than the inanimate determiner *ka*, in comparison to their peers in the No Explicit Knowledge group [interaction between Knowledge and Determiner, $\beta = 4.57$, SE = 1.46, model comparison: $\chi^2(1) = 9.42$, $p = 0.002$, Fig.10D(i)]. Two distinct generalized linear mixed-effects models on each knowledge group revealed that the Explicit Knowledge group chose the animate image more often when they heard the animate determiner *ko* than the

inanimate determiner *ka* [$\beta = 0.15$, SE = 4.31, model comparison: $\chi^2(1) = 26.14$, $p < 0.001$, Cohen's d = 1.88, means in Fig. 10D(i)] and that the same was true for the No Explicit Knowledge group [main effect of determiner: $\beta = 1.31$, SE = 0.35, model comparison: $\chi^2(1) = 15.59$, $p < 0.001$, Cohen's d = -0.61, means in Fig. 10D(i)].

The aggregated data thus provided further evidence that explicit knowledge of the rule governing determiner use boosts generalization, across metrics. This is not, in itself, surprising: when one uses a rule explicitly, one can apply it consistently across-the-board. Importantly, however, the explicit choice data provided evidence of generalization even in the absence of being able to state or apply an explicit rule. The post-experiment questionnaire revealed that, on average, adults who had not discovered the rule chose a corresponding image based on what sounded or felt right.

D. Summary Experiments 1-3

Together, these results indicate that the same evidence set does not elicit generalization across development. Infants and adults both appear to generalize, but not pre-school children. Infants appear to generalize early, but do not have a strong trial-wide preference for one of the images. Adults on the other hand are very strong generalizers, showing consistent signs of generalization across metrics. In stark contrast, pre-school children do not show signs of generalization with any metric used in this study. It is entirely likely however that pre-school children would indeed generalize given enough time. They have been shown to generalize grammatical rules in their native language with ease (e.g., Pozzan et al, 2016; Fisher, Gertner, Scott, & Yuan, 2010). Importantly, if we had only used preferential looking (aggregated over the whole duration of the test trials) to evaluate generalization, we would have only been able to conclude that adults generalize. These results illustrate the importance of using different behavioural indexes of cognitive processing.

Moreover, these results demonstrate that generalization of a new structure can occur in the absence of explicit knowledge. Even without explicit knowledge about the rule governing determiner use, adults generalize when asked to make a choice.

3.1.2 Discussion

Our experiment demonstrates a non-linear pattern of generalization across development: infants and adults generalize a novel grammatical rule, but not pre-school children. If we zoom in on just the results from infants and pre-school children or just the results from pre-school children and adults, the same patterns emerge as those observed in the literature. The youngest will generalize a broader set of structures (here, a structure not present in French) than their older peers, and adults will generalise sparse categories (here, animates and inanimates) but not pre-school children. If we had just tested infants and pre-school children, we could easily have said that pre-school children are not generalizing because prior knowledge is constraining their hypothesis space; and if we had just tested pre-school children and adults, we could easily have said that diffuse attention is impeding generalization. However, when we zoom back out, the observed pattern no longer seems consistent with explanations about prior knowledge and cognitive development. Adults' prior knowledge is not impeding generalization nor is infants' diffuse attention. Though it is clear that what the mind knows (prior knowledge) and how the mind processes incoming stimuli (maturity of cognitive processing) *can* affect generalization, it is less evident *how* exactly these two factors affect generalization. Broadly, any factor can

spur or curb generalization on one of three levels: extraction of an abstract structure, weighting of the likelihood of an abstract structure to be correct, or regulation of the threshold at which an individual generalizes. Our data do not allow us to determine whether pre-school children were unable to extract the animate-inanimate structure, whether they were just unsure of the structure or whether they had a higher generalization threshold. Nevertheless, what we know about how children go about learning can point us in a highly plausible direction. Studies have shown that even if children extract the correct structure and even if they know it is the most likely, they will continue to entertain other options (e.g., Deng & Sloutsky, 2015). By extension we could surmise that the pre-school children in our experiment may actually have had a higher generalization threshold than their younger and older peers.

It may very well be the case that the generalization threshold fluctuates with development, but it could too be the case that it fluctuates on a case-by-case basis. One of the most important, and often overlooked, characteristics about generalization is that it can subserve two, not necessarily distinct, functions: action and learning. For example, when an individual encounters a snake, she will need to generalize optimally in the moment whether the specimen in front of her is a (nearly) harmless garter snake or a more dangerous viper; in contrast, an individual learning about reptiles in a biology class will have to generalize optimally at the end of the semester on her exam. The generalization threshold is thus inextricable from temporality: generalizing correctly right here right now versus generalizing correctly in the long run. A recent study showed that when individuals are in a highly complex learning environment, generalization is essential to guide learning (Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018). However, there is one caveat. Generalizing parsimoniously, though not optimal in the short term, is best for learning in the long run (Wu et al, 2018). In other words, generalizing right here and right now allows an individual to make sense of the world but generalizing later comes with a greater guarantee that an individual will generalize correctly. If this is so, then it could be the case that pre-school children, who performed poorly at test, may actually have been the most optimal learners in the long run.

When an individual generalizes early, she runs the risk of positing a structure that does not correspond to the way the world is. Adults, in particular, seem prone to generalize themselves into a corner, so to speak: when adults generalize, they have trouble re-adjusting the extracted structure in light of new but contradictory evidence (Best et al, 2013). At the same time, an individual cannot have access to (or process, for that matter) all the stimuli in the universe to determine whether the world is actually so: she needs to have a highly, but not necessarily perfectly, reliable generalization tipping point.[16] We argue here that the point at which an individual believes she has enough evidence to generalize from an extracted structure, the tipping point, may depend on how much one knows (and thus by correlation, how far along one is in development). This proposition is founded on a primordial purpose for generalizing: the urge to know. Whether one needs to know the kind of snake one has encountered to act immediately or one needs to know species of reptiles for a future exam, an individual will generalize to, very broadly, make sense of her environment. Accordingly, how quickly one needs to make sense of one's environment, this overlooked temporality of generalization, will determine the generalization threshold. There are some sets of evidence that seem ostensibly equal, and yet individuals will generalize more quickly from a set that is surprising than one that is not (Gerken et al, 2015). If an English learning infant hears '*bebo'*, she may not generalize from this one exemplar the rule: 'repeated consonant'; in contrast, if that same infant hears '*zhezho'* ('zh' as in vi**s**ion or deci**s**ion), she may generalize right away (Gerken et al,

[16] This claim does not deny the presence of innate baggage. Part of the regulation of the tipping point may be influenced by hardwired 'biases'.

2015). The difference between these two utterances is that one is standard and the other surprising (Gerken et al, 2015). In English, consonant 'b' often occurs at the onset of a word (e.g., book, baby), but 'zh' does not. While hearing 'zh' at the beginning of a word is not downright outlandish, it is unusual. The authors of this study propose that surprisal may guide generalization: the need to fit an element into an existent structure. This idea of surprisal could very well be tied more broadly the idea of a baseline urge to know. An individual who knows a substantial amount about some part of the world, let us say her native language, may have a very strong desire to make sense of the sudden presence of a new element, such as novel determiners. In other words, the new element may perturb the structure the individual has built up: either it fits into the structure as is or the structure needs to be revamped. Accordingly, in our experiment, adults may have had a particularly strong desire to make sense of the novel determiners with respect to an extant structure, and as such to generalize. Along the same lines, an individual who knows very little about some part of the world, like her native language, may have a potent desire to make sense of any part of it, including novel elements, such as the determiners *ko* and *ka*. In other words, an individual with little understanding of her world will have an urge to understand anything she can lay her hands on, to lay a foundation for knowledge. As such, in our experiment, infants may have been motivated to make sense of the novel determiners to better understand their language, and accordingly generalize. In contrast, an individual who has some but incomplete knowledge about a part of her world, like a language, may not be particularly pressed to figure out the role of a novel element, like a determiner. This individual has enough knowledge to understand and act, to comprehend and produce language, so she has the time to study the novel element and generalize parsimoniously but optimally. In our experiment, pre-school children may have felt they had time to correctly establish the use of the novel determiners; understanding the videos or more broadly the language did not hinge on this knowledge. They therefore may as well generalize when it is optimal to do so. Broadly, the urge to know may be the baseline driving factor behind the generalization tipping point.

The urge to know could also account for patterns observed in the literature. Younger infants, who have little knowledge of the world, may generalize outlandish structures in an attempt to begin to make sense of the so-called 'blooming buzzing confusion' (James, 1890) that is their external environment. Adults, who have a lot of knowledge of the world, may generalize faster in an attempt to classify novel elements into an extant internal psychological space. The urge to know could potentially be a general parameter underlying generalization. While our data provide a zoomed out picture of generalization over the lifespan, they do not allow us to conclude that the quantity of knowledge indeed was the driving factor behind the observed non-linear trajectory of generalization through development. Further research is needed to determine directly whether the amount of knowledge modulates the rate of generalization and whether this parameter can improve computational models of generalization in a developmental context.

Fluctuations of the generalization tipping point that are anchored in the quantity of knowledge, rather than the development of cognition, further explain why pre-school children are not on average reticent generalizers. Studies that evaluate generalization of structures that are present in a child's native language demonstrate a linear progression of generalization (e.g., Fisher et al, 2010; Pozzan et al, 2016). Older children generalize more than younger children, and adults more than older children (Pozzan et al, 2016). For example, while two-year-olds do not reliably distinguish the two sentences 'The rabbit *bamoules* the pig' and 'The rabbit and the pig *bamoule'*, three to four-year-olds do. Interestingly, adults distinguish the two sentences so well, that they interpret the sentence 'The rabbit and the pig *bamoule'* as a scene in which the rabbit and the pig are performing the same action. Yet, that sentence is also consistent with

scenes in which the rabbit and pig are performing two distinct actions in the context of a broader common goal (e.g., playing, cleaning, working). These kinds of verbs may however be less likely candidates, in the wider scope of the language (e.g., there may be fewer distinct-action verbs) or for competent speakers (e.g., competent speakers are likely to already know general verbs like 'play'). Though potentially efficient at interpreting a known structure, adults' staunch generalization may be a bane when it comes to learning irregular structures or structures that differ greatly from known structures. It could be one of the factors that make learning a second language later in life ostensibly so difficult. Even though some adults attain native-like proficiency in second language grammar, how well an adult acquires second language grammar will depend on a slew of factors, such as verbal reasoning skills (DeKeyser, 2000) and her native language grammar (e.g. Paradis, 2018). Nevertheless, in this study, adults did correctly generalize the rule governing determiner use, even if they were not aware of said rule. One prospective, although seemingly counterintuitive, use for this paradigm could actually be to boost second language learning. Lab experiments have found that a learner who has acquired one structure will generalize knowledge to a related structure (Gomez, Gerken, & Schvaneveldt, 2000), and examinations of language acquisition in the wild have found that learners of a second language who have a like native language acquire grammar faster and to a higher degree (Paradis, 2018; Paradis, Tulpar, & Arppe, 2016). Thus, introducing an odd animate-inanimate determiner into French, for example, may aid an individual to acquire the animate-inanimate distinction on the verb 'to be' in Japanese (e.g., *Neko ga* ***iru*** There **is** a cat, versus *Isu ga* ***aru*** There **is** a chair). Further investigation is necessary to determine whether this protocol could take advantage of adults' native language knowledge to boost second language acquisition. More generally, research is needed to determine the differences (or similaries) between generalization of 'hypothesized' (e.g., *ko* seems to precede animals) and 'indubitable' (e.g., *ko* precedes animals) structures across development.

An interesting result, hidden amongst the null results, in the pre-school children's data was that pre-school children had a strong bias for animates. While this bias may reflect just that, a bias, it could also reflect a strategic tactic. If children have a tendency to explore uncertain options (Deng & Sloutsky, 2015; Schultz et al, 2019), one possible option would be that the inanimate determiner *ka* is simply a determiner for everything, including animates. Thus, if children had a hunch that the animate determiner *ko* was for animates and an uncertainty about *ka* being for everything, we would observe the same pattern, namely an animate bias. Our results do not allow us to disentangle these two causes of the observed effect, but we hope that other, perhaps novel, experimental protocols and measures will be able to isolate the driving factor (or factors).

These data, lastly, point the importance of understanding indices that we use as metrics of cognition. An individual who generalizes is certainly entertaining a structure, but what exactly it is 'to entertain' is harder to pinpoint. At the other end of the spectrum, an individual who is not generalizing could very well be entertaining a structure (e.g., maybe *ko* precedes animals) or not at all; at first glance, it seems to tell us little about what is going on in the mind. Ostensibly counterintuitively however an individual who does not generalize actually reveals a lot about what *could* be going on in the mind. Herein lies the force of null results: they ask of us to think broadly about what could be going on in the mind, rather than focusing narrowly on what is going on. To draw a parallel, null results prevent us from generalizing right on the spot, but they may just be what gets us to generalize correctly in the long run.

Here, we demonstrate that one set of evidence can incite generalization early in development and at maturity, but not in-between: infants and adults readily generalize, but pre-

school children do not. This ostensibly odd pattern raises important questions about generalization itself. While some knowledge, however rudimentary, is necessary for generalization, it may not be sufficient. We thus advance the hypothesis that perhaps generalization is more than a simple marker of knowledge: it may reflect a primordial urge to know.

3.1.3 Materials and Methods

This is a pre-registered study (https://osf.io/s3z59/). All materials, methods and analyses are as pre-registered, unless otherwise stated.

Materials.

Novel determiners. The novel determiners, *ko* (/ko/) and *ka* (/ka/), were the same as those in Barbir et al. They had been created such that they were phonotactically possible in French and that they resembled in form to French singular determiners, which are monosyllabic and have roughly similar phonological forms for masculine and feminine variants. In French, determiners are marked for grammatical gender in the singular form.
Definite masculine: *le* /lə/
Definite feminine: *la* /la/
Indefinite masculine: *un* /ɛ̃/
Indefinite feminine: *une* /yn/)

Training videos. The training video from Barbir et al was adapted to minimize any task-related noise:

1. Total exposure to the novel determiners was augmented (i.e., number of times a participant will hear *ko* and *ka*). Increasing exposure time has been shown to increase the probability that both children and adults will generalize correctly (e.g., Gebhart, Newport, & Aslin, 2009).
2. The variability of exposure was augmented (i.e., the number of different nouns with which *ko* and *ka* are paired). Increasing types, rather than tokens, has been shown to promote generalization (e.g., Gerken & Bollt, 2008; Xu & Tenenbaum, 2007).
3. Simple referential scenes were added (i.e., static images are presented with a neutral sentence: 'Oh, look at *ka* ball', Fig. 2C). When a novel grammatical distinction is based on meaning, studies often present isolated image-utterance pairs rather than stories (e.g., Culbertson, Smith, Jarvinen, & Haggarty, 2018).
4. Images of real animals and objects were added (i.e., a real butterfly, rather than a stuffed animal, Simple Scenes: Fig. 2C). Though infants can categorize animal toys from object toys and even blobs with eyes from plain blobs (McDonough & Mandler, 1998, and Jones, Smith, & Landau, 1991, respectively), the distinction between real animals and objects may be more salient than for toy animals and objects.
5. Novel (pseudo-) nouns and referents were added (Fig. 2B). Though infants are actively learning words, pre-school children and, to a greater extent, adults learn words less frequently. Novel noun-referent pairs (e.g., *maligour* + a yeti-like monster) put older participants in, so to speak, 'word learning mode', mimicking roughly the task they would have to do during the test phase.
6. Participants were familiarized with the test phase nouns (without indicating the referent, see below). Though all novel test items (and novel nouns) have successfully been used in word learning tasks with children (e.g., Dautriche, Fibla, Fiévet, & Christophe, 2018),

familiarization with the novel nouns allowed us to guarantee that participants were not confused because they labeled the items using existent French words (e.g., that thing looks a bit like a rabbit, it ought to be called 'rabbit').

There were a total of 4 training videos. During the videos, a woman acted out stories with stuffed animals and toy objects, using child-directed speech (Fig. 1A). The stories were entirely in French, except for the novel determiners and a handful of novel nouns. The training videos were composed of a mix of stories (A-D below). To provide participants with repetition and variation, some stories repeated across videos, once in each video for a total of 4 repetitions (same French stories and Familiarization stories), but other stories differed (different Word-leaning stories and Simple scenes). There were 4 kinds of stories (Fig. S2):

A. French stories (Fig. S2A): The woman acts out stories, in which only the novel determiners are not in French. These stories were designed to provide participants with an ecologically valid context from which to glean the characteristics of the novel determiners. These were the same scenes used in the training video from Barbir et al.
B. Word-learning stories (Fig. S2B): The woman acts out stories with novel items. In these stories, everything is in French except novel nouns (which refer to the novel items) and novel determiners. These stories were designed to make the task as similar as possible for the three age groups. Infants were likely to not know some French nouns or grammatical elements in the videos, and would thus be actively listening/watching and using any knowledge they have to understand the videos. Pre-school children and adults, however, were likely to have a solid understanding of the French words used in the videos: to put them, somewhat, in infants' shoes, we presented them with some novel nouns. Accordingly, all three age groups would have been (though to different degrees) in, what we call, 'word learning mode'.
C. Simple scenes (Fig. S2C): Participants saw a static image on the screen and the woman telling the stories would name the item, using neutral sentences containing the novel determiner (e.g., Oh, look at *ka* ball! Do you see *ka* ball?). These scenes were designed to present participants with just the essential elements, without extraneous linguistic or visual distractions: an item (animal or object), a French noun, and a novel determiner.
D. Familiarization stories (Fig. S2D): The woman acts out stories with the novel test items and nouns. The first part of the stories was designed to familiarize the participants with the items: the toys are described, but not named (e.g., 'Look! She is pink all over!'). It was the same as in the training video from Barbir et al. A second part was created specifically so that older participants (pre-school children and adults) would not spuriously associate a real French word with one of the novel items (e.g., that thing looks a bit like a rabbit, it must be called 'rabbit'). The second part of the stories was thus designed to familiarise the participants with the nouns (e.g., Here's *ko bamoule, ko bradole, ka pirdale* and *ka doripe,* Fig. 2D*),* without indicating which item is the referent of which noun.

The novel determiners are each presented 141 times during the training videos (*ko* x 141, *ka* x 141, across the 4 videos). They are paired with 11 distinct French animal nouns (*lapin* rabbit, *poule* chicken, *cochon* pig, *chien* dog, *chat* cat, *souris* mouse, *canard* duck, *grenouille* frog, *papillon* butterfly, *baleine* whale, *babouin* baboon) or object nouns (*livre* book, *tracteur* tractor, *biberon* bottle, *poussette* stroller, *voiture* car, *chaussure* shoe, *ballon* ball, *gateau* cake, *train* train, *bateau* boat, *banane* banana). The nouns were chosen based on previous CDI reports, such that 20-month-old French-learning infants would be likely to know them. The only exceptions were two animal nouns, whale and baboon, which were specifically chosen for an additional task for pre-school children and adults (Anticipation task, see below). The nouns

were chosen such that half of the animal nouns were masculine and half feminine, and the same for object nouns. As such, the novel determiners could not also be marking grammatical gender. The determiners were also paired with 2 distinct novel animal nouns (*maligour* /maliguʁ/ and *nuve* /nyv/) or novel object nouns (*jigoulate* /ʒigulat/ and *fomme* /fɔm/). Novel nouns were created such that they were phonotactically possible in French. Each novel noun was paired with an original animal or object (a blue yeti-like monster, axolotl-like creature, a water-based thermometer, and a fancy spray-bottle). The novel nouns were assigned a grammatical gender, such that one animal noun was feminine and the other masculine, and likewise for object nouns. All nouns began with a consonant so as to avoid cliticization that occurs when nouns begin with a vowel (*le+avion = l'avion*). The novel determiners functioned in the same way structurally as French determiners. For example, a set of highly frequent adjectives in French appear most often between the determiner and noun (e.g., *le joli chat,* the cute cat). This was thus replicated with the novel determiners: From time to time, a highly frequent adjective was added to the determiner-noun pairing (e.g., ***ka*** *grande chaussure*). Adjective use served to facilitate segmentation of the determiner-noun sequence, such that it would not be perceived as one new noun (e.g., *kachaussure*) or a proper noun. To aid categorization, story scenes were constructed to involve interaction between one animate and one inanimate (highlighting dissimilarity) or between two animates/two inanimates (highlighting similarity).

Test nouns and items. The test nouns and items were the same as in Barbir et al. During the test phase, novel determiners are presented with one of four novel nouns (*bamoule* /bamul/, *pirdale* /piʁdal/, *doripe* /doʁip/, *bradole* /bʁadɔl/). Novel nouns were created such that they are phonotactically possible in French. Each novel noun was paired with one novel original item, an animal or an object. Novel animals were a pink stuffed animal with a big head and many short feet and a mouse-like animal with rabbit ears and an anteater's trunk; while novel items were a round colourful xylophone-like musical toy and a standing top. These novel items and novel nouns have been used in previous studies investigating vocabulary acquisition, and 20-month-olds have been successful at learning the item-noun pairings (e.g., Dautriche, Fibla, Fievet, & Christophe, 2018; Barbir et al). To control for item effects, 4 combinations of novel noun and item pairings were created using a Latin square design. A pairing was randomly assigned to each participant. Each novel noun appeared thus with *ko* for roughly half the participants, and with *ka* for the other half. Accordingly, there were two versions of the videos, with opposite animacy attribution (in one version of the videos *pirdale* and *doripe* were animate nouns and *bamoule* and *bradole* were inanimate nouns, and the other version of the videos the opposite was true).

Test phase. The test phase was the same as in Barbir et al. To familiarise participants with the testing procedure, just prior to the test phase, participants see two training trials, with words and toys seen during the videos (e.g., The participant sees an image of the rabbit and tractor from the video and hears '*Oh regarde* ***ko*** *lapin !'* Oh look at ***ko*** rabbit!). The test phase is composed of two kinds of trials: test trials with the novel words and filler trials with French words. Each of the novel words was tested twice for a total of 8 test trials. 8 filler trials were interspersed during the test phase so that a participant would not see more than 2 test trials in a row. There were two kinds of filler trials, 4 of each kind: *Seen* filler trials consisted of nouns presented in the training videos and an image of a different exemplar of that noun, not seen in the videos (*souris* mouse, *cochon* pig, *biberon* bottle, *chaussure* shoe); *known* filler trials consisted of nouns that were not presented during the training video, but that were likely to be known by 20-month-old infants based on previous CDI reports (*poisson* fish, *cheval* horse, *vélo* bike, *chapeau* hat).

Anticipation trials. After the test phase, adults and pre-school children had 8 additional trials. They were not included for infants because of time constraints (the in-lab experiment was quite long as is for young infants) and because some of the words were not likely to be known. Anticipation nouns all shared a common first syllable (***ba****leine* whale, ***ba****bouin* baboon, ***ba****teau* boat, ***ba****nane* banana). There were 8 anticipation trials: half determiner informative and half determiner uninformative (Fig. S1). Informative trials were composed of one animate noun and one inanimate noun, while uninformative trials were composed of two animate nouns or two inanimate nouns. During informative trials, the determiner provides information as to which of the two images is the target. For example, during informative trials, a participant may see one image of a whale and another of a boat; the moment she hears 'Oh, look at *ko* ba…*',* she can direct her gaze to the animate image, namely the whale (if she has learned the association between *ko* and animates, or *ko* and 'whale'), because the boat is inanimate and thus preceded by *ka* (Fig. S1A). In contrast, during uninformative trials, a participant may see one image of a whale and another of a baboon; the moment she hears 'Ok, look at *ko* ba…*'*, she will not be able to know which of the two images is the target, because both are animate and thus preceded by *ko* (Fig. S1B). She will have to wait until she hears the noun itself. Therefore, participants ought to direct their gaze to the target earlier for informative than uninformative trials. Anticipation of an upcoming noun has been observed when determiners are marked for grammatical gender, in adults and even children as young as 28-months (Dahan, Swingley, Tanenhaus, & Magnuson, 2000; Van Heugten, Dahan, Johnson, & Christophe, 2012; Lew-Williams & Fernald, 2007). This extra task was included to investigate whether participants were using the novel determiners in online comprehension. Informative and uninformative trials were interspersed such that there would not be more than two informative or uninformative trials in a row.

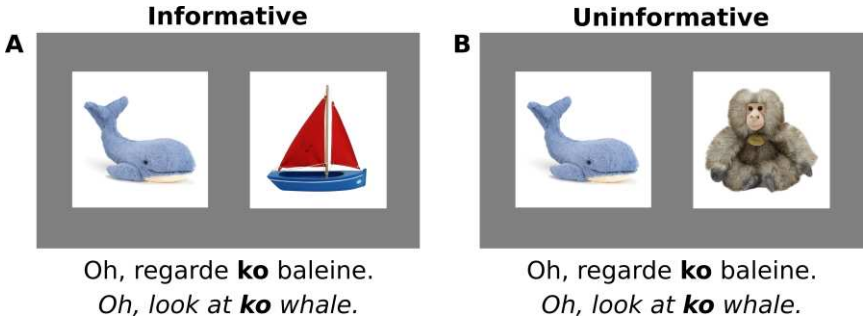


Fig. S1. Anticipation task. Both images were of items whose label began with the same first syllable ('ba'). **(A) Informative Trials.** Informative trials had one animate and one inanimate image. **(B) Uninformative Trials.** Uninformative trials had two animate images or two inanimate images.

Procedure Overview. Before coming to the lab, participants watched the training videos at home, one a day, for three consecutive days. For example, if the test session was on a Saturday, participants would watch the video at home on Wednesday, Thursday, and Friday.

At the lab, participants were tested individually in a sound-proof booth. The participants' gaze was recorded with an Eyelink 1000 eye-tracker at a frequency of 500 Hz. A five-point infant-friendly calibration was used (the code will be available on the study's OSF page at time of publication, currently embargoed). After the calibration, the experiment began.

Participants first viewed the training video (for the fourth time). The test phase followed, in which participants were presented with two images on the screen, one on the left and one on the right (each about 30cm x 30cm). Each trial had one animate and one inanimate image. Presentation side and animacy of images, as well as the presentation side and animacy of the

target were counterbalanced. Images were presented in silence for 2s, then a prompting sentence began to play, asking the participant to look at one of the two images. 1s after the prompting sentence, the sentence was repeated. For infants, 4s after the end of the second repetition, the trial ended. For pre-school children and adults, 2s after the end of the second repetition an orange frame appeared around each of the two images: it was the prompting signal for participants to point to the image that corresponded to the sentence. Older age groups were additionally asked to point, as a secondary measure of performance (for a case in which gaze results were not significant, but explicit choice was, see Babineau et al, submitted). Trial duration was chosen to match approximately across groups that pointed (2 seconds after the end of the second sentence plus the time to point, pre-school children and adults) and the group that did not (4 seconds after the end of the second sentence, infants). The test phase began with two training trials and then continued to a mix of test and filler trials. In the middle of the test phase, there was a short interlude video (~30s), during which the woman played with toys but did not name them or use the novel determiners. The interlude video served as a motivational break for participants. For pre-school children and adults, the anticipation phase followed the test phase.

The in-lab experiment lasted approximately 15 minutes.

Analyses Overview.

We used two eye-tracking measures to assess performance: looking-while-listening and preferential looking (Fernald et al, 2008). These two measures allow for a complementary and comprehensive analysis, with both specific information as to when precisely participants orient their gaze to the correct image (looking-while-listening), as well as a way to capture general gaze patterns in light of individual variation (preferential looking). For example, some participants might look right away to the target image, others may take more time and only do so toward the end of the trial. All analyses on gaze data were performed on the subset of trials that had enough data (the trials during which participants fixated the screen for at least 50% of the total time, from the moment the two images were presented on the screen to the end of the trial, 4s after the second repetition of the prompting sentence). Analyses were run on the subset of time in a trial from the moment a decision was possible (the moment when the determiner, *ko* or *ka*, was presented) to 2s after the second repetition of the auditory sentence (i.e., the moment when the prompt to point appeared for pre-school children and adults), a span of 5850ms. This time-window was pre-registered.

We used an additional behavioural measure to assess performance of pre-school children and adults: explicit choice, as indexed by pointing. Pre-school children and adults, but not infants, were prompted to point to the image that corresponded to the sentence they just heard, at the end of each trial.

Looking-while-listening.

We examined the proportion of looks to the animate image when participants heard the animate determiner *ko* followed by a novel noun (e.g., *ko bamoule*) and when they heard the inanimate determiner *ka* followed by a novel noun (e.g., *ko pirdale*) on a moment-to-moment basis during the trial. We used a cluster-based permutation (Maris & Oostenveld, 2007) via the eyetrackingR package in R (Dink & Ferguson, 2016). Gaze data sampled at 500 Hz, was down-sampled to 50Hz for analyses by binning adjacent time-points. The analysis ran a t-test at each time-point comparing the arcsine-transformed proportion of looks to the animate image when the determiner was *ko* versus when it was *ka*.[17] It grouped the adjacent time-points with a t-value

[17] We had pre-registered a cluster-based permutation analysis that ran a mixed-effects model for the anticipation trials. For the sake of consistency with the other cluster-based permutation analyses and those performed in Barbir et al, we ran a cluster-based permutation analysis based on t-tests for the anticipation results too.

greater than the predefined threshold of 1.5 into a cluster. 1000 permutations were run. A cluster was declared significant whenever clusters of the same size or larger were found in less than 5% of the permutations (the size of the time-clusters was computed as the sum of the adjacent t-values within the cluster). Significant 'clusters' indicate that participants were looking significantly longer to the animate image in one condition than the other, and the moment in the trial when this is so.

Preferential looking.
We also examined the proportion of looks to the animate image when participants heard the animate determiner *ko* and when they heard the inanimate determiner *ka,* averaged together over the whole trial. We used a linear mixed-effects model (via the *lmer* function in R) with the proportion of looks to the animate image during each trial as the dependent variable, determiner (*ko* vs. *ka*) as the independent variable, and participant as a random effect.

Explicit choice.
Choice accuracy on test trials, as indexed by pointing, was evaluated using a generalized linear mixed-effects model, with proportion of animate image choices was the dependent variable, determiner (*ko* vs. *ka*) was the independent variable, and participant was a random effect. The model was computed using the *glmer* function in R.[18]

Aggregated data analyses.
We run the same analyses on aggregated data from multiple like experiments. As we had not pre-registered this analysis, we preferred to be as conservative as possible and not reduce the time-window for gaze analyses: gaze data was thus analysed from to the onset of the determiner to the end of trial (0-7500ms). The alpha was Bonferroni-corrected for multiple comparisons ($\alpha = 0.05$/the number of analyses).

We use an alpha of $\alpha = 0.05$, unless otherwise stated.

Experiment 1: Infants.

Participants.
Infants were recruited from the lab database. All were monolingual French-learning infants, who heard less than 10% of another language. A total of 27 infants were included in the analyses (mean: 20.00 months; range: 19.20-20.09 months; 12 girls, 15 boys). Sample size ($n = 24$) was chosen based on previous studies investigating grammar learning with a like paradigm (e.g., Barbir et al). Recruiting however is done in batches: any 'extra' participants are not excluded (as pre-registered). Infants were excluded for excessive fussiness ($n = 4$), crying ($n = 1$), being ill ($n = 2$), moving the eye-tracker during the experiment ($n = 2$), not having watched the training video at home for 3 days consecutively ($n = 3$), having a vision related problem ($n = 1$), not having enough eye-tracking data (at least 50% of gaze data-points for 2 trials of each determiner, $n = 2$), and technical error ($n = 1$). Not enough eye-tracking data ($n = 2$) was the result of: hair falling in front of the target sticker necessary for eye-tracking ($n = 1$) or the infant putting her hand in front of her eyes ($n = 1$). Written informed consent was obtained from each child's

[18] We had pre-registered an analysis of correct target choices when participants heard the animate or inanimate determiners for the explicit choice data, and an analysis of the proportion of looks to the animate image for the gaze data. In this study, we were interested in the distinction between the two determiners, and thus a difference in animate image choices rather than the total correct target choices. We thus ran an analysis on the proportion of animate choices on the explicit choice data, rendering the explicit choice analyses and gaze analyses descriptively comparable. In the final paper, for transparency, both analyses will be presented.

parents prior to the experiment. All research was approved by the local ethical board: CER Paris Descartes.

Procedure.
Parents received instructions to show the videos to their child, but to stay as quiet and neutral as possible. In addition, parents were told not to refer to the content of the videos between viewings (for instructions, see Text. S1).

Before coming to the lab, parents were asked to fill out an online questionnaire related to the videos (26/27 parents filled out the questionnaire; for questionnaire, see Annex S2). 25/26 parents reported that their child was attentive for the whole duration of each video. Parents reported that over the three days the child's interest in the video increased (n = 8), stayed the same (n = 12), decreased (n = 5), or they did not respond to the question (n = 1). Children knew on average 18.77/22 of the nouns (min: 7; max: 22) used in the videos. All children knew: dog, cat, book, bottle, shoe. More than 20/26 children knew: rabbit, chicken, pig, duck, frog, ball, banana, boat, train, cake, stroller, car. Between 15 and 20/26 children knew: butterfly, mouse, whale, tractor. 6/26 children knew: baboon.

Parents also filled out the short French version of the MacArthur Communicative Development Inventory for 18-month-olds (Kern, Langue, Zesiger, & Bovet, 2010). Short French CDIs are available for 12-, 18-, and 24-month-olds: we chose the one that was the closest match in age, 18 months. We asked parents to fill out the short CDI rather than the long version (~600 words; Kern, 2007), to lighten their workload: they already allocated time to watching videos at home for three consecutive days and to fill out an online questionnaire. Vocabulary ranged from 18/97 to 97/97 words (mean = 79.4/97). There was no significant correlation between mean looking time to target and vocabulary size ($r(25) = 0.27$, $p = 0.18$).

At the lab, infants were seated on their parent's lap at a distance of 65cm from a 42' screen. Parents wore headphones and listened to a neutral music-mix during the experiment. They could not hear the stimuli presented to the infant.

The in-lab experiment lasted approximately 12 minutes.

Aggregated infant data analyses (current experiment and Barbir et al).
We aggregated data from two like infant experiments (the current experiment and the experiment from Barbir et al). The only difference between the two experiments was the content of the videos. The current experiment had more exposure to the novel determiners, both in quantity (number of uses) and variety (number of different associated nouns, e.g., *ko* rabbit, *ko* fish, *ko* dog). Infants were tested in the same conditions. A total of 51 infants were included in the analyses (mean: 19.29 months; range: 19.15-20.11 months; 18 girls, 23 boys). A total of 336 trials could thus be analysed.

Experiment 2: Pre-school children.

Participants.
All children were recruited from the lab database. 29 monolingual French children participated in the study. The criteria for being monolingual was hearing 90% French or more on a daily basis, from birth. Four children were excluded from the final sample for the following reasons: failing to correctly respond to the two training trials (n = 1), not having watched the video for three consecutive days (n = 1), having a hearing related problem (n = 1), and not having enough eye-tracking data (at least 50% of data for 2 trials of each determiner, n = 1). The remaining 25 children were included in the analyses (mean: 4;8 years; range: 4;4-4;11 years; 13 girls, 12 boys). Written informed consent was obtained from each child's parents prior to the experiment. All research was approved by the local ethics board: CER Paris Descartes.

Procedure.
Parents received instructions to show the videos to their child, but to stay as quiet and neutral as possible. In addition, parents were told not to refer to the content of the videos between viewings (for full instructions see, Annex S1).

Before coming to the lab, parents were asked to fill out an online questionnaire related to the videos (24/25 parents filled out the questionnaire; for questionnaire, see Annex S3). All parents reported that their child was attentive for the whole duration of each video. Parents reported that over the three days the child's interest in the video increased (n = 2), stayed the same (n = 16), or decreased (n = 6). 3/24 children asked a question about *ko*/*ka* (parents were instructed not to discuss the content of the videos with their child).

At the lab, children were tested individually in a sound-proof booth. They were seated at a distance of 65cm from a 27' screen, and wore headphones. The experimenter was seated next to the child, and provided encouragement and gave instructions. The experimenter could see the screen, but could not hear the audio. The experimenter told the child to listen carefully to stories, because there were clues that would help her during the game afterward (i.e., test phase).

Aggregated pre-school children data analyses (current experiment, supplemental experiment 1, and two supplemental experiments from Barbir et al).
We aggregated all the data from 4 experiments with pre-school children with the same experimental condition (*ko* and *ka*) and like stimuli and protocol (training videos): the main experiment, the supplemental experiment, and two supplemental experiments from Barbir et al. The main difference between the current experiment and the other three experiments was the quantity of exposure: whereas the current experiment was a four-day experiment (pre-school children watched videos at home for three days and then once in the lab), the other three experiments were one-day experiments (children saw the video once and proceeded directly to the test phase). Pre-school children were tested in comparable conditions (in lab or in a quiet room at their school). In total, we had gaze data from three experiments (children n = 74; total of 535 trials, mean: 4;9 years; range: 4;2-6;0 years; 36 girls, 38 boys) and explicit choice data from four experiments (children n = 100; mean: 4;6 years; range: 3;1-6;0 years; 47 girls, 55 boys). The aggregated data allowed us to investigate whether children could generalize from newly acquired determiners to novel words with more power, hopefully reducing any noise that would have occulted a weak trend.

Experiment 3: Adults.

Participants.
Adults were recruited from the lab database and near-by universities. A total of 28 French native speakers participated in the study. Two participants were excluded for failing to comply with the instructions for the experiment: not having watched the videos on three consecutive days (n = 1) and watching the videos while performing another activity (n = 1). The remaining 26 adults were included in the analysis (mean: 27;8 years; range: 19;3-39;2 years; 16 females, 9 males). Adults had learned on average 2.3 languages (min: 1, max: 8) in addition to French.[19] 6/26 adults reported having heard another language from birth: Arabic (n = 1), English (n = 1), Jahanka (n = 1), French creole (n = 1), Malagasy (n = 1), Spanish (n = 1). 24/26 reported enjoying learning languages. Written informed consent was obtained from each participant prior to the experiment. All research was approved by the local ethics board: CER Paris Descartes.

[19] Knowledge of classical languages (e.g., Latin or Ancient Greek) is included in the count.

Procedure.
Adults received instructions to watch the videos attentively but not reflect explicitly on or discuss the content of the videos between viewings (for instructions, see: Annex S4). 11/26 reported not explicitly thinking about the context of the films, 4/26 reporting thinking a little bit about the context, 10/26 reported explicitly thinking about the content, 1/26 did not respond to the question. To ensure that adults were indeed watching the videos attentively, each video included a written password displayed for 2 seconds at a randomly selected point in the video. Adults had to report the passwords upon arrival at the lab.

At the lab, adults were tested individually in a sound-proof booth. They were seated at a distance of 65cm from a 42' screen.

The in-lab experiment lasted approximately 15 minutes.

After the experiment, a questionnaire was administered regarding language-learning experience and experiment-related strategies (for questionnaire, see: Annex S5). The questionnaire was used to determine which adults could explicitly state the rule (Explicit Knowledge group) and which could not (No Explicit Knowledge group). 14/26 adults could explicitly state the rule governing the use of the grammatical element. 9/14 discovered the rule during the first video, 4/14 during the second, and 1/14 during the test phase in the lab. The questionnaire revealed, descriptively, that adults in the Explicit Knowledge group were generally more confident (on a scale of 1 'I guessed' to 10 'I'm sure') in their interpretations of the novel words (*bamoule, pirdale, doripe, bradole;* 9.64/10) than those in the No Explicit Knowledge group (6/10). The confidence reports provided an estimate of how adults felt about their answers. Statistical analyses were not performed on the confidence reports.

Aggregated adult data analyses (current experiment and supplemental experiment from Barbir et al).
We aggregated the data from the adult study in Barbir et al (adults $n = 20$) and the current study (adults $n = 26$, total: $n = 46$), to investigate first whether more power would reveal effects weakened by noise in each individual experiment, and second whether the slight methodological variations between the two experiments led to differences in performance. Both studies had the same experimental condition (*ko* and *ka*) and like stimuli and protocol (training videos). The main difference between the current experiment and the experiment from Barbir et al was the quantity of exposure: whereas the current experiment was a four-day experiment, the supplemental study from Barbir et al was a one-day experiment (adults saw the video once and proceeded directly to the test phase). They thus had less exposure, less varied exposure, and did not sleep between training exposure and test. There were a total of 24 participants who could explicit state the rule governing determiner use (Four Day: $n = 14$; One Day: $n = 10$) and 22 who could not (Four Day: $n = 12$; One Day: $n = 10$). The alpha for subgroup analyses (Explicit Knowledge/No Explicit Knowledge) was corrected to $\alpha = 0.05/2$.

References

Arunachalam, S., & Waxman, S. R. (2010). Meaning from syntax: Evidence from 2-year-olds. *Cognition*, *114*(3), 442-446.

Babineau, M., de Carvalho, A., Trueswell, J. & Christophe, A. (submitted). "Look! Ko pig wants to play with ko pretty turtle": Familiar words can serve as a semantic seed for syntactic bootstrapping.

Barbir, M., Babineau, M., Fiévet, A.-C., & Christophe, A. Infants quickly use newly learned grammar to guide acquisition of novel words. (In prep).

Best, C. A., Yim, H., & Sloutsky, V. M. (2013). The cost of selective attention in category learning: Developmental differences between adults and infants. *Journal of Experimental Child Psychology*, *116*(2), 105-119.

Culbertson, J., Smith, K., Jarvinen, H., & Haggarty, F. (2018). Do children privilege phonological cues in noun class learning?. *PsyArXiv*, d4yxr.

Cutler, A. (1993). Phonological cues to open-and closed-class words in the processing of spoken sentences. *Journal of Psycholinguistic Research*, *22*(2), 109-131.

Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. S. (2000). Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language*, *42*(4), 465-480.

Dautriche, I., Fibla, L., Fievet, A. C., & Christophe, A. (2018). Learning homophones in context: Easy cases are favored in the lexicon of natural languages. *Cognitive Psychology*, *104*, 83-105.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in second language acquisition*, *22*(4), 499-533.

Deng, W. S., & Sloutsky, V. M. (2015). The development of categorization: Effects of classification and inference training on category representation. *Developmental Psychology*, *51*(3), 392.

Dink, J. & B. Ferguson, B. (2016). _eyetrackingR_. R package version 0.1.6. Retrieved from http://www.eyetrackingR.com

Dudley, R., Orita, N., Hacquard, V., & Lidz, J. (2015). Three-year-olds' understanding of know and think. In *Experimental perspectives on presuppositions* (pp. 241-262). Springer, Cham.

Endress, A. D. (2013). Bayesian learning and the psychology of rule induction. *Cognition*, *127*(2), 159-176.

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*(2), 234-248.

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

Fisher, C., Gertner, Y., Scott, R. M., & Yuan, S. (2010). Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(2), 143-149.

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*(3), 360-371.

Gebhart, A. L., Newport, E. L., & Aslin, R. N. (2009). Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. *Psychonomic Bulletin & Review*, *16*(3), 486-490.

Gerken, L., Balcomb, F. K., & Minton, J. L. (2011). Infants avoid 'labouring in vain'by attending more to learnable than unlearnable linguistic patterns. *Developmental Science*, *14*(5), 972-979.

Gerken, L., & Bollt, A. (2008). Three exemplars allow at least some linguistic generalizations: Implications for generalization mechanisms and constraints. *Language Learning and Development*, *4*(3), 228-248.

Gerken, L., Dawson, C., Chatila, R., & Tenenbaum, J. (2015). Surprise! Infants consider possible bases of generalization for a single input example. *Developmental Science*, *18*(1), 80-89.

Gerken, L., & Knight, S. (2015). Infants generalize from just (the right) four words. *Cognition*, *143*, 187-192.

Gerken, L., & Quam, C. (2017). Infant learning is influenced by local spurious generalizations. *Developmental Science*, *20*(3), e12410.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of child language*, *32*(2), 249-268.

Gertner, Y., & Fisher, C. (2012). Predicted errors in children's early sentence comprehension. *Cognition*, *124*(1), 85-94.

Gervain, J., & Endress, A. D. (2017). Learning multiple rules simultaneously: Affixes are more salient than reduplications. *Memory & Cognition*, *45*(3), 508-527.

Gervain, J., Nespor, M., Mazuka, R., Horie, R., & Mehler, J. (2008). Bootstrapping word order in prelexical infants: A Japanese–Italian cross-linguistic study. *Cognitive Psychology*, *57*(1), 56-74.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3-55.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109-135.

Gomez, R. L., Gerken, L., & Schvaneveldt, R. W. (2000). The basis of transfer in artificial grammar learning. *Memory & Cognition*, *28*(2), 253-263.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Cambridge, Massachusetts: Harvard University Press.

He, A. X., & Lidz, J. (2017). Verb learning in 14-and 18-month-old English-learning infants. *Language Learning and Development*, *13*(3), 335-356.

James, W. *(*1890*). The principles of psychology (Vol. 1). Cambridge, MA: Harvard University Press.*

Jones, S. S., Smith, L. B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development*, *62*(3), 499-516.

Kern, S. (2007). Lexicon development in French-speaking infants. *First Language*, *27*(3), 227-250.

Kern, S., Langue, J., Zesiger, P., & Bovet, F. (2010). Adaptations françaises des versions courtes des inventaires du développement communicatif de MacArthur-Bates. *Approche Neuropsychologique des Apprentissages chez l'Enfant*, *107*(108), 217-228.

Kloos, H., & Sloutsky, V. M. (2008). What's behind different kinds of kinds: Effects of statistical density on learning and representation of categories. *Journal of Experimental Psychology: General*, *137*(1), 52.

Lempert, H. (1978). Extrasyntactic factors affecting passive sentence comprehension by young children. *Child Development*, 694-699.

Leung, J. H., & Williams, J. N. (2012). Constraints on implicit learning of grammatical form-meaning connections. *Language Learning*, *62*(2), 634-662.

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, *18*(3), 193-198.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods*, *164*(1), 177-190.

Mandler, J. M., & McDonough, L. (1993). Concept formation in infancy. *Cognitive Development*, *8*(3), 291-318.

McDonough, L., & Mandler, J. M. (1998). Inductive generalization in 9-and 11-month-olds. *Developmental Science*, *1*(2), 227-232.

Messenger, K., & Fisher, C. (2018). Mistakes weren't made: Three-year-olds' comprehension of novel-verb passives provides evidence for early abstract syntax. *Cognition*, *178*, 118-132.

Paradis, J. (2018, November). English L2 acquisition from childhood to adulthood. At the *43rd Boston University Conference on Language Development*. Association for Psychological Science, Paris.

Paradis, J., Tulpar, Y., & Arppe, A. (2016). Chinese L1 children's English L2 verb morphology over time: Individual variation in long-term outcomes. *Journal of Child Language*, *43*(3), 553-580.

Pozzan, L., Gleitman, L. R., & Trueswell, J. C. (2016). Semantic ambiguity and syntactic bootstrapping: The case of conjoined-subject intransitive sentences. *Language Learning and Development*, *12*(1), 14-41.

Quinn, P. C., & Bhatt, R. S. (2005). Learning perceptual organization in infancy. *Psychological Science*, *16*(7), 511-515.

Schulz, E., Wu, C. M., Ruggeri, A., & Meder, B. (2019). Searching for rewards like a child means less generalization and more directed exploration. *BioRxiv*, 327593.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323.

Strickland, B. (2017). Language Reflects "Core" Cognition: A New Theory About the Origin of Cross-Linguistic Regularities. *Cognitive science*, *41*(1), 70-101.

Van Heugten, M., Dahan, D., Johnson, E. K., & Christophe, A. (2012). Accommodating syntactic violations during online speech perception. In *Poster presented at the 25th Annual CUNY Conference on Human Sentence Processing, New York, NY*.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature human behaviour*, *2*(12), 915.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245.

3.2 Supplemental Information

3.2.1 Supplemental Analyses

Analysis S1. Known words.
We examined performance on filler trials with French words, in the same way we did for test trials.

Experiment 1: Infants.

Looking-while-listening.
A cluster-based permutation analysis showed that infants correctly identified the target word in the filler trials, looking more to the animate image when they heard 'the animate determiner *ko*+a French animate noun' than when they heard 'the inanimate determiner *ka*+a French inanimate noun' (960-2440ms time-window, $p < 0.001$; and 2900-5360ms time-window, $p < 0.001$, Fig. S1A).

Preferential looking.
A mixed-effects regression confirmed that infants correctly identified the target, with infants looking longer to the animate image when they heard '*ko*+French noun' (M = 0.6, SD = 0.09), then when they heard '*ka*+French noun' (M = 0.41, SD = 0.14): $\beta = 0.22$, SE = 0.03, $t = 7.29$, $p < 0.001$, Cohen's d = 0.96.

Infants thus appeared to correctly interpret known French nouns when preceded by novel determiners.

Experiment 2: Pre-school children.

Looking-while-listening.
A cluster-permutation analysis revealed that pre-school children looked longer to the animate image when they heard '*ko*+French word' than when they heard '*ka*+French word' from 520ms after the onset of the determiner to the prompt to point (5850ms, $p < 0.001$, Fig. S2B).

Looking preference.
A mixed-effects regression analysis confirmed that children looked significantly longer overall during the trial to the animate image when they had heard '*ko*+French word' (M = 0.7, SD = 0.19) than when they had heard '*ka*+French word' (M = 0.3, SD = 0.17) [$\beta = 0.4$, SE = 0.04, $t = 8.99$, $p < 0.001$, Cohen's d = 2.11].

Explicit choice.
Mean pointing accuracy on filler trials was 98% ± 0.99.

Taken together, these results suggest that pre-school children did not have difficulties interpreting known French words when preceded by novel determiners.

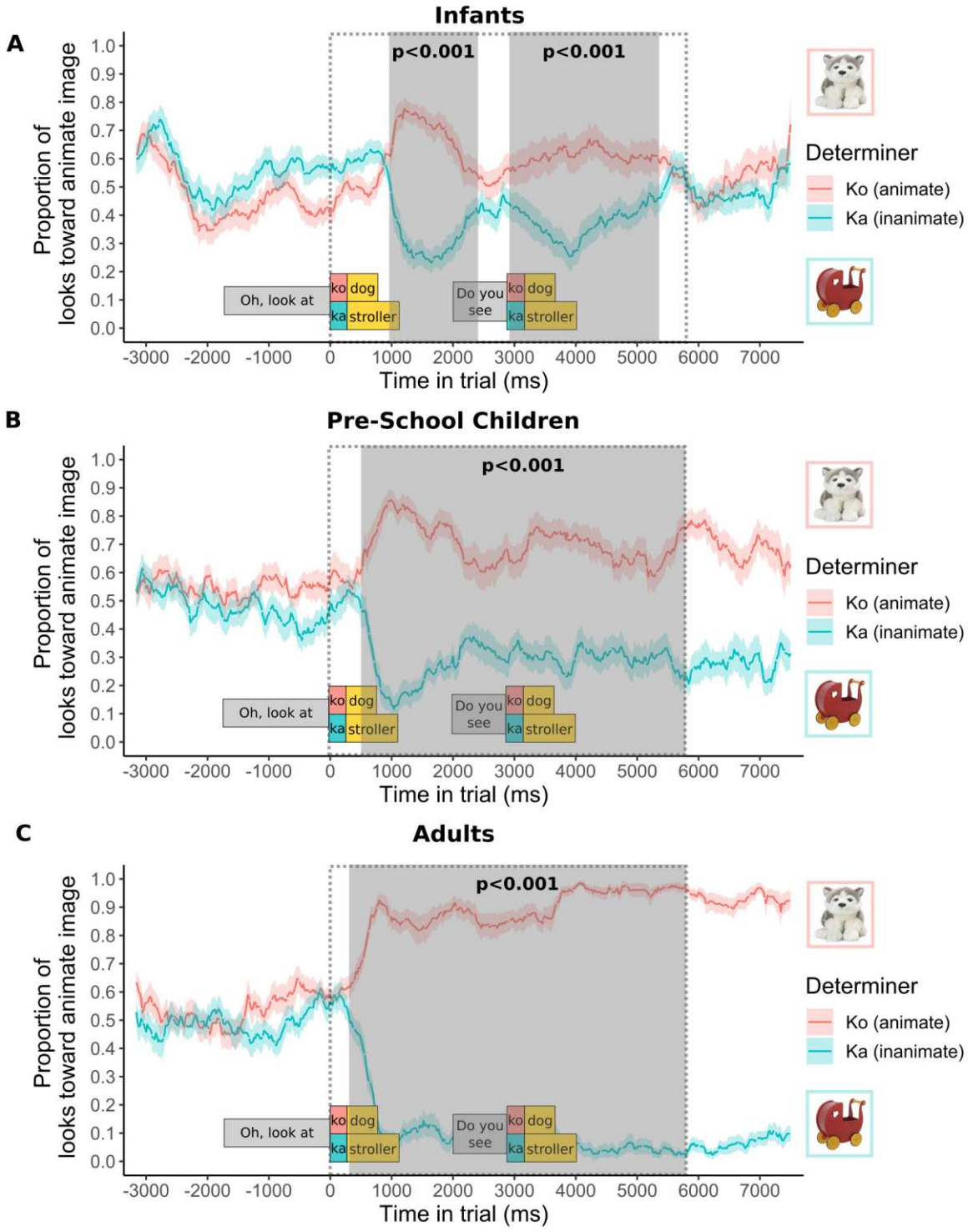


Figure S2. Gaze time-course for known words. (A) Experiment 1: Infants. Proportion of looks toward the animate image at each point in time during the trial, when the participants heard the animate determiner *ko* followed by a French noun (e.g., *ko* dog) in pink and when they heard the inanimate determiner *ka* followed by a French noun (e.g., *ka* stroller) in blue. Dark lines represent mean proportion across participants, and light shading the SEM (95% confidence intervals of the mean). The grey dotted box indicates the window from the onset of the determiner to the moment older participants were prompted to point. Grey shading indicates the time-window during which two conditions diverge significantly (960-2440ms ($p < 0.001$), and 2900-5360ms ($p < 0.001$); cluster-based permutation analysis). **(B) Experiment 2: Pre-school children.** Significant time-window: 520-5850ms ($p < 0.001$). **(C) Experiment 3: Adults.** Significant time-window: 300-5850ms ($p < 0.001$). Other visible divergences were not significantly different than what might be found by chance ($p > 0.1$).

Experiment 3: Adults.

Looking-while-listening.
A cluster-permutation analysis revealed that adults looked more to the animate image when they heard '*ko*+French word' than when they heard '*ka*+French word' from 300ms after the onset of the determiner to the end of the trial (5850ms, $p < 0.001$, Fig. S2C).

Looking preference.
A mixed-effects regression analysis confirmed that adults looked significantly longer overall during the trial to the animate image when they had heard '*ko*+French word' (M = 0.87, SD = 0.1) than when they had heard '*ka*+French word' (M = 0.13, SD = 0.11) [$\beta = 0.75$, SE = 0.04, $t = 19.22$, $p < 0.001$, Cohen's d = 5.02].

Explicit choice.
Mean pointing accuracy on filler trials was 99% ± 0.01.

Together, these results suggest that adults did not have major difficulties interpreting known French words when preceded by novel determiners.

Analysis S2: Anticipation.

To probe whether participants were using the novel determiners during online processing, we ran a cluster-based permutation analysis comparing looks to target during informative trials (where the determiner could be used to orient gaze to the target image, Fig. S1A) and uninformative trials (where only the noun can be used to orient gaze to the target image, Fig. S1B). Previous studies show that adults and children can orient their gaze based on information carried by determiners; however, these results pertain to existent (and thus well-known) determiners (e.g., Dahan et al, 2000; Van Heugten et al, 2012). This task aimed to investigate whether newly-learned determiners could, quickly, be used during online processing. If participants look faster to the target image in the informative than uninformative trials, then newly-learned determiners can quickly be used in online processing; however, if participants look just as fast to the target image in both the informative and uninformative trials, either the new determiners are not used this early on in learning for online processing or our measure was not fine-grained enough.

Experiment 2: Pre-school children.
We ran a cluster-based permutation analysis comparing gaze deployment during informative and uninformative trials. The analysis revealed no significant difference in looks to target in informative versus uninformative trials (Fig. S3B). Pre-school children thus did not appear to be orienting their gaze differently when the determiner was informative in comparison to when the determiner was uninformative. These results do not allow us to pinpoint whether this is because pre-school children were not interpreting the novel determiners online or whether our measure was just too noisy.

Experiment 3: Adults.
We ran the same cluster-based permutation analysis on adult data. The analysis revealed no significant difference in looks to target during informative versus uninformative trials (Fig. S3B). Just like pre-school children, adults too did not appear to be orienting their gaze differently when the determiner was informative in comparison to when the determiner was

uninformative. Again, these results do not allow us to pinpoint whether this is because of a lack of processing or a too coarse-grained measure.

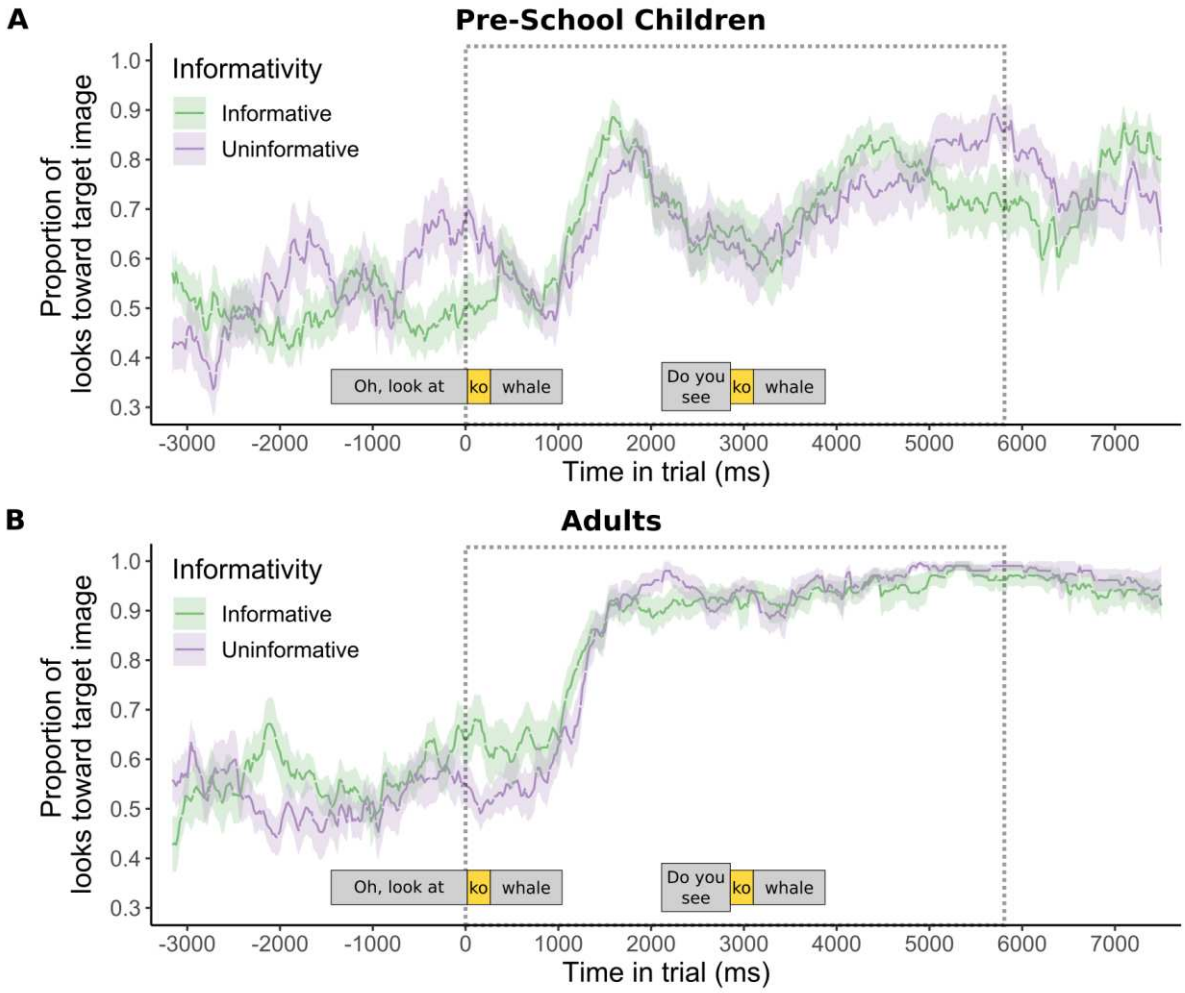


Figure S3. Gaze time-course for anticipation trials. (A) Experiment 2: Pre-school Children. Proportion of looks toward the target image at each point in time during the trial, when the participants saw an informative visual scene (one animate and one inanimate) in green versus when they saw an uninformative visual scene (two animates or two inanimates) in purple. Dark lines represent mean proportion across participants, and light shading the SEM (95% confidence intervals of the mean). Grey dotted box indicates the window from the onset of the determiner to the moment participants were prompted to point. Children did not orient their gaze differently on informative and uninformative trials. **(B) Experiment 3: Adults.** Adults did not orient their gaze differently on informative and uninformative trials.

Adults however seemed to be using distinct strategies during the task (applying a rule versus following intuition). Therefore, an effect of anticipation could have been diluted because one group was anticipating but not the other. We thus ran two additional post-hoc cluster-based permutation analyses on the group of adults that had explicit knowledge of the rule governing determiner use and on the group that did not. We hypothesized that the Explicit Knowledge group may show an effect because of solid knowledge of determiner use or, inversely, that the No Explicit Knowledge group may show an effect because of implicit processing of the determiners rather than top-down rule application. The alpha was Bonferroni-corrected for multiple comparisons, $\alpha = 0.05/2$.

The cluster-based permutation analyses revealed that neither adults who had explicit knowledge nor adults who did not have explicit knowledge oriented their gaze significantly earlier to the target in the informative trials. Further research is needed to determine whether use in online comprehension requires more exposure or whether our measure was just too coarse-grained.

Analysis S3: Across experiment comparison (Adults).

To determine whether there was a significant difference in adults' performance between the current study (Four-day) and the one in Barbir et al (One-day), we compared the two groups. We hypothesized that participants in the Four-day study may have higher overall performance, because they had more exposure.

Looking-while-listening.
A cluster-based permutation analysis comparing looks to target in each of the two studies showed that participants did not orient their gaze differently in two studies (Fig. S4A).

However, adults were using two substantially different strategies during the test phase (explicit rule application and intuition following), and the kind of study may have uniquely affected one of the two strategies but not the other: the Explicit Knowledge group may have been more sure of the rule they extracted, or the No Explicit Knowledge group may have had more time to build up implicit representations. We thus ran two separate cluster-based permutation analyses for each knowledge kind (Explicit Knowledge/No Explicit Knowledge). Again, there were no significant differences in deployment of gaze to target between the two studies for adults in the Explicit Knowledge group (Fig. S4B) nor adults in the No Explicit Knowledge group (Fig. S4C).

Preferential looking.
Adults did not look more to the animate image when they heard the animate determiner than the inanimate determiner, in the Four-Day or One-Day study [no interaction between Study and Determiner: $\beta = 0.07$, SE = 0.1, $t = 0.71$, $p = 0.48$, Fig. S4A]. Moreover, the experiment did affect the two knowledge groups' proportions of animate image choices differently [no interaction between Study, Determiner and Knowledge: $\beta = -0.25$, SE = 0.21, $t = -1.21$, $p = 0.23$].

Explicit choice.
A generalized linear mixed-effects analysis revealed that adults in the two experiments did not differ in the proportion of the animate image choices when they heard the animate and inanimate determiners (Four-day: *ko*: M = 0.81, SD = 0.4; *ka*: M = 0.31, SD = 0.46; One-day: *ko*: M = 0.73, SD = 0.45; *ka*: M = 0.23, SD = 0.42) [no interaction between Study and Determiner: $\beta = -1$, SE = 1.64, model comparison: $\chi^2(1) = 0.37$, $p = 0.54$]. Furthermore, the experiment did not affect the proportion of animate image choices differently for the two knowledge groups: Explicit Knowledge and No Explicit Knowledge [no interaction between Study, Determiner and Knowledge: $\beta = -0.8$, SE = 1.05, model comparison: $\chi^2(1) = 0.58$, $p = 0.45$].

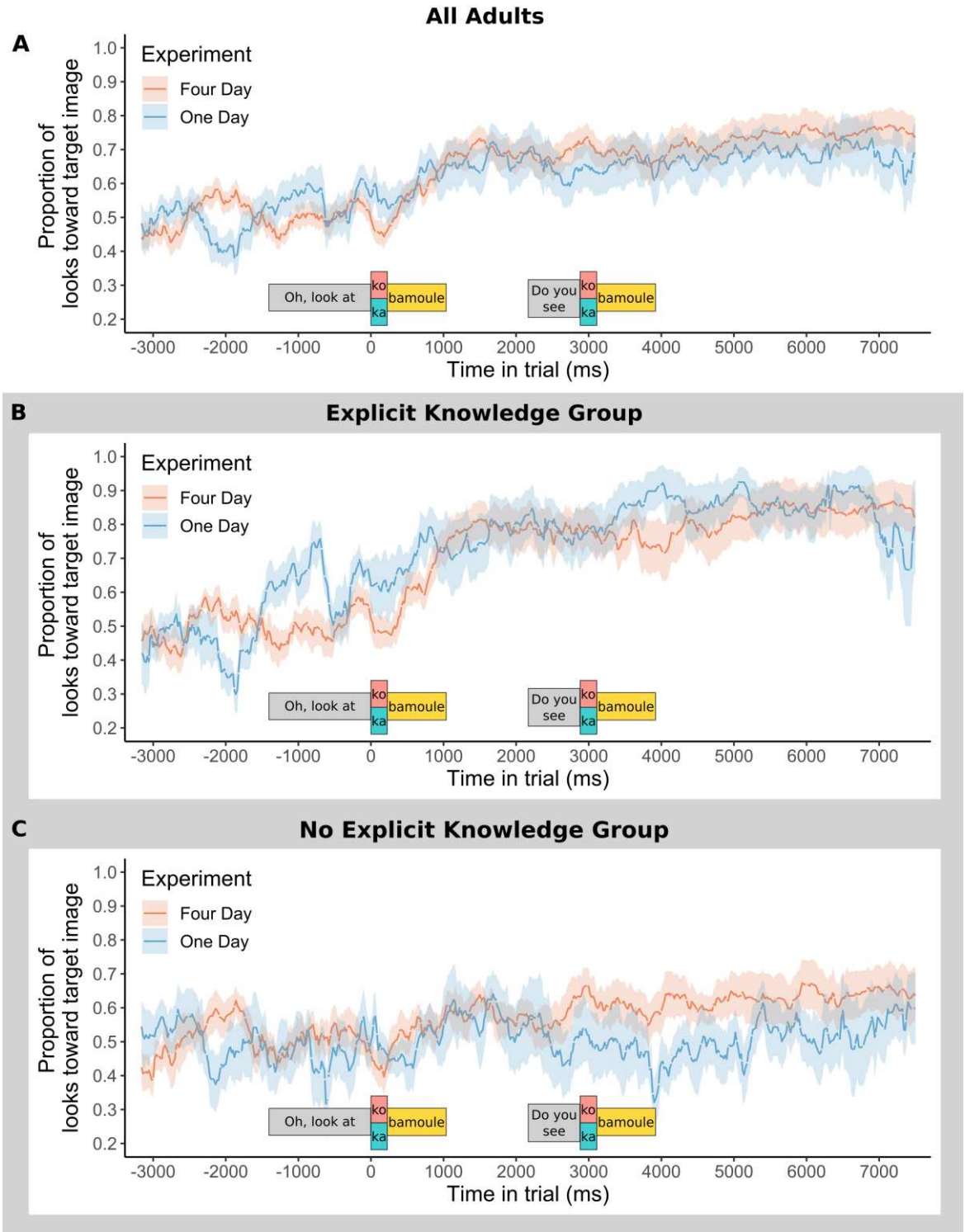


Figure S4. Gaze time-course across studies. Current study (four-day training phase) & supplemental study from Barbir et al (one-day training phase). **(A) Four-day vs. One-day.** Proportion of looks to the target image at each point in time during the trial, for participants in the Four-day Study in orange versus participants in the One-day Study in blue. Dark lines represent mean proportion across participants, and light shading the SEM (95% confidence intervals of the mean). There were no significant differences in gaze orientation to target between the two studies. **(B) Explicit Knowledge group: Four-day vs. One-day.** Subset of participants who had explicit knowledge of the rule governing determiner use. No differences. **(C) No Explicit Knowledge group: Four-day vs. One-day.** Subset of participants who did not have explicit knowledge of the rule governing determiner use. No differences.

In sum, there did not seem to be a significant difference in performance between adults in the current study (Four-day) and in the supplemental study from Barbir et al (One-day). There was also no significant difference between studies for adults who had discovered the rule governing determiner use or for those who were following intuitions. Further research is needed to determine whether infants too could potentially extract the rule governing determiner use in a one-day study. Such results would point to an incredibly potent language-learning strategy early in life.

3.2.2 Supplemental Experiment.

Experiment S1: Pre-school children (one-day study).

We first tested the novel training videos (adapted from Barbir et al) with a group of pre-school children, without the four-day training session. This experiment was carried out prior to the main experiment. It was not included in the pre-registration.

Participants.
Pre-school children were recruited at a municipal kindergarten. All were monolingual French speaking children, who did not hear another language at home. A total of 30 children were tested in two classes. Children were excluded for failing to correctly respond to the two training trials (n = 1), not being monolingual (n = 3), and technical error (n = 1). The remaining 25 children were included in the analyses (mean: 4;9 years; range: 4;2-5;3 years; 10 girls, 15 boys). Written informed consent was obtained from each child's parents prior to the experiment. All research was approved by the local ethical board: CER Paris Descartes.

Materials.
The materials were the same as in the main experiment, except for differences noted below.

Training video. The (one) training video included all the stories, presented once, except the one familiarizing participants with novel nouns (total ~10mins long). There were no anticipation items.

Procedure. The procedure was the same as the main experiment, except for the following differences.

Children were tested individually in a quiet room at their school.

There was no visual prompting signal to make an explicit choice (i.e., no orange frame). Instead, the experimenter prompted the child the point 10s into the trial (roughly the total length of each trial).

There was no anticipation phase.

The experiment lasted approximately 15 minutes.

Parents did not fill out any questionnaires.

Analyses.

Test items.

Looking-while-listening.

A cluster-based permutation analysis revealed that children did not look significantly more to the animate image when they heard *ko* than when they had heard *ka*, during any time-window (Fig. S5A).

Preferential looking.

A mixed-effects regression analysis showed that children did not look significantly longer overall during the trial to the animate image when they had heard *ko* than when they had heard *ka*: $\beta = -0.003$, SE = 0.04, t = -0.09, p = 0.93, Cohen's d = -0.01 (means shown in Fig. S5B(i)).

Explicit choice.

A generalized linear mixed-effects model revealed that children did not choose the animate image significantly more often when they heard the determiner *ko* than when they heard the determiner *ka*: $\beta = 0.39$, SE = 0.49, p = 0.42, model comparison: $\chi^2(1) = 0.64$, p = 0.43, Cohen's d = 0.16 (means displayed in Fig. S5C(i)).

Known words.

Looking-while-listening.

A cluster-based permutation analysis showed that children looked longer to the animate image when they heard '*ko*+French word' than when they heard '*ka*+French word' during the time-windows from 320ms to the prompt to point (5850ms, $p < 0.001$).

Preferential looking.

A mixed-effects regression analysis confirmed that children looked significantly longer overall during the trial to the animate image when they had heard '*ko*+French word' (M = 0.74, SD = 0.2) than when they had heard '*ka*+French word' (M = 0.24, SD = 0.17) [$\beta = 0.5$, SE = 0.04, t = 14.23, $p < 0.001$, Cohen's d = 2.67].

Explicit choice.

Mean pointing accuracy on filler trials was 96.5% ± 1.3.

The data from test trials and filler trials indicate that pre-school children did not seem to use the novel determiners to interpret new nouns. However, this was not a because children were overall confused with the task: children did not appear to have major difficulties interpreting known French words when preceded by novel determiners. For filler trials (i.e., known words), the effect onset for the significant time-window (when children start to look to animate image, 320ms) is, actually, comparable to the effect onset time observed for adults (300ms, Fig. S2C)

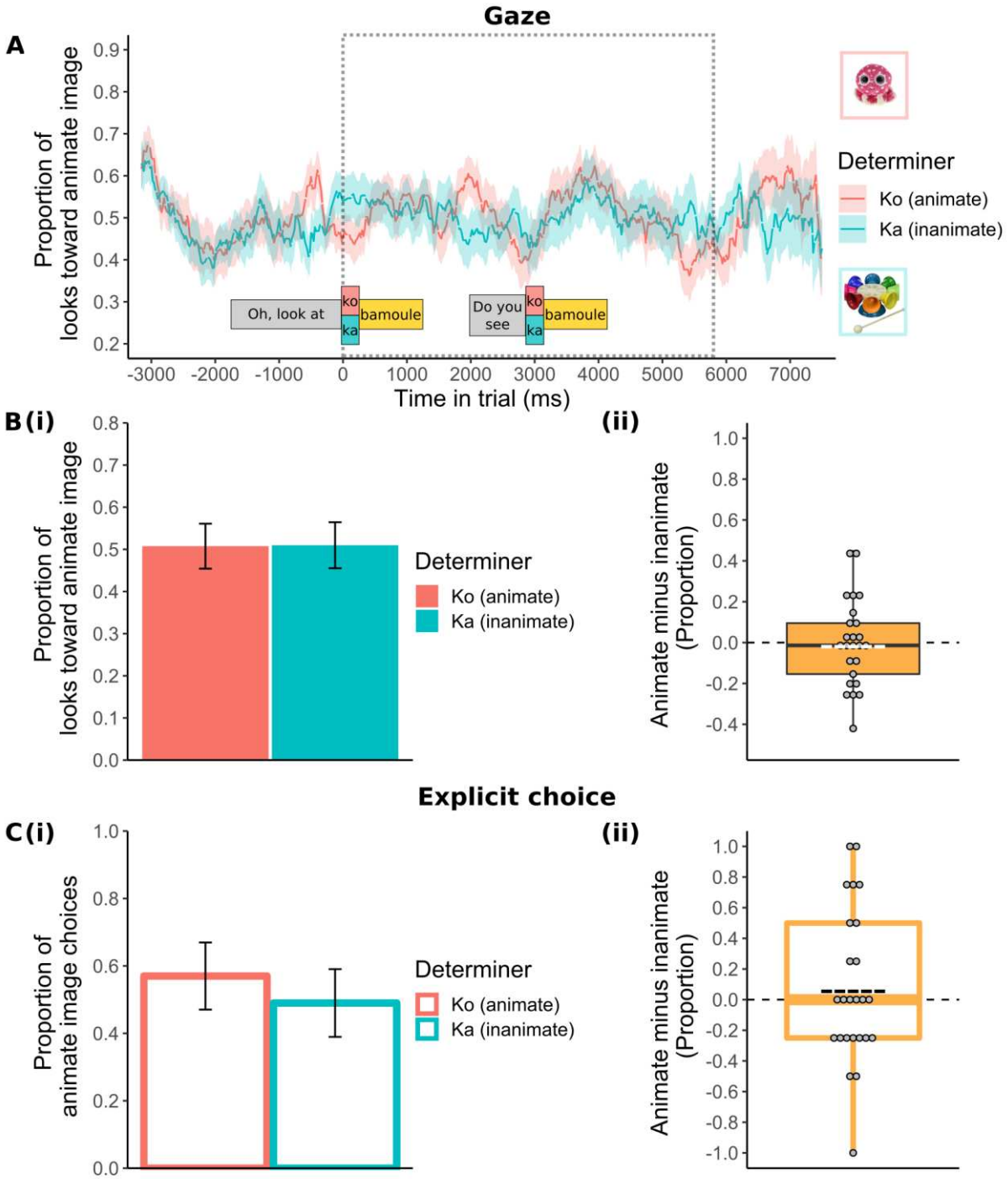


Fig. S5. Supplemental Experiment Results (Pre-school children, one-day study). (A) Time-course. Pre-school children did not look more to the animate image when they heard the animate determiner *ko* than the inanimate determiner *ka* at any time-point during the trial. **(B(i)) Overall looking preference.** Pre-school children did not look significantly longer to the animate image when they heard the animate determiner. (**B(ii)**) **Overall looking preference per participant.** The per participant difference between the proportion of looks to the animate image choices when participants heard the animate and inanimate determiners. **(C(i)) Explicit choice.** Pre-school children did not choose the animate image more often when they heard the animate determiner. **(C(ii)) Explicit choice per participant.** The per participant difference between the proportion of animate image choices when participants heard the animate and inanimate determiners.

3.2.3 Annex

Annex S1. Instructions for parents (Experiment 1: Infants and 2: Pre-school children).

Instructions for Parents (English)

Dear Ms./Mr. X.

In preparation for your visit to our BabyLab this Saturday (May 11th), here are the links to the short films that we would like you to show to X, once each day (at any moment of the day) as follows:

- **Tomorrow: Wednesday, May 8th**
  Link
- **The day after tomorrow: Thursday, May 9th**
  Link
- **Friday, May 10th**
  Link

(please contact me if you have technical problems with the videos)

Instructions

- It is preferable to show the child the videos in *FULL SCREEN* mode, on a TV or a computer screen. It is best if the child is seated facing the screen, at about 50cm and 1m from the screen, and attentive.
- It is important to show each video only once.
- If your child's attention drifts off or she/he starts looking elsewhere, don't hesitate to pause the video and take a short break. Encourage the child to refocus her/his attention on the screen, without making any references to the content of the video. When your child is looking back at the screen, you can continue to watch the video.
- Please stay as neutral and silent as possible during the video, without making comments.
- Lastly, it is essential not to refer to the content of the videos once the child has seen them.

Thank you very much for your commitment.
Best,
X

Instructions pour parents (Français)

Bonjour Madame, bonjour Monsieur,

En prévision de votre venue de ce samedi 11 mai à notre BabyLab, voici les liens vers les petits films que nous vous demandons de montrer à X, une fois chaque jour (peu importe le moment de la journée) soit :

- ***demain, mercredi 8 mai** :*
  Lien
- ***après-demain, jeudi 9 mai** :*
  Lien
- ***vendredi, le 10 mai** :*
  Lien

(n'hésitez pas à me contacter si vous rencontrez un quelconque problème technique)

20mois

Voici quelques consignes :

- *Il est souhaitable de lui montrer les vidéos en mode PLEIN ÉCRAN sur une télévision ou sur un écran d'ordinateur. L'idéal est que le bébé soit assis sur vos genoux, situé entre 50 cm et 1 m de l'écran, qu'il soit disponible et en forme.*

- *Il est très important de ne montrer chaque vidéo qu'une seule fois.*

- *Si votre bébé devient moins attentif ou regarde ailleurs, n'hésitez pas à faire une pause. Encouragez-le si besoin a refocaliser son attention sur l'écran mais sans faire référence au contenu de la vidéo. Lorsque votre bébé regarde de nouveau, vous pouvez relancer la vidéo.*

- *Vous, parents, soyez le plus neutre et silencieux possible, sans faire de commentaires.*

- *Dernière chose : Il est important de ne pas faire référence au contenu des vidéos une fois qu'elles ont été visionnées.*

4 ans

Voici quelques consignes :

- *Il est souhaitable de lui montrer les vidéos en mode PLEIN ÉCRAN sur une télévision ou sur un écran d'ordinateur. L'idéal est que votre enfant soit assis en face de l'écran, à une distance comprise entre 50 cm et 1 mètre.*

- *Si possible, merci de privilégier les moments de la journée où il/elle est disponible et en forme.*

- *Il est très important de ne montrer chaque vidéo qu'une seule fois.*

- *Si votre enfant devient moins attentif ou regarde ailleurs, n'hésitez pas à faire une pause. Encouragez-le/la si besoin a refocaliser son attention sur l'écran mais sans faire référence au contenu de la vidéo. Lorsque votre enfant regarde de nouveau ou qu'il/elle réclame la suite, vous pouvez relancer la vidéo.*

- *Vous, parents, soyez le plus neutre et silencieux possible, sans faire de commentaires.*

- *Dernière chose : Il est important de ne pas faire référence au contenu des vidéos une fois qu'elles ont été visionnées.*

Merci beaucoup par avance et bonne fin de journée, bien cordialement,

X

Annex S2. Questionnaire (Experiment 1: Infants).

1. Did the child watch the entire video?
   1st day: Yes No
   2nd day: Yes No
   3rd day: Yes No

2. Between the 1st and 3rd viewings, the child's attention…
   Increased Stayed the same Decreased

3. Does the child know the following words:

Dog:	Understands & Produces	Understands	Doesn't understand
Cat:	Understands & Produces	Understands	Doesn't understand
Book:	Understands & Produces	Understands	Doesn't understand
Bottle:	Understands & Produces	Understands	Doesn't understand
Shoe:	Understands & Produces	Understands	Doesn't understand
Rabbit:	Understands & Produces	Understands	Doesn't understand
Chicken:	Understands & Produces	Understands	Doesn't understand
Pig:	Understands & Produces	Understands	Doesn't understand
Duck:	Understands & Produces	Understands	Doesn't understand
Frog:	Understands & Produces	Understands	Doesn't understand
Ball:	Understands & Produces	Understands	Doesn't understand
Banana:	Understands & Produces	Understands	Doesn't understand
Boat:	Understands & Produces	Understands	Doesn't understand
Train:	Understands & Produces	Understands	Doesn't understand
Cake:	Understands & Produces	Understands	Doesn't understand
Stroller:	Understands & Produces	Understands	Doesn't understand
Car:	Understands & Produces	Understands	Doesn't understand
Butterfly:	Understands & Produces	Understands	Doesn't understand
Mouse:	Understands & Produces	Understands	Doesn't understand
Whale:	Understands & Produces	Understands	Doesn't understand
Tractor:	Understands & Produces	Understands	Doesn't understand
Baboon:	Understands & Produces	Understands	Doesn't understand

Annex S3. Questionnaire (Experiment 2: Pre-school Children).

1. Did the child watch the entire video?
   1st day: Yes No
   2nd day: Yes No
   3rd day: Yes No

2. Between the 1st and 3rd viewings, the child's attention…
   Increased Stayed the same Decreased

3. Did the child ask questions about the words 'ko' and 'ka'?
   Yes No

4. Does the child know the following words?
   Squid: Yes No
   Top: Yes No

Annex S4: Instructions for participants (Experiment 3: Adults).

Adult Instructions (English)

You are about to participate in an experiment we also run with children. Before coming to the lab, you have three videos to watch at home, one per day, for three consecutive days. The video contains clues that will be important during the test at the lab.

The videos are essential for the experiment. The time necessary to watch them is included in the sum you will receive for your participation.

It is imperative to watch the videos:

1. Only once
   - do not re-watch certain parts even if you feel you missed something, everything must watch the videos in the same manner and duration.
2. In one sitting
   - do not stop the video to jot things down, reflect or do something else, for example answering as phone call
   - please wait until the video is full loaded before watching it, this is prevent it from stopping to load during the viewing
3. On a big screen (e.g., computer)
4. Without distractions (e.g., music, reading emails, etc.)

This experiment is for adults and children. Because adults understand experiments differently than children, please:

1. Be attentive to the video from the beginning to the end, as much as possible (even if you have the impression that you are watching the same scene for the thousandth time)
2. Do not jot down content from the video
3. Between viewings, try not to explicitly think about the content of the videos and did not discuss the content of the videos with other people before coming to the lab

Some problems you may encounter:

1. Q. Oh dear! I forgot to watch the video the second day!
   A. Watch the video as soon as you can, and call us, we will try to push the day of the test back.
2. Q. Oh, I made a mistake and I watched the same video two days in a row. What should I do?
   A. Do no worry. Watch the next video the day after, and inform us when you come to the lab.

The videos are similar, certain scenes are the same and others change day to day.

To ensure that everyone has watched the videos with the upmost attention: there is a *password* in each video. The passwords, one for each video, are written on the bottom of the screen and are visible. Take note, we will ask you for them when you come to the lab. If you saw the password, but you did not have time to write it down, do not go back, do not pause the video: tell us, we will ask you another question about the video.

Instructions pour adultes (Français)

Vous allez participer à une expérience que nous menons avec des enfants. Avant de faire l'expérience au labo, vous aurez trois vidéos à regarder à la maison, une par jour, pendant trois jours consécutifs. La vidéo contient des indices qui seront utiles pendant le test au labo.

Les vidéos constituent une partie importante de l'expérience, le temps nécessaire pour les regarder est inclut dans le montant que vous allez recevoir après l'expérience.

Il est important de regarder les vidéos :

1. *Uniquement une fois*
   - *n'allez pas en arrière même si vous sentez que vous avez raté quelque chose, nous avons besoin que tout le monde regarde la vidéo de la même façon et dans les mêmes temps*
2. *En une seule fois*
   - *sans arrêter la vidéo pour noter des choses, réfléchir ou faire autre chose, comme par exemple répondre au téléphone*
   - *il faut attendre qu'elle soit totalement chargée avant de la regarder, afin d'éviter qu'elle s'interrompe*
3. *Sur un grand écran (e.g., ordinateur)*
4. *Sans distractions (e.g., musique, lecture d'emails, etc.)*

Ceci est une expérience pour adultes et enfants. Comme les adultes ont une différente manière d'approcher les expériences que les enfants, il est impératif que vous :

1. *Prêtiez attention à la vidéo de début à la fin, autant que possible (même si vous avez l'impression de regarder la même scène pour la nième fois)*
2. *Ne prenez pas des notes pendant la vidéo*
3. *Entre les visionnements de la vidéo, ne réfléchissez pas explicitement au contenu du film et ne discutez pas du contenu de la vidéo avec des autres personnes avant de venir au labo*

Problèmes que vous pouvez rencontrer :

1. *Q. Oh, mais j'ai oublié de regarder la vidéo le deuxième jour !*

R. Regardez la vidéo aussi tôt que possible, et appelez-nous, nous essaierons de décaler votre rdv d'un jour.

2. *Q. Oh, je me suis trompé.e et j'ai regardé la même vidéo deux jours de suite, qu'est-ce que je fais ?*
   R. Pas de panique. Regardez simplement la vidéo suivante le lendemain, et précisez-le nous lors de votre venue au labo.

Les vidéos sont similaires, certaines scènes sont les mêmes et d'autres changent d'un jour à l'autre.

*Pour s'assurer que tout le monde regarde les vidéos avec attention : Il y a un *mot de passe* dans chaque vidéo. Ces mots de passes, différents pour chaque vidéo, seront affichés en bas de l'écran, de manière bien visible. Notez-les, nous allons vous les demander lors de votre venue au labo. Si vous avez vu un mot de passe, mais vous n'avez pas eu le temps de le noter, n'allez pas en arrière, ne mettez pas la vidéo en pause : dites-le nous, nous allons vous poser une autre question à propos de la vidéo.*

Annex S5. Post-experiment questionnaire (Experiment 3: Adults).

Questionnaire (English)

I.
Languages
Languages heard at birth:

Languages spoken or understood (level and date at which you started learning the langauge):

Do you like learning languages?

II.
Experiment: At home
Did you watch the three videos in their entirety?

How attentive were you while watching the videos?

Did you think about the videos between viewings?

III.
Experiment: In-lab
1. a) During the test phase, how did you go about responding to new words ?

b) How confident are you on a scale of 1 (I guessed) to 10 (I am sure)?

c) Do you remember the words? Please tell us those your remember.

IV.
Experiment: The Rule
1. a) Did you notice ‘ko’s and ‘ka’s during the video ? If so, any did you think they were there?

b) Did you notice a rule governing their use? How did you discover the rule?

c) When did you discover the rule (which video, what point in the video, during test, or after experiment)?

Questionnaire (Français)

I.
Langues
Langues entendues à la naissance :

Langues parlées (niveau + date d’acquisition) :

Est-ce que vous aimez apprendre des langues :

II.
A la maison
Avez-vous regardé les trois vidéos à la maison ?

A quel point étiez-vous attentif/attentive ?

Avez-vous pensé à la vidéo entre les visionnements ?

III.
Au labo
1. a) Pendant le test, comment avez-vous fait pour répondre aux questions avec des nouveaux mots ? (si vous aviez une stratégie, décrivez la succinctement)

b) A quel point étiez-vous confiant.e dans vos réponses pour les nouveaux objets/mots pendant le test ? Estimez votre niveau de confiance sur l’échelle de 1 (*j’ai deviné au hasard*) à 10 (*je suis sûr.e*).

c) Est-ce que vous vous souvenez de certains/de la totalité de ces mots ? Notez ceux dont vous pensez vous souvenir.

IV.
La règle
2. a) Avez-vous remarqué des « ko » et « ka » pendant le film ? Si oui, à quoi pensez-vous qu’ils servaient ?

b) Pensez-vous avoir observé une règle qui déterminerait l'usage de ces mots ? Comment pensez-vous avoir procédé pour découvrir cette règle ?

c) Quand avez-vous remarqué la règle ? (Quel visionnement, quel moment de la vidéo, pendant le test, ou après l'expérience)

4. LEARNING KNOWLEDGE

Our results from Chapter Three show that the same set of evidence can spur generalization early in development and at maturity, but not in-between. We proposed that this non-linear trend may reflect learning strategies, rather than learning capacity. Early in learning, a lower generalization threshold could allow the mind to begin making sense of its external environment; further into the learning process, a higher generalization threshold could allow the mind to attain solid knowledge. All learners would be unified in seeking indubitable knowledge, but would have diverging moment-to-moment knowledge thresholds, depending how close they could get to that ideal with the evidence at hand. The evidence set that we have considered so far has been composed of external stimuli and some internal prior knowledge. External stimuli, however, can come in two kinds: messy raw evidence from the external world and knowledge-state evidence from other minds. In other words, amongst the 'blooming, buzzing confusion' (James, 1890) that is the external world, there are numerous other minds trying to make sense of the confusion, extracting knowledge as best they can. A learner could thus turn to other minds for already processed and certified information. Yet, knowledge-state evidence, created by finite minds, represents how the world is likely to be, and not how it is. A learner may thus gain somewhat disparate knowledge by processing a raw evidence set or by assimilating knowledge-state information. In a series of pre-registered experiments, we compare the effect supervised learning (i.e., learning direct translations of vocabulary: knowledge-state evidence) and unsupervised learning (i.e., gleaning information from uncertain contexts: raw evidence) have on performance and confidence. We test three groups that vary in age and literacy level: (literate) adults, literate first-grade children, and pre-literate first-grade children. Our results reveal that supervision universally boosts confidence, but has variable effects on performance, ranging from positive to downright negative. We advance the conclusion that knowledge-state evidence may be appealing insofar as it provides finite minds with the indubitability they seek in their interactions with the external world.

***

This study originally had one group of adults (n = 48) and one group of children (n = 48). Its goal was to compare supervised and unsupervised learning across age. We had pre-registered analyses examining the effect of literacy, which we planned to run if we had approximately balanced groups of pre-literate and literate children. In the end, the groups were balanced, and we discovered an interaction between literacy and learning condition: supervision was a boon for literate children, but a bane for pre-literate children. We will thus extend this study, doubling the number of children (total = 96; 48 literate and 48 pre-literate children), investigating both the effect of age and literacy. This extension, as well as precise analyses, will be added as an amendment to our existent pre-registration.

4.1 Why we prefer suboptimal language learning strategies: Supervision distorts the relationship between decisions and confidence

Barbir, M., Havron, N., Recht, S., Fiévet, A.-C., & Christophe, A.

Acquiring a new capacity, like speaking a language, requires accumulating enough evidence to sway a decision in the right way. The human mind can accumulate evidence spontaneously, in a highly unsupervised way, or, in contrast, the process of evidence accumulation can be bolstered with varying degrees of supervision, such as feedback. Humans, from infancy through to adulthood, have been shown to be highly adept at learning language in unsupervised settings. Diverse aspects of language, ranging from phonology (Maye, Werker, & Gerken, 2002), to word segmentation (Saffran, Aslin, & Newport, 1996; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997), and word meanings (Yu & Smith, 2007; Smith & Yu, 2008), and even grammar (Gomez & Gerken, 1999), can be gleaned off of evidence without supervision. Even more impressively, this kind of learning can be quick (in as little as just 2 minutes, Saffran et al, 1996; Gomez & Gerken, 1999) and efficient (with as few as just one exemplar of a category, Gerken, Dawson, Chatila, & Tenenbaum, 2015). However, what can be acquired this way in a short time and with few examples, will depend on the element that needs to be learned (Peña, Bonatti, Nespor, & Mehler, 2002; Gerken & Bollt, 2008) and the subset of the evidence presented (Gerken & Knight, 2015). Some elements, such as abstract structures (Peña et al, 2002), and some subsets of evidence, such as those embedded in ambiguous environments (Gillette, Gleitman, Gleitman, & Lederer, 1999), are acquired more easily (or at least quickly enough to be observed in the lab) with supervision. Supervision, though, is often considered as little more than a means to an end, a sort of evidence 'cheat sheet' to help a learner sway her decision the right way. Beyond the end attained and the skill acquired, supervision may leave an imprint on the mind. Acquiring a new skill with or without supervision, then, may have a disparate effect on processing, and, accordingly, on how a learner makes decisions.

Supervision can take many forms. It may involve altering the evidence: for example, slightly emphasizing some elements, like gaps between words to promote word segmentation (e.g., hearing *biboba* or hearing *bibo* *short silence* *ba*, Peña et al, 2002); or adding critical evidence: for instance, providing sentential information to aid in the acquisition of a novel word meaning in an ambiguous visual context (e.g., observing an ocean scene with a multitude of fish and coral, and hearing '*Bibo'* or hearing 'The *bibo* swims so fast', Gillette et al, 1999). Supervision also includes presenting the evidence differently: instead of presenting a learner with many examples from which she can deduce a regularity, the learner could, simply, be presented with an explanation of the regularity. This kind of supervision should ring a bell; it is the way a large chunk of classroom language learning is done.

Explanations are a form of testimonial evidence. A learner of French can be told, by a reliable source, that *chat* means 'cat' in her native language English, and this evidence will be enough to allow her to make a future decision correctly, right off the bat. The very next time she hears *chat*, as long as she remembers the translation, she will be able to interpret the utterance as one about felines, and the very next time she wants to say something about cats, she will be able to say *chat*. In stark contrast, the unsupervised learner may not be able to glean the meaning of a word from just one use of it in a sentence (e.g., '*Le chat est mignon*.'), especially if she knows few other words in the language. To be able to understand and use the word *chat*, she will likely have to hear multiple sentences, and make deductions progressively, step-by-step, using both linguistic and sensorial evidence. She may hear: '*Le chat ronronne', 'Un chat dort', 'Le chat joue'*, and notice that *chat* co-occurs with words such as *le* (the,

masculine), *un* (a, masculine), *ronronne* (purrs), *dort* (sleeps), *joue* (plays). She could thus glean off of what she knows about those words (e.g., only animates play, nouns are preceded by articles, etc.) that *chat* is likely an animate noun, and a feline. The transition from not knowing what *chat* means to having enough evidence to understand and use the word may thus take more time for the unsupervised learner than for her supervised counterpart. Supervision can thus either serve as an accelerator, speeding up learning, or as a bolster, sparking learning in unfavourable situations. In the presence of a cat, both learners will, eventually, be able to correctly utter *chat*; yet the two will have converged at the same endpoint after vastly diverging itineraries, notably in terms of length. Accordingly, supervised learning is ostensibly appealing. It is faster; it is simpler. However, the evidence used to sway the supervised learner's decision (e.g., identifying the thing over there as *chat*), and thus her future decisions (e.g., using *chat* when speaking of cats), is likely to be radically different from that of the unsupervised learner.

At first glance, a difference in kind of evidence may appear innocuous as long as the supervised and unsupervised learners both acquire a near-identical scope of meaning, as long as the ensuing decision is identical. Yet, supervised learning is funneled into a particular direction, and will thus bypass evidence that would have otherwise been taken into consideration. Furthermore, supervised learning by way of testimony may be considered infallible primarily because it was supervised and not because of a lack of counter-evidence in the world (e.g., noticing that *chat* is always used to refer to cats, and never to dogs). Thus, even though both learners may make similar decisions, interpreting *chat* as 'cat', the supervised learner will likely overlook the systematic co-occurrence of *le* and *chat,* or the co-occurrence of *ronronne* (purrs or hums) with both *chat* and *moteur* (motor). As such, supervision may be an attractive shortcut, but it may inadvertently alter the way in which language is processed and stored in the mind. Here, we thus asked whether cognitive processing of language differs after supervised or unsupervised learning.

Supervision may have a two-fold effect on cognitive processing. First, it will likely only present the learner with one dimension of the evidence (however essential this dimension may be). Second, the medium of supervision itself may contribute to the evidence: the act of being supervised, of being told by a reliable third-party that the world is a certain way, may be a kind of second-order evidence, evidence about the evidence. Broadly, supervision may distort the evidence a learner will rely on to make her decision, and, in accordance, how the evidence is processed by the mind.

Though the aim of supervision is to accelerate or bolster learning, it need not be based on how the mind learns. In the case of learning that *chat* = cat, it is taken for granted that *chat* is an isolable, self-standing linguistic item that has an intrinsic meaning associated to it, a sort of dictionary entry. Intuitively, words appear to bear the weight of their meaning on their own, and yet semantic processing, the cognitive processing of meaning, is not as clear cut as a word's boundaries. On one hand, cognitive processing relies on a word's sentential context, almost as if the meanings of words were steeped into co-occurring words and grammar. A word in a sentential context is interpreted faster than in isolation (e.g., hearing 'I fed the *cat'* or hearing '*cat'*, Lieberman, 1963), and this is so even when the context is seemingly neutral, such as 'Look at the *cat'*, Fernald & Hurtado, 2006). Furthermore, it has been shown that a word's context aids online processing, from a very young age (Dahan, Swingley, Tanenhaus, & Magnuson, 2000; van Heugten & Christophe, 2014). A listener who hears '*le',* the French masculine article, will anticipate a masculine noun (e.g., *chat*). Sentential contexts provide an unsuspected amount of information about a word's meaning, and reliably so. They are one of the tools infants use to bolster word learning (Gleitman, 1990; He & Lidz, 2017). Even just one

function word, like 'the/a' or 'she/he/they' can delimit the possible meaning of a word: if one hears 'The *bamoule'*, one will know right away that *bamoule* is likely an object; in contrast, if one hears 'They *bamoule'*, one will be able to deduce that *bamoule* is an action. On the other hand, some aspects of grammar too seem to be imbedded in individual words. For example, when a speaker produces an isolated word, she will activate associated grammatical features, such as the grammatical gender of a noun (e.g., a French speaker who hears '*chat'* will automatically activate the feature 'masculine', Wang, Chen, Schiller, 2019). Thus, cognitive processing of meanings spreads beyond predefined limits of a word into co-occurring words and grammar; and processing of isolated words appears to include more than the word's own dictionary definition, such as its grammatical features. Broadly, a word's meaning is inextricable from its context, and the context too is inextricable from the word's meaning. It is likely that the development of metalinguistic capacities that entrenches this perception that words are self-standing isolable linguistic items: both pre-literate children and illiterate adults, who can use a word correctly and understand its meaning, are not as nimble at isolating it from certain contexts (e.g., inversing the words in the utterance 'look up' to 'up look'; Havron & Arnon, 2017a; Havron & Arnon, 2017b). 'Cat' is therefore just one dimension of *chat*, one that overlooks the context it appears in (*le, un, ronronne, joue*) and its associated grammatical features (e.g., masculine noun, see Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005 for an in-depth analysis of the interplay between meaning, grammar, and context).

Though supervision can be designed to reflect cognitive processing as closely as possible, the aim of supervision is nonetheless to guide learning in the right direction. Thus, supervision may switch focus to the associations between words and grammar, for example, by presenting the two together (*le* + *chat* = cat), but this focus too will inadvertently neglect other dimensions of *chat*, such as its co-occurrence with certain verbs (*ronronne*, *joue*). In other words, no matter how closely supervision mimics cognitive processing of language, it will likely still alter how language is processed in the mind. Nevertheless, comparisons between different kinds of supervision shed light on how the scope of evidence affects processing. For instance, when supervised learning begins either with isolated words or with words in their grammatical context (e.g., *chat* vs. *le chat*), it can determine how well the grammatical element is acquired: learning words in isolation impedes (at least in the scope of a lab visit) future acquisition of associated grammar (Arnon & Ramscar, 2012). A proposed explanation for this observed effect is that of 'blocking' from classic conditioning studies (Rescorla,1968). Once a learner has acquired an association between a word (*chat*) and its referent (cat), other associations (e.g., the masculine article *le* and cat) are considered redundant and are not learned. In other words, if a word provides enough evidence on its own as to its meaning, a learner need not look elsewhere for further evidence. The same may hold for supervision in general. When supervision provides enough evidence for a learner to make a decision, then a learner may no longer need to explore further evidence, and discounting other evidence may affect future processing and acquisition. However, few studies have investigated how processing compares between supervised and unsupervised learning, beyond simply evaluating whether supervision can bolster learning.

The evidence used to make a decision may thus affect processing, for example, by focusing it on certain elements. Yet, the *way* in which evidence is presented can, in turn, affect how, and if at all, the evidence is processed. The reliability of a person conveying evidence, for example, has been shown to affect learning (Scofield & Behrend, 2008; Koenig & Harris, 2005). Children are less likely to acquire novel words (e.g., the word 'blicket' for a rare contraption) from speaker who had called a common item by a different word, such as calling a ball 'keys' (unreliable speaker), than a speaker who had labeled the item correctly, ball as 'ball' (reliable

speaker; Scofield & Behrend, 2008). At first glance, reliability appears to directly affect what is considered as evidence: a reliable speaker communicates evidence and an unreliable speaker does not. We could say that the way evidence is presented is a sort of second-order evidence: evidence about the evidence itself. Counterintuitively, even unreliable speakers can in certain experimental situations sway decisions (e.g., Heyman, Sritanyaratana, Vanderbilt, 2013; Jaswal, Croft, Setra, & Cole, 2010). For example, children will continue to believe an agent (e.g., the big bad wolf), even if the agent consistently gives wrong advice (e.g., the big bad wolf tells them that a sticker is hidden under a box, when it is not). In contrast, if children are just given a symbolic prompt (e.g., an arrow pointing in the wrong direction), they will stop believing it. One explanation for this behaviour is 'natural pedagogy' (Csibra & Gergely, 2009): the theory that humans (or any social agents) communicate to transmit knowledge. Therefore, evidence communicated by an intentional agent ought not be discounted, at least not completely and not right away. Children have been shown to go as far as willingly abandoning personally accumulated evidence in favour of opposing testimonial evidence (Jaswal, Pérez-Edgar, Kondrad, Palmquist, Cole & Cole, 2014). If a child and an adult both watch a cat go into the blue house, but the adult says the cat went into the yellow house, when asked into which house the cat went, the child will have a tendency to say it was the yellow house, in accordance with the adult's incorrect report. Therefore, the way in which evidence is presented can aid processing by identifying good sources of evidence (e.g., from a reliable person and it can also impede processing when it is used as a heuristic (e.g., people communicate knowledge, so this person communicates knowledge). Regardless of the kind of second-order evidence the medium provides, it appears to be a considered element during processing. Though much of the research on the interaction between testimonial evidence and subsequent decisions has been centered on young children, the general findings may not be limited to a certain age. Young children may be more likely to accept opposing (and incorrect) testimonial evidence, but a more subtle distinction between equally reliable unsupervised and supervised learning may reveal that adults and children both consider the medium as second-order evidence. Here we propose that supervision, when it consists in providing a learner with explicit guidance, may serve as second-order evidence, over and above, any information it provides. Supervision may thus modify processing of a first-order evidence set: we may choose to put more or less weight on first-order evidence because of how it had been presented.

At first glance, supervision appears to bolster or speed up acquisition. However, there may be much more happening on a cognitive level than meets the eye: supervision may not be entirely beneficial for a learner. Understanding the effects supervision may have on processing is important to elucidate why a mind that is so adept at learning in an unsupervised manner turns to supervision.

The goal of this study is to dissect the effects of supervision on cognitive processing of language. We focus on one form of supervision that is prevalent in language learning classrooms: testimony. We pit learning via testimony against unsupervised learning, to investigate three main factors: rate and degree of acquisition of a target item, degree of acquisition of associated items, and metacognition. We aimed at comparing two ecologically valid language learning methods: one that reflects how a native language is learned early in life (unsupervised) and one that reflects how second languages are often learned in classrooms (supervised via testimony). Moreover, in order to draw substantial conclusions about supervision, we test three groups of participants who vary along two dimensions: age and literacy. Any number of learner-internal factors can influence acquisition, and interact with the external language learning method. We focus on two learner-internal factors: the capacity to reflect on one's decision (metacognition) and the capacity to reflect on language more

specifically (metalinguistic awareness; e.g., isolating words in a sentence or isolating sounds within a word). Metacognition has been shown to develop through childhood. While the capacity to monitor performance is present from infancy (Goupil, Romand-Monnier, & Koudier, 2016), children tend to have a bias for over-confidence (Salles, Ais, Semelman, Sigman, & Calero, 2016; Finn & Metcalfe, 2014; van Loon, de Bruin, Leppink, & Roebers, 2017). Notably, children's over-confidence has been shown to depress reviewing of incorrect items (e.g., if a child is confident in an answer, she will not go back and review it, Destan & Roebers, 2015). Acting less on one's decisions (e.g., not correcting incorrect responses) may distort how language is learned at different ages, and indirectly affect the difference in how language is acquired in supervised and unsupervised learning tasks. On the other hand, metalinguistic awareness develops when a child learns to read (e.g., Brunswick, Martin, & Rippon, 2012), and remains largely immature in illiterate adults (for a review: Morais & Kolinsky, 2002). Accordingly, the capacity to reflect on linguistic elements may change, broadly, how language is learned, and whether a learner can take advantage of evidence provided in supervised and unsupervised learning tasks. We therefore tested adults (n = 48, mean age: 24;3 years; range: 18;1-34;11 years) and two groups of 5-6 year old children (mean age: 6;3 years, total: n = 49), pre-literate (n = 22, mean age: mean 6;1 years, range: 5;8-6;11 years) and literate (n = 27, mean age: 6;5 years; range: 5;10-6;11 years).[20] The 5-6-year-old age range was chosen because this is the age at which children begin learning to read, and as such children of the same age can either be pre-literate or literate. Pre-literate and literate groups were matched as closely as possible for age. Parents rated their child's reading capacity on a 6 point scale from (1) 'Has not started learning to read' to (6) 'Can read a whole book on her/his own'. Testing three groups, who vary along age and literacy dimensions, allowed us to better isolate the fundamental effect of supervision and to identify any interactions with other cognitive capacities.

We designed an artificial language learning task that simulates how languages are often learned. The goal of language learning classrooms is to provide learners with keys to understanding and communicating 'in the wild'. In other words, an entire language cannot be acquired via supervision, and at some points a learner will have to continue learning in an unsupervised manner. Therefore, only the first phase of the task differed between conditions: participants in the Supervised Condition heard translations of vocabulary items in the artificial language (e.g., *Pirdale* means 'book'), while those in the Unsupervised Condition did a control exercise (e.g., a picture-naming or picture-identification task in their native language; Teaching Phase, Fig. 1A). Supervision thus involved testimonial evidence, provided by an alien avatar, of the meaning of each noun prior to an unsupervised learning phase common to both groups (Learning Phase, Fig. 1B). The Learning phase was an adapted version of the cross-situational learning paradigm (Yu & Smith, 2007; Trueswell, Medina, Hafri, & Gleitman, 2013), during which participants heard a sentence in the artificial language (e.g., *Dol ka pirdale*) and saw two images (e.g., Fig. 1B(ii)). They chose which of two images corresponded to the sentence. Each sentence was composed of a carrier phrase (*dol*), a grammatical element (*ko* or *ka*) and a noun (e.g., *pirdale, doripe, jigoulate*). The grammatical element corresponded to the animacy of the noun (e.g., *ko* before all animates, *ka* before all inanimates). Participants in the Supervised Condition had already been exposed to the translation of the vocabulary but not the carrier phrase or grammar, while participants in the Unsupervised Condition heard the language for the first time. The two-phase artificial language learning design (Teaching then Learning) allowed us to later measure acquisition of the nouns (learned with supervision or without) and the associated grammatical items (i.e., *ko* and *ka*; learned without supervision) in the Test phase

[20] After the replication, we will have 96 children in total (48 pre-literate and 48 literate).

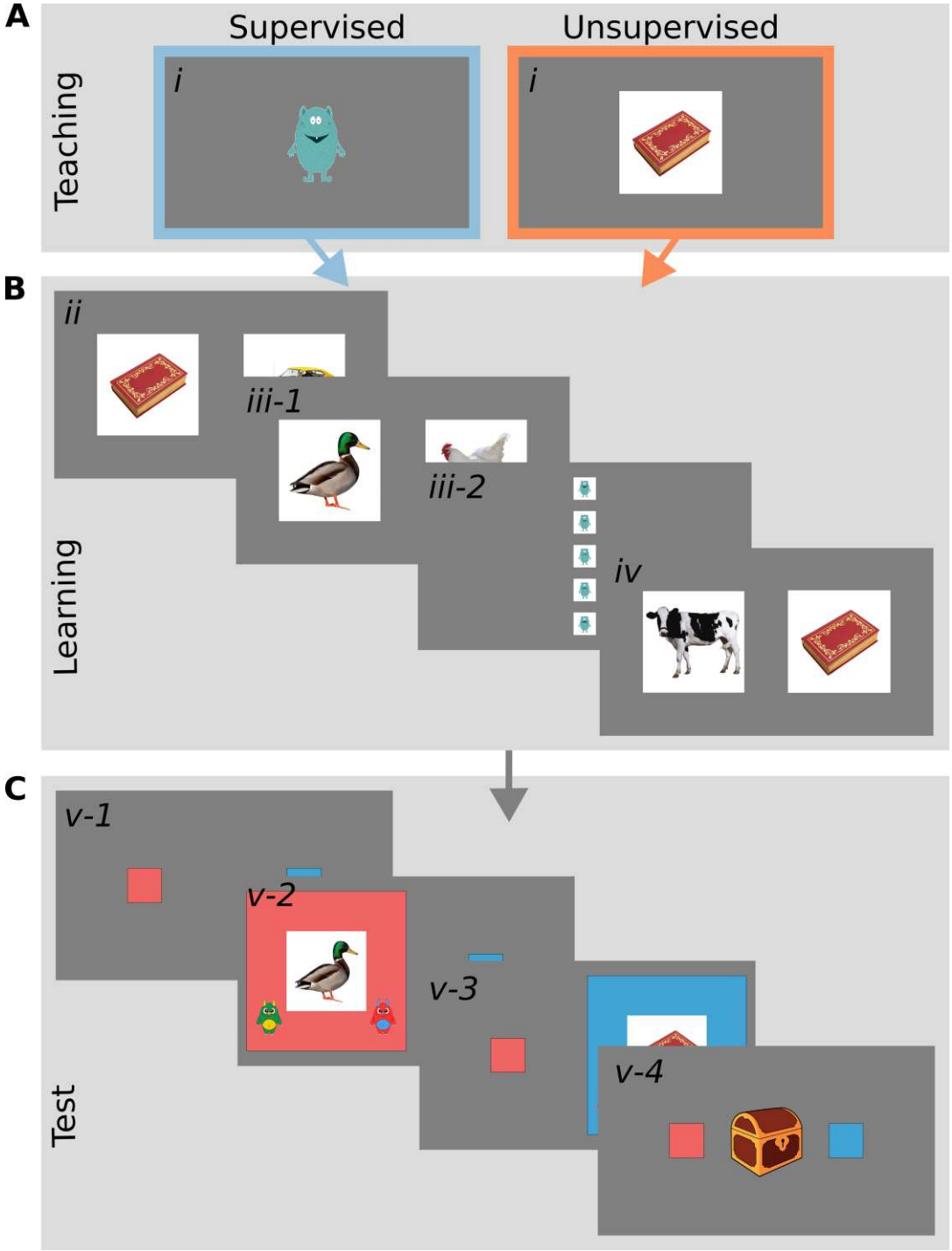


Fig. 1. Experimental protocol. (A) Teaching Phase. Participants in the Supervised Condition in blue hear translations of artificial language vocabulary (e.g., *Pirdale* means book). (i) *Example trial.* Participants see an alien avatar, who tells them a translation. They have to repeat the translation. Participants in the Unsupervised Condition in orange see images of items with no audio (e.g., a book). (i) *Example trial.* Participants see an image of a book, then adults have to identify the item and children have to name the picture out loud in their native language. **(B) Learning Phase.** Participants do a cross-situational learning task, in which they see two images and hear a prompting sentence in the artificial language (e.g., *Dol ka pirdale*). They have to choose the image that corresponds to the sentence. (ii-iv) *Example trials*. (iii-1–2) *Example confidence trial.* A cross-situational trial is followed by a confidence scale rating. **(B) Test Phase.** Participants do a nested two-alternative forced choice task evaluating performance and confidence. (v) *Example trial.* (v-1) Participants see two squares. (v-2) A square opens, revealing an image. Two monsters each say a sentence (one correct and one incorrect), and participants have to choose who said it best. (v-3). The other square opens with another question. (v-4). Participants choose the answer for which they were most confident, and put it in the treasure chest.

(Fig. 1C). Acquisition was evaluated using a two-alternative forced choice (2AFC) paradigm, in which participants saw an image and heard two sentences, one correct and one incorrect (e.g., Fig. 1C(v-2)). The sentences differed in just one critical element: (a) vocabulary (different content words): *Dol ko **jigoulate*** vs. *Dol ko **pirdale***; (b) grammar (different grammatical elements): *Dol **ko** bamoule* vs. *Dol **ka** bamoule*; (c) generalization (different grammatical elements, novel content word and image): *Dol **ko** nuve* vs. *Dol **ka** nuve.* Participants had to choose the sentence that best corresponded to the image. Grammar and generalization questions allowed us to measure both the acquisition of the associations between grammatical elements and vocabulary ((b) grammar sentences) and the capacity to extend the grammatical rule to novel, hitherto unheard, vocabulary ((c) generalization sentences). Participants were asked to report how confident they were in their decisions during the Learning and Test phases (Fig. 1B(iii-2) and Fig. 1C(v-4), respectively).

We compared acquisition of the mapping between sentences and corresponding images and metacognition across the two conditions (Supervised and Unsupervised) and the two groups of children (Pre-literate and Literate). We interpreted the results from the two age groups separately, because children completed a shorter version of the experimental design than adults (see Materials and Methods). We hypothesised that supervision would increase and/or speed up acquisition of nouns, while impeding acquisition of grammatical elements, and that it would increase metacognitive ability and bias. In other words, supervision would aid acquisition of a target item (noun) while impairing acquisition of associated items (grammar and generalization of the grammatical rule), as has been suggested in previous literature (Arnon & Ramscar, 2012; Paul & Grüter, 2016). We thought supervision might increase metacognitive ability (the capacity to monitor performance) and bias (the general feeling of confidence), because it may provide learners with information beyond the meaning of the word itself. First, it offers learners a direct correspondence between the artificial language and something they know, namely their native language. Intuitively, this structural correspondence may allow for better monitoring of performance (higher metacognitive ability) because learners have something concrete with which to compare performance. Second, the medium of testimony itself may serve as second-order evidence. Being explicitly told a meaning may increase the weight of evidence in favour of a decision, and thus increase baseline confidence (confidence bias). Alternatively, supervision may not have an effect on acquisition or metacognition. Most studies that compare supervised and unsupervised learning do so because learners were unable to acquire a target item in an unsupervised fashion (within a lab visit); in contrast, we expect both adults and children in the Unsupervised condition to learn the associations between sentences and images, because the paradigm we use has been shown to elicit robust word learning across development (from infancy to adulthood, Smith & Yu, 2008; Suanda, Mugwanya & Namy, 2014; Yu & Smith, 2007). Though we present participants with sentences, it ought to still be possible to learn the sentence-image mappings. Thus, supervision may not bolster acquisition. It is also possible that supervision may not have an effect on metacognition. Supervision may add to the weight of evidence for the first-order decision (i.e., the link between a sentence and an image), but not for second-order decisions (i.e., metacognition).

We used the behavioural measure of explicit choice to assess performance and metacognition. A difference in cognitive processing between Supervised and Unsupervised Conditions would be evidenced by a difference in the proportion of correct choices (performance) or in the confidence level choice (metacognition). We ran additional analyses on gaze, using two eye-tracking measures, one to determine performance: direction of gaze (gaze focus); and one to determine confidence: number of fixation switches (gaze distribution; Krajbich, Armel, & Rangel, 2010; Cavanagh, Wiecki, Kochar, & Frank, 2014; Folke, Jacobsen,

Fleming, & De Martino, 2017). Our results reveal that supervision does not procure an across-the-board benefit for performance, but rather uniformly sows an overconfidence bias.

4.1.1 Results

A. Learning phase: Performance

We first examine the rate and degree of artificial language acquisition during the learning phase.

Explicit choice data.

As a measure of rate and degree of acquisition, we examined the proportion of correct explicit choices (versus incorrect choices) for each age group across the four Learning phase blocks. Means for each age group per block are presented in Fig. 2(i). On average, the proportion of correct choices increased across blocks for both adults [main effect of Block: $\beta = 0.72$, SE = 0.05; model comparison: $\chi^2(1) = 59.61$, $p < 0.001$] and children [main effect of Block: $\beta = 0.24$, SE = 0.05; model comparison: $\chi^2(1) = 21.65$, $p < 0.001$]. This result indicates that both age groups were able to learn, at least some, sentence-image pairs, and that broadly this task was accessible for both ages.

Adults.

Adults in the Supervised Condition made more correct explicit choices than adults in the Unsupervised Condition [main effect of Condition: $\beta = 1.79$, SE = 0.25; model comparison: $\chi^2(1) = 34.77$, $p < 0.001$]. This effect, however, is reducible to a difference in the amount of exposure: the Supervised group had been exposed to the language for two blocks (Teaching phase) prior to the Unsupervised group. When we matched exposure (the first and second blocks of the Learning phase for the Supervised group and the third and fourth blocks of the Learning phase for Unsupervised group), participants' performance in both conditions did not differ significantly [main effect of Condition: $\beta = 0.67$, SE = 0.51; model comparison: $\chi^2(1) = 1.72$, $p = 0.19$]. This result indicates that supervision did not boost performance on our task (i.e., matching sentences to images). This finding is consistent with the possibility that adults in the Supervised Condition may have outperformed those in the Unsupervised Condition on a different task (e.g., producing words or providing translations), or vice versa.

Children.

Children made a comparable amount of correct choices in the two learning conditions [no main effect of Condition: $\beta = 0.1$, SE = 0.14; model comparison: $\chi^2(1) = 0.49$, $p = 0.48$]. Literate and pre-literate children, too, had similar proportions of correct choices [no main effect of Literacy: $\beta = -0.05$, SE = 0.14; model comparison: $\chi^2(1) = 0.15$, $p = 0.7$]. However, learning condition affected a child inversely depending on her level of literacy (Fig. 3A): literate children made more correct choices when they were in the Supervised Condition, while the opposite was true for pre-literate children, who made more correct choices when they were in the Unsupervised Condition [interaction between Condition and Literacy: $\beta = 0.67$, SE = 0.28; model comparison: $\chi^2(1) = 5.47$, $p = 0.019$]. To pinpoint the driving factor behind the cross-over interaction in the absence of main effects, we investigated the impact of learning condition on literate and pre-literate children separately. Literate children made significantly more correct choices in the Supervised Condition than in the Unsupervised Condition [main effect of Condition: $\beta = 0.43$, SE = 0.19; model comparison: $\chi^2(1) = 4.49$, $p = 0.03$]; while pre-literate children did not make significantly more correct choices in the Unsupervised Condition [main effect of Condition: $\beta = -0.26$, SE = 0.17; model comparison: $\chi^2(1) = 2.2$, $p = 0.14$]. Thus, after supervised teaching, literate children, similarly to adults, had higher performance on our task. Pre-literate children, however, did not seem to be able to take advantage of the evidence that supervision had

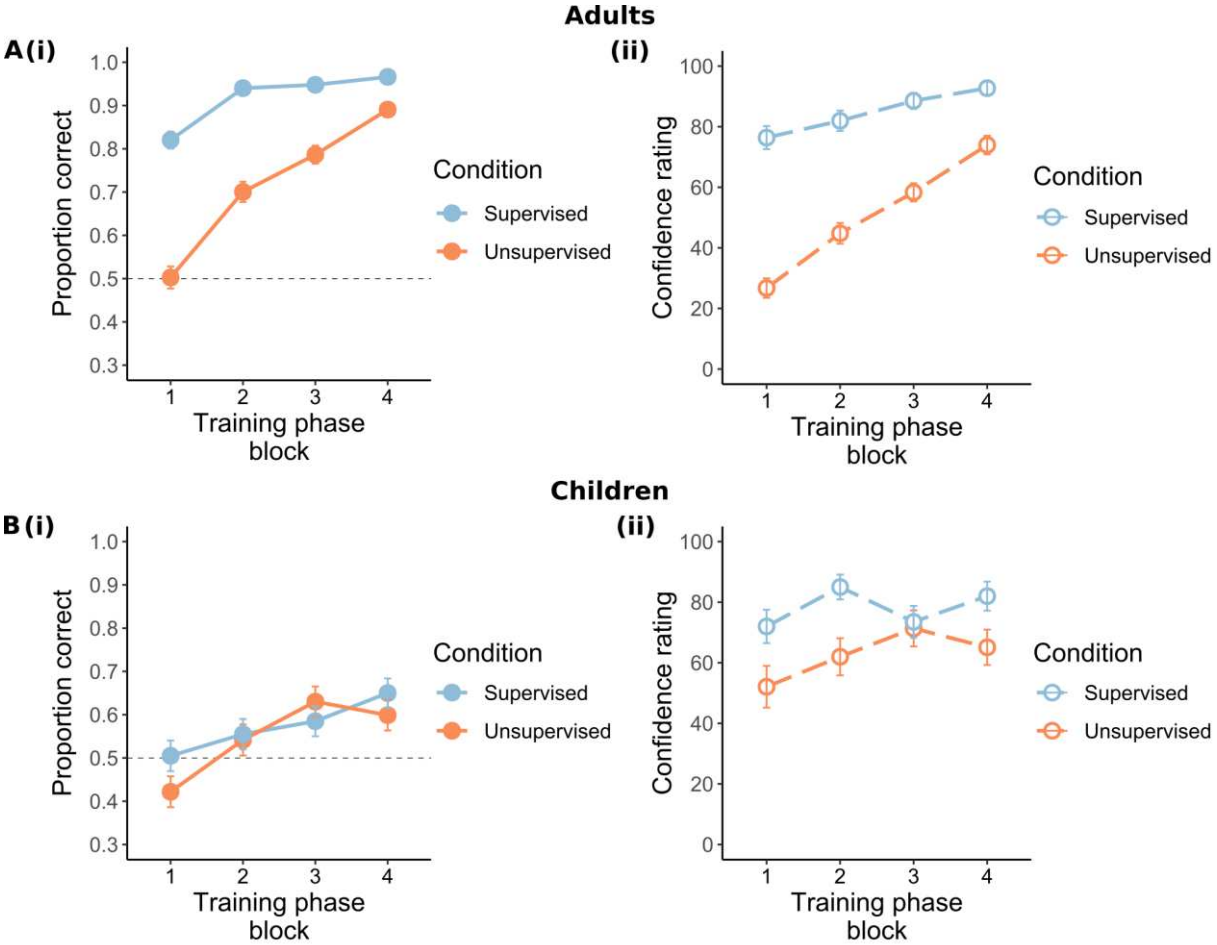


Fig. 2. Learning phase explicit choice results. (i) Performance. Mean proportion of correct image choices per block, for the Supervised Condition in blue and the Unsupervised Condition in orange. Error bars represent SEM (95% confidence intervals). Dotted black line indicates chance. **(ii) Confidence.** Mean confidence rating on a scale from 0 (I guessed) to 100 (I'm sure), for the Supervised Condition in blue and the Unsupervised Condition in orange. Error bars represent SEM (95% confidence intervals). **(A) Adults.** Performance increased across blocks ($p < 0.001$), and participants in the Supervised Condition had an overall higher performance ($p < 0.001$). Adults' performance was not significantly different when Blocks 1-2 for the Supervised Condition and Blocks 3-4 for the Unsupervised Condition were compared ('performance matched'). Participants in the Supervised Condition were more confident than those in the Unsupervised Condition, even when performance did not differ significantly ($p < 0.001$). **(B) Children.** Performance increased across block ($p < 0.001$). There were not significant differences in performance between conditions. Participants in the Supervised Condition were more confident than those in the Unsupervised Condition despite comparable levels of performance ($p < 0.001$).

provided: the direct translations of the artificial language words did not boost performance. They may have had more trouble interpreting the correspondences between isolated words in the artificial language and isolated words in their native language, perhaps because this task involves a certain level of metalinguistic reflection. This result suggests that what can be used as evidence will depend on cognitive capacities.

Just as we had for adults, we examined the proportion of correct choices when exposure to the language was matched (Blocks 1-2 for the Supervised Condition, and 3-4 for the Unsupervised Condition). Interestingly, children made significantly more correct choices after two blocks of unsupervised learning (Unsupervised Condition, Learning blocks 3 and 4) than two blocks of supervised learning (Supervised Condition, Learning blocks 1 and 2) [main effect

of Condition: $\beta = -0.4$, SE = 0.17; model comparison: $\chi^2(1) = 5.29$, $p = 0.02$]. Literacy did not significantly mediate this effect [no interaction between Condition and Literacy: $\beta = 0.56$, SE = 0.34; model comparison: $\chi^2(1) = 2.72$, $p = 0.1$]. Thus, children appeared to benefit more from two blocks of unsupervised learning than from two blocks of supervised learning, in the context of our task.

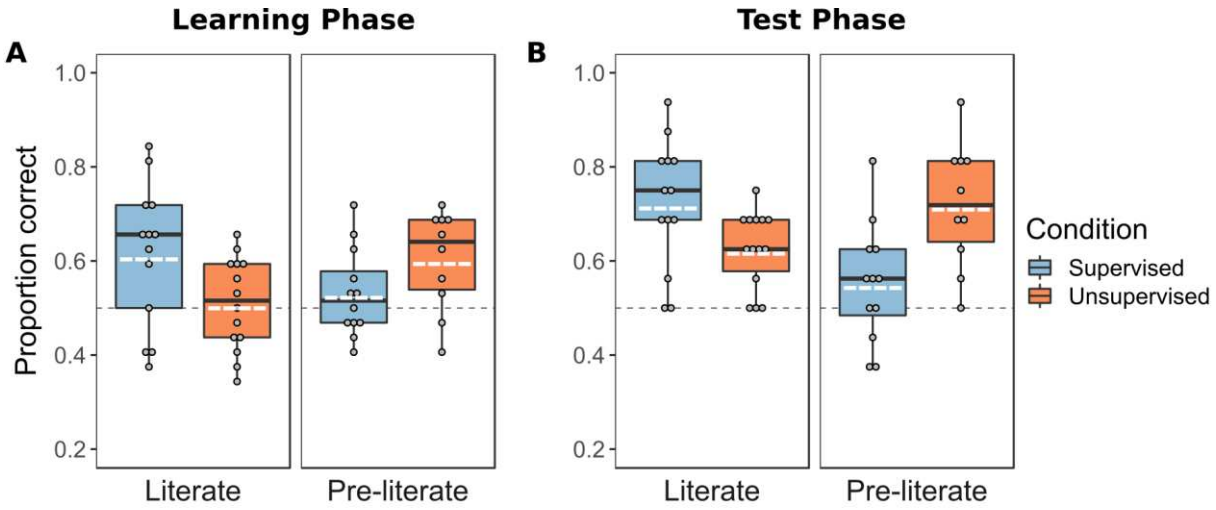


Fig. 3. Results by literacy level. Proportion of correct answers per literacy level. Literate group on the right, pre-literate group on the left. Supervised Condition in blue, Unsupervised Condition in orange. Dots indicate participants. Dashed white line indicates mean. Upper and lower regions of the box indicate the first and third quartiles (25^{th} to 75^{th} percentiles). The upper whisker represents the third quartile up to the 1.5 interquartile smallest value, while the lower whisker the 1.5 interquartile smallest value to the first quartile. **(A) Learning Phase.** The learning condition interacted with literacy level: supervision boosted acquisition in literate children but impeded acquisition in pre-literate children ($p = 0.03$). **(B) Test Phase.** The learning condition continued to interact with literacy level ($p < 0.001$).

Summary.
Together, our results indicate that supervision does not necessarily procure a benefit on performance: when exposure to the language was matched, the learning methods were comparable for adults, and unsupervised learning was more advantageous for children. They also suggest that some forms of supervision may require specific cognitive capacities, not necessarily available to all at any point in development.

Gaze data.
Next we examined the proportion of looks to the target over the course of a whole trial (gaze focus), as an index of acquisition of sentence-image mappings. Gaze focus has been shown to reflect performance (e.g., Krajbich, Armel, & Rangel, 2010), and is a measure commonly used to evaluate the performance of young children (Golinkoff, Hirsh-Pasek, Cauley, & Gordon, 1987).

The proportion of looks to the target image increased across blocks, for adults [$\beta = 0.054$, SE = 0.003, $p < 0.001$], and children [$\beta = 0.016$, SE = 0.006, $p = 0.004$, Fig. 4(i)]. The gaze results thus corroborate explicit choice results, reflecting an increase in performance across blocks.

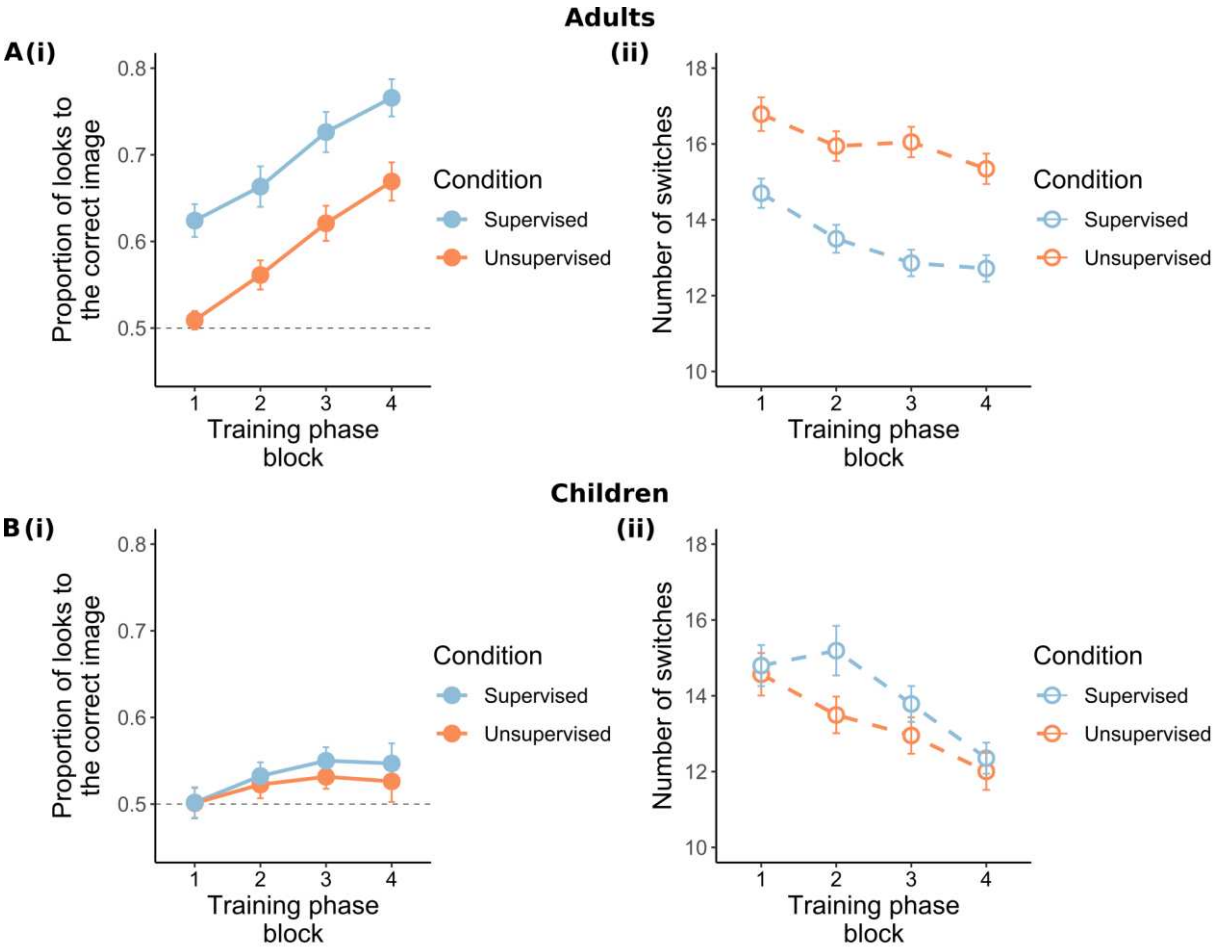


Fig. 4. Learning phase gaze results. (i) Performance. Mean proportion of looks to the target image per block, for the Supervised Condition in blue and the Unsupervised Condition in orange. Dotted black line indicates chance. **(ii) Confidence.** Mean number of gaze switches between the target and foil per block, for the Supervised Condition in blue and the Unsupervised Condition in orange (fewer switches index high confidence). **(A) Adults.** Performance increased across blocks ($p < 0.001$), and participants in the Supervised Condition had an overall higher performance ($p < 0.001$). Performance did not differ significantly between conditions on Blocks 1-2 for the Supervised Condition and Blocks 3-4 for the Unsupervised Condition. The number of gaze switches decreased across blocks, indicating an increase in confidence during the Learning phase ($p < 0.001$). Participants in the Supervised Condition gaze-switched less than those in the Unsupervised Condition, even when performance did not differ significantly ($p = 0.017$). **(B) Children.** Performance increased across blocks ($p = 0.004$), and did not differ between Supervised and Unsupervised Conditions. The number of gaze switches decreased across blocks ($p < 0.001$), but did not differ significantly between the two learning conditions.

Adults.

On average, adults in the Supervised Condition looked more to the target image than those in the Unsupervised Condition [main effect of Condition: $\beta = 0.105$, SE = 0.003, $p < 0.001$]. However, when the amount of exposure to the artificial language was matched (Learning blocks 1-2 for the Supervised Condition, 3-4 for the Unsupervised Condition), there was no longer a significant difference in the proportion of looks to the target image between learning conditions [no main effect of Condition: $\beta = -0.001$, SE = 0.026, $p = 0.98$]. For adults, gaze focus data, thus, provide convergent evidence that supervision procures little benefit for performance (in our task).

Children.

Overall, children in the Supervised Condition did not look more to the target image than those in the Unsupervised Condition [no main effect of Condition: $\beta = 0.006$, SE = 0.016, $p = 0.69$].

However, unlike the explicit choice results, there was not a significant interaction between Literacy and Condition [$\beta = 0.04$, SE = 0.032, $p = 0.22$]. When the amount of exposure was matched (Learning blocks 1-2 for the Supervised Condition, 3-4 for the Unsupervised Condition), there was likewise no significant difference in the proportion of looks to target between learning conditions [no main effect of Condition: $\beta = -0.03$, SE = 0.02, $p = 0.19$]. Though children's gaze focus data did not appear to converge on all points with explicit choice data, globally, both measures reveal convergent evidence that supervision does not provide an advantage for performance.

Summary.
Both explicit choice and gaze focus data, across age groups, suggest that supervised learning of one linguistic element (vocabulary meanings) does not procure an advantage for performance on a broader language-level task (mapping sentences to their corresponding meanings). This was so even though knowledge of vocabulary meanings, the items taught explicitly in the Supervised Condition (e.g., *pirdale* means 'book'), was alone sufficient for a learner to succeed at the unsupervised language-level task (e.g., *Dol ka pirdale. Pirdale* means 'book', so the corresponding image is the book image).

B. Learning phase: Metacognition

Next, we investigate metacognition (metacognitive ability and bias) during artificial language acquisition.

Explicit choice data: Metacognitive ability.
During the Learning phase, certain trials included a confidence judgement: after choosing one of the two images, participants were asked to rate how confident they were in the decision they just made on a five-point scale from 'I'm sure' to 'I guessed' (Fig. 1B(iii-2), means displayed in Fig. 2A(ii) and Fig. 2B(ii)). We evaluated two facets of confidence: metacognitive ability and metacognitive bias. Metacognitive ability was indexed by a difference in confidence ratings for correct and incorrect answers: if participants are able to monitor performance, then they ought to provide higher confidence ratings for correct responses than incorrect responses (Fig. 5(i)).

Adults.
Adults, on the whole, provided higher confidence ratings for correct answers than incorrect answers [main effect of Confidence: $\beta = 0.572$, SE = 0.12; model comparison: $\chi^2(1) = 23.07$, $p < 0.001$]. The difference in confidence ratings for correct and incorrect choices was greater for adults in the Supervised Condition than in the Unsupervised Condition [interaction between Confidence and Condition: $\beta = 1.193$, SE = 0.25; model comparison: $\chi^2(1) = 25.52$, $p < 0.001$]. Even when exposure to the language (and therefore also mean performance) was matched, adults in the Supervised Condition had a greater difference in confidence ratings than adults in the Unsupervised Condition [interaction between Confidence and Condition: $\beta = 1.455$, SE = 0.35; model comparison: $\chi^2(1) = 21.62$, $p < 0.001$, Fig. 5A(i)]. These results indicate that for adults, supervision may boost metacognitive ability: the capacity to monitor performance.

Children.
Children did not provide higher confidence ratings for correct responses than incorrect responses [main effect of Confidence: $\beta = 0.03$, SE = 0.08; model comparison: $\chi^2(1) = 0.14$, $p = 0.71$, Fig. 5B(i)]. This effect was not moderated by learning condition [no interaction between Confidence and Condition: $\beta = -0.07$, SE = 0.16; model comparison: $\chi^2(1) = 0.18$, p =

0.67]. However, unlike adults, who were almost at ceiling toward the end of the Learning phase, children were just above chance (56% correct answers all blocks combined, and 61% correct answers in the fourth block of the Learning phase, Fig. 2(ii)). When accuracy is very low, even adults have difficulties tracking their performance (e.g., Schwiedrzik, Singer, & Melloni, 2011; Kruger & Dunning, 1999). Therefore, our results are consistent with two broad explanations: either children understand the link between their choices and the confidence scale, but do not have metacognitive ability (perhaps because of near-chance performance), or children do not understand the link between their choices and the confidence scale, and are using ratings at random or following a different rule. To disentangle the two possible explanations, we investigated whether confidence ratings during the Learning phase were correlated with metacognitive ability in the Test phase in a post hoc analysis. Mean confidence during the Learning phase reliably predicted metacognitive ability in the Test phase [$F(1, 35) = 4.901$, $p = 0.03$, $R^2 = 0.1$]. In other words, children who were on average more confident during the Learning phase, were also better at monitoring their performance in the Test phase. Importantly, this result suggests that children understood and were using the confidence scale meaningfully.

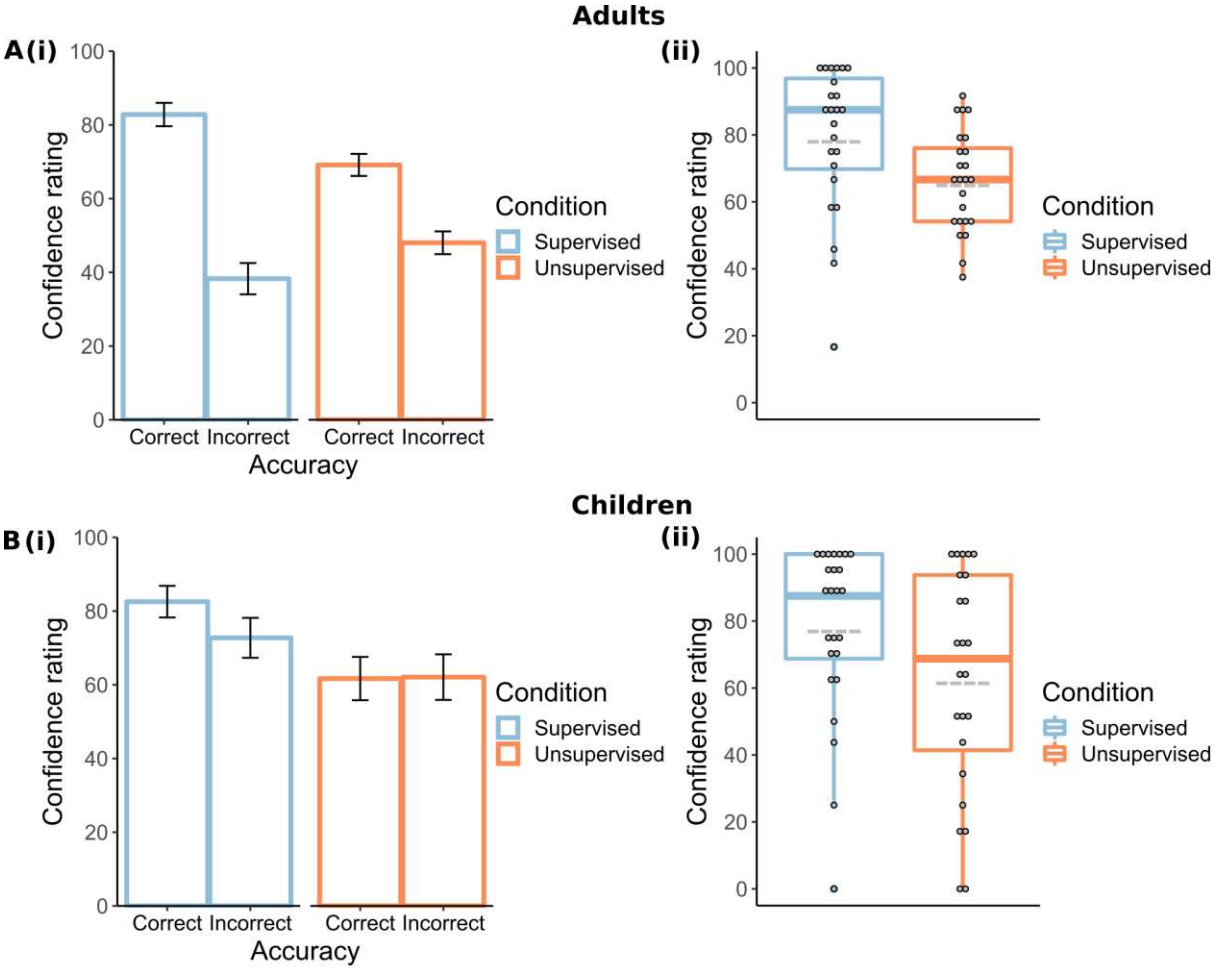


Fig. 5. Learning phase metacognition. Graphs show metacognition when performance is matched. Supervised Condition in blue, Unsupervised Condition in orange. **(i) Metacognitive ability.** Mean confidence rating for correct and incorrect answers. Error bars represent SEM. **(ii) Metacognitive bias.** Mean confidence rating. **(A) Adults.** Adults gave higher confidence ratings for correct answers than incorrect answers (metacognitive ability, $p < 0.001$), and those in the Supervised Condition had a higher metacognitive ability ($p < 0.001$). Adults in the Supervised Condition felt more confident on average than those in the Unsupervised Condition ($p < 0.001$). **(B) Children.** Children did not give significantly higher confidence ratings for correct answers than incorrect answers (low metacognitive ability). Children in the Supervised Condition felt more confident on average than those in the Unsupervised Condition ($p < 0.001$).

Thus, children were likely not able to accurately track performance during the Learning phase, regardless of learning condition, because performance was just above chance.

Explicit choice data: Metacognitive bias.

Metacognitive bias was indexed by a baseline difference in mean confidence ratings on a group level, when mean performance is matched. In other words, when learners in both conditions have the same level of performance (e.g., 75% correct choices), there may be a difference in baseline confidence (metacognitive bias). We thus examined mean confidence ratings for the subset of Learning phase blocks where performance was matched. We further investigated confidence ratings when exposure was matched. If a metacognitive bias is observed when exposure is matched, it cannot be reduced to a simple difference in the quantity of exposure (and therefore of evidence in favour of a decision). For adults, performance was matched on Blocks 1 and 2 for the Supervised Condition and 3 and 4 for the Unsupervised Condition, which also corresponded to the blocks with equal exposure to the language. For children, performance was matched across all four Learning blocks. Mean confidence ratings are displayed in Fig. 5(ii).

Adults.

Adults in the Supervised Condition were more confident than those in the Unsupervised Condition ($p < 0.001$, Fig. 5A(ii)). These results reveal a metacognitive bias for adults in the Supervised Condition: they feel more confident in their choices.

Children.

We then examined the mean confidence ratings for children when performance was equal (all four blocks) and when exposure to the language was equal (Blocks 1 and 2 for the Supervised Condition and 3 and 4 for the Unsupervised Condition). Children in the Supervised Condition were more confident than those in the Unsupervised Condition when performance was matched ($p < 0.001$, Fig. 5A(ii)), and even when exposure was matched ($p < 0.001$). Interestingly, when exposure to the language was equal, children in the Unsupervised Condition had significantly higher performance (explicit choices) than children in the Supervised Condition; yet, those in the Supervised Condition felt more confident. Moreover, this effect held for both literate ($p = 0.024$) and pre-literate children ($p < 0.001$). Thus, children, like their adult peers, feel more confident learning in a supervised manner than an unsupervised manner.

Summary.

Together, these results suggest that supervision may have an across-the-board effect on metacognition, rather than performance. Both children and adults felt more confident in their choices when they had learned with supervision. Additionally, for adults, supervision augmented metacognitive ability, perhaps because it offered adults something concrete with which to compare performance (namely, their native language).

Gaze data: Metacognition (metacognitive ability or bias).

As a marker of metacognition, we examined the number of gaze switches between a target and foil. Gaze distribution has been shown to reflect confidence judgements: many switches reflect low confidence, and vice versa (Folke et al, 2017). This is likely because low confidence fosters higher rates of exploration (Boldt, Blundell, & De Martino, 2019). The mean number of switches per learning condition, across blocks is presented in Fig. 4(ii). Overall, the number of switches decreased during the Learning phase, for adults [main effect of block: $\beta = -0.53$, $SE = 0.16$; model comparison: $\chi^2(1) = 10.23$, $p = 0.001$, Fig. 4A(ii)], and for children [main effect of block: $\beta = -0.86$, $SE = 0.15$; model comparison: $\chi^2(1) = 31.35$, $p < 0.001$, Fig. 4B(ii)].

Adults.
In general, adults in the Supervised Condition gaze-switched less than those in the Unsupervised Condition [main effect of Condition: $\beta = -2.893$, SE = 1.27; model comparison: $\chi^2(1) = 4.89$, $p = 0.027$]. These findings reflect performance, and are thus consistent with two interpretations: gaze-switches are another index of performance (i.e., the more one knows, the less one gaze-switches) or gaze-switches are an index of metacognition (i.e., when one knows that one knows more, one gaze-switches less). To disentangle the two, we computed gaze switches when performance was matched (in adults, Blocks 1-2 for the Supervised Condition and 3-4 for the Unsupervised Condition). When performance was equal, adults in the Supervised Condition still made fewer gaze switches than those in the Unsupervised Condition [main effect of Condition: $\beta = -3.69$, SE = 1.51; model comparison: $\chi^2(1) = 5.72$, $p = 0.017$, Fig. 6A]. Thus, fewer gaze-switches appear to indeed reflect higher confidence. Both metrics of metacognition (the number of switches and explicit confidence ratings) reveal a metacognitive bias: supervision increases confidence. Moreover, this analysis confirmed that the confidence bias was present on the whole set of Learning phase trials, and not just the ones that were accompanied by a confidence judgement.

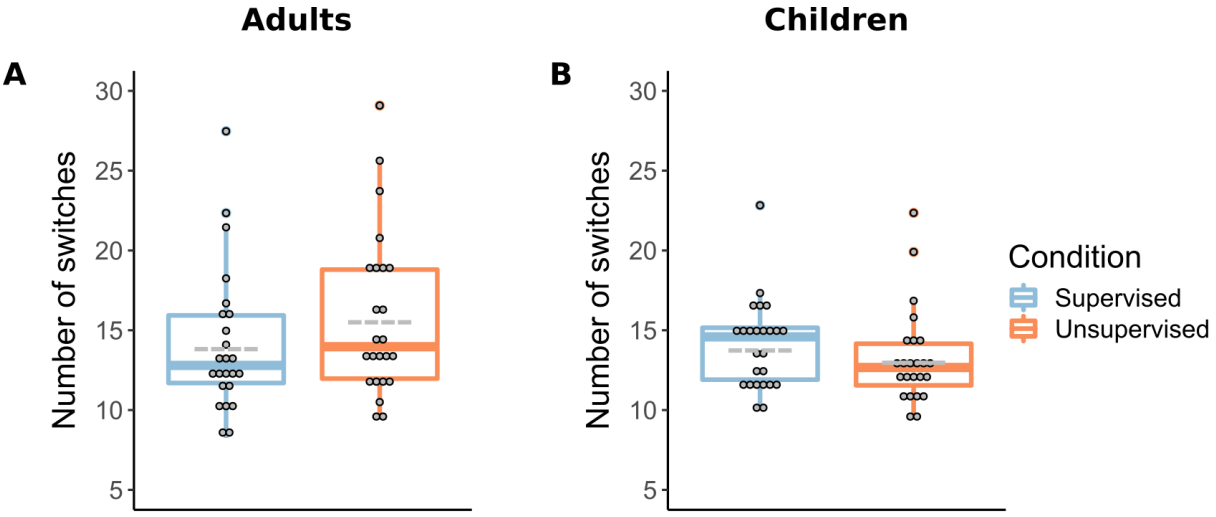


Fig. 6. Number of gaze switches during Learning phase. Mean number of gaze switches between target and foil in each trial (average over the whole Learning phase). **(A) Adults.** Adults in the Unsupervised Condition gaze-switch more, reflecting lower confidence ($p = 0.017$). **(B) Children.** Children on average gaze-switch at similar levels in both leaning conditions. However, there was an interaction between Literacy and Condition, with literate children gaze-switching more in the Unsupervised Condition and pre-literate children more in the Supervised Condition ($p = 0.019$).

Children.
Though the number of gaze switches has been shown to reflect low confidence in adults, it is unknown whether the same holds for children (the number of gaze switches has been shown to correlate with future performance, Yu, Yurovsky, & Xu, 2012). Children (between 6 and 10 years old) have been shown to stop searching for information much later than adults: even when children have enough evidence to make a decision, they continue exploring (Ruggeri, Lombrozo, Griffiths, & Xu, 2016). Less is known, however, about the cognitive mechanisms that drive this exploration. Having an explicit measure of confidence (confidence rating scale) allows us to compare, descriptively, whether the number of switches reflects reported confidence. When performance was matched, children in the Unsupervised Condition were less confident, and thus if children continue exploring because of low confidence, then they should

also have a greater number of gaze-switches. On the other hand, if children continue exploring despite high confidence, then there may not be a difference between the two conditions.

Children did not switch gaze more in one learning condition than the other [no main effect of Condition: $\beta = 0.98$, SE = 0.78; model comparison: $\chi^2(1) = 1.56$, $p = 0.21$, Fig. 6B]. There was also no significant difference in the number of switches by literate and pre-literate children [no main effect of Literacy: $\beta = -0.52$, SE = 0.78; model comparison: $\chi^2(1) = 0.44$, $p = 0.51$]. However, learning condition had an inverse effect on literate and pre-literate children's gaze distribution: literate children gaze seemed to switched less in the Supervised Condition, but pre-literate children more in the Supervised Condition [interaction Condition and Literacy: $\beta = -3.75$, SE = 1.56; model comparison: $\chi^2(1) = 5.48$, $p = 0.019$]. To correctly interpret the cross-over interaction in the absence of main effects, we examined each group of children separately, literate and pre-literate: literate children in the Supervised Condition did not gaze-switch significantly less than those in the Unsupervised Condition [no main effect of Condition: $\beta = -0.79$, SE = 1.04; model comparison: $\chi^2(1) = 0.55$, $p = 0.46$]; but pre-literate children in the Supervised Condition did switch significantly more than those in the Unsupervised Condition [main effect of Condition: $\beta = 2.32$, SE = 0.94; model comparison: $\chi^2(1) = 5.1$, $p = 0.02$]. The number of switches, thus, appeared to echo performance, but it did not appear to corroborate subjective confidence ratings. We did not find positive evidence in favour of the theory that children continue to explore even when they have enough evidence in favour of a decision: children appeared to explore visual scenes in accordance to performance. Nevertheless, children's overall performance was quite low, and persistent exploration may only be noticeable when task performance is high. Though children's gaze distribution yoked performance, as it had for adults, it did not reflect a metacognitive bias, as it also had for adults. This dissociation is interesting both for understanding the cognitive processes influenced by supervision and children's metacognition. It will be addressed in detail in the discussion section.

Summary.
For adults, gaze distribution, thus, provided convergent evidence for a metacognitive bias: supervision increases confidence (and reduces the number of gaze switches). For children, on the other hand, there was a dissociation between subjective confidence ratings and gaze distribution. However, gaze distribution closely followed performance, and thus likely children's level of uncertainty: supervision boosted performance and reduced the number of switches for literate children, but impeded performance and increased the number of switches for pre-literate children.

Learning phase results summary.
The Learning phase was designed to reflect an ecologically valid language learning situation: a learner may find herself interpreting a language 'in the wild' after some classroom teaching (Supervised Condition) or with no prior practice (Unsupervised Condition). Our results did not find a supervision advantage for acquisition. Moreover, supervised information appeared to only be integrated as evidence if a learner knew how to read, when she had developed necessary metalinguistic capacities. This was likely because supervision involved mapping isolated words in one language to isolated words in another. Furthermore, when the evidence provided by supervised and unsupervised learning was matched, the two methods were at par for adults' performance, and unsupervised learning came out on top for children's performance. Nevertheless, supervision had a clear impact on metacognition across age groups: it increased confidence. Namely, when performance and even amount of exposure was matched, supervision made learners more confident. Intuitively, this could be because the Learning task was easier from the get-go for participants in the Supervised Condition: they came into the task

with some knowledge. Though this may be why adults and literate children felt more confident, it does not explain why pre-literate children felt more confident: their performance had not benefitted from supervision. These findings point to a more general bias, perhaps along the lines of 'It was explained to me so I ought to know'. In the next section, we will look at the influence of supervision on more specific elements of the artificial language, including the very element participants in the Supervision Condition were trained on: vocabulary.

C. Test phase: Performance

After looking at the initial learning phase, we now examine the degree of acquisition of target (nouns) and associated items (grammatical elements) after the Learning phase.

In the Test phase, we investigated the effect of supervision on performance of taught items (vocabulary), and untaught but associated items (grammar/generalization of grammatical rule). In the first test block (Test Phase 1), we measured performance on vocabulary and associated grammar, and in the second test block (Test Phase 2), associated grammar and generalization of the grammatical rule. On average, adults and children were quite good at the Test task. Adults had above chance performance on trained items, vocabulary [$M = 0.95$, $SD = 0.22$, $\beta = 1.26$, $SE = 0.14$, $p < 0.001$] and grammar [$M = 0.75$, $SD = 0.42$, $\beta = 4.01$, $SE = 0.49$, $p < 0.001$], as well as on novel items, generalization [$M = 0.62$, $SD = 0.48$, $\beta = 0.64$, $SE = 0.17$, $p = 0.001$]. Children performed above chance on trained items, vocabulary [$M = 0.65$, $SD = 0.48$, $\beta = 0.5$, $SE = 0.12$, $p < 0.001$] and grammar [$M = 0.66$, $SD = 0.47$, $\beta = 0.66$, $SE = 0.08$, $p < 0.001$], but were at chance for novel items, generalization [$M = 0.49$, $SD = 0.5$, $\beta = -0.05$, $SE = 0.1$, $p = 0.61$]. Means are presented in Fig. 7.

Adults.

For adults, the two learning conditions did not significantly influence performance overall [no effect of Condition in Test Phase 1: $\beta = 2.72$, $SE = 0.48$; model comparison: $\chi^2(1) = 1.28$, $p = 0.26$; no effect of Condition in Test Phase 2: $\beta = 0.2$, $SE = 0.33$; model comparison: $\chi^2(1) = 0.36$, $p = 0.55$]. Interestingly, by the Test Phase, learners in the Unsupervised Condition had caught up to their supervised peers: unsupervised learners had two fewer blocks of exposure to the language (Unsupervised Condition: 4 blocks of unsupervised learning, versus Supervised Condition: 2 blocks of supervised learning followed by 4 blocks of unsupervised learning, for a total of 6 blocks). This finding may however be specific to the task at hand, and is consistent with the possibility that participants in the Supervised Condition had more evidence in favour of vocabulary. For instance, if we had elicited production (asked participants to say the vocabulary item that corresponded to the images), participants in the Supervised Condition may have had higher performance.

In general, adults showed higher performance on vocabulary than grammar [main effect of Trial type in Test Phase 1: $\beta = 0.42$, $SE = 0.37$; model comparison: $\chi^2(1) = 49.88$, $p < 0.001$], and at grammar than generalization [main effect of Trial type in Test Phase 2: $\beta = -0.76$, $SE = 0.23$; model comparison: $\chi^2(1) = 0.15$, $p = 0.002$]. In the scope of the learning task, it is not surprising that vocabulary was learned better than grammar: vocabulary reliably predicted the corresponding image, while grammar only did so for half the trials (grammar informative trials, see Materials and Methods). Accordingly, an efficient learner ought to focus most on vocabulary.

Moreover, learning condition did not have an inverse effect on acquisition of taught (vocabulary) and associated items (grammar), as has been suggested in the literature [no

interaction between Condition and Trial type: β = 1.02, SE = 0.68; model comparison: $\chi^2(1) = 2.21$, p = 0.14]. Thus, supervision did not appear to bolster acquisition of taught items (vocabulary) nor impede acquisition of associated items (grammar). In Test Phase 2, there was a trend for an interaction between Condition and Trial type (participants in the Unsupervised Condition being slightly better at grammar but worse on generalization than those in the Supervised Condition), but it did not reach significance [no interaction between Condition and Trial type: β = 0.59, SE = 0.31; model comparison: $\chi^2(1) = 3.28$, p = 0.07]. Overall, learning condition did not appear to have a significant effect on vocabulary and grammar acquisition, nor on generalization to novel items.

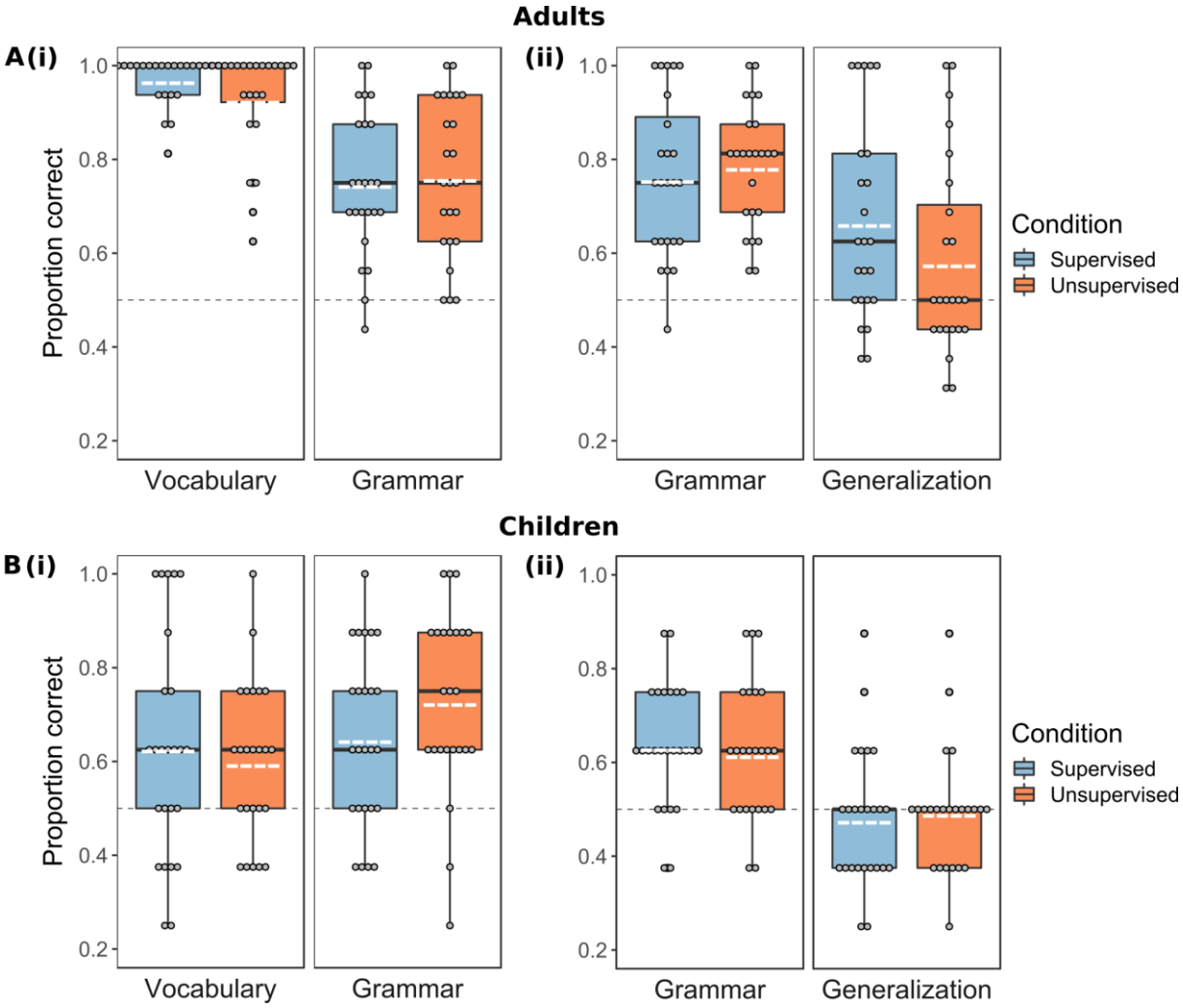


Fig. 7. Test phase results. Proportion of correct answers per trial type (Vocabulary, Grammar, Generalization). Supervised Condition in blue, Unsupervised Condition in orange. **(i) Test Phase 1. (ii) Test Phase 2. (A) Adults.** Adults are significantly better at vocabulary than grammar ($p < 0.001$), and likewise at grammar than generalization ($p < 0.001$). Condition did not significantly influence performance. **(B) Children.** Children are just as good at vocabulary as grammar, but they are significantly better at grammar than generalization ($p < 0.001$). Condition only interacted with literacy (Fig. 2B).

Children.

In Test Phase 1, the learning condition did not have an across-the-board positive or negative effect on children's performance [no main effect of Condition: β = -0.16, SE = 0.16; model comparison: $\chi^2(1) = 1.04$, p = 0.31]. Rather, supervision boosted performance for literate children, while it stalled performance for pre-literate children [interaction between Condition

and Literacy: $\beta = 1.16$, SE = 0.33; model comparison: $\chi^2(1) = 11.99$, $p < 0.001$, Fig. 2B; see Fig. S5 for a breakdown of performance on each trial type by literacy]. To better quantify the interaction between Condition and Literacy in the absence of main effects, we analysed the data from literate and pre-literate children separately. Literate children were significantly better at test when they had learned with supervision [main effect of Condition: $\beta = -0.75$, SE = 0.24; model comparison: $\chi^2(1) = 8.25$, $p = 0.004$]; but pre-literate children were better when they had learned without supervision [main effect of Condition: $\beta = 0.46$, SE = 0.22; model comparison: $\chi^2(1) = 4.33$, $p = 0.037$]. These results suggest that supervision can have a negative effect on acquisition, when compared to unsupervised learning. In Test Phase 2, learning condition did not affect performance [no main effect of Condition: $\beta = -0.02$, SE = 0.15; model comparison: $\chi^2(1) = 0.04$, $p = 0.84$; no interaction between Condition and Literacy: $\beta = 0.32$, SE = 0.29; model comparison: $\chi^2(1) = 1.21$, $p = 0.27$]. Together, the results from the two test phases indicate that supervision bolstered acquisition of trained items (vocabulary and grammar) for literate children, and impeded acquisition for pre-literate children, and that it did not procure a visible effect on interpretation of untrained items (generalization).

In contrast to adults, in Test Phase 1, there was a general trend for children to have a higher performance on grammar than vocabulary, but it did not reach significance [no main effect of trial type: $\beta = -0.35$, SE = 0.19; model comparison: $\chi^2(1) = 3.22$, $p = 0.07$]. However, like adults, in Test Phase 2, children were better at grammar than at novel items, generalization [main effect of trial type: $\beta = -0.58$, SE = 0.15; model comparison: $\chi^2(1) = 15.83$, $p < 0.001$]. There was no interaction between Trial type and Condition, nor Trial type and Literacy [no interaction between Condition and Trial type in Test Phase 1: $\beta = 0.54$, SE = 0.38; model comparison: $\chi^2(1) = 1.96$, $p = 0.16$; no interaction between Condition and Trial type in Test Phase 2: $\beta = -0.08$, SE = 0.29; model comparison: $\chi^2(1) = 0.08$, $p = 0.78$; no interaction between Literacy and Trial type in Test Phase 1: $\beta = 0.33$, SE = 0.38; model comparison: $\chi^2(1) = 0.75$, $p = 0.39$; no interaction between Literacy and Trial type in Test Phase 2: $\beta = -0.14$, SE = 0.29; model comparison: $\chi^2(1) = 0.22$, $p = 0.63$]. Accordingly, for children, like adults, supervision did not appear to procure an advantage for explicitly taught items (vocabulary) nor a disadvantage for associated items (grammar).

Summary.

It appears that, across age groups, supervision did not bolster performance of taught items (vocabulary), while hindering acquisition of associated items (grammar), as we had expected from the literature. However, supervision affected literate and pre-literate children differently: supervision was a boon for literate children, but a bane for pre-literate children. It would have been interesting to see whether this advantage and disadvantage plateaued after a certain performance threshold (as it had for adults) or whether the bolstering or hindering persisted well into the learning process. Interestingly, there was a progressive change in what is learned best, across age and literacy: while adults had learned vocabulary better than grammar, literate children were as good at both [$\beta = -0.19$, SE = 0.3; model comparison: $\chi^2(1) = 0.41$, $p = 0.52$], and pre-literate children were best at grammar [$\beta = -0.5$, SE = 0.24; model comparison: $\chi^2(1) = 4.33$, $p = 0.04$]. This result may solely reflect literacy skills (adults have high literacy skills, literate children medium literacy skills, and pre-literature children low literacy skills), or an interaction between literacy and the development of other cognitive capacities (e.g., adults may be more 'efficient' learners, focusing on what is most relevant for the learning task, here vocabulary). It would be interesting to see what children who have a couple years of reading experience (e.g., 10-12-year-olds) would be best at. Together, our results provide evidence that supervision may not be universally beneficial for acquisition.

D. Test phase: Metacognition

Next, we investigate language-level metacognition. We wanted to know how well learners could track performance across different linguistic elements (e.g., vocabulary versus grammar, metacognitive ability) and whether they felt more confident in some of the elements (when trial-pair accuracy was equal, metacognitive bias). In the first test block (Test Phase 1), we assessed metacognition when vocabulary was pit against associated grammar, and in the second test block (Test Phase 2), we measured metacognition when associated grammar was pit against interpreting novel nouns (generalization of the grammatical rule).

Metacognitive ability.

Adults.

Adults were able to track their performance in Test Phase 1 [main effect of Confidence: $\beta = 5.17$, SE = 0.85; model comparison: $\chi^2(1) = 69.55$, $p < 0.001$], but not Test Phase 2 [no effect of Confidence: $\beta = 0.41$, SE = 0.26; model comparison: $\chi^2(1) = 2.41$, $p = 0.12$]. This capacity was not moderated by learning condition [no interaction between Condition and Confidence in Test Phase 1: $\beta = 1.8$, SE = 1.64; model comparison: $\chi^2(1) = 1.47$, $p = 0.22$; no interaction between Condition and Confidence in Test Phase 2: $\beta = 0.2$, SE = 0.52; model comparison: $\chi^2(1) = 0.14$, $p = 0.71$]. Thus, adults appeared to know when they were right and when they were wrong when comparing Vocabulary and Grammar questions, but not Grammar and Generalization questions. Additionally, supervision did not seem to provide an advantage for metacognitive ability, and this is likely because performance on vocabulary trials (for which there ought to be an advantage for supervised learners, if there were one) is very high for both conditions.

Children.

Children, like adults, were able to monitor performance in Test Phase 1 [main effect of Confidence: $\beta = 0.49$, SE = 0.23; model comparison: $\chi^2(1) = 4.55$, $p = 0.03$], but not Test Phase 2 [main effect of Confidence: $\beta = -0.06$, SE = 0.22; model comparison: $\chi^2(1) = 0.1$, $p = 0.75$]. The effect of metacognitive ability was not mediated by learning condition in Test Phase 1 [no interaction between Confidence and Condition: $\beta = -0.73$, SE = 0.45; model comparison: $\chi^2(1) = 2.62$, $p = 0.11$]. However, there was an interaction between Confidence and Condition in Test Phase 2 [interaction between Confidence and Condition: $\beta = -1.63$, SE = 0.44; model comparison: $\chi^2(1) = 14.36$, $p < 0.001$]. To determine the source of this interaction in the absence of main effects, we ran the same analysis separately on each of the two learning conditions. The analysis revealed that children in the Unsupervised Condition had metacognitive ability [$\beta = 0.74$, SE = 0.32; model comparison: $\chi^2(1) = 5.5$, $p = 0.02$]: they knew when they had responded correctly and when they had responded incorrectly. In stark contrast, children in the Supervised Condition had, what we will call here, 'inverse metacognitive ability' [$\beta = -0.88$, SE = 0.29; model comparison: $\chi^2(1) = 9.32$, $p = 0.002$]: they classified incorrect answers as 'high confidence', and correct answers as 'low confidence'. Inverse-metacognitive ability has been attested in studies investigating children's confidence judgements (ages 6-10; Baer, Gill, & Odic, 2018). Children explain these inverse-confidence choices as 'challenges' (Baer et al, 2018). This is an interesting result that, most importantly, may indicate that children have metacognitive ability (be it correct or opposite) in Test Phase 2. We will first see if we replicate these results with a second group of children (on the whole set of children we plan to test, $n = 96$).

Metacognitive Bias.

Adults.
On average, adults had a metacognitive bias for vocabulary questions in Test Phase 1 [above chance intercept: $\beta = -1.17$, SE = 0.16, $p < 0.001$, Fig. 8A(i)] and this effect was stronger for participants in the Supervised Condition [main effect of Condition: $\beta = -1.06$, SE = 0.33; model comparison: $\chi^2(1) = 10.51$, $p = 0.001$, means are presented in Fig. 8A(i)]. They also had a bias for grammar questions in Test Phase 2 [above chance intercept: $\beta = 1.42$, SE = 0.18, $p < 0.001$, Fig. 8A(ii)], but this bias did not differ between conditions [no main effect of Condition: $\beta = 0.09$, SE = 0.37; model comparison: $\chi^2(1) = 0.06$, $p = 0.8$]. Thus, adults appear to feel more confident about vocabulary than associated grammar, and more confident about associated grammar then generalization.

Children.
Children did not have a metacognitive bias in Test Phase 1 [non-significant intercept: $\beta = -1.12$, SE = 0.14, $p = 0.38$, Fig. 8B(i)] nor in Test Phase 2 [non-significant intercept: $\beta = -0.16$, SE = 0.15, $p = 0.26$, Fig. 8B(ii)]. Learning condition did not modulate metacognitive bias [no main effect of Condition in Test Phase 1: $\beta = 0.34$, SE = 0.28; model comparison: $\chi^2(1) = 1.47$, $p = 0.23$; no main effect of Condition in Test Phase 2: $\beta = -0.17$, SE = 0.29; model comparison: $\chi^2(1) = 0.32$, $p = 0.57$] nor did literacy [no main effect of Literacy in Test Phase 1: $\beta = 0.04$, SE = 0.28; model comparison: $\chi^2(1) = 0.02$, $p = 0.87$; no effect of Literacy in Test Phase 2: $\beta = -0.12$, SE = 0.29; model comparison: $\chi^2(1) = 0.16$, $p = 0.69$]. Thus, children, in contrast to adults, appear to not be biased in favour of one trial type, even when they had higher overall performance on it. Moreover, even though participants in the Supervised Condition had more evidence in favour of vocabulary trials (two blocks extra exposure), there was no effect on metacognitive bias.

Summary.
Our results demonstrate that adults and children are able to track performance when comparing vocabulary and grammar (broadly, trained items), but less so when comparing grammar versus generalization (broadly, when faced with a new element). Interestingly, adults were biased towards the items they knew best, feeling more confident in vocabulary than grammar and grammar than vocabulary, but children were not. Furthermore, supervision appeared to exaggerate adults' bias for vocabulary. There thus seem to be convergences and divergences in metacognition for adults and children: both age groups have greater ease tracking performance of trained items (vocabulary and associated grammar), but only adults are biased to one linguistic item over another.

Test phase results summary.
The Test phase was designed to probe the effect of supervision on performance of individual linguistic elements, and metacognition across linguistic elements. A comparative measure of metacognition allowed us to evaluate how to the mind estimates confidence in ecologically valid situations, when performance and familiarity of items varies. The Test phase results found a supervision advantage for acquisition for literate children, but not for adults or pre-literate children.

By the Test phase, performance for adults who had a head start (Supervised Condition) and those who did not (Unsupervised Condition) had converged, at least on the elements measured by our task. Notably, performance on associated items (grammar/generalization) was not impaired by supervision on a target item (vocabulary). The Learning phase, common to

both learning conditions, was an unsupervised learning task, and unlike varying forms of supervision, may inevitably be accompanied by a degree of uncertainty. A learner who knows that *Pirdale* means 'book' will, at least in the very beginning, have to consider the role of *Dol ka* in the sentence '*Dol ka pirdale*' in order to determine whether the sentence corresponds to the image of book (just like the word *pirdale* does), or whether it now somehow corresponds to the other image, that of a cat (*Dol ka* modifying *pirdale*). However, a vestige of supervision persisted: adults in the Supervised Condition felt more confident for vocabulary than adults in the Unsupervised Condition. This result is consistent with two explanations: adults in the Supervised Condition had more evidence in favour of vocabulary (and thus would have performed better on a different, more difficult, task) or they felt they had more evidence in favour of vocabulary (but their performance would have been similar regardless of the task).

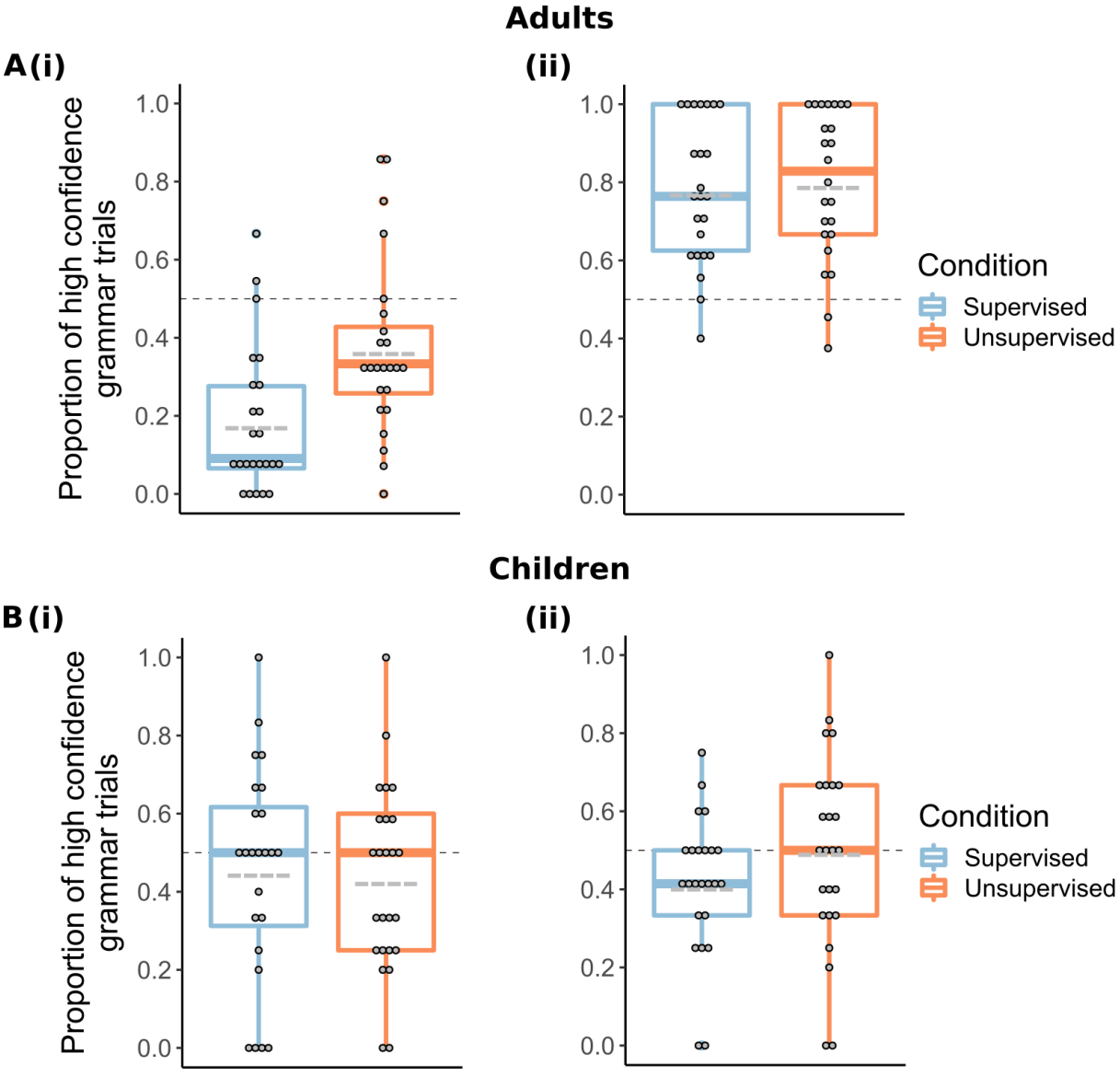


Fig. 8. Test phase metacognitive bias. Proportion of high confidence grammar trials when performance was the same on a trial-pair (both correct, both incorrect). Dashed black line indicates an equal proportion of high confidence grammar and vocabulary trials (Test Phase 1) or high confidence grammar and generalization trials (Test Phase 2). **(i) Test Phase 1. (ii) Test Phase 2. (A) Adults.** In Test phase 1 (i), adults had higher confidence for vocabulary trials than grammar trials ($p < 0.001$), and adults in the Supervised Condition were more biased than those in the Unsupervised Condition ($p = 0.001$). In Test Phase 2 (ii), adults had higher confidence for grammar trials than generalization trials ($p < 0.001$). **(B) Children.** Children were not biased to one trial kind in either phase of test.

Supervision, however, continued to affect children's performance well into the Test phase: it bolstered literate children's performance, but setback pre-literate children's performance. The effect on performance was not limited to vocabulary, but also extended to associated grammar (see Fig. S4). At first glance, this finding appears to suggest that children may consider the different elements of the sentences as less separable than adults. It could nonetheless also be reducible to a task-related artefact: even though the association between a vocabulary item and a grammatical element (e.g., *pirdale* and *ka*) did not rely on knowing the correspondence between sentence and image (e.g., *Dol ka pirdale* = book), in our task participants were asked which of the two sentences (correct grammar *Dol ka pirdale* and incorrect grammar *Dol ko pirdale*) best described the image. Thus, to a certain extent, grammar was, artificially, tied to sentential meaning. If a child though that 'book' was *Dol ka nuve*, then choosing between *Dol ka pirdale* and *Dol ko pirdale* would have been difficult. Adults, who were at nearly at ceiling for vocabulary, rarely found themselves in this situation. Supervision, nonetheless, did not affect children's metacognition of the different aspects of the artificial language; this suggests that the boost in confidence for those in the Supervised Condition observed in the Learning phase may very well have been a broad language- or task-level confidence boost: children just felt more confident about their acquisition of the artificial language and/or more confidence doing the task.

The boost in general confidence for the Supervised Condition observed in the Learning phase seems to thus have been linked to higher confidence for vocabulary in adults and higher language-level confidence for children.

4.1.2 Discussion

Our experiment demonstrates a dissociation between supervision's actual and felt impact on performance: supervision is not universally helpful, and yet it is perceived as such. The effect supervision has on performance can be unpacked into three major trends. First, supervision does not necessarily procure a learning advantage in comparison to unsupervised learning. Two blocks of supervised learning were at best (only for adults) just as good as two blocks of unsupervised learning on a later language-level task. Second, supervision may only have a positive effect on learning if there is an alignment between how supervision presumes cognitive processing to function and how it functions in reality. Supervision, in our task, was composed of direct translations of vocabulary, and it assumed that vocabulary items were considered by the mind to be self-standing, isolable elements. Adults and literate children both benefited from two extra blocks of supervised learning, but pre-literate children did not. Even though adults were more familiar with the concept of direct translation, literate children also procured a benefit from them. It seems therefore that literate and pre-literate learners' conception, and more broadly processing, of language may be quite different (in accordance with previous research, e.g., Brunswick et al, 2012; Havron & Arnon, 2017a, 2017b). Accordingly, when supervision is based on an idea of cognitive processing that misses the mark, it may not have a positive effect on a learner. Lastly, supervision can have a downright negative affect on acquisition. Pre-literate children who had two extra blocks of supervision performed worse than those who had only learned without supervision. Therefore, unusable evidence can nonetheless affect acquisition, sending a learner down the wrong track, and impeding acquisition. Yet, despite the vastly different effects supervision had on performance, learners universally felt more confident.

This work provides insight as to why a mind that is acutely adept at learning in an unsupervised manner inevitably turns to supervision when it can: supervision boosts confidence. Learning with supervision is not necessarily easier or faster, but it does seem to be that way. At first glance, this bias appears to be reducible to a general trend in supervised learning: supervision may not have uniformly aided performance in our task, but it may generally aid performance. Yet, no one learner could have acquired the same skill with and without supervision, and have observed that one way is easier or faster. She may have compared across tasks, but even then, there are few skills one acquired entirely with or without supervision. We argue, rather, that supervision may mimic aspects of a 'knowledge state', and it may be this that makes it so appealing. One of these aspects may be certainty: when one knows something, one is sure of said thing (e.g., I know *pirdale* means 'book'). Supervision, by definition, conveys indubitable evidence: *pirdale* means 'book' (not: (a) *pirdale* may mean book, (b) *pirdale* likely means book, (c) *pirdale* either means book or lion or shoe, (d) *pirdale* is some sort of object). In other words, supervision offers already-processed evidence; it provides the learner with knowledge, not simply evidence. Thus, supervision's main appeal may be the second-order evidence it provides: certainty about the first-order evidence.

Certainty, as we define it here, relies on a distinction between above-chance performance and knowledge. Learning basic associations may be enough for above-chance performance (e.g., choosing 'book' 60% of the time, when one hears *Dol ka pirdale*), and perhaps for hunches or inklings. However, only when a learner has gleaned a more abstract structure about the evidence (e.g., *pirdale* means 'book'), she will know what she knows. More broadly, we propose that this may be one of the fundamental bases of metacognitive ability: knowing that, and what, one knows relies on comparing one's performance to some standard, an abstract structure, however rudimentary. This abstract structure may thus serve as a standard, a ruler to accurately measure performance. This would explain why generally metacognitive ability increases as performance increases, but is warped at near-chance levels (e.g., Kunimoto, Miller, & Pashler, 2001; Schwiedrzik et al, 2011; Kruger & Dunning, 1999; for a review about how metacognition and performance correlate, Harvey, 1997). Early in learning there may not yet be an abstract structure in place with which to compare performance. If supervision provides an abstract structure, then we ought to observe greater metacognitive ability early in learning. Adults, in our study, had greater metacognitive ability in the Supervised than Unsupervised Condition in the Learning phase, when performance was matched. Supervision may have offered learners a distilled evidence, an abstract structure, with which to compare performance, something that would take longer to do (and a higher mean performance) for learners in the Unsupervised Condition. This difference disappeared by the Test phase, when performance for both groups was very high. Likely at this point, learners in the Unsupervised Condition had extracted the necessary structure, distilled the evidence themselves. For children, we did not observe a significant difference in metacognitive ability between groups. However, there was a very interesting correlation between mean confidence in the Learning phase and metacognitive ability in the Test phase: children who were more confident early in learning, were later better at judging their performance. Indirectly, most of the children who had high confidence had learned with supervision, and thus it seems that they were also better (though not significantly) at monitoring performance. We hope that the added power of a second group of children we plan to test (such that we have 24 rather than 12 pre-literate children in each condition, and 24 rather than 12 literate children in each condition, total 96) will clear up some of the noise in children's data, and perhaps allow us to make a conclusion rather than a claim. What may therefore be so appealing about supervision is this already-processed status of evidence: evidence as knowledge.

Our results revealed a broad metacognitive bias in favour of supervision, and the design of our experiment elegantly allowed us to compare the effect of supervision on metacognitive bias when performance was equal, when the total amount of exposure was equal, and when the task was easy or hard. For adults, when performance and exposure were equal, supervised learners felt more confident. For children, when performance was equal, supervised learners felt more confident; when exposure was equal (and unsupervised learners had higher performance), supervised learners nonetheless felt more confident; when pre-literate learners in the Supervised Condition procured no advantage for the task (and thus could not have simply felt more confident because the task was easy), they nonetheless felt more confident. Thus, the metacognitive bias cannot be reduced to having more first-order evidence or task-ease. Furthermore, it cannot be reduced to perceived task-ease: our gaze distribution results revealed that pre-literate children gaze switched more when they were in the Supervised Condition. This result suggests that pre-literate children were likely quite uncertain during the decision-making process. As such, the metacognitive bias cannot be reduced to supervision eliciting a happier or happy-go-lucky learning state, which has been shown to boost confidence (e.g., Koellinger & Treffers, 2015; Massoni, 2014). The high confidence bias for supervised learning thus appears to directly stem from the interpretation of the kind of evidence learners had, supervised versus unsupervised, and how they think they ought to feel.

Our results raise two broad questions about the nature of the supervision bias: first, its universality, and second, its origin. In our study, supervision took the form of testimonial evidence. This kind of supervision may be a very potent teaching method (see for example, Csibra & Gergely, 2009). Further research is needed to determine whether other kinds of explicit supervision that may be less testimonial (e.g., seeing an image of a book on a screen and hearing *pirdale*) could elicit a bias, and if so, whether there are degrees of bias; and moreover whether implicit supervision (e.g., exaggerating target elements) can too elicit bias. Moreover, our study, though spanning two age groups, looked at the effect of supervision on a population that has been exposed to classroom teaching (children all attended either a pre-school or grade-school at the time of test). Classroom teaching may promote a supervision bias. Extending this kind of investigation to age groups who have not yet begun attending school and perhaps to other non-Westernized societies (to match learners in age and compare them on schooling experience) would provide important clues as to the origins of this bias.

Beyond ascertaining the effect of supervision, this experiment also answers fundamental questions about learning 'in the wild'. Our task was designed to be an ecologically valid mini-language learning scenario. In particular, we recreated two kinds of linguistic elements: one that is sufficient to predict the meaning of the sentence (i.e., vocabulary), and one that is only sufficient to predict the meaning of the sentence for a subset of trials, but allows to generalize to new vocabulary items (i.e., grammar). In the context of communication, uttering a sentence with the wrong vocabulary item impedes communication (e.g., Uttering 'Look at the *cat'* in the presence of a book), but a sentence with the wrong grammatical element, ostensibly, less so (e.g., Uttering 'Look *from* the book' in the presence of a book; see for an example of cases in which grammatical errors go almost unregistered by the mind, Caffarra & Martin, 2019). In other words, though a word's context carries clues to its meaning, and even bears the burden of some of the meaning, because it is not always sufficient to pinpoint a referent, it appears to be less important for communication than the word itself. We propose that grammar may not be inherently more difficult to learn than vocabulary, but it may receive less weight during learning. Second-language grammar is often notably less well acquired by adult learners than vocabulary (DeKeyser, 2000; Johnson & Newport, 1989), and yet there is enormous variability in how well second-language grammar is acquired by an individual learner. There are many factors that

mediate how well second-language grammar is learned, such as a learner's native language, but our study tackles one specific factor: the way a learner goes about acquiring a language. Adults are known to be focused, and thus seemingly efficient learners: they are quick to zoom in on the essential features to succeed at a task, ignoring others (Best, Yim, & Sloutsky, 2013; Deng & Sloutsky, 2015; Ruggeri et al, 2016). Children, on the other hand, appear to take into consideration seemingly irrelevant features, even when they have identified the essential feature (Best et al, 2013; Deng & Sloutsky, 2015; Decker, Otto, Daw, & Hartley, 2016; Ruggeri et al, 2016). At first glance, this may appear to be an inefficient way of learning, and of allocating cognitive resources. However, if we apply these previous findings to understand the results of our ecologically valid learning task, children may actually be the more efficacious learners: what is seemingly unimportant for one task, may not be for another. In the Learning phase, grammar was relatively unimportant, because vocabulary reliably predicted the corresponding image of each sentence. Thus, adults, efficient learners as they are, likely focused on the vocabulary, and therefore at test, were significantly better at vocabulary than grammar. Children, on the other hand, likely spread attention out more broadly throughout the whole sentence. They were just as good at vocabulary and grammar. Less task-oriented learning may be particularly beneficial for acquisition of skills that have elements of varying ostensible importance, like in language.

Literate children appeared to benefit from supervision, while pre-literate children did not. A possible explanation for this effect is that literate children may have developed the necessary metalinguistic skills to transfer knowledge about one isolated word to another. Pre-literate children, on the other hand, may perhaps be more likely to consider sentences in their native language as a whole, and may not have known which properties to attribute to an isolated word and thus which to transfer to its translation. They would thus also have been more likely to interpret sentences in the artificial language as wholes. Calling the book '*Dol ko pirdale*' instead of '*Dol ka pirdale*' would have been, for a pre-literate child, like calling it '*pordale'* instead of '*pirdale',* for literate children or adults. This way of processing meaning as a whole could have been the reason why pre-literate children were best at grammar, while literate children were just as good at both and adults better at vocabulary. Pre-literate children perhaps could not chose to allocate more or less attention to one of the elements, because to them, there was but one whole. It may have been this incapacity to isolate, and compare, the different elements (in both the artificial language and the child's native language) that made supervised learning such a bane. The distinct effect supervision had on the two groups of children cannot be reduced to a baseline difficulty with the sentence 'A means B'. This sentence, though rare in everyday speech, it quite common in second-language instruction. At pre-school in France, children normally have had a few English classes, and thus literate and pre-literate children were likely exposed to similar degrees of second language instruction. Any difficulty with the sentence itself ought to thus have engendered a processing cost in both groups of children. Accordingly, it is likely that metalinguistic awareness aided children in taking advantage of the supervised evidence, while a lack thereof resulted in better learning of grammar. Importantly, these findings indicate that caution must be taken when creating classroom language teaching techniques for pre-literate children, as well as when testing a group of children that straddles the learning-to-read age.

Children may thus have a different conception of language, than adults, drawing nearer to an adult-like conception when they learn to read. A divergent understanding of language may also have led to a different impression of the novel items presented in the generalization trials. Children may have perceived the generalization trials as less novel than adults because of a general feeling of familiarity with the language, rather than of specific elements of the language.

Adults on the other hand may have had a stronger impression of novelty because the key element, the vocabulary item, was novel. This may have been one of the reasons for which children were not biased in favour of items they were familiar with and better at. Even though average performance on grammar trials was significantly higher than for generalization trials, children's relative confidence (grammar versus generalization) was at par for the two. However, this lack of bias is also consistent with a not-yet-adult-like way of computing confidence. Children may not keep track of overall confidence (e.g., overall confidence for familiar items) or may not value overall confidence as much. Previous studies have shown that children do keep track of evidence over the course of a task (e.g., having had more evidence in favour A over the course of the task, Decker et al, 2016), and thus likely also tally overall confidence. However, studies have also found that children may base decisions on evidence from a just-preceding trial, rather than a whole set of previous trials (Decker et al, 2016). In other words, children may be less prone than adults to suppose that the way the world had been up to time t, is how it will continue to be in the future. This may be because they have fewer, and less strong, priors about how the world is. Further research is needed to determine how children's and adult's confidence calculations diverge and converge, as well as the potential consequences of unbiased relative confidence on learning.

Ecologically valid learning situations present unique challenges for metacognition. A learner may know one element very well (e.g., vocabulary), another less well (e.g., grammar), and mixed into all of that, there may be novel elements a learner never encountered before (e.g., generalization). On top of a difference in mean performance, elements vary also in their nature (e.g., vocabulary versus grammar). A learner must thus keep track of her performance across different levels of knowledge and different kinds of elements. Our task was specifically designed to test how the human mind navigates these 'metacognitively noisy' environments. Adults have been shown to accurately compare performance on different tasks: one can accurately compare performance on a visual discrimination task and an auditory discrimination task (De Gardelle, Le Corre, & Mamassian, 2016; de Gardelle & Mamassian, 2014). However, this comparison is most often measured when performance is matched on both tasks. One study, which investigated metacognition when performance across tasks differed, found that adults appear to be biased in favour of the element they are on average better at, and that this bias reduced metacognitive ability (Maniscalco, Peters, & Lau, 2016). Instead of objectively assessing performance on each trial, adults' confidence was influenced by average performance. Insofar as metacognition ability is thought to serve as a guide for acquisition, this bias could have an effect on future learning (e.g., Hainguerlot, Vergnaud, & De Gardelle, 2018). In our study, adults felt more confident in decisions for trial types where they had, on average, higher performance: they felt more confident in vocabulary than grammar, and grammar than generalization. This bias may have been what reduced metacognitive ability in the second block (grammar versus generalization trials), even though adults had well above chance performance on both trial types. Adults may have been so biased in favour of grammar trials, that they favoured grammar trials even when trial-pair accuracy was not equal (e.g., generalization trial correct, but grammar trial incorrect). This bias for elements that one is generally better at may disrupt acquisition. First, it may cloak metacognitive ability for an element one is good at, such that a learner may, counterintuitively, have high confidence in errors for that linguistic element, preferring it in any situation over a lesser known element. Second, it may dissimulate potential metacognitive ability for a novel item; a learner may take longer to realise she has actually learned a novel item if she consistently feels less confident for it than other items she knows better. Though intuitively odd, metacognition can fail to track performance even when performance was high. Dissociations between objective and subjective accuracy (i.e., confidence and performance) have been observed (e.g. Rahnev, Maniscalco, Graves, Huang,

de Lange, & Lau, 2011; Recht, Mamassian & De Gardelle, 2019). Notably, relying on task-level, rather than trial-level confidence induces a dissociation between confidence and performance (Rahnev, Koizumi, McCurdy, D'Esposito, & Lau, 2015). Further research is needed to pinpoint whether, and how, a bias that adds noise to metacognitive ability steers future learning, and more broadly how minds navigate metacognitively noisy situations.

In our study, we used three measures of metacognition: confidence scale, two-alternative forced-choice confidence and gaze distribution. Importantly, we were able to compare these measures between adults and children. Adults' confidence scale ratings converged with gaze distribution data (as had been shown before, Folke et al, 2017), but this was not the case for children. There are two possible explanations for this dissociation: the first has to do with the possible interpretations of the task and the second is a hypothesis about the underlying cognitive mechanisms. The adult task and child task differed ever so slightly in how they were presented. This was done in an attempt to make the task pragmatically plausible and engaging for both age groups. While the adults were asked to report their confidence, children were told that the little alien avatar (the one who was teaching them her language) wanted to know how confident they were in the answer they just gave. As such, for children, the confidence report could have been perceived as a much more social act. They may have felt a greater pressure to show confidence (e.g., she told me the answers, I ought to feel confident). Gaze-switch would have, under this interpretation, corresponded to internal personal confidence, and the confidence rating to external social confidence.

The observed dissociation between gaze switch and confidence scale reports in children could also tell us something about the cognitive mechanisms underlying metacognition. The two measures may have tracked different aspects of confidence: trial-specific certainty that a decision is correct and task-level aggregate probability that a decision is correct. When a learner has to make a decision (e.g., choosing the image that corresponds to a sentence), she will monitor how likely she is to make the right choice (choice-specific confidence); however, she may also aggregate information from her decision with other priors (e.g., how often she chose correctly before; choice-independent confidence) to refine her confidence in her choice. Gaze switches may have reflected children's choice-specific confidence, while the confidence scale may have reflected children's task-level confidence. Broadly, trial-specific confidence would correspond to uncertainty and task-level confidence to an aggregate probably of being correct. Intuitively, uncertainty and aggregate probably of being correct appear to be two sides of the same confidence coin. Though confidence is believed to closely track uncertainty, what confidence is built upon will depend on one's school of thought. There are two broad schools of thought as to what is sufficient and necessary to make a confidence judgement (Fleming & Daw, 2017). The first posits a one-to-one relationship between confidence and uncertainty (e.g., Meyniel, Sigman, & Mainen, 2015). In other words, confidence is just a mirror of uncertainty. In contrast, the second posits that there is more to confidence than just tracking uncertainty. Confidence is rather an aggregate of multiple sources of information, including uncertainty, to best reflect the probability that a decision is correct given the evidence (e.g., Pouget, Drugowitsch & Kepecs, 2016). Though it is difficult to dissociate confidence from certainty, precisely because one tracks the other, studies have found that confidence can both yoke uncertainty and spread out to incorporate multiple sources of information (Bang & Fleming, 2018). Children thus may have an internal measure of uncertainty from early on, and refine what other information is used to evaluate confidence through development. This would also explain why metacognitive ability is observed from a young age, but remains slightly divergent until adolescence (Goupil et al, 2016, Salles et al, 2016; Finn & Metcalfe, 2014; van Loon et al, 2017). By adulthood, aggregate task-level information may directly feed into trial-specific

uncertainty, rendering the two difficult to dissociate. Studying metacognition in children could thus allow us to better understand the foundations of confidence. Importantly, the dissociation between gaze switches and confidence scale reports in children highlights a potentially distinct internal metric of uncertainty and external metric of confidence early in development.

In sum, our results demonstrate that supervision can have diverse effects on performance, ranging from a benefit to a bane, but that it uniformly augments confidence. We propose that supervision may pre-package evidence into 'knowledge state' bundles, and that even when this evidence is detrimental for performance a learner discordantly feels that she has knowledge. We show that supervision may need to align the evidence that it provides with cognitive processing in order to be a potentially potent learning method. This differential effect, which supervision has on learning, has important consequences for classroom teaching, and in particularly teaching of populations that do not necessarily process the world with an adult mind. More broadly, our study couples an ecologically valid learning situation with diverse measures of performance and metacognition, across different points in development. The data reveals convergences and divergences that shed new light on how the mind learns. Notably, we show a fundamental difference between comparative metacognition in adults and children and a dissociation between internal uncertainty and confidence scale reports in children. We hope that our data and discussion motivate future studies exploring how the mind learns, and, importantly, why it learns the way it does.

4.1.3 Materials and Methods

This is a pre-registered study (https://osf.io/s87e5/). All sample sizes, materials, methods, and analyses were pre-registered unless otherwise stated.

Experiment 1: Adults.

Participants.

Adults were recruited from the lab database and near-by universities. A total of 48 French native speakers participated in the study (24 for the Supervised Condition; 24 for the Unsupervised Condition). Sample size was chosen based on previous studies investigating grammar and vocabulary learning (e.g. Seigleman & Arnon, 2015). Randomization across conditions was determined by participant number (prior to the experiment, participant numbers were randomly assigned to one of the two conditions such that there were 24 in one condition and 24 in the other). The experimenter was not blind to the condition. Participants had learned on average 2.5 languages, to varying levels, in addition to French (min: 1; max: 6). 16/48 reported having heard another language from birth: Arabic (n = 3), Bambara (n = 2), Mandarin Chinese (n = 3), Portuguese (n = 2), Comorian (n = 1), Kabyle (n = 1), Occitan (n = 1), Spanish (n = 1), Turkish (n = 1), Ukrainian (n = 1), Russian (n = 1). 45/48 reported enjoying learning languages. All 48 adults were included in the analysis (mean: 24;3 years; range: 18;1-34;11 years; 36 females, 12 males). Written informed consent was obtained from each participant prior to the experiment. All research was approved by the local ethics board: CER Paris Descartes.

Materials.

Experimental protocol. The task was composed of three phases: Teaching phase, a Learning phase, and a Test phase. Participants completed either the supervised or unsupervised versions of the task. For the Supervised Condition, the Teaching phase consisted of being exposed to translations of to-be-learned words; for the Unsupervised Condition, the Teaching phase consisted of exposure to the visual environment of the experiment, namely the images. For both

conditions, the Learning and Test phases were the same. The learning phase was a cross-situational learning task, in which participants had to learn to associate sentences with the image to which they referred. The test phase was a two-alternative forced choice task, in which participants had to choose which of two sentences matched the image. No feedback was provided at any point during the experiment.

Participants were tested individually in a sound proof booth at the lab. They were seated at a distance of 65cm from a 27' screen. Gaze was recorded with an Eyelink 1000 eye-tracker at a frequency of 500 Hz. A five-point child-friendly calibration was used. After calibration, the experiment began (for written instructions presented to participants, see Annex S1). The experiment was presented in video game form: as participants progressed from trial to trial, a little monster avatar would collect coins. At the end of each block, a screen with a treasure chest would appear and the collected coins would fall into the treasure chest. When all the coins had fallen into the treasure chest, an upbeat sound sequence would play and the treasure chest would grow a little bigger. Just after, a video (~30s) unrelated to the task would play. The videos were chosen to be entertaining for both children and adults, and served as both an encouragement and break from the task. Participants completed multiple blocks of the Teaching, Learning and Test phases, in that order. After the experiment, participants filled out a questionnaire, half about previous language learning experiences, and half about the strategies they had used during the experiment.

The experiment lasted approximately 90 minutes.

The artificial language. The artificial language contained 1 carrier phrase (*dol* /dɔl/), 2 grammatical elements (*ko* /ko/ and *ka* /ka/) and 24 nouns (e.g., *pirdale* /piʁdal/, *doripe* /doʁip/, *maligour* /maliguʁ/). All words in the language were phonotactically possible in French. The majority of the words were chosen from previous experiments (e.g., Barbir, Babineau, Fiévet, & Christophe, in prep; Shi & Melançon, 2010; Pallier, Devauchelle, & Dehaene, 2011); the rest were created (the software Lexique was used to check that the created words did not exist in French, New & Pallier, 1999). Nouns could be mono-, bi-, or trisyllabic. Each sentence in the language began with the carrier phrase, was followed by a grammatical element, and ended with a noun (e.g., *Dol ko pirdale*). Grammatical elements distinguished animacy: *ko* for animates and *ka* for inanimates. This distinction was chosen because it does not exist in French, but is nevertheless present in other languages of the world (for discussion, see Strickland, 2017); it is likewise a highly salient distinction, distinguishable early in life (e.g., McDonough & Mandler, 1998). To ensure that participants did not conflate the artificial grammatical elements with grammatical gender (which is marked in French), half of all animate nouns were feminine and half masculine, and likewise for inanimate nouns.

Nouns were randomly matched to meanings (i.e., corresponding images) for each participant. Twenty-four nouns and images were chosen from a larger pool, such that not all participants saw each image or heard each noun from the pool. Images were pictures of common nouns which are learned early in life (e.g., book, cat, car, dog; based on MacArthur Communication Development Inventories from infant experiments in the lab). An equal number of images (12/24) belonged to each of the two categories to which corresponded the grammatical elements: animate or inanimate.

All sentences were produced by a native French speaker and had natural French prosody (that corresponded roughly to a sentence such as 'Look at the cat').

Teaching phase. The Teaching phase consisted of two blocks of 16 trials (Fig. 1). In the Supervised Condition, participants heard translations of the artificial language nouns (e.g., '*Pirdale'* means 'book', Fig. 1A(i)) and had to repeat them into a microphone. Translations were preceded by an attention grabbing phrase (e.g., 'You know what?', 'Hey!'), which was

randomly assigned for each trial. When participants had repeated a translation, they then clicked the spacebar to move on to the next trial. During the post-experiment questionnaire, participants reported having memorized on average 7.26/16 translations (min: 1; max: 16). In the Unsupervised Condition, participants saw images and were instructed to observe them carefully. Images were shown for 2s. When participants had seen the image, they clicked the spacebar to move on to the next trial. 11/24 participants reported trying to memorise the order of the images. Each noun (Supervised Condition) or the corresponding image (Unsupervised Condition) was presented twice, once in each block. The order of presentation of nouns or images was randomised. Unbeknownst to the participants, the nouns or images would be the same ones they would hear or see in the learning phase.

Learning phase. The Learning phase consisted of four blocks of 16 trials, one for each sentence. Two images were presented, one on each side of the screen (~20cm x 20cm), for 2s (Fig. 1B(ii)). Then participants heard a sentence in the artificial language (e.g., *Dol ko pirdale*). After an inter-stimulus interval of 1s, the sentence was replayed. 1s after the end of the second reiteration an orange box appeared around each image, and participants chose the image they thought corresponded to the sentence by clicking the left or right arrow on the keyboard. When a response was made, the orange box around the participant's choice would flash. After the response, the participant would either pass directly to the next trial, or would be asked to report how confident she was in the response she just made, on a five-point Likert scale from 'I'm sure' to 'I guessed' (3/16 per block, Fig. 1B(iii-2)).

There were two kinds of trials: grammatically informative and grammatically uninformative (Fig. S1). Grammatically informative trials were composed of target-foil pairs belonging to different grammatical categories (animate target & inanimate foil; inanimate target & animate foil, Fig. S1A), while grammatically uninformative trials were composed of target-foil image pairs belonging to the same category (animate target & animate foil; inanimate target & inanimate foil). In grammatically informative trials, the grammatical element (e.g., *ko*) in the sentence could serve as the distinguishing element between the two images (e.g., *ko*+animate, therefore the target is the animate). The mix of uninformative and informative trials allowed us to measure whether the grammatical element had an explicit effect on the pattern of acquisition. Half the trials were grammatically informative and half were grammatically uninformative.

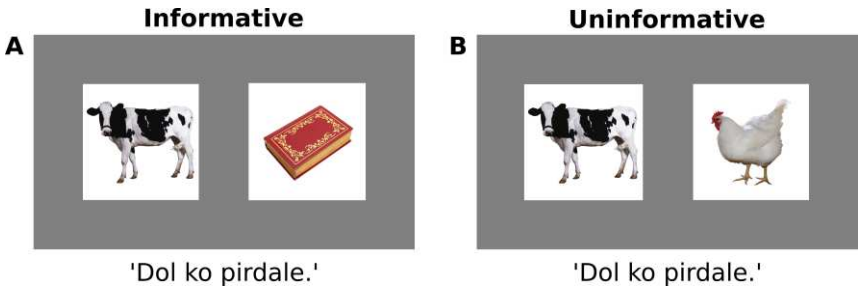


Fig. S1. Learning phase trial kinds. (A) Grammatically informative trails. The two images belonged to different grammatical categories: one animate and one inanimate. The grammatical element alone was sufficient to determine the target (e.g., *ko*+animate, therefore the cow). **(B) Grammatically uninformative trails.** The two images belonged to the same grammatical category: two animates or two inanimates. The grammatical element alone was not sufficient to determine the target (e.g., *ko*+animate, therefore the cow or the chicken).

Order and pairing of targets and foils as well as the selection of trials to be accompanied by a confidence report was pseudorandomized. Grammatical category and presentation side of targets and foils was counterbalanced and sequences were limited to four in a row (e.g., no more

than four animate images on the left side, for a complete list see Annex S2). To avoid rendering the task trivially easy, a target or foil from a trial could not reappear as either a target or foil in the very next trial.

Test Phase. During the Test Phase, participants performed a nested Two-Alternative Forced Choice task (2AFC, Fig. 1C(v-1–v-4)). Each trial began with the presentation of two squares (one blue, one red), one on each side of the screen (~7cm x 7cm, Fig. 1C(v-1)). Each of these squares contains a 2AFC task. First, the left square would grow bigger, and an image would appear. After 1s, two little monsters would appear, one on each side of the image. The monster on the left would say a sentence and after an inter-stimulus interval of 1s repeat it, then 1s later the monster on the right would do the same (Fig. 1C(v-2)). Participants selected the monster who had said the sentence that referred to the image. The sentences differed in just one critical element: **vocabulary** (different nouns): *Dol ka pirdale vs. Dol ka doripe*; **grammar** (different grammatical element): *Dol ka pirdale* vs. *Dol ko pirdale*; **generalization** (different grammatical element as well as novel noun and image): *Dol ka nuve* vs. *Dol ko nuve.* Vocabulary and grammar trials probed acquisition of each element during the training phase (i.e., did participants learn that *pirdale* was a book and did they learn that *ka* was used before the word *pirdale*). Generalization trials probed abstraction of the grammatical rule. The only way to solve the generalization task was to use knowledge about the grammatical element. If participants had learned that *ko* precedes animate nouns and *ka* inanimate nouns, then when faced with a novel image, they would be able to know which grammatical element corresponds to that image (e.g., image of a frog, therefore the animate determiner *ko*), and thus which sentence corresponds to that image (e.g., '*Dol ko nuve*' is the correct sentence and '*Dol ka nuve*' is incorrect). When a response was made, the selected monster would cartwheel, and the square would shrink to its original state. The right square would then grow bigger, and reveal a different image (Fig. 1C(v-3)). Again, the participant would have to choose the monster who said the sentence that referred to the image. After the response, the right square would close, returning to its original state. A treasure chest would appear between the two squares and participants would have the chose the answer for which they were most confident (the first/left square or the second/right square, Fig. 1C(v-4)). When a response was made, the chosen square would move into the treasure chest, while the other square would fall downwards and disappear. For each trial, participants thus responded to three questions overall: two measuring accuracy (one in each square, the red and blue one), and one measuring the relative confidence between the decisions performed in each of the squares.

The Test Phase consisted of two blocks of 16 trials, each consisting of two accuracy questions (and one relative confidence question). In the first block, Test Phase 1, each trial was composed of one Vocabulary question and one Grammar question. It evaluated acquisition of associations during the Leaning phase (i.e., noun with corresponding image and grammatical element with noun). Each trained sentence appeared once as the correct answer to one Vocabulary question and one Grammar question. In the second block, Test Phase 2, each trial was composed of one Grammar question and one Generalization question. It measured acquisition of the abstract rule governing use of the grammatical element (i.e., grammatical element and feature of the image, animate or inanimate). Each trained sentence appeared once as the correct answer to one Grammar question, and each new sentence appeared twice as the correct answer for a Generalization question (there were 8 new sentences overall, for 16 generalization questions). Generalization trials were created from 8 untrained noun-image pairs (half animate, half inanimate). This blocked design allowed us to compare participants' relative confidence between vocabulary and grammar and likewise between grammar and generalization (Fig. 1C(v-4)).

The order of questions and pairing of targets and foils was pseudorandomized. Presentation side and colour of boxes/identity of the monsters (one monster was green and yellow and the other was red and blue) was counterbalanced and sequences were limited to four-in-a-row (e.g., four animate target sentences on the left side, for a complete list see Annex S2). As in the Learning Phase, to avoid rendering the task trivially easy, a target or foil from one question could not reappear as either a target or foil in the very next question.

Questionnaire. At the end of the experiment, the experimenter administered a questionnaire. Questions related to participants experience learning languages and strategies during the experiment (see Annex S3). The questionnaire revealed that adults interpreted the grammatical element, roughly equally, in three different ways (1) distinct from the carrier phrase and noun: (a) '*Dol* + *ko/ka* + noun' or (b) '*Dolk* + *o/a* + noun'; or (2) distinct from the noun: *Dolko/Dolka* + noun. It also revealed that only 10/48 participants (5 in the Supervised Condition and 5 in the Unsupervised Condition) were able to explicitly state the grammatical rule: 4/48 deduced the rule during the Learning phase (2 in the Supervised Condition and 2 in the Unsupervised Condition), 4/48 during the Test phase (3/24 in the Supervised Condition and 1/24 in the Unsupervised Condition), and 2/48 were undecided between two rules, one of which was correct (2/48 in the Unsupervised Condition).

Experiment 2: Children.

Participants.
All children were recruited from the lab database. 55 monolingual French children participated in the study. The criteria for being monolingual was hearing 90% French or more on a daily basis, from birth. Six children were excluded from the final sample for the following reasons: having a language learning impairment (n = 1), refusing to point (n = 1) and explicitly stating that the rules of the game were different than those presented by the experimenter (n = 4). The four children who opted to play by their own rules, which consisted of choosing the answer on just one side of the screen (n = 3) or choosing the answer via a counting-out game (n = 1, e.g., eeney meeney miney moe), were reminded of the actual rules for playing the game (listening carefully to the auditory stimuli and then choosing a match). The child was excluded if she persisted with the made-up rule after the reminder. The remaining 49 children were included in the analyses (mean: 6;3 years; range: 5;8-6;11 years; 22 girls, 27 boys). Children who knew how to read (literate) and who did not know how to read (pre-literate) were matched for age as closely as possible [Literate, mean: 6;5 years; range: 5;10-6;11 years; n = 27; 14 girls, 13 boys; Pre-literate, mean: 6;1 years, range: 5;8-6;11 years; n = 22; 9 girls, 14 boys]. Older children were nevertheless more likely to be literate ($F(1, 47) = 130.2$, $p = 0.005$, $R^2 = 0.13$). Age was thus controlled for in analyses comparing literate and pre-literate children. It did not have a significant effect on any aspect of performance or confidence. Written informed consent was obtained from each child's parents prior to the experiment. All research was approved by the local ethics board: CER Paris Descartes.

Materials.
Children completed a shorter version of the adult task. Instructions were adapted to be intuitive and engaging for children. Only differences are noted below.

Experimental protocol. The experimenter was present in the sound-proof booth with the child. The experimenter provided oral instructions before each task, and encouraged the child. Children were told that the experimenter does not understand the little alien's language and thus

does not know the correct responses. No feedback was given at any point. The experimenter could see the visual stimuli, but could not hear the auditory stimuli. The experimenter was not blind to the condition, but was blind to the child's literacy level.

After the experiment, parents filled out a questionnaire about the child's literacy level. The experiment lasted approximately 45 minutes.

The artificial language. Twelve nouns and images (in comparison to twenty-four for adults) were chosen from the pool. Eight were used during training, and four were kept for generalization.

Teaching phase. The Teaching phase consisted of two blocks of 8 trials. In the Supervised Condition, when children had repeated a translation, the experimenter clicked on the spacebar to move on to the next trial. On average, children were quite good at repeating the artificial language words and their translations. Children were not excluded if they pronounced an artificial language word incorrectly (e.g., *dorite* instead of *doripe*). Incorrect pronunciations were not corrected. In the Unsupervised Condition, children saw images and were instructed to name them. Images were shown for 2s. When participants had correctly named the image, the experimenter clicked the spacebar to move on to the next trial. In general, children were quite good at the naming task. Children were not excluded if they could not name an image or named an image incorrectly: the experimenter told them the correct label for the image (example correction: "Hm. I think that's… a mouse. What do you think? Yeah, it must be a mouse. Good work.").

Learning phase. The Learning Phase consisted of four blocks of 8 trials, one for each sentence. Children were told that the little alien would tell them to look at one of the two images, but because she did not speak French well, she would say it in her language. Children were told that the experimenter wanted to play too, she wanted to know which image the little alien had told them to look at, and they were prompted to point to the chosen image when an orange box appeared around each of the images. The experimenter clicked the left or right arrow on the keyboard corresponding to the child's choice. After the response, just like adults, the child would either pass directly to the next trial, or would be asked to report how confident she was in the response she just made, on a five-point Likert scale from 'I'm sure' to 'I guessed' (2/8 per block). Children were told that the little alien sometimes wanted to know how confident they were in the choice they just made. The experimenter reminded the child of the trial and her choice (e.g., "There was a car and a book, and you chose the book"), and then asked the child to rank how correct she thought her answer was. Each level on the scale was clearly described to the child. The child was prompted to point to level that corresponded to her confidence. The experimenter clicked on the level with a mouse.

Test phase. On performance trials, children were prompted to point to the alien who said it best. The experimenter clicked the left/right arrow to select the corresponding alien. On confidence trials, children were reminded of the two trials they just completed (e.g., Experimenter points to the left box: "First, there was the dog, and you chose the yellow alien." Experimenter points to the right box: "Then, there was the book, and you chose the blue alien."). They were then asked to choose the answer for which they were most confident to put in the treasure check. They were reminded to think carefully about which answer they thought was most correct.

The test phase consisted of two blocks of 8 trials, each consisting of two accuracy questions (and one relative confidence question). Generalization trials were created from 4 untrained noun-image pairs (half animate, half inanimate).

Analyses overview.

Explicit choice data

Explicit choice data was not always normally distributed so any binary data (explicit choices: correct/incorrect) was analysed using generalized linear mixed-effects models (via the *glmer* function in R); continuous data (number of high confidence determiner choices) was analysed using linear mixed-effects models (via the *lmer* function in R); and differences in means (metacognitive bias) were computed using a bootstrap.

All mixed-effects models had the maximal random effect structure that allowed them to converge (Barr, Levy, Scheepers, & Tily, 2013); p-values were calculated by comparing models. Binary fixed effects were contrast coded by hand to center the intercept at 0 (-0.5, 0.5); children's age was centered at 6 years old. To control for the effect of age, all children's models included age as a covariate factor. Item (noun) was included as a random effect for analyses with 16 items; we did not include it for analyses with just 8 items so as to avoid overfitting the model (Babyak, 2004).

Boostraps shuffled each participant's learning condition (Supervised vs. Unsupervised) and then calculated a difference in means. 10 000 permutations were run.

Performance.

Accuracy (correct/incorrect) was measured as a binary dependent variable in a generalized linear mixed-effects model (glmm). Four models were run, one for each age group, and for each phase (Learning and Test). The glmm for the Learning phase included condition (Supervised/Unsupervised) as an independent variable; the one for the Test phase included learning condition and trial type (Vocabulary/Grammar or Grammar/Generalization), as well as the interaction between the two (Condition:Trial type). The glmms for children's data were the same, but they included Literacy (Literate/Pre-literate) as an independent variable, as well as all interactions (Literacy:Condition, Literacy:Trial type, Literacy:Condition:Trial type). Post-hoc glmms were run on literate and pre-literate children's data only when there was a interaction with literacy (e.g., Condition:Literacy) in the absence of main effects. These models allowed us to quantify the driving factors behind the interaction.

For the performance analyses, we had preregistered some t-tests and ANOVAs, but we could not run these analyses because our data did not have a normal distribution. We thus ran mixed-effects models.

Metacognition.

All Learning phase confidence data was included in both the metacognitive ability and bias analyses. Adults were presumed to know how to use a five-point confidence scale. However, we double-checked children's confidence ratings on our scale to ensure that only data from children who understood the scale were included (even though previous research shows that children as young as 3 years old are able to report subjective confidence on a Likert-scale, Lyons & Ghetti, 2011). Any children who always reported being 100% sure (n = 12) on the scale, were considered to have not understood the scale, and were thus excluded from the analysis. The analyses on metacognition in the Learning phase were thus performed on the data from the remaining 37 children: Literate: n = 19 (Supervised condition: n = 9, Unsupervised condition: n = 10); Preliterate: n = 18 (Supervised condition: n = 9, Unsupervised condition: n = 9).

Test phase confidence data was split into 'Different performance' and 'Same performance' trial pairs. The two-alternative forced choice confidence paradigm was performed on trial-pairs: participants made two first-order decisions, and then chose which of the two they felt more confident in. Each first-order decision was thus categorised as 'high confidence' or 'low confidence'. To measure metacognitive ability, the capacity to monitor performance, we analysed the 'Different performance' trial-pairs (one correct, one incorrect). To discern metacognitive bias, the baseline tendency to feel more confident in an item, we analysed the 'Same performance' trials (both correct, or both incorrect). Each data set for adults had a minimum of 414 data points, for a minimum of 41 participants; and for children a minimum of 187 data points, for a minimum of 24 participants (all analyses had 49 participants, except for one that had 24 participants: metacognitive ability in Test Phase 2; we interpret those results with caution).

The relationship between confidence and performance (metacognitive ability) was quantified using a generalized linear mixed-effects model with accuracy (correct/incorrect) as the dependent variable and the subjective confidence report (confidence scale rating or 2-alternative forced-choice confidence rating) as the independent variable. This analysis allowed for a neutral investigation of metacognitive ability, one that did not make any prior assumptions about the nature of the relationship between confidence and performance (in comparison to other standard measures of metacognition, such as the Meta-d' analysis proposed by Maniscalco & Lau, 2012). An assumption-free analysis was particularly important for interpreting children's metacognition, which may not be matured to an adult level. Four models were run. The glmm for the Learning phase included confidence (scale from 0 to 100) and learning condition (Supervised/Unsupervised) as an independent variable, as well as the interaction (Confidence:Condition). A main effect of confidence indicated that there was general metacognition ability, while the interaction provided insight into the differences in metacognition between conditions. The advantage of calculating the interaction (the slope between accuracy and confidence for each condition) was that we did not need to subset blocks in which performance was equal. All four blocks were included in the analysis. The glmm for the Test phase included confidence (high/low) and learning condition as independent variables, as well as the interaction (Confidence:Condition). Trial type (Vocabulary/Grammar or Grammar/Generalization) and the interaction between trial type and condition (Trial type:Condition) were included as covariates to account for variance, but were not analysed as independent variables. Trial type was not independent from confidence (when one trial type was classified as 'high confidence' the other was inevitably 'low confidence'). The glmms for children's data included Literacy as an independent variable and any associated interactions (Literacy:Condition, Literacy: Confidence). The triple interaction Literacy:Confidence:Condition was not included because any effect would have been based on too few data points. Post-hoc glmms were run on literate and pre-literate children's data only when there was a interaction with literacy (e.g., Confidence:Literacy) in the absence of main effects. These models allowed us to quantify the driving factors behind the interaction.

Metacognitive bias was computed using a bootstrap in the Learning phase, and a glmm in the Test phase. Data from the Learning phase blocks where performance was matched was included in the analysis (adults: Blocks 1 and 2 in the Supervised condition and Blocks 3 and 4 in the Unsupervised condition; children: all blocks). The bootstrap determined whether the difference in the mean level of confidence between the two conditions was higher than would be found by chance. A post-hoc bootstrap was run on each group of children (Literate and Pre-literate) as well as on the blocks where exposure was matched (Blocks 1 and 2 in the Supervised condition and Blocks 3 and 4 in the Unsupervised condition). Data from Test phase 'Same

performance' trials were used in the glmm analysis. The number of high confidence grammar trial choices was the dependent variable and condition was the independent variable. Children's glmms included literacy as an independent variable, and the interaction between literacy and condition.

We had preregistered a Meta-d' analysis for metacognitive ability and a t-test for metacognitive bias, however, we had too few data-points for a Meta-d' and our data was not normally distributed to allow for a t-test. We thus ran a mixed-effects model and a bootstrap.

Gaze data

Gaze data were measured using two metrics: gaze focus and gaze distribution. Gaze focus (i.e., looking time to target) has been shown to correspond to performance, while gaze distribution (i.e., switching gaze from a target to foils, and vice versa) to metacognition (Folke et al, 2017). We were thus able to evaluate performance and metacognition using two behavioural measures: gaze and explicit choices. Furthermore, it allowed us to compare gaze and explicit choices, descriptively. Gaze may, for instance, provide a more fine-grained measure of processing.

All analyses on gaze data were performed on the subset of trials that had enough data (the trials during which participants had looked at the screen for at least 50% of the total time, from the moment the two images were presented on the screen to the moment participants were prompted to make an explicit decision) and on the subset of time in a trial during which a decision was possible (the moment when the grammatical element was heard to the moment participants were prompted to make an explicit decision, a span of 5300ms).

Gaze focus.

Gaze focus was evaluated broadly using the 'preferential looking' method which sums overall gaze time to a target during the trial, and on a more fine-grained level via the 'looking-while-listening' method which encodes the evolution of gaze-to-target patterns from moment to moment during the trial. The preferential looking analysis was conducted using a linear mixed-effects model (via the *lmer* function in R) with percentage of looks to the target during each trial as the dependent variable. The looking-while-listening analysis was done using a cluster-based permutation analysis (Maris & Oostenveld, 2007) via the eyetrackingR package in R (Dink & Ferguson, 2016). Gaze data was sampled at 500 Hz, and was down-sampled for the analysis to 50 Hz. The analysis ran a t-test with the arcsine-transformed proportion of looks toward the target image.[21] It then grouped the adjacent time-points with a t-value greater than the predefined threshold of 1.5 into a cluster. 1000 permutations were run.

Gaze distribution.

Gaze distribution was quantified by measuring the number of switches between the target and distractor images. Data were analysed using linear mixed-effects models (via the *lmer* function in R).

Data comparisons

Relationships between two variables (e.g., age and literacy level) were computed using a linear regression (via the *lm* function in R). Data comparisons had not been pre-registered. They were used to investigate confounding factors, post-hoc.

[21] We had preregistered a cluster-based permutation analysis running mixed-effects models. However, given the size of the data, each analysis would have taken multiple days to compute. We thus ran cluster-based permutation analyses running t-tests for the dissertation, but will run the pre-registered analysis for the paper.

All analysis scripts were programmed using R 3.4.4.

References

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122*(3), 292-305.

Babyak, M. A. (2004). What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models, *Psychosomatic Medicine, 66*, 411-421.

Baer, C., Gill, I. K., & Odic, D. (2018). A domain-general sense of confidence in children. *Open Mind*, *2*(2), 86-96.

Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, *115*(23), 6082-6087.

Barbir, M., Babineau, M., Fiévet, A.-C., & Christophe, A. Infants quickly use newly learned grammar to guide acquisition of novel words. (In prep).

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Best, C. A., Yim, H., & Sloutsky, V. M. (2013). The cost of selective attention in category learning: Developmental differences between adults and infants. *Journal of Experimental Child Psychology*, *116*(2), 105-119.

Boldt, A., Blundell, C., & De Martino, B. (2019). Confidence modulates exploration and exploitation in value-based learning. *Neuroscience of Consciousness*, *2019*(1), niz004.

Brunswick, N., Martin, G. N., & Rippon, G. (2012). Early cognitive profiles of emergent readers: A longitudinal study. *Journal of Experimental Child Psychology*, *111*(2), 268-285.

Caffarra, S., & Martin, C. D. (2019). Not all errors are the same: ERP sensitivity to error typicality in foreign accented speech perception. *Cortex*, *116*, 308-320.

Cavanagh, J. F., Wiecki, T. V., Kochar, A., & Frank, M. J. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General*, *143*(4), 1476.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148-153.

Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. S. (2000). Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language*, *42*(4), 465-480.

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological Science*, *27*(6), 848-858.

de Gardelle, V., & Mamassian, P. (2014). Does confidence use a common currency across two visual tasks?. *Psychological Science*, *25*(6), 1286-1288.

De Gardelle, V., Le Corre, F., & Mamassian, P. (2016). Confidence as a common currency between vision and audition. *PLoS One*, *11*(1), e0147901.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*(4), 499-533.

Deng, W. S., & Sloutsky, V. M. (2015). The development of categorization: Effects of classification and inference training on category representation. *Developmental Psychology*, *51*(3), 392.

Destan, N., & Roebers, C. M. (2015). What are the metacognitive costs of young children's overconfidence?. *Metacognition and Learning*, *10*(3), 347-374.

Dink, J. & B. Ferguson, B. (2016). _eyetrackingR_. R package version 0.1.6. Retrieved from http://www.eyetrackingR.com

Fernald, A., & Hurtado, N. (2006). Names in frames: Infants interpret words in sentence frames faster than words in isolation. *Developmental Science*, *9*(3), F33-F40.

Finn, B., & Metcalfe, J. (2014). Overconfidence in children's multi-trial judgments of learning. *Learning and Instruction*, *32*, 1-9.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91.

Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2017). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, *1*(1), 0002.

Gerken, L., & Bollt, A. (2008). Three exemplars allow at least some linguistic generalizations: Implications for generalization mechanisms and constraints. *Language Learning and Development*, *4*(3), 228-248.

Gerken, L., Dawson, C., Chatila, R., & Tenenbaum, J. (2015). Surprise! Infants consider possible bases of generalization for a single input example. *Developmental Science*, *18*(1), 80-89.

Gerken, L., & Knight, S. (2015). Infants generalize from just (the right) four words. *Cognition*, *143*, 187-192.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135-176.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3-55.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, *1*(1), 23-64.

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*(1), 23-45.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109-135.

Goupil, L., Romand-Monnier, M., & Kouider, S. (2016). Infants ask for help when they know they don't know. *Proceedings of the National Academy of Sciences*, *113*(13), 3492-3496.

Hainguerlot, M., Vergnaud, J. C., & De Gardelle, V. (2018). Metacognitive ability predicts learning cue-stimulus associations in the absence of external feedback. *Scientific Reports*, *8*(1), 5602.

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Sciences*, *1*(2), 78-82.

Havron, N., & Arnon, I. (2017a). Minding the gaps: literacy enhances lexical segmentation in children learning to read. *Journal of Child Language*, *44*(6), 1516-1538.

Havron, N., & Arnon, I. (2017b). Reading between the words: The effect of literacy on second language lexical segmentation. *Applied Psycholinguistics*, *38*(1), 127-153.

He, A. X., & Lidz, J. (2017). Verb learning in 14-and 18-month-old English-learning infants. *Language Learning and Development*, *13*(3), 335-356.

Heyman, G. D., Sritanyaratana, L., & Vanderbilt, K. E. (2013). Young children's trust in overtly misleading advice. *Cognitive Science*, *37*(4), 646-667.

Jaswal, V. K., Croft, A. C., Setia, A. R., & Cole, C. A. (2010). Young children have a specific, highly robust bias to trust testimony. *Psychological Science*, *21*(10), 1541-1547.

Jaswal, V. K., Pérez-Edgar, K., Kondrad, R. L., Palmquist, C. M., Cole, C. A., & Cole, C. E. (2014). Can't stop believing: Inhibitory control and resistance to misleading testimony. *Developmental Science*, *17*(6), 965-976.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*(1), 60-99.

Koellinger, P., & Treffers, T. (2015). Joy leads to overconfidence, and a simple countermeasure. *PloS One*, *10*(12), e0143263.

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development*, *76*(6), 1261-1277.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, *10*(3), 294-340.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, *6*(3), 172-187.

Lyons, K. E., & Ghetti, S. (2011). The development of uncertainty monitoring in early childhood. *Child Development*, *82*(6), 1778-1787.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422-430.

Maniscalco, B., Peters, M. A., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception, & Psychophysics*, *78*(3), 923-937.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177-190.

Massoni, S. (2014). Emotion as a boost to metacognition: How worry enhances the quality of confidence. *Consciousness and Cognition*, *29*, 189-198.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101-B111.

McDonough, L., & Mandler, J. M. (1998). Inductive generalization in 9-and 11-month-olds. *Developmental science*, *1*(2), 227-232.

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian probability: From neural origins to behavior. *Neuron*, *88*(1), 78-92.

Morais, J., & Kolinsky, R. (2002). Literacy effects on language and cognition. In L. Bäckman & C. von Hofsten (Eds.), *Psychology at the turn of the millennium, Vol. 1. Cognitive, biological, and health perspectives* (pp. 507-530). Hove, England: Psychology Press/Taylor & Francis (UK).

New, B., & Pallier C. (1999). Lexique (3.8) [Online software]. Retrieved from http:// http://www.lexique.org/

Pallier, C., Devauchelle, A. D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, *108*(6), 2522-2527.

Paul, J. Z., & Grüter, T. (2016). Blocking effects in the learning of Chinese classifiers. *Language Learning*, *66*(4), 972-999.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*(5593), 604-607.

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366.

Rahnev, D., Maniscalco, B., Graves, T., Huang, E., de Lange, F. P., & Lau, H. (2011). Attention induces conservative subjective biases in visual perception. *Nature Neuroscience*, 14(12), 1513–1515.

Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M., & Lau, H. (2015). Confidence leak in perceptual decision making. *Psychological Science*, *26*(11), 1664-1680.

Recht, S., Mamassian, P., & de Gardelle, V. (2019). Temporal attention causes systematic biases in visual confidence. *Scientific Reports*, 9(1), 11622.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology*, *66*(1), 1.

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2016). Sources of developmental change in the efficiency of information search. *Developmental Psychology*, *52*(12), 2159.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*(2), 101-105.

Salles, A., Ais, J., Semelman, M., Sigman, M., & Calero, C. I. (2016). The metacognitive abilities of children and adults. *Cognitive Development*, *40*, 101-110.

Schwiedrzik, C. M., Singer, W., & Melloni, L. (2011). Subjective and objective learning effects dissociate in space and in time. *Proceedings of the National Academy of Sciences*, *108*(11), 4506-4511.

Scofield, J., & Behrend, D. A. (2008). Learning words from reliable and unreliable speakers. *Cognitive Development*, *23*(2), 278-290.

Shi, R., & Melançon, A. (2010). Syntactic categorization in French-learning infants. *Infancy*, *15*(5), 517-533.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558-1568.

Strickland, B. (2017). Language Reflects "Core" Cognition: A New Theory About the Origin of Cross-Linguistic Regularities. *Cognitive Science*, *41*(1), 70-101.

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *Journal of Experimental Child Psychology*, *126*, 395-411.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126-156.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological science*, *18*(5), 414-420.

Yu, C., Yurovsky, D., & Xu, T. (2012). Visual data mining: An exploratory approach to analyzing temporal patterns of eye movements. *Infancy*, *17*(1), 33-60.

van Heugten, M. & Christophe, A. (2014). Learning to accommodate syntactic violations during online speech perception. International Conference on Infant Studies, Berlin, 3-7 Juillet 2014.

van Loon, M., de Bruin, A., Leppink, J., & Roebers, C. (2017). Why are children overconfident? Developmental differences in the implementation of accessibility cues when judging concept learning. *Journal of Experimental Child Psychology*, *158*, 77-94.

Wang, M., Chen, Y., & Schiller, N. O. (2019). Lexico-syntactic features are activated but not selected in bare noun production: Electrophysiological evidence from overt picture naming. *Cortex*, *116*, 294-307.

4.2 Supplemental Information

4.2.1 Supplemental Analyses

Analysis S1. Performance on grammar trials across test phases (pre-registered).

We examined whether performance on associated grammar improved during the Test Phase: there were grammar questions in Test Phase 1 and Test Phase 2. Though vocabulary was sufficient to make correct choices during the Learning phase, it was not for the grammar trials during the Test phase. Participants who had overlooked the grammatical item during the Learning phase could thus refocus their attention on it during the Test phase, and as a result improve performance. It was possible to continue learning the associated grammar, because vocabulary trials were composed of two sentences which both had the correct grammatical item (i.e., *Dol ka pirdale* vs. *Dol ka nuve*). However, the Test phase also contained an element that could have made learning (and recalling) the grammatical element difficult: grammatically incorrect sentences (e.g., *Dol ko pirdale**). Therefore, performance could increase because of attention orientation, decrease because of incorrect evidence, or just stay the same across test phases.

Adults.
A generalized linear mixed-effects model revealed that adults' performance did not improve significantly between the two test phases: no main effect of Test Phase; $\beta = -0.1$, SE = 0.26; model comparison: $\chi^2(1) = 0.47$, $p = 0.5$ (Fig. S2A). Furthermore, performance between the two test phases was not modulated by learning condition: no interaction of Condition and Test Phase; $\beta = 0.07$, SE = 0.29; model comparison: $\chi^2(1) = 0.05$, $p = 0.82$. Adults' performance on grammatical elements thus did not significantly improve or diminish between test phrases. These results however do not allow us to conclude that a reorientation of attention to grammar would not have improved performance, had the first block of the Test phase been followed by a short Learning phase for instance, or that incorrect sentences would not have affected performance in the long run.

Children.
A generalized linear mixed-effects model showed that there was a trend for children's performance to dwindle across phases, but it did not reach significance: no main effect of Test Phase; $\beta = 0.29$, SE = 0.15; model comparison: $\chi^2(1) = 3.55$, $p = 0.06$ (Fig. S2B). Performance between the two test phases did not significantly differ depending on learning condition: no interaction of Condition and Test Phase; $\beta = -0.44$, SE = 0.31; model comparison: $\chi^2(1) = 2.1$, $p = 0.15$. Though there was a tendency for performance to decrease, the tendency may simply have been caused by fatigue: the last test phase began around 35 minutes into the experiment.

Summary.
Together, our results show that performance on grammatical items remained relatively constant over the two test phases. We cannot however rule out an influence of attention to the grammatical item, of incorrect sentences, or more generally of fatigue.

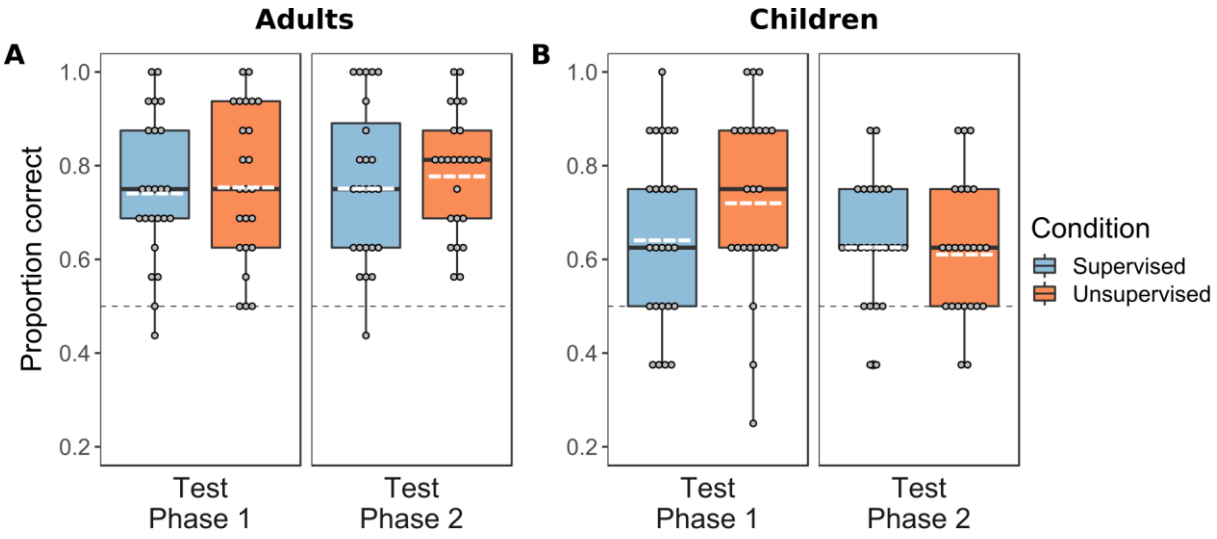


Fig. S2. Performance on grammatical elements across test phases. Proportion of correct answers on grammar trials in Test Phase 1 and 2. Supervised Condition in blue, Unsupervised Condition in orange. Dots indicate participants. Dashed white line indicates mean. Upper and lower regions of the box indicate the first and third quartiles (25^{th} to 75^{th} percentiles). The upper whisker represents the third quartile up to the 1.5 interquartile smallest value, while the lower whisker the 1.5 interquartile smallest value to the first quartile. **(A) Adults. (B) Children.** Performance does not significantly differ across phases for adults or children.

Analysis S2. Informativity: Use of grammatical element in the Learning Phase (pre-registered).

As an index of grammar processing, we compared participants' performance (explicit choice accuracy) and gaze deployment (looking-while-listening) on trials in which the grammatical element was informative (one animate and one inanimate image) and when it was uninformative (two animate or two inanimate images). We hypothesized that if participants were initially using the grammatical element to guide word learning, then we may observe different patterns of performance or gaze orientation on informative trials (higher initial accuracy and faster initial gaze orientation) than uninformative trials (lower initial accuracy and slower initial gaze orientation). We expected participants in the Unsupervised Condition to show a greater effect of informativity then those in the Supervised Condition, who had had their attention focused on the vocabulary items.

Explicit choice data.

Adults.
A generalized linear mixed-effects model showed that performance on informative trials (M = 0.82, SD = 0.39) was not higher than on uninformative trials (M = 0.82, SD = 0.38): [no main effect of informativity: β = -0.012, SE = 0.12; model comparison: $\chi^2(1) = 0.012$, p = 0.91]. Furthermore, learning condition did not modulate the effect (or lack thereof) of informativity: [no interaction between Informativity and Condition: β = 0.07, SE = 0.23; model comparison: $\chi^2(1) = 0.097$, p = 0.76]. Though an effect of informativity on performance would have revealed that participants were using the grammatical element to guide acquisition, a lack of effect does not reveal the inverse. Our measure may not have detected the use of the grammatical element

to guide learning because it was too coarse-grained; or, the grammatical element may not have been the only or predominant cue used to guide learning, diluting a potential effect.

Children.
A generalized linear mixed-effects model revealed no significant difference between performance on informative trials (M = 0.54, SD = 0.5) and uninformative trials (M = 0.58, SD = 0.49): [no main effect of Informativity: $\beta = -0.12$, SE = 0.11; model comparison: $\chi^2(1) = 1.3$, $p = 0.25$]. Moreover, learning condition did not interact with informativity [no interaction between Informativity and Condition: $\beta = 0.45$, SE = 0.5; model comparison: $\chi^2(1) = 0.45$, $p = 0.5$]. There was however a cross-over interaction between Informativity and Literacy, informativity seemed to have an inverse effect on literate and pre-literate children: [$\beta = -0.57$, SE = 0.21; model comparison: $\chi^2(1) = 7.35$, $p = 0.007$]. We refrained from running additional analyses on each literacy group to determine the driving factor behind this interaction, because this would have left few data-points per cell. We will perform these additional analyses when we have run the second group of children, and thus when we have a total of 48 literate and 48 pre-literate.

Gaze data.

Adults.
Cluster-based permutation analyses revealed no significant differences in gaze deployment overall, nor for each condition analysed separately. These results reflect the explicit choice results, and again do not tell us that learners were not using the grammatical elements. Our analysis may still have been too coarse-grained or learners may have been using a set of clues (grammar and vocabulary) which may have tempered an observable effect.

Children.
A cluster-based permutation analysis on children's data also revealed no significant clusters between informative and uninformative trials. This result held also for both learning conditions, supervised and unsupervised, and for both groups of children, literate and pre-literate. These data thus reflect explicit choice results, and adult results.

Summary.
Overall, our experimental design did not allow for an easy extrapolation of an effect grammatical items may have had on learning from the effect that vocabulary may have had on learning. These null results do not tell us that learners were not using the grammatical elements, and could have been caused by task related or measure related noise.

Analysis S3. Gaze deployment during learning (pre-registered).
As a more fine-grained index of processing during acquisition, we investigated whether moment-to-moment gaze deployment during the trial differed between learning conditions (looking-while-listening procedure). This measure allowed us to map any differences in online processing, and to compare them descriptively with results from our other two metrics: overall looking time and explicit choice.

Adults.
Cluster-based permutation analyses revealed that participants in the Supervised Condition deployed their gaze differently than those in the Unsupervised Condition, in each Learning

phase block. Participants in the Supervised Condition looked more to the target image than those in the Unsupervised Condition at each significant time-window:

Block 1:

520-1940ms ($p = 0.018$)

2880-5300ms ($p < 0.001$)

Block 2:

2600-5300ms ($p = 0.002$)

Block 3:

1040-5300ms ($p < 0.001$)

Block 4:

3560-4780ms ($p = 0.036$)

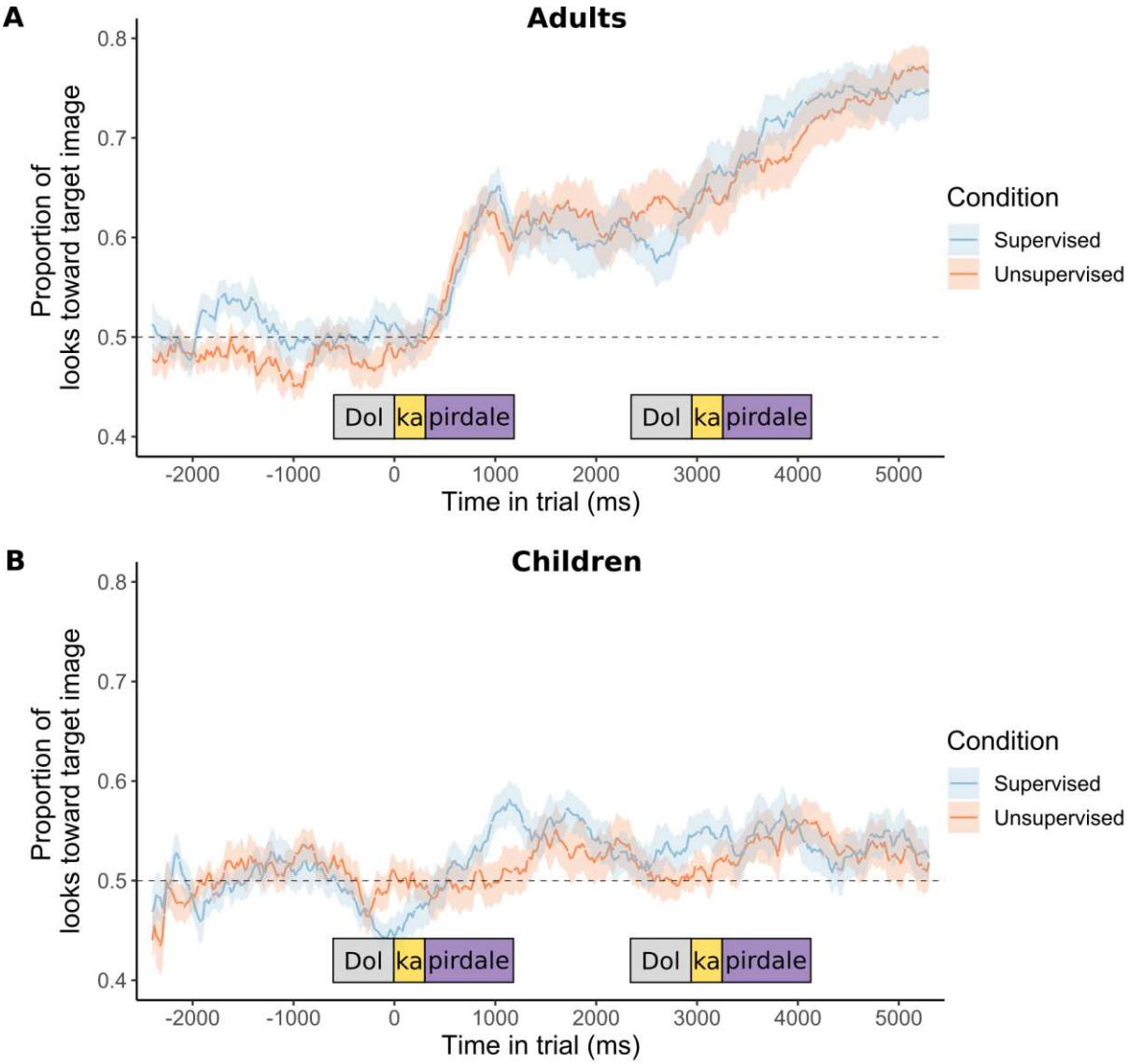


Fig. S3. Gaze time-course when explicit choice performance is matched (Learning phase). Proportion of looks toward the target image at each point in time during the trial when explicit choice performance was matched (Adults: Blocks 1-2 for the Supervised Condition and Blocks 3-4 for the Unsupervised Condition; Children: All blocks). Supervised Condition in blue, Unsupervised Condition in orange. Dark lines represent mean across participants, and light shading the SEM (95% confidence intervals of the mean). Dotted line represents chance. **(A) Adults. (B) Children.** There were no significant differences in gaze deployment between the two conditions when explicit choice performance was matched.

We then controlled for the amount exposure by subsetting the blocks where the amount of exposure was matched (Blocks 1-2 for the Supervised condition, and Blocks 3-4 for the Unsupervised Condition). The cluster-permutation analysis revealed no significant clusters, demonstrating that the deployment of gaze was comparable even at a moment-to-moment resolution (Fig. S3A). Therefore, two blocks of supervised and unsupervised learning had a comparable effect on gaze deployment, indicating that processing may be quite similar across learning conditions.

Children.
Cluster-based permutation analyses revealed that children did not deploy their gaze differently in the two conditions (Fig. S3B). This was also true when the analysis was run on the subset of just literate or just pre-literate children. The pattern of gaze focus over the course of a trial, thus, corroborated the total gaze focus results: supervision did not appear aid learners at orienting their gaze, and literacy did not seem to affect gaze patterns.

Summary.
Moment-to-moment gaze deployment appears to reflect both overall looks to a target and explicit choice. When explicit choice performance is matched, so is gaze deployment. We may have expected that participants in the unsupervised condition would perhaps interpret the determiner and thus show an early onset effect. These results point to the possibility that participants in both conditions were using the same cognitive processes to learn, despite different teaching methods.

Analysis S4. Metrics for testing performance in children (not pre-registered).

At first glance, choosing the referent of a sentence (referent choice) and choosing which sentence best corresponds to a referent (sentence choice) appear to tap into the same knowledge: sentence-referent mapping. Nevertheless, the former represents an act we do often, we scope out the referents of people's utterances (e.g., "Look at the cute kitty!"), but the latter perhaps less so, at least explicitly. We are rarely asked "is 'Look at the kitty' or 'Look at the doggy' better?", even though we may make such decision implicitly when we speak. This difference in tasks ought not to affect performance of adults, but children may find one more intuitive than the other. To make sure that we were using a task adapted for children and one that taps into similar knowledge, we predicted performance at test (sentence choice) based on performance during training (referent choice) with a linear regression. The analysis revealed that a child's level of performance during the training phase predicted performance during the test, despite task differences: [$F(1, 47) = 14.79$, $p < 0.001$, $R^2 = 0.22$].

Analysis S5. Interpreting element-specific metacognitive ability (descriptive analysis).

Adults in our experiment appeared to have metacognitive ability in Test Phase 1 (Vocabulary and Grammar questions) but did not appear to have metacognitive ability in Test Phase 2 (Grammar and Generalization questions). Our design was specifically a comparative one to investigate how the mind interprets messy metacognitive situations. One could however wonder for which specific items participants had had metacognitive ability.

Together, metacognitive ability and bias results suggest that adults likely have metacognition for vocabulary and grammar but not for generalization. We base our reasoning on the following

premise: if there is no bias, then having metacognitive ability for just one trial type may be enough to observe an effect of metacognition.

Test Phase 1.
If participants had metacognitive ability for vocabulary but not grammar and no bias, they would have been able to track performance on vocabulary trials but not on grammar trails.

Hypothetical case: Metacognitive ability for vocabulary but not grammar

Vocabulary trial correct, grammar trial incorrect:
Participants can track performance on vocabulary trials (I responded correctly), but not grammar trials (I guessed). When they have to compare the two, they will thus (correctly) feel more confident in the vocabulary trial.

Vocabulary trial incorrect, grammar trial correct:
Participants can track performance on vocabulary trials (I do not know, I guessed), but not on grammar trials (I guessed). When they compare the two, they feel just as unconfident in both. To choose the one they feel more confident in, they can choose at random or use other information to guide their decision (for example, confidence from preceding trials).
Option 1: If they choose at random, then we would observe metacognitive ability for half the trials (vocab correct/grammar incorrect) but not the other (vocab incorrect/grammar correct). There may still be an overall effect of metacognitive ability (but a little more noisy), driven purely by metacognitive ability for vocabulary.
Option 2: If they choose based on previous trial confidence (what items have I been good at in the past), they would choose (incorrectly) the vocabulary trial as a 'high confidence' trial. Therefore, we will not observe an effect of metacognitive ability, because participants would correctly track performance on half the trials (vocab correct/grammar incorrect), but incorrectly track of performance on the other half (vocab incorrect/grammar correct). In other words, they would always chose the vocabulary trial as the 'high confidence trial'.
Metacognitive bias results indicate that adults were biased in favour of vocabulary when trial-pair performance was the same (vocab correct/grammar correct or vocab incorrect/grammar incorrect).
Therefore, to observe comparative metacognitive ability in Test Phase 1, as we did, participants would have needed to have some metacognitive ability for grammar. It is as such likely that adults had metacognitive ability for both vocabulary and associated grammar.

Let us now look at why then we do not observe an effect of metacognitive ability in Test Phase 2.

Test Phase 2.
If participants had metacognitive ability for grammar but not generalization and no bias, they will be able to track performance on grammar trials but not on generalization trails.

Hypothetical case: Metacognitive ability for grammar but not generalization

Grammar trial correct, generalization trial incorrect:
Participants can track performance on grammar trials (I responded correctly), but not generalization trials (I guessed). When they have to compare the two, they will thus (correctly) feel more confident in the grammar trial.

Grammar trial incorrect, generalization trial correct:
Participants can track performance on grammar trials (I do not know, I guessed), but not on generalization trials (I guessed). When they compare the two, they feel just as unconfident in both. To choose the one they feel more confident in, they again choose at random or use other information to guide their decision (for example, confidence from preceding trials).
Option 1: If they choose at random, then we would observe metacognitive ability for half the trials (grammar correct/generalization incorrect) but not the other (grammar incorrect/generalization correct). There may still be an comparative effect of metacognitive ability if metacognitive ability on grammar trials is high enough (though perhaps this effect may be more noisy), driven purely by metacognitive ability for grammar.
Option 2: If they choose based on previous trial confidence (what items have I been good at in the past), they would choose (incorrectly) the grammar trial as a 'high confidence' trial. Therefore, we will not observe an effect of metacognitive ability, because participants would correctly track performance on half the trials (grammar correct/generalization incorrect), but incorrectly track of performance on the other half (grammar incorrect/generalization correct). In other words, they would always chose the grammar trial as the 'high confidence trial'.
Metacognitive bias results indicate that adults were biased in favour of grammar when trial-pair performance was the same (vocab correct/grammar correct or vocab incorrect/grammar incorrect).
Therefore, the lack of metacognitive ability in Test Phase 2, was thus likely caused by a lack of metacognitive ability for generalization trials coupled with a bias for grammar trials.

Though this is the intuitive explanation, it is also possible that a bias could be so strong that it could even occult emergent metacognitive ability. For instance, if we had tested the metacognitive ability of generalization trials alone (not comparatively as we had, but perhaps with a confidence scale), we may have observed some metacognitive ability. Further research is needed to determine whether there is a cloaking of metacognitive ability for new items when they are compared with familiar items, and the effect this has on learning.

Analysis S6. Interpreting the source of metacognitive bias (descriptive analysis).

Our results revealed that, when trial-pair performance is the same (both trials correct or both incorrect), adults have a metacognitive bias for vocabulary over grammar, and grammar over generalization. These results are consistent with a Bayesian approach to decision making: the posterior probability of being correct is the product of the evidence for the current trial (likelihood) and the prior probability (in our study, any aggregated evidence). Accordingly, when participants decide the trial for which they are most confident, they may weigh evidence for the current trial (I feel sure about the vocabulary trial and I feel sure about the grammar trial) and the prior probability (I have felt sure about vocabulary trials often, more so than grammar trials). Previous experiments have found that adults include prior probabilities into confidence decisions, and this influence of trial-external factors have been referred to as the 'confidence leak' (Rahnev et al, 2015). In our experiment, priors could have included first-order evidence in favour of one decision or second-order evidence of overall confidence for one trial type. Our data do not allow us to distinguish between the two, but this would be an interesting line of further research: determining the priors adults use in confidence judgements in messy metacognitive situations.

4.2.2 Supplemental Figure

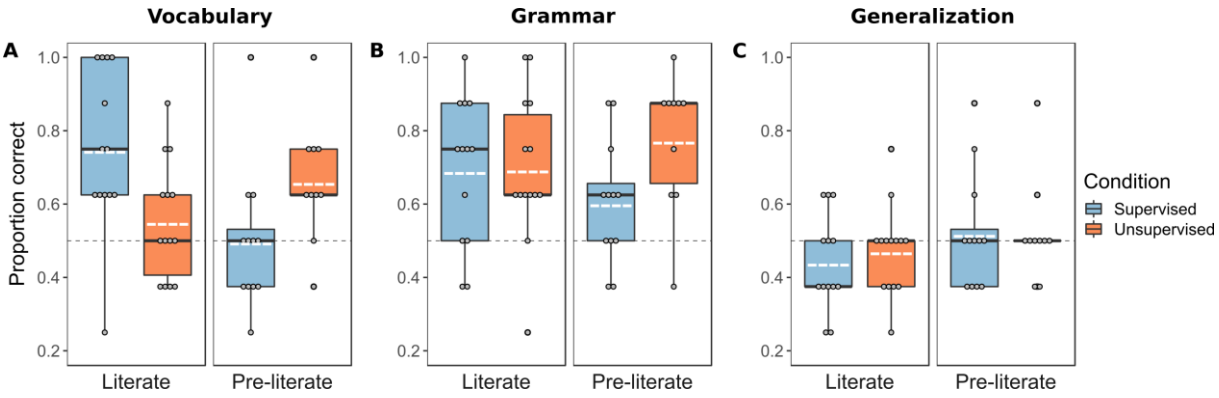


Fig. S4. Test phase results by literacy for each trial type. Mean proportion of correct answers for **(A) Vocabulary, (B) Grammar,** and **(C) Generalization.** There is a significant interaction between learning condition and literacy for Vocabulary trials ($p < 0.001$), but no significant triple interaction between learning condition, literacy and trial type in Test Phase 1. We hope the doubled sample size will allow us to get a clearer image of how the learning condition affects each element.

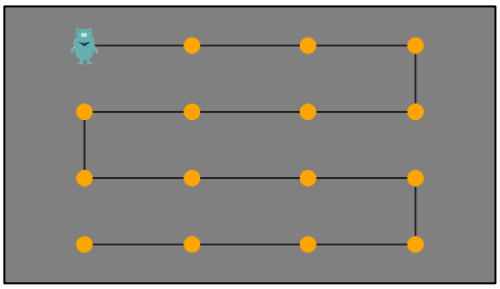
4.2.3 Annex

Annex S1. Experiment 1 (Adults): Instructions.

Adults saw one of two versions of the instructions: the one for the Supervised Condition or the one for the Unsupervised Condition. Here, they are presented on the same page to save space.

(English)
Let's learn Martian!
Instructions

With our dear friends, the Martians, we have created a game to learn the Martian language. In this game, you will collect gold coins (virtual, of course!) at each trial. At the end of each game level, these gold coins will allow you to access a surprise video. The video is not part of the learning game, it is a reward! Sit back and enjoy. At the beginning of each game level you will see the following path. To begin the level click the spacebar.



The game is divided into three parts: clues, training and test.

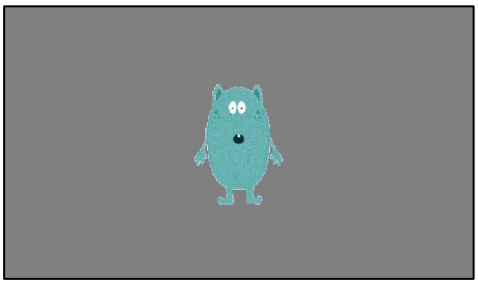
Before each trial, you will see this fancy circle. You have to look at it to begin the trial. If you look at it, but the trial does not start, come find us just outside the booth.

(Unsupervised) Part 1 – Clues:

You will get some clues that will help you during training. You will see one image. Look at it carefully. Once you have seen what it is, click the spacebar to continue. Note: the image will be displayed for just a short period of time.

(Supervised) Part 1 – Clues:

You will get some clues that will help you during training. A Martian will translate some Martian words into French (yeah, she's a Martian-French interpreter).



« sslkjghjl » means « computer ».

Listen to what she says carefully. Do not interrupt her.

Your task is to repeat the phrase she said, clearly into the microphone. If you are not sure about how to pronounce the Martian word, it's normal, you're a beginner learner, just try your best.

Once you have repeated the phrase, move on to the next trial by clicking the spacebar.

Part 2 – Training:

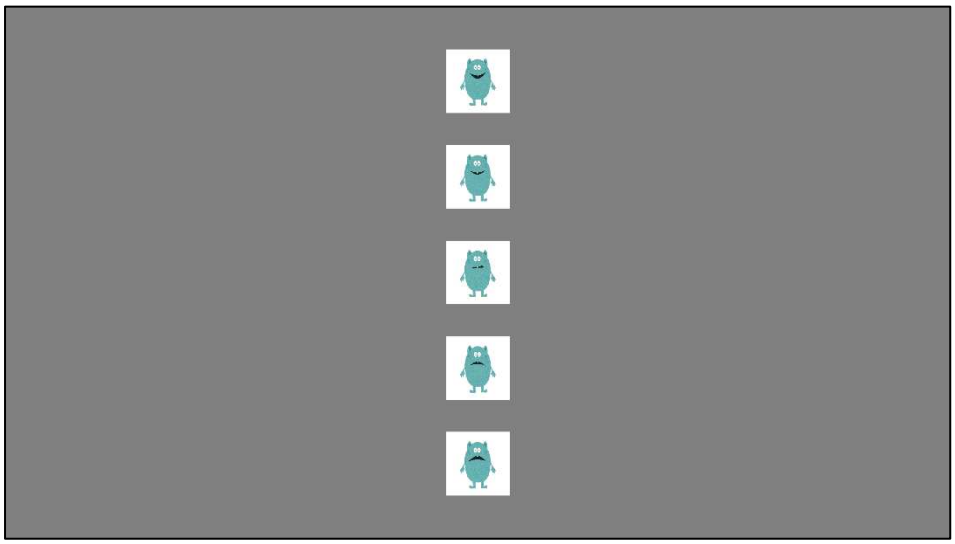
You will now hear an utterance in Martian, and your task is to find out which image it corresponds to. Listen carefully to the utterance and look carefully at the image. There are many many training trials, do not stress!

"smlkhfdlksqhfdldsq"

During this part of the experiment, we will record where you are looking on the screen with a camera. Please look at the screen and do not block your eyes with your hands.

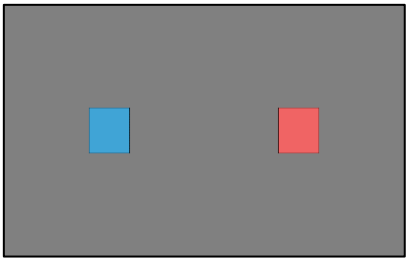
When the two images turn orange, you can choose which image corresponds to that utterance, by clicking ‘arrow left’ (for the left image) or ‘arrow right’ (for the right image). You cannot indicate your choice before the images turn orange. Your response will automatically move you on to the next trial.

From time to time, a trial will be followed by a confidence rating. We want to know how confident you are in your response for the trial you just did.



You will see 5 monsters each with an expression ranging form very happy (=I am very confident) to very sad (=I guessed). Use the mouse to click on the monster that corresponds to your level of confidence. The response will move you on directly the next trial.

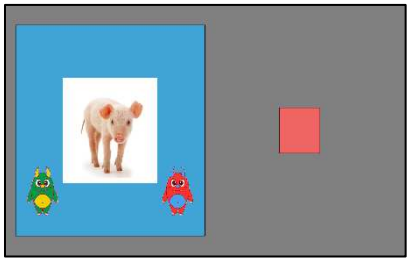
Part 3 – Test:



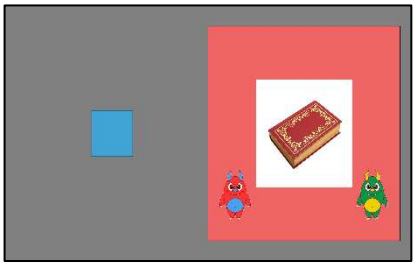
And here we are, you’re ready! Yes, we’re sure!

At each trial, you will see two boxes, one on the left and another on the right.

Inside each box, there is an image. On box will open, and you will see an image.

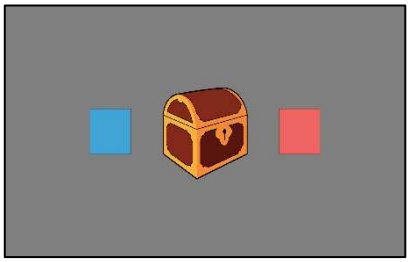


Next, two Martians will appear. One by one, the Martians will rise up to the image and produce an utterance in Martian. Your task is to choose the monster who said the utterance that corresponds to the image, by clicking ‘arrow left’ (for the left Martian) or ‘arrow right’ (for the right Martian). You can respond when both Martians have moved down to their original positions.



The box will close, and the other will open. The task is the same.

Once the second box is closed, a treasure chest will open. You now have to choose which answer you want to keep for points by clicking ‘arrow left’ (for the left box) or ‘arrow right’ (for the right box). The response will move you on to the next trial.

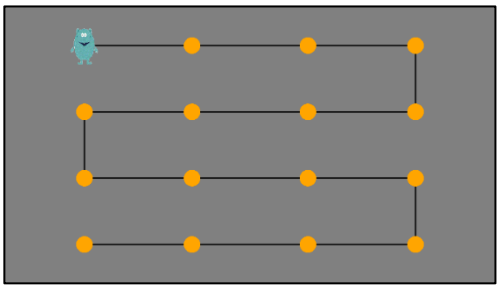


The end:

When you see a duck cartwheeling (yep, that’s the kind of stuff that goes down on Mars!), then you have finished the game. Come find us so we can explain the experiment!

(Français)
Apprenons la langue martienne !
Instructions

Avec nos amis les martiens, nous avons créé un jeu pour apprendre la langue martienne. Dans ce jeu, vous allez cumuler des pièces d’or (virtuelles bien sûr !) à chaque essai. A la fin de chaque niveau, ces pièces d’or vont vous permettre d’ouvrir une vidéo surprise. Cette vidéo ne fait pas partie du jeu, c’est un cadeau ! Reposez-vous, et profitez bien ! Au début de chaque niveau vous allez voir le plan ci-dessous. Pour démarrer le niveau cliquez sur la touche espace.



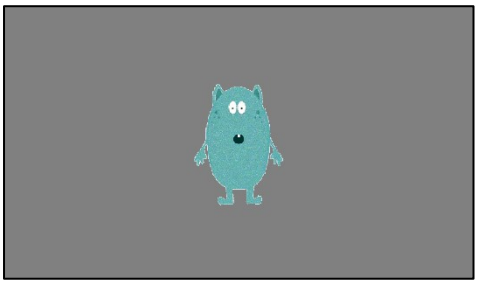
Le jeu est composé de trois parties : des indices, un entraînement, et un test.

Avant chaque essai, vous allez voir un joli cercle. Il faut le fixer pour démarrer l’essai. Si vous le fixez mais l’essai ne démarre pas, venez nous chercher.

(Non-supervisé) Partie 1 – Indices:

Vous allez recevoir des indices qui vont vous aider pendant l'entraînement. Vous allez voir une image. Regardez la bien. Une fois que vous avez bien vu ce que c'est, cliquez sur la touche « espace » pour passer à l'essai suivant. Attention : l'image reste sur l'écran durant très peu de temps.

(Supervisé) Partie 1 – Indices:
Vous allez recevoir des indices qui vont vous aider pendant l'entraînement. Une martienne va vous donner des traductions des mots martiens en français (eh oui, elle est interprète martien-français).



« sslkjghjl » ça veut dire « ordinateur ».

Il faut bien écouter la phrase ! Ne l'interrompez pas.

Votre tâche est de répéter la phrase qu'elle a dite, clairement dans le micro. Si vous n'êtes pas sûr.e de la meilleure façon de prononcer le mot en martien, c'est normal, vous êtes débutant en martien, faites de votre mieux.

Une fois que vous avez répété la phrase, passez à l'essai suivant avec la touche « espace ».

Partie 2 – Entraînement :

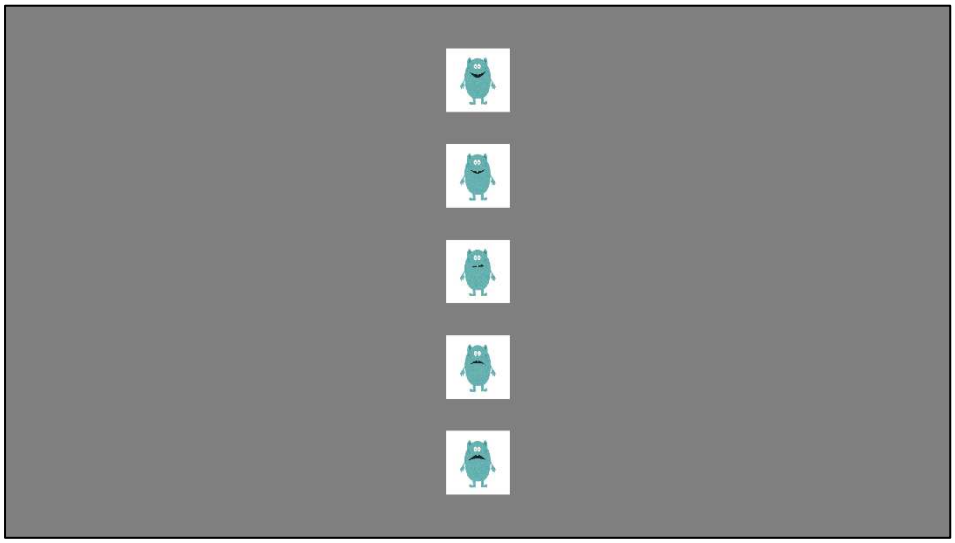
Vous allez maintenant entendre une phrase en martien, et votre tâche est de découvrir à quelle image elle correspond. Au début, la tâche paraîtra un peu difficile, vous ne saurez peut-être pas la réponse, c'est normal, vous êtes débutant. Prêtez bien attention à la phrase et aux images. Il y aura beaucoup d'essais pour apprendre, ne vous stressez pas !

"smlkhfdlksqhfdldsq"

Pendant cette partie, nous allons enregistrer votre regard avec une caméra, alors regardez bien l'écran et ne bloquez pas vos yeux avec vos mains.

Une fois que vous voyez que les deux images sont devenues oranges, vous pouvez indiquer quelle image correspond à cette phrase, en cliquant sur la flèche gauche (=image de gauche) ou la flèche droite (=image de droite). Vous ne pouvez pas indiquer une réponse avant que les images quittent leur couleur orange. Votre réponse vous fera automatiquement passer à l'essai suivant.

De temps en temps, un essai sera suivi d'une évaluation de confiance. Nous voulons savoir à quel point vous êtes confiants dans votre réponse à l'essai que vous venez de terminer.

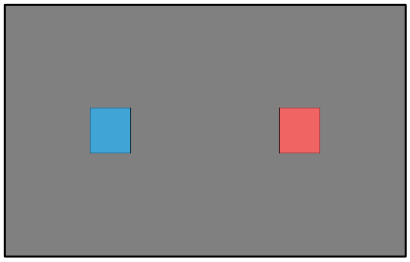


Vous allez voir 5 monstres qui ont des expressions allant de très heureux (= je suis très confiant) à très triste (= j'ai deviné au hasard). Utilisez la souris pour cliquer sur le monstre qui correspond à votre niveau de confiance dans votre réponse. La réponse va vous faire passer directement à l'essai suivant.

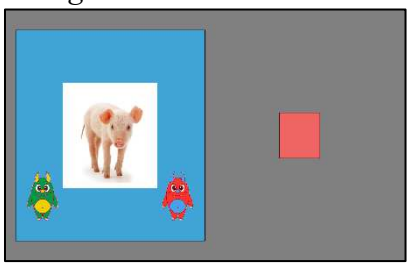
Partie 3 – Test :

Et voilà, vous êtes prêt.e ! Oui, oui, nous sommes sûrs.

A chaque essai, vous allez voir deux boites, une à gauche et l'autre à droite.

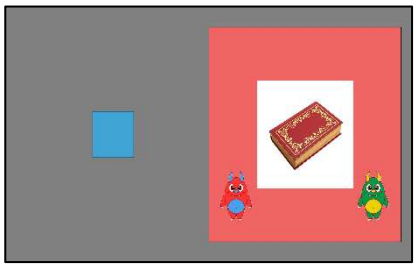


A l'intérieur de chaque boite, il y a une image. Une boite va s'ouvrir, et vous allez voir l'image.

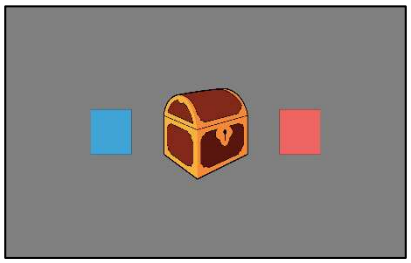


Ensuite, deux monstres vont apparaître. Un par un, les monstres vont s'approcher de l'image et dire une phrase en martien. Votre tâche est de choisir le monstre qui a dit la phrase qui correspond à l'image, en cliquant sur la flèche gauche (= monstre de gauche) ou la flèche droite (= monstre de droite). Vous pouvez indiquer votre réponse dès que les deux monstres sont revenus à leur position initiale.

Cette boite va se fermer, et l'autre va s'ouvrir. La tâche est la même.



Une fois que la deuxième boite s'est fermée, une malle au trésor va apparaître. Il faut choisir quelle réponse vous voulez garder pour des points : la boite de gauche (=flèche gauche) ou la boite de droite (=flèche droite). Cette réponse vous fera passer directement à l'essai suivant.



La fin :

Lorsque vous verrez un canard en plastique faire la roue (et oui, ça se passe aussi sur Mars !), le jeu est fini. Venez nous chercher pour que nous vous expliquions notre expérience !

Annex S2. Randomization constraints.

Training.

I. To avoid bias:
1. Half the targets are on the left and half on the right
2. Maximum 4 targets on the same side in a row
3. Maximum 4 targets of the same animacy in a row (e.g. cat, dog, fish, horse)
4. Maximum 4 animate images on the same side of the screen in a row

II. To prevent the animate/inanimate distinction to be evident:
1. Half the trials with one inanimate and one animate image (e.g., book and horse), one fourth the trials with two animates, and one fourth the trials with two inanimates.

III. To prevent the task from being trivially easy:
1. Neither the target nor foil can appear on consecutive trials

2. The same target-foil pair cannot appear twice (e.g., if cat-left image and horse-right image appear with cat as a target, horse-left image and cat-right image cannot appear later with cat or horse as a target).

IV. Confidence trial choice:
1. For adults: confidence trials will be randomly distributed such that the first is at the 4th, 5th, or 6th trial, the second is at the 9th, 10th, or 11th trial, and the third is at the 14th, 15th, or 16th trial
2. For children: confidence trials will be randomly distributed such that the first one is at the 2nd, 3rd, or 4th trial and the second at the 6th, 7th or 8th trial.

Test.

I. Vocabulary questions:
1. Maximum 4 of the same animacy in a row
2. The same target or distractor cannot appear in a row (as a target or distractor)
3. Each noun is a target once and a distractor once.

II. Grammar questions:
1. Maximum 4 of the same animacy in a row

III. Generalization questions:
1. Maximum 4 of the same animacy in a row.
2. The same target will not be able to appear twice in a row.

IV. Question pairings:
1. Half in which the image in one question's target is animate and the other inanimate, one fourth in which both are inanimate, and one fourth both animate
2. The same target cannot be the target or distractor in both questions in a pair
3. Half of one question kind (e.g., Determiner) on one side, half on the other.

V. Overarching constraints:
1. Maximum 4 of the same kind of questions (e.g., Determiner) on one side (left or right on the screen) in a row
2. Maximum 4 animate or inanimate images on one screen side in a row
3. Maximum 4 animate or inanimate targets of the same kind of question in a row
4. No two consecutive targets or distractors in a row

VI. Counterbalancing:
1. The colour of the boxes (half red and half blue per side)
2. The question kind in each box (half of one kind and half of another in each box)
3. The monster that says the correct answer in each trial (half left and half right)
4. The side the monster who says the right answer in each box is on (1/4 in which the left monster in both says the right answer, 1/4 in which the right monster in both says the right answer, 1/4 in which the left monster says the right answer in the left box and the right in the right box, and 1/4 in which the right monster says the right answer in the left box and the left in the right box)
5. Half of the trials the left monster is the red one and half it is the green one
6. The colour of the monster who is on each side in each box (1/4 in which the red monster is on the right in both boxes, 1/4 in which the green monster is on the right in both box, 1/4 in which the red monster is on the left in the left box and the green on the left in right box, and 1/4

in which the green monster is on the left in the left box and the red is on the left in the right box)
7. The colour of the monster that says the right answer in each box (1/4 in which the red monster in both says the right answer, 1/4 in which the green monster in both says the right answer, 1/4 in which the red monster says the right answer in the left box and the green in the right box, and 1/4 in which the green monster says the right answer in the left box and the green in the right box)

Annex S3. Experiment 1 (Adults): Post-experiment questionnaire.

Questionnaire (English)

I.
Languages

1. Languages heard at birth

2. Languages spoken or understood (level and date at which you started learning the language)

3. Do you like learning languages?

II.
Experiment
1. a) If Supervised, did you memorise the word translations? How many?
1. b) If Unsupervised, what did you do during the first task (images without sound)? Did that help you during the experiment?

2. During training (two images and one utterance), how did you go about associating images and Martian utterances?

3. During the test, did you notice utterances that weren't present during training? How did you go about responding?

Questionnaire (Français)

I.
Langues
1. Langues entendues à la naissance

2. Langues parlées (niveau + date d'acquisition)

3. Est-ce que vous aimez apprendre de nouvelles langues ?

II.
Expérience
1. a) Si Supervisé, avez-vous réussi à mémoriser les mots ? Combien ?

1. b) Si Non-supervisé, qu'est-ce que vous avez fait pendant la première tâche ? Est-ce que cela vous a aidé pendant l'expérience ?

2. Pendant l'entraînement, comment avez-vous fait pour associer les images avec le martien ?

3. Pendant le test, avez-vous remarqué des mots qui ne faisaient pas parti de l'entraînement ? Comment avez-vous fait pour répondre ?

5. GENERAL DISCUSSION

Over a series of chapters, we have explored what it is to know, from the subjective perspective of the learner. In the first chapter, we investigated how comparably little evidence can fuel acquisition of grammatical contexts in an ecologically valid setting. Our results revealed that just a handful of words can spur grammatical context acquisition in infants. Next, in the second chapter, we probed whether the same set of evidence sparks generalization across development, from infancy into childhood and through to adulthood. Our data showed that infants and adults generalize a novel grammatical context to new words, but pre-school children do not. We interpreted our results in the context of extant literature, proposing that the generalization tipping point may be adjustable. Finally, in the third chapter, we examined why some learning methods, notably supervised ones, may be preferred over others, across age and reading skill. We found that supervision universally boosts confidence, but had variable effects on performance, ranging from comparable with unsupervised learning to downright detrimental. We advance the conclusion that scanty evidence can kindle knowledge, that the threshold for what counts as knowledge may be mutable, and that learners may covet knowledge-state information, seeking indubitable knowledge. Broadly, we propose that a learner may aim for knowledge from the point at which she finds herself in the learning process.

The information that a mind has about the external world may count as knowledge when it attains a mutable standard, rather than an immutable ideal. This standard may be a criterion akin to the decision criterion used for decision making. To make a decision, an individual needs to weigh the evidence she has in favour of a decision in respect to a criterion, the decision criterion (e.g., Bogacz, 2007). For instance, an individual who is going to a bakery may only remember approximately where the bakery is: somewhere to the left after the traffic light. At this point, the individual may not have enough evidence in favour of the decision to turn left. A quick search on a map may confirm the individual's intuition, adding sufficient evidence in favour of the decision to turn left. The individual thus had a criterion at which she would be able to make the decision to turn left. This criterion has been shown to be adjustable in real-time (Cassey, Evens, Bogacz, Marshall, & Ludwig, 2013). In other words, the bakery-seeking individual may adjust her decision criterion as she is walking down the street. Knowledge may likewise rely on the same basic principle: enough information in respect to a criterion.

Knowledge however is not reducible to decision. An individual may have enough evidence for a decision without knowing that she has enough evidence (lack of metacognitive ability) or without having enough evidence for knowledge specifically (e.g., having some lingering uncertainty). A native English speaker learning French, for instance, may be able to correctly classify the grammatical gender of the word *banane* (banana) as a feminine, 65% of the time. Though the learner may have above chance accuracy, she may not feel she knows the grammatical gender of *banane*. When speaking about a banana, she may waver between attributing it the feminine determiner *la* or the masculine determiner *le*. Intuitively, we expect high performance and high confidence to index a state of knowledge. This conception of knowledge as indubitable may, however, be an ideal, that for which individuals aim. Rather, knowledge may have a mutable threshold, and this threshold may depend on how pressed one

is to know, the 'urge to know' that we proposed in the second chapter. Broadly, knowledge may represent a stabilized state[22] of 'knowing enough', in the learning process.

One of the consequences of a mutable standard of knowledge is that everything need not be learned to the same level. For instance, a native English speaker learning French may one day want to point out something about a banana. There will be elements of the sentence that are more important to know precisely than others. Ideally, the learner of French would want to utter '*La banane*'. However, being a learner, she may make a mistake. If she utters the incorrect noun (e.g., *La canapé*), her interlocutor will have trouble figuring out that she wants to talk about a banana; in contrast, if she utters the incorrect determiner (e.g., *Le banane*), her interlocutor likely will not struggle to determine the referent.[23] The learner could adjust her knowledge thresholds to different levels for determiners (low threshold, precision not vital) and nouns (high threshold, precision vital). A low knowledge threshold allows the learner to stabilize any extant information before being sure it represents how the world is, and this can, perhaps counterintuitively, be beneficial for acquisition. A low threshold allows the learner to begin using knowledge to fuel learning earlier and it allows the learner to allocate cognitive resources efficiently, spending more effort acquiring precise knowledge for some elements of the world and less for others. For an adult learning a second language, it may thus be more important to acquire nouns than grammatical contexts. This may be in part because nouns pick out their referents more reliably than grammatical contexts and in part because adults view words as isolable, meaning bearing elements. In contrast, for an infant learning her native language, it may be more important to acquire grammatical contexts. Grammatical contexts provide reliable cues for word meanings, and figuring out their use can be invaluable for further learning. A adjustable standard of knowledge could thus allow a learner to adapt learning in accordance to the situation at hand, carefully weighing learning time and resources. Consequently, a flexible standard may be one of the techniques a finite mind can use to best understand the external world.

In the sections that follow, we set forth future directions for research, in light of the results and discussion in this dissertation.[24]

5.1 Knowing grammar but not words
How grammar may be processed in the mind

The grammatical contexts we used in our studies were associative contexts, mimicking determiners in natural languages (e.g., *la+banane*). There is evidence that associative contexts may contribute to word meaning processing and that the features marked on associative contexts, such as grammatical gender, are evoked by isolated words (e.g., Dahan, Swingley, Tanenhaus, & Magnuson, 2000; and Wang, Chen, & Schiller, 2019, respectively). There may thus be a reciprocal relationship between associative contexts and content words (e.g., *la* and *banane*). Though there has been much research into how the brain encodes grammatical structure, few studies have investigated how the brain encodes associative context patterns (e.g.,

[22] A stabilized state may be one in which a plausible hypothesis is consider highly likely and thus tentatively adopted as the way the world likely is. A stabilized state is not a solidified state, which cannot easily be revised.
[23] Recent research shows that incorrect gendered articles do not impair processing accuracy in native speakers, and moreover do not elicit a P600 event-related potential associated with grammatical errors (Caffarra & Martin, 2019).
[24] In this general discussion, we interpret our results in respect to the main question of the dissertation and provide broad implications and directions for future research. Detailed discussions are presented in the discussion section of each chapter.

la + feminine noun; for a broad review on grammar processing in the brain, see Dehaene, Meyniel, Wacongne, Wang, & Pallier, 2015). The results from the series of studies in Chapter One and Two may provide some clues.

Our infant data revealed an early effect of the novel grammatical element. When infants heard '*ko+bamoule'* they looked more to the animate image than when they heard '*ka+bamoule'* from 140ms[25] after the onset of the determiner (*ko* or *ka*). This effect requires two elements: a fast oculomotor response and fast cognitive processing.

Looking-while-listening data is sometimes analysed from a short period of time after the onset of the target word (e.g., Zangl & Fernald, 2007). In infant research, this moment is centered around the gold standard of 367ms (e.g., Fernald, Swingley, & Pinto, 2001; Zangl & Fernald, 2007). It is considered to be the time necessary for an infant to deploy a saccade, and is derived from a study investigating gaze deployment in 3.5-month-old infants (Haith, Wentworth, & Canfield, 1993). It is important to first note that our analysis took into consideration all trials, and not only those in which participants had to make a saccade. Because our test paradigm was a two-choice task, participants could be looking at either one of the two images on the screen when they heard the determiner. Accordingly, half the participants would have already been looking at the target image just by chance (i.e., effect onset time of 0ms), and it is thus only the other half who would have to deploy a saccade. Infant gaze deployment time could thus be estimated at approximately 280ms (half the infants: 0ms; half the infants: 280ms; therefore general effect: 140ms). Additionally, it was possible to begin making a decision just a few milliseconds after the moment we call 'determiner onset', because we considered the determiner onset to be the moment of the consonant burst rather than the closure.[26] It may thus be realistic for older infants, like our 20-month-olds, to orient their gaze in less than 300ms, particularly in certain tasks, such as a two-option task (where the target image was either on the left or on the right, and where infants have the time to observe the images before hearing the auditory stimulus). This fast motor response, however, requires fast underlying cognitive processing.

Intuitively, we may expect adults to be faster at processing the novel determiners than infants, and yet, this is not what we observe. The effect for adults begins at 540ms. However, the looking-while-listening effect is carried, for the most part, by the Explicit Knowledge group of adults, and when we only look at that subset of adults, those who could explicitly state the rule governing determiner use, the effect begins a bit earlier, at 360ms.

There are two, not necessarily mutually exclusive, mechanistic explanations for the observed effect onset times. Onset times may reflect how the association between determiners and nouns is processed and may reflect what is being processed. Infants may have formed low-level associations between visual features and determiner use (e.g., eyes = *ko*). The images may have begun to evoke associative features, in a similar way that isolated words do for adults (Wang et al, 2019). Prior to hearing the determiner, the infants may have looked at each image,

[25] Aggregated data from the two studies. For an accurate measure of gaze orientation, we would need to run analysis specifically on the first look to target, on the subset of trials during which infants were looking at the foil at the moment of the determiner onset. Cluster-based permutation analyses, in contrast, reveal significant effect time-windows; the cluster construction method involves an arbitrary threshold (here, t=1.5), the value of which will change the exact boundaries of the time-cluster (lower values will give rise to larger clusters). Here, we use the onset of the time-cluster as an approximate index of the time-course of cognitive processes.

[26] Plosive consonants, such as 'k', 'p', 't', require an initial closure of the vocal apparatus before the quick 'bursty' sound we hear. As such, each plosive is preceded by a short silence. This silence could serve as a clue that the upcoming word may start with 'k', but it could not provide clues as to whether the word would be *ko* or *ka*.

remarked the essential feature or features (e.g., eyes), which then activated the grammatical association (e.g., *ko*). When the infant heard *ko* or *ka*, her attention could have quickly, and almost automatically, been drawn to the associated image. Under this interpretation infants would react to associations via attention attracted to sensory events automatically, and they would process a determiner-feature relationship, rather than a determiner-'noun meaning' relationship (e.g., *ko* therefore *bamoule* refers an animal). Infants may not have been constructing novel noun meanings *per se*; instead, they may have been constructing utterance-target object links. The link between determiner and target object (*ko* + thing with eyes) may be a precursor to a link between noun and target object (*bamoule* + this thing with eyes), and thus to noun meanings (e.g., *bamoule* means this creature). In other words, the association between determiners and target object (and indirectly the noun) may be accomplished largely bottom-up (the how) and may be a sensory feature relationship (the what). In contrast, adults who demonstrate a significant looking-while-listening effect are also those who can explicit state the rule governing determiner use. When an adult heard *ko* and *ka*, she may thus have oriented her attention by applying a rule: if *ko* then animal, if *ka* then object. Moreover, this rule may even have been an more complicated nested counterfactual: if, if *ko* then animal, then 'word' means 'that specific animal'.[27] Under this interpretation, adults would have oriented their attention voluntarily based on cognitive expectation, and they would process a determiner-'abstract meaning' relationship or a determiner-'counterfactual meaning' relationship. In other words, the associations between determiners and target object (and thus noun) may be accomplished largely top-down (the how) and may be a rule-based or meaning-based relationship (the what). Broadly, then, there may have been a difference in how participants oriented their gaze to the target image, automatic versus voluntary attention, and a difference in what the participants were processing, feature associations or meanings. These mechanistic explanations can account for the differences in effect onset time. Automatic attention orientation is faster than voluntary attention orientation (for an in depth discussion about different kinds of attention orientation, see Coull, Frith, Büchel, & Nobre, 2000; for an example of just how quick automatic attention can be, see Roye, Schröger, Jacobsen, & Gruber, 2010). A relationship between determiner and feature (e.g., *ko* + eyes) may be quicker to process than a relationship between determiner and meaning (e.g., *ko* + animal, so *bamoule* = animal). A determiner-meaning relationship could take longer because it is a multi-step reasoning process from determiner to meaning (e.g., determiner then concept then noun meaning) or could simply take longer because meanings themselves may be more cognitively complex than pure links between utterances and referents (e.g., *bamoule* is an animal versus *bamoule* is that thing). Therefore, the grammatical context may drive meaning acquisition in both infants and adults, via distinct cognitive mechanisms.

Interestingly, while the observed effect onset time for known (e.g., *ko*+rabbit) and novel nouns (*ko*+*bamoule*) was similar for adults (300ms and 360ms, respectively), it was not for infants (960ms and 140ms, respectively).[28] At first glance, we may have expected the effect onset time to be similar for known and novel nouns, or, even, later for novel nouns (which should be harder to process). In both kinds of trials, the determiner was sufficient to pick out the target, and the main difference was that in known trials participants could confirm their

[27] A complex counterfactual may best reflect some adults' reports during the post-experiment questionnaire. Some adults reported being frustrated while watching the training videos, because they knew that '*ko bamoule*' was an animal, but they could not know which (e.g., if, if '*ko'* is an animate, then '*bamoule'* is *one of those animals*).

[28] The effect onset times for known words in our study (e.g., *ko*+rabbit) are comparable with native language processing results (e.g., the+rabbit) reported in previous literature (e.g., adults: 300ms, Dahan et al, 2000; 18- and 21-month-old infants, 1000ms, Fernald, Swingley, & Pinto, 2001). This comparison serves only as an approximate reference point, because the protocols and analyses in previous studies differ from those in our study.

interpretation when they heard the noun (e.g., rabbit), but in novel trials they could not (e.g., *bamoule*). Yet, this interpretation assumes that a determiner evokes a category meaning (e.g., animate noun) and then a noun evokes a precise meaning (e.g., rabbit). In other words, the determiner just narrows down the possible meaning space of an upcoming word: it helps, but it is not necessary. Participants would still have been able to look at the rabbit if they had just heard 'rabbit'. However, meanings, if they spread over word boundaries, may be wholes, and not clear-cut aggregates. A determiner before a known word may thus contribute to processing differently than a determiner before a novel word. This distinction can arise because determiners have two associations: specific associations with upcoming nouns (e.g., *ko* precedes rabbit, *ko* precedes *bamoule*), and general associations with sensory or abstract features (e.g., animate noun). When the determiner precedes a known noun, it contributes to meaning processing: it offers its associated features to the meaning.[29] However, when a determiner precedes a novel noun, it may offer its associated features directly to processing. A learner could use these features to narrow down the possible meaning of the novel noun through some form of deduction (e.g., *ko*+eyes, therefore '*ko bamoule*' is the thing with the eyes), or these features could foster, what we call here, favourable learning situations. The features associated with a determiner, or any grammatical context, may draw closer together a word and its referent. When an infant hears a determiner, her attention may be automatically drawn to associated sensory features in her environment. Accordingly, her attention will be in the right kinds of places when she hears the novel word. She will be attending to nouns when she hears 'the *jigoulate'* and animals when she hears '*ko bamoule*'. Indirectly, she will begin forming associations between '*ko bamoule*' and the animal or animals she sees, nourishing the meaning of a once novel noun. A recent study found that attention drives decision making in adults, and the same may be true for infants (Smith & Krajbich, 2019). Therefore, infants may have perceived known word trials and novel word trials as fundamentally different. In the presence of pictures of known objects, the information provided by the determiner may have been put to use to access a meaning, a process which may inherently take some time, especially early in development; by contrast, in the presence of pictures of unknown objects, the information provided by the determiner may have remained in its raw form, eliciting mere feature associations. The difference between meaning and feature processing may be particularly pronounced in infants, because of heterogeneous cognitive maturation patterns early in development. While infants demonstrate longer decision-making latencies than adults well into their second year of life, by three months of age infant visual feature processing latencies are comparable to those of adults (Dehaene-Lambertz & Spelke, 2015). Our task was composed of two clear-cut situations: known words and unknown words. In the wild, so to speak, an infant may be less likely to know whether an upcoming word is one she knows or not, and there may thus be an early effect of grammatical context even before known words. Taken together, the results of our studies indicate that the learning processes underlying meaning acquisition in infants may be very basic and cognitively light, but highly potent.

In contrast to infants, adults may have interpreted both known and novel word trials as belonging to the same kind: a meaning access task. In known trials, adults will have accessed the meaning of the known noun, and in novel trials, they may have accessed a rule-based meaning, provided by the determiner. This explicit rule-based meaning, however, was not available to adults in the No Explicit Knowledge group. Yet, adults in the No Explicit

[29] In our study, we did not observe a behavioural effect of the determiner on processing of known word meanings for adults (anticipation trials). Our data do not allow us to pinpoint whether our behavioural measure was too noisy or whether there was just no determiner contribution to processing. Nevertheless, studies have found that native language determiners are used during online processing in toddlers as young as 28 months old (Van Heugten, Dahan, Johnson, & Christophe, 2012).

Knowledge group, who did not have access to an explicit rule, were able to correctly select, via pointing, the corresponding image when they heard novel words (e.g., *ko bamoule*). They must therefore have been able to glean some information about determiner use, implicitly, from their environment. However, this process may not have been the same as for infants, and more generally adults may not have been at the same point in the learning process as infants. It would be particularly interesting to investigate whether very early in learning, perhaps after 30s of exposure, adults demonstrate an early onset effect comparable to the one observed in infants. Early in acquisition, learners may form utterance-'sensory feature' links almost implicitly. Alternatively, this effect may be less pronounced in adults, because of a tendency to voluntarily orient attention and explicitly problem-solve: explicit problem solving has been shown to partially occult implicit acquisition (Fletcher et al, 2004). Moreover, explicit problem solving may have specifically drawn adults' attention to the novel noun forms, where meaning is thought to reside, focusing attention away from the determiner. In sum, adults may have interpreted novel noun meanings with any information at hand, be it an explicit rule or an intuition; importantly, however, there is evidence that adults may also have formed productive low-level links between determiners and sensory or abstract features.

We suggest, in light of our data, that the young mind may encode associative context patterns as context-noun associations (e.g., *ko bamoule*) and context-feature associations (e.g., *ko* + eyes), and that these associations can guide acquisition by creating 'favourable learning situations' (i.e., orienting attention to a subset of the sensory environment). An important consequence of this learning method is that the infant will be able to focus her cognitive resources on the pertinent subset of the total evidence, rather than equally diffuse her resources across the whole messy set. Future experimental and computational modeling research studies are needed test this hypothesis.

5.2 Learning grammar but not words
How to value grammar processing

Adults, in our studies, were able to acquire novel grammatical contexts, in a short span of time. However, the results in Chapter Three indicate that they may not be acquiring each linguistic element to the same degree. Adults were significantly better at acquiring nouns then determiners (see also, Arnon & Ramscar, 2012). Children, in contrast, were just as good at acquiring both. These results reflect long-standing intuitions about and investigations into second language acquisition in adulthood and in childhood: while both groups acquire noun meanings to high proficiency levels, adult learners acquire grammar to varying levels (e.g., DeKeyser, 2000; Johnson & Newport, 1989). The level to which an adult learner will acquire non-native grammar has been shown to depend on a number of factors, such as her native language (e.g. Paradis, 2018; Paradis, Tulpar, & Arppe, 2016) and metalinguistic capacity (e.g., DeKeyser, 2000). Often, however, lower proficiency levels after a lengthy learning period, are interpreted as difficulties in acquisition. Earlier, we proposed that different levels of proficiency could reflect a learning technique minds use to optimally allocate finite resources in their quest for knowledge. The learner's standard for knowledge may be adjustable for each element at a time t, as she advances in the process of acquisition. A learner may thus initially focus on vocabulary acquisition, as this element of language may seem the most important: uttering the correct vocabulary item will allow for basic communication (e.g., uttering just *banane* versus *la*). Further into the learning process, she may extend her focus to some grammatical features, as these elements may appear essential for communicating with precision (e.g., writing *bananes*

versus *banane*). A tractable standard of knowledge could engender observed discrepancies in proficiency: knowing that *banane* refers to bananas is vital, knowing that '*s*' marks the plural form important, and knowing that *banane* is a feminine noun almost trivial. One way to spur acquisition to a higher level of proficiency could be to render the ostensibly trivial, vital.

Linguistic elements begin to appear trivial when words encapsulate meaning, when each element is a stand-alone entity. Some elements can be overlooked because they are extraneous to a word, and thus to meaning. Yet, if meaning spreads beyond word boundaries, each linguistic element in an utterance is potentially important for meaning. Considering meanings as extending beyond word boundaries does not preclude that some elements may be more important for meaning than others; it simply entails that no element is trivial. In the study in Chapter Three, children may have considered the whole sentence to pick out the visual referent (e.g., '*Dol ko pirdale*' = rabbit), rather than the element unique to the visual referent (e.g., *pirdale*). Broadly, a word may be necessary for meaning, but not sufficient: a word may need to be present for a meaning to be about, let us say, rabbits, but it may be the larger sentential context that operationalizes meaning. In this dissertation, we have presented evidence that cognitive processing of meaning may extend beyond word boundaries (e.g., Lieberman, 1963; Dahan et al, 2000), and that the perception that words are isolable, self-standing items may not be universal (e.g., Morais & Kolinsky, 2002). The perception that a word is sufficient for meaning may arise when an individual learns to read, when she begins to explicitly isolate and manipulate words. Adult-learners may thus conflate necessity with sufficiency, perceiving some elements as vital and other as trivial. To equalize the extent to which all elements are acquired, perception would need to be re-centered: words would need to appear insufficient for meaning and contexts necessary.

In each study in this dissertation, adults were able to acquire a novel grammatical element not present in their native language, French. In Chapter One and Two, adults were able to extend a novel grammatical rule to interpret word meanings, even in the absence of explicit knowledge of the rule; in Chapter Three, they were able to acquire the associations between vocabulary and grammatical elements (e.g., ***ko*** *pirdale*, not ***ka*** *pirdale*), even though knowing the vocabulary item was sufficient to succeed at the learning task.[30] These results are consistent with two, non-exclusive, mechanistic explanations. First, the grammatical distinction animate-inanimate may be easy to acquire (perhaps because it is salient) or it may be easy to acquire for native French speakers (perhaps because French determiners are marked for a grammatical distinction, albeit a different one, feminine-masculine). Second, our learning tasks may have coaxed adults into paying attention to grammatical contexts. In Chapter One and Two, the novel grammatical element was introduced into the learners' native language. Native language processing of meanings has been shown to extend beyond word boundaries and to include grammatical elements (e.g., Dahan et al, 2000; Caffarra & Martin, 2019). Adults may thus have assimilated the novel element into an existent structure, one that places weight on grammatical contexts. In Chapter Three, adults were placed in an uncertain learning situation, a situation in which they could not discount the grammatical element from the get go. When adults first heard a sentence, such as, '*Dol ko pirdale',* and saw two images, a rabbit and a book, they could not know right away that *pirdale* was sufficient to correctly determine the referent. Even adults who had learned the direct translations of vocabulary items had to at least take into consideration the possibility that the preceding context may add or modify the meaning: '*Dol ko*' could specify grey rabbits or it could be a negation 'Not the'. Importantly, adults who had some supervision and those who did not had comparable performance. Therefore, even when

[30] In Chapter Three, we did not analyse whether adults without explicit knowledge of the rule governing the grammatical element's use were able to generalize.

attention is focused on vocabulary, it need not later block acquisition of associated elements.[31] An uncertain referential situation may have forced adults pay attention to the whole utterance, at least initially. Our data, however, do not allow us to disentangle the role the grammatical distinction may have had and the role the learning method may have had on acquisition. Further research is need to determine how these paradigms fare when the grammatical distinction is different (e.g., less salient) and when the participants' native language is different (e.g., English speakers learning a grammatical distinction on a determiner). Yet, the broader underlying ideas of harnessing native language knowledge and of instilling referential uncertainty may be viable ways in which to re-center the perception of meanings, and acquisition of new languages.

Intuitively, introducing a novel element into a learner's native language ought not have an effect on acquisition of another language. Yet, a learner's native language has been shown to influence how well she acquires grammar in a second language (Paradis, 2018). A learner whose native language and second language grammatical systems are vastly different (e.g., inflectional and lexical grammatical systems, e.g., an inflection, English: arrived vs. a lexical element, Mandarin Chinese: *dao le*) is less likely to attain full proficiency in second language grammar (Paradis, 2018). Moreover, knowledge has been shown to transfer, such that learning a complicated pattern is easier if one has learned a similar but simpler pattern first (Lany, Gómez, & Gerken, 2007). We propose that introducing a novel element into a learner's native language may fuel acquisition of a similar element in her second language. For instance, the participants in Chapter One and Two, who acquired an animate-inanimate distinction in their native language, may be more quick to acquire the animate-inanimate distinction on the verb 'to be' in Japanese (e.g., *neko ga iru* There is a cat vs. *isu ga aru* There is a chair). These non-native distinctions can be adapted to any existent structures in the native language. For example, a determiner distinction can be taught on other grammatical elements in languages that do not have determiners. A definite-indefinite determiner distinction (the vs. a) can be presented as a particle distinction in Japanese:

Existent Japanese subject-marking particle *ga:*
(a) *Neko ga iru* (Cat subject-marking particle there is; There is cat)

New particles indefinite-particle *se* and definite-particle *zu:*
(b) *Neko se iru* (Cat a (indefinite particle) there is; There is *a* cat)
(c) *Neko zu iru* (Cat the (definite particle) there is; There is *the* cat)

Adults may be more perceptive of a novel grammatical element in their native language, for a number of reasons: native language meaning processing may naturally stretch beyond word boundaries and include grammatical items; a surprising new item in an already known structure may spur acquisition (Gerken, Dawson, Chatila, & Tenenbaum, 2015); it may be easier to build-up knowledge on an existent foundation (e.g., knowledge that particles indicate the role the noun plays in the sentence, Lany et al, 2007). Importantly, if it is easier to acquire a novel grammatical element in one's native language, this knowledge may be able to transfer, fueling second-language acquisition. This learning method may appear convoluted, but it may force adult-learners to consider a grammatical element as necessary, learning it to a higher degree.

[31] It has been suggested that learning vocabulary does not block acquisition of grammatical elements that have a strict rule governing use (e.g., Siegelman & Arnon, 2015). It would be interesting to see how well the grammatical element is acquired in our paradigm when the rule to be learned is less regular. It could mimic seemingly arbitrary Mandarin Chinese classifiers (words used in the presence of a numeral), such as *tiao* which is used with fish, pants and noodles (e.g., '*san **tiao** yu'* three **classifier** fish).

Referential uncertainty could too incite adults to consider grammatical elements as necessary. Broadly, uncertainty has been shown to promote exploration (Boldt, Blundell, & De Martino, 2019). Therefore, an uncertain learning task like a cross-situational learning paradigm (Yu & Smith, 2007) may in and of itself instill a more extensive exploration of the linguistic and visual elements. In Chapter Three, uncertainty may have coaxed adults into considering the grammatical elements. Yet, until the test phase, vocabulary was sufficient to succeed at the task, to determine the visual referent. The task could however be modified such that vocabulary is not sufficient for reference, or broadly meaning. Some trials could present novel vocabulary and new referents; on these trials the only way to solve the task would be to use the grammatical elements (e.g., a picture of a tiger and a spoon, with a prompting sentence '*Dol ko nuve*'). These trials would only appear once (i.e., the learner would not see a tiger or a spoon or hear '*nuve'* again). Participants would quickly realize that to pick the correct referent, they could not rely on waiting to hear the vocabulary item again in a different visual context: they would have to use the evidence at hand (namely the grammatical element). By making grammar necessary for meaning and words insufficient, performance may be equalized.

We propose that the divergent levels of proficiency attained on non-native vocabulary and grammar may reflect an ostensibly efficient learning strategy: focusing on what is most important. Yet, when words are thought to encapsulate meaning, some linguistic elements may appear trivial. These elements may be set aside during learning. To re-center perception of words as necessary but not sufficient for meaning, words could to be presented as insufficient for meaning and grammatical contexts as necessary. These situations could be enacted by harnessing native language processing and creating uncertain referential situations. Further research is needed to investigate whether and how a re-centering of meaning perception could even out acquisition of linguistic elements.

5.3 Meaning in the mind

In this dissertation, we have argued that the gap between how we picture meaning to be and how meaning may actually be processed in the mind can provide clues as to how an ostensibly insufficient evidence set gives rise to word-meaning links. Our argument is based on two premises: meaning may extend beyond word boundaries and the insufficiency of evidence may be a general problem for cognition. We used these two premises to demonstrate how the '*gavagai* problem' may only appear to be a problem from the standpoint of a competent speaker, but not one from the standpoint of a learner.

If meaning is not neatly bundled into each word, but stretches across word boundaries, a learner could rely on the sentential context to fix meaning. In other words, word 'meanings' could be vague, and they likely are so. Word 'meanings' need to have enough precision to pick out only one *kind* of referent: 'mouse' needs to be able to refer to field mice, wood mice, Jerry and Mickey Mouse, but not hedgehogs. Thus, word 'meanings' may be akin to perceptual kinds (Shepard, 1987), and may as such be defined by an internal metric of similarity: 'mouse' refers to anything that is 'mousey'. Linguistic contexts, then, could serve to delimit the scope of these vague meanings. 'The mouse' would refer to a specific mouse; 'The mouse is swift' would refer to a specific mouse that has the attribute of being agile; and so on. In other words, the vagueness of word 'meanings' would be precisely what allows words to be so productive. However, when an individual learns to read, she will begin to extract and manipulate words explicitly, to see words as isolable, self-standing entities. As words become isolable and self-standing, so does meaning: meanings that spread across word boundaries are organized and neatly attributed to

words. An adult may thus perceive meanings from the point of view of a competent and literate speaker, but this perception may not be shared with not-yet-competent speakers or pre-literate speakers. Perhaps tellingly, the learner faced with the '*gavagai* problem' was a linguist, an adult trying to translate a hitherto unknown language (Quine, 1960). The learner was thus an already competent speaker of a language, and a literate one at that. If individual word 'meanings' are vague, but utterance meanings precise, the combination of words and sentential contexts could attenuate the '*gavagai* problem'.

The sentential context may narrow down a possible meaning, but even this more precise meaning will be consistent with multiple interpretations. *Gavagai*, uttered in plethora of linguistic and visual contexts, could be interpreted as 'rabbit' or 'a set of undetached rabbit parts'. A learner who knows that *ga* is a prefix indicating one lone thing, will be able to interpret *gavagai* as one rabbit, rather than just rabbit. Yet, knowing that the utterance is about one lone thing does not inform the learner as to whether *gavagai* refers to one rabbit or one set of undetached rabbit parts. The indeterminacy of reference thus applies to both words and sentences (Quine, 1960). Though here we refer to the Quine's *gavagai* example as the '*gavagai* problem', Quine did not see it as problem *per se*. For Quine, as long as elicited behaviour is the same, the number of possible interpretations of meaning is irrelevant. Our view here is slightly different. We acknowledge the problematic nature of the indeterminacy of meaning, but propose that the mind has a solution and that this solution is one that it applies broadly to make sense of the external world.

Minds have to make sense of a vast, uncertain external world, but they only have access to (and can only process) a subset of the total evidence present in the world. Minds thus have to build up knowledge based on the evidence set at hand. In other words, a mind will have to take into account what the evidence set attests (by way of an analysis of the evidence) and how likely it is that this evidence set represents the way the world is (by way of a simulation of the evidence on a world scale). Yet, the mind will not be able to know with certainty that her interpretation of the world reflects how the world is; it will not be able to check built up knowledge against an answer sheet. Knowledge therefore will not be indubitable nor infallible. Here, we could perhaps defend the fallibility of knowledge, but positing that as long as this kind of knowledge is good enough for minds, knowledge need not be indubitable (see James, 1897); or, we could perhaps argue that fallible knowledge is actually better for mind-world interaction than indubitable knowledge, because it allows for dynamic interaction (see Peirce, 1905). However, to defend fallibly is to not explain how the mind interacts with the world when its vision of the world could be flawed. We, therefore, propose that the mind could reduce the gap between its interpretation of the world and the way the world is by *aiming* for indubitability.

A mind could aim for indubitability broadly by overcoming its limitations. It could do so by efficiently allocating its finite resources and by expanding its resource capacity. In Chapter Two, we suggested that the urge to know could modulate the knowledge threshold, lowering it when an individual needs to act right away, and elevating it when an individual has time to act. The urge to know would also elevate the threshold when an individual needs to act with precision. For instance, an individual may only need approximate knowledge of what mice are like, and may thus have a low knowledge threshold. However, if this individual begins a wildlife photography hobby, she may need to differentiate voles from mice, and may thus need a higher knowledge threshold. Broadly, the urge to know in tandem with a mutable knowledge standard can allow for knowledge to tend to indubitability, when need be. Importantly, a mutable standard could allow an individual to oscillate between learning and knowing, between actively accumulating and analysing evidence and possessing stabilized knowledge. The

individual could, as such, allocate finite learning resources efficiently, aiming for indubitability when necessary.

A mind could also attempt to increase its resource capacity –accessing and processing more evidence. It could do so by combining the finite capacities of multiple minds. In Chapter Three, we proposed that a mind could access other minds' pre-packaged knowledge-states by learning via explicit teaching. However, other minds' knowledge could be also be transmitted largely implicitly. For instance, John McDowell proposed that language may be a receptacle of knowledge (McDowell, 1996). Language is defined and redefined to reflect knowledge by innumerable minds at any one moment, but also across generations (McDowell, 1996). The word 'mouse', for example, would mean anything 'mousy' as determined by all the minds, and thus by all the evidence to which they had access. Language could, as such, allow an individual mind to harness the processing power of innumerable minds, pushing knowledge up toward indubitability. Moreover, other minds' pre-packaged knowledge states may be inherently easier to make sense of. A phenomenon that has been observed in the literature is that young children appear to learn better in communicative situations (e.g., Ferguson & Waxman, 2016). When 6-month-old infants are presented with associations between beeps and visual stimuli, they do not learn them (in the span of a lab visit); in contrast, when they are presented with the same associations after hearing that beeps are used to communicate, they learn the associations (Ferguson & Waxman, 2016). This phenomenon has be attributed to a desire to understand communication. However, it could also reflect a propensity to take advantage of that which is easily learnable to scaffold acquisition of how the world is. In other words, infants may have noticed that language may be less confusing than all the blooming and buzzing in the external world. Therefore, the mind could augment finite resources by harnessing the power of other minds, and in so doing, get a step closer to indubitable knowledge.

The mind could thus tackle both the fallibility of knowledge and the indeterminacy of meaning, by reducing its limitations. It could use a mutable standard of knowledge in tandem with the urge to know to make the most of its own resources; and it could locate rich evidence sets, from which extracting reliable knowledge requires few resources. In all likelihood then utterance meanings are but a little shy of determinacy. With further research we may discover more ingenious methods minds use to learn in the 'blooming, buzzing confusion' (James, 1890) that is the external world.

***

A threshold of knowledge likely governs the fine line between learning and knowing. If this threshold is mutable, as we have proposed, individuals could learn then know, know then learn, honing the precision with which they act in the world. However, an adjustable knowledge threshold could add noise to behavioural observations: two individuals, with the same extracted information and the same level of certainty that that information is correct, may not generalize said information. Behaviour may thus appear noisy, but this noise may be able to reveal how minds interact with the world. We hope that future research will take more and more interest in null effects, perhaps to the same level as significant effects, and that one day we may discover that all that noise, well, it was just a cacophony of harmonies.

***

> Jane turned back to the car. Leaning against the front door, she perused the photos she had managed to take before the bamoule disappeared into the scenery. On the

photos, the bamoule was like all the others she had seen in school textbook pictures. She swiped through the photos, watching the bamoule pirdale and unpirdale, to a saccadic rhythm. It almost made her want to pirdale herself. Looking up from the camera, she squinted her gaze into the distance. The landscape, coated by an opaque curtain of heat, stood silent. I'll stay just a bit longer, she thought. You never know. Perhaps she would have a second fortuitous sighting of a bamoule. Better yet, a jigoulate nuving. Now wouldn't that be something![32]

32 Gloss for novel words: *bamoule* (armadillo), *to pirdale* (to roll into a ball), *jigoulate* (cactus), *to nuve* (scientific term: to have flowers blossom all over)

References

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122*(3), 292-305.

Bogacz, R. (2007). Optimal decision-making theories: linking neurobiology with behaviour. *Trends in Cognitive Sciences*, *11*(3), 118-125.

Boldt, A., Blundell, C., & De Martino, B. (2019). Confidence modulates exploration and exploitation in value-based learning. *Neuroscience of Consciousness*, *2019*(1), niz004.

Caffarra, S., & Martin, C. D. (2019). Not all errors are the same: ERP sensitivity to error typicality in foreign accented speech perception. *Cortex*, *116*, 308-320.

Cassey, T. C., Evens, D. R., Bogacz, R., Marshall, J. A., & Ludwig, C. J. (2013). Adaptive sampling of information in perceptual decision-making. *PloS One*, *8*(11), e78993.

Coull, J. T., Frith, C. D., Büchel, C., & Nobre, A. C. (2000). Orienting attention in time: behavioural and neuroanatomical distinction between exogenous and endogenous shifts. *Neuropsychologia*, *38*(6), 808-819.

Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. S. (2000). Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language*, *42*(4), 465-480.

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., & Pallier, C. (2015). The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. *Neuron*, *88*(1), 2-19.

Dehaene-Lambertz, G., & Spelke, E. S. (2015). The infancy of the human brain. *Neuron*, *88*(1), 93-109.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*(4), 499-533.

Fletcher, P. C., Zafiris, O., Frith, C. D., Honey, R. A. E., Corlett, P. R., Zilles, K., & Fink, G. R. (2004). On the benefits of not trying: Brain activity and connectivity reflecting the interactions of explicit and implicit sequence learning. *Cerebral Cortex*, *15*(7), 1002-1015.

Ferguson, B., & Waxman, S. R. (2016). What the [beep]? Six-month-olds link novel communicative signals to meaning. *Cognition*, *146*, 185-189.

Fernald, A., Swingley, D., & Pinto, J. P. (2001). When half a word is enough: Infants can recognize spoken words using partial phonetic information. *Child Development*, *72*(4), 1003-1015.

Gerken, L., Dawson, C., Chatila, R., & Tenenbaum, J. (2015). Surprise! Infants consider possible bases of generalization for a single input example. *Developmental Science*, *18*(1), 80-89.

Haith, M. M., Wentworth, N., & Canfield, R. L. (1993). The formation of expectations in early infancy. *Advances in Infancy Research*, *8,* 251-297.

James, W. (1890). *The Principles of Psychology* (Vol. 1). Cambridge, MA: Harvard University Press.

James, W. (1897). *The Will to Believe and Other Essays in Popular Philosophy*. Cambridge, MA and London: Harvard University Press.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*(1), 60-99.

Lany, J., Gómez, R. L., & Gerken, L. A. (2007). The role of prior experience in language acquisition. *Cognitive Science*, *31*(3), 481-507.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, *6*(3), 172-187.

McDowell, J. (1996). *Mind and world*. Cambridge, Massachusetts : Harvard University Press.

Morais, J., & Kolinsky, R. (2002). Literacy effects on language and cognition. In L. Bäckman & C. von Hofsten (Eds.), *Psychology at the turn of the millennium, Vol. 1. Cognitive, biological, and health perspectives* (pp. 507-530). Hove, England: Psychology Press/Taylor & Francis (UK).

Paradis, J. (2018, November). English L2 acquisition from childhood to adulthood. At the *43rd Boston University Conference on Language Development*. Association for Psychological Science, Paris.

Paradis, J., Tulpar, Y., & Arppe, A. (2016). Chinese L1 children's English L2 verb morphology over time: Individual variation in long-term outcomes. *Journal of Child Language*, *43*(3), 553-580.

Peirce, C. S. (1905). What Pragmatism is. *The Monist*, *15*(2), 161–181.

Quine, W. V. O. (1960). *Word and object*. Cambridge, Massachusetts: MIT press.

Roye, A., Schröger, E., Jacobsen, T., & Gruber, T. (2010). Is my mobile ringing? Evidence for rapid processing of a personally significant sound in humans. *Journal of Neuroscience*, *30*(21), 7310-7313.

Siegelman, N., & Arnon, I. (2015). The advantage of starting big: Learning from unsegmented input facilitates mastery of grammatical gender in an artificial language. *Journal of Memory and Language*, *85*, 60-75.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323.

Smith, S. M., & Krajbich, I. (2019). Gaze amplifies value in decision making. *Psychological Science*, *30*(1), 116-128.

Van Heugten, M., Dahan, D., Johnson, E. K., & Christophe, A. (2012). Accommodating syntactic violations during online speech perception. In *Poster presented at the 25th Annual CUNY Conference on Human Sentence Processing, New York, NY*.

Wang, M., Chen, Y., & Schiller, N. O. (2019). Lexico-syntactic features are activated but not selected in bare noun production: Electrophysiological evidence from overt picture naming. *Cortex*, *116*, 294-307.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414-420.

Zangl, R., & Fernald, A. (2007). Increasing flexibility in children's online processing of grammatical and nonce determiners in fluent speech. *Language Learning and Development*, *3*(3), 199-231.

On croit que ce n'est pas commencé

et c'est commencé.

On croit qu'on s'achemine vers une solution,

on se retourne,

et voilà qu'elle est derrière soi.

Marguerite Duras, Le square[33]

[33] Duras, M. (1987). *Le square* (Ed. nrf). Paris: Gallimard. (p 49).

RÉSUMÉ

Apprendre, c'est extraire et condenser de l'information pertinente à partir d'un certain nombre d'indices concordants. Pourtant, les indices auxquels un apprenant a, ou pourrait avoir, accès, peuvent sembler insuffisants pour ce à quoi il veut parvenir. Le caractère partiel ou insuffisant des indices disponibles est souvent évoqué dans le contexte de l'acquisition du langage : le sens des mots et la grammaire que les individus maîtrisent semblent nécessiter davantage d'information que celle présente dans leur environnement. Si une grande partie de la recherche se focalise sur l'identification de sources d'information négligées ou sous-estimées jusqu'alors, ici nous déplaçons le point de vue pour explorer plutôt les connaissances que ces sources d'information doivent permettre de construire. Nous proposons dans ce travail que les informations disponibles restent souvent insuffisantes pour atteindre le niveau de connaissance qu'un locuteur compétent pense avoir; toutefois, la même information peut être suffisante pour qu'un apprenant extraie l'information nécessaire au traitement cognitif du langage. Plus généralement, il pourrait y avoir un décalage entre ce que l'on pense apprendre et ce que l'on apprend véritablement.

Dans l'introduction de ce travail, nous commençons par une présentation des études qui suggèrent l'existence d'un décalage entre notre intuition de ce qu'est le sens des mots, et la manière dont le sens des mots est effectivement traité par le cerveau. Nous explorons ce décalage dans le contexte plus général de la manière dont les individus, qui ne disposent que d'informations limitées, tentent de comprendre le monde. Ensuite, dans le premier chapitre, nous examinons comment une faible quantité d'information peut malgré tout favoriser l'acquisition de la grammaire, en utilisant un paradigme écologique. Nos résultats démontrent que la connaissance d'une petite poignée de mots peut générer un cercle vertueux dans l'acquisition de la grammaire et du vocabulaire chez le bébé. Puis, dans le deuxième chapitre, nous tâchons de découvrir si cet ensemble réduit d'informations peut suffire à amorcer l'acquisition d'un savoir productif qui permettra ensuite de généraliser cette connaissance à des situations nouvelles, à travers les âges : du bébé à l'adulte en passant par l'enfant d'âge scolaire. Nos données suggèrent que les bébés et les adultes généralisent les caractéristiques d'un nouvel élément grammatical afin de comprendre de nouveaux mots, mais que les enfants d'âge scolaire ne généralisent pas. Nous expliquons nos résultats dans le contexte de la recherche existante sur la généralisation, en soulignant la possibilité que ce qui est considéré comme savoir est peut-être défini par rapport à un seuil, plutôt que par rapport à une définition idéale objective. Finalement, dans le troisième chapitre, nous étudions si l'information qui donne l'impression de refléter directement l'état des connaissances de quelqu'un d'autre (i.e., entendre une traduction directe : 'Bamoule'*, ça veut dire 'chat') est véritablement plus utile, ou seulement plus attractive. Nos résultats suggèrent que ce type d'information 'clé-en-main' augmente systématiquement la confiance de l'apprenant, mais a des effets variables sur la performance objective. A travers trois chapitres et dix expériences, nous proposons une série de principes qui définissent la connaissance : (1) une petite quantité d'information peut mener loin, (2) combien d'information suffit semble dépendre d'un seuil adaptable, et (3) le cerveau paraît être avide de certitude. Nous suggérerons que l'ensemble des informations à la disposition d'un individu peut être suffisant pour générer des connaissances, mais pas forcément le type de connaissances que l'on a l'intuition d'avoir. Ainsi, pour mieux comprendre comment on apprend, nous devons étudier ce que signifie vraiment 'savoir' pour l'apprenant.*

MOTS CLÉS

Acquisition du langage, généralisation, métacognition, acquisition du vocabulaire, apprentissage trans-situationnel, acquisition des langues artificielles

ABSTRACT

To learn is to extract and distill pertinent information from a set of evidence. Yet, the evidence a learner has, or could have, at hand may seem insufficient for what she is aiming to acquire. The insufficiency of the evidence is often evoked in respect to language acquisition: the word meanings and grammar that individuals know appear to require more evidence than that to which they have access in their environments. While the brunt of research has focused on identifying overlooked sources of evidence and widening the evidence set, here we switch gears and probe what exactly it is that the evidence set needs to support. We propose that the evidence a learner has may be insufficient to provide her with the knowledge competent language users think they have; however, the very same evidence may be sufficient to provide a learner with the information needed for cognitive processing of language. Very broadly, there may be a gap between what we think we acquire and what we really acquire.

In the introductory section of this dissertation, we begin by presenting evidence of a gap between what we feel words mean and how meanings are processed in the mind. We frame this gap in the broader context of how minds, with finite access to evidence, make sense the external world. Then, in the first chapter, we investigate how comparably little evidence can fuel acquisition of grammar in an ecologically valid setting. Our results reveal that just a handful of words can spur a virtuous cycle of grammar and vocabulary acquisition in infants. Next, in the second chapter, we examine whether the same set of evidence gives rise to productive knowledge or generalization (i.e., the capacity to use prior knowledge to interpret novel situations) across development, from infancy into childhood and through to adulthood. Our data show that infants and adults generalize a novel grammatical context to new words, but pre-school children do not. We interpret our results within the extant literature, pointing to the live possibility that what counts as knowledge may depend on where an individual places her 'knowledge threshold' rather than an immutable ideal. Finally, in the third chapter, we probe whether evidence that reflects a knowledge-state (e.g., explicitly hearing a direct translation: '*Bamoule'* means 'cat') is inherently more informative, or merely more appealing, for a learner. Our results demonstrate that pre-packaged knowledge-state evidence boosts confidence, but has variable effects on performance. Across three chapters and ten experiments, we build up a set of fundamental features about what it is 'to know': (1) a little evidence can go a long way, (2) how much evidence is considered to be enough may depend on a modifiable threshold, and (3) the mind may crave certainty. We advance the conclusion that the set of evidence available to an individual can be sufficient to foster knowledge. It may just not be sufficient to foster the kind of knowledge we think we have. Therefore, to better understand the way we learn, we need to investigate what is 'to know' from the point of view of the learner.

KEYWORDS

Language acquisition, generalization, metacognition, vocabulary acquisition, cross-situational learning, artificial language acquisition